Machine Learning and Causal Inference

Theory, Examples, and Computational Results

by

Andrew Herren

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2023 by the
Graduate Supervisory Committee:

P. Richard Hahn, Chair
Ming-Hung Kao
Hedibert Lopes
Robert McCulloch
Shuang Zhou

ARIZONA STATE UNIVERSITY

May 2023

ABSTRACT

This dissertation covers several topics in machine learning and causal inference. First, the question of "feature selection," a common byproduct of regularized machine learning methods, is investigated theoretically in the context of treatment effect estimation. This involves a detailed review and extension of frameworks for estimating causal effects and in-depth theoretical study. Next, various computational approaches to estimating causal effects with machine learning methods are compared with these theoretical desiderata in mind. Several improvements to current methods for causal machine learning are identified and compelling angles for further study are pinpointed. Finally, a common method used for "explaining" predictions of machine learning algorithms, SHAP, is evaluated critically through a statistical lens.

ACKNOWLEDGMENTS

To my family — Mom, Dad, Greg, and Laura. You all mean the world to me. You've carried me in more ways than you could ever know. Thank you for being rock-solid when I need support and for being awesome people when I just want to hang. My Ph.D. journey owes so much to all of you.

To my advisor, Richard. I remember talking to you about research when I was applying to graduate school and thinking, "he seems like someone I could work well with." Thank you for making that statement true, and for patiently teaching me so much of what I now know about statistics. I look forward to our continued collaboration.

To my graduate committee — Drs. Kao, Lopes, McCulloch, and Zhou. I've learned so much from you, in the classroom, in seminars, by reading your work, or through your feedback on my work. Thank you for making the statistics group at ASU a stellar research community, and for making me genuinely excited to present my dissertation. I am lucky to have each of you on my committee.

To my close friends and colleagues at ASU — Demetri, Bryce, Chelsea, Nikolay, and Sam — thank you for being you. You've all left a mark on me and have made my experience in graduate school amazing. I can't wait to see what the future holds for all of you.

This work was partially supported by NSF Grant DMS-1502640.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

This dissertation is an exploration of many themes around machine learning and causal inference. Chapter 2 reviews, and draws connections between, several frameworks for estimating causal effects. These formalisms help to establish principles for estimating average treatment effects when coarsening the available adjustment set.

Chapter 3 derives a "minimal" adjustment set for estimating the average treatment effect, which we call $\lambda(X)$. We then show in precise terms when it is desirable to reduce variance by "refining" this minimal adjustment set. We discuss the problem of bias in feature selection, in particular the notion of "regularization-induced confounding" (Hahn *et al.* (2020)) for treatment effect estimation. We show that certain estimation and regularization conventions can mitigate this problem.

Chapter 4 is a thorough empirical investigation of machine-learning-based causal effect estimators. We review Accelerated Bayesian Causal Forests (XBCF, Krantsevich *et al.* (2022)), Generalized Random Forests (GRF, Athey *et al.* (2019)), and several "meta-algorithms" (Künzel *et al.* (2019), Kennedy (2022)). We articulate our principles of simulation-driven methods development, in which simulations are used to "stress test" estimators, to identify weak points or failure modes, and to develop solutions. We combine this empirical strategy with the theoretical insights of Chapter 3 to design simulation experiments and to propose updates to some of the algorithms mentioned above.

Chapter 5 investigates the theme of "machine learning explainability," a field of research that dovetails with many concepts in causal inference. We discuss a method called SHAP for scoring machine learning predictions according to features' contri-

bution to a prediction. We argue that the intended applications of this explanation algorithm are not always well-served by the numerical output of the algorithm and that "counterfactual" approaches to explaining machine learning models deserve more attention and research.

Chapter 2

MEAN REGRESSION FOR CAUSAL EFFECT ESTIMATION

2.1 Notation

Let $Y \in \mathbb{R}$ be a continuous outcome variable, and $Z \in \{0, 1\}$ be a binary treatment variable. X is a p-vector of covariates defined on vector space $\mathcal{X}$. We are interested in the *causal* effect of Z on Y. In order to discuss this causal effect rigorously, we must introduce several frameworks for performing causal inference, from which we can formalize the estimand of interest. We refer to observations with $Z = 1$ as the *treated* group and we refer observations with $Z = 0$ as the *control* or *untreated* group.

2.2 Three Frameworks for Formalizing Causal Inference

2.2.1 Potential Outcomes

The "Potential Outcomes" framework for causal inference is most closely associated with Jerzy Neyman and Donald Rubin and reviewed in depth in Imbens and Rubin (2015). Define counterfactual random variables Y^0 and Y^1 in which Z is counterfactually fixed at $Z = 0$ and $Z = 1$, respectively, but the distributions of X and the rest of the data are unchanged. The distributions of Y^1 and Y^0 are not generally equivalent to $Y \mid Z = 0$ and $Y \mid Z = 1$ because X may influence both Y and Z so that $X \mid Z = 1$ and $X \mid Z = 0$ have different distributions from X.

The *average treatment effect* is formally defined using Potential Outcomes as

$$\bar{\tau} = \mathbb{E}\left[Y^1 - Y^0\right]$$

or the expected difference between the treated and control potential outcomes. We

will discuss the identifying assumptions that enable unbiased estimation of this counterfactual estimand later in this chapter, but first we review two other "frameworks" for inferring causal effects.

2.2.2 Causal Directed Graphical Models

The causal graphical model perspective is most commonly associated with the work of Judea Pearl and collaborators and is summarized in great detail in Pearl (2009). For an intuitive overview focused on causal effect estimation, we refer readers to Shalizi (2021). We introduced X above as a p-vector of covariates, but we can also split the vector into its p component random variables and write $X = \{X_1, \dots X_p\}$. The joint probability distribution of (Y, Z, X) may be written compositionally as

$$\mathrm{P}(Y, Z, X_1, \dots, X_p) = \mathrm{P}(Y \mid Z, X_1, \dots, X_p)\mathrm{P}(Z \mid X_1, \dots, X_p)\mathrm{P}(X_1, \dots, X_p).$$

$\mathrm{P}(X_1, \dots, X_p)$ can be further decomposed (non-uniquely) as

$$\mathrm{P}(X_1 \mid X_2, \dots, X_p)\mathrm{P}(X_2 \mid X_3, \dots, X_p) \dots \mathrm{P}(X_p)$$

or

$$\mathrm{P}(X_p \mid X_1, \dots, X_{p-1})\mathrm{P}(X_{p-1} \mid X_1, \dots, X_{p-2}) \dots \mathrm{P}(X_1)$$

or any other compositional arrangement of the covariates $\{X_1, \dots, X_p\}$.

For any given composition of $\mathrm{P}(Y, Z, X_1, \dots, X_p)$, it may be the case that some of the conditional probability terms simplify due to conditional independence, which allows that $\mathrm{P}(A \mid B, C) = \mathrm{P}(A \mid B)$ if $A \perp\!\!\!\perp C \mid B$. A compositional probability distribution can be expressed as a *directed graph* $\mathcal{G}$, with nodes $\{Y, Z, X_1, \dots, X_p\}$ and directed edges extending from one node A to another node B if A appears in the conditioning set of the probability distribution of B.

The terminology of directed graphical models mirrors that of genealogy, where nodes $A, B, C, \dots$ that appear in the conditioning set of the probability distribution

of another node T are *parents* of node T. Any nodes that can reach T through available directed paths are *ancestors* of T. Similarly, nodes reachable by a single directed edge from node U are *children* of U and nodes reachable by any directed path from U are *descendants* of U.

This representation of probability distributions by directed acyclic graphs (DAGs), reviewed in great depth by Koller and Friedman (2009), is entirely agnostic of causal relationships. We now define a *causal directed acyclic graph* (CDAG) as a directed graphical model in which every directed edge represents a causal relationship. Pearl (2009) highlights three patterns of connection among sets of three nodes in a CDAG, which we introduce here for later discussion.

- *Chain*: $A \longrightarrow B \longrightarrow C$
- *Fork*: $A \longleftarrow B \longrightarrow C$
- *Collider*: $A \longrightarrow B \longleftarrow C$

Finally, before we discuss functional causal models, it is important to note that the graphical causal model has its origins in the work of Sewell Wright, whose "path analysis" methods were an early linear Gaussian precursor to the more general modern theory expounded by Pearl. Interested readers should refer to Wright (1918), Wright (1920), and Wright (1921).

Functional Causal Models

Before proceeding, we caveat that while the notation and many of the results presented below largely follow Pearl (2009), our exposition may differ in parts due to the focus of this dissertation. We are primarily concerned with causal graphs compatible with the variables outlined in Section 2.1. Specifically, the outcome Y is continuous,

the treatment Z is binary, and the covariates X are arbitrary, though they will correspond to the notion of "covariates" in applied settings as "pre-treatment variables" (see, for example, Imbens and Rubin (2015)). We do *not* consider or discuss in detail graphs in which variables in X are descendants of either Z or Y. We will also assume throughout this chapter and the following chapter that the available covariates X are all of the variables in a specified causal model except for Z and Y

A CDAG $\mathcal{G}$ embeds causal relationships between variables but says nothing of their magnitude, complexity and functional form. Letting V_x refer to all of the causal parents of variable X in a CDAG $\mathcal{G}$, we define the functional causal model of X as

$$X \longleftarrow F(V_x, \epsilon_X)$$

where ϵ_X is an exogenous (but not necessarily univariate) random variable and F is the function that completely determines X from V_x and ϵ_X.

To make this concept more concrete before we proceed, suppose that X has no causal parents ($V_x = \varnothing$) and has a standard normal distribution. In this case, X can be generated by plugging uniform random variables ϵ_X into the standard normal inverse CDF, so that $F(V_x, \epsilon_X) = \Phi^{-1}(\epsilon_X)$.

While a CDAG $\mathcal{G}$ implies functional models for each of the nodes in the graph, the functional model of greatest interest is that which generates Y from its parents, which in applications typically include Z and a number of covariates X:

$$Y \longleftarrow F(X, Z, \epsilon_Y).$$

We can use this functional model to write the average treatment effect estimand as

$$\bar{\tau} = \mathbb{E}\left[F(X, Z = 1, \epsilon_Y) - F(X, Z = 0, \epsilon_Y)\right]$$

where the expectation is taken over (X, ϵ_Y) jointly.

This mathematical representation of an "intervention" is worth highlighting for its unity with the two potential outcomes. $F(X, Z = z, \epsilon_Y)$ fixes $Z = z$ mechanistically while allowing X and ϵ_Y to vary randomly without modification. This is quite different from $F(X, Z, \epsilon_Y) \mid Z = z$, which induces observed distributions of $X \mid Z = z$ and $\epsilon_Y \mid Z = z$. Though the latter carries the same distribution as ϵ_Y, it is not generally true that $X \overset{d}{\sim} X \mid Z = z$. From here, it is clear that $Y^1 \leftarrow F(X, Z = 1, \epsilon_Y)$ and $Y^0 \leftarrow F(X, Z = 0, \epsilon_Y)$, as noted, for example, in Richardson and Robins (2013).

2.2.3 Mean Decomposition of Potential Outcomes

A "mean decomposition" approach to causal inference, which prioritizes identification and estimation target the means of counterfactual random variables is often associated with the work of James Heckman (see for example Heckman (1996)). To illustrate this approach, we define a decomposition

$$\begin{aligned}
\mu(X) &= \mathbb{E}\left[F(X, Z = 0, \epsilon_Y) \mid X\right] \\
\tau(X) &= \mathbb{E}\left[F(X, Z = 1, \epsilon_Y) \mid X\right] - \mathbb{E}\left[F(X, Z = 0, \epsilon_Y) \mid X\right] \\
\nu(X, \epsilon_Y) &= F(X, Z = 0, \epsilon_Y) - \mu(X) \\
\delta(X, \epsilon_Y) &= \{F(X, Z = 1, \epsilon_Y) - F(X, Z = 0, \epsilon_Y)\} - \tau(X)
\end{aligned}$$

where the expectations above are evaluated over ϵ_Y, so that

$$Y \leftarrow \mu(X) + \nu(X, \epsilon_Y) + Z\left[\tau(X) + \delta(X, \epsilon_Y)\right].$$

We can use this model to write our target estimand as

$$\bar{\tau} = \mathbb{E}\left[\tau(X)\right]$$

where the expectation is taken over X. This decomposition of Y into two "mean terms," $\mu(X)$ and $\tau(X)$ and two "error terms" will prove important in discussions of feature selection for causal effect estimation.

This decomposition allows us to represent the counterfactual potential outcomes in terms of two "mean functions" — $\mu(X)$ and $\tau(X)$ — and two "error terms." We will have much more to say about this decomposition later, for now, we define $\mu(X)$ as the *prognostic function* and $\tau(X)$ as the *treatment effect function.*

2.3 Identifying Assumptions for Causal Effect Estimation

The estimand $\bar{\tau}$ written in any of the three frameworks in Section 2.2 involves querying an unknown entity, whether Y^z, F, or $\tau(X)$. In practice, analysts are left with the observed data (Y, Z, X) and must make assumptions in order to estimate $\bar{\tau}$.

Identifying assumptions allow us to obtain unbiased estimate of $\bar{\tau}$ using calculable functions of observed data. We review each of the assumptions here.

1. **Consistency**: This concept refers to the alignment of observed treatment and assigned treatment and is most naturally expressed using Potential Outcomes notation:

$$Y_i = Y_i^1 Z_i + Y_i^0 (1 - Z_i)$$

2. **No Interference**: This concept refers to the independence of one unit's treatment assignment and another unit's potential outcomes

$$(Y_i^1, Y_i^0) \perp\!\!\!\perp Z_j; \forall j \neq i$$

3. **Overlap**: This refers to every unit having nonzero probability of treatment

$$0 < \mathrm{P}(Z_i = 1 \mid X_i) < 1$$

4. **Conditional Unconfoundedness**: This requires that the counterfactual potential outcome distributions are rendered independent of Z given covariates X

$$(Y_i^1, Y_i^0) \perp\!\!\!\perp Z_i \mid X_i$$

Note that due to the unity $Y^z = F(X, Z = z, \epsilon_Y)$, Assumptions 1, 2, and 4 can be similarly articulated using the functional causal model:

$$Y_i = Z_i F(X_i, 1, \epsilon_Y) + (1 - Z_i) F(X_i, 0, \epsilon_Y)$$

$$(F(X_i, 1, \epsilon_Y), F(X_i, 0, \epsilon_Y)) \perp\!\!\!\perp Z_j; \forall j \neq i$$

$$(F(X_i, 1, \epsilon_Y), F(X_i, 0, \epsilon_Y)) \perp\!\!\!\perp Z_i \mid X_i$$

and conditional unconfoundedness can also be framed using the decomposed potential outcome model:

$$(\nu(X_i, \epsilon_Y), \delta(X_i, \epsilon_Y)) \perp\!\!\!\perp Z_i \mid X_i$$

2.3.1 Graphical Identification Criteria

The functional causal model allows for a simple mathematical unity with the potential outcomes framework. However, much of the machinery and insight of the Causal DAG framework are built around the direct use and manipulation of CDAGs, rather than functional models. In particular, the conditional unconfoundedness assumption may be expressed as a graphical criterion known as "d-separation," which we discuss in this section largely mirroring the presentation of Pearl (2009). As in Pearl (2009), we let a "path" in graph $\mathcal{G}$ refer to an ordered list of edges, which connect successive nodes. For example, a path from node A to C might look like $\{(A, B), (B, C)\}$, $\{(A, B), (B, D), (D, E)(E, C)\}$, or (A, C), depending on the graph. In many graphs, two nodes are connected by several possible paths.

In a CDAG, we are interested in paths from Z to Y, which may include the direct path $Z \longrightarrow Y$ as well as a number of "backdoor paths" that reflect causal relationships between and among Z, Y and covariates X. Pearl (2009) defines a path between two nodes, Z and Y, as "d-separated," or "blocked" by a set of variables S if the following is true:

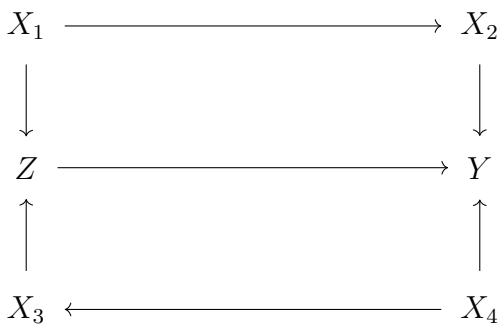

Figure 2.1: The "box graph", which encodes two backdoor paths from Z to Y.

1. The path contains a *chain* or *fork* whose middle variable is in S

2. The path contains a *collider* which is not in S, nor are any of its descendants

The process of blocking every path between two variables Z and Y with a set S is *also* defined by Pearl (2009) as "d-separating" variables Z and Y with S.

We are not interested in blocking a direct path from Z to Y, and this leads to Pearl's definition of the "backdoor criterion," which a set S satisfies if it d-separates every *backdoor* path between Y and Z and includes no descendants of Z (sometimes referred to as "post-treatment variables"). It is worth highlighting that this criterion does not imply a *unique* adjustment set S. Rather, it provides a set of desiderata for evaluating an adjustment set in a graph. For example, conditioning on a collider opens a backdoor path, which can be closed by conditioning on either of the collider's parents.

To make this discussion more concrete, consider the CDAG presented in Figure 2.1 which we will refer to from now on as the "box graph." This graph represents a causal model whose probability distribution factors into

$$\mathrm{P}\left(Y \mid Z, X_2, X_4\right) \mathrm{P}\left(Z \mid X_1, X_3\right) \mathrm{P}\left(X_2 \mid X_1\right) \mathrm{P}\left(X_3 \mid X_4\right) \mathrm{P}\left(X_1\right) \mathrm{P}\left(X_4\right)$$

The box graph implies a functional causal model $Y \longleftarrow F(Z, X_2, X_4, \epsilon_Y)$ from which the modified CDAG presented in Figure 2.2 can be constructed. Since both Y^1

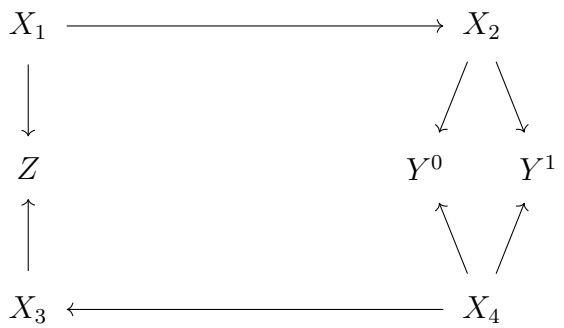

Figure 2.2: The "box graph" with Y replaced by counterfactual random variables (Y^1, Y^0)

and Y^0 are counterfactual random variables constructed by $F(1, X_2, X_4, \epsilon_Y)$ and $F(0, X_2, X_4, \epsilon_Y)$, respectively, the modified graph does not contain edges from Z to either Y^1 or Y^0.

In this modified box graph, there are two paths from Z to Y^z:

1. $\{(Z, X_1), (X_1, X_2), (X_2, Y^z)\}$

2. $\{(Z, X_3), (X_3, X_4), (X_4, Y^z)\}$

Each of these paths contains a chain and a fork. The middle node of the fork in path 1 is X_1 and the middle node of the chain in path 1 is X_2. The middle node of the fork in path 2 is X_4 and the middle node of the chain in path 2 is X_3. Including either *or both* of $\{X_1, X_2\}$ and $\{X_3, X_4\}$ in a conditioning set S d-separates Z and (Y^1, Y^0) in this graph, demonstrating the non-uniqueness of valid control sets that can be determined by causal graphs.

For illustrative purposes, consider another CDAG presented in Figure 2.3 which we will refer to from now on as the "M graph." This graph represents any probability distribution that factors into

$$\mathrm{P}(Y \mid Z, X_3)\,\mathrm{P}(Z \mid X_1)\,\mathrm{P}(X_2 \mid X_1, X_3)\,\mathrm{P}(X_1)\,\mathrm{P}(X_3)$$

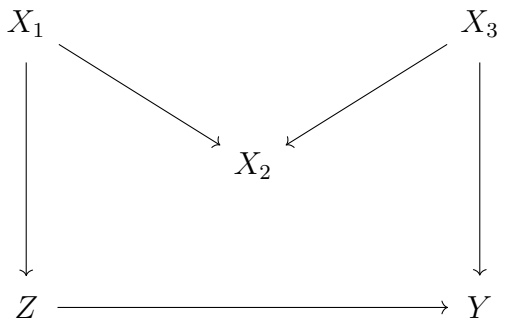

Figure 2.3: The "M graph", which contains a collider among its covariate set $X = (X_1, X_2, X_3)$.

Performing the same counterfactual modification to Y in this graph, we see that there is one path from Z to Y^z:

$$\{(Z, X_1), (X_1, X_2), (X_2, X_3), (X_3, Y^z)\}.$$

This path contains a collider $(X_1 \longrightarrow X_2 \longleftarrow X_3)$ so that the empty adjustment set $S = \varnothing$ satisfies the backdoor criterion with respect to Z and Y. Conditioning on X_2 alone opens an undirected path from X_1 to X_3, and the criterion presented above implies that such a path may be "blocked" by conditioning on either or both of X_1 and X_3. The adjustment sets S that satisfy the backdoor criterion with respect to Z and Y are thus:

- $S = \varnothing$
- $S = \{X_1, X_2\}$
- $S = \{X_2, X_3\}$
- $S = \{X_1, X_2, X_3\}$

This example is instructive for two reasons. First, it underscores the non-uniqueness of valid adjustment sets as in the discussion of the box graph. Second, it shows that conditioning on some subset of variables in a graph may "open up" paths that

were previously blocked, thus necessitating further conditioning to d-separate the treatment and outcome.

Each of the two specific graphs share a common feature: a control set $S = X$, consisting of all of the covariates in the causal models under consideration in this dissertation, satisfies the backdoor criterion with respect to Z and Y. In the next section, we will articulate the relationship between the backdoor criterion and conditional unconfoundedness.

2.3.2 Conditional Unconfoundedness in the Three Causal Frameworks

For the remaining sections of this chapter (and for most of Chapter 3) we will establish most theoretical results assuming that X is discrete. These insights will inform the computational results in Chapter 4 which include continuous covariates. We will also reflect on extensions to / analogues with continuous covariates at various points in the following two chapters.

The **conditional unconfoundedness** assumption has a direct correspondence between the potential outcomes and mean decomposition formalisms, namely that the two expressions below are exactly equivalent:

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

$$(\nu(X, \epsilon_Y), \delta(X, \epsilon_Y)) \perp\!\!\!\perp Z \mid X$$

This can be demonstrated as follows. Since,

$$Y^1 \leftarrow \mu(X) + \nu(X, \epsilon_Y) + [\tau(X) + \delta(X, \epsilon_Y)] = \mu(X) + \tau(X) + \nu(X, \epsilon_Y) + \delta(X, \epsilon_Y),$$

$$Y^0 \leftarrow \mu(X) + \nu(X, \epsilon_Y),$$

and both $\mu(X)$ and $\tau(X)$ are constant conditional on X, $(\mu(X), \tau(X)) \perp\!\!\!\perp Z \mid X$ holds trivially, so that

$$(Y^1, Y^0) \perp\!\!\!\perp Z \mid X \iff (\nu(X, \epsilon_Y), \delta(X, \epsilon_Y)) \perp\!\!\!\perp Z \mid X.$$

What relation do these two equivalent statements hold to the graphical criterion that X d-separates Z and Y? From a CDAG $\mathcal{G}$, we define a modified graph $\tilde{\mathcal{G}}$ such that the outcome $Y \longleftarrow F(Z, X, \epsilon_Y)$ is replaced with two counterfactual nodes:

$$Y^1 \longleftarrow F(1, X, \epsilon_Y), \text{ and}$$
$$Y^0 \longleftarrow F(0, X, \epsilon_Y).$$

Theorem 1.2.4 of Pearl (2009) shows that if X d-separates Z and (Y^1, Y^0) then $(Y^1, Y^0) \perp\!\!\!\perp Z \mid X$. The converse, however, is not true without a further assumption of *faithfulness* (sometimes also called *stability* in the literature, as in Pearl (2009)). Faithfulness asserts that the conditional independence relationships computable from a functional causal model do not vary depending on the parameters of the model, a phenomenon that Shalizi (2021) refers to heuristically as a "conspiracy among the parameters" and which we illustrate below with a simple example. The details of this equivalence (and the various departures) between graphs and potential outcomes are discussed in detail in Richardson and Robins (2013).

Example of a violation of the faithfulness assumption

Consider the following data-generating process (DGP):

$$\begin{aligned}
X &\sim \text{Categorical}\left(p_1, p_2, p_3\right) \\
\pi(X) &= \gamma_1 \mathcal{I}\left\{X = 1\right\} + \gamma_2 \mathcal{I}\left\{X = 2\right\} + \gamma_3 \mathcal{I}\left\{X = 3\right\} \\
Z &\sim \text{Bernoulli}\left(\pi(X)\right) \\
\mu(X) &= \alpha_1 \mathcal{I}\left\{X = 1\right\} + \alpha_2 \mathcal{I}\left\{X = 2\right\} + \alpha_3 \mathcal{I}\left\{X = 3\right\} \\
\tau(X) &= \beta \\
Y &\sim \text{Bernoulli}\left(\mu(X) + \tau(X) Z\right)
\end{aligned} \tag{2.1}$$

with the constraint that $\max_x \mu(x) + \beta \in (0, 1)$.

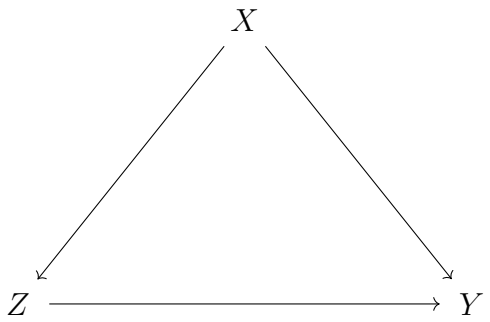

Figure 2.4: The "triangle graph", which contains a single "fork" variable X.

Note that the Bernoulli CDF $F(Y \leq y) = 0 + (1 - \mu(X) + \tau(X)Z)\mathcal{I}\{y = 0\} + \mathcal{I}\{y = 1\}$ has a discrete image, so inverse transform sampling cannot be used to generate Y values from X, Z and a uniform error ϵ_Y, however the simple function $\mathcal{I}(\epsilon_Y < \mu(X) + \tau(X)Z)$ for standard uniform ϵ_Y will generate Bernoulli random variables with probability $\mu(X) + \tau(X)Z$. Thus, we can express the functional model for Y as

$$Y \longleftarrow F(X, Z, \epsilon_Y) = \mathcal{I}\{\epsilon_Y < \mu(X) + \tau(X)Z\}$$

The graph that corresponds to this DGP is presented in Figure 2.4. We see that satisfying the backdoor criterion with respect to Y and Z requires a control set of $S = X$. Note however, that there are particular values of the data-generating parameters for this model that render $(Y^1, Y^0) \perp\!\!\!\perp Z$ unconditional of X. In particular, let

$$\begin{aligned}(p_1, p_2, p_3) &= \left(\frac{16}{60}, \frac{28}{60}, \frac{16}{60}\right) \\ (\gamma_1, \gamma_2, \gamma_3) &= \left(\frac{5}{8}, \frac{5}{14}, \frac{5}{8}\right) \\ (\alpha_1, \alpha_2, \alpha_3) &= \left(\frac{1}{4}, \frac{2}{4}, \frac{3}{4}\right)\end{aligned} \tag{2.2}$$

In this case,

$$\mathrm{P}(Y^1 = 1 \mid Z = 1) = \frac{\sum_{x=1}^{3}(\mu(x) + \beta)\pi(X)p_x}{\sum_{x=1}^{3}\pi(X)p_x} = \frac{1}{2} + \beta$$
$$\mathrm{P}(Y^1 = 1 \mid Z = 0) = \frac{\sum_{x=1}^{3}(\mu(x) + \beta)\pi(X)p_x}{\sum_{x=1}^{3}\pi(X)p_x} = \frac{1}{2} + \beta$$

so that $\mathrm{P}(Y^1 = 1 \mid Z = 0) = \mathrm{P}(Y^1 = 1 \mid Z = 1)$ and thus $Y^1 \perp\!\!\!\perp Z$. Similar derivations hold for Y^0 so that in this particular model, conditional unconfoundedness is satisfied, while in the causal graph in Figure 2.4 that includes all models described by Equation 2.1, (Y^1, Y^0) are not d-separate from Z. The faithfulness assumption rules out sets of parameters as in Equation 2.2 that give spurious independence relationships that conflict with the specific class of independences corresponding to the general CDAG of a causal model. This assumption allows for conditional unconfoundedness ($(Y^1, Y^0) \perp\!\!\!\perp Z \mid X$) to imply that X d-separates (Y^1, Y^0) from Z.

2.4 Identifying Assumptions for Causal Effect Estimation with a Subset of Variables

The discussion in Section 2.3 focuses on the question of identification of a causal effect of Z on Y given an entire covariate set X. Without any further assumptions, we saw that d-separation is a stronger assumption than conditional unconfoundedness, but that the two assumptions are equivalent if we assume faithfulness. We now turn to the question of identification conditional not on X but on functions of the covariates $s(X)$. While the identity $s(X) = X$ is technically such a function, the focus of this dissertation will be on functions $s : \mathcal{X} \longrightarrow \mathcal{S}$ where $\mathcal{S} \subset \mathcal{X}$; in other words, functions that induce some measure of coarsening or dimension reduction of the covariates.

First, note that $s(X)$ here is intended to refer to arbitrary functions of X, including variable selection operations (which return only a subset of the variables in X), nonlinear operations (i.e. $s(X) = X_1X_2 + \mathcal{I}\{X_4 > 0\}$), and simple linear combina-

tions (i.e. $s(X) = X_1 - X_2$). Broadly, $s(X)$ plays the same "feature selection" role as do nonparametric machine learning methods. The following section gives several illustrative examples of such functions with discrete and continuous covariates.

2.4.1 Feature Selection as a Function of Covariates

To make the concept of a "feature selection function" concrete, consider a simple example in which X is composed of 3 independent binary covariates, so that the covariate space $\mathcal{X}$ has 8 elements. In this case any selection function is essentially a stratification function, which must map each of these 8 elements to a (possibly collapsed) "stratum." As we will see when we review estimators in Section 2.6, the values of these labels do not matter for estimation purposes — what matters is *which levels of X are grouped together*.

In Table 2.1, we give four examples of such operations. The first, s_1, maps each level of X to its own stratum and is thus equivalent to the identity function in that conditioning on $s_1(X)$ has exactly the same effect as conditioning on X. The next two functions, s_2 and s_3, perform traditional "variable selection" by collapsing unique levels of X_1 and X_2, respectively, into the same strata. Finally, s_4 performs a more elaborate feature construction, expressable as

$$s_4(X) = 2 + 2X_3 - \mathcal{I}\{X_1 = X_2\},$$

which returns a "non-axis-aligned" adjustment set.

To give one more example of a feature selection function, this time on continuous covariates, note that decision trees construct mutually exclusive bases in the form of repeated logical statements. Consider that the decision tree in Figure 2.5 partitions X into three leaves, and the conditions that define these leaves can be expressed as a

X_1	X_2	X_3	$s_1(X)$	$s_2(X)$	$s_3(X)$	$s_4(X)$
0	0	0	1	1	1	1
1	0	0	2	1	2	2
0	1	0	3	2	1	2
1	1	0	4	2	2	1
0	0	1	5	3	3	3
1	0	1	6	3	4	4
0	1	1	7	4	3	4
1	1	1	8	4	4	3

Table 2.1: Visualization of several "feature selection" operations performable by functions of the unique levels of X

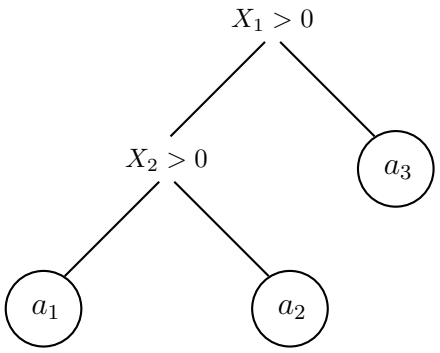

Figure 2.5: A simple decision tree with splits on X_1 and X_2

3-vector of basis functions,

$$s(X) = \left[\mathcal{I}\left\{X_1 > 0\right\}\mathcal{I}\left\{X_2 > 0\right\}, \mathcal{I}\left\{X_1 > 0\right\}\mathcal{I}\left\{X_2 \leq 0\right\}, \mathcal{I}\left\{X_1 \leq 0\right\}\right].$$

2.4.2 Identification of the Average Treatment Effect using $s(X)$

In potential outcomes, $s(X)$ satisfying conditional unconfoundedness is expressed naturally as

$$(Y^1, Y^0) \perp\!\!\!\perp Z \mid s(X).$$

To express this identification criterion in the "mean decomposition" approach, we must first define a new decomposition

$$\begin{aligned}
\mu(s(X)) &= \mathbb{E}\left[Y^0 \mid s(X)\right] = \mathbb{E}\left[\mathbb{E}\left[Y^0 \mid s(X)\right] \mid s(X)\right] = \mathbb{E}\left[\mu(X) \mid s(X)\right] \\
\tau(s(X)) &= \mathbb{E}\left[Y^1 \mid s(X)\right] - \mathbb{E}\left[Y^0 \mid s(X)\right] \\
&= \mathbb{E}\left[\tau(X) \mid s(X)\right] + \mathbb{E}\left[\mu(X) \mid s(X)\right] - \mathbb{E}\left[\mu(X) \mid s(X)\right] = \mathbb{E}\left[\tau(X) \mid s(X)\right] \\
\nu(s, X, \epsilon_Y) &= Y^0 - \mathbb{E}\left[Y^0 \mid s(X)\right] = Y^0 - \mathbb{E}\left[\mu(X) \mid s(X)\right] \\
&= \mu(X) + \nu(X, \epsilon_Y) - \mathbb{E}\left[\mu(X) \mid s(X)\right] = (\mu(X) - \mu(s(X))) + \nu(X, \epsilon_Y) \\
\delta(s, X, \epsilon_Y) &= \left(Y^1 - Y^0\right) - \left(\mathbb{E}\left[Y^1 \mid s(X)\right] - \mathbb{E}\left[Y^0 \mid s(X)\right]\right) \\
&= (\tau(X) - \tau(s(X))) + \delta(X, \epsilon_Y)
\end{aligned}$$

From here, we follow the same line of reasoning as in Section 2.3: $\mu(s(X))$ and $\tau(s(X))$ are constant given $s(X)$ so that

$$(\nu(s, X, \epsilon_Y), \delta(s, X, \epsilon_Y)) \perp\!\!\!\perp Z \mid s(X) \iff (Y^1, Y^0) \perp\!\!\!\perp Z \mid s(X).$$

The structural model view of conditional unconfoundedness exposes two separate conditions in the case of coarse functions s of X. The first is exactly as demanded by conditional unconfoundedness given all of X,

$$(\nu(X, \epsilon_Y), \delta(X, \epsilon_Y)) \perp\!\!\!\perp Z \mid s(X).$$

The second is that

$$(\mu(X), \tau(X)) \perp\!\!\!\perp Z \mid s(X).$$

Unconditionally, both $\nu(X, \epsilon_Y)$ and $\mu(X)$ (as well as $\delta(X, \epsilon_Y)$ and $\tau(X)$) are random variables. Conditional on X, both $\mu(X)$ and $\tau(X)$ are constants and the conditional unconfoundedness assumption simply stipulates that any residual variation in Y^1 and Y^0 is independent of Z given X. Conditional on $s(X)$, both $\mu(X)$ and $\tau(X)$ are potentially non-constant random variables, and conditional unconfoundedness

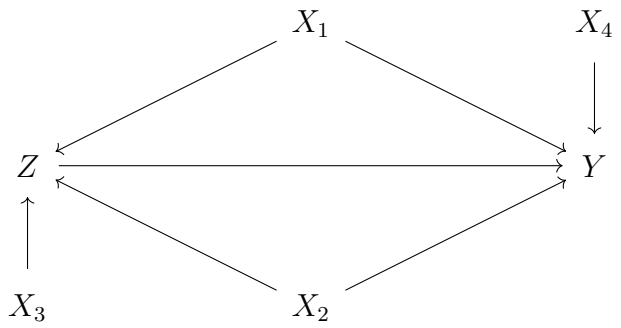

Figure 2.6: Causal graph in terms of original covariates

demands that the residual variation in $\mu(X)$ and $\tau(X)$ (not controlled by $s(X)$) be independent of Z given $s(X)$. We will have more to say about this assumption as it relates to the average treatment effect in the following section, but we first must discuss how $s(X)$ influences identification in causal graphs.

To discuss identification of a graphical model given $s(X)$, we must consider transformed graphs which include the variable(s) $s(X)$ as nodes. Depending on the associated functional model and the nature of $s(X)$ the new graph may be causal in its transformed variables or may contain probabilistic (non-causal) edges. We review examples of each case below.

First, consider this DGP, corresponding to the graph in Figure 2.6.

$$
\begin{aligned}
X_1, X_2, X_3, X_4 &\sim \text{Bernoulli}\,(p) \\
Z &\sim \text{Bernoulli}\,(\gamma_0 + \gamma_1\,(2X_1X_2 - X_1 - X_2 + 1) + \gamma_2 X_3) \\
Y &\sim \mathcal{N}\,(\tau Z + \alpha_0 + \alpha_1\,(2X_1X_2 - X_1 - X_2 + 1) + \gamma_2 X_4)
\end{aligned}
$$

Defining $W \longleftarrow F(X_1, \epsilon_W) = 2(X_1 - 1)(\epsilon_W - 1)$ as a random variable generated by a functional transformation of X_1 and $\epsilon_W \sim \text{Bernoulli}\,(p)$, we can write a modified graph that is still causal in each of its variables as in Figure 2.7.

Letting $s(X) = W$, we see that $s(X)$ d-separates all backdoor paths from Z to Y in the transformed causal graph, just as did (X_1, X_2) in the original causal graph. Consider now the box diagram presented in Figure 2.1. X_2 has no direct

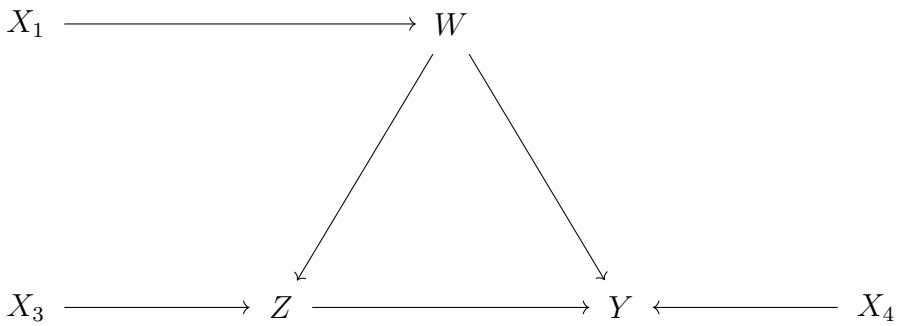

Figure 2.7: Causal graph under transformed covariates

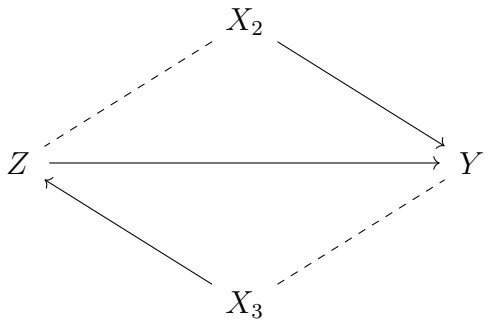

Figure 2.8: The "box graph" represented only by (X_2, X_3, Z, Y)

causal relationship on Z, and X_3 has no direct causal relationship with Y, but if we write the graph solely in terms of these four variables, we must represent the probabilistic relationship between Z and X_2 (unconditional of X_1) and that between Y and X_3 (unconditional of X_4) which we do using a dashed line in Figure 2.8. In this graph, there are two backdoor paths between Z and Y: $\{(Z, X_2), (X_2, Y)\}$ and $\{(Z, X_3), (X_3, Y)\}$, so that an adjustment set $s(X)$ which "selects" X_2 and X_3 d-separates Z and (Y^1, Y^0).

Perhaps it is worth checking in to pose (and answer) the question: what is the purpose of this lengthy elaboration of identifying conditions using three causal inference frameworks? Some concepts are more easily expressed in certain frameworks. For example, the causal graph framework, which encourages expressing causal relationships in graphs, allows practitioners to more readily reason about variable selection

for causal inference. By contrast, the assumptions of no interference and consistency are more naturally expressed in the potential outcomes verbiage, which more directly represents observations in a sample. Finally, the concept of separating counterfactual random variable Y^1 and Y^0 into "mean terms" $\mu(X)$ and $\tau(X)$ and "error terms" $\nu(X, \epsilon_Y)$ and $\delta(X, \epsilon_Y)$ we be useful in reasoning about average treatment effect identifying criteria in the following section.

2.5 Identifying Assumptions for Average Treatment Effect

The purpose of the identifying assumptions presented above, broadly, is to enable computation of counterfactual estimands (which by their very nature are not directly observable) via observable random variables. Consider again the average treatment effect defined in Section 2.2.1:

$$\bar{\tau} = \mathbb{E}\left[Y^1 - Y^0\right].$$

2.5.1 Identification Based on Conditional Unconfoundedness

We can show step-by-step how the identifying assumptions are used to convert this counterfactual estimand into an observable estimand. First, the law of iterated expectations shows that

$$\bar{\tau} = \mathbb{E}\left[Y^1 - Y^0\right] = \mathbb{E}\left[\mathbb{E}\left[Y^1 - Y^0 \mid X\right]\right] = \mathbb{E}\left[\mathbb{E}\left[Y^1 \mid X\right] - \mathbb{E}\left[Y^0 \mid X\right]\right].$$

Conditional unconfoundedness implies that both

$$\mathbb{E}\left[Y^1 \mid X\right] = \mathbb{E}\left[Y^1 \mid Z = 1, X\right]$$
$$\mathbb{E}\left[Y^0 \mid X\right] = \mathbb{E}\left[Y^0 \mid Z = 0, X\right]$$

Conditional on Z, we can rewrite

$$\mathbb{E}\left[Y^1 \mid Z = 1, X\right] = \mathbb{E}\left[Y^1 Z + Y^0(1 - Z) \mid Z = 1, X\right]$$
$$\mathbb{E}\left[Y^0 \mid Z = 0, X\right] = \mathbb{E}\left[Y^1 Z + Y^0(1 - Z) \mid Z = 0, X\right]$$

and by consistency we have that

$$\mathbb{E}\left[Y^1 \mid Z = 1, X\right] = \mathbb{E}\left[Y^1 Z + Y^0(1-Z) \mid Z = 1, X\right] = \mathbb{E}\left[Y \mid Z = 1, X\right]$$
$$\mathbb{E}\left[Y^0 \mid Z = 0, X\right] = \mathbb{E}\left[Y^1 Z + Y^0(1-Z) \mid Z = 0, X\right] = \mathbb{E}\left[Y \mid Z = 0, X\right]$$

Finally, we require that both $\mathbb{E}\left[Y \mid Z = 1, X\right]$ and $\mathbb{E}\left[Y \mid Z = 0, X\right]$ be nondegenerate across the entire support of X, so that both $\mathbb{E}\left[\mathbb{E}\left[Y \mid Z = 1, X\right]\right] = \mathbb{E}\left[Y^1\right]$ and $\mathbb{E}\left[\mathbb{E}\left[Y \mid Z = 0, X\right]\right] = \mathbb{E}\left[Y^0\right]$. This can be achieve by ensuring that the conditioning sets $(Z = 1, X)$ and $(Z = 0, X)$ are not measure zero whenever X is not measure zero, which is true because positivity

$$0 < \mathrm{P}(Z = 1 \mid X) < 1$$

holds across the support of X.

Thus we have that

$$\bar{\tau} = \mathbb{E}\left[Y^1 - Y^0\right] = \mathbb{E}\left[\mathbb{E}\left[Y \mid Z = 1, X\right] - \mathbb{E}\left[Y \mid Z = 0, X\right]\right],$$

where the conditional expectations in the final estimand are expressed in terms of observable random variables Y, Z, and X.

There are many ways to convert this estimand to an estimator, which will be discussed in more depth in Section 2.6. For now, we simply note that the "no interference" assumption, which stipulates unconditional independence of potential outcomes and treatment assignment *between* observations, is required to ensure that any of the estimators has expected value $\mathbb{E}\left[\mathbb{E}\left[Y \mid Z = 1, X\right] - \mathbb{E}\left[Y \mid Z = 0, X\right]\right]$.

2.5.2 Minimal Identifying Assumptions for the Average Treatment Effect

Consider now an adjustment set $s(X)$ where $s(\mathcal{X}) \subset \mathcal{X}$. The above derivations hold, except that we now express the law of iterated expectations in term of $s(X)$:

$$\bar{\tau} = \mathbb{E}\left[Y^1 - Y^0\right] = \mathbb{E}\left[\mathbb{E}\left[Y^1 - Y^0 \mid s(X)\right]\right] = \mathbb{E}\left[\mathbb{E}\left[Y^1 \mid s(X)\right] - \mathbb{E}\left[Y^0 \mid s(X)\right]\right].$$

This thus requires only one modification to the assumptions as discussed in Section 2.4:

$$(Y^1, Y^0) \perp\!\!\!\perp Z \mid s(X).$$

Note, however, that the derivation above does not make use of full conditional independence of the random variables (Y^1, Y^0) and Z given $s(X)$, just that

$$\mathbb{E}\left[Y^1 \mid s(X)\right] = \mathbb{E}\left[Y^1 \mid Z = 1, s(X)\right]$$
$$\mathbb{E}\left[Y^0 \mid s(X)\right] = \mathbb{E}\left[Y^0 \mid Z = 0, s(X)\right]$$

a condition known more generally in probability theory as "mean conditional independence." This point – that certain average causal effects such as the ATE are identified without requiring full conditional independence – has been made in several places in the literature, most notably in Heckman (1996) and Cameron and Trivedi (2005).

2.5.3 Identifying Assumptions via Mean Terms

The assumption stated above is "minimal" but it is also not particularly illuminating. It asserts simply that the smallest adjustment set that identifies the average treatment effect is the set $s(X)$ that can remove Z from the conditioning sets in $\mathbb{E}[Y^0 \mid Z = 0, s(X)]$ and $\mathbb{E}[Y^1 \mid Z = 1, s(X)]$. Analysts are left to reason about the counterfactual random variables Y^0 and Y^1. Above, we showed that these two variables can be expressed in terms of "mean functions" and "error terms,"

$$Y^0 = \mu(X) + \nu(X, \epsilon_y)$$
$$Y^1 = \mu(X) + \tau(X) + \nu(X, \epsilon_y) + \delta(X, \epsilon_y).$$

Depending on the application, which variables impact the prognostic and treatment effect functions may be easier to hypothesize than which variables impact the entire distributions of Y^0 and Y^1.

We now show that a very similar assumption which we call *mean conditional unconfoundedness*, based on $\mu(X)$ and $\tau(X)$ alone identifies the average treatment effect in exactly the same manner as in Section 2.5.2:

$$\mu(X) \perp\!\!\!\perp Z \mid s(X)$$

$$\tau(X) \perp\!\!\!\perp Z \mid s(X)$$

These criteria offer a way of reasoning about "feature selection" – any function s of X that satisfies mean conditional unconfoundedness identifies the average treatment effect. Any variation in X not accounted for by $s(X)$ may well impact the shape or magnitude of $\nu(s, X, \epsilon_y)$ and $\delta(s, X, \epsilon_y)$, but these terms are mean zero and vanish when evaluating the average treatment effect.

Now, we demonstrate that this assumption identifies the average treatment effect, assuming as stated in Section 2.3.2 that X is discrete. The applications of consistency, positivity and no interference are exactly as in Section 2.5, so we omit them here and focus on showing that

$$\mathbb{E}\left[Y^1 \mid Z = 1, s(X)\right] = \mathbb{E}\left[Y^1 \mid s(X)\right]$$

$$\mathbb{E}\left[Y^0 \mid Z = 0, s(X)\right] = \mathbb{E}\left[Y^0 \mid s(X)\right]$$

First, observe that

$$\begin{aligned}
\mathbb{E}\left[Y^0 \mid X, Z\right] &= \mathbb{E}\left[F(X, 0, \epsilon_y) \mid X, Z\right] = \int_{\epsilon_y} F(X, 0, \epsilon_y) dP(\epsilon_y \mid X, Z) \\
&= \int_{\epsilon_y} F(X, 0, \epsilon_y) dP(\epsilon_y) = \mu(X) = \mathbb{E}\left[Y^0 \mid X\right] \\
\mathbb{E}\left[Y^1 \mid X, Z\right] &= \mathbb{E}\left[F(X, 1, \epsilon_y) \mid X, Z\right] = \int_{\epsilon_y} F(X, 1, \epsilon_y) dP(\epsilon_y \mid X, Z) \\
&= \int_{\epsilon_y} F(X, 1, \epsilon_y) dP(\epsilon_y) = \mu(X) + \tau(X) = \mathbb{E}\left[Y^1 \mid X\right]
\end{aligned}$$

where the second line of each derivation holds because $\epsilon_y \perp\!\!\!\perp (X, Z)$.

Now, we can write

$$\begin{aligned}
\mathbb{E}\left[Y^0 \mid s(X), Z\right] &= \mathbb{E}\left[\mathbb{E}\left[Y^0 \mid X, Z\right] \mid s(X), Z\right] = \mathbb{E}\left[\mu(X) \mid s(X), Z\right] \\
&= \mathbb{E}\left[\mu(X) \mid s(X)\right] = \mathbb{E}\left[Y^0 \mid s(X)\right] \\
\mathbb{E}\left[Y^1 \mid s(X), Z\right] &= \mathbb{E}\left[\mathbb{E}\left[Y^1 \mid X, Z\right] \mid s(X), Z\right] = \mathbb{E}\left[\mu(X) + \tau(X) \mid s(X), Z\right] \\
&= \mathbb{E}\left[\mu(X) + \tau(X) \mid s(X)\right] = \mathbb{E}\left[Y^1 \mid s(X)\right]
\end{aligned}$$

where the second line of each derivation holds because $\mu(X) \perp\!\!\!\perp Z \mid s(X)$ and $\tau(X) \perp\!\!\!\perp Z \mid s(X)$, and the result holds.

Finally, the mathematical statement and intuition behind *mean conditional unconfoundedness* differs from that of mean conditional independence, but we do not contend that the substance of the assumptions are different. Rather, it is plausible that an exact mathematical equivalence exists, though we do not explore the issue in this dissertation. This research will largely be concerned with applications of mean conditional unconfoundedness, rather than any claim to its mathematical novelty.

2.6 Estimators for the Average Treatment Effect

Note that the identification derivations above always convert $\mathbb{E}\left[Y^1 - Y^0\right]$ into $\mathbb{E}\left[\mathbb{E}\left[Y \mid s(X), Z = 1\right] - \mathbb{E}\left[Y \mid s(X), Z = 0\right]\right]$ for some conditioning set $s(X)$ (which could simply be the identity $s(X) = X$). Estimators for the ATE rely on sample based approximations of these conditional expectations. First, we note that when covariates are discrete, there is an exact equivalence between three common ATE estimators, which we review below.

2.6.1 Equivalence of Three Common Estimators when X is Discrete

Let n be the number of observed samples of treatment Z, outcome Y and covariates X. We assume that X is discrete and can thus be transformed into a matrix of

binary "regression contrast offsets" which we refer to as $\tilde{X}$ and we let p refer to the number of columns in $\tilde{X}$.

We define the inverse propensity weighting (IPW), stratification, and regression estimators using the entire covariate set X below

$$\begin{aligned}
\bar{\tau}_{IPW} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i Z_i}{\hat{p}(X_i)} - \frac{Y_i(1-Z_i)}{1-\hat{p}(X_i)}\right) \\
\bar{\tau}_{strat} &= \sum_{x\in\mathcal{X}}\frac{n_x}{n}\left(\bar{Y}_{x,Z=1} - \bar{Y}_{x,Z=1}\right) \\
\bar{\tau}_{reg} &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_{Z=1,X=x_i} - \hat{Y}_{Z=0,X=x_i}\right)
\end{aligned}$$

where the propensity score is defined empirically,

$$\hat{p}(x) = \frac{N_{x,Z=1}}{n_x} \qquad \begin{aligned} N_{x,Z=1} &= \sum_{i=1}^{n}\mathbf{1}(X_i = x, Z = 1) \\ n_x &= \sum_{i=1}^{n}\mathbf{1}(X_i = x) \end{aligned}$$

and the regression fit for $\bar{\tau}_{reg}$ is a fully saturated linear model

$$Y = \left(\alpha_0 + \alpha_1\tilde{X}_1 + \cdots + \alpha_p\tilde{X}_p\right) + Z\left(\beta_0 + \beta_1\tilde{X}_1 + \cdots + \beta_p\tilde{X}_p\right) + \epsilon$$

Each of these estimators require the condition that every unique value of X has at least one treated and at least one control unit (that is $N_{x,Z=1} > 0$ and $N_{x,Z=0} > 0$ for all x). Now, observe that the IPW and stratification estimators are exactly equal in this case

$$\begin{aligned}
\bar{\tau}_{IPW} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i Z_i}{\hat{p}(X_i)} - \frac{Y_i(1-Z_i)}{1-\hat{p}(X_i)}\right) \\
&= \frac{1}{n}\sum_{x\in\mathcal{X}}\left(\frac{n_x N_{x,Z=1}\bar{Y}_{x,Z=1}}{N_{x,Z=1}} - \frac{n_x N_{x,Z=0}\bar{Y}_{x,Z=0}}{N_{x,Z=0}}\right) \\
&= \frac{1}{n}\sum_{x\in\mathcal{X}}\left(n_x\bar{Y}_{x,Z=1} - n_x\bar{Y}_{x,Z=0}\right) \\
&= \sum_{x\in\mathcal{X}}\frac{n_x}{n}\left(\bar{Y}_{x,Z=1} - \bar{Y}_{x,Z=0}\right) = \bar{\tau}_{strat}
\end{aligned}$$

Chapter 17 of Imbens and Rubin (2015) gives a similar result comparing "subclassification" estimators (stratification estimators that use ranges of the propensity score as strata) and the IPW estimator.

Similarly, $\hat{Y}_{z,x}$ in the regression estimator reduces to the cell mean $\bar{Y}_{x,z}$, so that the same equivalence holds

$$\begin{aligned}
\bar{\tau}_{reg} &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_{Z=1,X=x_i} - \hat{Y}_{Z=0,X=x_i}\right) \\
&= \sum_{x\in\mathcal{X}}\frac{n_x}{n}\left(\hat{Y}_{Z=1,X=x} - \hat{Y}_{Z=0,X=x}\right) \\
&= \sum_{x\in\mathcal{X}}\frac{n_x}{n}\left(\bar{Y}_{x,Z=1} - \bar{Y}_{x,Z=1}\right) = \bar{\tau}_{strat}
\end{aligned}$$

For a subset $s(X)$, the estimators are defined in the exact same manner, simply using the unique levels of $s(X)$ rather than X.

$$\begin{aligned}
\bar{\tau}^{s}_{IPW} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i Z_i}{\hat{p}(s(X_i))} - \frac{Y_i(1-Z_i)}{1-\hat{p}(s(X_i))}\right) \\
\bar{\tau}^{s}_{strat} &= \sum_{j\in s(\mathcal{X})}\frac{n_j}{n}\left(\bar{Y}_{j,Z=1} - \bar{Y}_{j,Z=1}\right) \\
\bar{\tau}^{s}_{reg} &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_{Z=1,s(X)=s(x_i)} - \hat{Y}_{Z=0,s(X)=s(x_i)}\right)
\end{aligned}$$

where the propensity score is defined empirically,

$$\hat{p}(s(x)) = \frac{N_{j,Z=1}}{n_j} \qquad \begin{aligned} N_{j,Z=1} &= \sum_{i=1}^{n}\mathbf{1}(s(X_i)=j, Z=1) \\ n_j &= \sum_{i=1}^{n}\mathbf{1}(s(X_i)=j) \end{aligned}$$

and the regression fit for $\bar{\tau}_{reg}$ is a fully saturated linear model on a similarly defined design matrix $\tilde{S}$

$$Y = \left(\alpha_0 + \alpha_1\tilde{S}_1 + \cdots + \alpha_p\tilde{S}_p\right) + Z\left(\beta_0 + \beta_1\tilde{S}_1 + \cdots + \beta_p\tilde{S}_p\right) + \epsilon$$

2.6.2 Flexible Estimators for the Average Treatment Effect with Arbitrary Covariates

Estimating conditional expectations is comparatively straightforward with a discrete conditioning set. The exact equivalence above vanishes when covariates are continuous, as does the relative simplicity of most of the estimators. The direct regression estimator requires a specification of the functional form of the regression equation — linearity in the covariates may be assumed, or more flexible bases, including polynomials and splines, may be constructed. For the IPW estimator, estimates of the propensity score face the same model specification question — will the propensity model be linear in the covariates? The stratification estimator faces a related challenge — without discrete covariates, exact stratification is impossible. One common alternative is to discretize an estimated propensity score and stratify on the discretized "strata" of a continuous propensity score (Imbens and Rubin (2015), Chapter 17).

We will have more to say about regression estimators with continuous covariates in Chapter 4. Chapter 3 investigates the theoretical considerations in selecting an adjustment set $s(X)$ from discrete covariates. These insights inform our computational investigations in Chapter 4.

It is worth mentioning another estimator which we do not discuss in depth in this dissertation — "matching" — in which treated-control pairs with similar covariate values are constructed and used as the basis for a regression adjustment. Readers interested in more in-depth coverage of matching are referred to Chapter 18 of Imbens and Rubin (2015) or Chapter 5 of Morgan and Winship (2015).

Chapter 3

FEATURE SELECTION IN CAUSAL INFERENCE USING DISCRETE COVARIATES: THEORY, DESIDERATA, AND EXAMPLES

3.1 Introduction and Notation

We assume throughout this chapter that n samples of (Y, Z, X) are observed and that X is discrete. Let K refer to the total number of unique levels of X and define a function $k : \mathcal{X} \longrightarrow \{1, \ldots, K\}$ that maps the unique levels of X to a univariate set of labels $\{1, \ldots, K\}$. We assume that $Z \mid X \sim \text{Bernoulli}(\pi(X))$ where $\pi(X) \in (0, 1)$ and we assume that $Y \mid Z, X \sim \mathcal{N}(\mu(X) + Z\tau(X), \sigma^2)$.

3.2 Theoretical Considerations in Selecting $s(X)$

Section 2.3 outlined four identifying conditions for causal effects of Z on Y, and Section 2.5.3 showed that the "conditional unconfoundedness" assumption can be supplanted by a comparatively weaker "mean conditional unconfoundedness" assumption, which demands of any conditioning set $s(X)$ that

$$\mu(X) \perp\!\!\!\perp Z \mid s(X), \text{ and}$$
$$\tau(X) \perp\!\!\!\perp Z \mid s(X).$$

This leaves open the question: how does an analyst go about selecting an adjustment set that satisfies mean conditional unconfoundedness? We review two prominent examples from the causal inference literature, before introducing general theory on how to "go beyond" either of these two adjustment sets.

3.2.1 Propensity Function

Rosenbaum and Rubin (1983) define the *propensity score* as the conditional probability of receiving treatment given covariates,

$$\pi(X) = \mathrm{P}\left(Z = 1 \mid X\right),$$

and show that the propensity score satisfies conditional unconfoundedness when X satisfies conditional unconfoundedness. This in turn implies that a valid estimate of $\mathrm{P}\left(Z = 1 \mid X\right)$ can serve as a conditioning set that identifies the average treatment effect, an insight that is especially helpful when X is high-dimensional as a one-dimensional projection can be used in its place.

We show briefly that the propensity score also satisfies the comparatively weaker *mean conditional unconfoundedness* assumption. Let $s(X) = \pi(X)$ and note that $Z \sim \text{Bernoulli}(\pi(X))$ so that, conditional on $\pi(X)$, Z is independent from any other function of X, including $\mu(X)$ and $\tau(X)$.

3.2.2 Generalized Prognostic Function

Hansen (2008) shows that when a treatment effect function $\tau(X)$ is constant and X satisfies conditional unconfoundedness, $\mathbb{E}\left[Y^0 \mid X\right]$ also satisfies conditional unconfoundedness. This conditional expectation corresponds to $\mu(X)$ in our structural model decomposition, so we use that notation here. Hansen (2008) also shows that $\mu(X)$, which he refers to as the "prognostic score," may be estimated as $\mathbb{E}\left[Y \mid X, Z = 0\right]$.

Furthermore, Hansen (2008) notes that in the more general case in which $\tau(X)$ is a non-constant function of X, the average treatment effect is only identified given both $\mu(X)$ and $\tau(X)$. We denote the union

$$s(\mathcal{X}) = \mu(\mathcal{X}) \cup \tau(\mathcal{X}),$$

and refer to $s(X)$ as the "generalized prognostic function." We can see that the generalized prognostic function satisfies mean conditional unconfoundedness because $\mu(X)$ and $\tau(X)$ are both constant given the generalized prognostic function and are therefore independent of Z.

3.2.3 Minimal Deconfounding Function

We have seen that two common "data-defined" approaches to computing a univariate conditioning set, the propensity score and the prognostic score, identify the average treatment effect. Each conditioning set provides the comparatively-strong phenomenon of conditional unconfoundedness. We have also noted that for our estimand of interest — the average treatment effect — only mean conditional unconfoundedness is necessary. A natural question is thus: what is the *smallest* possible conditioning set that provides mean conditional unconfoundedness? We define

$$\lambda(X) = \mathbb{E}\left[Z \mid \mu(X), \tau(X)\right],$$

and first show that $\lambda(X)$ satisfies mean conditional unconfoundedness. Observe that the distribution of $Z \mid X$ is fully characterized by

$$\mathrm{P}\left(Z = 1 \mid X\right) = \pi(X) = \mathbb{E}\left[Z \mid X\right].$$

We must thus show that

$$\mathbb{E}\left[Z \mid \lambda(X)\right] = \mathbb{E}\left[Z \mid \lambda(X), \mu(X), \tau(X)\right].$$

First, observe that $\lambda(\mathcal{X})$ is a projection onto $\mu(\mathcal{X}) \cup \tau(\mathcal{X})$ so that

$$\mathbb{E}\left[Z \mid \lambda(X), \mu(X), \tau(X)\right] = \mathbb{E}\left[Z \mid \mu(X), \tau(X)\right].$$

Now, by the law of total expectation, we have that

$$\begin{aligned}\mathbb{E}\left[Z \mid \lambda(X)\right] &= \mathbb{E}\left[\mathbb{E}\left[Z \mid \lambda(X), \mu(X), \tau(X)\right] \mid \lambda(X)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[Z \mid \mu(X), \tau(X)\right] \mid \lambda(X)\right] \\ &= \mathbb{E}\left[\lambda(X) \mid \lambda(X)\right] = \lambda(X) = \mathbb{E}\left[Z \mid \lambda(X), \mu(X), \tau(X)\right] \\ &= \mathbb{E}\left[Z \mid \mu(X), \tau(X)\right]\end{aligned}$$

Now, we show that $\lambda(X)$ is the *smallest* adjustment set that satisfies mean conditional unconfoundedness.

Proposition 1. *Let $\mathcal{X}$ be a discrete space supporting the random variable X. There exists no function s for which it is both true that*

1. $|s(\mathcal{X})| < |\lambda(\mathcal{X})|$
2. $\mu(X) \perp\!\!\!\perp Z \mid s(X)$ *and* $\tau(X) \perp\!\!\!\perp Z \mid s(X)$

Proof. We assume that there exists an s satisfying conditions 1 and 2 above. There thus exist distinct x, x' such that $s(x) = s(x')$ and $\lambda(x) \neq \lambda(x')$. If $\lambda(x) \neq \lambda(x')$, then it must be the case that both $(\mu(x), \tau(x)) \neq (\mu(x'), \tau(x'))$ and $\mathbb{E}\left(Z \mid \mu(x), \tau(x)\right) \neq \mathbb{E}\left(Z \mid \mu(x'), \tau(x')\right)$. Let $c = s(x) = s(x')$ and observe that condition 2 above implies

$$\begin{aligned}\mathbb{E}\left[Z \mid s(X) = c\right] &= \mathbb{E}\left[Z \mid s(X) = c, \mu(x), \tau(x)\right] = \mathbb{E}\left[Z \mid \mu(x), \tau(x)\right] \\ &= \mathbb{E}\left[Z \mid s(X) = c, \mu(x'), \tau(x')\right] = \mathbb{E}\left[Z \mid \mu(x'), \tau(x')\right].\end{aligned}$$

This is a contradiction as this would imply that $\lambda(x) = \lambda(x')$. □

3.3 Variance Considerations in Selecting an Adjustment Set

Thus far, we have reviewed several adjustment sets that satisfy mean conditional unconfoundedness and identify the ATE. It is worth pausing to reflect on this discussion for a moment. Given the propensity score identifies the ATE and is computable

from the (Z, X) data, why did we bother discussing the generalized prognostic function and $\lambda(X)$? The primary reason is variance reduction, which will be the subject of Proposition 2. We will make this precise below and in the ensuring discussion, but for now consider a data-generating process with many variables that influence only Z. Estimating and adjusting for the propensity score will effect a stratification on many *instrumental* variables, which only influence the treatment assignment. Since these variables were not necessary for identifying the average treatment effect and they do not help predict the outcome, adding them to an adjustment set increases the variance of the resulting ATE estimator.

We consider the DGP introduced in Section 3.1. Since there is a one-to-one correspondence between X and $k(X)$, for notational simplicity, we let X refer to $k(X)$, a mapping from the unique values of $\mathcal{X}$ to numeric stratum indices. Define subset-specific sample sizes as follows:

- N_x: the number of observations with $X = x$,
- $N_{x,z}$: the number of observations with $X = x$ and $Z = z$.

We assume that $N_x > 0$ and $N_{x,z} > 0$ for all x and $z \in \{0, 1\}$. The stratification estimator introduced in Section 2.6 can be expressed here using a stratification function $s(\mathcal{X})$, which returns $J \leq K$ discrete function values. We compute the average difference in outcomes between the treated and control groups separately for individuals

in each of the J strata, so that

$$\begin{aligned}
\bar{\tau}^{s}_{strat} &= \sum_{j \in s(\mathcal{X})} \frac{N_j}{n} \left(\bar{Y}_{j,1} - \bar{Y}_{j,0} \right) \\
N_{j,0} &= \sum_{i=1}^{n} \mathbf{1}\left\{ s(X_i) = j \right\} \mathbf{1}\left\{ Z_i = 0 \right\} \\
\bar{Y}_{j,0} &= \frac{1}{N_{j,0}} \sum_{i=1}^{n} Y_i \mathbf{1}\left\{ s(X_i) = j \right\} \mathbf{1}\left\{ Z_i = 0 \right\} \\
N_j &= \sum_{i=1}^{n} \mathbf{1}\left\{ s(X_i) = j \right\} \\
N_{j,1} &= \sum_{i=1}^{n} \mathbf{1}\left\{ s(X_i) = j \right\} \mathbf{1}\left\{ Z_i = 1 \right\} \\
\bar{Y}_{j,1} &= \frac{1}{N_{j,1}} \sum_{i=1}^{n} Y_i \mathbf{1}\left\{ s(X_i) = j \right\} \mathbf{1}\left\{ Z_i = 1 \right\}
\end{aligned}$$

Note that if we choose the trivial stratification $s(x) = x$, we stratify completely on all K unique levels of $\mathcal{X}$.

Below we present a proposition that articulates when an unbiased estimator that stratifies beyond the minimal adjustment set, $\lambda(X)$, improves upon the minimal estimator by reducing estimator variance.

Proposition 2. *Consider an adjustment set $s(X)$ which satisfies mean conditional unconfoundedness and, for at least two $x, x' \in \mathcal{X}$, $s(x) \neq s(x')$ while $\lambda(x) = \lambda(x')$. Define $\bar{\tau}^{\lambda}_{strat}$ as a stratification estimator based on the unique levels of $\lambda(X)$ and $\bar{\tau}^{s}_{strat}$ as a stratification estimator based on the unique levels of $s(X)$. Then $Var\left(\bar{\tau}^{s}_{strat}\right) <$*

$Var\left(\bar{\tau}^{\lambda}_{strat}\right)$ *if* $\nu < \eta$ *where*

$$m(j) = |\{s(x) : x \in \mathcal{X} \text{ such that } \lambda(x) = j\}|$$

$$\mathcal{B} = \{j \in \lambda(\mathcal{X}) : m(j) > 1, \textit{sub-strata means and variances are constant}\}$$

$$\mathcal{C} = \{j \in \lambda(\mathcal{X}) : m(j) > 1, \textit{sub-strata means and variances are non-constant}\}$$

$$\nu = \sum_{b \in \mathcal{B}} \left[Var\left(\frac{N_b}{n} \left(\bar{Y}_{b,1} - \bar{Y}_{b,0} \right) \right) - Var\left(\sum_{\ell=1}^{m(b)} \frac{N_{b\ell}}{n} \left(\bar{Y}_{b\ell,1} - \bar{Y}_{b\ell,0} \right) \right) \right]$$

$$\eta = \sum_{c \in \mathcal{C}} \left[Var\left(\sum_{\ell=1}^{m(c)} \frac{N_{c\ell}}{n} \left(\bar{Y}_{c\ell,1} - \bar{Y}_{c\ell,0} \right) \right) - Var\left(\frac{N_c}{n} \left(\bar{Y}_{c,1} - \bar{Y}_{c,0} \right) \right) \right]$$

and $Var(\bar{\tau}^{s}_{strat}) \geq Var\left(\bar{\tau}^{\lambda}_{strat}\right)$ *otherwise.*

A detailed proof is provided in Appendix A, but here we offer a sketch of the proof to build intuition. In comparing two stratifications, λ and s, across discrete covariates X, we can partition the level sets of the two stratfication functions as follows:

1. $\mathcal{A}$: values of $x \in \mathcal{X}$ for which both λ and s agree
2. $\mathcal{B}$: values of $x \in \mathcal{X}$ for which s substratifies λ but the mean and variance of $Y \mid Z$ are constant across substrata formed by s
3. $\mathcal{C}$: values of $x \in \mathcal{X}$ for which s substratifies λ and either the mean of $Y \mid Z$, the variance of $Y \mid Z$, or both vary across substrata formed by s

We ignore $\mathcal{A}$ and focus on $\mathcal{B}$ and $\mathcal{C}$. In the case of $\mathcal{B}$, s performs "unnecessary" stratification, estimating and re-aggregating conditional means which are the same in the underlying data generating process, and thus incurs additional variance over the λ stratification estimator. On the other hand, when we consider $\mathcal{C}$, λ incurs additional variance over s by failing to control for differences in $Y \mid Z$.

We note that many of the core insights of Proposition 2 are known to researchers and presented in various forms in the literature. Rotnitzky *et al.* (2010) show that

marginal structural models can achieve variance reduction by removing instruments from the propensity score. Henckel *et al.* (2022) derive and prove theorems about efficient adjustment sets in linear causal graphs. Rotnitzky and Smucler (2020) derive similar results for nonparametric models on causal graphs. Witte *et al.* (2020) develop an graph-based algorithm for learning efficient adjustment sets from data. Cinelli *et al.* (2020) shows using graphs that what we call "prognostic" adjustment can reduce variance and what we call "instrumental" adjustment can increase variance, and Hernan and Robins (2022) explain a similar phenomenon. The literature on sampling theory is also clear about when stratified sampling schemes (different from the type of post-stratified estimator we study here) can reduce estimator variance for population parameters (see for example Lohr (2019)). We do not present this proposition for the sake of "claiming" its content. Rather, we seek to express several well-understood insights about adjustment in causal inference in the context of our problem of interest in a way that is not done precisely in any of the above citations.

3.4 Bias-Variance Tradeoff and Regularization-Induced Confounding

The previous section compared two unbiased estimators, addressing the question of when an estimator with a larger adjustment set might yield a variance reduction. A common insight in statistics and machine learning is that biased estimators can effect a large enough variance reduction to attain a favorable estimator mean-squared error (see for example Chapter 4 of Murphy (2022)). To what extent can this insight be put to use for our purposes? We argue that the mean squared error (MSE) of $\tau(X)$ should be minimized. In practical settings, in which X is continuous, or X is discrete and high-dimensional, some manner of regularized machine learning must be employed to stabilize estimation of the ATE. Considering only estimators that are strictly unbiased can yield intolerably high variance, as we will see in Chapter 4.

Great care must be taken in this case to avoid a phenomenon described as "regularization induced confounding" (RIC) in Hahn *et al.* (2018) and Hahn *et al.* (2020). This phenomenon is specific to the problem of regularized *estimation*, not prediction (for a thorough discussion of the difference, see Efron (2020)). If an observed sample of $(\tau(X), X)$ pairs were available, many supervised learning methods for predicting $\tau(X)$ from X with regularization are available and can be tuned to minimize the MSE of $\tau(X) \mid X$. In such cases, focusing on the prediction MSE of $\tau(X)$ introduces a direct tradeoff between its bias and variance. In causal inference problems, however, $\tau(X)$ is an unobserved function, and it must be estimated in some fashion from (Y, Z, X) pairs.

There are many approaches to nonparametric estimation of $\tau(X)$, many of which (explicitly or implicitly) estimate both $\mu(X)$ and $\tau(X)$. We will review machine learning estimators for causal inference in more depth in the next chapter, and it is important to underscore that there *are* methods that estimate $\tau(X)$ using a different approach. This presentation simply illustrates the phenomenon of RIC for a common class of methods that estimate $\mu(X)$ and $\tau(X)$ jointly. We refer to estimators of μ and τ as $\hat{\mu}(X)$ and $\hat{\tau}(X)$. Since the true functions $\mu(X)$ and $\tau(X)$ are unobserved, estimation typically proceeds by exploiting the fact that setting $\hat{Y} = \hat{\mu}(X) + Z\hat{\tau}(X)$ allows $\hat{Y}$ to be "scored" against Y.

If X is discrete, then both $\mu(X)$ and $\tau(X)$ may be estimated from the entire covariate set X, but as we have seen in this chapter, it may be desirable to consider a "coarsened" adjustment set $s(X)$. If X is continuous, then an adjustment set $s(X)$ typically must be specified in terms of basis functions used to estimate $\mu(X)$ and $\tau(X)$ (setting $s(X) = X$ as a linear basis with continuous X, for example, is a stronger modeling assumption than the "no coarsening" effect of setting $s(X) = X$ in the discrete case).

In many machine learning methods, the procedure for defining $s(X)$ is a crucial aspect of the regularization offered. In the case of discrete covariates, when $s(X)$ constitutes a stratification of the unique levels of X, estimation of $\hat{\mu}(s(X))$ and $\hat{\tau}(s(X))$ may proceed straightforwardly from a saturated linear model of Y on Z interacted with $\tilde{S}$, a transformation of the strata of $s(X)$ into a regression matrix (assuming that every unique level of $s(X)$ has at least one treated and one untreated observation).

Thus, $\hat{\mu}(s(X))$ and $\hat{\tau}(s(X))$ are completely characterized by the specification of $s(X)$, and we refer to the combined outcome prediction as $\hat{Y}_s = \hat{\mu}(s(X)) + Z\hat{\tau}(s(X))$. This estimation problem can thus be specified as an optimization problem with respect to s, where we seek

$$s^* = \arg\min_s \mathbb{E}\left(\hat{Y}_s - Y\right)^2.$$

One insight of regression theory (see for example Rencher and Schaalje (2008)) is that the MSE of a regression model is nonincreasing in the number of covariates, so that the objective above is minimized by stratifying on *all of* X. Consider instead attempting to minimize $\mathbb{E}\left(\hat{Y}_s - Y\right)^2 + \alpha|s(X)|$, mirroring the LASSO penalty of Tibshirani (1996). This will favor small adjustment sets s that do not substantially increase $\mathbb{E}\left(\hat{Y}_s - Y\right)^2$. The expansion of $\mathbb{E}\left(\hat{Y}_s - Y\right)^2$ below (derived in detail in Appendix C) shows that a simple "size-based regularization" can have very unpredictable effects on the MSE of $\tau(X)$.

$\hat{\tau}(s(X))$ can be decomposed into its mean, the stratification estimator $\hat{\tau}_s$, and an offset $\hat{t}(s(X))$. We can also decompose the "true" $\tau(X)$ into its mean, the average treatment effect $\bar{\tau}_x$ and an offset $t(X)$. Defining $\hat{Y}_s$ as above, we have that $\mathbb{E}\left(\hat{Y}_s - Y\right)^2$

decomposes into

$$\begin{aligned}
\mathbb{E}\left(\hat{Y}_s - Y\right)^2 = \mathbb{E}\left(\pi(X)\right)\left[\text{Var}\left(\hat{\tau}_s\right) + \text{ Bias}\left(\hat{\tau}_s\right)^2\right] \\
&+ \text{Var}\left(\hat{\mu}(s(X))\right) + \text{ Bias}\left(\hat{\mu}(s(X))\right)^2 \\
&+ \mathbb{E}\left[Z\left(\hat{t}(s(X)) - t(X)\right)^2\right] \\
&+ 2\mathbb{E}\left[Z\left(\hat{\mu}(s(X)) - \mu(X)\right)\left(\hat{\tau}(s(X)) - \tau(X)\right)\right] \\
&+ 2\mathbb{E}\left[Z\left(\hat{\tau}_s - \bar{\tau}_x\right)\left(\hat{t}(s(X)) - t(X)\right)\right] \\
&+ 2\mathbb{E}\left[Z\left(\mu(X) + Z\tau(X) - Y\right)\left(\hat{\mu}(X) + Z\hat{\tau}(X) - \mu(X) - Z\tau(X)\right)\right] \\
&+ \sigma^2
\end{aligned}$$

Here the constrained search for small s adjustment sets will aim to minimize this sum and will implicitly prioritize controlling the values of the terms with the largest magnitude. Thus, rather than simply trading off bias and variance in $\hat{\tau}_s$, we trade off the bias of $\hat{\tau}_s$, with the variance of $\hat{\tau}_s$, the bias and variance of $\hat{\mu}(s(X))$, and several conditional product expectations involving $\hat{\mu}(s(X))$, $\hat{\tau}_s$, and $\hat{t}(s(X))$. While this example covers only one method of regularization, it shows that naive application of regularized prediction to a causal effect estimation problem can result in an estimator whose bias and variance depend on the properties of the true data generating process, in particular, the functions $\mu(X)$, $\tau(X)$, and $\pi(X)$. *This* is the core of the "regularization-induced confounding" phenomenon — a mode of regularization can have a completely benign impact on estimators of $\bar{\tau}$ or a disastrous impact, depending on aspects of the data which are unknown *a priori* to the analyst.

To give two concrete examples of how a data generating process may encourage estimators that penalize $|s(X)|$ to select adjustment sets that confound $\bar{\tau}_s$, consider:

- $\mu(X)$ is large in magnitude to $\tau(X)$: in this case, selecting strata that reduce the variance of $\hat{\mu}(s(X))$ may decrease $\mathbb{E}\left(\hat{Y}_s - Y\right)^2$ by much more than selecting strata that reduce $\left[\text{Var}\left(\bar{\tau}_s\right) + \text{ Bias}\left(\bar{\tau}_s\right)^2\right]$ (and thus help to deconfound $\bar{\tau}_s$)

- A confounded $s(X)$ makes $\mathbb{E}\left[Z\left(\hat{\mu}(s(X)) - \mu(X)\right)\left(\hat{\tau}(s(X)) - \tau(X)\right)\right]$ negative: in this case, trading off "deconfounding" stratification with such an $s(X)$ will depend on the magnitude of $\mathbb{E}\left[Z\left(\hat{\mu}(s(X)) - \mu(X)\right)\left(\hat{\tau}(s(X)) - \tau(X)\right)\right]$ compared to that of $\text{Var}\left(\hat{\tau}_s\right) + \text{ Bias}\left(\hat{\tau}_s\right)^2$.

The first case is discussed at length in Hahn *et al.* (2020). These examples are not a purely analytical curiosity and have many plausible occurrences in real life settings. For example, in health and social science settings, outcomes such as blood pressure or test scores are marked by high noise and heterogeneity, and common interventions may have a comparatively small effect.

3.5 Adjusting for Different Control Sets in Estimation of $\mu(X)$ and $\tau(X)$

Our definition of mean conditional unconfoundedness in Section 2.5.3 assumed that a common adjustment set would be used in estimating $\mu(X)$ and $\tau(X)$ on the entire set of (Y, Z, X) observations. In fact, we can show that the ATE is identifiable without using the exact same adjustment set $s(X)$ for estimating both $\mu(X)$ and $\tau(X)$. Let $s_1(X)$ refer to the adjustment set used for $\mu(X)$ and $s_2(X)$ refer to the adjustment set used for $\tau(X)$. A sufficient condition for identifying the ATE given both s_1 and s_2 is that

$$\mu(X) \perp\!\!\!\perp Z \mid s_1(X)$$

$$\tau(X) \perp\!\!\!\perp Z \mid s_2(X)$$

The first condition implies that

$$\begin{aligned}\mathbb{E}\left[Y \mid Z = 0, s_1(X)\right] &= \mathbb{E}\left[\mu(X) + \nu(X, \epsilon_y) \mid Z = 0, s_1(X)\right] \\ &= \mathbb{E}\left[\mu(X) + \nu(X, \epsilon_y) \mid s_1(X)\right] = \mathbb{E}\left[\mu(X) \mid s_1(X)\right],\end{aligned}$$

after setting $R = Y - \mathbb{E}\left[\mu(X) \mid s_1(X)\right]$, we have by the second condition that

$$\begin{aligned}
\mathbb{E}\left[R \mid Z = 1, s_2(X)\right] &= \mathbb{E}\left[\mu(X) + \tau(X) + \nu(X, \epsilon_y) + \delta(X, \epsilon_y) \mid Z = 1, s_2(X)\right] \\
&\quad - \mathbb{E}\left[\mathbb{E}\left[\mu(X) \mid s_1(X)\right] \mid Z = 1, s_2(X)\right] \\
&= \mathbb{E}\left[\mu(X) - \mathbb{E}\left[\mu(X) \mid s_1(X)\right] \mid Z = 1, s_2(X)\right] \\
&\quad + \mathbb{E}\left[\tau(X) + \nu(X, \epsilon_y) + \delta(X, \epsilon_y) \mid s_2(X)\right] \\
&= \mathbb{E}\left[\mu(X) - \mathbb{E}\left[\mu(X) \mid Z = 1, s_1(X)\right] \mid Z = 1, s_2(X)\right] \\
&\quad + \mathbb{E}\left[\tau(X) \mid s_2(X)\right].
\end{aligned}$$

and thus

$$\mathbb{E}\left[\mathbb{E}\left[R \mid Z = 1, s_2(X)\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\tau(X) \mid s_2(X)\right]\right] = \mathbb{E}\left[\tau(X)\right]$$

3.5.1 Two Stage Estimation of the Average Treatment Effect

This introduces the possibility of two-stage nonparametric estimation procedure for the ATE. A $\mu(X)$ model can be fit (and cross-validated) to (Y, X) on the control dataset using any nonparametric method. The resulting model can be used to obtain estimates $\hat{\mu}(X)$ on the treated dataset and then a $\tau(X)$ model can be fit to $(Y - \hat{\mu}(X), X)$ on the treated dataset.

This is similar at first glance to the "T-Learner" meta-algorithm which we will discuss in the following chapter, but a key difference is that the T-Learner estimates $\mu(X)$ using (Y, X) on the control dataset, then estimates $\mu(X) + \tau(X)$ using (Y, X) on the treated dataset, and estimates $\tau(X)$ as the difference in predictions. The residualized approach in this section estimates $\tau(X)$ directly in the second model, allowing analysts to separate regularization of $\mu(X)$ and $\tau(X)$. The desirability of separately regularizing μ and τ is discussed at length in Hahn *et al.* (2020) and it stems from the possibility of the two functions varying drastically in magnitude and complexity.

3.5.2 Separate Regularization of $\mu(X)$ and $\tau(X)$

In the simplified example presented in Section 3.4, we can use this insight about separate adjustment sets for μ and τ to partially address regularization-induced confounding. One of the scenarios discussed above, in which RIC can be extreme, involves a $\mu(X)$ term which is much larger in magnitude than the $\tau(X)$ term. In applications where this is suspected to be the case, analysts can apply a comparatively weak penalty to $|s_1(X)|$ than to $|s_2(X)|$. This parallels the approach taken in BCF (Hahn *et al.* (2020)), in which $\mu(X)$ and $\tau(X)$ are fit using BART models with different degrees of regularization on their respective tree ensembles.

Chapter 4

PRACTICAL MACHINE LEARNING APPLICATIONS OF FEATURE SELECTION INSIGHTS

4.1 Causal Effect Estimation via Nonparametric Function Estimation

Künzel *et al.* (2019) provide a detailed overview of several "meta-algorithms" for estimating the conditional average treatment effect function, $\tau(X)$. We review each such method along with several modifications briefly here. First, they introduce and label two methods that have been explored in the literature. The "S-Learner" learns a function $f(X, Z)$ using all sample observations (Y_i, X_i, Z_i) and then estimates $\hat{\tau}(X)$ as $f(X, 1) - f(X, 0)$. For one concrete example of such an estimator, Hill (2011) estimate the nonparametric f term using Bayesian Additive Regression Trees (BART, Chipman *et al.* (2010)). The "T-Learner" learns two functions: $f_0(X)$ using control observations $(Y_i, X_i) \mid Z_i = 0$ and $f_1(X)$ using treated observations $(Y_i, X_i) \mid Z_i = 1$, and then estimates $\tau(X)$ as their difference $f_1(X) - f_0(X)$.

The "X-Learner" originates with Künzel *et al.* (2019) and it follows several estimation steps:

1) Estimate $f_0(X)$ using control observations $(Y_i, X_i) \mid Z_i = 0$ and $f_1(X)$ using treated observations $(Y_i, X_i) \mid Z_i = 1$

2) Estimate $h_0(X)$ using control observations $(\hat{f}_1(X_i) - Y_i, X_i) \mid Z_i = 0$ and $h_1(X)$ using treated observations $(Y_i - \hat{f}_0(X_i), X_i) \mid Z_i = 1$

3) Estimate $\tau(X)$ by finding a weighting function $g(X)$ to minimize the variance of $g(X)h_0(X) + (1 - g(X))h_1(X)$ (the authors note that setting $g(X)$ equal to the

propensity score appears to work well empirically)

Kennedy (2022) defines and introduces a similar method called the "DR-Learner." This method splits a dataset at random into two folds and fits models of $Y|X, Z = 1$, $Y|X, Z = 0$, and $Z|X$ on the first fold. We refer to estimates from this first step as $\hat{Y}_1$, $\hat{Y}_0$, and $\hat{\pi}$, respectively. On the second fold, a "pseudo-outcome" is constructed as

$$\tilde{Y} = \frac{Z - \hat{\pi}}{\hat{\pi}(1 - \hat{\pi})} \left(Y - Z\hat{Y}_1 - (1 - Z)\hat{Y}_0 \right) + \hat{Y}_1 - \hat{Y}_0,$$

and $\tau(X)$ is estimated on this fold via a model of $\tilde{Y} \mid X$.

Künzel *et al.* (2019) also introduce the "F-Learner," which estimates $\pi(X)$ on (Z_i, X_i) and then estimates $\tau(X)$ on $\left(\frac{Z_i - \hat{\pi}(X_i)}{\hat{\pi}(X_i)(1 - \hat{\pi}(X_i))} Y_i, X_i \right)$, and the "U-Learner" estimates $f(X)$ on (Y_i, X_i) then estimates $\pi(X)$ on (Z_i, X_i) and then estimates $\tau(X)$ on $\left(\frac{Y_i - \hat{f}(X_i)}{Z_i - \hat{\pi}(X_i)}, X_i \right)$. Our experimentation revealed these estimators to be quite unstable on data-generating processes with strong selection (due to the inclusion of π or $Z - \pi$ in the denominator), so we do not discuss these methods in great depth. We do note, however, that one aspect of the U-Learner — fitting a marginal $\hat{Y}$ model and a marginal $\hat{Z}$ model — is a key component of several common causal machine learning estimators. Specifically, this includes "double machine learning" (DML, Chernozhukov *et al.* (2022)) and the causal forest (Athey and Wager (2019)).

We will incorporate the causal forest into many of the experiments in this chapter, so we describe it here at length. The line of work culminating in the R implementation of the causal forest algorithm has its origins in the causal tree method and theory introduced in Athey and Imbens (2015). Wager and Athey (2018) develop a modified random forest algorithm and asymptotic efficiency theory for conditional average treatment effects. Athey *et al.* (2019) introduce the "generalized random forest" (GRF), of which the causal forest is one of several application-focused imple-

mentations. First, GRF substitutes $\tilde{Y} = Y - \hat{m}(X)$ as outcome and $\tilde{Z} = Z - \hat{\pi}(X)$ as treatment, where $\hat{m}(X)$ is an estimate of $\mathbb{E}[Y \mid Z]$. Next, GRF trains a random forest (Breiman (2001)), and in each tree, it splits the sample dataset $(\tilde{Y}, \tilde{Z}, X)$ into two folds. The first fold is used to grow a tree recursively on a "pseudo-outcome" re-defined at each node as

$$R_\nu = (\tilde{Z}_i - \bar{Z}_\nu)((\tilde{Y}_i - \bar{Y}_\nu) - \bar{\tau}_\nu(\tilde{Z}_i - \bar{Z}_\nu)),$$

where ν refers to a node, $\bar{Z}_\nu$ is the average $\tilde{Z}$ within node ν, $\bar{Y}_\nu$ is the average $\tilde{Y}$ within node ν, and $\bar{\tau}_\nu$ is the coefficient of a regression of $\tilde{Y} - \bar{Y}_\nu$ on $\tilde{Z} - \bar{Z}_\nu$ within the samples in node ν. Finally, weights are computed based on shared membership in the tree leaves using the second fold of each tree, and those weights define a kernel used to estimate $\tau(X)$.

The "Bayesian causal forest" (BCF) method of Hahn *et al.* (2020) defines a non-parametric Bayesian model of $\tau(X)$, which jointly estimates functions $f(X)$, $g(X)$ and parameters a, b_0, and b_1 on (Y_i, Z_i, X_i) so that

$$\begin{aligned}\mathbb{E}[Y_i \mid Z_i, X_i] &= af(X_i) + (b_0(1 - Z_i) + b_1 Z_i)\, g(X_i) \\ &= [af(X_i) + b_0 g(X_i)] + (b_1 - b_0)\, g(X_i) Z_i,\end{aligned}$$

with both f and g sampled using BART. After simulating posterior samples of f, g, a, b_0 and b_1, we estimate $\tau(X)$ as $(b_1 - b_0)\, g(X)$. BCF is typically sampled with $\pi(X)$ included as a covariate in either or both of the $f(X_i)$ and $g(X_i)$ models, which is particularly useful in cases of "targeted selection," where treatment is assigned based on the expected untreated outcome. Krantsevich *et al.* (2022) introduce a modified BCF estimator — XBCF — that replaces the BART models for f and g with XBART, a fast algorithmic approximation of BART (He and Hahn (2021)).

We do not claim that these estimators are "exhaustive" of approaches for estimating $\tau(X)$ using machine learning. Other approaches not reviewed in depth in this

dissertation include Wager *et al.* (2016), Bradic *et al.* (2019), Oberst *et al.* (2021), Johansson *et al.* (2020), Johansson *et al.* (2016), Ju *et al.* (2019), Zhang *et al.* (2022), Shi *et al.* (2019). Many of these methods are focused on high-dimensional penalized linear regression, making better use of propensity scores in high dimensions, or defining neural network architectures to help estimate causal effects. This chapter is focused more narrowly on understanding the inductive biases, explicit or implicit, of different causal inference methods on realistic (but simulated) low-n tabular data settings. Tree ensembles are common and highly performant nonparametric method for prediction with tabular data (Grinsztajn *et al.* (2022)), so we focus on comparing methods which either directly use tree ensembles, such as XBCF and GRF, or can be made to use tree ensembles, such as the S–, T–, X–, and DR–Learners.

4.2 Enhanced Regularization Targeting the Average Treatment Effect

As noted in Section 2.5.3, the average treatment effect is identified with the narrower assumption of *mean conditional unconfoundedness*, or that both $\mu(X)$ and $\tau(X)$ are independent of Z given an adjustment set $s(X)$. This desired feature selection may be articulated in the context of nonparametric estimation as follows. BCF estimates functions $f(X)$ and $g(X)$ which together approximate $\mu(X)$ and $\tau(X)$. This estimation method allows for separate regularization of the two structural components of the response surface. Hahn *et al.* (2020) suggest modest regularization on $g(X)$ and comparatively weaker regularization on $f(X)$. They use BART models to estimate f and g and they control the respective regularization through the number of trees used in each ensemble as well as the tree depth parameters, α and β. They also recommend including $\pi(X)$ as a covariate in the $f(X)$ model to manage targeted selection — if treatment selection is informed by expected untreated outcomes, so that $\pi(X)$ and $\mu(X)$ are closely correlated, then inference on $\mu(X)$ can be improved

by including $\pi(X)$ as a covariate.

The use of $\pi(X)$ may improve inferences on $\mu(X)$ in cases of targeted selection, but an insight of Chapter 3 is that $\pi(X)$ is a strictly larger adjustment set than is necessary to identify the average treatment effect. In the discussion of Hahn *et al.* (2020), Ray *et al.* (2020) show that a DGP with strong instruments can lead BCF to underperform other methods, since $\pi(X)$ is not informative about $\mu(X)$ when it includes several strong instruments. Identifying and removing instruments is challenging in a data-adaptive machine learning setting. With a priori knowledge of $\mu(X)$, $\tau(X)$, and $\pi(X)$, we could attempt to identify instruments by conducting conditional incremental independence tests on each of the variables in a dataset. However, in practice, we must typically estimate $\mu(X)$ and $\tau(X)$, even if $\pi(X)$ is known because of an experimental design.

Estimating $\mu(X)$ and $\tau(X)$, so that we can identify instruments and remove them from $\pi(X)$ and then estimate $\mu(X)$ and $\tau(X)$ with this modified $\pi(X)$ is problematic for several reasons. First, in many settings including the Bayesian approach undertaken in BCF, it requires care to avoid feedback between the estimation procedures for $\pi(X)$ and any feature selection procedure used as a precursor to subsequent estimation procedures for the average causal effect (as discussed in Zigler *et al.* (2013), Zigler and Dominici (2014), and Wang *et al.* (2015)). Second, if one were able to construct initial models of $\mu(X)$ and $\tau(X)$ that were accurate and precise enough to provide reliable representations that enable valid instrument selection, then there is little practical need to proceed to the second stage of estimate the treatment effect using a dimension-reduced adjustment set! In short, one specific case in which estimation of $\mu(X)$ and $\tau(X)$ is difficult in finite samples due to the presence of strong instruments is also the case in which identification and removal of those instruments using a first-stage model will be unreliable.

What alternatives are available for removing instruments? The theory discussed Chapter 3 helps us to understand *that* the general goal of causal feature selection is to remove instruments without removing confounders or prognostic variables, but it is silent on *how*. Several causal variable selection methods suggest a conservative approach which selects instruments, confounders and prognostic variables (Belloni *et al.* (2014), Shortreed and Ertefaie (2017)). De Luna *et al.* (2011) propose a method for selecting confounding variables via a series of conditional independence tests. Vansteelandt *et al.* (2012) define a stochastic search procedure designed to screen confounders while controlling the estimator MSE. Zigler and Dominici (2014) proposes a Bayesian method for selecting confounders that uses model averaging to address "feedback" issues with Bayesian approaches that jointly model the outcome and propensity score. Wilson and Reich (2014) introduce a penalized regression method to select confounders in joint linear models of treatment and outcome. Schnitzer *et al.* (2016) propose an iterative method to screen instruments by forward selection of a propensity model using a collaborate targeted minimum-loss estimation procedure (C-TMLE). Häggström (2018) learn a graphical model from observed data and select confounders based on the resulting graph.

We note however that many of these approaches are either focused on axis-aligned variable selection (thus limiting the available feature selection functions, $s(X)$), are very computationally demanding, or require large datasets. Because we are interested in algorithms that "learn" basis functions which minimize the MSE of $\tau(X)$, we focus our potential improvements on refinements of the XBCF / BCF family of estimators.

Rather than attempt to deliberately test for and identify instruments, a computationally feasible alternative that is the subject of our early experiments is to compute $\pi(s(X))$ using multiple subsets of the full covariate set, each of which we denote as $s(X)$. We achieve this by first sampling a number k between 1 and $p-1$ and then

sampling k covariates from $\{1, \ldots, p\}$ without replacement. Letting S_k refer to this sampled set of covariates, we compute $\pi(S_k) = \mathbb{E}[Z \mid S_k]$ and include $\pi(S_k)$ in the matrix of covariates.

We let Π_X refer to a matrix of propensity scores including the "full propensity" $\pi(X)$ and a user-defined number of "propensity submodels" $\pi(s(X))$ for different feature subsets s. We define the "Multiple Propensity XBCF" (XBCF-MP) estimator as an XBCF model where Π_X augments X as the covariate matrix in both the f and g terms, so that

$$\mathbb{E}[Y \mid X, Z] = [af(X, \Pi_X) + b_0 g(X, \Pi_X)] + (b_1 - b_0) g(X, \Pi_X) Z$$

4.3 Principles for Comparing Regularized Causal Effect Estimators

Several of the methods introduced in Section 4.1, notably GRF, the X-Learner, and the DR-Learner, are presented by their authors alongside large-sample theory demonstrating either semiparametric efficiency or pointwise MSE convergence. These results are impressive, not only for their technical sophistication but also in articulating the specific mathematical conditions under which an estimator performs optimally. Given how much these results vary in what they prove about each estimator, it would be challenging and, in our opinion, not very illuminating to attempt to compare estimators based on their large sample theory. Furthermore, the sample sizes required for such asymptotic results to hold are typically so large that the benefits of different regularization or feature selection schemes are not obvious.

But what are the alternatives to large sample study of nonparametric estimators? Finite sample results are difficult to establish, though there is an emerging literature of such results using conformal inference (Lei and Candès (2021)). This literature is promising and we suspect that, as these methods develop, they will become part of the applied toolkit for researchers and analysts working doing causal effect estimation.

The focus of this dissertation, however, will be to compare methods — and guide future method development — using simulation studies.

It is worth taking time to dwell on this topic and its philosophical importance in this dissertation. It is certainly true that simulation studies do not offer conclusive "proof" of a method's performance beyond the specific data generating processes under review. Indeed, simulation studies can be designed (even "cherry-picked") to give one method an advantage over other methods. Care must be taken to design meaningful and informative simulation studies, for which the goal is experimentation — using the simulations to test hypotheses and offer insights about the "inductive biases" of different methods.

The discussion of Hahn *et al.* (2020) highlights the importance of designing simulation studies with real world characteristics in mind. Motivated by applications in the social and health sciences, they focus on data generating processes with:

1. **High noise**: variation in Y accounted for by $\mu(X) + Z\tau(X)$ is small
2. **Large relative prognostic effects**: variation in $\mu(X) + Z\tau(X)$ accounted for by $\tau(X)$ is small
3. **Targeted selection**: $\mu(X)$ is closely related to $\pi(X)$.

This is important as it motivates the development of simulation studies in which treatment effects are neither "too easy" to learn, in which case most methods would present similar results, nor "too hard" to learn, in which case most methods would also present similar results. Along these lines, Knaus *et al.* (2021) develop an "empirical monte carlo" strategy of evaluating estimators using simulations informed by real data. Curth and van der Schaar (2021) and Curth and van der Schaar (2023) discuss issues in comparing estimators against common "benchmark" causal inference

datasets (such as ACIC and IHDP) and offer guidance and examples on designing simulation studies with experimentation in mind.

This dissertation contends that carefully crafted simulation studies are useful in comparing methods, and they are also useful in *designing* new causal effect estimation methods. Broadly, the progression of simulation studies in the next few sections is such that we will run simulations, observe some results, offer some hypotheses, and evaluate the hypotheses in the following section, typically by making changes to one or several estimators. The end result is several practical recommendations for improving estimators "out-of-the-box" as well as promising research directions for low-level implementations of new methods.

4.4 Data Generating Processes for Experimentation

We consider several data-generating processes (DGPs) for simulation studies comparing the estimators introduced above.

4.4.1 Targeted Selection with Small Treatment Effects and Large Prognostic Effects

We define a modified version of the "targeted selection" DGP reviewed in Hahn *et al.* (2020), in which the selection probability is closely influenced by the expected untreated potential outcome, $\mu(X)$, and the magnitude of the treatment effect, $\tau(X)$, is smaller in comparison to the prognostic effect. This mirrors the reality of many

observational causal inference problems in the health and social sciences.

$$
\begin{aligned}
X_1, X_3, X_6, \ldots, X_p &\sim \mathcal{N}(0,1) \\
X_2, X_5 &\sim \text{Bernoulli}\left(\frac{1}{2}\right) \\
X_3 &\sim \text{Categorical}\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \\
\mu(X) &= 1 + 4X_1(2X_2 - 1) - 4\mathcal{I}\left(X_3 = 1\right) + 4\mathcal{I}\left(X_3 = 3\right) \\
\tau(X) &= \frac{1}{2} + \frac{1}{2}X_1 + \frac{1}{2}(2X_2 - 1) \\
\pi(X) &= \Phi\left(\frac{-\bar{\mu} + \mu(X) + 2X_4 - 2(2X_5 - 1)}{9}\right) \\
Z &\sim \text{Bernoulli}\left(\pi(X)\right) \\
Y &\sim \mathcal{N}(\mu(X) + \tau(X)Z, (\kappa s)^2)
\end{aligned}
$$

where $\bar{\mu}$ is the sample mean of $\mu(X)$, s^2 is the sample variance of $\mu(X) + \tau(X)Z$ and κ is a scale multiple that determines the noise-to-signal ratio of the DGP. This DGP has several features that make it ideal for experimental comparison of causal estimators:

1. Strong confounding: X_1, X_2 and X_3 have a strong confounding effect between $\pi(X)$ and $\mu(X)$ and a modest confounding effect between $\pi(X)$ and $\tau(X)$

2. Targeted selection: $\pi(X)$ is a noisy function of $\mu(X)$, so that the distribution of $\mu(X)$ learnable from $(Y, X) \mid Z = 0$ may differ strongly from the distribution of $\mu(X)$ learnable from $(Y, X) \mid Z = 1$

3. Weak instruments: X_4 and X_5 are both modest instrumental variables, whose effect is not as strong as the confounding effects of X_1 through X_3

4. Noise variables: X_6 through X_p (p itself will vary in simulation experiments) have no relationship with Y or Z and should thus be removed from the adjustment set if possible.

We will refer to this DGP as "DGP 1" for shorthand in the presentation of results.

4.4.2 Strong Instruments, Strong Prognostic Effects, Modest Confounding

We now consider a DGP with strong instruments, strong prognostic effects, comparatively weak confounding. This DGP underscores the importance of necessary dimension reduction — adjusting for instrumental variables will inflate variance with no deconfounding effect and failing to adjust for prognostic effects will inflate variance.

$$
\begin{aligned}
X_1, \ldots, X_p &\sim \text{Uniform}\,(0,1) \\
\mu(X) &= 3\sin(2\pi X_1) + 6\cos(2\pi X_2) + 6\sin(2\pi X_3) \\
\tau(X) &= \frac{1}{2} + 2\left(X_4 - \frac{1}{2}\right) + \frac{1}{2}\left(X_5 - \frac{1}{2}\right) \\
\pi(X) &= \Phi\left((X_1 - 0.5) + (X_4 - 0.5) + 5\,(X_6 - 0.5) + 5\,(X_7 - 0.5) + 5\,(X_8 - 0.5)\right) \\
Z &\sim \text{Bernoulli}\,(\pi(X)) \\
Y &\sim \mathcal{N}(\mu(X) + \tau(X)Z, (\kappa s)^2)
\end{aligned}
$$

where s^2 is the sample variance of $\mu(X) + \tau(X)Z$ and κ is a scale multiple that determines the noise-to-signal ratio of the DGP. We will refer to this DGP as "DGP 2" for shorthand in the presentation of results.

4.4.3 Homogeneous treatment effect, linear confounding, and prognostic functions

We now consider a DGP with modest confounding, no instruments, and modest prognostic effects. The treatment effect is homogeneous and all selection and prognostic functions are linear in the covariates. This DGP serves as a counterbalance to the complicated response surfaces presented above, seeking to address the question

of how each of these approaches perform when the effects are "easy" to estimate.

$$
\begin{aligned}
X_1, \ldots, X_p &\sim \text{Uniform}\,(0,1) \\
\mu(X) &= 3X_1 - 3X_2 \\
\tau(X) &= \frac{1}{2} \\
\pi(X) &= \frac{1 + 2X_1}{4} \\
Z &\sim \text{Bernoulli}\,(\pi(X)) \\
Y &\sim \mathcal{N}(\mu(X) + \tau(X)Z, (\kappa s)^2)
\end{aligned}
$$

where s^2 is the sample variance of $\mu(X) + \tau(X)Z$ and κ is a scale multiple that determines the noise-to-signal ratio of the DGP.

4.5 Simulation Results

4.5.1 Computational Details: Model Fitting and Treatment Effect Evaluation

In each simulation study, we run the following 7 CATE estimation method: XBCF, XBCF-MP, GRF, DR-Learner, S-Learner, T-Learner, X-Learner. Most of these methods provide estimates of $\tau(X)$, which are then converted to estimates of the average treatment effect by computing their sample average.

The code for these experiments is available at `https://github.com/andrewherren/dissertation-experiments`. Wherever an estimate of $\pi(X)$ is required, `xgboost` (Chen and Guestrin (2016)) is used with a modest degree of regularization (50 trees, learning rate of 0.05, max depth of 2 for each tree in the ensemble, and early stopping after 10 boosting rounds). When a propensity function is assumed known (as it is in some of the simulations below), $\pi(s(X))$ is evaluated by numerically integrating any variables not included in $s(X)$ out of the true $\pi(X)$ function. Whenever an estimate of $Y \mid X$ is required, XBART (He and Hahn (2021)) is used with 40 trees for 30

iterations (after 30 "burn-in" iterations) and the predictions are averaged across each of these iterations for a single set of average predicted values.

XBCF is fit with 30 trees in the $f(X)$ model, with node splitting probabilities (see Chipman *et al.* (2010)) set to $\alpha = 0.95$ and $\beta = 1.25$, and 10 trees in the $g(X)$ model, with node splitting probabilities set to $\alpha = 0.25$ and $\beta = 3$. This corresponds to a high degree of regularization on the $\tau(X)$ term and a comparatively low degree of regularization on the $\mu(X)$ term. Additionally, $\pi(X)$ is included in each of the $f(X)$ and $g(X)$ models. The same training instructions are used for XBCF-MP; the only difference is that the $\pi(X)$ used is now a matrix of many "propensity submodels." Estimates of $\tau(X)$ are obtained in each case by running the XBCF sampler for 30 iterations after 30 "burn-in" samples and averaging the $\tau(X)$ estimates.

GRF, except in the case of one of our later simulation studies, is fit using its defaults. The marginal $\hat{Y}$ model is first estimated using a random forest, and, in simulations where the propensity score is assumed unknown, the marginal $\hat{\pi}$ model is also estimated using a random forest. Then the causal forest algorithm described in Section 4.1 is fit to $Y - \hat{Y}$ and $Z - \hat{\pi}$.

The meta-algorithms (S–, T–, X–, DR–Learner) are all estimated using XBART with 40 trees for 30 iterations after 30 "burn-in" iterations. First-stage estimates that must be plugged into second-stage estimators are averaged across all 30 samples. Similarly, second-stage prediction samples are converted to estimates of $\tau(X)$ by averaging pointwise.

Our simulation studies consider each of the three DGPs in Section 4.4 with $n = 500$. For each DGP, we evaluate 1,000 simulations each across the grid of $p \in \{10, 50\}$ and $\kappa \in \{0.25, 0.5, 1, 2\}$. These evaluations range from "low dimensionality, low noise" $(p = 10, \kappa = 0.25)$ to "moderate dimensionality, high noise" $(p = 50, \kappa = 2)$. For each combination of DGP, p, and κ, we evaluate each method's root mean squared error

across the 1,000 simulations, as

$$\text{RMSE} = \sum_{j=1}^{1{,}000} \left(\bar{\tau}_j - \bar{\tau}\right)^2,$$

where $\bar{\tau}_j$ is the ATE estimate for simulation j and $\bar{\tau}$ is the true ATE.

4.5.2 Simulation Results using Method "Defaults"

Table B.1 shows the simulation results in which the true propensity function is assumed known. Here, we can see that the meta-algorithms are not competitive with XBCF and GRF. Furthermore, two things become apparent in the comparison of XBCF and the other two methods. First, the "multiple propensity" update to XBCF tends to attain a lower RMSE, likely by being able to condition on instrument free version of the propensity score in estimating $\mu(X)$. Second, GRF tends to attain an even lower RMSE in many cases. A hypothesis that we will investigate in the next round of simulations is that conditioning on an estimate of $\mathbb{E}\left[Y \mid X\right] = \mu(X) + \pi(X)\tau(X)$ allows the tree ensemble to focus on learning offsets from $\hat{Y}$ rather than learning the entire response surface and decoupling $\mu(X)$ and $\tau(X)$ at the same time.

Table B.2 shows the simulation results in which the propensity function must be estimated from a fixed sample. Here, we see that XBCF and XBCF-MP far outperform any of the other methods, and also that XBCF-MP outperforms XBCF in DGP 2, which has strong instruments. This presents the curious finding that GRF outperforms XBCF in the case where the propensity function is known (as in the case of a randomized experiment, for example) but is not competitive when regularized estimates of the propensity function are used.

4.5.3 Simulation Results with $\hat{Y}$ Estimates as XBCF Covariates

Motivated by the results of Section 4.5.2, in which GRF outperforms XBCF when the true propensity function is known, we modify the XBCF estimator to include estimates of $\mathbb{E}[Y \mid s(X)]$ for different random subsets of features drawn as $s(X)$ (including the "full model," $\mathbb{E}[Y \mid X]$, as a default). Table B.3 shows the simulation results in which the true propensity function is assumed known. Here, we see as before that the meta-algorithms are not competitive with XBCF and GRF. We now also see that XBCF attains a lower RMSE than GRF and XBCF-MP does not offer nearly the same degree of RMSE reduction as it did in Section 4.5.2.

Table B.4 shows the simulation results in which the propensity function must be estimated from a fixed sample. These results largely mirror previous results, in which regularized estimation of propensity scores impacts the performance of GRF and the meta-learners much more than that of XBCF.

4.5.4 Simulation Results with $\hat{Y}$ as XBCF Covariates and Regularized Margins in GRF

Observing that GRF takes a large performance hit when using estimated propensity score, we next wonder, is it possible to improve the performance of GRF by providing XBART estimates of $\hat{Y}$ and xgboost estimates of $\hat{\pi}$? Table B.3 shows these results when the true propensity function is assumed known. Here, we see that GRF attains a lower RMSE than XBCF. Table B.4 shows the same results in which the propensity function must be estimated from a fixed sample. Here, we see that GRF performs better than in Section 4.5.2 and 4.5.3, but still attains a higher RMSE than XBCF, suggesting that the method is highly sensitive to "good" estimates of the propensity score.

4.6 Discussion of Results and Future Directions

4.6.1 How Effective are Multiple Propensities at Reducing RMSE?

In Section 4.2, we motivated the use of multiple propensity features as targeting the *removal* of instruments. We reviewed the literature on removing instruments and noted that most approaches did not address our goals. We reasoned that estimating multiple propensity scores on randomly drawn "subsets" of features would give the tree ensembles "hints" as to propensity scores that better predict the outcome. This is a case in which our use of XBCF is beneficial, as the greedy training structure encourages splitting on the specific $\hat{\pi}(s(X))$ columns that fit the outcome well.

We can understand this effect by studying the results in Tables B.1 and B.2. When the propensity function is *known*, XBFC-MP attains a universal reduction in RMSE over XBCF, as instruments can be completely integrated out of submodels. When propensities must be estimated from (Z, X) pairs, the picture is less clear. For DGP 2, which contains several very strong instruments, XBCF-MP attains the same completely univeral RMSE reduction. For DGP 1, XBCF attains a lower RMSE (sometimes considerably lower) when the noise level (governed by κ) is low. In this case, the "full propensity" score does contain instruments, but is associated strongly enough with $\mu(X)$ that screening instruments may not be as beneficial as conditioning on the values of full $\pi(X)$ in estimating $\mu(X)$. Finally, on DGP 3, in which the estimation problem is straightforward, XBCF-MP attains a higher RMSE, likely a simple byproduct of training an algorithm with a large feature space.

This suggests that the "instrument-screening" benefits of conditioning on multiple $\pi(s(X))$ models accrue in DGPs where instruments make estimation extremely challenging (as in the example presented in Ray *et al.* (2020)).

4.6.2 What is the Effect of Conditioning on $\hat{Y}$?

Noting that

$$\mathbb{E}[Y \mid X] = \mathbb{E}[\mathbb{E}[Y \mid Z, X] \mid X] = \mathbb{E}[\mu(X) + Z\tau(X) \mid X] = \mu(X) + \pi(X)\tau(X),$$

we can see that even a perfect model for $\hat{Y}$ is not "instrument-free" in the sense that variation in $\pi(X)$ leads to variation in $\mathbb{E}[Y \mid X]$. However, we observe in Table B.1 that GRF, which uses the transformed outcome $\tilde{Y} = Y - \hat{Y}$ and the transformed treatment $\tilde{Z} - \pi(X)$, outperforms XBCF and XBCF-MP, sometimes substantially. This gap disappears in Table B.2, in which estimated propensity scores are used. Given the extreme values of the propensities in DGPs 1 and 2, estimating the propensity function with regularization may preserve the shape and ordering of the $\pi(X)$ but not the exact values. For methods like XBCF which use $\hat{\pi}(X)$ as covariates in a tree ensemble, preserving the sort order of the true $\pi(X)$ is likely more important than preserving its specific values. GRF's comparatively sharp decline in performance when using estimated propensities suggests that it depends more proximately on the actual values of $\pi(X)$.

While it is reasonable to study the performance implications of GRF's use of $\pi(X)$ compared to XBCF, we note that XBCF by default does not make *any* use of marginal $\hat{Y}$ estimates. This presented an avenue for experimentation, particularly on the studies for which GRF outperformed XBCF. Could XBCF's performance be improved through the use of $\hat{Y}$? We followed the same approach as "multiple propensity" XBCF — fitting multiple models of $Y \mid s(X)$ for different subsets s of the variables and including them as covariates in both XBCF and XBCF-MP.

We see in Table B.3 that when the true propensity function is known, conditioning on multiple $\hat{Y}$ submodels substantially narrows the performance gap between XBCF and XBCF-MP. This suggests (though it does not prove conclusively) that the in-

strument removal benefits of XBCF-MP are also satisfied by conditioning on multiple $\hat{Y}$ values. Furthermore, Table B.4 shows that XBCF and XBCF-MP with estimated propensities exhibit a considerably less conclusive performance comparison, though XBCF-MP tends to perform slightly better in low noise simulations for DGPs 1 and 2.

Finally, it is worth asking the question — what justifies using an "estimate" of Y in an XBCF model that ultimately attempts to fit Y given Z and X? We believe this procedure is appropriate for several reasons. First, the XBCF model predicts Y through two partitioned nonparametric terms $f(X)$ and $g(X)$ which combine with Z to predict Y. In some sense, the estimation task is to learn two "offsets" from the conditional mean of $Y \mid X$, so that conditioning each of these learners on $\hat{Y}$ is more akin to "centering" the estimates than "using the data twice." Second, while our simulations averaged over samples of $\hat{Y} \mid X$ estimated using XBART, it would be feasible and indeed desirable in future cases to incorporate the uncertainty inherent in our samples of $\hat{Y}$ into the second stage XBCF model. This is akin to using Jeffrey's Rule (Diaconis and Zabell (1986)), in which a probability distribution is updated by conditioning on, and then integrating out, the values of another probability distribution, as opposed to linking the two distributions formally using a Bayesian update. While this justification of XBCF-MP is well-motivated, it does preclude the ability to construct a "fully Bayesian" extension of the method (as in the correspondence of BCF and XBCF).

4.6.3 The Benefits of Access to "Unlabeled" Data

One common theme of the results in Appendix B is that estimators using true propensities outperform estimators that use estimated propensities. In the case of GRF in particular, the difference is typically quite stark. What practical implications

does this have for analysts that don't have access to the randomization protocol of an experiment? Herren and Hahn (2020) note that "unlabeled data" of treatment-covariate pairs (Z, X) can be used to estimate the propensity function with greater fidelity. These simulations demonstrate the thought experiment at the extreme limit of that phenomenon, in which an unlimited amount of treatment-control data is available to learn the propensity score perfectly.

We also note that other forms of unlabeled data can be incorporated into several of the nonparametric estimators introduced into this chapter. A large amount of "untreated" outcome-covariate pairs (Y, X), which are plausibly more abundant in many medical databases, can be used in XBCF and the T-Learner to obtain better estimates of $\mu(X)$.

4.6.4 Future Directions

These experiments have provided empirical insight on the behavior of regularized estimators on extreme-but-realistic DGPs. Here, we discuss several research directions — inspired by this work — that we are currently pursuing. First, we note that we were able to improve GRF's performance over its default (and over that of XBCF) by conditioning on regularized XBART-based estimates of $\hat{Y}$. This leads us to wonder whether aspects of the GRF estimation procedure, in particular its node-wise regression adjustment, could be incorporated into the XBCF sampling algorithm. As we see in Table B.5, GRF attains a lower RMSE than XBCF in the low noise ($\kappa < 1$) settings, so we suspect it may be the case that any modified XBCF sampler should incorporate estimates of the residual variance in deciding whether to employ such an adjustment. Nonetheless, this is an exciting avenue for future research.

Furthermore, a considerable benefit of XBCF highlighted in Krantsevich *et al.* (2022) is the ability to "warm-start" the MCMC sampler of BCF and attain better

interval coverage of $\tau(X)$ than XBCF at a much faster rate than BCF. We did not explore interval coverage of the estimators in this study, but it remains important future work to investigate the ability of XBCF (and its many modifications), GRF, and the meta-learners in quantifying uncertainty about $\tau(X)$.

Chapter 5

FUNCTIONAL ANOVA FOR MODEL EXPLAINABILITY

5.1 Introduction

Algorithmic approaches to "explaining" model predictions have proliferated as machine learning methods gain in popularity. We refer interested readers to Molnar (2022) or Arrieta *et al.* (2020) for in-depth surveys. For the purposes of this paper, we simply note that there are many high-level approaches to explaining model predictions and we focus solely on SHAP (Lundberg and Lee (2017)). SHAP is a popular "local feature attribution" method, which means it attempts to explain a model by "scoring" input feature contributions for a specific prediction. Other common examples of local attribution methods include LIME (Ribeiro *et al.* (2016)), Integrated Gradients (Sundararajan *et al.* (2017)), and GradCAM (Selvaraju *et al.* (2017)). While each of these methods deserve detailed study, this paper is a thorough investigation of the statistical properties of SHAP.

SHAP applies the Shapley value from game theory (Shapley (1953)) to model explanation by considering features as "players" in a cooperative game. Lundberg and Lee (2017) approximate Shapley values for each feature using a weighted least squares regression, where the regression weights are a transformation of the original Shapley value weights. They refer to this method (and the accompanying python library[1]) as SHAP, which is the focus of this paper. The idea of explaining a model through Shapley values has also appeared several times in earlier literature. Both Štrumbelj and Kononenko (2014) and Datta *et al.* (2016) discuss approximations to the Shapley

[1] `https://github.com/slundberg/shap`

value for explaining specific predictions. In the literature on variance-based sensitivity analysis, Shapley values have been used to allocate global model variance to specific features (Owen (2014), Song *et al.* (2016), Owen and Prieur (2017)).

SHAP is popular in industry (Bhatt *et al.* (2020)) and its reach has motivated an active literature of debates and proposed improvements. Several papers have presented modifications to SHAP that make use of the correlations between features in the training set (Aas *et al.* (2019), Frye *et al.* (2020)). Others have proposed algorithms that augment Shapley value computation with user-specified knowledge of causal patterns in the data (Datta *et al.* (2016), Frye *et al.* (2019), Wang *et al.* (2020)). Kumar *et al.* (2020) argue that many of the above methods have problems that make Shapley values an awkward fit for the problem of machine learning interpretability. Kaur *et al.* (2020) note that many professional data scientists misinterpret Shapley values. Chen *et al.* (2020) respond to many of the concerns noted above, arguing that it is up to users to figure out which variety of Shapley value is useful for their problem, but that there is nothing wrong with the general approach of using Shapley values.

This paper seeks to clarify this debate by studying the statistical properties of SHAP, where we note connections to the literature on computer experiments and sensitivity analysis. Specifically, SHAP can be represented as a partition of the components of a model's functional ANOVA decomposition (Hoeffding (1948)). This functional ANOVA lens enables a formal investigation of two problems that occur frequently in the SHAP literature:

- How many of the 2^p conditional expectations to calculate when approximating Shapley values

- The choice of a reference, or "baseline," distribution for each of the conditional

expectations

While it is well known that there are many possible variants of Shapley value (Sundararajan and Najmi (2020)), this paper clarifies the technical decisions that a SHAP user must make (or does make implicitly by using the `shap` library). Rather, we hope that this paper helps practitioners move past this debate and proceed cautiously in using SHAP for their specific model.

5.2 SHAP Overview and Notation

This section introduces both the "Shapley value" from game theory and the SHAP method in machine learning explainability. We show that SHAP attempts to define the Shapley value in the context of model interpretation and then show that this definition can be expressed as a linear combination of functional ANOVA components. We will typically use "Shapley value" in context to refer either to the original game theoretic concept, or to the estimand approximated by SHAP.

5.2.1 Shapley Value

Shapley (1953) considered cooperative games with n players, each of whom can join a coalition with 0 or more other players who will receive a collective score. In the economics literature, these scores are often utilities or monetary values, but the mathematics of cooperative games only requires that some numeric score be associated with each coalition. Let Ω refer to the set of n players and 2^Ω be the set of all possible subsets of Ω. Let S refer to an arbitrary coalition, so that $S \in 2^\Omega$. The score of a given coalition S is determined by the game's *characteristic function*, $\nu : 2^\Omega \to \mathbb{R}$. Shapley introduced the following formula, which has come to be known as the "Shapley value,"

$$\phi_i(\nu) = \sum_{S \subseteq \Omega \setminus \{i\}} \frac{(|S|)!\,(n - |S| - 1)!}{n!} \left[\nu(S \cup \{i\}) - \nu(S)\right]$$

for player i and characteristic function ν. Broadly speaking, the Shapley value is a weighted average of the contribution that player i makes to each of the 2^{n-1} coalitions that do not include player i. Shapley showed that this formula produces a unique score which satisfies the following axioms:

1. Symmetry: if $\nu(S \cup \{i\}) = \nu(S \cup \{j\})$ for all $S \in 2^{\Omega}$ and $i \neq j$, then $\phi_i(\nu) = \phi_j(\nu)$
2. Efficiency: $\sum_{i=1}^{n} \phi_i(\nu) = \nu(\Omega) - \nu(\varnothing)$
3. Additivity: For two games with characteristic functions ν and μ, $\phi_i(\nu + \mu) = \phi_i(\nu) + \phi_i(\mu)$

While it has its origins in theoretical microeconomics, the Shapley value has proven useful in modeling a variety of phenomena. We refer interested readers to Roth (1988) for a detailed discussion of the broader impact of the Shapley value.

5.2.2 SHAP: Modified Shapley Values for Model Explainability

This description of the SHAP algorithm largely follows its presentation in Lundberg and Lee (2017) and its implementation in the `shap` python library. Much of this investigation was conducted based on the SHAP codebase as implemented in 2020 and 2021. A more recent review of the codebase suggests many of the described implementation details are current, but we focus this chapter on the reviewed version of `shap`, allowing that some defaults or procedures may have changed since 2021.

Definition of Players and Coalitions

Lundberg and Lee (2017) apply Shapley's formula to model explanation by redefining of a game's players and characteristic function in terms of a trained machine learning

model. In SHAP, the p features of a model's training set are considered "players" and the characteristic function is a call to the model's prediction function. While the characteristic function in its traditional formulation is a set function, which operates on subsets of players of a game, most model prediction functions require a value for every feature that was used in training. At a high level, SHAP's solution to the problem of "including" or "excluding" features from a prediction call is to switch between a "target value" for a given feature, which would indicate that the feature was included in a coalition, and a "reference value" used for non-included features.

To see how this works in more detail, we first introduce some helpful notation and terminology. The *target* is the specific prediction that a modeler seeks to explain, and the *baseline* is a "background" covariate vector which will replace the target value for features that are "excluded" from a coalition. Since the construction of coalitions requires switching between baseline and target values, we let z refer to a binary vector where 1 indicates use of the target value. Thus, we can map a coalition S to a synthetic covariate vector, $x_{synthetic}$, as follows:

$$x_{baseline} = (b_1, b_2, ..., b_p)$$

$$x_{target} = (t_1, t_2, ..., t_p)$$

$$g_i(S) = \begin{cases} 1 & i \in S \\ 0 & i \notin S \end{cases}$$

$$z = g(S) = (g_1(S), ..., g_p(S))$$

$$h(z, x_{baseline}, x_{target}) = x_{baseline} \times (1 - z) + x_{target} \times z$$

$$x_{synthetic} = h(g(S), x_{baseline}, x_{target})$$

where the multiplication terms in the expression $x_{baseline} \times (1 - z) + x_{target} \times z$ are both element-wise.

To see this more concretely, suppose a model's training set has 3 real-valued

features. We define $x_{target} = (t_1, t_2, t_3)$ as the vector of covariates corresponding to the *target* and $x_{baseline} = (b_1, b_2, b_3)$ as the vector of covariates corresponding to the *baseline*. Let f be the prediction function for a trained machine learning model. We observe the model predictions for each of the baseline and target as $f(b_1, b_2, b_3) = a$ and $f(t_1, t_2, t_3) = b$ and we seek to explain the difference, $b - a$, in terms of each of the three features. If we consider the three features as "players," then the set Ω is equal to $\{1, 2, 3\}$ and we can construct a mapping from each of the "coalitions" to valid vectors in $\mathbb{R}^3$ as follows.

S	z	$x_{synthetic}$	S	z	$x_{synthetic}$
$\varnothing$	$(0,0,0)$	(b_1, b_2, b_3)	$\{1,2\}$	$(1,1,0)$	(t_1, t_2, b_3)
$\{1\}$	$(1,0,0)$	(t_1, b_2, b_3)	$\{1,3\}$	$(1,0,1)$	(t_1, b_2, t_3)
$\{2\}$	$(0,1,0)$	(b_1, t_2, b_3)	$\{2,3\}$	$(0,1,1)$	(b_1, t_2, t_3)
$\{3\}$	$(0,0,1)$	(b_1, b_2, t_3)	$\{1,2,3\}$	$(1,1,1)$	(t_1, t_2, t_3)

Now we can define the Shapley characteristic function as $\nu(S) = f(x_{synthetic}) = f\left(h(g(S), x_{baseline}, x_{target})\right)$. Figure 5.1 visualizes the synthetic samples created in service of SHAP estimation.

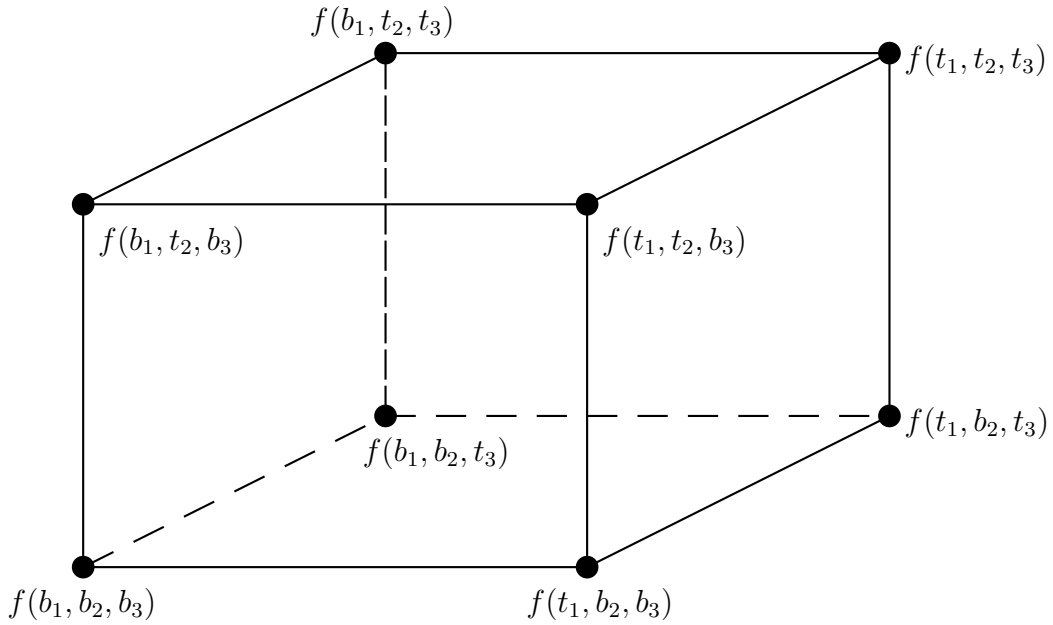

Figure 5.1: Hypercube view of SHAP model evaluation with one baseline value and one target value

We can see that the Shapley values for each of the three features are

$$
\begin{aligned}
\phi_1(\nu) &= \sum_{S \in \Omega \setminus \{i\}} \frac{(|S|)!\,(3-|S|-1)!}{3!} \left[\nu(S \cup \{i\}) - \nu(S)\right] \\
&= \frac{0!\,(3-1)!}{3!} \left[\nu(\varnothing \cup \{1\}) - \nu(\varnothing)\right] + \frac{1!\,(3-2)!}{3!} \left[\nu(\{2\} \cup \{1\}) - \nu(\{2\})\right] + \\
&\quad \frac{1!\,(3-2)!}{3!} \left[\nu(\{3\} \cup \{1\}) - \nu(\{3\})\right] + \frac{2!0!}{3!} \left[\nu(\{2,3\} \cup \{1\}) - \nu(\{2,3\})\right] \\
&= \frac{1}{3}\left[f(t_1, b_2, b_3) - f(b_1, b_2, b_3)\right] + \frac{1}{6}\left[f(t_1, t_2, b_3) - f(b_1, t_2, b_3)\right] + \\
&\quad \frac{1}{6}\left[f(t_1, b_2, t_3) - f(b_1, b_2, t_3)\right] + \frac{1}{3}\left[f(t_1, t_2, t_3) - f(b_1, t_2, t_3)\right] \\
\phi_2(\nu) &= \frac{1}{3}\left[f(b_1, t_2, b_3) - f(b_1, b_2, b_3)\right] + \frac{1}{6}\left[f(t_1, t_2, b_3) - f(t_1, b_2, b_3)\right] + \\
&\quad \frac{1}{6}\left[f(b_1, t_2, t_3) - f(b_1, b_2, t_3)\right] + \frac{1}{3}\left[f(t_1, t_2, t_3) - f(t_1, b_2, t_3)\right] \\
\phi_3(\nu) &= \frac{1}{3}\left[f(b_1, b_2, t_3) - f(b_1, b_2, b_3)\right] + \frac{1}{6}\left[f(t_1, b_2, t_3) - f(t_1, b_2, b_3)\right] + \\
&\quad \frac{1}{6}\left[f(b_1, t_2, t_3) - f(b_1, t_2, b_3)\right] + \frac{1}{3}\left[f(t_1, t_2, t_3) - f(t_1, t_2, b_3)\right]
\end{aligned}
$$

S	z_1	z_2	z_3	z_4	x_1	x_2	x_3	x_4	$\binom{4}{\lvert S\rvert}$	$\lvert S\rvert$
$\varnothing$	0	0	0	0	b_1	b_2	b_3	b_4	1	0
$\{1\}$	1	0	0	0	t_1	b_2	b_3	b_4	4	1
$\{2\}$	0	1	0	0	b_1	t_2	b_3	b_4	4	1
$\{3\}$	0	0	1	0	b_1	b_2	t_3	b_4	4	1
$\{4\}$	0	0	0	1	b_1	b_2	b_3	t_4	4	1
$\{1,2\}$	1	1	0	0	t_1	t_2	b_3	b_4	6	2
$\{1,3\}$	1	0	1	0	t_1	b_2	t_3	b_4	6	2
$\{1,4\}$	1	0	0	1	t_1	b_2	b_3	t_4	6	2
$\{2,3\}$	0	1	1	0	b_1	t_2	t_3	b_4	6	2
$\{2,4\}$	0	1	0	1	b_1	t_2	b_3	t_4	6	2
$\{3,4\}$	0	0	1	1	b_1	b_2	t_3	t_4	6	2
$\{1,2,3\}$	1	1	1	0	t_1	t_2	t_3	b_4	4	3
$\{1,2,4\}$	1	1	0	1	t_1	t_2	b_3	t_4	4	3
$\{1,3,4\}$	1	0	1	1	t_1	b_2	t_3	t_4	4	3
$\{2,3,4\}$	0	1	1	1	b_1	t_2	t_3	t_4	4	3
Ω	1	1	1	1	t_1	t_2	t_3	t_4	1	4

Table 5.1: Powerset of coalitions

Multiple Baseline Values

Section 5.2.2 discusses Shapley value estimation in the context of a single baseline value. In practice, SHAP users commonly evaluate their target prediction in reference to multiple baselines. We show that notation of the previous sections can be extended to cover multiple baselines quite straightforwardly by averaging the Shapley values calculated for each individual baseline.

Let Z be a $2^p \times p$ binary matrix whose rows are coalitions (1 indicates a feature's inclusion). Z is contructed for $p = 4$ in Table 5.1. We let n refer to the number of baseline vectors under evaluation. We refer to the i-th baseline as $b^{(i)}$ and the target vector as t. Let $X^{(i)} = Zt + (1 - Z)b^{(i)}$ be a $2^p \times p$ matrix of "synthetic" predictors, determined by the baseline and target vectors. Note that for any baseline, $b^{(i)}$, the Shapley values $\phi^{(i)}$ can be estimated by regressing $f(X^{(i)})$ on Z as detailed in Section 5.2.2. Now, we can express the SHAP regression problem with multiple baselines as

$$\begin{pmatrix} Z \\ \cdots \\ Z \end{pmatrix} \phi^* = \begin{pmatrix} f(X^{(1)}) \\ \cdots \\ f(X^{(n)}) \end{pmatrix}$$

$$\begin{pmatrix} I & \cdots & I \end{pmatrix} \begin{pmatrix} Z \\ \cdots \\ Z \end{pmatrix} \phi^* = \begin{pmatrix} I & \cdots & I \end{pmatrix} \begin{pmatrix} f(X^1) \\ \cdots \\ f(X^n) \end{pmatrix}$$

$$nZ\phi^* = f(X^{(1)}) + \cdots + f(X^{(n)})$$

$$nZ\phi^* = Z\phi^{(1)} + \cdots + Z\phi^{(n)}$$

$$\phi^* = \frac{\phi^{(1)} + \cdots + \phi^{(n)}}{n}$$

Thus, the solution to the SHAP regression problem with multiple baselines is simply the average of the SHAP estimates for each of the individual baselines. Since expectation is a linear operator, we see that these Shapley values can alternatively be computed as the solution to a regression of the average synthetic predictions on Z ($Z\phi^* = \frac{1}{n}\left[f(X^{(1)}) + \cdots + f(X^{(n)})\right]$). If multiple baselines are selected to approximate the sampling distribution of X, then we can use hypercube notation of Section 5.2.2 to write

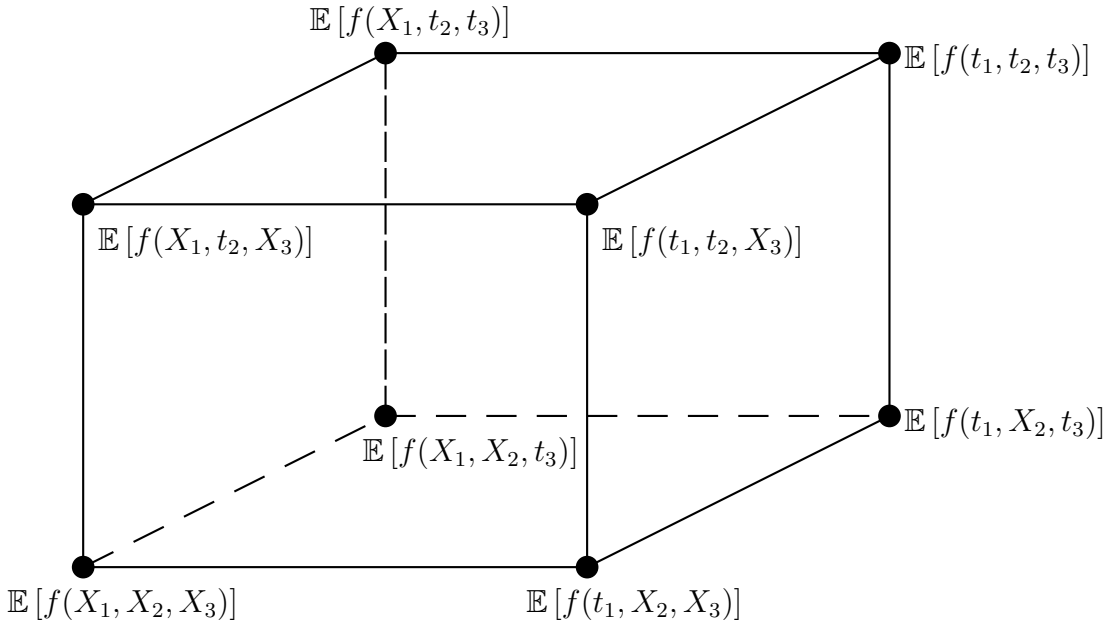

Figure 5.2: Hypercube view of SHAP model evaluation with multiple baselines. Note that all corners of this hypercube except one require computing (or estimating) a conditional expectation of a function over subsets of the random variable X.

Functional ANOVA Representation of SHAP

Functional ANOVA refers to a decomposition of function evaluations into the 2^p powerset of "effects" attributable to subsets of features. Much has been written about the functional ANOVA. We introduce the notation necessary to draw connections to SHAP and refer the interested reader to Hoeffding (1948), Stone (1994), Hooker (2004), Hooker (2007), and Liu and Owen (2006) for more detail.

Hooker (2004) define the functional ANOVA recursively in terms of a subset $u \subseteq \{1, \ldots, p\}$ of feature indices,

$$f_u = \mathbb{E}\left[f(X) - \sum_{v \subset u} f_v \mid X_u = x_u \right]$$

where X_u refers to the variables in X indexed by indices u and x_u refers to a specific realization of X_u. Since each component f_v of f_u has the same recursive contrast-

based structure as f_u, we can rewrite f_u as

$$f_u = \mathbb{E}\left[f(X) \mid X_u = x_u\right] + \sum_{s=0}^{|u|-1} \sum_{v \subset u : |v| = s} (-1)^{|u|-s} \, \mathbb{E}\left[f(X) \mid X_v = x_v\right]$$

and the function evaluation $f(x)$ can be represented as

$$f(x) = \sum_{u \subseteq \{1,\ldots,p\}} f_u$$

We now illustrate this decomposition with a specific example. Assume that all features X_i are independently distributed $\mathrm{U}\left[0,1\right]$, so that each marginal density $p(X_i) = 1$ and that $p = 3$. In this case, $\Omega = \{1,2,3\}$, $X = (X_1, X_2, X_3)$, and the power set of feature combinations is given by

$$2^{\Omega} = \{\varnothing, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$$

So the value of a function f evaluated at some realization x of the random variable

X can be decomposed as

$$
\begin{aligned}
f(x) &= f_{\varnothing} + f_1 + f_2 + f_3 + f_{12} + f_{13} + f_{23} + f_{123} \\
f_{\varnothing} &= \mathbb{E}\left[f(X_1, X_2, X_3)\right] \\
f_1 &= \mathbb{E}\left[f(X_1, X_2, X_3) \mid X_1 = x_1\right] - f_{\varnothing} = \mathbb{E}\left[f(x_1, X_2, X_3)\right] - \mathbb{E}\left[f(X_1, X_2, X_3)\right] \\
f_2 &= \mathbb{E}\left[f(X_1, X_2, X_3) \mid X_2 = x_2\right] - f_{\varnothing} = \mathbb{E}\left[f(X_1, x_2, X_3)\right] - \mathbb{E}\left[f(X_1, X_2, X_3)\right] \\
f_3 &= \mathbb{E}\left[f(X_1, X_2, X_3) \mid X_3 = x_3\right] - f_{\varnothing} = \mathbb{E}\left[f(X_1, X_2, x_3)\right] - \mathbb{E}\left[f(X_1, X_2, X_3)\right] \\
f_{12} &= \mathbb{E}\left[f(X_1, X_2, X_3) \mid X_1 = x_1, X_2 = x_2\right] - f_1 - f_2 - f_{\varnothing} \\
&= \mathbb{E}\left[f(x_1, x_2, X_3)\right] - \mathbb{E}\left[f(x_1, X_2, X_3)\right] - \mathbb{E}\left[f(X_1, x_2, X_3)\right] + \mathbb{E}\left[f(X_1, X_2, X_3)\right] \\
f_{13} &= \mathbb{E}\left[f(X_1, X_2, X_3) \mid X_1 = x_1, X_3 = x_3\right] - f_1 - f_3 - f_{\varnothing} \\
&= \mathbb{E}\left[f(x_1, x_2, X_3)\right] - \mathbb{E}\left[f(x_1, X_2, X_3)\right] - \mathbb{E}\left[f(X_1, X_2, x_3)\right] + \mathbb{E}\left[f(X_1, X_2, X_3)\right] \\
f_{23} &= \mathbb{E}\left[f(X_1, X_2, X_3) \mid X_2 = x_2, X_3 = x_3\right] - f_2 - f_3 - f_{\varnothing} \\
&= \mathbb{E}\left[f(X_1, x_2, x_3)\right] - \mathbb{E}\left[f(X_1, x_2, X_3)\right] - \mathbb{E}\left[f(X_1, X_2, x_3)\right] + \mathbb{E}\left[f(X_1, X_2, X_3)\right] \\
f_{123} &= \mathbb{E}\left[f(X_1, X_2, X_3) \mid X_1 = x_1, X_2 = x_2, X_3 = x_3\right] \\
&\quad - f_{12} - f_{13} - f_{23} - f_1 - f_2 - f_3 - f_{\varnothing} \\
&= f(x_1, x_2, x_3) - \mathbb{E}\left[f(x_1, x_2, X_3)\right] - \mathbb{E}\left[f(x_1, X_2, x_3)\right] - \mathbb{E}\left[f(X_1, x_2, x_3)\right] \\
&\quad + \mathbb{E}\left[f(x_1, X_2, X_3)\right] + \mathbb{E}\left[f(X_1, x_2, X_3)\right] + \mathbb{E}\left[f(X_1, X_2, x_3)\right] - \mathbb{E}\left[f(X_1, X_2, X_3)\right]
\end{aligned}
$$

Some arithmetic shows that in this case

$$\begin{aligned}
\phi_1(f) &= \frac{1}{3}\left[\mathbb{E}\left[f(x_1, X_2, X_3)\right] - \mathbb{E}\left[f(X_1, X_2, X_3)\right]\right] \\
&\quad + \frac{1}{6}\left[\mathbb{E}\left[f(x_1, x_2, X_3)\right] - \mathbb{E}\left[f(X_1, x_2, X_3)\right]\right] + \\
&\quad \frac{1}{6}\left[\mathbb{E}\left[f(x_1, X_2, x_3)\right] - \mathbb{E}\left[f(X_1, X_2, x_3)\right]\right] \\
&\quad + \frac{1}{3}\left[f\left(x_1, x_2, x_3\right) - \mathbb{E}\left[f(X_1, x_2, x_3)\right]\right] \\
&= \frac{1}{3}\left(f_1\right) + \frac{1}{6}\left(f_{12} + f_1\right) + \frac{1}{6}\left(f_{13} + f_1\right) + \frac{1}{3}\left(f_{123} + f_{12} + f_{13} + f_1\right) \\
&= f_1 + \frac{1}{2}\left(f_{12} + f_{13}\right) + \frac{1}{3}\left(f_{123}\right) \\
\phi_2(f) &= f_2 + \frac{1}{2}\left(f_{12} + f_{23}\right) + \frac{1}{3}\left(f_{123}\right) \\
\phi_3(f) &= f_3 + \frac{1}{2}\left(f_{13} + f_{23}\right) + \frac{1}{3}\left(f_{123}\right)
\end{aligned}$$

This allows a straightforward interpretation of SHAP as an equal division of functional ANOVA terms for a given "target" value x. More broadly, with p features we can write the SHAP estimate for feature i and function f as

$$\phi_i(f) = \sum_{j=1}^{p} \frac{1}{j} \sum_{S \subseteq 2^{\Omega}: i \in S, |S| = j} f_S$$

Note that this equivalence is not new to the sensitivity analysis literature. Owen (2014) decomposes the global Shapley value as above using the variances of the functional ANOVA terms. For examples of references in the context of individual Shapley values, see Keevers (2020), Hiabu *et al.* (2022), and Bordt and von Luxburg (2022). We do not present the equivalence here as a novel finding, but rather to motivate empirical statistical issues that arise in the estimation of Shapley values

Connection to Design of Experiments

A central challenge in computing Shapley values in the p-dimensional model interpretation setting is that the power set expansion of 2^p terms is computationally prohibitive for large p. This problem is addressed in the model interpretation literature

by choosing a small subset of the full power set for evaluation. Before discussing the details of this approximation in SHAP (Section 5.2.2), we introduce the closely related problem of *fractional factorial* design.

Dean *et al.* (2017) define a *contrast* vector as an m-vector of coefficients c such that $\sum_{j=1}^{m} c_j = 0$ which define an estimator cy as a linear combination of y values. In multi-baseline SHAP, the y values correspond to 2^p corners ($\mathbb{E}\left[f(X) \mid X_S = x_S\right]$) of the hypercube in Figure 5.2 and the contrast coefficients c corresponds to SHAP weights determined by the formula $\frac{|S|!(p-|S|-1)!}{p!}$ if $i \notin S$ and $\frac{(|S|-1)!(p-|S|-2)!}{p!}$ otherwise.

Thus, SHAP estimates for feature i correspond to a contrast in $\mathbb{E}\left[f(X) \mid X_S = x_S\right]$ that estimate $\phi_i(f) = \sum_{j=1}^{p} \frac{1}{j} \sum_{S \subseteq 2^{\Omega}: i \in S, |S|=j} f_S$. Traditional experimental design literature introduces the notion of a *fractional factorial* design, which economizes the number of experiment runs at the expense of estimating some higher order interactions. In the SHAP framing, we can see the utility of this approach in the following example. Suppose $f_S = 0$ for all S with $|S| > 1$, $\phi_i(f)$ can be evaluated for each i as $\mathbb{E}\left[f(X) \mid X_i = x_i\right] - \mathbb{E}\left[f(X)\right]$, requiring $2p$ conditional expectation calculations rather than 2^p.

Of course, exact knowledge of the nonzero interaction terms is rare. In traditional experiments, fractional factorial designs are often created with careful integration of domain knowledge and statistical expertise so that interactions are omitted if prior scientific knowledge suggests factors are not related. In the SHAP use case, domain knowledge can also play a role, though it may be difficult in high dimensional problems to identify specific interactions that can be excluded. Instead, users may choose sampling plans according to specific hypotheses. One example is the hypothesis of *factor sparsity* (Box and Meyer (1986)), which posits that only a small subset of features and their higher-order interactions are active. Another example is the hypothesis that higher-order interactions are rare.

We note briefly that in traditional design of experiments (DoE) problems, the number of factors being studied, p, might be relatively manageable, but each sample might be time-consuming or costly to collect. In SHAP experiments, each sample collection is simply a call to a machine learning prediction API, which is typically efficient on modern computers. The problem with a full factorial design in SHAP is typically that 2^p is an impossibly large number on high-dimensional models. Superficially, both approaches involve collecting $n < 2^p$ samples. However, the reasons for doing so are different enough that many approaches which are used in DoE (Gaussian process surrogates, Gramacy (2020), for example) are not always tractable or applicable in explainability. In the following section, we show the standard sampling procedure employed by the `shap` library in approximating Shapley values.

SHAP Sampling

We see that the exact Shapley value formula includes 2^{p-1} differences in model conditional expectations, for a total of 2^p conditional expectations of f. This makes Shapley value estimation intractable for large p. To overcome this, Lundberg and Lee (2017)'s method approximates these values by deliberately sampling coalitions in descending order of $|p/2 - |S||$ and approximating the values via weighted linear regression. Covert and Lee (2021) discuss some convergence properties of this approximation method and introduce an alternative. We will have more to say on SHAP approximations when p is large, but first, we introduce the weighted least squares SHAP approximation of Lundberg and Lee (2017).

First, note using the formula above that the sum of Shapley values for a target x across every feature is $\sum_i \phi_i = f(x) - f_\varnothing$. Thus, any attempt to estimate ϕ_i via linear regression introduces the condition that $\sum_i \phi_i = f(x) - \mathbb{E}\left[f(X)\right]$, so that these two estimates cannot be omitted during sampling-based estimation. To illustrate the

sampling scheme, we now consider a model with $p = 4$ features, given in Table 5.1.

We group coalitions in Table 5.1 by size (for example $S_1 = \{1\}$ and $S_2 = \{2\}$ are different coalitions but $|S_1| = 1 = |S_2|$). Observe that every coalition with $|S| = 1$ has an inverse coalition with with $|S| = 3$ ($\{2, 3, 4\}$ is the inverse of $\{1\}$, etc...). For $|S| = 2$, on the other hand, there is no inverse coalition. In general, with p features, there are $\lfloor (p-1)/2 \rfloor$ matching blocks. If p is odd, then every block of coalitions with $|S| = a$ for some a will have an inverse block, and if p is even, then there will be 1 "center block."

SHAP attempts to enumerate the entire power set, starting with the outermost blocks (1 and 3 in Table 5.1). The sampling process is iterative and at each step, SHAP determines whether to enumerate an entire block. Suppose a user has specified that they would like to draw $m < 2^p$ samples. We defer discussion of the regression weights to the next section, but for now we note that the regression weights imply a frequency distribution of samples from each block that is proportional to the block's size. During sampling, SHAP uses this implied distribution as a stage gate. SHAP iterates from $i = 1$ to $\lceil (p-1)/2 \rceil$ and at each i, SHAP looks at the implied frequency of a block, the number of samples that can be allocated, and determines whether to allocate all the samples from block i (and $p - i$ if $i \neq p/2$). Let $k \leq m$ be the number of remaining samples, j be the size of block i, and w be the target share of samples from block i. SHAP will enumerate the entire block if $w \geq \frac{j}{k}$. Once $w < \frac{j}{k}$, SHAP samples from the remaining blocks uniformly with replacement.

SHAP Regression

Once a subset of the SHAP coalitions has been sampled, the Shapley values are estimated using weighted linear regression. The equivalence between weighted least squares estimation and Shapley values is established in the supplement to Lundberg

and Lee (2017) [2], but we derive it below using a slightly different approach which will make clear the relation to experimental design. At a high level, the goal is to fit a regression model to the synthetic data whose coefficients are the exact Shapley values when the full powerset of coalitions is observed and will approximate the Shapley values when the data are a subsample of coalitions.

Let Z be a binary matrix whose rows are coalitions (1 indicates a feature's inclusion). Assume for now that Z is a $(2^p - 2) \times p$ matrix, so that all of the coalitions with $0 < |S| < p$ have been sampled. The weighting function Lundberg and Lee (2017) use for the SHAP regression is

$$w_i(Z) = \frac{p-1}{\binom{p}{s} s\,(p-s)} = \frac{(p-1)\,(p-s-1)!\,(s-1)!}{p!}$$
$$s = \sum_{j=1}^{p} Z_{ij}$$

This function is undefined when $Z_{i\cdot}$ is the one or zero vector (corresponding to $f(x)$ and $\mathbb{E}\left[f(X)\right]$, respectively), which is why the regression matrix Z only has $2^p - 2$ rows. However, we still incorporate these two vectors in estimation via the side condition expressed above.

Let $y_t = f(x)$ and $y_b = \mathbb{E}\left[f(X)\right]$ and we express the regression model below. First, we define the vector of conditional expectations corresponding to hypercube corners in Figure 5.2. For each row Z_i in the Z matrix, let S_i refer to the columns j for which $Z_{ij} = 1$. Now define $y_i = \mathbb{E}\left[f(X) \mid X_{S_i} = x_{S_i}\right]$. Letting j be a p-dimensional vector of ones, we can approximate y via a linear model with side conditions.

$$y = Z\beta + \varepsilon$$
$$j'\beta = y_t - y_b$$

[2]The supplemental files can be accessed at `https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html`

We also define a matrix of weights as

$$W = \begin{pmatrix} w_1 & 0 & 0 & ... & 0 \\ 0 & w_2 & 0 & ... & 0 \\ 0 & 0 & w_3 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & w_{(2^p-2)} \end{pmatrix}$$

We estimate the Shapley values by fitting a constrained weighted least squares regression using the synthetic Z matrix, the predictions y, the weight matrix W, $f(x)$ and $f_\varnothing$. The derivation of the regression solution and its equivalence to Shapley values is given in detail in Appendix C. Defining I as the p-dimension identity matrix and J as a $p \times p$ matrix of all ones, we note that the regression solution is given by

$$\hat{\beta} = \left(\frac{p}{p-1} I - \frac{1}{p-1} J \right) Z' W y + \frac{j \left(y_t - y_b \right)}{p}$$

which corresponds to a weighted contrast estimate of the 2^p predictions. It is worth noting that the derivations in Appendix C simply demonstrate that the SHAP regression estimator returns Shapley values when the entire power set of coalitions is available. Lundberg and Lee (2017) show through a simulation study that the SHAP regression estimates can converge to the exact Shapley values with $m << 2^p$ samples. Covert and Lee (2021) study the convergence to exact Shapley values analytically.

5.3 Estimation Decisions

It was shown above that SHAP is a partition of contrasts of functional ANOVA terms. Since there are 2^p functional ANOVA terms, when p is large, users cannot typically compute 2^p conditional expectations. Thus, users must select some subset of the full power set of feature interactions. This is the first significant choice a user

must make in estimating Shapley values for their model, although most users make this decision implicitly by using the `shap` library which samples as in Section 5.2.2.

In addition to the sampling challenges outlined above, there is also the question of how to approximate the requisite conditional expectations. In real world scenarios, it would be highly unusual to have access to an analytical formula for the joint distribution, $p(X)$, of the features. Thus, even if a model provides parameters that can be converted into an analytical formula $y = f(X)$, analytically computing the expectation of $f(X) \mid X_u$ for any subset u of feature interactions is impossible without further assumptions. $p(X)$ must be estimated from the data or assumed directly by the user, introducing the second significant user decision.

Consider approximating an expected value with a sample X of size n and evaluating $S \subseteq 2^{\Omega}$ of the functional ANOVA effects. These two decisions thus correspond to the choice of S and then choice of X.

5.3.1 Selecting Interaction Terms to Evaluate

The default sampling scheme for conditional expectations implemented in the `shap` library is the paired sampling approach described in Section 5.2.2.

The Current Implementation of Sampling in the `shap` Library

The `shap` sampling procedure first draws subsets s with $|s| = 1$ and $|s| = p - 1$ and proceeds "inwards" in decreasing order of $|p/2 - |s||$. We show below that this procedure is generally effective if lower-order interactions dominate, because the first and second order interactions are properly aliased (according to the share of interactions described in Section 5.2.2) after only $2p$ samples are taken.

We demonstrate this empirically for one example. Consider the case of $p = 6$ in

which we sample a 12×6 matrix of subsets s with $|s| = 1$ and $|s| = 5$

$$X_r = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

The matrix of second order interaction terms can be split into 5 sub-matrices indexed

as X_{ji} with j as the leading interaction variable.

$$X_{1i} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad X_{2i} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

$$X_{3i} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \quad X_{4i} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \quad X_{5i} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

The weight terms for the regression matrix X_r are all $\frac{p-1}{\binom{p}{|s|}|s|(p-|s|)} = 1/6$. One way to impose the side condition that $\sum_i \phi_i = f(x_t) - f(x_b)$, is to subtract one of the X_r columns from the others. Performing this operation defines a new $2p \times p - 1$ main effect matrix which we denote X_r^* and a new set of interaction matrices X_{ji}^* with the p-th column of X_r subtracted. Observe that

$$
\begin{aligned}
X_r^{*'} W X_r^* &= (1/3)I + (1/3)J \\
\left(X_r^{*'} W X_r^*\right)^{-1} &= 3I - (1/2)J \\
X_r^{*'} W X_{1i}^* &= \begin{pmatrix} 1/2 & 1/2 & 1/2 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/3 & 1/3 & 1/6 \\ 1/3 & 1/2 & 1/3 & 1/3 & 1/6 \\ 1/3 & 1/3 & 1/2 & 1/3 & 1/6 \\ 1/3 & 1/3 & 1/3 & 1/2 & 1/6 \end{pmatrix} \\
X_r^{*'} W X_{2i}^* &= \begin{pmatrix} 1/3 & 1/3 & 1/3 & 1/6 \\ 1/2 & 1/2 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/3 & 1/6 \\ 1/3 & 1/2 & 1/3 & 1/6 \\ 1/3 & 1/3 & 1/2 & 1/6 \end{pmatrix}
\end{aligned}
$$

And thus, for these first two sets of interactions, the alias matrix is given by

$$\left(X_r^{*\prime} W X_r^*\right)^{-1} X_r^{*\prime} W X_{1i}^* = \begin{pmatrix} 1/2 & 1/2 & 1/2 & 1/2 & 1/2 \\ 1/2 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 \end{pmatrix}$$

$$\left(X_r^{*\prime} W X_r^*\right)^{-1} X_r^{*\prime} W X_{2i}^* = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 1/2 & 1/2 \\ 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \end{pmatrix}$$

which is exactly the alias pattern given by the full design matrix with $2^p - 2$ samples. Thus, with only $2p$ samples, Shapley regression estimates will be "exact" in the sense of properly allocating interactions if the model consists of first and second order interactions. In this sense, the `shap` sampling strategy can be an effective way to sample from the 2^p possible rows of the design matrix.

An Alternative Sampling Approach: Numeric Interaction Testing

While the `shap` sampling approach appears to work well upon empirical investigation (and is explored in theoretical terms in Covert and Lee (2021)), the question may remain: to what extent is the model dominated by low-order interaction terms? From the derivations in Section 5.2, we can see that if $f_S = 0$ for all S with $|S| > 1$, then $\phi_i = f_i$ and $\sum_i f_i = f(x) - f_\varnothing$. Of course, since $f_S \in \mathbb{R}$, the converse is not true. Hooker (2004) defines $\sigma_S^2(f_S) = \mathbb{E}\left[f_S^2\right]$ where the expectation is taken with respect to features X_S which were not conditioned in f_S. We can see that if $\sigma_S^2(f_S) = 0$ and

$\mathbb{E}[f_S] = 0$ (as is commonly established in the functional ANOVA decomposition), then $f_S = 0$ for all x.

One approach to numerically screening interactions terms introduced in Hooker (2004) is to search for the smallest set of functional ANOVA terms $\mathcal{S}$ such that $\sum_{s \in \mathcal{S}} \sigma_s^2(f_s)$ accounts for a pre-specified share of the variance of $f(X)$. Hooker (2004) introduces two different algorithms representing different inductive biases: a "breadth-first" algorithm which assumes higher order interaction effects are more likely to be null and a "depth-first" algorithm which assumes factor sparsity. We present a modified version of the breadth-first algorithm below and show how it can be used in conjunction with SHAP and a suitable interaction scoring criteria $\psi(S)$. Note that the substance of the algorithm below is as introduced in Hooker (2004), but the presentation and notation is updated here to reflect the intended application and previously-introduced notation. Let $\varepsilon \in [0, 1]$ be specified by the user.

Algorithm 1: Breadth-first search for low-order functional ANOVA terms

Result: List of functional ANOVA terms to include in SHAP approximation

$V_t = \mathrm{Var}(f(X))$;

$V_s = 0$;

$U = \{\}$;

for j = *1 to* p **do**

  $\mathcal{S} = \{S \subseteq 2^{\Omega} : |S| = j\}$;

  $\Psi = \{\psi(s) \text{ for } s \text{ in } \mathcal{S}\}$;

  $I = \mathrm{argsort}(\Psi)$;

  for k = *1 to* $\binom{p}{j}$ **do**

    $s = \mathcal{S}_{I_k}$;

    $U = U \cup s$;

    $V_s = V_s + \hat{\sigma}_s(f_s)$;

    if $V_s > V_t(1 - \varepsilon)$ **then**

      break;

    else

      continue;

    end

  end

end

Since the procedure above involves the global $\hat{\sigma}_s(f_s)$, it can be used to identify a sufficient set of features for computing Shapley values on any target value for a given model. In particular, it provides a approximate empirical measure of "convergence" since $V_s/V_t > 1 - \varepsilon$. This is of course only helpful and desirable if the resulting U is a considerably reduced subset of 2^{Ω}. However, if the algorithm fails to stop at a small subset of 2^{Ω}, that can also provide the user with valuable information about the

quality of their Shapley value estimates with less than 2^p conditional expectations. Hooker (2004) also notes that when p is large, this algorithm can be configured to stop at a given interaction order, thus providing a tractable way of signaling that higher order interactions may be present and the user should take care in interpreting Shapley values. Variants of this algorithm are appealing in that they enable a numeric search that can be tailored to the user's inductive bias. For example, Hooker (2004) provides an depth-first algorithm which corresponds to the factor sparsity hypothesis. With the computational infrastructure in place to compute $G_s(x)$ for subsets s of 2^Ω, researchers can modify the search algorithm to prioritize a hypothesized feature structure.

If the algorithm has been run successfully and a subset U of 2^Ω has been identified, we can reduce the computational burden of SHAP estimation in either of two ways. First, noting that $\phi_i(f) = \sum_{i=1}^{p} \frac{1}{i} \sum_{S \subseteq \Omega : |S|=i} f_S$, we can simply compute f_S for every $S \in U$ and combine their estimates according to the formula for $\phi_i(f)$. Second, we can construct a regression matrix Z whose rows correspond only to the coalitions $S \in U$ and estimate ϕ_i via weighted least squares.

Now, we discuss one approach to estimating an interaction importance measure $\psi(S)$. For $p = 3$, we can visualize the SHAP terms as a power set, organized in rows according to $|s|; s \subseteq 2^\Omega$ as in Figure 5.3.

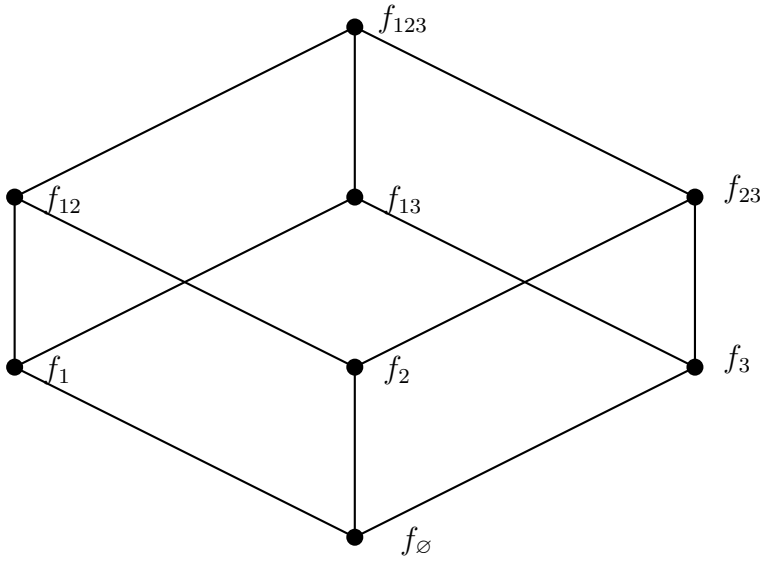

Figure 5.3: Hasse diagram of connections between functional ANOVA terms. Excluding an interaction or main effect from the evaluation set involves pruning the estimable functional ANOVA terms in the above diagram. When a small subset of this lattice explains most of the variability of $f(X)$, Shapley values may be approximated well with less computational burden.

Now, consider the effect of removing $s = \{1\}$ from the lattice above, which leaves a new structure defined in Figure 5.4 below. Thus, the effect of removing $s = \{1\}$ is that $f(x)$ must be approximated as $f_\varnothing + f_2 + f_3 + f_{23} = \mathbb{E}\left[f(X_1, x_2, x_3)\right]$ If $f(x) = f_\varnothing + f_2 + f_3 + f_{23}$, then the approximation is perfect so that $\mathbb{E}\left[\left(f(x) - \mathbb{E}\left[f(X_1, x_2, x_3)\right]\right)^2\right] = 0$. Using the notation and verbiage of Hooker (2004), we let $G_s(x)$ denote the approximation of $f(x)$ with s and all of its supersets removed from the functional ANOVA lattice and we refer to $\mathbb{E}\left[\left(f(x) - G_s(x)\right)^2\right]$ as the $\mathcal{L}_2$ cost of exclusion (L2COE) of set s. In this example, $G_s(x) = \mathbb{E}\left[f(X_1, x_2, x_3)\right]$ and $\text{L2COE}(s) = \mathbb{E}\left[\left(f(x) - \mathbb{E}\left[f(X_1, x_2, x_3)\right]\right)^2\right]$. We can use the L2COE as an interaction importance criteria in Algorithm 1, replacing $\psi(s) = \text{L2COE}(s)$. Liu and Owen (2006) present a "pick-freeze" algorithm for estimating L2COE. First, select a sample Z of size $n \times p$. Then, select a sample X of size $n \times |s|$ where $|s|$ is the size of the interaction term s.

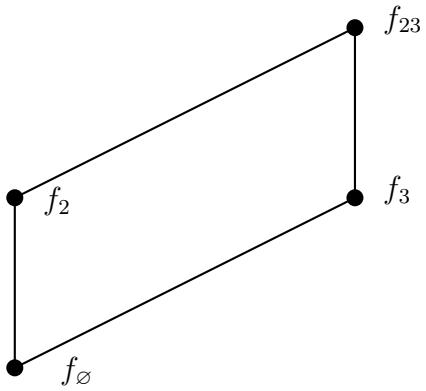

Figure 5.4: Pruned Hasse diagram of functional ANOVA terms with feature 1 removed. Approximating the Shapley value with these terms requires $4 < 2^3 = 8$ conditional expectations.

This gives an estimate

$$\text{L2COE} = \frac{1}{n2^{|s|}} \sum_{i=1}^{n} \left(\sum_{r \subseteq s} f(x_i^r, z_i^{-r}) \right)^2$$

where x_i^r refers to the i-th sample of x for the features indexed by interaction s and z_i^{-r} refers to the i-th sample of z for the features **not** indexed by interaction s.

To illustrate this algorithm, we implement a simple example where the active subsets in this case are $\{1\}$, $\{2\}$, $\{3\}$ and $\{2,3\}$.

$$X = (X_1, X_2, X_3)$$

$$y = f(X) = X_1 + X_2 + X_3 + X_2 X_3$$

To test this algorithm, we conduct 100 simulations with $n = 500$ using a custom R implementation of the algorithm[3]. We approximate $V_t = \sum_{s \subseteq 2^\Omega} \hat{\sigma}_s(f_s)$ where $\hat{\sigma}_s(f_s) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}_s^2(x_i)$ and the expected values in $\hat{f}_s(x_i)$ are also calculated as empirical expectations with respect to the matrix $\mathbf{x} = (x_1, \cdots, x_n)'$. X_2 and X_3 have the same L2COE, and we observe that the correct ranking of either $\{\{2\}, \{3\}, \{1\}, \{2,3\}\}$ or $\{\{3\}, \{2\}, \{1\}, \{2,3\}\}$ in 100% of simulations.

[3]The code is available at `https://github.com/andrewherren/shap-anova-examples`

Finally, with a list $\mathcal{S}$ of low-order interaction terms returned by the above algorithm, Shapley values can be estimated by computing the functional ANOVA terms for each of the terms in $\mathcal{S}$ and combining them according to

$$\phi_i(f) = \sum_{j=1}^{p} \frac{1}{j} \sum_{S \subseteq 2^{\Omega}: i \in S, |S|=j} f_S.$$

Measures of Effective Dimensionality

The above search procedure is helpful in identifying structure of the model, but it could be costly when the inductive bias is incorrect. Saltelli *et al.* (2010) outline a sensitivity analysis technique that can be used to determine both factor sparsity as well as interaction order. Kucherenko *et al.* (2009) introduce two definitions of "effective dimensionality" which correspond roughly to the principles of effect sparsity and dominance by low-order interactions. The first measure, *truncation* dimensionality, is defined as the value of d_T such that

$$\sum_{u \subseteq \{1,\ldots,d_T\}} \sigma_u^2 \geq (1-\varepsilon)\,\sigma^2$$

which corresponds broadly to the minimum number of factors required to explain $(1-\varepsilon)\%$ of the total variance. The second measure, *superposition* dimensionality, is defined as the value of d_S such that

$$\sum_{0<|u|<d_S} \sigma_u^2 \geq (1-\varepsilon)\,\sigma^2$$

which corresponds to the minimum interaction size required to explain $(1-\varepsilon)\%$ of the total variance.

Saltelli *et al.* (2010) show that two common variance-based sensitivity analysis measures can be used to assess d_T and d_S. Common estimands for the first order and

total variance-based sensitivities for feature i are given by

$$S_i = \frac{\mathrm{Var}_{X_i}\left(\mathbb{E}_{X_{-i}}\left(f(X) \mid X_i\right)\right)}{\mathrm{Var}\left(f(X)\right)}$$
$$S_{T_i} = \frac{\mathbb{E}_{X_{-i}}\left(\mathrm{Var}_{X_i}\left(f(X) \mid X_{-i}\right)\right)}{\mathrm{Var}\left(f(X)\right)} = 1 - \frac{\mathrm{Var}_{X_{-i}}\left(\mathbb{E}_{X_i}\left(f(X) \mid X_{-i}\right)\right)}{\mathrm{Var}\left(f(X)\right)}$$

There are a number of approaches to computing these indices using random or low-discrepancy samples for X_i and X_{-i} (see, for example, Sobol (2001) and Saltelli (2002)). The R package `sensitivity` (Iooss *et al.* (2021)) provides a convenient interface to compute both S_i and S_{T_i}. Once these indices are computed, the importance of interactions can be determined by the magnitude of S_i/S_{T_i} for each i and the relative sparsity (or varying feature importance) can be determined by the distribution of S_i or S_{T_i}. In addition to providing a computationally efficient gauge of sparsity and interaction importance, in some cases these indices could be used on their own to replace the numeric search algorithm in Section 5.3.1. For example, if $\sum_{i\in\{1,2,5\}} S_i \geq 1 - \varepsilon$, then the SHAP estimates can be computed with only 7 conditional expectations.

5.3.2 The Impact of Choosing a Baseline Distribution

The SHAP estimand can be loosely characterized as "the average effect of setting feature i equal to its target value," which raises the question: average with respect to which distribution? In Section 5.2, we referred to "individual" and "multiple" baselines as two different approaches to SHAP. However, we can unify these two approaches using the functional ANOVA notation. In the functional ANOVA, each of the f_u terms are a contrast of conditional expectations, which may be taken with respect to any distribution $p(X)$. In the "multiple baseline" scenario commonly employed by SHAP users, the implied distribution $p(X)$ is the empirical distribution of each feature, treated independently from other features. Similarly, the "single

baseline" scenario constitutes a degenerate distribution for which $p(X) = \mathbf{1}\{X = x_{baseline}\}$.

We examine the impact of the choice of baseline distribution by comparing Shapley values with $p = 3$ and $x_{target} = (1, 1, 1)$ with three different baseline distributions:

(a) **Global independent**: $p(X) \sim \mathcal{N}((0, 0, 0), I)$

(b) **Global correlated**: $p(X) \sim \mathcal{N}\left((0, 0, 0), \begin{pmatrix} 1 & 0.9 & 0.5 \\ 0.9 & 1 & 0.75 \\ 0.5 & 0.75 & 1 \end{pmatrix}\right)$

(c) **Local independent**: $p(X) \sim \mathcal{N}(x_{target}, 0.25^2 I)$

(d) **Single baseline**: $p(X_i) = \mathbf{1}\{X_i = x_{baseline,i}\}$

We compute Shapley values for x_{target} using the above three distributions on four functions presented in the table below.

The results (Table 5.2) show that while each of these baselines could be viewed as reasonable or plausible in different circumstances, they are far from interchangeable. With strong correlation in the data, using the correlated conditional distribution to compute Shapley values has a profound effect on the estimated Shapley values relative to treating the features as independent. Similarly, a local baseline distribution centered around the target only estimates nonzero Shapley values for functions with strong evenness or nonlinearity. The single baseline estimates are perhaps the most intuitive to grasp, but they raise a crucial question of where to place the one representative baseline. Google's AI Explanations Whitepaper (Google (2020)) suggests the single baseline only when there is an "informative reference" value for comparison.

$f(X)$	Baseline	X_1	X_2	X_3
$-2X_1 + 1.5X_2 + 0.5X_3$	A	-2.00	1.50	0.50
	B	-0.39	-0.03	0.41
	C	0	0	0
	D	-2.00	1.50	0.50
$-2X_1 + 1.5X_2 + 0.5X_3 - 2X_2X_3$	A	-2.00	0.50	-0.50
	B	-0.47	-0.01	-0.02
	C	0	0	0
	D	-2.00	0.50	-0.50
$-2\sin(X_1) + 1.5\lvert X_2\rvert + 0.125X_3^2$	A	-1.68	0.31	-0.00
	B	-0.78	-0.48	-0.13
	C	-0.05	0.00	-0.01
	D	-1.68	1.50	0.12
$-2\sin(X_1) + 1.5\lvert X_2\rvert + 0.125X_3^2 + \cos(X_2X_3)$	A	-1.69	0.22	-0.09
	B	-0.80	-0.45	-0.22
	C	-0.05	0.02	0.01
	D	-1.68	1.27	-0.10

Table 5.2: Shapley values with different baseline distributions

5.4 Discussion

This work has explored the relationship between SHAP and the functional ANOVA. A result of this connection is that decades of literature on computer experiments and function approximation may be brought to bear on questions of model interpretability. However, this connection also dashes any hope of settling the question of "which Shapley values to use" due to the influence of the baseline distribution which is embedded in the estimand itself!

While SHAP is typically regarded as a tool for making modeling decisions "interpretable," this interpretability is not free. As noted above, the estimand can be loosely messaged to stakeholders as "the average effect of setting a feature equal to its target value," but this conceals decisions about the underlying distribution used in computing those averages.

The connection between sensitivity analysis and model explainability is fascinating. We believe that computational methods used in sensitivity analysis (for example, quasi-monte carlo or low-discrepancy sampling) might be profitably applied to improve convergence of existing explainability methods.

An interesting future line of research would be to study the goals of model explainability and determine for which purposes SHAP is best suited. Modern model explainability tools are increasingly being used for legal or compliance purposes, such as algorithmic recourse (Karimi *et al.* (2020)). Based on this review of SHAP, it is not immediately obvious that a series of functional ANOVA terms can provide the necessary information for algorithmic recourse. At the very least, this topic deserves further study, contrasting SHAP with methods such as counterfactual explanations (Wachter *et al.* (2017)).

Chapter 6

CONCLUSION

This dissertation has been a fascinating investigation into nonparametric causal inference and machine learning explainability. We have seen that each of three common frameworks for estimating causal effects offer complementary abstractions, each of which is useful for framing the questions we pose about feature selection for causal inference.

We showed that it is possible to express a minimal adjustment set (though computing this may be impossible) and explained when a larger adjustment set is warranted for its variance reduction properties. We discussed the possibility of using separate adjustment sets for $\mu(X)$ and $\tau(X)$, either in a "two-stage" estimator or in methods like XBCF (Krantsevich *et al.* (2022)).

We presented and advocated for a simulation-based approach to developing new statistical methods. We tested several improvements that can be made straightforwardly to treatment effect estimation methods like XBCF and identified new avenues for implementation and experimentation.

Finally, we explored the statistical issues that arise in "machine learning explainability," particularly as revealed through the popular SHAP method. We make connections to the sensitivity analysis literature, arguing that explainability practitioners could make better use of algorithms and concepts from this literature. We also advocate for further research into counterfactual explainability methods for applications such as algorithmic recourse.

Aas, K., M. Jullum and A. Løland, "Explaining Individual Predictions when Features are Dependent: More Accurate Approximations to Shapley Values", arXiv preprint arXiv:1903.10464 (2019).

Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI", Information Fusion **58**, 82–115 (2020).

Athey, S. and G. W. Imbens, "Machine learning methods for estimating heterogeneous causal effects", stat **1050**, 5, 1–26 (2015).

Athey, S., J. Tibshirani and S. Wager, "Generalized random forests", Annals of Statistics **47**, 2, 1148–1178 (2019).

Athey, S. and S. Wager, "Estimating treatment effects with causal forests: An application", Observational Studies **5**, 2, 37–51 (2019).

Belloni, A., V. Chernozhukov and C. Hansen, "Inference on treatment effects after selection among high-dimensional controls", The Review of Economic Studies **81**, 2, 608–650 (2014).

Bhatt, U., A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura and P. Eckersley, "Explainable Machine Learning in Deployment", in "Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency", pp. 648–657 (2020).

Bordt, S. and U. von Luxburg, "From Shapley Values to Generalized Additive Models and back", arXiv preprint arXiv:2209.04012 (2022).

Box, G. E. and R. D. Meyer, "An Analysis for Unreplicated Fractional Factorials", Technometrics **28**, 1, 11–18 (1986).

Bradic, J., S. Wager and Y. Zhu, "Sparsity double robust inference of average treatment effects", arXiv preprint arXiv:1905.00744 (2019).

Breiman, L., "Random forests", Machine Learning **45**, 1, 5–32 (2001).

Cameron, A. C. and P. K. Trivedi, *Microeconometrics: methods and applications* (Cambridge university press, 2005).

Chen, H., J. D. Janizek, S. Lundberg and S.-I. Lee, "True to the Model or True to the Data?", arXiv preprint arXiv:2006.16234 (2020).

Chen, T. and C. Guestrin, "Xgboost: A scalable tree boosting system", in "Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining", pp. 785–794 (2016).

Chernozhukov, V., W. K. Newey and R. Singh, "Automatic debiased machine learning of causal and structural effects", Econometrica **90**, 3, 967–1027 (2022).

Chipman, H. A., E. I. George, R. E. McCulloch *et al.*, "Bart: Bayesian additive regression trees", The Annals of Applied Statistics **4**, 1, 266–298 (2010).

Cinelli, C., A. Forney and J. Pearl, "A crash course in good and bad controls", Available at SSRN **3689437** (2020).

Covert, I. and S.-I. Lee, "Improving KernelSHAP: Practical Shapley Value Estimation Using Linear Regression", in "International Conference on Artificial Intelligence and Statistics", pp. 3457–3465 (PMLR, 2021).

Curth, A. and M. van der Schaar, "Doing great at estimating cate? on the neglected assumptions in benchmark comparisons of treatment effect estimators", arXiv preprint arXiv:2107.13346 (2021).

Curth, A. and M. van der Schaar, "In search of insights, not magic bullets: Towards demystification of the model selection dilemma in heterogeneous treatment effect estimation", arXiv preprint arXiv:2302.02923 (2023).

Datta, A., S. Sen and Y. Zick, "Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems", in "2016 IEEE symposium on security and privacy (SP)", pp. 598–617 (IEEE, 2016).

De Luna, X., I. Waernbaum and T. S. Richardson, "Covariate selection for the nonparametric estimation of an average treatment effect", Biometrika **98**, 4, 861–875 (2011).

Dean, A., D. Voss and D. Draguljić, *Design and Analysis of Experiments*, vol. 2 (Springer, 2017).

Diaconis, P. and S. Zabell, "Some alternatives to bayes's rule", in "Information Pooling and Group Decision Making", edited by B. Grofman and G. Owen, pp. 25–38 (J.A.I. Press, 1986).

Efron, B., "Prediction, estimation, and attribution", International Statistical Review **88**, S28–S59 (2020).

Frye, C., D. de Mijolla, L. Cowton, M. Stanley and I. Feige, "Shapley-based Explainability on the Data Manifold", arXiv preprint arXiv:2006.01272 (2020).

Frye, C., I. Feige and C. Rowat, "Asymmetric Shapley Values: Incorporating Causal Knowledge into Model-Agnostic Explainability", arXiv preprint arXiv:1910.06358 (2019).

Google, "AI Explanations Whitepaper", `https://storage.googleapis.com/cloud-ai-whitepapers/AI%20Explainability%20Whitepaper.pdf` (2020).

Gramacy, R. B., *Surrogates: Gaussian Process Modeling, Design and Optimization for the Applied Sciences* (Chapman Hall/CRC, Boca Raton, Florida, 2020), `https://bobby.gramacy.com/surrogates/`.

Grinsztajn, L., E. Oyallon and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?", arXiv preprint arXiv:2207.08815 (2022).

Häggström, J., "Data-driven confounder selection via markov and bayesian networks", Biometrics **74**, 2, 389–398 (2018).

Hahn, P. R., C. M. Carvalho, D. Puelz and J. He, "Regularization and confounding in linear regression for treatment effect estimation", Bayesian Analysis **13**, 1, 163–182 (2018).

Hahn, P. R., J. S. Murray and C. M. Carvalho, "Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion)", Bayesian Analysis **15**, 3, 965–1056 (2020).

Hansen, B. B., "The prognostic analogue of the propensity score", Biometrika **95**, 2, 481–488 (2008).

He, J. and P. R. Hahn, "Stochastic tree ensembles for regularized nonlinear regression", Journal of the American Statistical Association pp. 1–20 (2021).

Heckman, J. J., "Identification of causal effects using instrumental variables: Comment", Journal of the American Statistical Association **91**, 434, 459–462 (1996).

Henckel, L., E. Perković and M. H. Maathuis, "Graphical criteria for efficient total effect estimation via adjustment in causal linear models", Journal of the Royal Statistical Society Series B: Statistical Methodology **84**, 2, 579–599 (2022).

Hernan, M. A. and J. M. Robins, *Causal Inference: What If* (Boca Raton: Chapman & Hall, CRC, 2022).

Herren, A. and P. R. Hahn, "Semi-supervised learning and the question of true versus estimated propensity scores", arXiv preprint arXiv:2009.06183 (2020).

Hiabu, M., J. T. Meyer and M. N. Wright, "Unifying local and global model explanations by functional decomposition of low dimensional structures", arXiv preprint arXiv:2208.06151 (2022).

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

Hoeffding, W., "A Class of Statistics with Asymptotically Normal Distribution", Annals of Mathematical Statistics **19**, 3, 293–325 (1948).

Hooker, G., "Discovering Additive Structure in Black Box Functions", in "Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining", pp. 575–580 (2004).

Hooker, G., "Generalized Functional ANOVA Diagnostics for High-dimensional Functions of Dependent Variables", Journal of Computational and Graphical Statistics **16**, 3, 709–732 (2007).

Imbens, G. W. and D. B. Rubin, *Causal inference in statistics, social, and biomedical sciences* (Cambridge University Press, 2015).

Iooss, B., S. D. Veiga, A. Janon, G. Pujol, with contributions from Baptiste Broto, K. Boumhaout, T. Delage, R. E. Amri, J. Fruth, L. Gilquin, J. Guillaume, M. I. Idrissi, L. Le Gratiet, P. Lemaitre, A. Marrel, A. Meynaoui, B. L. Nelson, F. Monari, R. Oomen, O. Rakovec, B. Ramos, O. Roustant, E. Song, J. Staum, R. Sueur, T. Touati and F. Weber, *sensitivity: Global Sensitivity Analysis of Model Outputs*, URL `https://CRAN.R-project.org/package=sensitivity`, r package version 1.25.0 (2021).

Johansson, F., U. Shalit and D. Sontag, "Learning representations for counterfactual inference", in "International conference on machine learning", pp. 3020–3029 (PMLR, 2016).

Johansson, F. D., U. Shalit, N. Kallus and D. Sontag, "Generalization bounds and representation learning for estimation of potential outcomes and causal effects", arXiv preprint arXiv:2001.07426 (2020).

Ju, C., R. Wyss, J. M. Franklin, S. Schneeweiss, J. Häggström and M. J. van der Laan, "Collaborative-controlled lasso for constructing propensity score-based estimators in high-dimensional data", Statistical methods in medical research **28**, 4, 1044–1063 (2019).

Karimi, A.-H., G. Barthe, B. Schölkopf and I. Valera, "A Survey of Algorithmic Recourse: Definitions, Formulations, Solutions, and Prospects", arXiv preprint arXiv:2010.04050 (2020).

Kaur, H., H. Nori, S. Jenkins, R. Caruana, H. Wallach and J. Wortman Vaughan, "Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning", in "Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems", pp. 1–14 (2020).

Keevers, T. L., "A Power Series Expansion of Feature Importance", Technical Report **DST-Group-TR-3743** (2020).

Kennedy, E. H., "Towards optimal doubly robust estimation of heterogeneous causal effects", arXiv preprint arXiv:2004.14497 (2022).

Knaus, M. C., M. Lechner and A. Strittmatter, "Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence", The Econometrics Journal **24**, 1, 134–161 (2021).

Koller, D. and N. Friedman, *Probabilistic graphical models: principles and techniques* (MIT press, 2009).

Krantsevich, N., J. He and P. R. Hahn, "Stochastic tree ensembles for estimating heterogeneous effects", arXiv preprint arXiv:2209.06998 (2022).

Kucherenko, S., M. Rodriguez-Fernandez, C. Pantelides and N. Shah, "Monte Carlo Evaluation of Derivative-based Global Sensitivity Measures", Reliability Engineering & System Safety **94**, 7, 1135–1148 (2009).

Kumar, I. E., S. Venkatasubramanian, C. Scheidegger and S. Friedler, "Problems with Shapley-value-based Explanations as Feature Importance Measures", arXiv preprint arXiv:2002.11097 (2020).

Künzel, S. R., J. S. Sekhon, P. J. Bickel and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning", Proceedings of the national academy of sciences **116**, 10, 4156–4165 (2019).

Lei, L. and E. J. Candès, "Conformal inference of counterfactuals and individual treatment effects", Journal of the Royal Statistical Society Series B: Statistical Methodology **83**, 5, 911–938 (2021).

Liu, R. and A. B. Owen, "Estimating Mean Dimensionality of Analysis of Variance Decompositions", Journal of the American Statistical Association **101**, 474, 712–721 (2006).

Lohr, S. L., *Sampling: design and analysis* (Chapman and Hall/CRC, 2019).

Lundberg, S. M. and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions", in "Advances in neural information processing systems", pp. 4765–4774 (2017).

Molnar, C., *Interpretable Machine Learning* (2022), 2 edn., URL `https://christophm.github.io/interpretable-ml-book`.

Morgan, S. L. and C. Winship, *Counterfactuals and causal inference* (Cambridge University Press, 2015).

Murphy, K. P., *Probabilistic Machine Learning: An Introduction* (MIT Press, 2022).

Oberst, M., N. Thams, J. Peters and D. Sontag, "Regularizing towards causal invariance: Linear models with proxies", arXiv preprint arXiv:2103.02477 (2021).

Owen, A. B., "Sobol' Indices and Shapley Value", SIAM/ASA Journal on Uncertainty Quantification **2**, 1, 245–251 (2014).

Owen, A. B. and C. Prieur, "On Shapley Value for Measuring Importance of Dependent Inputs", SIAM/ASA Journal on Uncertainty Quantification **5**, 1, 986–1002 (2017).

Pearl, J., *Causality* (Cambridge University Press, 2009).

Ray, K., S. Botond and A. van der Vaart, "Contributed discussion to "bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects"", Bayesian Analysis **15**, 3, 1026–1028 (2020).

Rencher, A. C. and G. B. Schaalje, *Linear Models in Statistics* (John Wiley & Sons, 2008).

Ribeiro, M. T., S. Singh and C. Guestrin, ""Why Should I Trust You?" Explaining the Predictions of any Classifier", in "Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining", pp. 1135–1144 (2016).

Richardson, T. S. and J. M. Robins, "Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality", Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper **128**, 30, 2013 (2013).

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

Roth, A. E., *The Shapley Value: Essays in Honor of Lloyd S. Shapley* (Cambridge University Press, 1988).

Rotnitzky, A., L. Li and X. Li, "A note on overadjustment in inverse probability weighted estimation", Biometrika **97**, 4, 997–1001 (2010).

Rotnitzky, A. and E. Smucler, "Efficient adjustment sets for population average causal treatment effect estimation in graphical models", The Journal of Machine Learning Research **21**, 1, 7642–7727 (2020).

Saltelli, A., "Making Best Use of Model Evaluations to Compute Sensitivity Indices", Computer physics communications **145**, 2, 280–297 (2002).

Saltelli, A., P. Annoni, I. Azzini, F. Campolongo, M. Ratto and S. Tarantola, "Variance Based Sensitivity Analysis of Model Output. Design and Estimator for the Total Sensitivity Index", Computer physics communications **181**, 2, 259–270 (2010).

Schnitzer, M. E., J. J. Lok and S. Gruber, "Variable selection for confounder control, flexible modeling and collaborative targeted minimum loss-based estimation in causal inference", The international journal of biostatistics **12**, 1, 97–115 (2016).

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization", in "Proceedings of the IEEE international conference on computer vision", pp. 618–626 (2017).

Shalizi, C., *Advanced data analysis from an elementary point of view* (Cambridge University Press, 2021), URL `https://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/`.

Shapley, L. S., "A Value for n-person Games", Contributions to the Theory of Games **2**, 28, 307–317 (1953).

Shi, C., D. M. Blei and V. Veitch, "Adapting neural networks for the estimation of treatment effects", arXiv preprint arXiv:1906.02120 (2019).

Shortreed, S. M. and A. Ertefaie, "Outcome-adaptive lasso: variable selection for causal inference", Biometrics **73**, 4, 1111–1122 (2017).

Sobol, I. M., "Global Sensitivity Indices for Nonlinear Mathematical Models and their Monte Carlo Estimates", Mathematics and computers in simulation **55**, 1-3, 271–280 (2001).

Song, E., B. L. Nelson and J. Staum, "Shapley Effects for Global Sensitivity Analysis: Theory and Computation", SIAM/ASA Journal on Uncertainty Quantification **4**, 1, 1060–1083 (2016).

Stone, C. J., "The Use of Polynomial Splines and their Tensor Products in Multivariate Function Estimation", The Annals of Statistics pp. 118–171 (1994).

Štrumbelj, E. and I. Kononenko, "Explaining Prediction Models and Individual Predictions with Feature Contributions", Knowledge and information systems **41**, 3, 647–665 (2014).

Sundararajan, M. and A. Najmi, "The Many Shapley Values for Model Explanation", in "International Conference on Machine Learning", pp. 9269–9278 (PMLR, 2020).

Sundararajan, M., A. Taly and Q. Yan, "Axiomatic Attribution for Deep Networks", arXiv preprint arXiv:1703.01365 (2017).

Tibshirani, R., "Regression shrinkage and selection via the lasso", Journal of the Royal Statistical Society: Series B (Methodological) **58**, 1, 267–288 (1996).

Vansteelandt, S., M. Bekaert and G. Claeskens, "On model selection and model misspecification in causal inference", Statistical methods in medical research **21**, 1, 7–30 (2012).

Wachter, S., B. Mittelstadt and C. Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR", Harv. JL & Tech. **31**, 841 (2017).

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

Wager, S., W. Du, J. Taylor and R. J. Tibshirani, "High-dimensional regression adjustments in randomized experiments", Proceedings of the National Academy of Sciences **113**, 45, 12673–12678 (2016).

Wang, C., F. Dominici, G. Parmigiani and C. M. Zigler, "Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models", Biometrics **71**, 3, 654–665 (2015).

Wang, J., J. Wiens and S. Lundberg, "Shapley Flow: A Graph-based Approach to Interpreting Model Predictions", arXiv preprint arXiv:2010.14592 (2020).

Wilson, A. and B. J. Reich, "Confounder selection via penalized credible regions", Biometrics **70**, 4, 852–861 (2014).

Witte, J., L. Henckel, M. H. Maathuis and V. Didelez, "On efficient adjustment in causal graphs", The Journal of Machine Learning Research **21**, 1, 9956–10000 (2020).

Wright, S., "On the nature of size factors", Genetics **3**, 4, 367 (1918).

Wright, S., "The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs", Proceedings of the National Academy of Sciences of the United States of America **6**, 6, 320 (1920).

Wright, S., "Correlation and causation", Journal of Agricultural Research **22**, 557–585 (1921).

Zhang, L., Y. Wang, M. Schuemie, D. Blei and G. Hripcsak, "Adjusting for indirectly measured confounding using large-scale propensity scores", arXiv preprint arXiv:2110.12235 (2022).

Zigler, C. M. and F. Dominici, "Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects", Journal of the American Statistical Association **109**, 505, 95–107 (2014).

Zigler, C. M., K. Watts, R. W. Yeh, Y. Wang, B. A. Coull and F. Dominici, "Model feedback in bayesian propensity score estimation", Biometrics **69**, 1, 263–273 (2013).

APPENDIX A

PROOFS

Proof of Proposition 2 in Section 3.3

We consider a sample of n i.i.d. observations of (X, Y, Z) from the data generating process in Section 3.1. We assume that there exists a function λ defined on $\mathcal{X}$ such that the ATE is identified conditional on $\lambda(X)$. We assume that $|\lambda(X)| = J < K = |\mathcal{X}|$ so that the unique values of $\lambda(X)$ define a non-trivial "coarsening" of X. Consider a function $s(X)$ such that there exists at least one pair $x, x' \in \mathcal{X}$ such that $s(x) \neq s(x')$ while $\lambda(x) = \lambda(x')$. We assume that $s(X)$ also identifies the ATE, so that conditioning on $s(X)$ does not induce collider bias.

For each $j \in \lambda(\mathcal{X})$, there exist $m(j) \geq 1$ unique values of $s(\mathcal{X})$, which we denote as $\{j_1, \ldots, j_m\}$. By the definition of $s(X)$, there exists at least one $j \in \lambda(\mathcal{X})$ such that $m(j) > 1$.

We define a stratification estimator of the ATE based on $\lambda(X)$ as:

$$
\begin{aligned}
\bar{\tau}^{\lambda}_{strat} &= \sum_{j \in \lambda(\mathcal{X})} \frac{N_j}{n} \left(\bar{Y}_{j,1} - \bar{Y}_{j,0} \right) \\
N_j &= \sum_{i=1}^{n} \mathbf{1} \left\{ \lambda(X_i) = j \right\} \\
N_{j,1} &= \sum_{i=1}^{n} \mathbf{1} \left\{ \lambda(X_i) = j \right\} \mathbf{1} \left(Z_i = 1 \right\} \\
N_{j,0} &= \sum_{i=1}^{n} \mathbf{1} \left\{ \lambda(X_i) = j \right\} \mathbf{1} \left(Z_i = 0 \right\} \\
\bar{Y}_{j,1} &= \frac{1}{N_{j,1}} \sum_{i=1}^{n} Y_i \mathbf{1} \left\{ \lambda(X_i) = j \right\} \mathbf{1} \left(Z_i = 1 \right\} \\
\bar{Y}_{j,0} &= \frac{1}{N_{j,0}} \sum_{i=1}^{n} Y_i \mathbf{1} \left\{ \lambda(X_i) = j \right\} \mathbf{1} \left(Z_i = 0 \right\}
\end{aligned}
$$

We define an additional stratification estimator based on $s(X)$ as

$$\begin{aligned}
\bar{\tau}^s_{strat} &= \sum_{j \in \lambda(\mathcal{X})} \left(\sum_{\ell=1}^{m(j)} \frac{N_{j\ell}}{n} \left(\bar{Y}_{j\ell,1} - \bar{Y}_{j\ell,0} \right) \right) \\
N_{j\ell} &= \sum_{i=1}^{n} \mathbf{1}\left\{ s(X_i) = j\ell \right\} \\
N_{j\ell,1} &= \sum_{i=1}^{n} \mathbf{1}\left\{ s(X_i) = j\ell \right\} \mathbf{1}\left\{ Z_i = 1 \right\} \\
N_{j\ell,0} &= \sum_{i=1}^{n} \mathbf{1}\left\{ s(X_i) = j\ell \right\} \mathbf{1}\left\{ Z_i = 0 \right\} \\
\bar{Y}_{j\ell,1} &= \frac{1}{N_{j\ell,1}} \sum_{i=1}^{n} Y_i \mathbf{1}\left\{ s(X_i) = j\ell \right\} \mathbf{1}\left(Z_i = 1 \right) \\
\bar{Y}_{j\ell,0} &= \frac{1}{N_{j\ell,0}} \sum_{i=1}^{n} Y_i \mathbf{1}\left\{ s(X_i) = j\ell \right\} \mathbf{1}\left(Z_i = 0 \right)
\end{aligned}$$

where the notation $s(X_i) = j\ell$ is overloaded here to denote an *equivalence*, rather than equality of the stratum labels. To explain this, consider some $j \in \lambda(\mathcal{X})$ with $m(j) > 1$ so that there are $m(j)$ unique levels of $s(\mathcal{X})$ within stratum j. We define an ordering of these sub-strata such that they can be uniquely identified by $j\ell$ for every $\ell \in \{1, \ldots, m(j)\}$. This ordering establishes a one-to-one correspondence between the values $s(x)$ for every x such that $\lambda(x) = j$ and the pairs (j, ℓ).

Now, we consider a $j \in \lambda(\mathcal{X})$ with $m(j) > 1$. We introduce some notation.

$$\begin{aligned}
\mu_{j,1} &= \mathbb{E}\left(Y \mid \lambda(X) = j, Z = 1 \right) & \mu_{j\ell,1} &= \mathbb{E}\left(Y \mid s(X) = j\ell, Z = 1 \right) \\
\mu_{j,0} &= \mathbb{E}\left(Y \mid \lambda(X) = j, Z = 0 \right) & \mu_{j\ell,0} &= \mathbb{E}\left(Y \mid s(X) = j\ell, Z = 0 \right) \\
\sigma^2_{j,1} &= \mathrm{Var}\left(Y \mid \lambda(X) = j, Z = 1 \right) & \sigma^2_{j\ell,1} &= \mathrm{Var}\left(Y \mid s(X) = j\ell, Z = 1 \right) \\
\sigma^2_{j,0} &= \mathrm{Var}\left(Y \mid \lambda(X) = j, Z = 0 \right) & \sigma^2_{j\ell,0} &= \mathrm{Var}\left(Y \mid s(X) = j\ell, Z = 0 \right)
\end{aligned}$$

By the law of iterated expectations and the law of total variance, it follows that

$$\begin{aligned}
\mu_{j,1} &= \mathbb{E}\left(\mathbb{E}\left(Y \mid s(X) = j\ell, Z = 1 \right) \mid \lambda(X) = j, Z = 1 \right) = \mathbb{E}\left(\mu_{j\ell,1} \mid \lambda(X) = j, Z = 1 \right) \\
\sigma^2_{j,1} &= \mathbb{E}\left(\mathrm{Var}\left(Y \mid s(X) = j\ell, Z = 1 \right) \mid \lambda(X) = j, Z = 1 \right) \\
&\quad + \mathrm{Var}\left(\mathbb{E}\left(Y \mid s(X) = j\ell, Z = 1 \right) \mid \lambda(X) = j, Z = 1 \right) \\
&= \mathbb{E}\left(\sigma^2_{j\ell,1} \mid \lambda(X) = j, Z = 1 \right) + \mathrm{Var}\left(\mu_{j\ell,1} \mid \lambda(X) = j, Z = 1 \right)
\end{aligned}$$

We denote

$$\begin{aligned}
\bar{\mu}_{j,1} &= \mathbb{E}\left(\mu_{j\ell,1} \mid \lambda(X) = j, Z = 1 \right) = \mu_{j,1} \\
\bar{\sigma}^2_{j,1} &= \mathbb{E}\left(\sigma^2_{j\ell,1} \mid \lambda(X) = j, Z = 1 \right) \\
\mathbf{v}\left(\mu_{j\ell,1} \right) &= \mathrm{Var}\left(\mu_{j\ell,1} \mid \lambda(X) = j, Z = 1 \right)
\end{aligned}$$

Conditioning on $\mathbf{N} = \{N_{j1,1}, \ldots, N_{jm,1}, N_{j1,0}, \ldots, N_{jm,0}\}$, we see for a given j that

$$\begin{aligned}
\text{Var}\left(\sum_{\ell=1}^{m(j)} N_{j\ell}\bar{Y}_{j\ell,1} \mid \mathbf{N}\right) &= \sum_{\ell=1}^{m(j)} N_{j\ell}^2 \text{Var}\left(\bar{Y}_{j\ell,1}\right) = \sum_{\ell=1}^{m(j)} N_{j\ell}^2 \frac{\sigma_{j\ell,1}^2}{N_{j\ell,1}} \\
&= \sum_{\ell=1}^{m(j)} \frac{(N_{j\ell,1} + N_{j\ell,0})^2}{N_{j\ell,1}} \sigma_{j\ell,1}^2 \\
\text{Var}\left(N_j\bar{Y}_{j,1} \mid \mathbf{N}\right) &= N_j^2 \text{Var}\left(\bar{Y}_{j,1}\right) = N_j^2 \frac{\sigma_{j,1}^2}{N_{j,1}} = \frac{(N_{j,1} + N_{j,0})^2}{N_{j,1}} \sigma_{j,1}^2 \\
&= \frac{\left(\sum_{\ell=1}^{m(j)} (N_{j\ell,1} + N_{j\ell,0})\right)^2}{\sum_{\ell=1}^{m(j)} N_{j\ell,1}} \sigma_{j,1}^2 \\
&= \frac{\left(\sum_{\ell=1}^{m(j)} (N_{j\ell,1} + N_{j\ell,0})\right)^2}{\sum_{\ell=1}^{m(j)} N_{j\ell,1}} \left(\bar{\sigma}_{j,1}^2 + \mathbf{v}\left(\mu_{j\ell,1}\right)\right)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left(\sum_{\ell=1}^{m(j)} N_{j\ell}\bar{Y}_{j\ell,1} \mid \mathbf{N}\right) &= \sum_{\ell=1}^{m(j)} N_{j\ell}\mathbb{E}\left(\bar{Y}_{j\ell,1}\right) = \sum_{\ell=1}^{m(j)} N_{j\ell}\mathbb{E}\left(Y \mid s(X) = j\ell, Z = 1\right) \\
&= \sum_{\ell=1}^{m(j)} N_{j\ell}\mu_{j\ell,1} \\
\mathbb{E}\left(N_j\bar{Y}_{j,1} \mid \mathbf{N}\right) &= N_j\mathbb{E}\left(\bar{Y}_{j,1}\right) = N_j\mathbb{E}\left(Y \mid \lambda(X) = j, Z = 1\right) = N_j\mu_{j,1} \\
&= \left(\sum_{\ell=1}^{m(j)} N_{j\ell,1}\right) \bar{\mu}_{j,1}
\end{aligned}$$

Thus, we have that

$$\begin{aligned}
\operatorname{Var}\left(\sum_{\ell=1}^{m(j)} N_{j\ell}\bar{Y}_{j\ell,1}\right) &= \mathbb{E}\left(\operatorname{Var}\left(\sum_{\ell=1}^{m(j)} N_{j\ell}\bar{Y}_{j\ell,1} \mid \mathbf{N}\right)\right) \\
&\quad + \operatorname{Var}\left(\mathbb{E}\left(\sum_{\ell=1}^{m(j)} N_{j\ell}\bar{Y}_{j\ell,1} \mid \mathbf{N}\right)\right) \\
&= \mathbb{E}\left(\sum_{\ell=1}^{m(j)} \frac{(N_{j\ell,1}+N_{j\ell,0})^2}{N_{j\ell,1}}\sigma^2_{j\ell,1}\right) + \operatorname{Var}\left(\sum_{\ell=1}^{m(j)} N_{j\ell}\mu_{j\ell,1}\right) \\
\operatorname{Var}\left(N_j\bar{Y}_{j,1}\right) &= \mathbb{E}\left(\operatorname{Var}\left(N_j\bar{Y}_{j,1} \mid \mathbf{N}\right)\right) + \operatorname{Var}\left(\mathbb{E}\left(N_j\bar{Y}_{j,1} \mid \mathbf{N}\right)\right) \\
&= \mathbb{E}\left(\frac{\left(\sum_{\ell=1}^{m(j)}(N_{j\ell,1}+N_{j\ell,0})\right)^2}{\sum_{\ell=1}^{m(j)} N_{j\ell,1}}\left(\bar{\sigma}^2_{j,1} + \mathbf{v}\left(\mu_{j\ell,1}\right)\right)\right) \\
&\quad + \operatorname{Var}\left(\left(\sum_{\ell=1}^{m(j)} N_{j\ell,1}\right)\bar{\mu}_{j,1}\right)
\end{aligned}$$

We can broaden this to entire set of observations:

$$\begin{aligned}
\operatorname{Var}\left(\bar{\tau}^s_{strat}\right) &= \operatorname{Var}\left(\sum_{j\in\lambda(\mathcal{X})}\sum_{\ell=1}^{m(j)} \frac{N_{j\ell}}{n}\left(\bar{Y}_{j\ell,1} - \bar{Y}_{j\ell,0}\right)\right) \\
&= \frac{1}{n^2}\left[\mathbb{E}\left(\sum_{j\in\lambda(\mathcal{X})}\sum_{\ell=1}^{m(j)}(N_{j\ell,1}+N_{j\ell,0})^2\left(\frac{\sigma^2_{j\ell,1}}{N_{j\ell,1}} + \frac{\sigma^2_{j\ell,0}}{N_{j\ell,0}}\right)\right)\right] \\
&\quad + \frac{1}{n^2}\left[\operatorname{Var}\left(\sum_{j\in\lambda(\mathcal{X})}\sum_{\ell=1}^{m(j)} N_{j\ell}\left(\mu_{j\ell,1} - \mu_{j\ell,0}\right)\right)\right] \\
\operatorname{Var}\left(\bar{\tau}^\lambda_{strat}\right) &= \operatorname{Var}\left(\sum_{j\in\lambda(\mathcal{X})} \frac{N_j}{n}\left(\bar{Y}_{j,1} - \bar{Y}_{j,0}\right)\right) \\
&= \frac{1}{n^2}\mathbb{E}\left(\sum_{j\in\lambda(\mathcal{X})}\left(\sum_{\ell=1}^{m(j)}(N_{j\ell,1}+N_{j\ell,0})\right)^2(v_1+v_0)\right) \\
&\quad + \frac{1}{n^2}\operatorname{Var}\left(\sum_{j\in\lambda(\mathcal{X})}\left(\sum_{\ell=1}^{m(j)} N_{j\ell}\right)(\bar{\mu}_{j,1} - \bar{\mu}_{j,0})\right)
\end{aligned}$$

where

$$v_1 = \frac{\bar{\sigma}^2_{j,1} + \mathbf{v}\left(\mu_{j\ell,1}\right)}{\sum_{\ell=1}^{m(j)} N_{j\ell,1}}$$
$$v_0 = \frac{\bar{\sigma}^2_{j,0} + \mathbf{v}\left(\mu_{j\ell,0}\right)}{\sum_{\ell=1}^{m(j)} N_{j\ell,0}}$$

Case I: Equal Sub-Strata Means and Variances

If $\sigma^2_{j\ell,i} = \bar{\sigma}^2_{j,i}$ and $\mu_{j\ell,i} = \bar{\mu}_{j,i}$ for all $\ell, i \in \{1, \ldots, m(j)\} \times \{0, 1\}$, then $\mathbf{v}\left(\mu_{j\ell,i}\right) = 0$ and the variance and expectation terms factor out of both expressions and we are left to compare the nonlinear sums of the strata cell sizes. Focusing on $\lambda(X) = j$ and $Z = 1$, we show by induction that $\sum_{\ell=1}^{m(j)} \frac{(N_{j\ell,1}+N_{j\ell,0})^2}{N_{j\ell,1}} \geq \frac{(\sum_{\ell=1}^{m(j)}(N_{j\ell,1}+N_{j\ell,0}))^2}{\sum_{\ell=1}^{m(j)} N_{j\ell,1}}$ for positive cell sizes $N_{j\ell,1}$.

For the base case, suppose that $m(j) = 2$ so that

$$\sum_{\ell=1}^{m(j)} \frac{(N_{j\ell,1} + N_{j\ell,0})^2}{N_{j\ell,1}} = \frac{(N_{j1,1} + N_{j1,0})^2}{N_{j1,1}} + \frac{(N_{j2,1} + N_{j2,0})^2}{N_{j2,1}}$$

and

$$\frac{(\sum_{\ell=1}^{m(j)}(N_{j\ell,1} + N_{j\ell,0}))^2}{\sum_{\ell=1}^{m(j)} N_{j\ell,1}} = \frac{((N_{j1,1} + N_{j1,0}) + (N_{j2,1} + N_{j2,0}))^2}{N_{j1,1} + N_{j2,1}}.$$

For ease of exposition, we let

$$a = N_{j_1,1} \qquad b = N_{j_1,0}$$
$$c = N_{j_2,1} \qquad d = N_{j_2,0}$$

and we thus compare $\frac{(a+b)^2}{a} + \frac{(c+d)^2}{c}$ with $\frac{(a+b+c+d)^2}{a+c}$ where $a, b, c, d > 0$

$$\begin{aligned}
0 &\leq [c(a+b) - a(c+d)]^2 \\
0 &\leq c^2(a+b)^2 + a^2(c+d)^2 - 2ac(a+b)(c+d) \\
2ac(a+b)(c+d) &\leq c^2(a+b)^2 + a^2(c+d)^2 \\
ac\left[(a+b) + (c+d)\right]^2 &\leq \left[c(a+b)^2 + a(c+d)^2\right](a+c) \\
\frac{\left[(a+b) + (c+d)\right]^2}{(a+c)} &\leq \frac{\left[c(a+b)^2 + a(c+d)^2\right]}{ac} \\
\frac{\left[(a+b) + (c+d)\right]^2}{(a+c)} &\leq \frac{(a+b)^2}{a} + \frac{(c+d)^2}{c}
\end{aligned}$$

Now, we proceed to the induction case. Assume that

$$\sum_{\ell=1}^{m(j)} \frac{(N_{j\ell,1} + N_{j\ell,0})^2}{N_{j\ell,1}} \geq \frac{(\sum_{\ell=1}^{m(j)}(N_{j\ell,1} + N_{j\ell,0}))^2}{\sum_{\ell=1}^{m(j)} N_{j\ell,1}}.$$

We consider a new stratum, indexed $m(j)+1$ and we see that

$$\sum_{\ell=1}^{m(j)} \frac{(N_{j\ell,1}+N_{j\ell,0})^2}{N_{j\ell,1}} \geq \frac{(\sum_{\ell=1}^{m(j)}(N_{j\ell,1}+N_{j\ell,0}))^2}{\sum_{\ell=1}^{m(j)} N_{j\ell,1}}$$
$$\sum_{\ell=1}^{m(j)} \frac{(N_{j\ell,1}+N_{j\ell,0})^2}{N_{j\ell,1}} + \omega \geq \frac{(\sum_{\ell=1}^{m(j)}(N_{j\ell,1}+N_{j\ell,0}))^2}{\sum_{\ell=1}^{m(j)} N_{j\ell,1}} + \omega$$
$$\sum_{\ell=1}^{m(j)+1} \frac{(N_{j\ell,1}+N_{j\ell,0})^2}{N_{j\ell,1}} \geq \frac{(\sum_{\ell=1}^{m(j)}(N_{j\ell,1}+N_{j\ell,0}))^2}{\sum_{\ell=1}^{m(j)} N_{j\ell,1}} + \omega$$

where

$$\omega = \frac{(N_{m(j)+1,1}+N_{m(j)+1,0})^2}{N_{m(j)+1,1}}$$

Letting

$$a = \sum_{\ell=1}^{m(j)} N_{j\ell,1} \qquad b = \sum_{\ell=1}^{m(j)} N_{j\ell,0}$$
$$c = N_{m(j)+1,1} \qquad d = N_{m(j)+1,0}$$

we see that

$$\begin{aligned}
&\frac{(\sum_{\ell=1}^{m(j)}(N_{j\ell,1}+N_{j\ell,0}))^2}{\sum_{\ell=1}^{m(j)} N_{j\ell,1}} + \frac{(N_{m(j)+1,1}+N_{m(j)+1,0})^2}{N_{m(j)+1,1}} \\
&\quad = \frac{(a+b)^2}{a} + \frac{(c+d)^2}{c} \\
&\quad \geq \frac{(a+b+c+d)^2}{a+c} = \frac{(\sum_{\ell=1}^{m(j)+1}(N_{j\ell,1}+N_{j\ell,0}))^2}{\sum_{\ell=1}^{m(j)+1} N_{j\ell,1}}
\end{aligned}$$

and the relationship follows by induction.

Thus, when $\mathbb{E}(Y \mid s(X)=j\ell, Z=1)$ and $\mathrm{Var}(Y \mid s(X)=j\ell, Z=1)$ are constant for all $j \in \lambda(\mathcal{X})$ and $\ell \in \{1,\ldots,m(j)\}$, it follows that $\mathrm{Var}\left(\bar{\tau}^{\lambda}_{strat}\right) \leq \mathrm{Var}\left(\bar{\tau}^{s}_{strat}\right)$.

We now let

$$\alpha_1 = \frac{1}{n^2}\mathbb{E}\left(\sum_{j\in\lambda(\mathcal{X})} \bar{\sigma}^2_{j,1}\left(\left(\sum_{\ell=1}^{m(j)} \frac{(N_{j\ell,1}+N_{j\ell,0})^2}{N_{j\ell,1}}\right) - \left(\frac{\left(\sum_{\ell=1}^{m(j)}(N_{j\ell,1}+N_{j\ell,0})\right)^2}{\sum_{\ell=1}^{m(j)} N_{j\ell,1}}\right)\right)\right)$$
$$\alpha_0 = \frac{1}{n^2}\mathbb{E}\left(\sum_{j\in\lambda(\mathcal{X})} \bar{\sigma}^2_{j,0}\left(\left(\sum_{\ell=1}^{m(j)} \frac{(N_{j\ell,1}+N_{j\ell,0})^2}{N_{j\ell,0}}\right) - \left(\frac{\left(\sum_{\ell=1}^{m(j)}(N_{j\ell,1}+N_{j\ell,0})\right)^2}{\sum_{\ell=1}^{m(j)} N_{j\ell,0}}\right)\right)\right)$$

so that $\alpha = \alpha_1 + \alpha_0$ is the degree to which $\mathrm{Var}\left(\bar{\tau}^{\lambda}_{strat}\right) \leq \mathrm{Var}\left(\bar{\tau}^{s}_{strat}\right)$ when all substrata of $\lambda(X)$ are constant. We see that this depends on $\bar{\sigma}^2_{j,1}$, $\bar{\sigma}^2_{j,0}$, and the distribution of $N_{j\ell,i}$ for each j and i.

Case II: Unequal Strata Means or Variances

We partition $\lambda(\mathcal{X})$ into three sets:

- $\mathcal{A}$: $m(a) = 1$ for all $a \in \mathcal{A}$
- $\mathcal{B}$: for all $b \in \mathcal{B}$, either
  - $m(b) > 1$
  - $\sigma^2_{b\ell,i}$ and $\mu_{b\ell,i}$ are constant for all $\ell \in m(b)$, $i \in \{0, 1\}$
- $\mathcal{C}$: For all $c \in \mathcal{C}$
  - $m(c) > 1$, and
  - $\sigma^2_{c\ell,i}$ or $\mu_{c\ell,i}$ is non-constant for some $i \in \{0, 1\}$

In the previous section $\mathcal{C} = \varnothing$, so that the variance comparison is uncomplicated: $s(X)$ was "overstratified" relative to $\lambda(X)$ and as a result $\text{Var}\left(\bar{\tau}^{\lambda}_{strat}\right) \leq \text{Var}\left(\bar{\tau}^{s}_{strat}\right)$.

In this case, $\mathcal{C} \neq \varnothing$ so that there may be variance reduction to stratification (see Lohr (2019) for one reference). We define several terms

$$\beta_1 = \frac{1}{n^2}\mathbb{E}\left(\sum_{b\in\mathcal{B}}\bar{\sigma}^2_{b,1}\left(\left(\sum_{\ell=1}^{m(b)}\frac{(N_{b\ell,1}+N_{b\ell,0})^2}{N_{b\ell,1}}\right) - \left(\frac{\left(\sum_{\ell=1}^{m(b)}(N_{b\ell,1}+N_{b\ell,0})\right)^2}{\sum_{\ell=1}^{m(b)}N_{b\ell,1}}\right)\right)\right)$$

$$\beta_0 = \frac{1}{n^2}\mathbb{E}\left(\sum_{b\in\mathcal{B}}\bar{\sigma}^2_{b,0}\left(\left(\sum_{\ell=1}^{m(b)}\frac{(N_{b\ell,1}+N_{b\ell,0})^2}{N_{b\ell,0}}\right) - \left(\frac{\left(\sum_{\ell=1}^{m(b)}(N_{b\ell,1}+N_{b\ell,0})\right)^2}{\sum_{\ell=1}^{m(b)}N_{b\ell,0}}\right)\right)\right)$$

$$c_1 = \frac{1}{n^2}\mathbb{E}\left(\sum_{c\in\mathcal{C}}\left(\sum_{\ell=1}^{m(c)}(N_{c\ell,1}+N_{c\ell,0})\right)^2\left(\frac{\bar{\sigma}^2_{c,1}+\mathbf{v}\left(\mu_{c\ell,1}\right)}{\sum_{\ell=1}^{m(c)}N_{c\ell,1}}\right)\right)$$
$$+\frac{1}{n^2}\mathrm{Var}\left(\sum_{c\in\mathcal{C}}\left(\sum_{\ell=1}^{m(c)}N_{c\ell}\right)(\bar{\mu}_{c,1})\right)$$

$$c_0 = \frac{1}{n^2}\mathbb{E}\left(\sum_{c\in\mathcal{C}}\left(\sum_{\ell=1}^{m(c)}(N_{c\ell,1}+N_{c\ell,0})\right)^2\left(\frac{\bar{\sigma}^2_{c,0}+\mathbf{v}\left(\mu_{c\ell,0}\right)}{\sum_{\ell=1}^{m(c)}N_{c\ell,0}}\right)\right)$$
$$+\frac{1}{n^2}\mathrm{Var}\left(\sum_{c\in\mathcal{C}}\left(\sum_{\ell=1}^{m(c)}N_{c\ell}\right)(-\bar{\mu}_{c,0})\right)$$

$$c_2 = \frac{1}{n^2}\mathbb{E}\left(\sum_{c\in\mathcal{C}}\sum_{\ell=1}^{m(c)}(N_{c\ell,1}+N_{c\ell,0})^2\left(\frac{\sigma^2_{c\ell,1}}{N_{c\ell,1}}+\frac{\sigma^2_{c\ell,0}}{N_{c\ell,0}}\right)\right)$$
$$+\frac{1}{n^2}\mathrm{Var}\left(\sum_{c\in\mathcal{C}}\sum_{\ell=1}^{m(c)}N_{c\ell}\left(\mu_{c\ell,1}-\mu_{c\ell,0}\right)\right)$$

$$\eta = c_1 + c_0 - c_2$$
$$\nu = \beta_1 + \beta_0$$

We see that $\mathrm{Var}\left(\bar{\tau}^{\lambda}_{strat}\right) > \mathrm{Var}\left(\bar{\tau}^{s}_{strat}\right)$ if $\eta > \nu$, where η refers to the reduction in variance by stratifying on $s(X)$ within $\mathcal{C}$ and ν refers to the increase in variance by "over-stratifying" on $\mathcal{B}$.

APPENDIX B

SIMULATION RESULTS

Simulation Results of Chapter 4

The tables below present ATE RMSEs of each method under consideration at varying values of p and κ. "XBCF" refers to the accelerated Bayesian Causal Forest (Krantsevich *et al.* (2022)) and "XBCF-MP" refers to our proposed modification of XBCF in which the covariates are augmented with multiple estimates of $\hat{\pi}(s(X))$ for different subsets of the covariates defined by a s. "GRF" refers to Generalized Random Forest (Athey *et al.* (2019)). "DRL" refers to the "DR-Learner" (Kennedy (2022)). "SL," "TL," and "XL" refer to the "S-Learner," "T-Learner," "X-Learner," respectively (Künzel *et al.* (2019)).

Known Propensities, Default XBCF

DGP	p	κ	XBCF	XBCF-MP	GRF	DRL	SL	TL	XL
1	10	0.25	0.27	0.23	0.44	2.22	0.84	1.31	0.64
1	10	0.50	0.42	0.40	0.57	3.05	1.32	2.09	1.29
1	10	1.00	0.84	0.75	0.87	4.47	2.41	3.73	2.73
1	10	2.00	1.88	1.44	1.55	5.07	3.59	4.93	4.16
1	50	0.25	0.30	0.24	0.48	3.68	1.16	1.82	1.15
1	50	0.50	0.47	0.42	0.57	4.54	2.08	2.99	2.23
1	50	1.00	0.94	0.77	0.88	5.24	3.61	5.00	4.30
1	50	2.00	2.63	1.69	1.56	5.67	4.02	5.81	5.33
2	10	0.25	0.30	0.30	0.48	2.98	2.84	2.78	2.63
2	10	0.50	0.52	0.52	0.64	3.49	3.16	3.65	2.96
2	10	1.00	1.04	0.98	0.95	4.13	3.38	4.56	3.43
2	10	2.00	2.32	1.99	1.69	4.66	3.52	5.30	4.06
2	50	0.25	0.34	0.32	0.54	3.95	3.25	3.92	3.28
2	50	0.50	0.62	0.56	0.67	4.40	3.48	4.63	3.65
2	50	1.00	1.23	1.05	0.98	5.04	3.87	5.81	4.69
2	50	2.00	3.26	2.51	1.75	5.54	4.20	6.52	5.65
3	10	0.25	0.03	0.04	0.04	0.06	0.03	0.04	0.03
3	10	0.50	0.07	0.07	0.07	0.10	0.06	0.07	0.06
3	10	1.00	0.13	0.14	0.13	0.18	0.12	0.14	0.12
3	10	2.00	0.25	0.25	0.25	0.37	0.22	0.27	0.24
3	50	0.25	0.03	0.03	0.05	0.06	0.03	0.04	0.03
3	50	0.50	0.07	0.07	0.07	0.11	0.06	0.08	0.07
3	50	1.00	0.14	0.14	0.14	0.20	0.14	0.16	0.13
3	50	2.00	0.25	0.24	0.25	0.39	0.26	0.30	0.25

Table B.1: Comparison of machine-learning-based ATE estimators with known propensities. RMSE of ATE across 1,000 simulations with $n = 500$.

Estimated Propensities, Default XBCF

DGP	p	κ	XBCF	XBCF-MP	GRF	DRL	SL	TL	XL
1	10	0.25	0.62	0.89	4.09	2.22	0.83	1.32	0.89
1	10	0.50	1.00	1.49	4.23	3.10	1.34	2.12	1.62
1	10	1.00	2.00	2.77	4.65	4.46	2.42	3.72	3.09
1	10	2.00	4.69	4.66	5.27	5.13	3.68	5.01	4.47
1	50	0.25	0.94	1.37	4.84	3.69	1.18	1.84	1.45
1	50	0.50	1.88	2.47	4.98	4.55	2.05	2.98	2.55
1	50	1.00	3.60	3.98	5.42	5.29	3.64	5.03	4.59
1	50	2.00	5.47	5.28	6.11	5.68	4.06	5.84	5.51
2	10	0.25	1.13	0.60	3.49	3.00	2.86	2.80	2.71
2	10	0.50	1.43	0.88	3.59	3.44	3.14	3.61	3.08
2	10	1.00	2.11	1.45	3.98	4.18	3.40	4.59	3.65
2	10	2.00	3.59	2.77	4.68	4.72	3.62	5.38	4.35
2	50	0.25	1.30	0.62	4.60	3.96	3.24	3.90	3.41
2	50	0.50	1.75	0.98	4.85	4.41	3.50	4.63	3.87
2	50	1.00	2.72	1.67	5.45	5.11	3.87	5.83	4.93
2	50	2.00	4.97	3.39	6.45	5.51	4.21	6.51	5.79
3	10	0.25	0.03	0.04	0.04	0.06	0.03	0.04	0.03
3	10	0.50	0.07	0.09	0.07	0.10	0.06	0.07	0.06
3	10	1.00	0.15	0.21	0.13	0.19	0.12	0.14	0.12
3	10	2.00	0.26	0.30	0.26	0.38	0.22	0.27	0.24
3	50	0.25	0.03	0.05	0.08	0.06	0.03	0.04	0.03
3	50	0.50	0.07	0.11	0.11	0.11	0.06	0.08	0.07
3	50	1.00	0.14	0.23	0.17	0.20	0.13	0.16	0.13
3	50	2.00	0.27	0.34	0.31	0.42	0.27	0.32	0.28

Table B.2: Comparison of machine-learning-based ATE estimators with estimated propensities. RMSE of ATE across 1,000 simulations with $n = 500$.

Known Propensities, XBCF with Multiple $\hat{Y}$ Estimates as Covariates

DGP	p	κ	XBCF	XBCF-MP	GRF	DRL	SL	TL	XL
1	10	0.25	0.30	0.28	0.47	2.20	0.83	1.31	0.63
1	10	0.50	0.41	0.38	0.57	3.09	1.35	2.12	1.32
1	10	1.00	0.68	0.60	0.85	4.49	2.39	3.70	2.70
1	10	2.00	1.13	1.07	1.55	5.14	3.63	4.96	4.20
1	50	0.25	0.33	0.32	0.47	3.70	1.15	1.82	1.15
1	50	0.50	0.49	0.43	0.57	4.52	2.06	3.00	2.23
1	50	1.00	0.72	0.61	0.85	5.27	3.59	4.98	4.28
1	50	2.00	0.98	0.97	1.53	5.71	4.05	5.83	5.35
2	10	0.25	0.28	0.26	0.49	2.97	2.82	2.77	2.61
2	10	0.50	0.36	0.36	0.59	3.46	3.15	3.62	2.96
2	10	1.00	0.59	0.60	0.95	4.15	3.43	4.62	3.48
2	10	2.00	1.14	1.11	1.71	4.68	3.52	5.29	4.05
2	50	0.25	0.29	0.27	0.53	3.97	3.26	3.93	3.28
2	50	0.50	0.37	0.36	0.65	4.38	3.48	4.62	3.64
2	50	1.00	0.59	0.58	0.99	5.07	3.87	5.82	4.71
2	50	2.00	1.23	1.06	1.75	5.53	4.21	6.53	5.65
3	10	0.25	0.12	0.12	0.04	0.06	0.03	0.04	0.03
3	10	0.50	0.12	0.12	0.06	0.10	0.06	0.07	0.06
3	10	1.00	0.18	0.18	0.13	0.18	0.12	0.14	0.12
3	10	2.00	0.28	0.27	0.24	0.37	0.21	0.27	0.23
3	50	0.25	0.17	0.16	0.05	0.06	0.03	0.04	0.03
3	50	0.50	0.18	0.18	0.07	0.11	0.06	0.08	0.06
3	50	1.00	0.25	0.25	0.13	0.21	0.13	0.16	0.13
3	50	2.00	0.32	0.32	0.25	0.37	0.25	0.31	0.26

Table B.3: Comparison of machine-learning-based ATE estimators with known propensities and marginal $\hat{y}$ models as covariates in the XBCF models. RMSE of ATE across 1,000 simulations with $n = 500$.

Estimated Propensities, XBCF with Multiple $\hat{Y}$ Estimates as Covariates

DGP	p	κ	XBCF	XBCF-MP	GRF	DRL	SL	TL	XL
1	10	0.25	0.30	0.28	4.09	2.19	0.83	1.32	0.89
1	10	0.50	0.43	0.40	4.23	3.08	1.35	2.11	1.61
1	10	1.00	0.67	0.60	4.62	4.50	2.41	3.72	3.09
1	10	2.00	1.12	1.10	5.30	5.17	3.70	5.04	4.51
1	50	0.25	0.32	0.29	4.83	3.68	1.17	1.82	1.43
1	50	0.50	0.48	0.39	4.98	4.49	2.07	2.99	2.55
1	50	1.00	0.73	0.55	5.38	5.32	3.61	4.99	4.56
1	50	2.00	1.25	1.97	6.05	5.68	4.01	5.79	5.44
2	10	0.25	0.23	0.26	3.47	2.97	2.83	2.77	2.68
2	10	0.50	0.36	0.43	3.60	3.48	3.16	3.63	3.10
2	10	1.00	0.58	0.70	3.98	4.18	3.42	4.60	3.66
2	10	2.00	1.10	1.22	4.63	4.64	3.55	5.31	4.28
2	50	0.25	0.23	0.26	4.60	3.96	3.25	3.92	3.43
2	50	0.50	0.33	0.38	4.80	4.36	3.47	4.59	3.83
2	50	1.00	0.60	0.62	5.41	5.07	3.85	5.81	4.90
2	50	2.00	1.18	1.04	6.40	5.44	4.15	6.47	5.76
3	10	0.25	0.12	0.11	0.04	0.06	0.03	0.04	0.03
3	10	0.50	0.12	0.10	0.07	0.10	0.06	0.07	0.06
3	10	1.00	0.17	0.15	0.13	0.19	0.12	0.14	0.12
3	10	2.00	0.27	0.25	0.25	0.37	0.21	0.27	0.24
3	50	0.25	0.16	0.15	0.08	0.06	0.03	0.04	0.03
3	50	0.50	0.17	0.14	0.11	0.11	0.07	0.08	0.07
3	50	1.00	0.23	0.18	0.18	0.21	0.14	0.16	0.14
3	50	2.00	0.31	0.26	0.30	0.37	0.25	0.31	0.26

Table B.4: Comparison of machine-learning-based ATE estimators with estimated propensities and marginal $\hat{y}$ models as covariates in the XBCF models. RMSE of ATE across 1,000 simulations with $n = 500$.

Known Propensities, XBCF with Multiple $\hat{Y}$ Estimates as Covariates, GRF Estimated with a First-Stage XBART Model of $\hat{Y}$ Rather than the Default Random Forest

DGP	p	κ	XBCF	XBCF-MP	GRF	DRL	SL	TL	XL
1	10	0.25	0.30	0.28	0.21	2.21	0.83	1.31	0.63
1	10	0.50	0.42	0.38	0.32	3.10	1.34	2.13	1.32
1	10	1.00	0.71	0.64	0.70	4.49	2.39	3.68	2.68
1	10	2.00	1.08	1.04	1.39	5.13	3.68	5.03	4.26
1	50	0.25	0.33	0.32	0.25	3.63	1.15	1.79	1.14
1	50	0.50	0.49	0.43	0.32	4.52	2.05	2.97	2.20
1	50	1.00	0.74	0.62	0.68	5.31	3.61	5.01	4.32
1	50	2.00	0.96	0.92	1.38	5.67	3.99	5.75	5.28
2	10	0.25	0.27	0.26	0.30	2.97	2.82	2.77	2.62
2	10	0.50	0.35	0.34	0.40	3.50	3.16	3.64	2.98
2	10	1.00	0.56	0.58	0.75	4.11	3.37	4.54	3.41
2	10	2.00	1.10	1.09	1.50	4.64	3.55	5.33	4.07
2	50	0.25	0.29	0.28	0.32	3.95	3.25	3.92	3.26
2	50	0.50	0.38	0.36	0.42	4.38	3.48	4.60	3.64
2	50	1.00	0.58	0.55	0.70	5.05	3.87	5.80	4.69
2	50	2.00	1.23	1.06	1.47	5.47	4.11	6.48	5.58
3	10	0.25	0.13	0.12	0.13	0.06	0.03	0.04	0.03
3	10	0.50	0.13	0.13	0.11	0.11	0.06	0.07	0.06
3	10	1.00	0.18	0.17	0.12	0.18	0.12	0.14	0.12
3	10	2.00	0.28	0.27	0.23	0.36	0.22	0.27	0.24
3	50	0.25	0.17	0.17	0.17	0.06	0.03	0.04	0.03
3	50	0.50	0.18	0.18	0.15	0.11	0.06	0.08	0.06
3	50	1.00	0.25	0.25	0.14	0.21	0.13	0.15	0.13
3	50	2.00	0.33	0.32	0.23	0.40	0.26	0.31	0.26

Table B.5: Comparison of machine-learning-based ATE estimators with known propensities, marginal $\hat{y}$ models as covariates in the XBCF models, GRF estimated with XBART for $\hat{y}$ and true propensities for $\hat{\pi}$. RMSE of ATE across 1,000 simulations with $n = 500$.

Estimated Propensities, XBCF with Multiple $\hat{Y}$ Estimates as Covariates, GRF Estimated with a First-Stage XBART Model of $\hat{Y}$ Rather than the Default Random Forest

DGP	p	κ	XBCF	XBCF-MP	GRF	DRL	SL	TL	XL
1	10	0.25	0.30	0.27	0.17	2.20	0.84	1.32	0.89
1	10	0.50	0.44	0.40	0.26	3.05	1.32	2.09	1.59
1	10	1.00	0.69	0.60	1.04	4.48	2.42	3.72	3.08
1	10	2.00	1.15	1.06	2.85	5.13	3.65	5.00	4.47
1	50	0.25	0.32	0.29	0.20	3.69	1.18	1.82	1.44
1	50	0.50	0.49	0.40	0.75	4.54	2.04	2.99	2.56
1	50	1.00	0.72	0.56	2.33	5.26	3.59	4.99	4.56
1	50	2.00	1.25	1.93	4.60	5.73	4.07	5.85	5.51
2	10	0.25	0.23	0.26	0.93	2.98	2.83	2.78	2.69
2	10	0.50	0.36	0.43	1.49	3.47	3.16	3.63	3.09
2	10	1.00	0.59	0.71	2.14	4.15	3.41	4.57	3.64
2	10	2.00	1.10	1.23	3.00	4.74	3.61	5.37	4.34
2	50	0.25	0.25	0.27	1.09	3.94	3.23	3.90	3.41
2	50	0.50	0.34	0.40	1.54	4.36	3.48	4.60	3.83
2	50	1.00	0.59	0.60	2.23	5.05	3.86	5.81	4.89
2	50	2.00	1.17	1.07	3.48	5.50	4.20	6.50	5.78
3	10	0.25	0.12	0.11	0.11	0.06	0.03	0.04	0.03
3	10	0.50	0.12	0.10	0.09	0.10	0.06	0.07	0.06
3	10	1.00	0.18	0.16	0.12	0.20	0.13	0.14	0.13
3	10	2.00	0.27	0.26	0.24	0.39	0.22	0.27	0.24
3	50	0.25	0.16	0.15	0.13	0.06	0.03	0.04	0.03
3	50	0.50	0.17	0.14	0.11	0.11	0.07	0.08	0.07
3	50	1.00	0.24	0.18	0.12	0.20	0.13	0.15	0.12
3	50	2.00	0.31	0.26	0.24	0.40	0.26	0.31	0.26

Table B.6: Comparison of machine-learning-based ATE estimators with estimated propensities, marginal $\hat{y}$ models as covariates in the XBCF models, GRF estimated with XBART for $\hat{y}$ and xgboost for $\hat{\pi}$. RMSE of ATE across 1,000 simulations with $n = 500$.

APPENDIX C

DERIVATIONS

Derivation of RIC from Prediction MSE in Section 3.4

As in the proof of Proposition 2, we assume that X is discrete. Fitting a model of $\mathbb{E}[Y \mid Z, X] = \mu(X) + Z\tau(X)$, we estimate the μ and τ terms nonparametrically using a saturated linear model based on some adjustment set $s(X)$. Estimation follows directly from specification of s, so cast the problem of estimating $\mathbb{E}[Y \mid Z, X]$ as an optimization problem with respect to s,

$$s^* = \arg\min_s ||\mu(s(X)) + Z\tau(s(X)) - Y||^2.$$

In order to solve this saturated linear regression, we encode $s(X)$ as a full-rank design matrix, which we call $\tilde{S}$. Estimating a saturated regression of Y on $\tilde{S}$ and Z gives two sets of coefficients:

- coefficients corresponding to the non-interacted columns of $\tilde{S}$, which we dub α, and
- coefficients corresponding to the columns of $\tilde{S}$ interacted with Z, which we dub β.

Thus, $\hat{\mu}(s(X)) = \tilde{S}\alpha$ and $\hat{\tau}(s(X)) = \tilde{S}\beta$.

Now, observe that for any estimate $\hat{\mu}(s(X))$ and $\hat{\tau}(s(X))$, we have

$$\begin{aligned}
(\hat{Y}_s - Y)^2 &= \left[\hat{\mu}(s(X)) + Z\hat{\tau}(s(X)) - \mu(X) - Z\tau(X) - \epsilon\right]^2 \\
&= \left[\hat{\mu}(s(X)) + Z\hat{\tau}(s(X)) - \mu(X) - Z\tau(X)\right]^2 + \epsilon^2 \\
&\quad - 2\epsilon\left[\hat{\mu}(s(X)) + Z\hat{\tau}(s(X)) - \mu(X) - Z\tau(X)\right] \\
&= \left(\hat{\mu}(s(X)) - \mu(X)\right)^2 + Z^2\left(\hat{\tau}(s(X)) - \tau(X)\right)^2 \\
&\quad + \epsilon^2 - 2\epsilon\left(\hat{\mu}(s(X)) - \mu(X)\right) - 2Z\left(\hat{\tau}(s(X)) - \tau(X)\right)\left[\epsilon + \left(\hat{\mu}(s(X)) - \mu(X)\right)\right]
\end{aligned}$$

Now, note that, given n observations of (X, Y, Z) and a stratification function s, the stratification estimator of the ATE (which is equivalent to the regression estimator and the empirically-weighted IPW estimator with discrete covariates) can be written as

$$\hat{\tau}_s = \sum_{s \in s(\mathcal{X})} \frac{n_s}{n}\left(\bar{Y}_{s,z=1} - \bar{Y}_{s,z=0}\right)$$

where $n_s = \sum_{i=1}^n \mathbf{1}\left(s(X_i) = s\right)$ for a sample of size n.

This targets an estimand of

$$\bar{\tau}_s = \mathbb{E}_{s(X)}\left[\mathbb{E}_Y\left[Y \mid s(X), Z = 1\right] - \mathbb{E}_Y\left[Y \mid s(X), Z = 0\right]\right] = \mathbb{E}_{s(X)}\left[\mathbb{E}\left[\tau(X) \mid s(X)\right]\right]$$

of which the fully-stratified estimand corresponds to the case in which $s(X) = X$

$$\tau_x = \mathbb{E}_X\left[\mathbb{E}_Y\left[Y \mid X, Z = 1\right] - \mathbb{E}_Y\left[Y \mid X, Z = 0\right]\right] = \mathbb{E}_X\left[\tau(X)\right]$$

We now note that $\hat{\tau}(s(X))$ estimated with the same n observations and adjustment set s as above, can be rewritten, for any x_i, as

$$\hat{\tau}(s(x_i)) = \bar{Y}_{s(x_i),1} - \bar{Y}_{s(x_i),0}.$$

We define, for each x_i, a "centered" $\hat{\tau}(s(x_i))$ as

$$\hat{t}(s(x_i)) = \hat{\tau}(s(x_i)) - \hat{\tau}_s$$

So that $\hat{\tau}(s(X)) = \hat{t}(s(X)) + \hat{\tau}_s$ and we can thus represent the predicted value of Y for any s by $\bar{\tau}_s$ and three other terms, which we define below

$$\hat{Y}_s = \hat{\mu}(s(X)) + Z\left(\hat{\tau}_s + \hat{t}(s(X))\right)$$

Similarly, the "true" $\tau(X)$ term of Y can be decomposed for any x

$$\mu(x) + Z\tau(x) = \mu(x) + Z\left(\tau_x + t(x)\right)$$
$$\tau_x = \mathbb{E}\left[\tau(X)\right] \qquad t(x) = \tau(x) - \mathbb{E}\left[\tau(X)\right]$$

This can also be written in terms of aggregated covariate strata $s(X)$

$$\tau_s = \mathbb{E}_{s(X)}\left[\mathbb{E}\left(\Delta_s \mid s(X)\right)\right]$$
$$t(s) = \mathbb{E}\left(\Delta_s \mid s(X) = s\right) - \mathbb{E}_{s(X)}\left[\mathbb{E}\left(\Delta_s \mid s(X)\right)\right]$$

where $\Delta_s = \mathbb{E}\left[Y \mid s(X) = s, Z = 1\right] - \mathbb{E}\left[Y \mid s(X) = s, Z = 0\right]$. If $s(X)$ does not satisfy mean conditional unconfoundedness, then τ_s is not necessarily equal to τ_x.

For any random vector (Y, X, Z) and adjustment set $s(X)$, we have that

$$\begin{aligned}
\left(\hat{Y}_s - Y\right)^2 &= \left(\hat{\mu}(s(X)) + Z\left(\hat{\tau}_s + \hat{t}(s(X))\right) - Y\right)^2 \\
&= \left(\hat{\mu}(s(X)) + Z\left(\hat{\tau}_s + \hat{t}(s(X))\right) - \mu(X) - Z\left(\tau_x + t(X)\right)\right)^2 && \text{(C.1)} \\
&\quad + \left(\mu(X) + Z\left(\tau_x + t(X)\right) - Y\right)^2 && \text{(C.2)} \\
&\quad + 2\left(\hat{Y}_s - \theta_Y\right)\left(\theta_Y - Y\right) && \text{(C.3)}
\end{aligned}$$

where $\theta_Y = \mathbb{E}\left[Y \mid Z, X\right] = \mu(X) + Z\left(\tau_x + t(X)\right)$. Note that C.1 constitutes the "squared prediction error" of $\hat{\mu}(s(X)) + Z\left(\hat{\tau}_s + \hat{t}(s(X))\right)$ with respect to the true structural model $\mu(X) + Z\left(\tau_x + t(X)\right)$, C.2 is a stratification-independent measure of the magnitude of the noise terms of Y, and C.3 is twice the product of the true noise term of Y and the prediction error of $\hat{\mu}(s(X)) + Z\left(\hat{\tau}_s + \hat{t}(s(X))\right)$.

We compare estimators based on different stratification functions $s(x)$ via their mean squared error, $\mathbb{E}\left(\hat{Y} - Y\right)^2$. Since $C.2$ does not depend on the choice of $s(X)$, we denote its expectation as σ^2. Similarly, C.3 is a constant multiple of the prediction error which is squared in equation C.1, so we focus our analysis on equation C.1.

$$\begin{aligned}
&\left(\hat{\mu}(s(X)) + Z\left(\hat{\tau}_s + \hat{t}(s(X))\right) - \mu(X) - Z\left(\tau_x + t(X)\right)\right)^2 \\
&= \left(\left(\hat{\mu}(s(X)) - \mu(X)\right) + Z\left(\bar{\tau}_s - \tau_x\right) + Z\left(\hat{t}(s) - t(x)\right)\right)^2 \\
&= \left(\hat{\mu}(s(X)) - \mu(X)\right)^2 + Z\left(\bar{\tau}_s - \tau_x\right)^2 + Z\left(\hat{t}(s) - t(x)\right)^2 \\
&+ 2Z\left(\hat{\mu}(s(X)) - \mu(X)\right)\left(\bar{\tau}_s - \tau_x\right) + 2Z\left(\hat{\mu}(s(X)) - \mu(X)\right)\left(\hat{t}(s) - t(x)\right) \\
&+ 2Z\left(\bar{\tau}_s - \tau_x\right)\left(\hat{t}(s) - t(x)\right) \\
&= \left(\hat{\mu}(s(X)) - \mu(X)\right)^2 + Z\left(\bar{\tau}_s - \tau_x\right)^2 + Z\left(\hat{t}(s) - t(x)\right)^2 \\
&+ 2Z\left(\hat{\mu}(s(X)) - \mu(X)\right)\left(\hat{\tau}(s(X)) - \tau(X)\right) + 2Z\left(\bar{\tau}_s - \tau_x\right)\left(\hat{t}(s) - t(x)\right)
\end{aligned}$$

We evaluate the expectation of this expression in parts. First, note that

$$\begin{aligned}
\mathbb{E}\left(\hat{\mu}(s(X)) - \mu(X)\right)^2 &= \mathbb{E}\left[\hat{\mu}(s(X)) - \mathbb{E}\left[\mu(X) \mid s(X)\right] + \mathbb{E}\left[\mu(X) \mid s(X)\right] - \mu(X)\right]^2 \\
&= \mathbb{E}\left[\left(\hat{\mu}(s(X)) - \mathbb{E}\left[\mu(X) \mid s(X)\right]\right)^2\right] \\
&\quad + \mathbb{E}\left[\left(\mathbb{E}\left[\mu(X) \mid s(X)\right] - \mu(X)\right)^2\right] \\
&\quad + \mathbb{E}\left[\left(\hat{\mu}(s(X)) - \mathbb{E}\left[\mu(X) \mid s(X)\right]\right)\left(\mathbb{E}\left[\mu(X) \mid s(X)\right] - \mu(X)\right)\right] \\
&= \text{Var}\left(\hat{\mu}(s(X))\right) + \text{ Bias}\left(\hat{\mu}(s(X))\right)^2 + 0 \\
\mathbb{E}\left[Z\left(\bar{\tau}_s - \tau_x\right)^2\right] &= \mathbb{E}\left[\mathbb{E}\left[Z\left(\left(\bar{\tau}_s - \tau_s\right) + \left(\tau_s - \tau_x\right)\right)^2 \mid Z\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Z\left(\bar{\tau}_s - \tau_s\right)^2 \mid Z\right]\right] + \mathbb{E}\left[\mathbb{E}\left[Z\left(\tau_s - \tau_x\right)^2 \mid Z\right]\right] \\
&\quad + 2\mathbb{E}\left[\mathbb{E}\left[Z\left(\bar{\tau}_s - \tau_s\right)\left(\tau_s - \tau_x\right) \mid Z\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Z\left(\bar{\tau}_s - \tau_s\right)^2 \mid Z\right]\right] + \mathbb{E}\left[\mathbb{E}\left[Z\left(\tau_s - \tau_x\right)^2 \mid Z\right]\right] + 0 \\
&= \mathbb{E}\left(\pi(X)\right)\left[\text{ Var}\left(\bar{\tau}_s\right) + \text{ Bias}\left(\bar{\tau}_s\right)^2\right]
\end{aligned}$$

These terms measure the bias and variance of the prognostic function and the average treatment effect computed on the data after stratifying by $s(X)$.

These derivations show that $\mathbb{E}\left(\hat{Y} - Y\right)^2$ can be decomposed into the bias and variance of $\hat{\tau}_s$, the bias and variance of $\hat{\mu}(s(X))$, the bias and variance of $\hat{t}(s(X))$, and expected conditional product terms between $\hat{\mu}$, $\hat{\tau}_s$, and $\hat{t}$.

Minimizing the MSE without any penalty will favor full stratification (i.e. $s(X) = X$). If we introduce a penalized objective function $\mathbb{E}\left(\hat{Y}_s - Y\right)^2 + \alpha|s(X)|$ for some $\alpha > 0$, the optimization problem trades off model fit ($\mathbb{E}\left(\hat{Y}_s - Y\right)^2$) and stratification size $|s(X)|$. While it is certainly possible to construct a bias-variance decomposition of $\hat{Y}_s$ such that $\mathbb{E}\left(\hat{Y}_s - Y\right)^2 = \mathbb{E}\left(\hat{Y}_s - \mathbb{E}(Y)\right)^2 + \left(\mathbb{E}(Y) - Y\right)^2$, this tradeoff is in $\hat{Y}$, not the estimator of interest, $\hat{\tau}_s$.

Derivation of the SHAP Regression

Following Rencher and Schaalje (2008), we can express the objective function as

$$L(\beta, \lambda) = (y - Z\beta)' W (y - Z\beta) + \lambda' (j'\beta - (y_t - y_b))$$

where λ is a Lagrange multiplier.

We can minimize this objective function by differentiating L with respect to β and λ, setting both partial derivatives equal to 0, and checking the determinant of the Hessian matrix. Solving for β gives

$$\begin{aligned}
\hat{\beta} &= \left(Z'WZ\right)^{-1}\left(I - jA^{-1}j'\left(Z'WZ\right)^{-1}\right)Z'Wy + \left(Z'WZ\right)^{-1} jA^{-1}\left(y_t - y_b\right) \\
A &= j'\left(Z'WZ\right)^{-1} j
\end{aligned}$$

We note, as do Lundberg and Lee (2017), that $Z'WZ$ is a symmetric matrix of the form $aJ + bI$, where I is the p-dimension identity matrix and J is a $p \times p$ matrix

of all ones. Solving for a and b, we get that

$$a = \sum_{i=1}^{p-1} \frac{(i-1)}{p\,(p-i)}$$
$$b = \frac{p-1}{p}$$

We can determine $(Z'WZ)^{-1}$ by setting $(Z'WZ)^{-1} = cJ + dI$ and solving

$$(Z'WZ)\,(Z'WZ)^{-1} = I$$
$$(aJ + bI)\,(cJ + dI) = acJ^2 + bcJ + adJ + bdI = pacJ + bcJ + adJ + bdI = I$$

This implies that $pac + bc + ad = 0$ and $bd = 1$ so that $d = 1/b = p/\,(p-1)$ and $c = -\left(\frac{a}{b}\right)\left(\frac{1}{pa+b}\right)$. Now observe that $j'\,(Z'WZ)^{-1} = (cp+d)j'$ and $j'\,(Z'WZ)^{-1}\,j = (cp+d)p$ and thus

$$\begin{aligned}
(j'\,(Z'WZ)^{-1}\,j)^{-1} &= \frac{1}{(cp+d)p} \\
j(j'\,(Z'WZ)^{-1}\,j)^{-1}j'\,(Z'WZ)^{-1} &= \frac{jj'\,(Z'WZ)^{-1}}{(cp+d)p} = J/p \\
(Z'WZ)^{-1}\left(I - jA^{-1}j'\,(Z'WZ)^{-1}\right) &= (cJ+dI)\left(I - \frac{1}{p}J\right) \\
&= dI + cJ - \frac{d}{p}J - cJ = dI - \frac{d}{p}J \\
&= \frac{p}{p-1}I - \frac{1}{p-1}J
\end{aligned}$$

Similarly,

$$(Z'WZ)^{-1}\,j(j'\,(Z'WZ)^{-1}\,j)^{-1} = \frac{(Z'WZ)^{-1}\,j'}{(cp+d)p} = j/p$$

Thus, the regression solution simplifies to

$$\begin{aligned}
\hat{\beta} &= (Z'WZ)^{-1}\left(I - jA^{-1}j'\,(Z'WZ)^{-1}\right)Z'Wy + (Z'WZ)^{-1}\,jA^{-1}\,(y_t - y_b) \\
&= \left(\frac{p}{p-1}I - \frac{1}{p-1}J\right)Z'Wy + \frac{j\,(y_t - y_b)}{p}
\end{aligned}$$

$Z'W$ is a $p \times (2^p - 2)$ matrix in which the columns correspond to coalitions in the Z matrix multiplied by the weight w_i of that coalition. $\frac{p}{p-1}I - \frac{1}{p-1}J$ is a square symmetric matrix. Letting s represent the number of nonzero entries in a given column of $Z'W$,

we see that the weights attached to those nonzero entires is $w(s) = \frac{(p-1)(p-s-1)!(s-1)!}{p!}$ so

$$-\frac{1}{p-1}JZ'W = \begin{pmatrix} -\frac{s(p-1)(p-s-1)!(s-1)!}{(p-1)p!} & \dots & -\frac{s(p-1)(p-s-1)!(s-1)!}{(p-1)p!} \\ \dots & \dots & \dots \\ -\frac{s(p-1)(p-s-1)!(s-1)!}{(p-1)p!} & \dots & -\frac{s(p-1)(p-s-1)!(s-1)!}{(p-1)p!} \end{pmatrix}$$
$$= \begin{pmatrix} -\frac{s(p-s-1)!(s-1)!}{p!} & \dots & -\frac{s(p-s-1)!(s-1)!}{p!} \\ \dots & \dots & \dots \\ -\frac{s(p-s-1)!(s-1)!}{p!} & \dots & -\frac{s(p-s-1)!(s-1)!}{p!} \end{pmatrix}$$

and

$$\frac{p}{p-1}IZ'W = \begin{pmatrix} \frac{p(p-1)(p-s-1)!(s-1)!}{(p-1)p!} & \dots & \frac{p(p-1)(p-s-1)!(s-1)!}{(p-1)p!} \\ \dots & \dots & \dots \\ \frac{p(p-1)(p-s-1)!(s-1)!}{(p-1)p!} & \dots & \frac{p(p-1)(p-s-1)!(s-1)!}{(p-1)p!} \end{pmatrix}$$
$$= \begin{pmatrix} \frac{p(p-s-1)!(s-1)!}{p!} & \dots & \frac{p(p-s-1)!(s-1)!}{p!} \\ \dots & \dots & \dots \\ \frac{p(p-s-1)!(s-1)!}{p!} & \dots & \frac{p(p-s-1)!(s-1)!}{p!} \end{pmatrix}$$

And thus the entries of matrix $B = \left(\frac{p}{p-1}I - \frac{1}{p-1}J\right)Z'W$ are

$$B_{ij} = \begin{cases} \frac{(p-s)!(s-1)!}{p!} & Z_{ji} = 1 \\ -\frac{s(p-s-1)!(s-1)!}{p!} = -\frac{(p-s-1)!s!}{p!} & Z_{ji} = 0 \end{cases}$$

which correspond to Shapley weights with and without the feature of interest included in a coalition. When $s = 0$ or $s = p - 1$, the Shapley weight is exactly $1/p$, which is the weight attached to y_b and y_t in the second term of regression coefficient solution. Thus, since Z includes all $2^p - 2$ synthetic coalitions, the first term of $\hat{\beta}$ is a linear combination of all of the synthetic predictions $f(Z)$ and the second term of $\hat{\beta}$ is a column vector of $y_t/p - y_b/p$.

Rather than view the regression weights as separate error and constraint terms, we can concatenate into one linear operation. Adding y_b to the beginning of the y vector and y_t to the end, we have $y_* = (y_b \quad y' \quad y_t)'$. Similarly, we can append the vector j'/p to both ends of B, getting $B_* = (-j'/p \quad B \quad j'/p)$. Then we see that $\hat{\beta} = B_* y_*$. Since each row of B_* is a set of positive and negative Shapley weights and y_* is the complete set of 2^p model predictions, we see that each entry i in the p rows of $\hat{\beta}$ correspond to the exact Shapley value, ϕ_i, for feature i.

To see this illustrated, we return to the example from Figure 5.1. We can see that

$$Z = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \qquad W = \begin{pmatrix} 1/3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/3 \end{pmatrix}$$

$$Z'WZ = \begin{pmatrix} 1 & 1/3 & 1/3 \\ 1/3 & 1 & 1/3 \\ 1/3 & 1/3 & 1 \end{pmatrix}$$

$$(Z'WZ)^{-1} = \begin{pmatrix} 12/10 & -3/10 & -3/10 \\ -3/10 & 12/10 & -3/10 \\ -3/10 & -3/10 & 12/10 \end{pmatrix}$$

$$I - jA^{-1}j'\left(Z'WZ\right)^{-1} = \begin{pmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{pmatrix}$$

$$\left(Z'WZ\right)^{-1}\left(I - jA^{-1}j'\left(Z'WZ\right)^{-1}\right) = \begin{pmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1 & -1/2 \\ -1/2 & -1/2 & 1 \end{pmatrix}$$

$$Z'W = \begin{pmatrix} 1/3 & 0 & 0 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 0 & 1/3 & 0 & 1/3 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \end{pmatrix}$$

$$\left(Z'WZ\right)^{-1}\left(I - jA^{-1}j'\left(Z'WZ\right)^{-1}\right)Z'W = \begin{pmatrix} 1/3 & -1/6 & -1/6 \\ -1/6 & 1/3 & -1/6 \\ -1/6 & -1/6 & 1/3 \\ -1/3 & 1/6 & 1/6 \\ 1/6 & -1/3 & 1/6 \\ 1/6 & 1/6 & -1/3 \end{pmatrix}'$$

$$\left(Z'WZ\right)^{-1}jA^{-1} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}$$

$$y = \begin{pmatrix} f\left(t_1, b_2, b_3\right) \\ f\left(b_1, t_2, b_3\right) \\ f\left(b_1, b_2, t_3\right) \\ f\left(t_1, t_2, b_3\right) \\ f\left(t_1, b_2, t_3\right) \\ f\left(b_1, t_2, t_3\right) \end{pmatrix}$$

$$y_b = f(b_1, b_2, b_3)$$

$$y_t = f(t_1, t_2, t_3)$$

If we conduct the same concatenation as described above, we get

$$y_* = \begin{pmatrix} f\left(b_1, b_2, b_3\right) \\ f\left(t_1, b_2, b_3\right) \\ f\left(b_1, t_2, b_3\right) \\ f\left(b_1, b_2, t_3\right) \\ f\left(t_1, t_2, b_3\right) \\ f\left(t_1, b_2, t_3\right) \\ f\left(b_1, t_2, t_3\right) \\ f\left(t_1, t_2, t_3\right) \end{pmatrix}$$

$$B_* = \begin{pmatrix} -1/3 & 1/3 & -1/6 & -1/6 & 1/6 & 1/6 & -1/3 & 1/3 \\ -1/3 & -1/6 & 1/3 & -1/6 & 1/6 & -1/3 & 1/6 & 1/3 \\ -1/3 & -1/6 & -1/6 & 1/3 & -1/3 & 1/6 & 1/6 & 1/3 \end{pmatrix}$$

And thus we find that

$$
\begin{aligned}
\phi &= \hat{\beta} = B_* y_* \\
\phi_1 &= \frac{1}{3}\left[f\left(t_1, b_2, b_3\right) - f\left(b_1, b_2, b_3\right)\right] + \frac{1}{6}\left[f\left(t_1, t_2, b_3\right) - f\left(b_1, t_2, b_3\right)\right] + \\
&\quad \frac{1}{6}\left[f\left(t_1, b_2, t_3\right) - f\left(b_1, b_2, t_3\right)\right] + \frac{1}{3}\left[f\left(t_1, t_2, t_3\right) - f\left(b_1, t_2, t_3\right)\right] \\
\phi_2 &= \frac{1}{3}\left[f\left(b_1, t_2, b_3\right) - f\left(b_1, b_2, b_3\right)\right] + \frac{1}{6}\left[f\left(t_1, t_2, b_3\right) - f\left(t_1, b_2, b_3\right)\right] + \\
&\quad \frac{1}{6}\left[f\left(b_1, t_2, t_3\right) - f\left(b_1, b_2, t_3\right)\right] + \frac{1}{3}\left[f\left(t_1, t_2, t_3\right) - f\left(t_1, b_2, t_3\right)\right] \\
\phi_3 &= \frac{1}{3}\left[f\left(b_1, b_2, t_3\right) - f\left(b_1, b_2, b_3\right)\right] + \frac{1}{6}\left[f\left(t_1, b_2, t_3\right) - f\left(t_1, b_2, b_3\right)\right] + \\
&\quad \frac{1}{6}\left[f\left(b_1, t_2, t_3\right) - f\left(b_1, t_2, b_3\right)\right] + \frac{1}{3}\left[f\left(t_1, t_2, t_3\right) - f\left(t_1, t_2, b_3\right)\right]
\end{aligned}
$$

and the Shapley values are the same as those calculated using the Shapley formula in Section 5.2.1. Thus, we see that the regression approximation yields exact Shapley values when the number of samples is exactly equal to 2^p.