Extending Targeted Function Balancing to Models without Linear Representations

---

A Thesis
Presented to
The Division of Mathematical and Natural Sciences
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

He Bai

May 2024

Approved for the Division
(Mathematics - Statistics)

Leonard Wainstein

Acknowledgements

I would like to thank my thesis advisor, Leonard Wainstein, who,

1. created the original TFB, a fantastic causal inference method that I am honored to generalize;

2. instead of increasing my level of thesis-induced insanity as advisors are known to do, actually drastically decreased it by being invariably kind, patient, and wise;

3. dutifully pointed out that including the word "Squidward" (Hillenburg et al., 1999-Present) in my thesis title is a risky move that may result in my perpetual disgrace and ostracization in academia.

I would like to thank the UCLA Practical Causal Inference reading group for giving me advice that I would not have been able to get anywhere else.

I would like to thank my partner, Leo Latimer, for discovering the secret door in the basement of Eliot Hall and making everything better simply by existing.

I would like to thank my favorite library circulation desk worker, Leandra Bruggink, for calling me "self-contained" and kindly offering her unofficial friendship before, while, and (I sincerely hope) long after the writing of this thesis.

I would like to thank Eli Franz for also calling me "self-contained", and, in his words, "being a weird little dude who lived in your house for a semester".

I would also like to thank all the other persons collectively known as "the Oreos" who were not present when I wrote my acknowledgements during spring break.

Finally, I would like to thank Kenai Burton-Heckman for writing a thesis on TFB that inspired me on what to include in mine.

List of Abbreviations

ACIC	Atlantic Causal Inference Conference
ATC	average treatment effect on the control
ATE	average treatment effect
ATT	average treatment effect on the treated
BART	Bayesian additive regression trees
CDF	cumulative distribution function
CEF	conditional expectation function
DGP	data generating process
DIM	difference in means
EBAL	entropy balancing
GTFB	generalized targeted function balancing
GLM	generalized linear model
IPW	inverse propensity weighting
KBAL	kernel balancing
MSE	mean squared error
NSW	National Supported Work
OLS	ordinary least squares
RCT	randomized control trial
RF	random forest
RMSE	root mean squared error
RSS	residual sum of squares
SATC	sample average treatment effect on the control
SATE	sample average treatment effect
SATT	sample average treatment effect on the treated
SE	standard error
SD	standard deviation
TFB	targeted function balancing
WDIM	weighted difference in means

List of Symbols

n	total number of observations
n_t	number of treated observations
n_c	number of control observations
D	treatment assignment
X	covariates
Y	observed outcome
$Y(0)$	potential outcome without treatment
$Y(1)$	potential outcome with treatment
$f_0(X)$	CEF of potential outcome without treatment
$\epsilon(0)$	error of potential outcome without treatment
w	weights
$\hat{\tau}_{\text{wdim}}$	weighted difference-in-means estimator
$\text{imbal}(w, f_0(X), D)$	imbalance in $f_0(X)$
$\text{imbal}(w, X, D)$	imbalance in X
$V_{\hat{f}_0}$	variance of estimated potential outcome without treatment
$Q_q(\cdot)$	qth quantile of $\cdot$
$\sigma(0)$	standard error of $\epsilon(0)$
β	coefficient in OLS and GLM
g	link function in GLM
g'	derivative of g

Table of Contents

List of Tables

List of Figures

Abstract

Targeted Function Balancing (TFB), proposed by Wainstein (2022), is a covariate balancing weight method for estimating the causal effect of a binary treatment on an outcome in the setting of observational studies. TFB linearly regresses an outcome on the covariates and balances functions that are probabilistically close to the linear model. This thesis proposes a generalized TFB (GTFB), which does not require models for the outcome to have linear representations. GTFB preserves TFB's philosophy of strategically leaving imbalance in the covariates to achieve higher efficiency, and seeks balance on functions of the covariates determined by the model specification to safeguard against modeling error. In addition, GTFB allows the use of any predictive model for the outcome, widening the range of data settings where the philosophy of TFB is applicable. This thesis demonstrates that GTFB performs well on simulated data in conjunction with ordinary least squares, probit regression, Bayesian additive regression trees, and boosting. This thesis also compares GTFB to other commonly used causal inference methods in two applications: the 2016 Atlantic Causal Inference Conference (ACIC) competition and the National Supported Work (NSW) dataset. The ACIC competition datasets, with real covariates but simulated treatment and outcomes, are designed to assess the performance of causal inference methods. The NSW dataset originates from a study on the effect of a subsidized labor training program on workers' income. The dataset includes a randomized control trial that provides an unbiased estimate for the causal effect and an observational study where causal inference methods can be tested.

Dedication

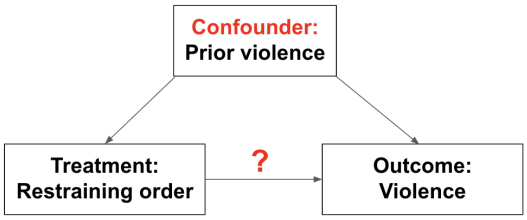
To my parents, who will probably show this thesis to many people but, very understandably, never attempt to read it themselves.

Introduction

An important concern in all scientific inquiry is the distinction between correlation and causation. Causal inference, a subfield of statistics, formalizes the search of causation from correlation.

Scientific studies aiming to estimate causal effects include two key elements: the treatment, which is the intervention administered, and the outcome, which is the quantity that potentially responds to the treatment. When treatment is binary, observations fall into one of two groups: the treated, which receives treatment, and the control, which does not. As an example, consider a study on the effect of restraining orders on the rate of violence experienced by survivors of stalking by former intimate partners. In this setting, the administration of retraining orders is the treatment, and the rate of violence is the outcome. Survivors of stalking who obtained restraining orders constitute the treated group, and those who did not constitute the control group.

Figure 1: Confounding



Many studies in the social sciences fall under the category of observational studies, where the treatment assignment mechanism is not random or known. This setting complicates the estimation of causal effects. Again consider the example on restraining orders. Suppose we observe that survivors of stalking who obtained restraining orders experience the same rate of violence as those who did not. It may be tempting to conclude that restraining orders fail to reduce the rate of violence, but as demonstrated in Figure 1, we may reasonably expect that survivors who have experienced violence before the end of their relationships may be more likely to (1) obtain re-

straining orders and (2) experience higher rates of violence after the end of their relationships. Therefore, upon observing equal rates of violence in survivors who obtained restraining orders and those who did not, we cannot conclude from this lack of difference an absence of causal effect, since it is plausible that restraining orders decrease the rate of violence, but survivors who obtain them are at a higher risk of violence due to prior instances of violence. We refer to prior violence as a confounding variable, which means it has an effect on both the treatment and the outcome. Confounding may lead us to spuriously conclude causal relationships or fail to observe actual causal relationships. We refer to observed potential confounding variables as covariates. If all confounders are observed, unbiased estimation of the causal effect is possible if we adjust for the potential confounding effect of the covariates. Weighting is one approach to adjusting for this effect.

Weighting methods incorporate a weight, which determines the importance of each observation, in the calculation of the causal effect. Since the impact of confounding on the causal estimate is driven by dissimilar distributions of the confounding variable in the treated and control groups, weights broadly seek to make the treated and control groups more similar in terms of some characteristic of the covariates. For example, inverse propensity score weights equate the distributions of the covariates in the treated and control groups. Balancing weights, however, equate the means of some function of the covariates. By seeking balance on some quantity that is assumed to contribute to bias in the causal estimate, balancing weights may lessen the impact of confounding. In our example about restraining orders, recall that we may expect survivors who obtained restraining orders experience a higher rate of prior violence than those who did not. Assuming that we want to make the control group more similar to the treated group, it is reasonable to give higher weight to survivors who did not obtain restraining orders but experienced prior violence, and lower weight to survivors who did not obtain restraining orders and did not experience prior violence. This weighting scheme would make the weighted rates of violence more similar in the treated and control groups. Thus, the weighted difference in rates of violence can no longer be explained by prior violence.

This thesis seeks to contribute to the methodological literature on balancing weights. Specifically, it extends Targeted Function Balancing (TFB), a novel covariate balancing weight method developed by Wainstein (2022), so that its philosophy can be used in conjunction with a wider range of modeling tools. TFB regresses the outcome on the covariates using a linear model, and finds weights that balance (1) the predicted outcomes according to the model and (2) the covariates. TFB has been

demonstrated to perform well on simulated and real data, but the requirement of linearity for the outcome model is restrictive. The extension proposed by this thesis, generalized targeted function balancing (GTFB), extends TFB to allow the use of models that are non-linear. For example, GTFB allows the use of generalized linear models and tree-based machine learning methods. This extension (1) enables the application of the TFB philosophy to a broader class of settings where the outcome may not be a linear function of the covariates, and (2) potentially improves the causal estimate by allowing more powerful models.

Chapter 1 of this thesis introduces the framework of causal inference in which GTFB operates. Chapter 2 mathematically formulates GTFB and uses concrete settings to explain the formulation. Chapter 3 demonstrates the performance of GTFB on simulated datasets. Chapter 4 applies GTFB to existing datasets to evaluate the performance of GTFB in more naturalistic settings.

Chapter 1

Background

This section formally introduces the notation and assumptions used in this thesis, the estimand of interest and its corresponding estimators, and the framework of weighting.

1.1 Notation and Setting

Experiments in the empirical sciences that involve a comparison between treated and control groups fall roughly into one of two categories: randomized control trials (RCTs) and observational studies. In RCTs, treatment assignment is random; in particular, treatment status is independent of the covariates. Randomization equates the distributions of covariates in the treated and control groups in expectation, and thus eliminates the effect that different covariate distributions may have on the outcome. Therefore, a simple difference in the mean outcomes may be a satisfactory causal estimate in the RCT setting. In contrast, in the more common paradigm of observational studies, treatment assignment is not random or known. For instance, the example on restraining orders is an observational study. In the case of observational studies, different covariate distributions in the treated and control groups may lead to different outcomes, complicating the estimation of causal effects. This thesis is concerned with the estimation of causal effects in the setting of observational studies.

Throughout this thesis, we assume that treatment is binary. Let $i \in \{1, ..., n\}$ index the units of observation. Let n_c denote the number of control units, and $n_t = n - n_c$ denote the number of treated units. Order the units of observation so that the first n_c units are control and the remaining n_t units are treated. Let $D = [D_1 \ \dots \ D_n]^\top$ denote the vector of treatment indicators, which means that $D_i = 0$ for $i \in \{1, \ \dots, \ n_c\}$ and $D_i = 1$ for $i \in \{n_{c+1}, \dots, n\}$. In the example on restraining orders, $D_i = 1$ indicates that the ith survivor in the sample obtained a

restraining order, and $D_i = 0$ indicates that they did not.

Let X_i denote the covariate vector for each unit i, and let X denote the matrix of the X_i's:

$$X_i = \begin{bmatrix} X_i^{(1)} \\ \vdots \\ X_i^{(p)} \end{bmatrix} \in \mathbb{R}^p, \quad X = \begin{bmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{bmatrix} \in \mathbb{R}^{n \times p},$$

where p is the number of observed covariates. As an example, X_i in the study on restraining orders could include whether survivor i experienced prior violence, the criminal record of the person who stalked them, and the survivor's socio-econimic status.

Let $Y = [Y_1, \ \dots \ Y_n]^\top \in \mathbb{R}^n$ denote the vector of observed outcomes. In the study on restraining orders, $Y_i = 1$ indicates that survivor i experienced violence during the study period, and $Y_i = 0$ indicates that they did not.

1.1.1 Potential Outcomes

The potential outcomes framework (Splawa-Neyman et al., 1990; Rubin, 1974) conceptualizes the causal effect by comparing the observed and counterfactual outcomes of each observation. Assuming binary treatment, each observation i has two potential outcomes: one with treatment, which we denote with $Y_i(1)$, and one without treatment, which we denote with $Y_i(0)$. Denote the observed outcome with Y_i. For each observation, $Y_i(D_i)$ is the observed outcome, and $Y_i(1 - D_i)$ is the unobserved, counterfactual outcome. The potential outcomes framework defines causal effect of the treatment for observation i as $Y_i(1) - Y_i(0)$. Again consider the study on the effect of restraining orders on violence, using Table 1.1 as a simple example. First consider Survivor 1. $D_1 = 1$ indicates that Survivor 1 obtained a restraining order, and $Y_1 = 0$ indicates that they did not experience violence. Note that $Y_1(0) = 0$, which means Survivor 1 would have experienced violence if they did not receive a restraining order. Thus, the restraining order was effective in preventing violence for Survivor 1. Now consider Survivor 2. $D_2 = 0$ indicates that Survivor 2 did not receive a restraining order, and $Y_2 = 0$ indicates that they did not experience violence. Since $Y_2(1) = 1$, a restraining order would have caused violence for Survivor 2. Lastly, we consider Survivor 3. $D_3 = 1$ indicates that Survivor 3 obtained a restraining order, and $Y_3 = 1$ indicates that they experienced violence despite having received a restraining order. Thus, in the case of Survivor 3, the restraining order failed to prevent violence.

Since treatment is either administered or not administered in real-life settings, an

Survivor index (i)	**Restraining order** (D)	**Experienced violence** (Y)		
		$Y(0)$	$Y(1)$	Y
1	Received ($D_1 = 1$)	Yes ($Y_1(0) = 1$	No ($Y_1(1) = 0$)	No ($Y_1 = 0$)
2	Not received ($D_2 = 0$)	No ($Y_2(0) = 0$)	Yes ($Y_2(1) = 1$)	No ($Y_2 = 0$)
3	Received ($D_3 = 1$)	No ($Y_3(0) = 1$)	Yes ($Y_3(1) = 1$)	Yes ($Y_3 = 1$)
4	Received ($D_4 = 1$)	Yes ($Y_4(0) = 1$)	No ($Y_4(1) = 0$)	No ($Y_4 = 0$)
5	Not received ($D_5 = 0$)	No ($Y_5(0) = 0$)	No ($Y_5(1) = 0$)	No ($Y_5 = 0$)

Table 1.1: The missing data problem

important problem is that only one of the potential outcomes is observed for each observation in the sample. Therefore, the causal effect cannot be directly calculated. Thus, the potential outcomes framework formulates the problem of causal inference as a problem of missing data. Table 1.1 is also a demonstration for the missing data interpretation of potential outcomes. Observe that for survivors who did not receive restraining orders, the observed outcome, Y_i, is equal to $Y_i(0)$; for survivors who received restraining orders, the observed outcome, Y_i, is equal to $Y_i(1)$. In order to compute the causal effect of restraining orders for each survivor, we need to know both $Y_i(0)$ and $Y_i(1)$, but one of these potential outcomes is always missing.

We define additional notation to help address the problem of missing data. Let $d \in \{0, 1\}$ denote particular instances of the treatment. Of particular importance to the focus of this thesis is the dth conditional expectation function of $Y_i(d)$ given X_i, which we denote with $f_d(X)$. Mathematically, we define

$$f_d(X_i) = \mathbb{E}[Y_i(d)|X_i],$$

where $\mathbb{E}[\cdot]$ denotes the expectation over probability density function $p(\cdot)$. With this notation, we can write

$$Y_i(d) = f_d(X_i) + \epsilon_i(d), \tag{1.1}$$

where $\epsilon_i(d)$ is the error term of $Y_i(d)$. Since $f_d(X_i)$ is the conditional expectation function of $Y_i(d)$, we know that

$$\mathbb{E}[\epsilon_i(d)|X_i] = 0,$$

and we can write $\text{Var}(\epsilon_i(d)|X_i) = \sigma_i^2(d)$ for some standard deviation $\sigma_i(d)$ specific to observation i.

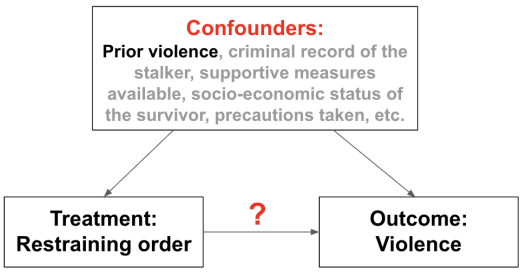
To approximate the estimand of interest for this thesis, GTFB involves estimating

$f_0(X_i)$ using various models. For the sake of generalizing $f_0(X_i)$ to allow functions without linear representations, we do not assume a functional form for $f_0(X_i)$. We introduce the models used in this thesis in Section 1.5.

1.2 Assumptions

We adopt the **Stable Unit Treatment Value Assumption (SUTVA)** associated with the potential outcomes framework. SUTVA stipulates that (1) the potential outcome $Y_i(d)$ is not a function of treatment D_j for all $i \neq j$, and (2) the same version of treatment is assigned to all treated units. In the example on restraining orders, SUTVA requires that (1) whether or not Survivor i obtains a restraining order has no impact on whether or not survivor j experiences violence, and (2) all survivors who received restraining orders received the exact same order. SUTVA is likely not satisfied in this study since different states may issue different types of restraining orders, violating the second requirement. Nevertheless, we adopt SUTVA since it facilitates the estimation of causal effects.

Figure 1.1: Conditional Ignorability



We also assume **conditional ignorability**:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i | X_i,$$

which states that *conditional on the observed covariates*, treatment assignment is independent of the potential outcomes. Define a **confounder** as a variable that influences both the treatment status and the potential outcomes. Following the example on restraining orders, we may expect that survivors who have experienced prior violence are more likely to (1) obtain restraining orders and (2) experience violence. Therefore, as a confounder in the study, prior violence may lead to a failure to observe a causal effect between restraining orders and the rate of violence. Conditional ignorability states that there is no unobserved confounder: all confounders are included

in X and can be adjusted for. Again consider the example on restraining orders, illustrated by Figure 1.1. If the only observed covariate is prior violence, conditional ignorability would state that prior violence is the only variable that influences both the administration of restraining orders and subsequent violence. Observe that this assumption is unrealistic: other potential confounders include the criminal record of the stalker, the socio-economic status of the survivor, supportive resources available to the survivor, and precautions taken by the survivor. Since potential causal relationships involve many confounders and studies can only observe a limited number of them, conditional ignorability is not likely to hold. However, it is an important simplifying assumption that facilitates methodological development.

1.3 Estimands and Estimators

In a given statistical analysis, we refer to the quantity of interest as the estimand. There are several common estimands in causal inference. The Average Treatment Effect,

$$\text{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)],$$

represents the mean effect of treatment across the entire population of interest. The Average Treatment Effect on the Treated,

$$\text{ATT} = \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1],$$

represents the mean effect of treatment for the treated units. The Average Treatment Effect on the Controls,

$$\text{ATC} = \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 0],$$

represents the mean effect of treatment for the units that did not receive treatment.

In the example on restraining orders, the ATT is the average effect of restraining orders in reducing violence for survivors who obtained restraining orders, the ATC is the average effect of restraining orders in reducing violence for survivors who did not obtain restraining orders, and the ATE is the average effect of restraining orders in reducing violence for all survivors in the population. In order to distinguish between these estimands, note that treatment effects can be homogeneous, where the expected causal effect is constant across all observations, or heterogeneous, where the true causal effect systematically varies for different observations. The three causal

estimands might take different values when (1) the distribution of covariates in the treated and control groups differ and (2) the treatment effect is systematically heterogeneous.

Continuing the example on restraining orders, it is possible that restraining orders are more effective in reducing violence for survivors who have not experienced violence previously, since the order could signal to the stalker that the legal system takes the safety of the survivor seriously and deter them from violence; in contrast, in situations where the prior violence was perpetrated, the order may aggravate the stalker and escalate the situation. In this case, if we assume that the treated group includes a higher proportion of survivors who have experienced prior violence than the control group and that the treatment effect is negative in both groups, the ATC would be greater in magnitude than the ATT. The ATE, which is a weighted mean of the ATT and ATC, always lies in between[1]. We focus on estimating the **ATT** in this thesis.

The ideal estimator for the ATT is the **Sample Average Treatment Effect on the Treated (SATT)**:

$$\text{SATT} = \frac{1}{n_t} \sum_{i:D_i=1} Y_i(1) - \frac{1}{n_t} \sum_{i:D_i=1} Y_i(0).$$

Since the SATT simply realizes the expectations in the ATT as sample means, it is unbiased for the ATT. However, $Y_i(0)$ is not observed for treated units, so the SATT cannot be calculated in the sample without stringent assumptions. One estimator that remedies this observability problem is the **Difference in Means (DIM)** estimator:

$$\text{DIM} = \frac{1}{n_t} \sum_{i:D_i=1} Y_i(1) - \frac{1}{n_c} \sum_{i:D_i=0} Y_i(0),$$

which simply replaces the mean $Y_i(0)$ in the treated group with the mean $Y_i(0)$ in the control group. The DIM can be calculated from the sample, but it is often biased in the presence of confounding. For example, simply taking the difference in the rates of violence for survivors who did versus did not obtain restraining orders does not address the fact that higher rates of prior violence can lead to both obtaining restraining orders

[1]By the law of total expectation, we can rewrite the ATE:

$$\begin{aligned}\mathbb{E}[Y_i(1) - Y_i(0)] &= \mathbb{E}\big[\mathbb{E}[Y_i(1) - Y_i(0)|D_i]\big] \\ &= p(D_i = 1)\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1] + p(D_i = 1)\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 0] \\ &= p(D_i = 1) \cdot \text{ATT} + p(D_i = 0) \cdot \text{ATC}.\end{aligned}$$

and experiencing violence. Therefore, if we observe the same rate of violence in the treated and control groups, we do not know whether this lack of difference is a genuine indication that there is no effect or a consequence of confounding. An improvement upon the DIM is the **Weighted Difference in Means** estimator:

$$\hat{\tau}_{\text{wdim}} = \frac{1}{n_t} \sum_{i:D_i=1} Y_i(1) - \frac{1}{n_c} \sum_{i:D_i=0} w_i Y_i(0),$$

where we define $w = [w_1 \cdots w_{n_c}]^\top \in \mathbb{R}^{n_c}$ to be a vector of non-negative weights for the control units. Observe that $\hat{\tau}_{\text{wdim}}$ replaces the unobserved $\frac{1}{n_t} \sum_{i:D_i=1} Y_i(0)$ with its weighted counterpart on the control group, $\frac{1}{n_c} \sum_{i:D_i=0} w_i Y_i(0)$, which is observed. The goal of weighting is to make the control group more "similar" to the treated group in terms of some function of the covariates. Again consider the example on restraining orders, illustrated by a simple dataset in Table 1.2. The top panel displays the treated group, and the bottom panel displays the control group. Observe that in this example, survivors who did not receive restraining orders are less likely to have experienced prior violence. Quantitatively, prior to weighting, we observe that

$$\frac{1}{n_c} \sum_{i:D_i=0} X_i = \frac{1}{4},$$

$$\frac{1}{n_t} \sum_{i:D_i=1} X_i = \frac{3}{4}.$$

Assume that prior violence is the only observed covariate and potential confounder. In order to address confounding, we give higher weight to the survivor who did not obtain restraining orders but experienced prior violence (Survivor 4) than survivors who did not obtain restraining orders and have not experienced prior violence (Survivors 1-3). After weighting,

$$\frac{1}{n_c} \sum_{i:D_i=0} w_i X_i = \frac{3}{4} = \frac{1}{n_t} \sum_{i:D_i=1} X_i.$$

Thus, these weights equate the covariate means in the treated and control groups in an attempt to mitigate the effect of confounding on the ATT estimate.

1.4 Weighting

We now derive the condition under which $\hat{\tau}_{\text{wdim}}$ is unbiased.

Since the SATT is the ideal estimator for the ATT, ideal weights equate $\hat{\tau}_{\text{wdim}}$

Survivor	Restraining order (D)	Violence (Y)	Prior violence (X)	Weight (w)
2	1	1	0	1
5	1	1	1	1
7	1	0	1	1
8	1	1	1	1
Survivor	Restraining order (D)	Violence (Y)	Prior violence (X)	Weight (w)
1	0	1	0	$\frac{1}{3}$
3	0	0	0	$\frac{1}{3}$
4	0	1	1	3
6	0	0	0	$\frac{1}{3}$

Table 1.2: Weighting

with the SATT in expectation. To evaluate the effectiveness of the weights, we take the difference between $\hat{\tau}_{\text{wdim}}$ and the SATT:

$$\begin{aligned}
\hat{\tau}_{\text{wdim}} - \text{SATT} &= \frac{1}{n_t}\sum_{i:D_i=1} Y_i(1) - \frac{1}{n_c}\sum_{i:D_i=0} w_i Y_i(0) - \left(\frac{1}{n_t}\sum_{i:D_i=1} Y_i(1) - \frac{1}{n_t}\sum_{i:D_i=1} Y_i(0)\right) \\
&= \frac{1}{n_t}\sum_{i:D_i=1} Y_i(0) - \frac{1}{n_c}\sum_{i:D_i=0} w_i Y_i(0) \\
&= \frac{1}{n_t}\sum_{i:D_i=1} \big(f_0(X_i) + \epsilon_i(0)\big) - \frac{1}{n_c}\sum_{i:D_i=0} w_i\big(f_0(X_i) + \epsilon_i(0)\big) \\
&= \big(\frac{1}{n_t}\sum_{i:D_i=1} f_0(X_i) - \frac{1}{n_c}\sum_{i:D_i=0} w_i f_0(X_i)\big) + \big(\frac{1}{n_t}\sum_{i:D_i=1} \epsilon_i(0) - \frac{1}{n_c}\sum_{i:D_i=0} w_i\epsilon_i(0)\big).
\end{aligned}$$

We assume that w is "honest", i.e., a function of X and D, but not Y. Thus, we have

$$\big(w_i \perp\!\!\!\perp Y_i(0)\big)|X, D,$$

which means that

$$\big(w_i \perp\!\!\!\perp \epsilon_i(0)\big)|X, D.$$

This implies that

$$\begin{aligned}
\mathbb{E}\big[w_i\epsilon_i(0)\big] &= \mathbb{E}\Big(\mathbb{E}[w_i\epsilon_i(0)|X,D]\Big) \\
&= \mathbb{E}\Big(\mathbb{E}[w_i|X,D]\mathbb{E}[\epsilon_i(0)|X,D]\Big) \\
&= \mathbb{E}\Big(\mathbb{E}[w_i|X,D]\Big)\mathbb{E}\Big(\mathbb{E}[\epsilon_i(0)|X,D]\Big) \\
&= \mathbb{E}[w_i]\mathbb{E}[\epsilon_i(0)] \qquad (1.2) \\
&= 0. \qquad (1.3)
\end{aligned}$$

Since $\mathbb{E}[\text{SATT}] = \text{ATT}$, we can rewrite the bias of $\hat{\tau}_{\text{wdim}}$:

$$\mathbb{E}[\hat{\tau}_{\text{wdim}} - \text{ATT}] = \mathbb{E}[\hat{\tau}_{\text{wdim}} - \text{SATT}] = \mathbb{E}[\frac{1}{n_t}\sum_{i:D_i=1} f_0(X_i) - \frac{1}{n_c}\sum_{i:D_i=0} w_i f_0(X_i)]. \quad (1.4)$$

Thus, the bias of $\hat{\tau}_{\text{wdim}}$ is primarily determined by the weighted difference in means of $f_0(X)$. For any function $h(X)$ of X, define the **imbalance** of $h(X)$ as the weighted difference in means of $h(X)$:

$$\text{imbal}(w, h(X), D) = \frac{1}{n_t}\sum_{i:D_i=1} h(X_i) - \frac{1}{n_c}\sum_{i:D_i=0} w_i h(X_i).$$

Then we can rewrite the bias of $\hat{\tau}_{\text{wdim}}$ as

$$\text{Bias}(\hat{\tau}_{\text{wdim}}) = \mathbb{E}[\text{imbal}(w, f_0(X), D)].$$

Existing methods attempt to achieve $\text{imbal}(w, f_0, D) = 0$ in different ways. For example, note that when $f_0(X) = X\beta$,

$$\text{imbal}(f_0(X)) = \text{imbal}(X)\beta,$$

which means that balance in X translates directly into balance in $f_0(X)$. One class of widely used methods, **balancing weights**, addresses $\text{imbal}(w, X, D)$, which is defined as

$$\text{imbal}(w, X, D) = \frac{1}{n_t}\sum_{i:D_i=1} X_i - \frac{1}{n_c}\sum_{i:D_i=0} w_i X_i.$$

Exact balancing weights set $\text{imbal}(w, X, D) = 0$, and approximate balancing weights set $|\text{imbal}(w, X, D)| < r$, where r is a low value specific to the method. Balancing weights then choose w that minimizes some function of w as a means to control the

variance of w. Balancing weights methods perform well when $f_0(X)$ is linear in X, but linearity is required to guarantee good performance. One important example of balancing weights methods is **entropy balancing (EBAL)**, proposed by Hainmueller (2012), which achieves exact balance on X while minimizing $\frac{1}{n_c}\sum_{i:D_i=0} w_i \log(w_i)$. We make extensive use of EBAL as a comparison estimator in Chapter 3.

Another comparison estimator we will use in this thesis is **kernel balancing (KBAL)**, proposed by Hazlett (2020). KBAL achieves approximate balance on very high dimensional non-linear functions of the covariates that can approximate any smooth function.

A different approach to debiasing $\hat{\tau}$ uses weights derived from the probability of treatment given the covariates. Define the propensity score for each X_i as

$$\pi(X_i) = p(D_i = 1|X_i).$$

Proposed by Rosenbaum & Rubin (1983), **inverse propensity weights (IPW)** for the ATT are chosen based on the propensity score:

$$w_i^{\text{IPW}} = \frac{n_0}{n_1}\frac{\pi(X_i)}{1-\pi(X_i)}.$$

To understand the utility of IPW, compute that

$$\begin{aligned}
\frac{n_0}{n_1}\frac{\pi(X_i)}{1-\pi(X_i)} &= \frac{n_0/n}{n_1/n}\frac{\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)} \\
&= \frac{p(D_i=0)}{p(D_i=1)}\frac{p(D_i=1|X_i)}{1-p(D_i=1|X_i)} \\
&= \frac{p(D_i=0)}{p(D_i=1)}\frac{p(D_i=1|X_i)}{p(D_i=0|X_i)} \\
&= \frac{p(D_i=0)}{p(D_i=1)}\frac{p(D_i=1,X_i)/p(X_i)}{p(D_i=0,X_i)/p(X_i)} \\
&= \frac{p(D_i=0)}{p(D_i=1)}\frac{p(D_i=1,X_i)}{p(D_i=0,X_i)} \\
&= \frac{p(D_i=1,X_i)/p(D_i=1)}{p(D_i=0,X_i)/p(D_i=0)} \\
&= \frac{p(X_i|D_i=1)}{p(X_i|D_i=0)}.
\end{aligned}$$

Thus, IPW equate the distributions of X in the treated and control groups. Now

compute that

$$\begin{aligned}
\text{Bias}(\hat{\tau}_{\text{wdim}}) &= \mathbb{E}[\text{imbal}(w, f_0(X), D)] \\
&= \mathbb{E}\big[\frac{1}{n_t}\sum_{i:D_i=1} f_0(X_i) - \frac{1}{n_c}\sum_{i:D_i=0} w_i f_0(X_i)\big] \\
&= \mathbb{E}[f_0(X_i)|D_i = 1] - \mathbb{E}[w_i f_0(X_i)|D_i = 0] \\
&= \int f_0(x)\, p(X = x|D = 1)\, dx \\
&\quad - \int w\, f_0(x)\, p(X = x, D = 0)\, dx \\
&= \int f_0(x)\, p(X = x|D = 1)\, dx \\
&\quad - \int f_0(x)\, \frac{p(X = x|D = 1)}{p(X = x|D = 0)}\, p(X = x, D = 0)\, dx \\
&= \int f_0(x)\, p(X = x|D = 1)\, dx \\
&\quad - \int f_0(x)\, p(X = x|D = 1)\, dx \\
&= 0.
\end{aligned}$$

Thus, $\hat{\tau}_{\text{IPW}}$ is unbiased for any f_0, regardless of whether it is linear in X. In practice, the true $\pi(X_i)$ is unknown, and we model D as a function of X and estimate $\pi(X_i)$ with prediction $\hat{\pi}(X_i)$. If $\hat{\pi}(X_i)$ is specified incorrectly or modeling error is high, $\hat{w_i}^{\text{IPW}}$ would be incorrect.

1.5 Models

This section briefly describes models that we use for f_0 throughout this thesis.

1.5.1 GLMs

Generalized linear models (GLMs) assume that for some link function g and its inverse g^{-1}, $g^{-1}\big(f_0(X_i)\big)$ is linear in X_i. Stated formally,

$$f_0(X_i) = g\big(X_i^\top \beta\big),$$

where $\beta = [\beta^{(0)} \ \dots \ \beta^{(p)}]^\top \in \mathbb{R}^p$ is a vector of coefficients.

This thesis features three commonly used link functions:

- the identity function, which formulates the GLM as

$$f_0(X_i) = X_i^\top \beta,$$

  which is equivalent to a linear model. The identity link applies to outcomes that are linearly related to the covariates;

- the logit function, which formulates the GLM as

$$f_0(X_i) = p(Y_i(0) = 1|X_i) = \frac{\exp(X_i^\top \beta_0)}{\exp(X_i^\top \beta_0) + 1}.$$

  This is known as logistic regression and commonly used to model binary outcomes;

- the probit function, which formulates the GLM as

$$f_0(X_i) = p(Y_i(0) = 1|X_i) = \Phi(X_i^\top \beta_0),$$

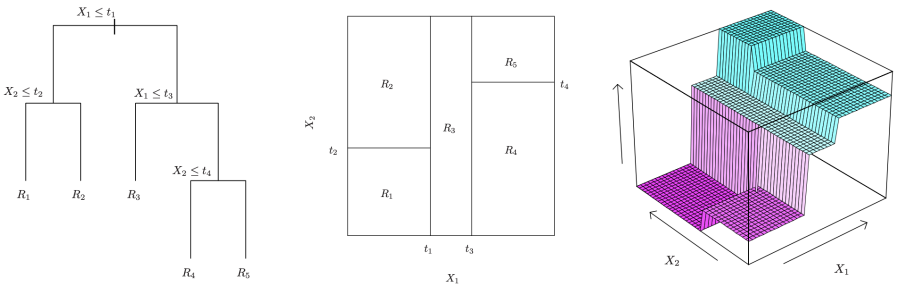
  where Φ is the cumulative distribution function of the standard normal distribution. This is known as probit regression and is an alternative method to modeling binary outcomes.

Probit regression often yields very similar predictions as logistic regression. However, Chambers & Cox (1967) found that the two models can be distinguished when the sample size is large and a substantial portion of observations take extreme values in an independent variable.

1.5.2 Random forest

Decision trees is a supervised machine learning method that generates predictions in a manner that mimics human decision-making. For simplicity, we assume the outcome is continuous while discussing decision trees and the related ensemble methods. Each decision tree starts with a root node and performs recursive binary splitting following the steps below:

1. Consider all covariates and possible thresholds that split the covariate space into two distinct regions. For each region, predict the outcome as the mean outcome in the region.



(a) Decision tree (b) Covariate space (c) Step function

Figure 1.2: Decision trees

(a) This figure comes from Chapter 8 of James et al. (2013).

2. Find the covariate and threshold that result in the greatest reduction in the residual sum of squares (RSS). Make the first split at this threshold of the covariate.

3. Repeat the above steps for each branch until a pre-specified stopping condition is met.

Consider an example with a continuous outcome and two continuous predictors, X_1 and X_2. Figure 1.2a from James et al. (2013) visually demonstrates the structure of a decision tree for this setting. Note that the tree has a root node that splits the covariate space by comparing the values of X_1 to a threshold t_1. The tree then proceeds to divide the two splits of the covariate space by comparing X_2 to t_2 if $X_1 \leq t_1$, and X_1 to t_3 if $X_1 > t_1$. After making one more split, the tree partitions the covariate space into five rectangular regions, $R_1, \ldots, R_5$, represented visually by Figure 1.2b. The tree then uses the mean outcome value of each region as the prediction for observations that belong to the region. Since the predictions are identical across all observations in each region, decision trees model the outcome as a step function, illustrated by Figure 1.2c, where the value of the predicted outcome for each region is represented by the height of the surface that lies directly above the region.

Random forest is an ensemble machine learning method making use of decision trees. In order to control the variance of predictions made by single decision trees, random forest takes bootstrap samples from the original data and fits a decision tree independently on each bootstrap sample. To further decrease the variance of the predictions, random forest decreases the correlation between decision trees by randomly selecting a subset of m covariates to consider for each split of each tree, where m is by default set to $\sqrt{p}$. The random forest then reports the mean of the

predictions across the bootstrap samples as the final prediction. Thus, random forest improves upon decision trees by reducing the variance of the predictions.

1.5.3 Boosting

Boosting is also an ensemble learning method making use of decision trees. In contrast to the random forest which relies on many independent trees, boosting fits decision trees that depend highly on one another. Central to the algorithm of boosting are the shrinkage parameter (λ) and the number of splits allowed in each tree (d). During the first iteration of the boosting algorithm, a tree with d splits is fit to the original data, and a residual is calculated as $y - \lambda\hat{f}(X)$. For each subsequent iteration of the algorithm, a tree with d splits is fit to the residual from the previous iteration, and the residual is updated by subtracting the shrunken version of the new prediction. The algorithm reports the sum of the shrunken predictions from all iterations.

Whereas the random forest focuses on decreasing the variance of the predictions, boosting decreases the bias. Since each tree is fit to the residual of previous trees, each iteration of the boosting algorithm is informed by all previous iterations and specializes in exploring directions that previous iterations have not adequately explored. Thus, the advantage of boosting lies in the metaphorical fact that each tree, metaphorically, stands on the metaphorical shoulders of the not-so-giant trees that came before.

1.5.4 Bayesian additive regression trees

Bayesian additive regression trees (BART) is again an ensemble method that makes use of decision trees. BART combines the strategies of both random forest and boosting, incorporating both randomness and information from existing trees while fitting each new tree. The BART algorithm starts with K root nodes which all predict the mean outcome divided by K, and runs for B iterations. For each tree in each subsequent iteration, a partial residual is computed by subtracting the sum of the latest predictions from all other trees. The trees are then randomly perturbed, where the probability of each possible perturbation is determined by its improvement on the tree's fit to the partial residual. Possible perturbations to a tree include growing an additional branch, pruning an existing branch, changing the method of splitting for a node, and changing the prediction of a node. BART reports B predictions, each of which is the sum of predictions from K trees. If a single prediction is desired, the mean can be taken from the B predictions.

In the Bayesian framework, each set of reported predictions from BART constitutes a sample from the posterior predictive distribution. Since the Bayesian posterior predictive distribution approximates independent samples from the population of interest, BART enjoys the desirable property that functions of the set of predictions directly estimate the corresponding functions of independent samples from the population. For example, because the variance of the BART predictions is an estimator for the variance of the single reported prediction, we can avoid the computationally intensive bootstrap for the variance of the reported prediction.

Chapter 2

Generalized Targeted Function Balancing

2.1 Original Targeted Function Balancing

This subsection gives a brief, non-technical overview of the original formulation of TFB by Wainstein (2022).

Recall that $\hat{\tau}_{\text{wdim}}$ is unbiased when the imbalance of $f_0(X)$ is 0. TFB is a method for finding weights that minimize imbal$(w, f_0(X_i), D)$. TFB first regresses $Y_i(0)$ on X_i in the control group to estimate a linear model $\hat{f}_0$ of f_0, and uses $\hat{f}_0$ to obtain predictions $\hat{f}_0(X_i)$ on the treated group. Since $\hat{f}_0(X)$ is an estimate of $f_0(X)$, and balance on $f_0(X)$ is desirable, it is tempting to find weights that minimize imbalance in $\hat{f}_0(X)$. However, error and uncertainty exist in every model, so balance in $\hat{f}_0(X)$ does not necessarily imply balance in $f_0(X)$. Therefore, in an attempt to capture the true $f_0(X)$ in a neighborhood around $\hat{f}_0(X)$, TFB minimizes the worst-case imbal$(w, f_0(X), D)$ for $f_0(X)$ that is "probabilistically near" $\hat{f}_0(X)$. The region of probabilistic nearness around $\hat{f}_0(X)$ is one where the true $f_0(X)$ falls with some high probability according to the estimated variance of $\hat{f}_0(X)$.

Since the TFB weights are completely determined by the model $\hat{f}_0$, TFB turns the question of how to balance the covariates into the question of how to model the outcome. Selecting a model for the outcome allows better use of the researcher's domain knowledge than selecting a weighting method. Additionally, model choice is a familiar topic to applied researchers and is relatively well documented. Therefore, modeling the outcome is more straightforward than selecting a weighting method from the depths of the causal inference literature. Thus, TFB reduces the burden of

decision making on the applied researcher.

TFB in its original form requires the model for f_0 to have a linear representation, meaning that we must be able to write $\hat{f}_0(X) = X\hat{\beta}$. X can be expanded to include non-linear and interaction terms, but the requirement of linear representation is still restrictive - it precludes the use of GLMs, many of the non-parametric machine learning methods, and many other model. This chapter of the thesis extends TFB to GTFB, which allows the use of any predictive models. We demonstrate that the original TFB is equivalent to the linear version of the generalized TFB up to a scalar after deriving the generalized TFB in subsection 2.2.

2.2 Deriving GTFB

This subsection derives the Generalized Targeted Function Balancing (GTFB), which allows any predictive model for $\hat{f}_0$.

We start by approximating the imbalance of $f_0(X)$. First rewrite the absolute value of this imbalance to incorporate the predicted outcomes, $\hat{f}_0(X)$:

$$\big|\text{imbal}(w, f_0(X), D)\big| = \big|\text{imbal}(w, f_0(X) - \hat{f}_0(X), D) + \text{imbal}(w, \hat{f}_0(X), D)\big|.$$

Following the rationale of the original TFB, GTFB seeks balance on $f_0(X)$ that is "probabilistically near" $\hat{f}_0(X)$. Note that $f_0(X) = [f_0(X_1)\ f_0(X_2)\ \cdots\ f_0(X_n)]^\top \in \mathbb{R}^n$, and $\hat{f}_0(X)$ is the n-dimensional estimate of $f_0(X)$. In order to operationalize probabilistic nearness, we define a small region S_q centered at the origin in the n-dimensional Euclidean space, in which, according to the estimated variance of $\hat{f}_0(X)$, $f_0(X) - \hat{f}_0(X)$ falls with some high probability q. We say that $f_0(X)$ is "probabilitstically near" $\hat{f}_0(X)$ if the difference between the two is within S_q. Now write the worst-case absolute imbalance of $f_0(X)$ probabilistically near $\hat{f}_0(X)$:

$$\max_{f_0 \in \hat{f}_0 - S_q} |\text{imbal}(w, f_0(X) - \hat{f}_0(X), D) + \text{imbal}(w, \hat{f}_0(X), D)|.$$

To conveniently delineate the region S_q, we impose the working assumption that $\hat{V}_{\hat{f}_0}^{-\frac{1}{2}}\big(\hat{f}_0(X) - f_0(X)\big) \xrightarrow{d} N(n_n, I_n)$. Denote with $\hat{V}_{\hat{f}_0}$ the estimated variance of $\hat{f}_0(X)$. We can then choose S_q such that after scaling by $\hat{V}_{\hat{f}_0}^{-\frac{1}{2}}$, S_q is a ball centered at 0_n that has probability q of containing $N(0_n, I_n)$:

$$S_q = \{\hat{V}_{\hat{f}_0}^{\frac{1}{2}} z \in \mathbb{R}^n : ||z||_2 \leq \sqrt{Q_q(\chi_n^2)}\},$$

and we refer to $\sqrt{Q_q(\chi^2_n)}$ as the Squidward Constant.

Write

$$u = \begin{bmatrix} -\frac{1}{n_c}w_1 \\ \vdots \\ -\frac{1}{n_c}w_{n_c} \\ \frac{1}{n_t} \\ \vdots \\ \frac{1}{n_t} \end{bmatrix} \in \mathbb{R}^n, \quad f_0(X) - \hat{f}_0(X) = \begin{bmatrix} f_0(X_1) - \hat{f}_0(X_1) \\ \vdots \\ f_0(X_{n_c}) - \hat{f}_0(X_{n_c}) \\ f_0(X_{n_{c+1}}) - \hat{f}_0(X_{n_{c+1}}) \\ \vdots \\ f_0(X_n) - \hat{f}_0(X_n) \end{bmatrix} \in \mathbb{R}^n.$$

Note that for any vector $v = [v_1 \cdots v_n]^\top$, $v^\top u = \text{imbal}(w, v, D)$. We can then rewrite the worst-case absolute imbalance of $f_0(X)$ probabilistically near $\hat{f}_0(X)$:

$$\begin{aligned}
&\max_{f_0 \in \hat{f}_0 - S_q} \left|(f_0(X) - \hat{f}_0(X))^\top u + \hat{f}_0(X)^\top u\right| \\
&\leq \max_{f_0 \in \hat{f}_0 - S_q} \left|(\hat{f}_0(X) - f_0(X))^\top u\right| + \left|\hat{f}_0(X)^\top u\right| \\
&= \max_{\hat{f}_0 - f_0 \in S_q} \left|(\hat{f}_0(X) - f_0(X))^\top u\right| + \left|\hat{f}_0(X)^\top u\right| \\
&= \max_{||z|| \leq \sqrt{Q_q(\chi^2_n)}} \left|\left|[\hat{V}^{\frac{1}{2}}_{\hat{f}_0} z]^\top\ u\right|\right| + \left|\hat{f}_0(X)^\top u\right| \\
&= \max_{||z|| \leq \sqrt{Q_q(\chi^2_n)}} \left|\left|z^\top (\hat{V}^{\frac{1}{2}}_{\hat{f}_0})^\top u\right|\right| + \left|\hat{f}_0(X)^\top u\right| \\
&= \max_{||z|| \leq \sqrt{Q_q(\chi^2_n)}} ||z^\top||\ ||(\hat{V}^{\frac{1}{2}}_{\hat{f}_0})^\top u|| + \left|\hat{f}_0(X)^\top u\right| \\
&= \max_{||z|| \leq \sqrt{Q_q(\chi^2_n)}} ||z||\ ||\hat{V}^{\frac{1}{2}}_{\hat{f}_0} u|| + \left|\hat{f}_0(X)^\top u\right| \\
&= \sqrt{Q_q(\chi^2_n)} \times \left|\left|\hat{V}^{\frac{1}{2}}_{\hat{f}_0} u\right|\right|_2 + \left|\hat{f}_0(X)^\top u\right| \\
&= \sqrt{Q_q(\chi^2_n)} \times \left|\left|\hat{V}^{\frac{1}{2}}_{\hat{f}_0} u\right|\right|_2 + \left|\text{imbal}(w, \hat{f}_0(X), D)\right|.
\end{aligned}$$

Thus, we conservatively approximate the imbalance in $f_0(X)$:

$$|\text{imbal}(w, f_0(x), D)| \leq \sqrt{Q_q(\chi^2_n)} \times \left|\left|\hat{V}^{\frac{1}{2}}_{\hat{f}_0} u\right|\right|_2 + \left|\text{imbal}(w, \hat{f}_0(X), D)\right|. \tag{2.1}$$

We proceed to derive GTFB weights using the mean squared error (MSE) of $\hat{\tau}_{\text{wdim}}$ for the SATT. Appendix B shows the derivation of the MSE, but we state the result

here:

$$\mathbb{E}\Big(\big(\tau_{\text{wdim}}(w) - \text{SATT}\big)^2 \;\Big|\; X, D\Big) = |\text{imbal}(w, f_0(X), D)|^2 + \frac{1}{n_c^2} \sum_{i:D_i=0} w_i^2 \sigma_i(0)^2 + C,$$

where C is not a function of w. Recall that we approximate the worst-case $|\text{imbal}(w, f_0(X), D)|$ using Equation 2.1 since $f_0(X)$ is not directly observed. GTFB chooses weights that minimize the approximated MSE:

$$\begin{aligned}
&\hat{w}^{\text{GTFB}} \\
&\approx \underset{w \geq 0}{\text{argmin}} \left[|\text{imbal}(w, f_0(X), D)|^2 + \frac{1}{n_c^2} \sum_{i:D_i=0} w_i^2 \sigma_i^2(0) \right] \\
&\approx \underset{w \geq 0}{\text{argmin}} \Bigg[\frac{1}{n_c^2} ||w\ \hat{\sigma}(0)^\top||_2^2 && \text{(term 1)} \\
&\quad + \Bigg(\Big|\text{imbal}(w, \hat{f}_0(X), D)\Big| && \text{(term 2)} \\
&\quad + \sqrt{Q_q(\chi_n^2)}\ \Big|\Big|\hat{V}_{\hat{f}_0}^{\frac{1}{2}} u\Big|\Big|_2 \Bigg)^2 \Bigg], && \text{(term 3)}
\end{aligned}$$

where $\sum_{i:D_i=0} w_i = n_c$ and $\hat{\sigma}^2(0) = [\hat{\sigma}_1^2(0) \ldots \hat{\sigma}_{n_c}^2(0)]^\top$, where $\hat{\sigma}_i^2(0) = \frac{1}{n_c} \sum_{i:D_i=0} (Y_i - \hat{f}_0(X_i))^2$ is an estimate of $\sigma_i^2(0)$.

2.3 Understanding GTFB

In order to understand the optimization problem presented in the previous section, we consider each of its three terms individually.

Term 1, $\frac{1}{n_c^2}||w\ \hat{\sigma}(0)^\top||_2^2$, controls the variance of w to minimize the MSE of $\hat{\tau}_{\text{wdim}}$.

Term 2, $\Big|\text{imbal}(w, \hat{f}_0(X), D)\Big|$, penalizes imbalance in $\hat{f}_0(X)$. This term relies on the estimate, $\hat{f}_0(X)$, to accurately approximate $f_0(X)$, in order to achieve the ultimate goal of penalizing imbalance in $f_0(X)$.

However, due to inevitable modeling error and estimation uncertainty, we cannot expect $\hat{f}_0(X) = f_0(X)$. Term 3 addresses this concern. As we demonstrate later in this section, Term 3 penalizes imbalance in a quantity in which $f_0(X)$ is assumed to be linear. Balance in this quantity translates into balance in $f_0(X)$. Thus, Term 3 bypasses the model $\hat{f}_0$ to safeguard against modeling error in $\hat{f}_0$. The importance of

this term is determined by $\hat{V}_{\hat{f}_0}^{\frac{1}{2}}$, which means that the more prediction uncertainty there is in $\hat{f}_0(X)$, the more emphasis GTFB places upon Term 3. Term 3 is scaled by the Squidward Constant.

To illustrate the functionality of Term 3, we consider linear models and GLMs as examples. In the setting of linear models, we assume that $f_0(X) = X\beta$ for some $\beta \in \mathbb{R}^p$. Estimate $\hat{f}_0(X) = X\hat{\beta}$. We can then simplify Term 3 by rewriting the variance of $\hat{f}_0(X)$ and obtain

$$\text{Term 3} = \sqrt{Q_q(\chi_n^2)}\big|\big|\hat{V}_{\hat{\beta}}^{\frac{1}{2}}\text{imbal}(w, X, D)\big|\big|_2.$$

Observe that the term now penalizes imbalance in X depending on the estimated variance of $\hat{\beta}$. Balance in X implies balance in any $f_0(X)$ that is a linear function of X, which yields an unbiased $\hat{\tau}_{\text{wdim}}$. Thus, GTFB seeks balance in X to safeguard against modeling error in $\hat{f}_0$. Linear GTFB differs from the original TFB only by the Squidward Constant, which involves a chi-squared distribution with p degrees of freedom in the original TFB. With p denoting the number of observed covariates and n denoting the sample size, we expect n to be far greater than p in typical observational studies. Therefore, GTFB places substantially greater emphasis on balancing X than TFB does. Section 3.1 demonstrates the effect of the difference in the Squidward Constant.

Now assume that f_0 can be modeled using a GLM. Let g be the inverse link function, so that

$$f_0(X) = g(X\beta).$$

Estimate $f_0(X)$ using $\hat{f}_0(X) = g(X\hat{\beta})$. Then we can write

$$\text{imbal}(w, \hat{f}_0(X), D) = \text{imbal}(w, g(X\hat{\beta}_0), D).$$

Using the Delta method to derive the variance of $\hat{f}_0(X)$, we can rewrite Term 3 as

$$\sqrt{Q_q(X_n^2)}\big|\big|\hat{V}_{\hat{\beta}_0}^{\frac{1}{2}}\text{imbal}([g']X)\big|\big|_2,$$

where $[g']$ is the diagonal matrix of derivatives,

$$[g'] = \begin{bmatrix} g'(X_1^\top\hat{\beta}) & 0 & ... & 0 \\ 0 & g'(X_2^\top\beta) & ... & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & ... & 0 & g'(X_n^\top\beta) \end{bmatrix}.$$

Appendix A shows the mathematical derivation. We can interpret the simplified version of Term 3 using first-order Taylor approximation:

$$\begin{aligned} g(X\beta) &\approx g(X_0\beta) + g'(X_0\beta)(X - X_0)\beta \\ &= g(X_0\beta) + g'(X_0\beta)X\beta - g'(X_0\beta)X_0\beta \\ &= \big(g(X_0\beta) - g'(X_0\beta)X_0\beta\big) + g'(X_0\beta)X\beta \end{aligned}$$

for some $X_0 \in \mathbb{R}^{n\times p}$. This approximation is linear in $g'(X_0\beta)X$. Since Taylor approximation requires X_0 to be close to X, so we can substitute X for X_0, further approximating $g(X\beta)$ as a linear function of $g'(X\beta)X$. Therefore, balance in $g'(X\beta)X$ translates to balance on $g(X\beta) = f_0(X)$. Thus, Term 3 in the minimization problem seeks balance on a quantity in which f_0 is linear to safeguard against modeling error in $\hat{f}_0$.

2.4 Variance estimation

Wainstein (2022) proposes a closed-form variance estimator for the original TFB.

For general weight $\hat{w}$ for $\hat{\tau}_{\text{wdim}}$, Wainstein (2022) proposes the variance estimator

$$\hat{V}_{\text{ATT}}(\hat{w}, \hat{f}_0) = \frac{1}{n_t^2} \sum_{i:D_i=1} \Big(Y_i(1) - \hat{f}_0(X_i) - \hat{\tau}_{\text{wdim}}(\hat{w})\Big)^2 + \frac{1}{n_c^2} \sum_{i:D_i=0} \hat{w_i}^2 \hat{e_i}^2(0), \quad (2.2)$$

where $\hat{e_i}^2(0) = Y_i(0) - \hat{f}_0(X_i)$. Theorem 2 from Wainstein (2022) states that under certain conditions, the proposed $\hat{V}_{\text{ATT}}(\hat{w}, \hat{f}_0)$ is consistent for $V_{\text{ATT}}(\hat{w})$. Since the estimated variance uses $\hat{f}_0(X)$ without assuming a linear structure for it, we use the same formulation for GTFB.

One condition that is not satisfied under the current formulation of GTFB is that $\hat{w}$ is a function of X and D, but not Y. Wainstein (2022) proposes sample splitting to address the problem. Sample splitting amounts to randomly splitting the sample into two halves of approximately equal size. Denote with superscripts $(s1)$ and $(s2)$ values from the two samples. GTFB then (1) builds a model for f_0 on sample 2, (2) uses the model to obtain predictions $\hat{f}_0(X)^{(s1)}$ and $\hat{V}_{\hat{f}_0}$ on sample 1, (3) uses these values to obtain weights $\hat{w}^{(s1)}_{\text{GTFB}}$, and (4) switches the samples and repeats the process. Sample splitting decorrelates $\hat{w}$ and Y within each sample, so the condition of $\hat{w}$ being a function of only X and D is satisfied for the estimator from each sample. These estimators are $\hat{\tau}^{(s1)}_{\text{wdim}}(\hat{w}^{(s1)}_{\text{GTFB}})$ and $\hat{\tau}^{(s2)}_{\text{wdim}}(\hat{w}^{(s2)}_{\text{GTFB}})$. Report the final estimate, $\hat{\tau}^{\text{GTFB}}_{\text{wdim}}$, as

the average of the estimates from the two samples:

$$\hat{\tau}^{\text{GTFB}}_{\text{wdim}} = \frac{\hat{\tau}^{(s1)}_{\text{wdim}}(\hat{w}^{(s1)}_{\text{GTFB}}) + \hat{\tau}^{(s2)}_{\text{wdim}}(\hat{w}^{(s2)}_{\text{GTFB}})}{2}. \tag{2.3}$$

Since the condition for "honest" weights is satisfied within each sample, we obtain variance estimators $\hat{V}^{(s1)}_{\text{ATT}}(\hat{w}^{(s1)}, \hat{f}_0^{\,(s2)})$ and $\hat{V}^{(s2)}_{\text{ATT}}(\hat{w}^{(s2)}, \hat{f}_0^{\,(s1)})$ by Equation 2.2. Then the variance estimator of the reported $\hat{\tau}^{\text{GTFB}}_{\text{wdim}}$ is

$$\text{Var}(\hat{\tau}^{\text{GTFB}}_{\text{wdim}}) = \frac{1}{4}\hat{V}^{(s1)}_{\text{ATT}}(\hat{w}^{(s1)}, \hat{f}_0^{\,(s2)}) + \frac{1}{4}\hat{V}^{(s2)}_{\text{ATT}}(\hat{w}^{(s2)}, \hat{f}_0^{\,(s1)}). \tag{2.4}$$

By the same logic as in Wainstein (2022), consistent estimation of $\text{Var}(\hat{\tau}^{\text{GTFB}}_{\text{wdim}})$ requires $\hat{\tau}^{(s1)}_{\text{wdim}}(\hat{w}^{(s1)}_{\text{GTFB}})$ and $\hat{\tau}^{(s2)}_{\text{wdim}}(\hat{w}^{(s2)}_{\text{GTFB}})$ to be asymptotically uncorrelated. Given correct specification of f_0, this condition is plausible.

Chapter 3

Demonstrations

This chapter demonstrates the performance of GTFB on simulated data, using correctly specified models for f_0 on three variations of the data generating process (DGP) detailed below. We compare the estimate $\hat{\tau}_{\text{wdim}}$ by GTFB to the true ATT value and estimates by other methods.

First generate the vector of covariates, $X_i = [X_i^{(1)} \ X_i^{(2)} \ X_i^{(3)} \ X_i^{(4)}]$. For $l \in \{1, \cdots, 4\}$, $X_i^{(l)}$ is simulated independently from a standard normal distribution:

$$X_i^{(l)} \sim N(0, 1).$$

Next generate the treatment status of the observations as a function of X_i. Let

$$\psi(X_i) = \frac{2.2\big(0.25[X_i^{(1)}]^3 + [X_i^{(2)}]^3 + [X_i^{(3)}]^3 + [X_i^{(4)}]^3\big)}{[X_i^{(1)}]^2 + 2},$$

and generate the propensity scores according to a probit model:

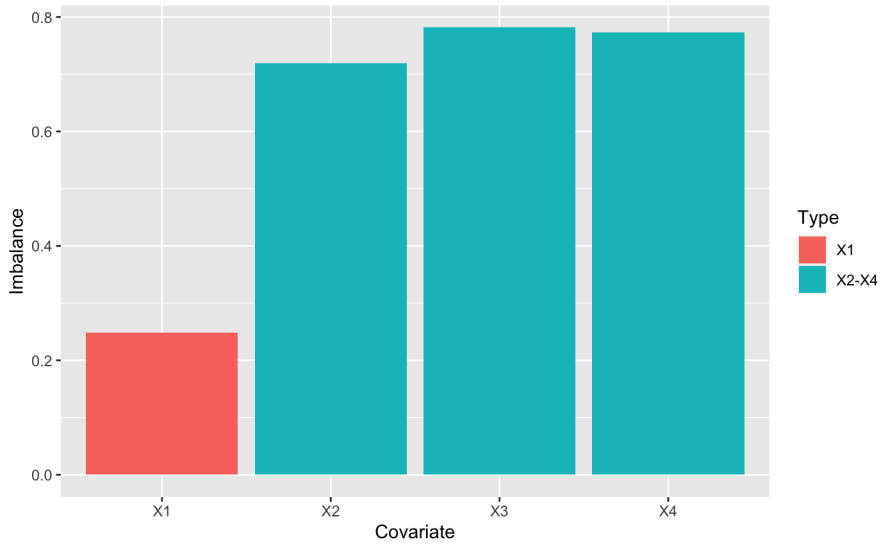
$$\pi(X_i) = p(D_i = 1|X_i) = \Phi(\psi(X_i)),$$

where Φ is the CDF of the standard normal distribution. Note that we generate $\psi(X_i)$ as a highly non-linear function of X_i to highlight differences between weighting methods in later demonstrations.

Figure 3.1 exhibits the starting imbalance of the covariates. Due to the lowest coefficient of $X_i^{(1)}$ in the numerator of $\psi(X_i)$, $X^{(1)}$ suffers from the least starting imbalance.

We then generate the outcome as some function of X_i, depending on the specific DGP. In particular, $X^{(1)}$ has the most influence on the outcome in each of the DGPs.

Figure 3.1: Starting imbalance of X



(a) This figure shows the starting imbalance of X for DGP 1. The imbalance is computed using one iteration of DGP 1 with $n = 8000$.

For each simulation, we take a large number ($M = 800$ for DGPs 1a and 1b, $M = 400$ for DGP 1c) of random samples from the DGP with sample size $n = 800$. Denote with $\hat{\tau}^{(m)}_{\text{wdim}}$ an arbitrary weighted estimator obtained in the mth iteration of the simulation. We evaluate the performance of the estimators using the bias and root mean squared error (RMSE):

$$\text{Bias} = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{\tau}^{(m)}_{\text{wdim}} - \text{ATT}\right), \tag{3.1}$$

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{\tau}^{(m)}_{\text{wdim}} - \text{ATT})^2}. \tag{3.2}$$

3.1 DGP 1a: GTFB in the linear setting

This subsection demonstrates the performance of GTFB in the linear setting, where GTFB strategically leaves imbalanace in the covariates to achieve high efficiency in addition to unbiasedness. For this DGP, we generate the outcome as a linear function

of X:

$$Y_i = 10X_i^1 + \sum_{l=2}^{4} X_i^{(l)} + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, 10). \tag{3.3}$$

Note that the variance of ϵ_i produces $R^2 = 0.5$. Observe that the coefficient of D_i is zero in Equation 3.3, which means Y_i is not a function of treatment assignment. Consequently, there is no treatment effect, and $Y_i = Y_i(0) = Y_i(1)$. This implies that the ATT, along with the ATC and ATE, is 0. Note that $X_i^{(1)}$ has the greatest coefficient in Equation 3.3, so it is the most influential in determining the outcome. Recall that $X^{(1)}$ also suffers from the least starting imbalance, as shown in Figure 3.1.

We consider four estimators for this simulation: GTFB, TFB, EBAL, and the unweighted DIM.

- For GTFB, we use the ordinary least squares (OLS) to model f_0 and the closed-form estimator to find the variance of $\hat{f}_0(X)$: $\hat{V}_{\hat{f}_0} = X\hat{V}_{\hat{\beta}}X^\top$.

- We use the same model and variance estimator for the original TFB. Recall that the linear GTFB differs from TFB in the degrees of freedom of the Squidward Constant. Specifically, due to a smaller degree of freedom ($p = 4 < 800 = n$), TFB is expected to place less emphasis on balancing X than GTFB. We expect the TFB methods to be unbiased because the specification of f_0 is correct.

- EBAL equates the means of X_i on the treated and control groups. We expect EBAL to be unbiased because f_0 is a linear function of X, and as noted in Section 1.4, when f_0 is linear in X, balance in X translates into balance in f_0.

- Finally, we expect the unweighted DIM to be biased because of the confounding effect of X which DIM does not attempt to address.

Figure 3.2 compares GTFB, TFB, EBAL, and DIM. As expected, GTFB, TFB, and EBAL are unbiased while DIM is clearly biased. TFB and GTFB also have smaller RMSE and therefore higher efficiency than EBAL.

The leftover imbalance after weighting explains the efficiency gain of the TFB methods. Figure 3.3 displays the remaining imbalance in $X^{(l)}$ after applying the weights. Observe that EBAL achieves perfect balance on all covariates. However, despite the fact that all covariates start with positive imbalance, GTFB and TFB both result in a small amount of negative imbalance in $X^{(1)}$. Since the TFB methods seek

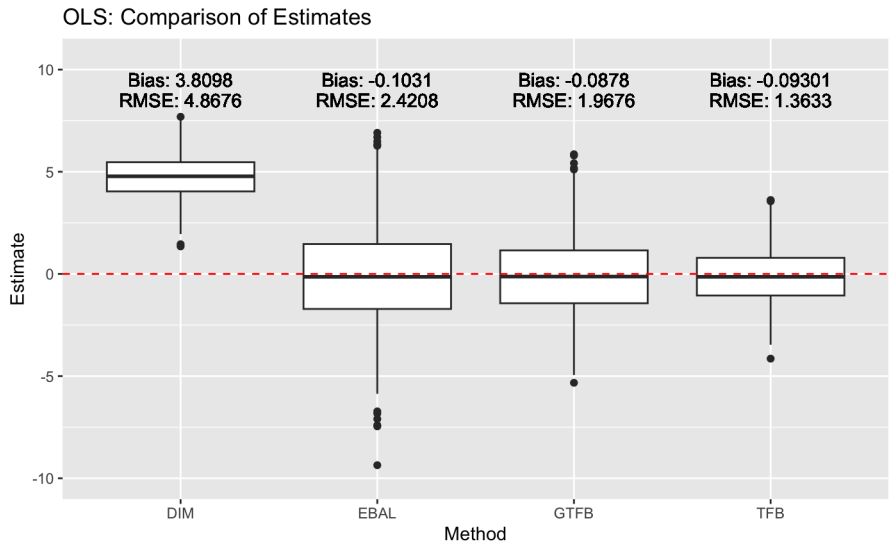
balance on $f_0(X)$ instead of X, this negative imbalance counteracts remaining positive imbalance in the other covariates. For example, suppose the remaining imbalance in $X^{(1)}$ is -0.06, and the remaining imbalance in $X^{(2)}, X^{(3)}$, and $X^{(4)}$ is each 0.2. Then by Equation 3.3,

$$\text{imbal}(w, f_0(X), D) = -0.06 \cdot 10 + 0.2 \cdot 3 = 0.$$

Since it is less challenging to strategically leave imbalances that counteract each other than to achieve exact balance on all covariates, the TFB methods achieve higher efficiency than EBAL.

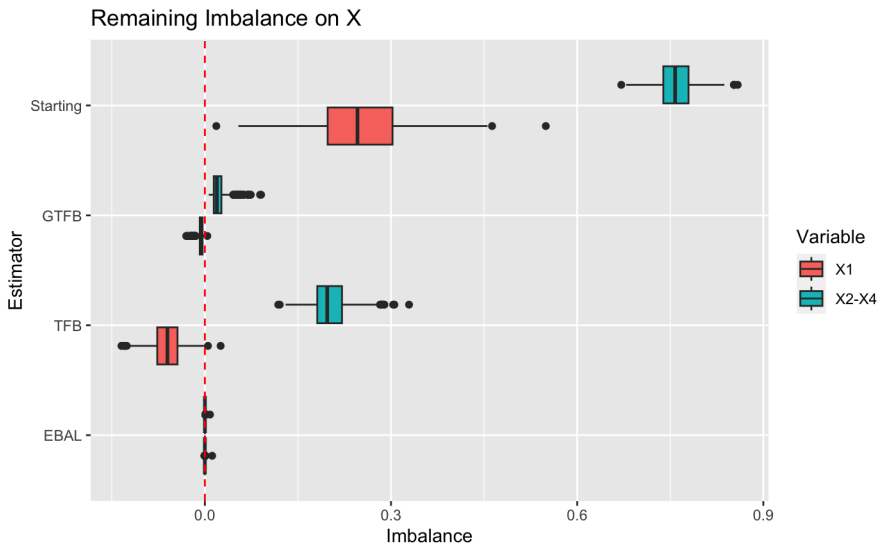
We also observe in Figure 3.2 that TFB is more efficient than GTFB. To explain this difference, recall from Section 2.3 that GTFB places greater emphasis on balancing X than TFB. Accordingly, we observe in Figure 3.3 that TFB leaves greater negative imbalance in $X^{(1)}$ and greater positive imbalance in the other covariates than GTFB. Therefore, TFB takes more advantage of the estimated outcomes, and thus achieves higher efficiency than GTFB.

Figure 3.2: Bias of estimates



(a) This figure compares distributions of ATT estimates by GTFB, TFB, EBAL, and the unweighted DIM. Results were aggregated across 800 draws from DGP 1a with $n = 800$. The bias and RMSE of the methods are also displayed. The horizontal dashed line represents the true causal effect, ATT = 0.

Figure 3.3: Leftover imbalance in X



(a) This figure compares the distributions of remaining imbalance in $X^{(1)}$ and mean remaining imbalance in $X^{(2)}$-$X^{(4)}$ after weighting by TFB, GTFB, and EBAL with the starting imbalance in X in the case of linear f_0. Results were aggregated across 800 draws from DGP 1a with $n = 800$. The vertical dashed line represents zero leftover imbalance, $\text{imbal}(w, X^{(l)}, D) = 0$.

3.2 DGP 1b: GTFB in the GLM setting

This subsection demonstrates the performance of GTFB that uses probit regression. Despite the non-linearity of f_0 as a function of X, GTFB leaves imbalance in a quantity in which f_0 is approximately linear to achieve unbiasedness and high efficiency. For this DGP, generate Y_i as a binary outcome variable using a process similar to that of DGP 1a:

$$p(Y_i = 1|X_i) = \Phi(X_i^{(1)} + \frac{1}{10}\sum_{l=2}^{4} X_i^{(l)} - 0.8), \tag{3.4}$$

where Φ is the CDF of the standard normal distribution. Note that $X_i^{(1)}$ is again most influential in determining Y_i. In this DGP, $Y_i = 1$ for around 50% of the observations. As in DGP 1a, the coefficient of D_i is zero in Equation 3.4, which means that Y_i is not a function of treatment assignment. Therefore, ATT$= 0$.

Since TFB is unavailable for GLMs, we compare the performance of GTFB, EBAL, and the unweighted DIM. For GTFB, we use probit regression to model f_0 and es-

timate the variance of $\hat{f}_0(X)$ with the estimator derived using the Delta method in Section 2.3: $\hat{V}_{\hat{f}_0} = [g']X\hat{V}_{\hat{\beta}_0}X^\top[g']$. Because the specification of f_0 is correct, we expect GTFB to be unbiased. We expect DIM to be biased since it fails to address confounding by X.

We now discuss the expected performance of EBAL. By Zhao & Percival (2017), EBAL is doubly robust, which is to say the estimator is unbiased for the ATT if at least one of two conditions is satisfied:

1. $\pi(X_i)$ is well modeled by logistic regression, (Condition 1)

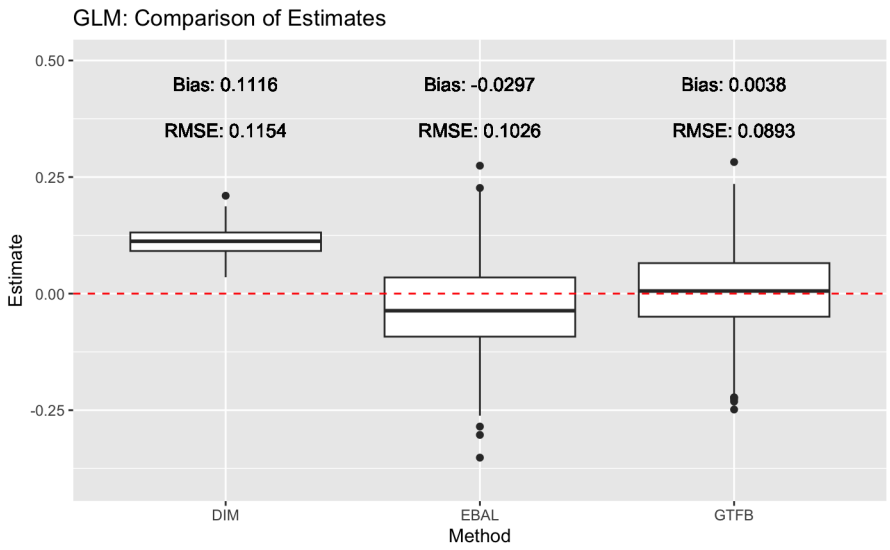
2. $f_0(X_i)$ is a linear function of X_i. (Condition 2)

Recall that in generating treatment for DGP 1, (1) the link function for $\pi(X_i)$ is the probit, and (2) $\psi(X_i)$ is not linear in X_i. This means $\pi(X_i)$ is not well modeled by logistic regression, so Condition (1) is not satisfied. Therefore, in order to guarantee unbiasedness of the EBAL estimate, Y_i must be linear in X_i. Since we generate Y_i using the probit link function, Y_i is not linear in X_i. Additionally, we subtract a small constant (i.e., 0.8) from $X_i^1 + \frac{1}{10}\sum_{l=2}^{4} X_i^{(l)}$ to further reduce the ability of linear models to approximate $f_0(X_i)$ in the interval $(0, 1)$. Thus, Condition (2) is also breached, and we expect EBAL to be biased.

Figure 3.4 shows that GTFB yields unbiased estimates for the ATT. EBAL and DIM are biased. As a consequence of biased, EBAL has a larger RMSE than GTFB.

Figure 3.5a displays the leftover imbalance in X after weighting. For GTFB, we no longer observe substantial negative imbalance in $X^{(1)}$ that we observed in the previous DGP. This absence of negative imbalance can be explained by Taylor expansion. As noted in Section 2.3, we can approximate $f_0(X)$ as a linear function of $g'(X\beta)X$. Therefore, since GTFB seeks balance on $f_0(X)$, we expect negative imbalance in $g'(X\beta)X^{(1)}$ instead of the untransformed $X^{(1)}$. Figure 3.5b confirms this expectation: GTFB leaves a small amount of negative imbalance in $g'(X\beta)X^{(1)}$ to counteract positive imbalance in the other $g'(X\beta)X^{(l)}$'s which are more challenging to balance. EBAL, in contrast, achieves perfect balance on the covariates, but does not attempt to balance $g'(X\beta)X$. Since f_0 is a linear function of $g'(X\beta)X$ and not the covariates, EBAL is biased for the ATT.

Figure 3.4: Bias of Estimates



(a) This figure compares the distributions of ATT estimates by GTFB, EBAL, and the unweighted DIM. The bias and RMSE of the methods are also displayed. Results were aggregated across 800 draws from DGP 1b with $n = 800$. The horizontal dashed line represents the true causal effect, ATT = 0.

3.3 DGP 1c: GFTB in the tree-based model setting

This subsection demonstrates the performance of GTFB in the setting of tree-based machine learning models: boosting and BART, both reviewed in Section 1.5. In this setting, $\hat{f}_0(X_i)$ is not a parametric function of X_i. GTFB again strategically leaves imbalance in the quantity in which $f_0(X_i)$ is linear, yielding unbiasedness and high efficiency.

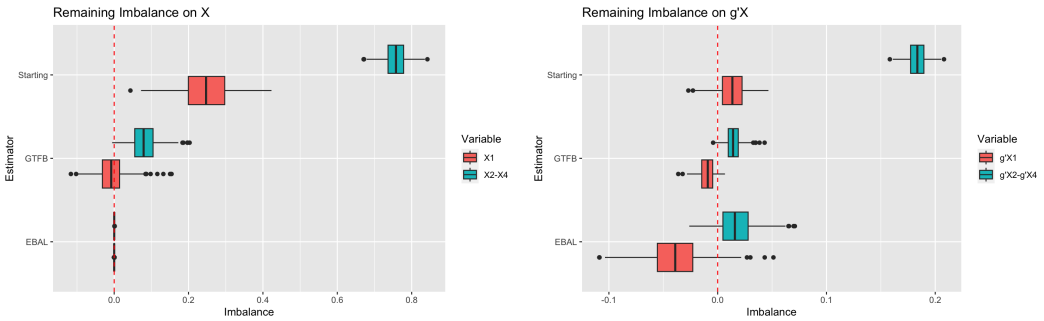
For this DGP, we generate Y_i as a linear function of indicator variables $Z_i^{(l)}$:

$$Y_i = 10Z_i^{(1)} + \sum_{l=2}^{4} Z_i^{(l)} + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} N(0, 10),$$

where

$$Z_i^{(l)} = \begin{cases} 1 & \text{if } X_i^{(l)} > 0, \\ -1 & \text{if } X_i^{(l)} \leq 0 \end{cases}$$

for $l \in \{1, \ldots, 4\}$, and the variance of ϵ_i is chosen so that $R^2 = 0.5$. Since Y_i is



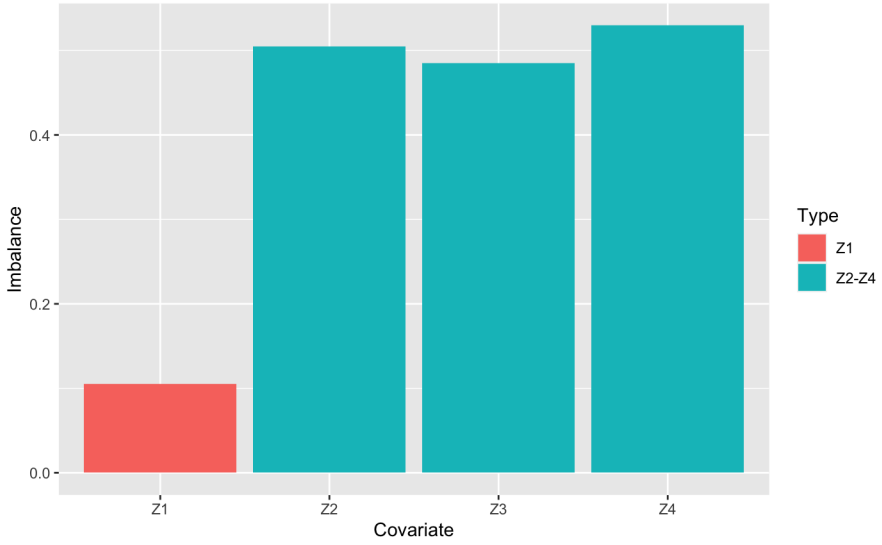
(a) Leftover imbalance in X (b) Leftover imbalance in $g'X$

Figure 3.5: Leftover imbalance

(a) This figure compares the remaining imbalance in X after weighting with the starting imbalance in X. Results were aggregated across 800 draws from DGP 1b with $n = 800$. The vertical dashed line represents the absence of leftover imbalance, $\text{imbal}(w, X^{(l)}, D) = 0$.

(b) This figure compares the remaining imbalance in $g'(X\beta)X$ after weighting and the starting imbalance in $g'(X\beta)X$. Results were aggregated across 800 draws from DGP 1b with $n = 800$. The horizontal dashed line represents the absence of leftover imbalance, $\text{imbal}(w, X^{(l)}, D) = 0$.

Figure 3.7: Starting imbalance of Z



(a) This figure shows the starting imbalance in Z in DGP 1c. The imbalance was computed using one iteration of DGP 1c with $n = 8000$.

now linear in Z_i, the indicator with the highest coefficient, $Z_i^{(1)}$, is now the most influential in determining the outcome. Figure 3.7 shows that $Z^{(1)}$, similar to $X^{(1)}$ in the previous DGPs, is easier to balance than the other indicators. Therefore, despite

the lack of a closed-form variance estimator that allows us to explicitly rewrite the optimization problem to show that GTFB seeks balance in Z, assuming that GTFB models f_0 well, we can still expect that GTFB will leave negative imbalance in $Z^{(1)}$ to offset positive imbalance in the other indicators. Again note that ATT$= 0$.

We use two tree-based machine learning methods to model f_0: boosting and BART. We again compare the performance of GTFB, EBAL, and DIM. We estimate the variance of $\hat{f}_0(X_i)$ with the bootstrap for the boosting model, and the posterior predictive distribution for the BART model. Tree-based machine learning methods model $f_0(X_i)$ as a step function, so f_0 is correctly specified by the models. Therefore, we expect GTFB to be unbiased for the ATT. Again, since $f_0(X_i)$ is no longer linear in X_i and the propensity score is not well modeled by logistic regression, we expect EBAL to be biased. Finally, we expect DIM to be biased since it does not attempt to adjust for the effect of confounding.

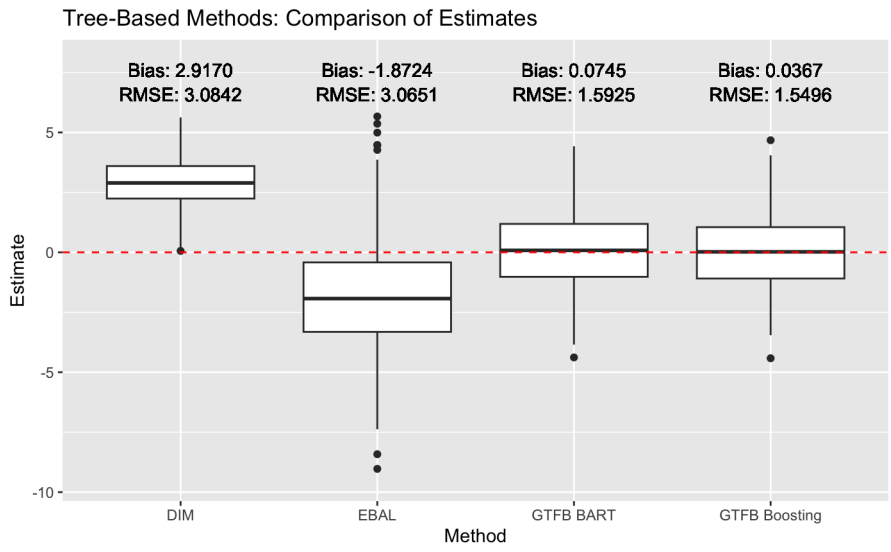
Figure 3.8 confirms these expectations by showing that GTFB, in conjunction with boosting and BART, yields unbiased estimates for the ATT, while EBAL and the unweighted DIM are biased. GTFB using BART and boosting yield very similar bias and variance as each other. With both substantial bias and greater variance, EBAL has greater RMSE than the GTFB methods.

In order to understand the smaller bias of GTFB than EBAL, we study Figure 3.9a, which shows the leftover imbalance in $X^{(l)}$ after weighting. We again do not observe substantial negative imbalance in $X^{(1)}$ for GTFB, which can be explained by the fact that Y_i is a linear function of Z_i instead of X_i. Figure 3.9b shows the leftover imbalance in Z after weighting. Again, despite positive starting imbalance in $Z^{(l)}$ for all $l \in \{1, \cdots, 4\}$, GTFB leaves slight negative imbalance in $Z^{(1)}$ to counteract positive imbalance in the indicators that are more challenging to balance. In particular, GTFB using BART leaves more negative imbalance in $Z^{(1)}$ and more positive imbalance in $Z^{(2)}$-$Z^{(4)}$. This difference may be attributable to the fact that boosting yields greater estimated variance for $\hat{f}_0(X_i)$ than BART since BART prevents overfitting through the use of regularization priors while boosting does not. In contrast to the GTFB methods, EBAL achieves near-perfect balance on X, but does not attempt to balance Z. Since $f_0(X_i)$ is linear in Z_i and not X_i, EBAL is biased for the ATT.

3.4 Variance estimation

This section demonstrates the performance of the variance estimator introduced in Section 2.4 on the three DGPs discussed above. Specifically, we obtain 95% confidence

Figure 3.8: Bias of estimates



(a) This figure compares the distributions of ATT estimates by (1) GTFB using boosting, (2) GTFB using BART, (3) EBAL, and (4) the unweighted DIM. The bias and RMSE of the methods are also displayed. Results were aggregated across 400 draws from DGP 1c with $n = 800$. The horizontal dashed line represents the true causal effect, ATT $= 0$.

intervals and evaluate the coverage probability of these intervals.

Assuming that the distribution of $\hat{\tau}^{\text{GTFB}}_{\text{wdim}}$ is approximately normal, for $\gamma \in (0, 1)$, we can obtain a $\gamma \times 100\%$ confidence interval

$$\hat{\tau}^{\text{GTFB}}_{\text{wdim}} \pm \Phi^{-1}\big(\frac{1+\gamma}{2}\big)\sqrt{\widehat{\text{Var}}(\hat{\tau}^{\text{GTFB}}_{\text{wdim}})},$$

where Φ is the CDF of the standard normal distribution.

Figure 3.11 displays the coverage probability of 95% confidence intervals of $\hat{\tau}^{\text{GTFB}}_{\text{wdim}}$ at five sample sizes for all the DGPs introduced in this section. We choose BART as a representative of the tree-based methods from DGP 1c because boosting, which requires the bootstrap for the variance estimate, is very computationally intensive. We see that coverage is consistently at or close to the nominal 95% when the sample size is 500 or greater.

Since the variance estimator from Section 2.4 relies on the asymptotic uncorrelatedness of estimates from the two sample splits, we check the correlations of $\hat{\tau}^{(s1)}_{\text{wdim}}(\hat{w}^{(s1)}_{\text{GTFB}})$ and $\hat{\tau}^{(s2)}_{\text{wdim}}(\hat{w}^{(s2)}_{\text{GTFB}})$ for different sample sizes to confirm the plausibility of asymptotic uncorrelatedness. Table 3.1 displays the correlations at different sample

(a) Leftover imbalance in X (b) Leftover imbalance in Z

Figure 3.9: Leftover imbalance

(a) This figure compares the remaining imbalance in X after weighting to the starting imbalance in X. Results were aggregated across 400 draws from DGP 1c with $n = 800$. The vertical dashed line represents the absence of leftover imbalance, $\text{imbal}(w, X^{(l)}, d) = 0$.

(b) This figure compares the remaining imbalance in Z after weighting to the starting imbalance in Z. Results were aggregated across 400 draws from DGP 1b with $n = 800$. The horizontal dashed line represents the absence of leftover imbalance, $\text{imbal}(w, Z^{(l)}, d) = 0$.

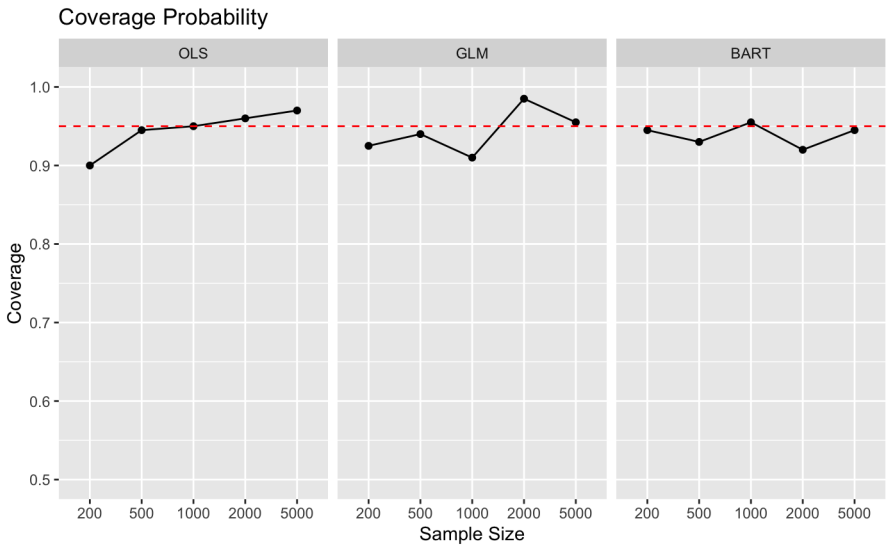
sizes. Observe that for DGP 1a, the correlation decreases as sample size increases, with significance observed only at the smallest sample size. For DGP 1b, the correlation is consistently low at all sample sizes, and there is no significance at any sample size. Table For DGP 1c, the correlation decreases slowly with the increase in sample size, and the correlation is significant at all but the largest sample size. A plausible explanation is that BART, as a tree-based method, does not converge as quickly to the true f_0 as OLS and GLM under correct specification, so the correlation disappears slowly. With satisfactory coverage at all sample sizes for all methods, finite-sample correlation is not of major concern.

Sample Size	**OLS**	**GLM**	**BART**
200	0.1702*	-0.0772	0.1536*
500	0.0119	0.0444	0.3260*
1000	-0.0452	0.0464	0.2183*
2000	0.0330	-0.0719	0.2491*
5000	0.0358	-0.0287	0.1214

Table 3.1: Correlations

(a) This table displays the correlations between $\hat{\tau}^{(s1)}_{\text{wdim}}(\hat{w}^{(s1)}_{\text{GTFB}})$ and $\hat{\tau}^{(s2)}_{\text{wdim}}(\hat{w}^{(s2)}_{\text{GTFB}})$ for GTFB in the settings of OLS, probit regression, and BART, across all five sample sizes considered. For each pair of estimates, we fit an OLS of $\hat{\tau}^{(s2)}_{\text{wdim}}(\hat{w}^{(s2)}_{\text{GTFB}})$ as a function of $\hat{\tau}^{(s1)}_{\text{wdim}}(\hat{w}^{(s1)}_{\text{GTFB}})$. * indicates statistical significance of the coefficient of $\hat{\tau}^{(s1)}_{\text{wdim}}(\hat{w}^{(s1)}_{\text{GTFB}})$ at $\alpha = 0.05$.

Figure 3.11: Coverage



(a) This figure shows the coverage probability of 95% confidence intervals of $\hat{\tau}^{\text{GTFB}}_{\text{wdim}}$ for DGPs 1a, 1b, and 1c. Results include five different sample sizes with 200 draws from each. The horizontal dashed line represents the nominal coverage probability, 95%.

Chapter 4

Applications

This chapter demonstrates the performance of GTFB on two applications: the partially simulated datasets from the 2016 Atlantic Causal Inference Conference (ACIC) competition (Dorie et al., 2019) and the empirical dataset from the National Supported Work (NSW) study (Dehejia & Wahba, 1999). This chapter compares the performance of GTFB to the performance of well-known methods in causal inference.

4.1 2016 ACIC competition

This section evaluates the performance of GTFB using the datasets designed by Dorie et al. (2019). The data originated from "Is Your SATT Where It's At?", a causal data analysis competition that took place during the 2016 Atlantic Causal Inference Conference. The competition served as a comparison study of numerous methods in causal inference in a variety of data settings that were not designed to highlight the strengths of particular methods.

The authors used covariates (X) from the Collaborative Perinatal Project (CPP) (Niswander et al., 1972), a massive longitudinal study on pregnancy and birth in the U.S., to ensure covariate distributions are realistic. The competition data mimics an investigation into the effect of birth weight ($D = 0$ for infants with low weight, $D = 1$ for infants with normal weight) on the infant's IQ (Y). The authors selected a subset of the CPP data that includes $n = 4802$ observations and $p = 58$ covariates that researchers may reasonably include as potential confounders in the proposed context. These covariates include 5 binary variables, 3 categorical variables, 27 count variables, and 3 continuous variables. The authors simulated the treatment assignment and outcomes using 77 different DGPs varying in

- degree of non-linearity of D and Y in X[1],
- percent of treated observations[2],
- overlap between treated and control X[3],
- number of confounders[4],
- treatment effect heterogeneity[5],
- magnitude of treatment effect[6].

The competition data includes 7700 datasets: 100 iterations of the 77 DGPs. Each of the datasets record

1. the treatment assignment: D;
2. the CEF of the potential outcomes: $f_0(X)$ and $f_1(X)$;
3. the potential outcomes: $Y(0)$ and $Y(1)$;
4. the observed outcomes: Y.

The in-sample treatment effect is computed using (2) on the treated group as the unbiased estimator for the ATT. The methods compared estimate the ATT using (4). Note that in real-world settings, only (1) and (4) are observable. Access to (2) in the competition data provides an unbiased estimator for the ATT to which other estimators can be compared, but this is only a convenience brought about by the simulation setting.

[1]Both D and Y could be linear, (second- or third-degree) polynomial, or step functions of X.

[2]The expected percentage of observations that are treated ranges from 35% to 65%.

[3]For each DGP, two possibilities exist: (1) the range of covariate values for the treated group completely overlaps with that of the control group; (2) the range of covariate values for the treated group is a subset of that of the control group, where some observations with extreme covariate values are assigned propoensity scores that are very close of 0, which excludes them from treated group.

[4]In some DGPs, not all potential covariates and transformations of covariates contribute to both D and Y. The covariates (and transformations) that contribute to one of D and Y but not the other are not confounders. The presence of non-confounding covariates highlights the efficacy of methods that select and target confounders.

[5]In generating the outcome, some covariate terms may interact with the treatment to cause treatment effect heterogeneity. In each DGP, treatment effect heterogeneity can be high, low, or non-existent. In DGPs with high treatment heterogeneity, approximately six terms that determine Y interact with D; in DGPs with low heterogeneity, three terms interact with D; in DGPs without heterogeneity, no term interacts with D.

[6]Variation in the above characteristics create variation in the magnitude of the treatment effect.

Ignoring for now the red box, Table (4.1) synthesizes results from (1) the 2016 ACIC competition (Dorie et al., 2019) and (2) a study on optimization-based methods using the ACIC competition data (Cousineau et al., 2023). The figure shows that almost all of the top performing methods make use of BART. These include the BART regression estimator proposed by Hill (2011) and other causal inference methods that incorporate modeling using BART. Important to the subsequent discussion is the **regression estimator** which estimates $\hat{f}_0$ on the control group using a particular model, applies $\hat{f}_0$ to the treated group to obtain predictions $\hat{f}_0(X)$, and takes the mean of $Y_i(1) - \hat{f}_0(X_i)$ for the treated units as an estimate of the ATT.

Due to the apparent suitability of the BART model for the competition data, we use BART to model f_0 in GTFB [7]. Due to limitations on computation time, this thesis tests GTFB on the first five (out of 100) iterations of each of the 77 DGPs. To evaluate the performance of the methods, we follow Cousineau et al. (2023) in standardizing $\hat{\tau}_{\text{wdim}}$ and the treatment effect by dividing the raw quantities by the standard deviation of the outcomes. For the kth DGP, denote with $\hat{\tau}_{k,j}$ the mean estimate of the ATT in the jth iteration, denote with τ_k the treatment effect, and denote with $\text{SD}_k(Y)$ the standard deviation of the outcomes. We compute the following metrics:

$$\text{Bias}_{k,j} = \frac{\hat{\tau}_{k,j} - \tau_k}{\text{SD}_k(Y)} \quad \text{for } k \in \{1, \ldots, 77\},\ j \in \{1, \ldots, 5\},$$

$$\text{RMSE} = \sqrt{\frac{1}{77 \cdot 5} \sum_{k=1}^{77} \sum_{j=1}^{5} \text{Bias}_{k,j}^2}.$$

The red box in Table 4.1 displays the result of GTFB from our investigation. Due to the small bias and RMSE of GTFB, we place the method in the second place out of the 32 methods compared. GTFB also exhibits good coverage probability for the 95% confidence interval, albeit slight over-coverage is observed.

This result provides important support for the performance and applicability of GTFB on a variety of data settings that are not designed to highlight the strengths of GTFB, and confirms that GTFB's performance is on par with other well-known methods in causal inference when they make use of the same model. In particular, this application provides evidence that GTFB performs well compared to the correspond-

[7] The percentage of treated units varies across different datasets, and for some iterations of high-treatment DGPs, below 33% of the observations belong to the control group. In order to ensure optimal performance of the BART model on the data, we build the model on the entire dataset (instead of only building the model on the control group, as was the case in Section 3.3) and include D as a predictor. To obtain predictions, we set all observations to have $D = 0$ so that they correspond to $\hat{f}_0(X)$

ing regression estimator (bart in Table 4.1). However, note that GTFB's improvement on the BART regression estimator is not drastic, so additional justification is needed for choosing GTFB when the simpler and more computationally efficient regression estimator is available. Some potential advantages of GTFB include:

- The coverage probability of confidence intervals by GTFB may be superior.

- The performance of regression estimators relies entirely on the accuracy of the model. As evidenced by the superior performance of BART-related methods, BART models the outcome well in the competition data. However, there is no guarantee that any particular model would perform well in real-world data settings. Therefore, over-reliance on modeling accuracy, exhibited by regression estimators, is inadvisable, and it may be more prudent to use methods that do not rely solely on the model fit. As an example, since EBAL seeks balance on the covariates, its performance does not rely on the accuracy of any outcome model. As a result, recall from Section 3.2 that although EBAL is biased for the ATT when $f_0(X_i)$ is not a linear function of X_i, it improves upon the unweighted DIM. The formulation of GTFB incorporates this philosophy by additionally seeking balance on a quantity in which the outcome is linear to bypass potential modeling error, and thus safeguards against the consequences of poor model fit.

- As a weighting method, GTFB does not extrapolate beyond what is observed in the data. Let $\hat{\tau}_{\text{greatest}}$ denote the greatest observed difference between the average outcome in the treated group and any outcome in the control group, and $\hat{\tau}_{\text{least}}$ denote the least observed difference between the average outcome in the treated group and any outcome in the control group. Under the constraints that the weights must (1) sum to a fixed number (i.e., the number of control units) and (2) be non-negative, the estimate of the ATT necessarily falls into the interval $[\hat{\tau}_{\text{least}}, \hat{\tau}_{\text{greatest}}]$. In contrast, regression estimators may extrapolate beyond the data. As mentioned by Ben-Michael (2020), since extrapolation is associated with high model dependence, models that are misspecified or suffer from high error can lead to unsatisfactory estimates. Additionally, high model dependence implies that the resulting estimator is likely to suffer from high variance.

Figure 4.1: ACIC Results

Method	Bias	SD	RMSE	Coverage
bart_on_pscore	0.001	0.014	0.014	88.4
gtfb_bart	0.001	0.015	0.015	98.2
bart_on_tmle	0.000	0.016	0.016	93.5
mbart_symint	0.002	0.017	0.017	90.3
bart_mchains	0.002	0.017	0.017	85.7
bart_xval	0.002	0.017	0.017	81.2
bart	0.002	0.018	0.018	81.1
sl_bart_tmle	0.003	0.029	0.029	91.5
h2o_ensemble	0.007	0.029	0.030	100.0
bart_iptw	0.002	0.032	0.032	83.1
sl_tmle	0.007	0.032	0.032	87.6
superlearner	0.006	0.038	0.039	81.6
calcause	0.003	0.043	0.043	81.7
tree_strat	0.022	0.047	0.052	87.4
balanceboost	0.020	0.050	0.054	80.5
adj_tree_strat	0.027	0.068	0.074	60.0
lasso_cbps	0.027	0.077	0.082	30.5
sl_tmle_joint	0.010	0.101	0.102	58.9
kbal	0.036	0.083	0.091	–
balancehd	0.041	0.099	0.107	–
cbps	0.041	0.099	0.107	99.7
sbw	0.041	0.102	0.110	–
cbps_exact	0.041	0.105	0.112	–
ebal	0.041	0.110	0.117	–
teffects_psmatch	0.043	0.099	0.108	47.0
linear_model	0.045	0.127	0.135	22.3
mhe_algorithm	0.045	0.127	0.135	22.8
teffects_ra	0.043	0.133	0.140	37.5
cbps_over	0.044	0.117	0.125	–
teffects_ipwra	0.044	0.161	0.166	35.3
genmatch	0.052	0.141	0.151	–
teffects_ipw	0.042	0.298	0.301	39.0

(a) This table synthesizes results from Dorie et al. (2019), Cousineau et al. (2023), and this thesis to produce an overall ranking. Methods in blue boxes make use of BART. The red box shows the results of GTFB that uses BART to model f_0, which is tested on five iterations of each of the 77 DGPs. Some methods do not have associated coverage probabilities because Cousineau et al. (2023) do not include coverage in their analysis.

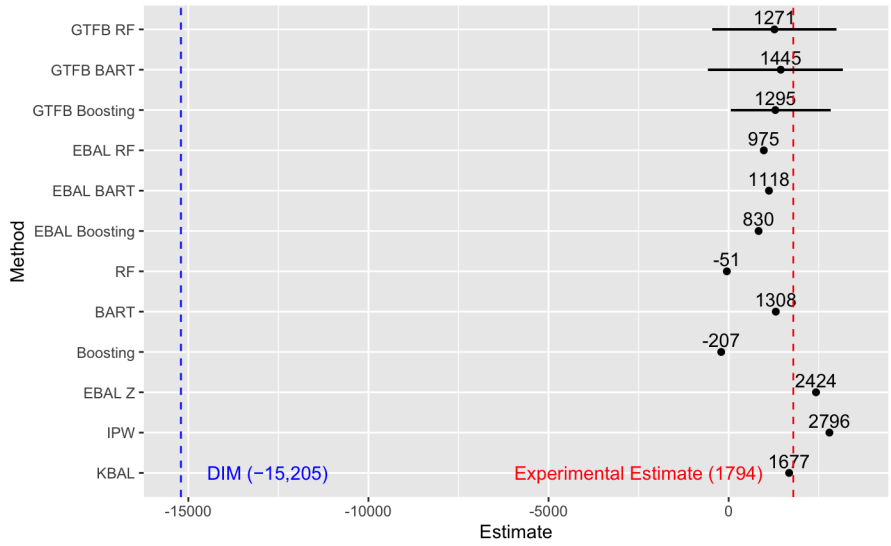
4.2 National Supported Work

This section evaluates the performance of GTFB on the National Supported Work (NSW) dataset. The NSW study, which took place in the mid 1970s, provided a subsidized labor training program as the treatment to workers who required employment assistance. The recorded outcome is earnings in 1978.

The NSW dataset is well-suited for testing causal inference methods because it includes an RCT of eligible candidates and a non-experimental comparison group of ineligible workers. Therefore, the data provides

1. an unbiased experimental estimate, which is a difference in means between the

Figure 4.2: NSW estimates



(a) This figure compares the ATT estimates for the NSW dataset by 12 methods: GTFB using RF, BART, and boosting to model f_0; EBAL on $\hat{f}_0$ using RF, BART, and boosting; the RF, BART, and boosting regression estimators; EBAL on the covariates; KBAL; and IPW. All models use the 10 untransformed covariates from the dataset. The reported GTFB estimates are the median estimates from 100 random splits. We estimate the variance of GTFB using Equation (4.1) and find the 95% confidence intervals using the normal approximation.

experimental treated group ($n_t = 185$) and the experimental control group ($n_c = 260$). This unbiased estimate serves as a benchmark with which the estimates obtained from methods of interest can be compared;

2. an "observational study" consisting of the experimental treated group ($n_t = 185$) and the non-experimental comparison group ($n_c = 2490$), where causal inference methods can be applied. We test GTFB on this "observational study".

We use all 10 observed covariates in the analysis without transformation: age, years of education, marital status, indicator for holding a high school degree, indicator for being Black, indicator for being Hispanic, earnings in 1974 and 1975, and employment status in the same years.

The experimental estimate from the RCT is \$1794, suggesting that the labor training program was effective in increasing the participants' income, while the naive

Covariate	RCT Treated	RCT Control	Non-RCT Comparison
Sample size	185	260	2490
Age	25.82 (7.1550)	25.05 (7.0577)	34.85 (10.4408)
Education	10.35 (2.0107)	10.09 (1.6143)	12.12 (3.0824)
No degree	0.71 (0.1145)	0.83 (0.1393)	0.31 (0.0133)
Married	0.19 (0.0355)	0.15 (0.0264)	0.87 (0.0510)
Black	0.84 (0.1705)	0.83 (0.1356)	0.25 (0.0116)
Hispanic	0.06 (0.0185)	0.11 (0.0215)	0.03 (0.0037)
Revenue 74	2095.57 (4886.62)	2107.03 (5687.91)	19428.75 (13406.88)
Revenue 75	1532.06 (3219.25)	1266.01 (3102.98)	19063.34 (13596.95)
Employment 74	0.29 (0.0472)	0.25 (0.0358)	0.91 (0.0652)
Employment 75	0.4 (0.0600)	0.32 (0.0421)	0.9 (0.0601)

Table 4.1: Covariate means

(a) This table displays the covariate means in the experimental treated, experimental control, and non-experimental comparison groups in the NSW data. Standard deviations of the covariates are displayed in parentheses next to the covariate means.

difference in means estimate from the observational study is -$15,205, indicating that participants of the labor training program earn less compared to non-participants after the program. The dramatic difference between these two estimates can be explained by the fact that the participants in the RCT are sampled from the target population of the subsidized labor training, a population of workers who experienced persistent employment difficulty. Therefore, eligible candidates can reasonably be expected to have substantially lower income than workers in the non-experimental comparison group who are ineligible for the NSW program, and subsidized work may not be able to completely bridge the gap even with an increase in income due to the intervention. Table 4.1 confirms this suspicion by comparing the mean covariate values in the experimental treated, experimental control, and non-experimental comparison groups. Observe that the means of all covariates are similar between the experimental treated and control groups, reflecting successful randomization in the RCT. The covariate means of these groups suggest lower socio-economic status than the comparison group: the eligible candidates are younger, have fewer years of education, are less likely to hold high school degrees and be married, are more likely to belong to a racial or ethnic minority, earn less during 1974 and 1975, and are more likely to be unemployed during these years.

We compare the performance of 12 methods:

- GTFB using RF, BART, and boosting to model f_0;

- EBAL on $\hat{f}_0(X)$ using RF, BART, and boosting;
- the RF, BART, and boosting regression estimators;
- EBAL on the covariates, KBAL, and IPW.

Note that EBAL on $\hat{f}_0(X)$ equates the mean $\hat{f}_0(X)$ and regression estimators take the mean difference between the predicted and actual potential outcomes for the treated group. All models are fit uding the 10 untransformed covariates.

We use 100 random splits of the NSW dataset to obtain estimates from GTFB and EBAL on $\hat{f}_0(X)$. Following the advice from Chernozhukov et al. (2018), we report the median of the estimates as final. Denote with $\hat{\tau}^{\text{GTFB}}_{\text{wdim},r}$ the ATT estimate by GTFB in the rth random split of the data, and denote with $\hat{\tau}^{\text{GTFB}}_{\text{wdim,med}}$ the median of the estimates. Again following Chernozhukov et al. (2018), we estimate the variance of $\hat{\tau}^{\text{GTFB}}_{\text{wdim,med}}$ as follows:

$$\text{Var}(\hat{\tau}^{\text{GTFB}}_{\text{wdim,med}}) = \text{median}\Big(\text{Var}(\hat{\tau}^{\text{GTFB}}_{\text{wdim},r}) + \frac{(\hat{\tau}^{\text{GTFB}}_{\text{wdim},r} - \hat{\tau}^{\text{GTFB}}_{\text{wdim,med}})^2}{n}\Big). \tag{4.1}$$

Figure 4.2 demonstrates the performance of the methods compared. For each outcome model, the GTFB method yields an ATT estimate that is closer to the experimental benchmark than (1) EBAL on $\hat{f}_0(X)$ and (2) the regression estimator. The 95% confidence intervals of all GTFB methods encompass the experimental estimate and estimates by commonly used causal inference methods: KBAL, IPW, and EBAL on the covariates.

The key message from this application is that GTFB improves substantially upon the naive DIM estimator, yielding estimates that fall roughly into the same neighborhood as the well-known causal inference methods. While evaluating the performance of estimators in comparison to the experimental benchmark, it is important to recognize that while the experimental estimate is unbiased for the true ATT, the available data constitutes only one realization of the data generating process, so the estimate need not represent the ATT numerically. Therefore, estimators that are closest to the experimental benchmark are not necessarily also closest to the true ATT value. Thus, GTFB's proximity to the experimental benchmark is a promising but inconclusive result, and the fact that other estimators may be closer is not cause for concern.

Conclusion

TFB, proposed by Wainstein (2022), is a covariate balancing weight method for estimating the causal effect of a binary intervention on an outcome in the setting of observational studies. TFB linearly regresses the outcome on the covariates, and seeks balance in functions near the resulting regression model. The original formulation of TFB requires these models to have linear representations, and this thesis extends TFB to GTFB, which allows the use of any predictive model. GTFB models the outcome as a function of the covariates, and seeks balance on the predicted outcomes and a quantity in which the outcome is assumed to be linear. GTFB performs well on simulated and real data in conjunction with a variety of parametric and non-parametric models.

4.1 Future work

Some future directions related to GTFB include:

- Examining the difference between GTFB and EBAL on the predicted outcomes. Recall that EBAL on the predicted outcomes perform very similarly to the corresponding GTFB methods in the NSW application in Section 4.2. EBAL on the predicted outcomes also yield very similar ATT estimates as GTFB in demonstrations not shown in this thesis. EBAL on the predicted outcomes, which achieves exact balance on the predicted outcomes, constitutes one of the three terms in the GTFB optimization problem. Observing a difference in performance between GTFB and EBAL on the predicted outcomes would demonstrate the functionality of GTFB as a whole. Since EBAL on the predicted outcomes does not attempt to safeguard against modeling error, the key to observing the difference may be studying cases where the outcome model is misspecified or has high error;

- Comparing the performance between GTFB using a specific model to predict

the outcome and the regression estimator using the same model. Comparison across different data settings helps inform the choice between GTFB and the regression estimators;

- Studying the effect of varying the degrees of freedom related to the Squidward Constant. The Squidward Constant differentiates GTFB in the linear setting from the original TFB and is important in determining the trade-off between balancing the predicted outcomes and safeguarding against modeling error. Section 3.1 demonstrates that a smaller degree of freedom increases the efficiency of the weighted difference in means estimator in the setting where the outcome is linear in the covariates. Therefore, it would be worthwhile to study the effect of lowering the Squidward Constant in different DGPs to inform the choice of a sensible constant in practical settings;

- Using more types of models to predict the outcome for GTFB. The strength of GTFB lies in its general applicability to different models, and testing the performance of GTFB in conjunction with a wider variety of models helps establish the utility of the generalization. Some worthwhile models include: neural networks, generalized kernel regularized least squares, Bayesian models, and more types of GLMs.

Appendix A

Estimating $V_{\hat{f}_0}$ in the GLM setting

This section of the appendix derives a closed-form variance estimator of $\hat{f}_0(X)$ in the setting of GLMs using the Delta method. We follow the convention associated with the Delta method and write $\theta = \beta$, $h(\theta) = g(X\beta)$. Impose the working assumption that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma). \tag{A.1}$$

Denote the variance estimator of $\hat{\theta}$ with $\hat{V}_{\hat{\theta}}$, Equation A.1 implies that

$$\hat{V}_{\hat{\theta}} \approx \frac{\Sigma}{n}. \tag{A.2}$$

We now apply the Delta method and derive that

$$\begin{aligned} h(\hat{\theta}) &\approx h(\theta) + \frac{\partial h(\theta)}{\partial \theta}(\hat{\theta} - \theta) \\ h(\hat{\theta}) - h(\theta) &\approx \frac{\partial h(\theta)}{\partial \theta}(\hat{\theta} - \theta) \\ \sqrt{n}(h(\hat{\theta}) - h(\theta)) &\approx \sqrt{n}\frac{\partial h(\theta)}{\partial \theta}(\hat{\theta} - \theta). \end{aligned}$$

Therefore, by the same logic as Equation A.1,

$$\sqrt{n}(h(\hat{\theta}) - h(\theta)) \xrightarrow{d} N(0, \frac{\partial h(\theta)}{\partial \theta}\Sigma\frac{\partial h(\theta)}{\partial \theta}^{\top}).$$

Denote with $\hat{V}_{h(\hat{\theta})}$ the variance estimator for $h(\hat{\theta})$. By the same logic as Equation A.2, we can approximate $\hat{V}_{h(\hat{\theta})}$ as

$$\hat{V}_{h(\hat{\theta})} \approx \frac{1}{n}\frac{\partial h(\theta)}{\partial \theta}\Sigma\frac{\partial h(\theta)}{\partial \theta}^\top \approx \frac{\partial h(\theta)}{\partial \theta}\hat{V}_{\hat{\theta}}\frac{\partial h(\theta)}{\partial \theta}^\top.$$

In order to obtain a closed form estimator for $\hat{V}_{h(\hat{\theta})}$, we now compute $\frac{\partial h(\theta)}{\partial \theta}$:

$$\begin{aligned}
\frac{\partial h(\theta)}{\partial \theta} &= \frac{\partial g(X\beta)}{\partial \beta} \\
&= \begin{bmatrix} \frac{\partial g(X_1^\top\beta)}{\partial \beta_1} & \frac{\partial g(X_1^\top\beta)}{\partial \beta_2} & \dots & \frac{\partial g(X_1^\top\beta)}{\partial \beta_p} \\ \dots & & & \\ \frac{\partial g(X_n^\top\beta)}{\partial \beta_1} & \frac{\partial g(X_n^\top\beta)}{\partial \beta_2} & \dots & \frac{\partial g(X_n^\top\beta)}{\partial \beta_p} \end{bmatrix} \\
&= \begin{bmatrix} g'(X_1^\top\beta)X_1^{(1)} & \dots & g'(X_n^\top\beta)X_1^{(p)} \\ \dots & & \\ g'(X_n^\top\beta)X_1^{(1)} & \dots & g'(X_n^\top\beta)X_n^{(p)} \end{bmatrix} \\
&= \begin{bmatrix} g'(X_1^\top\beta) & 0 & \dots & 0 \\ 0 & g'(X_1^\top\beta) & \dots & 0 \\ \dots & & & \\ 0 & \dots & 0 & g'(X_n^\top\beta) \end{bmatrix} X \\
&= [g']X,
\end{aligned}$$

where

$$[g'] = \begin{bmatrix} g'(X_1^\top\beta) & 0 & \dots & 0 \\ 0 & g'(X_2^\top\beta) & \dots & 0 \\ \dots & & & \\ 0 & \dots & 0 & g'(X_n^\top\beta) \end{bmatrix}.$$

Now compute $\hat{V}_{\hat{f}_0}$:

$$\begin{aligned}
\hat{V}_{\hat{f}_0} &= \hat{V}_{h(\hat{\theta})} \\
&\approx \frac{\partial h(\theta)}{\partial \theta}\hat{V}_{\hat{\theta}}\frac{\partial h(\theta)}{\partial \theta}^\top \\
&= [g']X\hat{V}_{\hat{\beta}}X^\top[g'].
\end{aligned}$$

Appendix B

Deriving the MSE of $\hat{\tau}_{\text{wdim}}$ for the SATT

In this section of the appendix we estimate the Mean Squared Error (MSE) of $\hat{\tau}_{\text{wdim}}$ for the SATT. Since w is the quantity of interest, we treat it as a constant throughout the derivation. Recall from Section 1.4 that

$$\hat{\tau}_{\text{wdim}} - \text{SATT} = \text{imbal}(w, f_0(X), D) + \big(\frac{1}{n_t}\sum_{i:D_i=1}\epsilon_i(0) - \frac{1}{n_c}\sum_{i:D_i=0} w_i\epsilon_i(0)\big),$$

so we can write

$$
\begin{aligned}
&\mathbb{E}\Big(\big(\hat{\tau}_{\text{wdim}}(w) - \text{SATT}\big)^2 \;\Big|\; X, D\Big) \\
&= \mathbb{E}\Big[\Big(\text{imbal}(w, f_0(X), D) + \big(\frac{1}{n_t}\sum_{i:D_i=1}\epsilon_i(0) - \frac{1}{n_c}\sum_{i:D_i=0} w_i\epsilon_i(0)\big)\Big)^2 \;\Big|\; X, D\Big] \\
&= \mathbb{E}[\text{imbal}(w, f_0(X), D)^2 \mid X, D] \\
&\quad + 2\mathbb{E}\big[\text{imbal}(w, f_0(X), D)\big(\frac{1}{n_t}\sum_{i:D_i=1}\epsilon_i(0) - \frac{1}{n_c}\sum_{i:D_i=0} w_i\epsilon_i(0)\big) \;\Big|\; X, D\big] \\
&\quad + \mathbb{E}\big[\big(\frac{1}{n_t}\sum_{i:D_i=1}\epsilon_i(0) - \frac{1}{n_c}\sum_{i:D_i=0} w_i\epsilon_i(0)\big)^2 \;\Big|\; X, D\big] \\
&= \text{imbal}(w, f_0(X), D)^2 \\
&\quad + 2\big(\text{imbal}(w, f_0(X), D)\big)\mathbb{E}\big[\big(\frac{1}{n_t}\sum_{i:D_i=1}\epsilon_i(0) - \frac{1}{n_c}\sum_{i:D_i=0} w_i\epsilon_i(0)\big) \;\Big|\; X, D\big] \\
&\quad + \mathbb{E}\big[\big(\frac{1}{n_t}\sum_{i:D_i=1}\epsilon_i(0)\big)^2\big] - 2\mathbb{E}\big[\big(\frac{1}{n_t}\sum_{i:D_i=1}\epsilon_i(0)\big)\big(\frac{1}{n_c}\sum_{i:D_i=0} w_i\epsilon_i(0)\big) \;\Big|\; X, D\big] \\
&\quad + \mathbb{E}\big[\big(\frac{1}{n_c}\sum_{i:D_i=0} w_i\epsilon_i(0)\big)^2 \;\Big|\; X, D\big].
\end{aligned}
$$

Assuming that observations are independent, for all $i \neq j$,

$$\mathbb{E}\big[\epsilon_i(0)\epsilon_j(0)|X, D\big] = \mathbb{E}\big[\epsilon_i(0)|X, D\big]\mathbb{E}\big[\epsilon_j(0)|X, D\big].$$

By conditional ignorability, we have that

$$\mathbb{E}[\epsilon_i(0)|X, D] = \mathbb{E}[\epsilon_i(0)|X] = 0,$$

so we know that

$$\mathbb{E}\big[\epsilon_i(0)\epsilon_j(0)|X, D\big] = 0.$$

Further, recall from Section 1.4 that

$$
\begin{aligned}
\mathbb{E}\big[w_i\epsilon_i(0)|X, D\big] &= \mathbb{E}[w_i|X, D]\mathbb{E}[\epsilon_i(0)|X, D] \\
&= \mathbb{E}[w_i|X, D] \cdot 0 \\
&= 0.
\end{aligned}
$$

Thus, we can further simplify the MSE of $\hat{\tau}_{\text{wdim}}$:

$$
\begin{aligned}
&\mathbb{E}\Big(\big(\tau_{\text{wdim}}(w) - \text{SATT}\big)^2 \;\Big|\; X, D\Big) \\
&= \text{imbal}(w, f_0(X), D)^2 + 2\big(\text{imbal}(w, f_0(X), D)\big) \cdot 0 \\
&\quad + \mathbb{E}\big[\frac{1}{n_t^2} \sum_{i:D_i=1} \epsilon_i^2(0) \;\Big|\; X, D\big] - 2\mathbb{E}\big[\big(\frac{1}{n_t} \sum_{i:D_i=1} \epsilon_i(0)\big)\big(\frac{1}{n_c} \sum_{i:D_i=0} w_i \epsilon_i(0)\big) \;\Big|\; X, D\big] \\
&\quad + \mathbb{E}\big[\frac{1}{n_c^2} \sum_{i:D_i=0} w_i^2 \epsilon_i^2(0) \;\Big|\; X, D\big] \\
&= \text{imbal}(w, f_0(X), D)^2 + \mathbb{E}\big[\frac{1}{n_t^2} \sum_{i:D_i=1} \epsilon_i^2(0) \;\Big|\; X, D\big] - 0 \\
&\quad + \mathbb{E}\big[\frac{1}{n_c^2} \sum_{i:D_i=0} w_i^2 \epsilon_i^2(0) \;\Big|\; X, D\big] \\
&= |\text{imbal}(w, f_0(X), D)|^2 + \mathbb{E}\big[\frac{1}{n_t^2} \sum_{i:D_i=1} \epsilon_i^2(0) \;\Big|\; X, D\big] + \mathbb{E}\big[\frac{1}{n_c^2} \sum_{i:D_i=0} w_i^2 \epsilon_i^2(0) \;\Big|\; X, D\big].
\end{aligned}
$$

Note that $\mathbb{E}\big[\frac{1}{n_t^2} \sum_{i:D_i=1} \epsilon_i^2(0) \;\Big|\; X, D\big]$ is not a function of w. Further recall from Section 1.1.1 that $\text{Var}\big(\epsilon_i(0)|X\big) = \sigma_i^2(0)$. Therefore, we can write

$$
\mathbb{E}\big[\epsilon_0^2(0) \;\Big|\; X, D\big] = \mathbb{E}\big[\epsilon_i^2(0) \;\Big|\; X, D\big] - \mathbb{E}\big[\epsilon_i(0) \;\Big|\; X, D\big]^2 = \text{Var}\big(\epsilon_i(0)\big) = \sigma_i^2(0).
$$

Thus, we can ultimately approximate the MSE as

$$
\begin{aligned}
&\mathbb{E}\Big(\big(\tau_{\text{wdim}}(w) - \text{SATT}\big)^2 \;\Big|\; X, D\Big) \\
&= |\text{imbal}(w, f_0(X), D)|^2 + \frac{1}{n_c^2} \sum_{i:D_i=0} w_i^2 \sigma_i^2(0) + C,
\end{aligned}
$$

where C is not a function of w.

References

Ben-Michael, E. E. (2020). *Why Weight?: Weighting Approaches for Causal Inference with Panel and Cross-Sectional Data*. University of California, Berkeley.

Chambers, E. A., & Cox, D. R. (1967). Discrimination between alternative binary response models. *Biometrika*, *54*(3-4), 573–578.

Chang, Q., & Goplerud, M. (2022). Generalized kernel regularized least squares. *arXiv preprint arXiv:2209.14355*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Cousineau, M., Verter, V., Murphy, S. A., & Pineau, J. (2023). Estimating causal effects with optimization-based methods: A review and empirical comparison. *European Journal of Operational Research*, *304*(2), 367–380.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, *94*(448), 1053–1062.

Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, *20*(1), 25–46.

Hainmueller, J., & Hazlett, C. (2014). Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis*, *22*(2), 143–168.

Hazlett, C. (2020). Kernel balancing. *Statistica Sinica*, *30*(3), 1155–1189.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240.

Hillenburg, S., Tibbitt, P., Cecarelli, M., & Waller, V. (1999-Present). SpongeBob SquarePants [TV series]. United Plankton Picturesuction Company, Nickelodeon Animation Studios, and Rough Draft Studiosuction Company.

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning*, vol. 112. Springer.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, (pp. 604–620).

Niswander, K. R., Gordon, M. J., & Gordon, M. (1972). *The women and their pregnancies: the Collaborative Perinatal Study of the National Institute of Neurological Diseases and Stroke*, vol. 73. National Institute of Health.

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Splawa-Neyman, J., Dabrowska, D. M., & Speed, T. P. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, (pp. 465–472).

Wainstein, L. (2022). Targeted function balancing. *arXiv preprint arXiv:2203.12179*.

Zhao, Q., & Percival, D. (2017). Entropy balancing is doubly robust.