NORTHWESTERN UNIVERSITY

Essays in Econometrics

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Economics

By

Max Tabord-Meehan

EVANSTON, ILLINOIS

June 2019

Abstract

Essays in Econometrics

Max Tabord-Meehan

This dissertation studies three distinct problems in econometrics. Chapter 1 proposes an adaptive randomization procedure for two-stage randomized controlled trials. The method uses data from a first-wave experiment in order to determine how to stratify in a second wave of the experiment, where the objective is to minimize the variance of an estimator for the average treatment effect (ATE). I consider selection from a class of stratified randomization procedures called *stratification trees*: these are procedures whose strata can be represented as decision trees, with differing treatment assignment probabilities across strata. By using the first wave to estimate a stratification tree, the method simultaneously selects which covariates to use for stratification, how to stratify over these covariates, as well as the assignment probabilities within these strata. My main result shows that using this randomization procedure with an appropriate estimator results in an asymptotic variance which minimizes the variance bound for estimating the ATE, over an optimal stratification of the covariate space. Moreover, by extending techniques developed in Bugni et al. (2018), the results presented are able to accommodate a large class of assignment mechanisms within strata, including stratified block randomization. I also present extensions of the procedure to the setting of multiple treatments, and to the targeting of subgroup-specific effects. In a simulation study,

I find that the method is most effective when the response model exhibits some amount of "sparsity" with respect to the covariates, but can be effective in other contexts as well, as long as the first-wave sample size used to estimate the stratification tree is not prohibitively small. The chapter concludes by applying the method to the study in Karlan and Wood (2017), where I estimate stratification trees using the first wave of their experiment.

Chapter 2 (which is joint work with Eric Mbakop) studies a new statistical decision rule for the treatment assignment problem. Consider a utilitarian policy maker who must use sample data to allocate one of two treatments to members of a population, based on their observable characteristics. In practice, it is often the case that policy makers do not have full discretion on how these covariates can be used, for legal, ethical or political reasons. We treat this constrained problem as a statistical decision problem, where we evaluate the performance of decision rules by their maximum regret. We focus on settings in which the policy maker may want to select amongst a *collection* of such constrained classes: examples we consider include choosing the number of covariates over which to perform best-subset selection, and model selection when approximating a complicated class via a sieve. We adapt and extend results from statistical learning to develop a decision rule which we call the Penalized Welfare Maximization (PWM) rule. We establish an oracle inequality for the regret of the PWM rule which shows that it is able to perform model selection over the collection of available classes. We then use this oracle inequality to derive relevant bounds on maximum regret for PWM. We illustrate the model-selection capabilities of our method with a small simulation exercise, and conclude by applying our rule to data from the Job Training Partnership Act (JTPA) study.

Chapter 3 studies inference in the linear model with dyadic data. Dyadic data are indexed by pairs of "units", for example trade data between pairs of countries. Because of

the potential for observations with a unit in common to be correlated, standard inference procedures may not perform as expected. I establish a range of conditions under which a t-statistic with the dyadic-robust variance estimator of Fafchamps and Gubert (2007) is asymptotically normal. Using these theoretical results as a guide, I perform a simulation exercise to study the validity of the normal approximation, as well as the performance of a novel finite-sample correction. The chapter concludes with guidelines for applied researchers wishing to use the dyadic-robust estimator for inference

Acknowledgements

This dissertation would not have been possible without the help of my advisors, family, peers, and friends. First, I would like to thank my advisor Ivan Canay, whose guidance and encouragement were pivotal in my development as a researcher. I would also like to thank my committee members Chuck Manski and Joel Horowitz, whose advice and wisdom greatly influenced my dissertation.

I am also grateful for the encouragement and support provided by my family, friends, and classmates. In particular, I would like to thank Susan Ou for her advice and tireless support.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Stratification Trees for Adaptive Randomization in Randomized Controlled Trials

1.1. Introduction to Chapter 1

This chapter proposes an adaptive randomization procedure for two-stage randomized controlled trials (RCTs). The method uses data from a first-wave experiment in order to determine how to stratify in a second wave of the experiment, where the objective is to minimize the variance of an estimator for the average treatment effect (ATE). We consider selection from a class of stratified randomization procedures which we call stratification trees: these are procedures whose strata can be represented as decision trees, with differing treatment assignment probabilities across strata.

Stratified randomization is ubiquitous in randomized experiments. In stratified randomization, the space of available covariates is partitioned into finitely many categories (i.e. strata), and randomization to treatment is performed independently across strata. Stratification has the ability to decrease the variance of estimators for the ATE through two parallel channels. The first channel is from ruling out treatment assignments which are potentially uninformative for estimating the ATE. For example, if we have information on the sex of individuals in our sample, and outcomes are correlated with sex, then performing stratified randomization over this characteristic can reduce variance (we present an example of this for the standard difference-in-means estimator in Appendix A.3.1). The second channel through which stratification can decrease variance is by allowing for differential treatment

assignment probabilities across strata. For example, if we again consider the setting where we have information on sex, then it could be the case that for males the outcome under one treatment varies much more than under the other treatment. As we show in Section 1.3.2, this can be exploited to reduce variance by assigning treatment according to the *Neyman Allocation*, which in this example would assign more males to the more variable treatment. Our proposed method leverages supervised machine-learning techniques to exploit both of these channels, by simultaneously selecting *which* covariates to use for stratification, *how* to stratify over these covariates, as well as the optimal assignment probabilities within these strata, in order to minimize the variance of an estimator for the ATE.

Our main result shows that using our procedure results in an estimator whose asymptotic variance minimizes the semi-parametric efficiency bound of Hahn (1998), over an "optimal" stratification of the covariate space, where we restrict ourselves to stratification in a class of decision trees. A decision tree partitions the covariate space such that the resulting partition can be interpreted through a series of yes or no questions (see Section 1.2.2 for a formal definition and some examples). We focus on strata formed by decision trees for several reasons. First, since the resulting partition can be represented as a series of yes or no questions, it is easy to communicate and interpret, even with many covariates. This feature could be particularly important in many economic applications, because many RCTs in economics are undertaken in partnership with external organizations (for example, every RCT described in (Karlan and Appel, 2016) was undertaken in this way), and thus clear communication of the experimental design could be crucial. Second, using partitions based on decision trees gives us theoretical and computational tractability. Third, as we will explain below, using decision trees allows us to flexibly address the additional goal of minimizing the variance of estimators for subgroup-specific effects. Lastly, decision trees naturally encompass the type

of stratifications usually implemented by practitioners. The use of decision trees in statistics and machine learning goes back at least to the work of Breiman (see Breiman et al., 1984; Györfi et al., 1996, for classical textbook treatments), and has seen a recent resurgence in econometrics (examples include Athey and Imbens, 2016; Athey and Wager, 2017).

An important feature of our theoretical results is that we allow for the possibility of so-called restricted randomization procedures *within* strata. Restricted randomization procedures limit the set of potential treatment allocations, in order to force the true treatment assignment proportions to be close to the desired target proportions (common examples used in a variety of fields include Antognini and Giovagnoli, 2004; Efron, 1971; Kuznetsova and Tymofyeyev, 2011; Wei, 1978; Zelen, 1974). Restricted randomization induces dependence in the assignments within strata, which complicates the analysis of our procedure. By extending techniques recently developed in Bugni et al. (2018), our results will accommodate a large class of restricted randomization schemes, including stratified block randomization, which as we discuss in Example 1.2.5 is a popular method of randomization.

Stratified randomization has additional practical benefits beyond reducing the variance of ATE estimators. For example, when a researcher wants to analyze subgroup-specific effects, stratifying on these subgroups serves as a form of pre-analysis registration, and as we will show, can help reduce the variance of estimators for the subgroup-specific ATEs. It is also straightforward to implement stratified randomization with multiple treatments. Although our main set of results apply to estimation of the global ATE in a binary treatment setting, we also present results that apply to settings with multiple treatments, as well as results for targeting subgroup-specific treatment effects.

The literature on randomization in RCTs is vast (references in (Athey and Imbens, 2017), (Cox and Reid, 2000), (Glennerster and Takavarasha, 2013), (Pukelsheim, 2006),

(Rosenberger and Lachin, 2015), and from a Bayesian perspective, (Ryan et al., 2016), provide an overview). The classical literature on optimal randomization, going back to the work of Smith (1918), maintains a parametric relationship for the outcomes with respect to the covariates, and targets efficient estimation of the model parameters. In contrast, this chapter follows a recent literature which instead maintains a non-parametric model of potential outcomes, and targets efficient estimation of treatment effects (see Remark 1.2.2 for a discussion about alternative objectives, in particular maximizing population welfare). This recent literature can be broadly divided into "one-stage" procedures, which do not use previous experiments to determine how to randomize (examples include Aufenanger, 2017; Barrios, 2014; Kallus, 2018; Kasy, 2016), and "multi-stage" procedures, of which our method is an example. Multi-stage procedures use the response information from previous experimental waves to determine how to randomize in subsequent waves of the experiment. We will call these procedures *response-adaptive*. Although response adaptive methods require information from a prior experiment, such settings do arise in economic applications. First, many social experiments have a pilot phase or multi-stage structure. For example, Simester et al. (2006), Karlan and Zinman (2008), and Karlan and Wood (2017) all feature a multi-stage structure, and Karlan and Appel (2016) advocate the use of pilot experiments to help avoid potential implementation failures when scaling up to the main study. Second, many research areas have seen a profusion of related experiments which could be used as a first wave of data in a response-adaptive procedure (see for example the discussion in the introduction of Hahn et al., 2011). The study of response-adaptive methods to inform many aspects of experimental design, including how to randomize, has a long history in the literature on clinical trials, both from a frequentist and Bayesian perspective (see for example

the references in Cheng et al., 2003; Hu and Rosenberger, 2006; Sverdlov, 2015), as well as in the literature on bandit problems (see Bubeck et al., 2012).

Two papers which propose response-adaptive randomization methods in a framework similar to ours are Hahn et al. (2011) and Chambaz et al. (2014). Hahn et al. (2011) develop a procedure which uses the information from a first-wave experiment in order to compute the propensity-score that minimizes the asymptotic variance of an ATE estimator, over a *discrete* set of covariates (i.e. they stratify the covariate space ex-ante). They then use the resulting propensity score to assign treatment in a second-wave experiment. In contrast, our method computes the optimal assignment proportions over a data-driven discretization of the covariate space. Chambaz et al. (2014) propose a multi-stage procedure which uses data from previous experimental waves to compute the propensity score that minimizes the asymptotic variance of an ATE estimator, where the propensity score is constrained to lie in a class of functions with appropriate entropy restrictions. However, their method requires the selection of several tuning parameters as well as additional regularity conditions, and their optimal target depends on these features in a way that may be hard to assess in practice. Their results are also derived in a framework where the number of experimental waves goes to infinity, which may not be a useful asymptotic framework for many settings encountered in economics. Finally, the results in both Hahn et al. (2011) and Chambaz et al. (2014) assume that assignment was performed completely independently across individuals. In contrast, we reiterate that our results will accommodate a large class of stratified randomization schemes.

The chapter proceeds as follows: In Section 1.2, we provide a motivating discussion, an overview of the procedure, and the formal definition of a stratification tree. In Section 1.3, we present the formal results underlying the method as well as several relevant extensions. In Section 1.4, we perform a simulation study to assess the performance of our method in finite

samples. In Section 1.5, we consider an application to the study in Karlan and Wood (2017), where we estimate stratification trees using the first wave of their experiment. Section 1.6 concludes.

1.2. Preliminaries

In this section we discuss some preliminary concepts and definitions. Section 1.2.1 presents a series of simplified examples which we use to motivate our procedure. Section 1.2.2 establishes the notation and provides the definition of a *stratification tree*, which is a central concept of the paper. Section 1.2.3 presents a high-level discussion of the proposed method.

1.2.1. Motivating Discussion

We present a series of simplified examples which we use to motivate our proposed method. First we study the problem of optimal experimental assignment without covariates. We work in a standard potential outcomes framework: let $(Y(1), Y(0))$ be potential outcomes for a binary treatment $A \in \{0, 1\}$, and let the observed outcome Y for an individual be defined as

$$Y = Y(1)A + Y(0)(1 - A) \ .$$

Let

$$E[Y(a)] = \mu_a, Var(Y(a)) = \sigma_a^2 \ ,$$

for $a \in \{0, 1\}$. Our quantity of interest is the average treatment effect

$$\theta := \mu_1 - \mu_0 \ .$$

Suppose we perform an experiment to obtain a size n sample $\{(Y_i, A_i)\}_{i=1}^n$, where the sampling process is determined by $\{(Y_i(1), Y_i(0))\}_{i=1}^n$, which are i.i.d, and the treatment assignments $\{A_i\}_{i=1}^n$, where exactly $n_1 := \lfloor n\pi \rfloor$ individuals are *randomly* assigned to treatment $A = 1$, for some $\pi \in (0, 1)$ (however, we emphasize that our results will accommodate other methods of randomization). Given this sample, consider estimation of θ through the standard difference-in-means estimator:

$$\hat{\theta}_S := \frac{1}{n_1} \sum_{i=1}^{n} Y_i A_i - \frac{1}{n - n_1} \sum_{i=1}^{n} Y_i (1 - A_i) \ .$$

It can then be shown that

$$\sqrt{n}(\hat{\theta}_S - \theta) \xrightarrow{d} N(\theta, V_1) \ ,$$

where

$$V_1 := \frac{\sigma_1^2}{\pi} + \frac{\sigma_0^2}{1 - \pi} \ .$$

In fact, it can be shown that under this randomization scheme V_1 is the finite sample variance of the normalized estimator, but this will not necessarily be true for other randomization schemes. Our goal is to choose π to minimize the variance of $\hat{\theta}$. Solving this optimization problem yields the following solution:

$$\pi^* := \frac{\sigma_1}{\sigma_1 + \sigma_0} \ .$$

This allocation is known as the *Neyman Allocation*, which assigns more individuals to the treatment which is more variable. Note that when $\sigma_0^2 = \sigma_1^2$, so that the variances of the potential outcomes are equal, the optimal proportion is $\pi^* = 0.5$, which corresponds to the standard "balanced" treatment allocation. In general, implementing π^* is infeasible without knowledge of σ_0^2 and σ_1^2. In light of this, if we had prior data $\{(Y_j, A_j)\}_{j=1}^m$ (either from a first-wave or a similar prior study), then we could use this data to estimate π^*, and then use

this estimate to assign treatment in a subsequent wave of the study. The idea of sequentially updating estimates of unknown population quantities using past observations, in order to inform experimental design in subsequent stages, underlies many procedures developed in the literatures on response adaptive experiments and bandit problems, and is the main idea underpinning our proposed method.

Remark 1.2.1. Although the Neyman Allocation minimizes the variance of the difference-in-means estimator, it is entirely agnostic on the welfare of the individuals in the experiment itself. In particular, the Neyman Allocation could assign the majority of individuals in the experiment to the inferior treatment if that treatment has a much larger variance in outcomes (see Hu and Rosenberger 2006 for relevant literature in the context of clinical trials, as well as Narita (2018) for recent work on this issue in econometrics). While this feature of the Neyman Allocation may introduce ethical or logistical issues in some relevant applications, in this paper we focus exclusively on the problem of estimating the ATE as accurately as possible. See Remark 1.2.2 for further discussion on our choice of optimality criterion. ■

Next we repeat the above exercise with the addition of a discrete covariate $S \in \{1, 2, ..., K\}$ over which we stratify. We perform an experiment which produces a sample $\{(Y_i, A_i, S_i)\}_{i=1}^n$, where the sampling process is determined by i.i.d draws $\{(Y_i(1), Y_i(0), S_i)\}_{i=1}^n$ and the treatment assignments $\{A_i\}_{i=1}^n$. For this example suppose that the $\{A_i\}_{i=1}^n$ are generated as follows: for each k, exactly $n_1(k) := \lfloor n(k)\pi(k) \rfloor$ individuals are randomly assigned to treatment $A = 1$, with $n(k) := \sum_{i=1}^n \mathbf{1}\{S_i = k\}$.

Note that when the assignment proportions $\pi(k)$ are not equal across strata, the difference-in-means estimator $\hat{\theta}_S$ is no longer consistent for θ. Hence we consider the following weighted

estimator of θ:

$$\hat{\theta}_C := \sum_k \frac{n(k)}{n} \hat{\theta}(k) \ ,$$

where $\hat{\theta}(k)$ is the difference-in-means estimator for $S = k$:

$$\hat{\theta}(k) := \frac{1}{n_1(k)} \sum_{i=1}^{n} Y_i A_i \mathbf{1}\{S_i = k\} - \frac{1}{n(k) - n_1(k)} \sum_{i=1}^{n} Y_i (1 - A_i) \mathbf{1}\{S_i = k\} \ .$$

In words, $\hat{\theta}_C$ is obtained by computing the difference in means for each k and then taking a weighted average over each of these estimates. Note that when $K = 1$ (i.e. when S can take on one value), this estimator simplifies to the difference-in-means estimator. It can be shown under appropriate conditions that

$$\sqrt{n}(\hat{\theta}_C - \theta) \xrightarrow{d} N(0, V_2) \ ,$$

where

$$V_2 := \sum_{k=1}^{K} P(S = k) \left[\left(\frac{\sigma_0^2(k)}{1 - \pi(k)} + \frac{\sigma_1^2(k)}{\pi(k)} \right) + \left(E[Y(1) - Y(0) | S = k] - E[Y(1) - Y(0)] \right)^2 \right] \ ,$$

with $\sigma_d^2(k) = E[Y(d)^2 | S = k] - E[Y(d) | S = k]^2$. The first term in V_2 is the weighted average of the conditional variances of the difference in means estimator for each $S = k$. The second term in V_2 arises due to the additional variability in sample sizes for each $S = k$. We note that this variance is the semi-parametric efficiency bound derived by Hahn (1998) for estimators of the ATE which use the covariate S. Following a similar logic to what was proposed above without covariates, we could use first-wave data $\{(Y_j, A_j, S_j)\}_{j=1}^{m}$ to form a sample analog of V_2, and choose $\{\pi^*(k)\}_{k=1}^{K}$ to minimize this quantity.

Now we introduce the setting that we consider in this paper: suppose we observe covariates $X \in \mathcal{X} \subset \mathbb{R}^d$, so that our covariate space is now multi-dimensional with potentially

continuous components. How could we practically extend the logic of the previous examples to this setting? A natural solution is to *discretize* (i.e. *stratify*) $\mathcal{X}$ into K categories (strata), by specifying a mapping $S : \mathcal{X} \to \{1, 2, 3, ..., K\}$, with $S_i := S(X_i)$, and then proceed as in the above example. As we argued in the introduction, stratified randomization is a popular technique in practice, and possesses several attractive theoretical and practical properties. In this paper we propose a method which uses first-wave data to estimate (1) the optimal stratification, and (2) the optimal assignment proportions within these strata. In other words, given first-wave data $\{(Y_j, A_j, X_j)\}_{j=1}^m$ from a randomized experiment, where $X \in \mathcal{X} \subset \mathbb{R}^d$, we propose a method which selects $\{\pi(k)\}_{k=1}^K$ *and* the function $S(\cdot)$, in order to minimize the variance bound in Hahn (1998). In particular, our proposed solution selects a randomization procedure amongst the class of what we call *stratification trees*, which we introduce in the next section.

Remark 1.2.2. Our focus on the minimization of asymptotic variance is in line with standard asymptotic optimality results for regular estimators (see for example Theorems 25.20 and 25.21 in Van der Vaart, 1998). However, accurate estimation of the ATE is not the only objective one could consider when designing an RCT. In particular, we could instead consider using an ATE estimator to construct a statistical decision rule, with the goal of maximizing population welfare (see Manski 2009 for a textbook discussion). If, as in Manski (2004), we evaluate decision rules by their maximum regret, then our optimality objective would be to design the randomization procedure in order to minimize the maximum regret of the decision rule. We remark that selecting a randomization procedure to minimize asymptotic variance may in fact reduce *pointwise* regret, when paired with an appropriate decision rule. In particular, Athey and Wager (2017) derive a bound on regret whose constant scales with the semi-parametrically efficient variance. Our method selects a randomization

procedure which minimizes this variance, and hence subsequently minimizes the constant in this bound. ■

1.2.2. Notation and Definitions

In this section we establish the notation of the paper and define the class of randomization procedures that we will consider. Let $A_i \in \{0,1\}$ be a binary variable which denotes the treatment received by a unit i (we consider the extension to multiple treatments in Section 1.3.2), and let Y_i denote the observed outcome. Let $Y_i(1)$ denote the potential outcome of unit i under treatment 1 and let $Y_i(0)$ denote the potential outcome of unit i under treatment 0. The observed experimental outcome for each unit is related to their potential outcomes through the expression:

$$Y_i = Y_i(1)A_i + Y_i(0)(1 - A_i) \ .$$

Let $X_i \in \mathcal{X} \subset \mathbb{R}^d$ denote a vector of observed pre-treatment covariates for unit i. Let Q denote the distribution of $(Y_i(1), Y_i(0), X_i)$ and assume that $\{(Y_i(1), Y_i(0), X_i)\}_{i=1}^n$ consists of n i.i.d observations from Q. We restrict Q as follows:

Assumption 1.2.1. *Q satisfies the following properties:*

- *$Y(a) \in [-M, M]$ for some $M < \infty$, for $a \in \{0,1\}$, where the marginal distributions $Y(1)$ and $Y(0)$ are either continuous or discrete with finite support.*
- *$X \in \mathcal{X} = \times_{j=1}^d [b_j, c_j]$, for some $\{b_j, c_j\}_{j=1}^d$ finite.*
- *$X = (X_C, X_D)$, where $X_C \in \mathbb{R}^{d_1}$ for some $d_1 \in \{0, 1, 2, ..., d\}$ is continuously distributed with a bounded, strictly positive density. $X_D \in \mathbb{R}^{d-d_1}$ is discretely distributed with finite support.*

Remark 1.2.3. The restriction that the $Y(a)$ are bounded is used several times throughout the proofs for technical convenience, but it is possible that this assumption could be

weakened. In applications it may be the case that X_C as defined above may not be continuous on $\times_j [b_j, c_j]$, but is instead censored at its endpoints; see for example the application considered in Section 1.5. Our results will continue to hold in this case as well. ■

Our quantity of interest is the average treatment effect (ATE) given by:

$$\theta = E[Y_i(1) - Y_i(0)] \ .$$

An experiment on our sample produces the following data:

$$\{W_i\}_{i=1}^n := \{(Y_i, A_i, X_i)\}_{i=1}^n \ ,$$

whose joint distribution is determined by Q, the potential outcomes expression, and the *randomization procedure*. We focus on the class of stratified randomization procedures: these randomization procedures first stratify according to baseline covariates and then assign treatment status independently across each of these strata. Moreover, we attempt to make minimal assumptions on how randomization is performed *within* strata, in particular we do *not* require the treatment assignment within each stratum to be independent across observations.

We will now describe the structure we impose on the class of possible strata we consider. For L a positive integer, let $K = 2^L$ and let $[K] := \{1, 2, ..., K\}$. Consider a function $S : \mathcal{X} \to [K]$, then $\{S^{-1}(k)\}_{k=1}^K$ forms a partition of $\mathcal{X}$ with K strata. For a given positive integer L, we work in the class $S(\cdot) \in \mathcal{S}_L$ of functions whose partitions form *tree partitions* of depth L on $\mathcal{X}$, which we now define. Note that the definition is recursive, so we begin with the definition for a tree partition of depth one:

Definition 1.2.1. *Let* $\Gamma_j \subset [b_j, c_j]$*, let* $\Gamma = \times_{j=1}^{d} \Gamma_j$*, and let* $x = (x_1, x_2, ..., x_d) \in \Gamma$*. A tree partition of depth one on* Γ *is a partition of* Γ *which can be written as*

$$\Gamma_D(j, \gamma) \cup \Gamma_U(j, \gamma) \ ,$$

where

$$\Gamma_D(j, \gamma) := \{x \in \Gamma : x_j \leqslant \gamma\} \ ,$$

$$\Gamma_U(j, \gamma) := \{x \in \Gamma : x_j > \gamma\} \ ,$$

for some $j \in [d]$ *and* $\gamma \in \Gamma_j$*. We call* $\Gamma_D(j, \gamma)$ *and* $\Gamma_U(j, \gamma)$ *leaves (or sometimes terminal nodes), whenever these are nonempty.*

Example 1.2.1. Figure 1.1 presents two different representations of a tree partition of depth one on $[0, 1]^2$. The first representation we call *graphical*: it depicts the partition on a square drawn in the plane. The second depiction we call a *tree representation*: it illustrates how to describe a depth one tree partition as a yes or no question. In this case, the question is "is x_1 less than or greater than 0.5?".

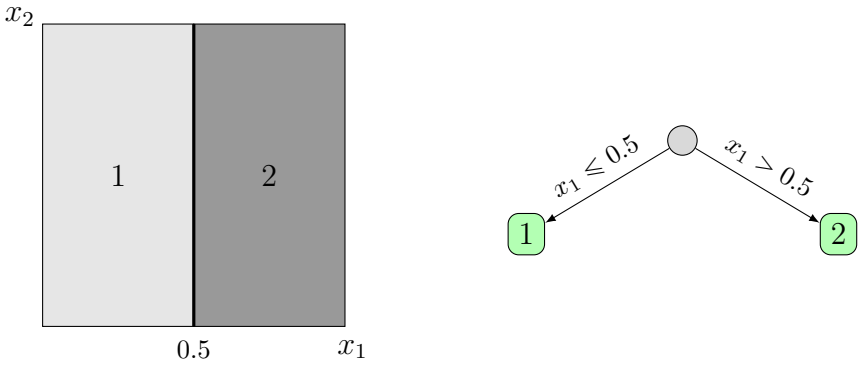

Figure 1.1. Two representations of a tree partition of depth 1 on $[0, 1]^2$. Graphical representation (left), tree representation (right).

Next we define a tree partition of depth $L > 1$ recursively:

Definition 1.2.2. *A tree partition of depth $L > 1$ on $\Gamma = \times_{j=1}^{d} \Gamma_j$ is a partition of Γ which can be written as $\Gamma_D^{(L-1)} \cup \Gamma_U^{(L-1)}$, where*

$$\Gamma_D^{(L-1)} \text{ is a tree partition of depth } L-1 \text{ on } \Gamma_D(j,\gamma) \ ,$$

$$\Gamma_U^{(L-1)} \text{ is a tree partition of depth } L-1 \text{ on } \Gamma_U(j,\gamma) \ ,$$

for some $j \in [d]$ and $\gamma \in \Gamma_j$. We call $\Gamma_D^{(L-1)}$ and $\Gamma_U^{(L-1)}$ left and right subtrees, respectively, whenever these are nonempty.

Example 1.2.2. Figure 1.2 depicts two representations of a tree partition of depth two on $[0,1]^2$.

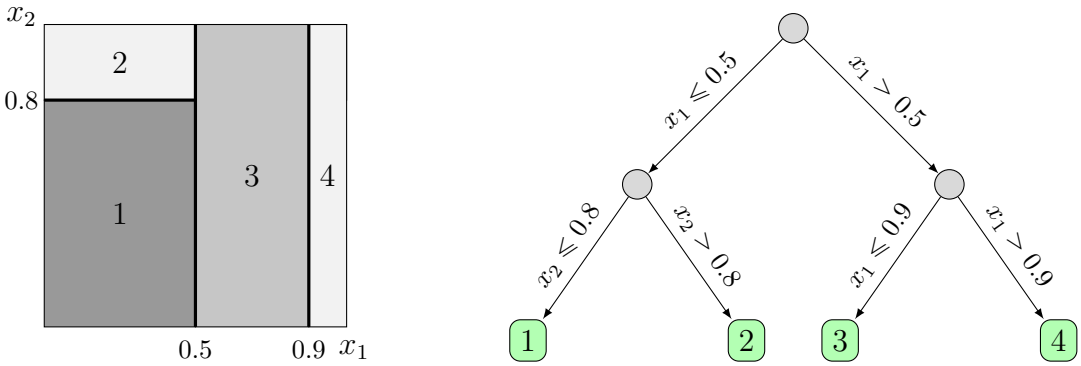

Figure 1.2. Two representations of a tree partition of depth 2 on $[0,1]^2$. Graphical representation (left), tree representation (right).

We focus on strata that form tree partitions for several reasons. First, these types of strata are easy to represent and interpret, even in higher dimensions, via their tree representations or as a series of yes or no questions. We argued in the introduction that this could be of particular importance in economic applications. Second, as we explain in Remark 1.3.4 and the Appendix, restricting ourselves to tree partitions gives us theoretical and computational

tractability. In particular, computing an optimal stratification is a difficult discrete optimization problem for which we exploit the tree structure to design an evolutionary algorithm. Third, the recursive aspect of tree partitions makes the targeting of subgroup-specific effects convenient, as we show in Section 1.3.2.

For each $k \in [K]$, we define $\pi := (\pi(k))_{k=1}^{K}$ to be the vector of target proportions of units assigned to treatment 1 in each stratum.

A *stratification tree* is a pair (S, π), where $S(\cdot)$ forms a tree partition, and π specifies the target proportions in each stratum. We denote the set of stratification trees of depth L as $\mathcal{T}_L$.

Remark 1.2.4. To be precise, any element $T = (S, \pi) \in \mathcal{T}_L$ is equivalent to another element $T' = (S', \pi') \in \mathcal{T}_L$ whenever T' can be realized as a re-labeling of T. For instance, if we consider Example 1.2.1 with the labels 1 and 2 reversed, the resulting tree is identical to the original except for this re-labeling. $\mathcal{T}_L$ should be understood as the quotient set that results from this equivalence. ■

Example 1.2.3. Figure 1.3 depicts a representation of a stratification tree of depth two. Note that the terminal nodes of the tree have been replaced with labels that specify the target proportions in each stratum.

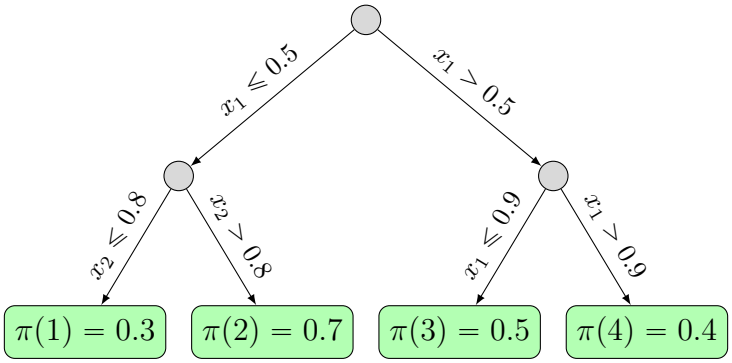

Figure 1.3. Representation of a Stratification Tree of Depth 2

We further impose that the set of trees cannot have arbitrarily small (nonempty) cells, nor can they have arbitrarily extreme treatment assignment targets:

Assumption 1.2.2. *We constrain the set of stratification trees $T = (S, \pi) \in \mathcal{T}_L$ such that, for some fixed $\nu > 0$ and $\delta > 0$, $\pi(k) \in [\nu, 1 - \nu]$ and $P(S(X) = k) > \delta$ whenever $S^{-1}(k) \neq \varnothing$.*

Remark 1.2.5. In what follows, we adopt the following notational convention: if $S^{-1}(k) = \varnothing$, then $E[W|S(X) = k] = 0$ for any random variable W. ■

Remark 1.2.6. The depth L of the set of stratification trees will remain fixed but arbitrary throughout most of the analysis. We return to the question of how to choose L in Section 1.3.2. ■

For technical reasons, we will impose one additional restriction on $\mathcal{T}_L$. We emphasize that this assumption is *only* used to avoid issues which may arise from the potential non-measurability of certain objects.

Assumption 1.2.3. *Let $\mathcal{T}_L^\dagger \subset \mathcal{T}_L$ be a countable, closed subset of the set of stratification trees*[1]*. We then consider the set of stratification trees restricted to this subset.*

Remark 1.2.7. A restriction similar to Assumption 1.2.3 was recently considered in Kitagawa and Tetenov (2018) in order to avoid measurability issues. Note that, in practice, restricting the set of stratification trees to a finite grid satisfies Assumption 1.2.3. However, our results also apply much more generally. ■

[1]Here "closed" is with respect to an appropriate topology on $\mathcal{T}_L$, see Appendix A.2 for details. It is possible that Assumption 1.2.3 could be eliminated by using the theory of weak convergence developed by Hoffman-Jorgensen, see Van der Vaart and Wellner (1996) for a textbook discussion.

Recall that we are interested in randomization procedures that stratify on baseline covariates and then assign treatment status independently across strata. For $T = (S, \pi)$, let $S_i := S(X_i)$ be the strata label for an individual i. For each $T \in \mathcal{T}_L$, and given sample of size n, an experimental assignment is described by a random vector $A^{(n)}(T) := (A_i(T))_{i=1}^n$ for each $T \in \mathcal{T}_L$. For our purposes a *randomization procedure* (or randomization scheme) is a family of such random vectors $A^{(n)}(T)$ for each $T = (S, \pi) \in \mathcal{T}_L$. The only assumptions that we require on the randomization procedure are that the assignments are exogenous conditional on the strata, and that the assignment proportions converge to the target proportions asymptotically. Assumptions 1.3.4 and 1.3.5 re-state these conditions formally. Examples 1.2.4 and 1.2.5 illustrate two such randomization schemes which are popular in economics, and many more schemes have been considered in the the literature on clinical trials: examples include Efron (1971), Wei (1978), Antognini and Giovagnoli (2004), and Kuznetsova and Tymofyeyev (2011).

Example 1.2.4. *Simple random assignment* assigns each individual within stratum k to treatment via a coin-flip with weight $\pi(k)$. Formally, for each T, $A^{(n)}(T)$ is a vector with independent components such that

$$P(A_i(T) = 1 | S_i = k) = \pi(k) \ .$$

Simple random assignment is theoretically convenient, and features prominently in papers on adaptive randomization. However, it is considered unattractive in practice because it results in a "noisy" assignment for a given target $\pi(k)$, and hence could be very far off the target assignment for any given random draw. Moreover, this extra noise increases the finite-sample variance of ATE estimators relative to other assignment procedures which target $\pi(k)$ more directly (see for example the discussion in Kasy, 2013).

Example 1.2.5. *Stratified block randomization* (SBR) assigns a fixed proportion $\pi(k)$ of individuals within stratum k to treatment 1. Formally, let $n(k)$ be the number of units in stratum k, and let $n_1(k)$ be the number of units assigned to treatment 1 in stratum k. In SBR, $n_1(k)$ is given by

$$n_1(k) = \lfloor n(k)\pi(k) \rfloor \ .$$

SBR proceeds by randomly assigning $n_1(k)$ units to treatment 1 for each k, where all

$$\binom{n(k)}{n_1(k)} \ ,$$

possible assignments are equally likely. This assignment procedure has the attractive feature that it targets the proportion $\pi(k)$ as directly as possible. An early discussion of SBR can be found in Zelen (1974). SBR has recently become a popular method of assignment in economics (for example, every RCT published in the Quarterly Journal of Economics in 2017 used SBR).

1.2.3. Overview of Procedure

In this section we provide an overview of our procedure, before stating the formal results in Section 1.3. Recall the setting from the end of Section 1.2.1: given first-wave data, our goal is to estimate a stratification tree which minimizes the asymptotic variance in a certain class of ATE estimators, which we now introduce. For a fixed $T \in \mathcal{T}_L$, consider estimation of the following equation by OLS:

$$Y_i = \sum_k \alpha(k)\mathbf{1}\{S_i = k\} + \sum_k \beta(k)\mathbf{1}\{A_i = 1, S_i = k\} + u_i \ .$$

Then our ATE estimator is given by

$$\hat{\theta}(T) = \sum_k \frac{n(k)}{n} \hat{\beta}(k) \ ,$$

where $n(k) = \sum_i \mathbf{1}\{S_i = k\}$. In words, this estimator takes the difference in means between treatments within each stratum, and then averages these over the strata. Given appropriate regularity conditions, the results in Bugni et al. (2018) imply the following result for a *fixed* $T = (S, \pi) \in \mathcal{T}_L$:

$$\sqrt{n}(\hat{\theta}(T) - \theta) \xrightarrow{d} N(0, V(T)) \ ,$$

where

$$V(T) = \sum_{k=1}^{K} P(S(X) = k) \Big[\left(E[Y(1) - Y(0)|S(X) = k] - E[Y(1) - Y(0)]\right)^2 + \left(\frac{\sigma_0^2(k)}{1 - \pi(k)} + \frac{\sigma_1^2(k)}{\pi(k)} \right) \Big] \ ,$$

and

$$\sigma_a^2(k) = E[Y(a)^2|S(X) = k] - E[Y(a)|S(X) = k]^2 \ .$$

Again we remark that this variance is the semi-parametric efficiency bound of Hahn (1998) amongst all (regular) estimators that use the strata indicators as covariates. We propose a two-stage adaptive randomization procedure which asymptotically achieves the minimal variance $V(T)$ across all $T \in \mathcal{T}_L$. In the first stage, we use first-wave data $\{(Y_j, A_j, X_j)\}_{j=1}^m$ to estimate some "optimal" tree $\widetilde{T}$ which is designed to minimize $V(T)$. More formally, what we require is that

$$|V(\widetilde{T}) - V^*| \xrightarrow{a.s} 0 \ ,$$

as $m \to \infty$, where V^* is the minimum of $V(T)$ in $\mathcal{T}_L$. We show in Proposition 1.3.1 that a straightforward way to construct such a $\widetilde{T}$ is to minimize an empirical analog of $V(T)$:

$$\widetilde{T}^{EM} \in \arg\min_{T \in \mathcal{T}_L} \widetilde{V}(T) \ ,$$

where $\widetilde{V}(\cdot)$ is an empirical analog of $V(\cdot)$ defined in Appendix A.4. In general, computing $\widetilde{T}^{EM}$ involves solving a complicated discrete optimization problem. In Appendix A.4, we describe an evolutionary algorithm that we use to solve this problem. In Section 1.3.2, we describe a version of this estimator that selects the appropriate depth L via cross-validation.

In the second stage, we perform a randomized experiment using stratified randomization with $A^{(n)}(\widetilde{T})$ to obtain second-wave data $\{(Y_i, A_i, X_i)\}_{i=1}^n$. Finally, to analyze the results of the experiment, we consider the use of two possible estimators. The first estimator we consider "pools" the first-wave and second-wave data together. To accomplish this, we stratify on the experimental waves; that is, we append an extra stratum which contains the first-wave data, indexed by $k = 0$, to $\widetilde{T}$. We call the resulting stratification tree an "augmented" tree, and denote it by $\hat{T}$, (see Example 1.2.6 for an illustration). We then use all of the available data when estimating the saturated regression. The resulting pooled estimator is denoted by $\hat{\theta}(\hat{T})$. The second estimator we consider uses only the second-wave data to estimate the ATE. We call this estimator the unpooled estimator and denote it by $\hat{\theta}(\widetilde{T})$. From now on, we state all of our results for the pooled estimator $\hat{\theta}(\hat{T})$, with the understanding that analogous results hold for the unpooled estimator as well (see Remark 1.3.1 for details).

Example 1.2.6. Figure 1.4 depicts a representation of an augmented tree. First the tree partitions the first-wave data into its own stratum indexed by $k = 0$, and then proceeds as before.

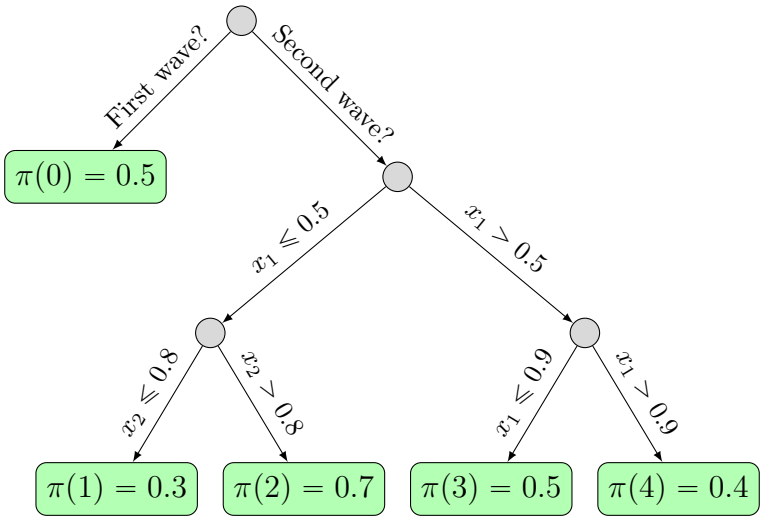

Figure 1.4. An Augmented Stratification Tree

Remark 1.2.8. In applications it may also be the case that the first-wave experiment was itself stratified. It would then be natural to incorporate this stratification into the specification of the augmented tree $\hat{T}$. Analogous results to what we derive in Section 1.3 will hold in this case as well. ■

From now on, to be concise, we will call data from the first-wave the *pilot* data, and data from the second-wave the *main* data. To summarize, the method proceeds as follows:

OUTLINE OF PROCEDURE

- Obtain pilot data $(Y_j, A_j, X_j)_{j=1}^m$.
- Use pilot data to construct $\widetilde{T}$ (either $\widetilde{T}^{EM}$ or the cross-validated version $\widetilde{T}^{CV}$ defined in Section 1.3.2).
- Perform a randomized experiment using $A^{(n)}(\widetilde{T})$ (as defined in Section 1.2.2) to obtain main data $(Y_i, A_i, X_i)_{i=1}^n$.

- Perform inference on the average treatment effect using $\hat{\theta}(\hat{T})$, where $\hat{T}$ is the augmented tree as described above.

In Section 1.3.1, we provide conditions under which

$$\sqrt{N}(\hat{\theta}(\hat{T}) - \theta) \xrightarrow{d} N(0, V^*) \ ,$$

where $N = m + n$, as $m, n \to \infty$. We also describe a consistent estimator of the asymptotic variance. In Section 1.3.2, we consider several extensions of the procedure: to multiple treatments, to the targeting of subgroup-specific effects, as well as to using cross-validation to select the depth L of the stratification tree.

Remark 1.2.9. It is common practice in the analysis of RCTs to estimate θ by running OLS on a linear regression with strata fixed effects:

$$Y_i = \beta A_i + \sum_k \delta(k) \mathbf{1}\{S_i = k\} + u_i \ .$$

If the assignment targets $\pi(k)$ are not equal across strata, as in this paper, then $\hat{\beta}$ is not a consistent estimator of θ. However, it can be shown that $\hat{\beta}$ is consistent when the assignment targets are equal across strata. Moreover, in the special case where assignment is conducted using a randomization procedure with "strong balance", such as SBR, this estimator has the same limiting distribution as $\hat{\theta}$ (see Bugni et al., 2018, for details). It can be shown that our results continue to hold with this alternative estimator, as long as the assignment proportions $\pi(k)$ are restricted to be equal, and SBR is used as the randomization procedure.

■

1.3. Results

In this section we derive the theoretical properties of our estimator. Section 1.3.1 presents the main result of the paper, that $\hat{\theta}(\hat{T})$ is asymptotically normal with minimal variance in $\mathcal{T}_L$, and describes a consistent estimator of its asymptotic variance. Section 1.3.2 presents several extensions: a cross-validation procedure to select the depth L of the stratification tree, as well as extensions for the targeting of subgroup specific effects and to multiple treatments.

1.3.1. Main Results

In this section we present the main theoretical properties of our method. In particular, we provide conditions under which $\hat{\theta}(\hat{T})$ is asymptotically normal with minimal variance in the class of estimators defined in Section 1.2.3, as well as provide a consistent estimator of its asymptotic variance. Recall that our goal is to use pilot data in order to estimate some "optimal" stratification tree $\widetilde{T}$, and then use this tree to perform the experimental assignment in a second wave of the experiment. To that end, we assume the existence of pilot data $\{W_i\}_{i=1}^m := \{(Y_i, X_i, A_i)\}_{i=1}^m$, generated from a randomized experiment performed on a sample from the same population as the main experiment, which we use to construct $\widetilde{T}$. Throughout the analysis of this section we consider the following asymptotic framework for the size of m (the size of the pilot) relative to the size of n (the size of the main study):

Assumption 1.3.1. *We consider the following asymptotic framework:*

$$\frac{m}{N} = o\left(\frac{1}{\sqrt{N}}\right) \ ,$$

where $N = m + n$, *as* $m, n \to \infty$.

Remark 1.3.1. Rate assumptions like Assumption 1.3.1 are only required to study the properties of the pooled estimator $\hat{\theta}(\hat{T})$. The properties of the unpooled estimator $\hat{\theta}(\widetilde{T})$ can be derived under the weaker assumption that $m \to \infty$ and $n \to \infty$ without any restrictions on their relative rates. In what follows, we state all of our results for the estimator $\hat{\theta}(\hat{T})$ only, with the understanding that analogous results will hold for $\hat{\theta}(\widetilde{T})$ under this weaker assumption. ■

Remark 1.3.2. The asymptotic framework introduced in Assumption 1.3.1 will ensure that the asymptotic variance of $\hat{\theta}(\hat{T})$ is not distorted. However, this asymptotic framework requires that m/N vanishes quite quickly, which may inaccurately reflect the finite sample behavior of our estimator in applications where the first wave of the experiment is large relative to the second: see for example the application considered in Section 1.5, where two waves of equal size were used. In Remark 1.3.5 we explain how our results would change in an asymptotic framework where we allow

$$\frac{m}{N} = \lambda + o\left(\frac{1}{\sqrt{N}}\right) \ ,$$

for $0 \leqslant \lambda \leqslant 1$. See Appendix A.3.2 or details. However, we emphasize here that this alternative framework does *not* change the mechanics of the procedure in any way. We also explore the effect of large pilot samples in the simulation study of Section 1.4. ■

In all of the results of this section, the depth L of the class of stratification trees is fixed and specified by the researcher. We return to the question of how to choose L in Section 1.3.2. Given a pilot sample $\{W_i\}_{i=1}^m$, we require the following high-level consistency property for our estimator $\widetilde{T}$:

Assumption 1.3.2. *The estimator $\widetilde{T}_m$ is a $\sigma\{(W_i)_{i=1}^m\}/\mathcal{B}(\mathcal{T}_L)$ measurable function of the pilot data*[2] *and satisfies*

$$|V(\widetilde{T}_m) - V^*| \xrightarrow{a.s} 0 \ ,$$

where

$$V^* = \inf_{T \in \mathcal{T}_L} V(T) \ ,$$

as $m \to \infty$.

Note that Assumption 1.3.2 does not imply that V^* is *uniquely* minimized at some $T \in \mathcal{T}_L$ and so we do not make any assumptions about whether or not $\widetilde{T}$ converges to any *fixed* tree. In Proposition 1.3.1, we show that a straightforward method to construct such a $\widetilde{T}$ is to solve the following empirical minimization problem:

$$\widetilde{T}^{EM} \in \arg\min_{T \in \mathcal{T}_L} \widetilde{V}(T) \ ,$$

where $\widetilde{V}(T)$ is an empirical analog of $V(T)$ (as defined in Appendix A.4) constructed using the pilot data. A nice feature of this choice of $\widetilde{T}$ is that it also corresponds to minimizing (an estimated version of) the *finite sample* variance of our estimator in the case of SBR. In Section 1.3.2, we consider an alternative construction of $\widetilde{T}$ which uses cross-validation to select the depth of the tree. We verify Assumption 1.3.2 for $\widetilde{T}^{EM}$ under the following assumption about the randomization procedure used in the pilot study (although we emphasize that this assumption is *not necessary* to establish such a result in general):

Assumption 1.3.3. *The pilot experiment was performed using simple random assignment (see Example 1.2.4).*

[2] $\mathcal{B}(\mathcal{T}_L)$ is the Borel-sigma algebra on $\mathcal{T}_L$ generated by an appropriate topology and $\sigma\{(W_i)_{i=1}^m\}$ is the sigma-algebra generated by the pilot data. See the appendix for details.

Proposition 1.3.1. *Let $\widetilde{T}^{EM}$ be a minimizer of the empirical variance. Under Assumptions 1.2.1, 1.2.2, 1.2.3, 1.3.1 and 1.3.3, Assumption 1.3.2 is satisfied.*

Next, we describe the assumptions we impose on the randomization procedure in the second-wave experiment. For $T = (S, \pi)$, let $S_i := S(X_i)$ and $S^{(n)} := (S_i)_{i=1}^n$ be the random vector of stratification labels of the observed data (note that, although $S(\cdot)$ is a deterministic function, X_i is a random variable and hence the resulting composition S_i is itself random). Let $p(k; T) := P(S_i = k)$ be the population proportions in each stratum. We require the following exogeneity assumption:

Assumption 1.3.4. *The randomization procedure is such that, for each $T = (S, \pi) \in \mathcal{T}_L$:*

$$\left[(Y_i(0), Y_i(1), X_i)_{i=1}^n \perp A^{(n)}(T)\right] \Big| S^{(n)} \ .$$

This assumption asserts that the randomization procedure can depend on the observables only through the strata labels.

We also require that the randomization procedure satisfy the following "consistency" property:

Assumption 1.3.5. *The randomization procedure is such that*

$$\sup_{T \in \mathcal{T}_L} \left| \frac{n_1(k; T)}{n} - \pi(k)p(k; T) \right| \xrightarrow{p} 0 \ ,$$

for each $k \in [K]$. Where

$$n_1(k; T) = \sum_{i=1}^{n} \mathbf{1}\{A_i(T) = 1, S_i = k\} \ .$$

This assumption asserts that the assignment procedure must approach the target proportion asymptotically, and do so in a uniform sense over all stratification trees in $\mathcal{T}_L$. Other

than Assumptions 1.3.4 and 1.3.5, we do not require any additional assumptions about how assignment is performed within strata. Bugni et al. (2018) make similar assumptions for a *fixed* stratification function and show that it is satisfied for a wide range of assignment procedures, including those introduced in Examples 1.2.4 and 1.2.5. In Proposition 1.3.2 below, we verify that Assumptions 1.3.4 and 1.3.5 hold for stratified block randomization, which is a common assignment procedure in economic applications.

Proposition 1.3.2. *Suppose randomization is performed through SBR (see Example 1.2.5), then Assumptions 1.3.4 and 1.3.5 are satisfied.*

Finally, we impose one additional regularity condition on the distribution Q when $(Y(0), Y(1))$ are continuous. We impose this assumption because of technical complications that arise from the fact that the set of minimizers of the population variance $V(T)$ is not necessarily a singleton:

Assumption 1.3.6. *Fix some a and k and suppose $Y(a)$ is continuous. Let $\mathcal{G}$ be the family of quantile functions of $Y(a)|S(X) = k$, for $S^{-1}(k)$ nonempty. Then we assume that $\mathcal{G}$ forms a pointwise equicontinuous family.*

Remark 1.3.3. To our knowledge this assumption is non-standard. In Lemma A.5.3 we show that a sufficient condition for Assumption 1.3.6 to hold is that the quantile functions be continuous (i.e. that the densities of $Y(a)|S(X) = k$ do not contain "gaps" in their support), and that the quantile functions vary "continuously" as we vary $S \in \mathcal{S}_L$. ■

We now state the main result of the paper: an optimality result for the pooled estimator $\hat{\theta}(\hat{T})$. In Remark 1.3.4 we comment on some of the technical challenges that arise in the proof of this result.

Theorem 1.3.1. *Given Assumptions 1.2.1, 1.2.2, 1.2.3, 1.3.1, 1.3.2, 1.3.4, 1.3.5, and 1.3.6, we have that*

$$\sqrt{N}(\hat{\theta}(\hat{T}) - \theta) \xrightarrow{d} N(0, V^*) \ ,$$

where $N = m + n$*, as* $m, n \to \infty$*.*

Remark 1.3.4. Here we comment on some of the technical challenges that arise in proving Theorem 3.3.1. First, we develop a theory of convergence for stratification trees by defining a novel metric on $\mathcal{S}_L$ based on the Frechet-Nikodym metric, and establish basic properties about the resulting metric space. In particular, we use this construction to show that a set of minimizers of $V(T)$ exists given our assumptions, and that $\widetilde{T}$ converges to this set of minimizers in an appropriate sense. For these results we exploit the properties of tree partitions for two purposes: First, we frequently exploit the fact that for a fixed index $k \in [K]$, the class of sets $\{S^{(-1)}(k) : S \in \mathcal{S}_L\}$ consists of rectangles, and hence forms a VC class. Second, as explained in Remark 1.2.4, every $T \in \mathcal{T}_L$ is in fact an equivalence class. Using the structure of tree partitions, we define a canonical representative of T (see Definition A.2.1) which simplifies our derivations.

Next, because Assumptions 1.3.4 and 1.3.5 impose so little on the dependence structure of the randomization procedure, standard central limit theorems cannot be applied. When the stratification is fixed, Bugni et al. (2018) establish asymptotic normality by essentially re-writing the sampling distribution of the estimator as a partial-sum process. In our setting the stratification is *random*, and so to prove our result we generalize their construction in a way that allows us to re-write the sampling distribution of the estimator as a *sequential empirical process* (see Van der Vaart and Wellner, 1996, Section 2.12.1 for a definition). We then exploit the asymptotic equicontinuity of this process to establish asymptotic normality (see Lemma A.1.1 for details). ■

We finish this subsection by constructing a consistent estimator for the variance V^*. Let $N(k) := m$ if $k = 0$ and $N(k) := n(k)$ otherwise. Let

$$\widehat{V}_H = \sum_{k=0}^{K} \frac{N(k)}{N} \left(\hat{\beta}(k) - \hat{\theta}\right)^2 ,$$

and let

$$\widehat{V}_Y = R' \hat{V}_{hc} R ,$$

where $\hat{V}_{hc}$ is the robust variance estimator for the parameters in the saturated regression, and R is following vector with $K + 1$ "leading" zeros:

$$R' = \left[0, 0, 0, \ldots, 0, \frac{N(0)}{N}, \frac{N(1)}{N}, \ldots, \frac{N(K)}{N}\right] .$$

We obtain the following consistency result:

Theorem 1.3.2. *Given Assumptions 1.2.1, 1.2.2, 1.2.3, 1.3.1, 1.3.2, 1.3.4, 1.3.5, and 1.3.6, then*

$$\widehat{V}(\hat{T}) \xrightarrow{p} V^* ,$$

where

$$\widehat{V}(T) = \widehat{V}_H(T) + \widehat{V}_Y(T) ,$$

as $m, n \to \infty$.

Remark 1.3.5. In Appendix A.3.2 we provide results under the "large pilot" asymptotic framework which we presented in Remark 1.3.2. Here we will briefly preview these results: under appropriate conditions it can be shown that in this alternative framework,

$$\sqrt{N}(\hat{\theta}(\hat{T}) - \theta) \xrightarrow{d} N(0, V_\lambda^*) ,$$

where

$$V_\lambda^* = \lambda V_0 + (1-\lambda)V^* \ ,$$

and

$$V_0 = \frac{\sigma_0^2(0)}{1-\pi(0)} + \frac{\sigma_1^2(0)}{\pi(0)} \ .$$

In words, we see that the pooled estimator $\hat{\theta}(\hat{T})$ now has an asymptotic variance which is a weighted combination of the optimal variance and the variance from estimation in the pilot experiment, with weights which correspond to their relative sizes. ■

1.3.2. Extensions

In this section we present some extensions to the main results. First we present a version of $\widetilde{T}$ whose depth is selected by cross-validation. Second, we explain how to accommodate the targeting of subgroup-specific effects. Finally, we explain how to extend our method to the setting with multiple treatments.

1.3.2.1. Cross-validation to select L. In this section we describe a method to select the depth L via cross-validation. The tradeoff in choosing L can be framed as follows: by construction, choosing a larger L has the potential to lower the variance of our estimator, since now we are optimizing in a larger set of trees. On the other hand, choosing a larger L will make the set of trees more complex, and hence will make the optimal tree harder to estimate accurately for a given pilot-data sample size. We suggest a procedure to select L with these two tradeoffs in mind. We proceed by first specifying some maximum upper bound $\bar{L}$ on the depth to be considered. For each $0 \leqslant L \leqslant \bar{L}$ (where we understand $L = 0$ to mean no stratification), define

$$V_L^* := \arg\min_{T \in \mathcal{T}_L} V(T) \ .$$

Note that by construction it is the case that $V_0^* \geqslant V_1^* \geqslant V_2^* \geqslant ... \geqslant V_{\bar{L}}^*$. Let $\widetilde{T}_L$ be the stratification tree estimated from class $\mathcal{T}_L$, then by Assumption 1.3.2, we have that

$$|V(\widetilde{T}_L) - V_L^*| \xrightarrow{a.s} 0 \ ,$$

for each $L \leqslant \bar{L}$. Despite the fact that $\widetilde{T}_L$ asymptotically achieves a (weakly) lower variance as L grows, it is not clear that, in finite samples, a larger choice of L should be favored, since we run the risk of estimating the optimal tree poorly (i.e. of overfitting). In order to protect against this potential for overfitting, we propose a simple cross-validated version of the stratification tree estimator. The use of cross-validation to estimate decision trees goes back at least to the work of Breiman (see Breiman et al., 1984). For an overview of the use of cross-validation methods in statistics in general, see Arlot et al. (2010).

The cross-validation procedure we propose proceeds as follows: let $\{W_i\}_{i=1}^m$ be the pilot data, and for simplicity suppose m is even. Split the pilot sample into two halves and denote these by $\mathcal{D}_1 := \{W_i\}_{i=1}^{m/2}$ and $\mathcal{D}_2 := \{W_i\}_{m/2+1}^m$, respectively. Now for each L, let $\widetilde{T}_L^{(1)}$ and $\widetilde{T}_L^{(2)}$ be stratification trees of depth L estimated on $\mathcal{D}_1$ and $\mathcal{D}_2$. Let $\widetilde{V}^{(1)}(\cdot)$ and $\widetilde{V}^{(2)}(\cdot)$ be the empirical variances computed on $\mathcal{D}_1$ and $\mathcal{D}_2$ (where, in the event that a cell in the tree partition is empty, we assign a value of infinity to the empirical variance). Define the following cross-validation criterion:

$$\widetilde{V}_L^{CV} := \frac{1}{2}\left(\widetilde{V}^{(1)}\left(\widetilde{T}_L^{(2)}\right) + \widetilde{V}^{(2)}\left(\widetilde{T}_L^{(1)}\right)\right) \ .$$

In words, for each L, we estimate a stratification tree on each half of the sample, compute the empirical variance of these estimates by using the *other* half of the sample, and then average the results. Intuitively, as we move from small values of L to large values of L, we would expect that this cross-validation criterion should generally decrease with L, and then

eventually increase, in accordance with the tradeoff between tree complexity and estimation accuracy. We define the cross-validated stratification tree as follows:

$$\widetilde{T}^{CV} = \widetilde{T}_{\hat{L}} \ ,$$

with

$$\hat{L} = \arg\min_{L} \widetilde{V}_{L}^{CV} \ ,$$

where in the event of a tie we choose the smallest such L. Hence $\widetilde{T}^{CV}$ is chosen to be the stratification tree whose depth minimizes the cross-validation criterion $\widetilde{V}_{L}^{CV}$. If each $\widetilde{T}_L$ is estimated by minimizing the empirical variance over $\mathcal{T}_L$, as described in Sections 1.2.2 and 1.3.1, then we can show that the cross-validated estimator satisfies the consistency property of Assumption 1.3.2:

Proposition 1.3.3. *Under Assumptions 1.2.1, 1.2.2, 1.2.3, 1.3.1 and 1.3.3, Assumption 1.3.2 is satisfied for* $\widetilde{T}^{CV} = \widetilde{T}^{EM}_{\hat{L}}$ *in the set* $\mathcal{T}_{\bar{L}}$*, that is,*

$$|V(\widetilde{T}^{CV}) - V^{*}_{\bar{L}}| \xrightarrow{a.s} 0 \ ,$$

as $m \to \infty$*.*

Remark 1.3.6. Our description of cross-validation above defines what is known as "2-fold" cross-validation. It is straightforward to extend this to "V-fold" cross-validation, where the dataset is split into V pieces. Breiman et al. (1984) find that using at least 5 folds is most effective in their setting (although their cross-validation technique is different from ours), and in many statistical applications 5 or 10 folds has become the practical standard. For our purposes, we focus on 2-fold cross validation because of the computational difficulties we face in solving the optimization problem to compute $\widetilde{T}^{EM}$. ■

In light of Proposition 1.3.3 we see that all of our previous results continue to hold while using $\widetilde{T}^{CV}$ as our stratification tree. However, Proposition 1.3.3 *does not* help us conclude that $\widetilde{T}^{CV}$ should perform any better than $\widetilde{T}_{\bar{L}}$ in finite samples. Although it is beyond the scope of this paper to establish such a result, doing so could be an interesting avenue for future work. Instead, we assess the performance of $\widetilde{T}^{CV}$ via simulation in Section 1.4, and note that it does indeed seem to protect against overfitting in practice. In Section 1.5, we use this cross-validation procedure to select the depth of the stratification trees we estimate for the experiment undertaken in Karlan and Wood (2017).

1.3.2.2. Stratification Trees for Subgroup Targeting. In this subsection we explain how the method can flexibly accommodate the problem of variance reduction for estimators of subgroup-specific ATEs, while still minimizing the variance of the unconditional ATE estimator in a restricted set of trees. It is common practice in RCTs for the strata to be specified such that they are the subgroups that a researcher is interested in studying (see for example the recommendations in Glennerster and Takavarasha, 2013). This serves two purposes: the first is that it enforces a pre-specification of the subgroups of interest, which guards against ex-post data mining. Second, it allows the researcher to improve the efficiency of the subgroup specific estimates.

Let $S' \in \mathcal{S}_{L'}$ be a tree of depth $L' < L$, whose terminal nodes represent the subgroups of interest. Suppose these nodes are labelled by $g = 1, 2, ..., G$, and that $P(S'(X) = g) > 0$ for each g. The subgroup-specific ATEs are defined as follows:

$$\theta^{(g)} := E[Y(1) - Y(0)|S'(X) = g] \ .$$

We introduce the following new notation: let $\mathcal{T}_L(S') \subset \mathcal{T}_L$ be the set of stratification trees which can be constructed as *extensions* of S'. For a given $T \in \mathcal{T}_L(S')$, let $\mathcal{K}_g(T) \subset [K]$ be

the set of terminal nodes of T which pass through the node g in S' (see Figure 1.5 for an example).

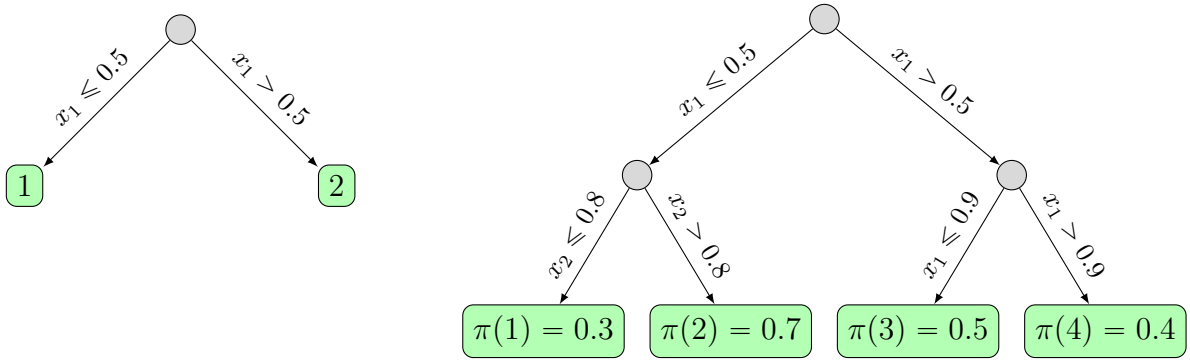

Figure 1.5. On the left: a tree S' whose nodes represent the subgroups of interest.
On the right: an extension $T \in \mathcal{T}_2(S')$. Here $\mathcal{K}_1(T) = \{1, 2\}, \mathcal{K}_2(T) = \{3, 4\}$

Given a tree $T \in \mathcal{T}_L(S')$, a natural estimator of $\theta^{(g)}$ is then given by

$$\hat{\theta}^{(g)}(T) := \sum_{k \in \mathcal{K}_g} \frac{n(k)}{n'(g)} \hat{\beta}(k) \ ,$$

where $n'(g) = \sum_{i=1}^{n} \mathbf{1}\{S'(X_i) = g\}$ and $\hat{\beta}(k)$ are the regression coefficients of the saturated regression over T. It is then straightforward to show using the recursive structure of stratification trees that choosing T as a solution to the following problem:

$$V^*(S') := \min_{T \in \mathcal{T}_L(S')} V(T) \ ,$$

will minimize the asymptotic variance of the subgroup specific estimators $\hat{\theta}^{(g)}$, while still minimizing the variance of the global ATE estimator $\hat{\theta}$ in the restricted set of trees $\mathcal{T}_L(S')$. Moreover, to compute a minimizer of $V(T)$ over $\mathcal{T}_L(S')$, it suffices to compute the optimal tree for each subgroup, and then append these to S' to form the stratification tree. Finally,

the appropriate analogues to Theorems 3.3.1 and 1.3.2 for the estimators $\hat{\theta}^{(g)}$ will also follow without any additional assumptions.

In Section 1.5 we illustrate the application of this extension to the setting in Karlan and Wood (2017). In their paper, they study the effect of information about a charity's effectiveness on subsequent donations to the charity, and in particular the treatment effect heterogeneity between large and small prior donors. For this application we specify S' to be a tree of depth 1, whose terminal nodes correspond to the subgroups of large and small prior donors. We then compute $\widetilde{T}$ for each of these subgroups and append them to S' to form a stratification tree which simultaneously minimizes the variance of the subgroup-specific estimators, while still minimizing the variance of the global estimator in this restricted class.

1.3.2.3. Extension to Multiple Treatments. Here we consider the extension to multiple treatments. Let $\mathcal{A} = \{1, 2, ..., J\}$ denote the set of possible treatments, where we consider the treatment $A = 0$ as being the "control group". Let $\mathcal{A}_0 = \mathcal{A} \cup \{0\}$ be the set of treatments including the control. Our quantities of interest are now given by

$$\theta_a := E[Y(a) - Y(0)] \ ,$$

for $a \in \mathcal{A}$, so that we consider the set of ATEs of the treatments relative to the control. Let $\theta := (\theta_a)_{a \in \mathcal{A}}$ denote the vector of these ATEs.

The definition of a stratification tree $T \in \mathcal{T}_L$ is extended in the following way: instead of specifying a collection $\pi = (\pi(k))_{k=1}^K$ of assignment targets for treatment 1, we specify, for each k, a *vector* of assignment targets for all $a \in \mathcal{A}_0$, so that $\pi = (\{\pi_a(k)\}_{a \in \mathcal{A}_0})_{k=1}^K$, where each $\pi_a(k) \in (0, 1)$ and $\sum_{a \in \mathcal{A}_0} \pi_a(k) = 1$. We also consider the following generalization of our

estimator: consider estimation of the following equation by OLS

$$Y_i = \sum_{k\in[K]} \alpha(k)\mathbf{1}\{S_i = k\} + \sum_{a\in\mathcal{A}}\sum_{k\in[K]} \beta_a(k)\mathbf{1}\{A_i = a, S_i = k\} + u_i \ ,$$

then our estimators are given by

$$\hat{\theta}_a(T) = \sum_k \frac{n(k)}{n}\hat{\beta}_a(k) \ .$$

Now, for a fixed $T \in \mathcal{T}_L$, the results in Bugni et al. (2018) imply that $\sqrt{n}(\hat{\theta}(T) - \theta)$ is asymptotically multivariate normal with covariance matrix given by:

$$\mathbb{V}(T) := \sum_k p(k;T)\left(\mathbb{V}_H(k;T) + \mathbb{V}_Y(k;T)\right) \ ,$$

with

$$\mathbb{V}_H(k;T) := \text{outer}\left[(E[Y(a) - Y(0)|S(X) = k] - E[Y(a) - Y(0)]) : a \in \mathcal{A}\right] \ ,$$

$$\mathbb{V}_Y(k;T) := \frac{\sigma_0^2(k)}{\pi_0(k)}\iota_{|\mathcal{A}|}\iota'_{|\mathcal{A}|} + \text{diag}\left(\left(\frac{\sigma_a^2(k)}{\pi_a(k)}\right) : a \in \mathcal{A}\right) \ ,$$

where the notation $v := (v_a : a \in \mathcal{A})$ denotes a column vector, $\text{outer}(v) := vv'$, and ι_M is a vector of ones of length M. Note that from the results in Cattaneo (2010), this is the semi-parametric efficiency bound in the multiple treatment setting for the discretization $S(\cdot)$.

Because we are now dealing with a covariance matrix $\mathbb{V}(T)$ as opposed to the scalar quantity $V(T)$, we need to be more careful about what criterion we will use to decide on an optimal T. The literature on experimental design has considered various targets (see Pukelsheim, 2006, for some examples). In this paper we will consider the following collection of targets:

$$V^* = \min_{T\in\mathcal{T}_L} ||\mathbb{V}(T)|| \ ,$$

where $||\cdot||$ is some matrix norm. In particular, if we let $||\cdot||$ be the Euclidean operator-norm, then our criterion is equivalent to minimizing the largest eigenvalue of $\mathbb{V}(T)$, which coincides with the notion of E-optimality in the study of optimal experimental design in the linear model (see for example Section 6.4 of Pukelsheim, 2006). Intuitively, if we consider the limiting normal distribution of our estimator, then any fixed level-surface of its density forms an ellipsoid in $\mathbb{R}^{|\mathcal{A}|}$. Minimizing $||\mathbb{V}(T)||$ in the Euclidean operator-norm corresponds to minimizing the longest axis of this ellipsoid.

If we consider the following generalization of the empirical minimization problem:

$$\widetilde{T}^{EM} = \arg\min_{T \in \mathcal{T}_L} ||\widetilde{\mathbb{V}}(T)|| \ ,$$

where $\widetilde{\mathbb{V}}(T)$ is an empirical analog of $\mathbb{V}(T)$, then analogous results to those presented in Section 1.3.1 continue to hold in the multiple treatment setting as well, under some additional regularity conditions (see Appendix A.3.3 for precise statements).

1.4. Simulations

In this section we analyze the finite sample behaviour of our method via a simulation study. We consider three DGPs in the spirit of the designs considered in Athey and Imbens (2016). For all three designs, the outcomes are specified as follows:

$$Y_i(a) = \kappa_a(X_i) + \nu_a(X_i) \cdot \epsilon_{a,i} \ .$$

Where the $\epsilon_{a,i}$ are i.i.d $N(0, 0.1)$, and $\kappa_a(\cdot)$, $\nu_a(\cdot)$ are specified individually for each DGP below. In all cases, $X_i \in [0,1]^d$, with components independently and identically distributed as $Beta(2,5)$. The specifications are given by:

Model 1: $d = 2$, $\kappa_0(x) = 0.2$, $\nu_0(x) = 5$,

$$\kappa_1(x) = 10x_1\mathbf{1}\{x_1 > 0.4\} - 5x_2\mathbf{1}\{x_2 > 0.4\} \ ,$$

$$\nu_1(x) = 10x_1\mathbf{1}\{x_1 > 0.6\} + 5x_2\mathbf{1}\{x_2 > 0.6\} \ .$$

This is a "low-dimensional" design with two covariates. The first covariate is given a higher weight than the second in the outcome equation for $Y(1)$.

Model 2: $d = 10$, $\kappa_0(x) = 0.5$, $\nu_0(x) = 5$,

$$\kappa_1(x) = \sum_{j=1}^{10}(-1)^{j-1}10^{-j+2}\mathbf{1}\{x_j > 0.4\} \ ,$$

$$\nu_1(x) = \sum_{j=1}^{10}10^{-j+2}\mathbf{1}\{x_j > 0.6\} \ .$$

This is a "moderate-dimensional" design with ten covariates. Here the first covariate has the largest weight in the outcome equation for $Y(1)$, and the weight of subsequent covariates decreases quickly.

Model 3: $d = 10$, $\kappa_0(x) = 0.2$, $\nu_0(x) = 9$,

$$\kappa_1(x) = \sum_{j=1}^{3}(-1)^{j-1}10 \cdot \mathbf{1}\{x_j > 0.4\} + \sum_{j=4}^{10}(-1)^{j-1}5 \cdot \mathbf{1}\{x_j > 0.4\} \ ,$$

$$\nu_1(x) = \sum_{j=1}^{3}10 \cdot \mathbf{1}\{x_j > 0.6\} + \sum_{j=4}^{10}5 \cdot \mathbf{1}\{x_j > 0.6\} \ .$$

This is a "moderate-dimensional" design with ten covariates. Here the first three covariates have similar weight in the outcome equation for $Y(1)$, and the next seven covariates have a smaller but still significant weight.

In each case, $\kappa_0(\cdot)$ is calibrated so that the average treatment effect is close to 0.1, and $\nu_0(\cdot)$ is calibrated so that $Y_i(1)$ and $Y_i(0)$ have similar unconditional variances (see Appendix

A.4 for details). In each simulation we test five different methods of stratification, where we estimate the ATE using the saturated regression estimator described in Section 1.2.3. In all cases, when we stratify we consider a maximum of 8 strata (which corresponds to a stratification tree of depth 3). In all cases we use SBR to perform assignment. We consider the following methods of stratification:

- No Stratification: Here we assign the treatment to half the sample, with no stratification.
- Ad-hoc: Here we stratify in an "ad-hoc" fashion and then assign treatment to half the sample in each stratum. To construct the strata we iteratively select a covariate at random, and stratify on the midpoints of the currently defined strata.
- Stratification Tree: Here we split the sample and perform a pilot experiment to estimate a stratification tree, we then use this tree to assign treatment in the second wave.
- Cross-Validated Tree: Here we estimate a stratification tree as above, while selecting the depth via cross validation.
- Infeasible Optimal Tree: Here we estimate an "optimal" tree by using a large auxiliary sample. We then use this to assign treatment to the entire sample (see Appendix A.4 for further details).

We perform the simulations with a sample size of $5,000$, and consider three different splits of the total sample for the pilot experiment and main experiment when performing our method (for all other methods all $5,000$ observations are used in one experiment). For all cases with a pilot, the pilot experiment was performed using simple random assignment without stratification. To estimate the stratification trees we minimize an empirical analog of the asymptotic variance as described in Appendix A.4.

We assess the performance of the randomization procedures through the following criteria: the empirical coverage of a 95% confidence interval formed using a normal approximation, the percentage reduction in average length of the 95% CI relative to no stratification, the power of a t-test for an ATE of 0, and the percentage reduction in root mean-squared error (RMSE) relative to no stratification. For each design we perform 5000 Monte Carlo iterations. Table 1.1 presents the simulation results for Model 1.

Sample Size			Criteria			
Pilot	Main	Stratification Method	Coverage	%ΔLength	Power	%ΔRMSE
100	4900	No Stratification	94.4	0.0	78.6	0.0
		Ad-Hoc	94.5	-7.0	83.8	-7.1
		Strat. Tree	94.4	0.0	77.1	2.0
		CV Tree	94.9	-5.1	81.3	-4.8
		Infeasible Tree	94.7	-19.0	91.4	-18.3
500	4500	No Stratification	94.6	0.0	78.3	0.0
		Ad-Hoc	94.3	-7.0	83.4	-6.8
		Strat. Tree	94.5	-13.5	88.1	-13.1
		CV Tree	94.8	-12.9	88.2	-13.2
		Infeasible Tree	94.1	-19.0	92.1	-18.3
1500	3500	No Stratification	94.4	0.0	77.4	0.0
		Ad-Hoc	94.4	-7.0	82.7	-7.0
		Strat. Tree	94.3	-12.0	86.2	-11.5
		CV Tree	94.3	-11.7	85.9	-11.9
		Infeasible Tree	94.4	-19.0	92.2	-19.6

Table 1.1. Simulation Results for Model 1

In Table 1.1, we see that when the pilot study is small (sample size 100), our method can perform poorly relative to ad-hoc stratification. However, the CV tree does a good job of avoiding overfitting, and performs only slightly worse than ad-hoc stratification for this design. When we consider a medium-sized pilot study (sample size 500), we see that both the stratification tree and the CV tree outperform ad-hoc stratification. To put these gains in perspective, the ad-hoc stratification procedure would require 500 additional observations to

match the performance of the stratification trees, and the no-stratification procedure would require 1500 additional observations. Finally, when using a large pilot study (sample size 1500), we see a small drop in performance for both trees. This drop in performance can be explained through the alternative "large-pilot" asymptotic framework that we introduced in Remark 1.3.5. Summarizing the results of Table 1.1, the CV tree seems to do a good job of preventing overfitting and in general performs as well or better than the stratification tree in all three scenarios. Overall, the stratification tree and CV tree display modest gains relative to ad-hoc stratification in this low-dimensional setting. Next we study the results for Model 2, presented in Table 1.2:

Sample Size			Criteria			
Pilot	Main	Stratification Method	Coverage	%ΔLength	Power	%ΔRMSE
100	4900	No Stratification	94.1	0.0	46.8	0.0
		Ad-Hoc	94.8	-1.8	48.2	-3.7
		Strat. Tree	94.4	7.0	42.1	6.5
		CV Tree	94.1	-7.7	53.2	-7.7
		Infeasible Tree	94.2	-19.6	64.4	-19.5
500	4500	No Stratification	94.2	0.0	46.1	0.0
		Ad-Hoc	94.4	-1.8	48.6	-2.1
		Strat. Tree	94.5	-12.7	58.0	-13.5
		CV Tree	94.5	-14.0	58.1	-13.7
		Infeasible Tree	94.3	-19.7	65.0	-19.4
1500	3500	No Stratification	93.9	0.0	46.6	0.0
		Ad-Hoc	94 .4	-1.8	49.0	-1.8
		Strat. Tree	94.0	-12.4	57.9	-11.7
		CV Tree	94.1	-12.1	58.9	-11.9
		Infeasible Tree	93.8	-19.7	65.9	-18.6

Table 1.2. Simulation Results for Model 2

In Table 1.2, we see that for a small pilot, we get similar results to Model 1, with the CV tree again doing a good job of avoiding overfitting. For a medium-sized pilot, both trees display sizeable gains relative to ad-hoc stratification. To put these gain in perspective, both

the ad-hoc stratification and the no-stratification procedures would require 1500 additional observations to match the performance of the stratification trees. To summarize the results of Table 1.2, we again have that the CV tree performs best across all three specifications. For small pilots it does a good job of preventing overfitting, and for larger pilots it displays sizeable gains relative to ad-hoc stratification. Finally, we study the results of Model 3, presented in Table 1.3.

Sample Size		Stratification Method	Criteria			
Pilot	Main		Coverage	%ΔLength	Power	%ΔRMSE
100	4900	No Stratification	95.4	0.0	30.9	0.0
		Ad-Hoc	95.1	-2.2	31.7	-0.6
		Strat. Tree	94.5	16.3	24.2	19.5
		CV Tree	94.8	1.0	30.4	2.1
		Infeasible Tree	94.6	-7.4	36.0	-5.5
500	4500	No Stratification	95.2	0.0	30.9	0.0
		Ad-Hoc	95.4	-2.2	32.2	-4.5
		Strat. Tree	94.4	-2.1	32.4	-1.1
		CV Tree	95.4	-1.9	31.7	-4.4
		Infeasible Tree	95.1	-7.4	35.0	-9.8
1500	3500	No Stratification	94.2	0.0	30.9	0.0
		Ad-Hoc	94.8	-2.2	31.9	-3.1
		Strat. Tree	94.6	-4.0	32.1	-4.7
		CV Tree	94.4	-3.5	32.1	-2.7
		Infeasible Tree	95.0	-7.4	35.2	-7.5

Table 1.3. Simulation Results for Model 3

In Table 1.3, we see very poor performance of our method when using a small pilot. However, as was the case for Models 1 and 2, the CV tree still helps to protect against overfitting. When moving to the medium and large sized pilots, we see that both trees perform comparably to ad-hoc stratification. We note that the gains from stratification in this design are quite small. For example, the no-stratification procedure would require

only 200 additional observations to match the performance of ad-hoc stratification, and approximately 500 additional observations to match the performance of the optimal tree.

Overall, we conclude that stratification trees can provide moderate to substantial improvements over ad-hoc stratification, with the greatest improvements coming from DGPs with some amount of "sparsity", as in Model 2. The cross-validation method seems most robust to the choice of pilot-study size, however, in general we caution against using the method with very small pilots.

1.5. An Application

In this section we study the behavior of our method in an application, using the experimental data from Karlan and Wood (2017). First we provide a brief review of the empirical setting: Karlan and Wood (2017) study how donors to the charity Freedom from Hunger respond to new information about the charity's effectiveness. The experiment, which proceeded in two separate waves corresponding to regularly scheduled fundraising campaigns, randomly mailed one of two different marketing solicitations to previous donors, with one solicitation emphasizing the scientific research on FFH's impact, and the other emphasizing an emotional appeal to a specific beneficiary of the charity. The outcome of interest was the amount donated in response to the mailer. Karlan and Wood (2017) found that, although the effect of the research insert was small and insignificant, there was substantial heterogeneity in response to the treatment: for those who had given a large amount of money in the past, the effect of the research insert was positive, whereas for those who had given a small amount, the effect was negative. They argue that this evidence is consistent with the behavioural mechanism proposed by Kahneman (2003), where small prior donors are driven by a "warm-glow" of giving (akin to Kahneman's System I decision making), in contrast to large prior donors, who are driven by altruism (akin to Kahneman's System II decision

making). However, the resulting confidence intervals of their estimates are wide, and often contain zero (see for example Figure 1 in Karlan and Wood, 2017). The covariates available in the dataset for stratification are as follows:

- Total amount donated prior to mailer
- Amount of most recent donation prior to mailer (denoted `pre gift` below)
- Amount of largest donation prior to mailer
- Number of years as a donor (denoted `# years` below)
- Number of donations per year (denoted `freq` below)
- Average years of education in census tract
- Median zipcode income
- Prior giving year (either 2004/05 or 2006/07) (denoted `p.year` below)

As a basis for comparison, Figure 1.6 depicts the stratification used in Karlan and Wood (2017).

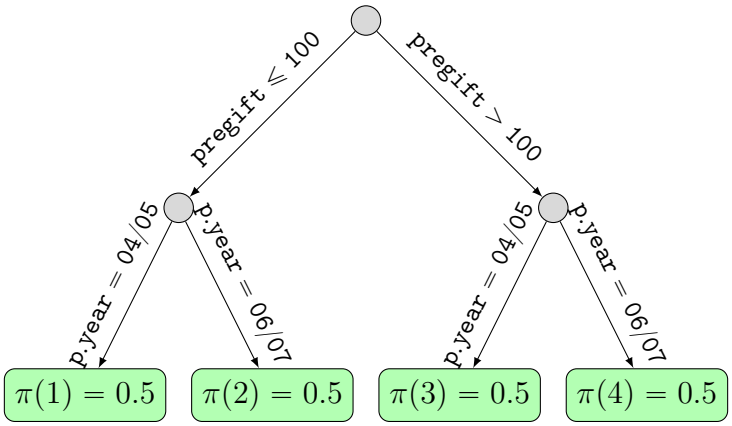

Figure 1.6. Stratification used in Karlan and Wood (2017)

We estimate two different stratification trees using data from the first wave of the experiment (with a sample size of $10,869$)[3], that illustrate stratifications which could have been

used to assign treatment in the second wave. We compute the trees by minimizing an empirical analog of the variance, as described in Sections 1.2.3 and 1.3.1. The first tree is fully unconstrained, and hence targets efficient estimation of the unconditional ATE estimator, while the second tree is constrained in accordance with Section 1.3.2 to efficiently target estimation of the subgroup-specific effects for large and small prior donors (see below for a precise definition). In both cases, the depth of the stratification tree was selected using cross validation as described in Section 1.3.2, with a maximal depth of $\bar{L} = 5$ (which corresponds to a maximum of 32 strata). When computing our trees, given that some of these covariates do not have upper bounds a-priori, we impose an upper bound on the allowable range for the strata to be considered in accordance with Remark 1.2.3 (we set the upper bound as roughly the 97th percentile in the dataset, although in practice this should be set using historical data).

Figure 1.7 depicts the unrestricted tree estimated via cross-validation. We see that the cross-validation procedure selects a tree of depth one, which may suggest that the covariates available to us for stratification are not especially relevant for decreasing the variance of the estimator. However, we do see a wide discrepancy in the assignment proportions for the selected strata. In words, the subgroup of respondents who have been donors for more than 16 years have a larger variance in outcomes when receiving the research mailer than the control mailer. In contrast the subgroup of respondents who have been donors for less than 16 years have roughly equal variances in outcomes under both treatments.

[3]Replication data is available by request from Innovations for Poverty Action. Observations with missing data on median income, average years of education, and those receiving the "story insert" were dropped.

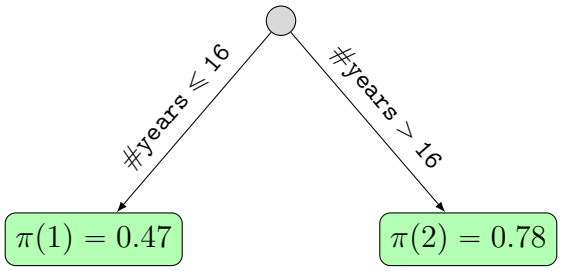

Figure 1.7. Unrestricted Stratification Tree estimated from Karlan and Wood (2017) data

Next, we estimate the restricted stratification tree which targets the subgroup-specific treatment effects for large and small prior donors. We specify a large donor as someone who's most recent donation prior to the experiment was larger than $100. We proceed by estimating each subtree using cross-validation. Figure 1.8 depicts the estimated tree. We see that the cross-validation procedure selects a stratification tree of depth 1 in the left subtree and a tree of depth 0 (i.e. no stratification) in the right subtree, which further reinforces that the covariates we have available may be uninformative for decreasing variance.

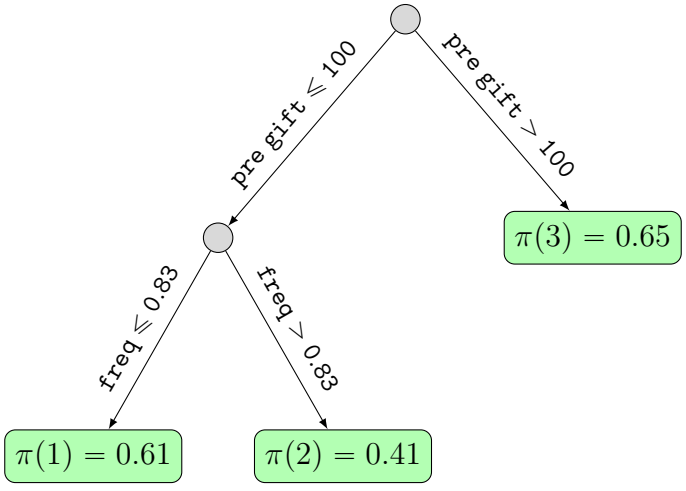

Figure 1.8. Restricted Stratification Tree estimated from Karlan and Wood (2017) data

The results of this exercise suggest a potential added benefit from using our method: when using cross-validation, the depth of the resulting tree could serve as a diagnostic tool to help assess the potential gains from stratification in a given application. In particular, if the procedure outputs a very shallow tree given a relatively large sample, this may suggest that the potential gains from stratification are small. To further assess the potential gains from stratification in this application, in Appendix A.4 we repeat the simulation exercise of Section 1.4 with an application-based simulation design, where we treat the sample data as the true DGP. There we find that using an "optimal" stratification tree of depth 2 results in an 8% reduction in RMSE and a 6% reduction in CI length relative to no stratification (using a CV tree with a maximum depth of 2 results in a 3% reduction in RMSE and a 2% reduction in CI length). This again reinforces that the gains from stratification may be fairly small in this setting.

1.6. Conclusion

In this paper we proposed an adaptive randomization procedure for two-stage randomized controlled trials, which uses the data from a first-wave experiment to assign treatment in a second wave of the RCT. Our method uses the first-wave data to estimate a stratification tree: a stratification of the covariate space into a tree partition along with treatment assignment probabilities for each of these strata. The main result of the paper showed that using our procedure results in an estimator with an asymptotic variance which minimizes the semi-parametric efficiency bound of Hahn (1998), over an optimal stratification of the covariate space. We also described extensions which accommodate multiple treatments, as well as to target subgroup-specific effects. In simulations, the method was most effective when the response model exhibited some amount of "sparsity" with respect to the covariates, but was

shown to be effective in other contexts as well, as long as the sample size of the pilot being used to estimate the stratification tree was not prohibitively small.

Going forward, there are several extensions of the paper that we would like to consider. First, many RCTs are performed as *cluster* RCTs, that is, where treatment is assigned at a higher level of aggregation such as a school or city. Extending the results of the paper to this setting could be a worthwhile next step. Another avenue to consider would be to combine our randomization procedure with other aspects of the experimental design. For example, Carneiro et al. (2016) set up a statistical decision problem to optimally select the sample size, as well as the number of covariates to collect from each participant in the experiment, given a fixed budget. It may be interesting to embed our randomization procedure into a similar decision problem. Finally, although our method employs stratified randomization, we assumed throughout that the experimental sample is an i.i.d sample. Further gains may be possible by considering a setting where we are able to conduct stratified *sampling* in the second wave as well as stratified randomization. To that end, Song and Yu (2014) develop estimators and semi-parametric efficiency bounds for stratified sampling which may be useful.

CHAPTER 2

Model Selection for Treatment Choice: Penalized Welfare Maximization

2.1. Introduction to Chapter 2

This paper develops a new statistical decision rule for the treatment assignment problem. A major goal of treatment evaluation is to provide policy makers with guidance on how to assign individuals to treatment, given experimental or quasi-experimental data. Following the literature inspired by Manski (2004) (a partial list in econometrics includes Armstrong and Shen, 2015; Athey and Wager, 2017; Bhattacharya and Dupas, 2012; Chamberlain, 2011; Dehejia, 2005; Hirano and Porter, 2009; Kasy, 2014; Kitagawa and Tetenov, 2018; Kock and Thyrsgaard, 2017; Schlag, 2007; Stoye, 2009, 2012; Tetenov, 2012), we treat the treatment assignment problem as a statistical decision problem of maximizing population welfare. Like many of the above papers, we evaluate our decision rule by its maximum regret.

Often, policy makers have observable characteristics at their disposal on which to base treatment, however, they may not always have full discretion on how these covariates can be used. For example, policy makers may face exogenous constraints on how they can use covariates for legal, ethical, or political reasons. Even in cases where policy makers have leeway in how they assign treatment, plausible modeling assumptions may imply certain

[0](continued from previous page), NASMES 2017, and the Bristol Econometrics Study Group for helpful comments, as well as Nitish Keskar for help in implementing EWM. This research was supported in part through the computational resources and staff contributions provided for the Social Sciences Computing Cluster (SSCC) at Northwestern University. All mistakes are our own.

restrictions on assignment. Kitagawa and Tetenov (2018) develop what they call the Empirical Welfare Maximization (or EWM) rule, whose primary feature is its ability to solve the treatment choice problem when exogenous constraints are placed on assignment. EWM will play an important role in the development of our rule, which we call the Penalized Welfare Maximization rule (PWM).

The PWM rule is designed to address situations in which the policy maker can choose amongst a *collection* of constrained allocations. To be concrete, suppose we have two treatments, and we represent assignment into these treatments by partitioning the covariate space into two pieces. We can then think of constraints on assignment as constraints on the allowable subsets that we can consider for the partitions. Kitagawa and Tetenov (2018) focus on deriving bounds on maximum regret of the EWM rule for a *fixed* class of subsets of finite VC dimension (see Györfi et al. (1996) for a definition). In this paper, however, we consider settings where the class of allowable subsets is "large". We approach the problem by approximating our class of allowable allocations by a sequence of subclasses of finite VC dimension. We establish an oracle inequality for the regret of the PWM rule which shows that it behaves as if we knew the "correct" class to use in the sequence. We then use this result to derive bounds on the maximum regret of the PWM rule in two empirically relevant settings.

The first setting that we consider is one where the class of feasible allocations has infinite VC dimension. In particular, we argue that economic modeling assumptions may sometimes put restrictions on the unconstrained optimum that naturally generate classes of infinite VC dimension. For example, plausible assumptions may only impose shape restrictions on the optimal allocation. To solve the optimal welfare assignment problem in this setting, we approximate these large classes of feasible allocations by sequences of classes of finite VC

dimension. The strength of the PWM rule in this setting will then be to provide a data-driven method by which to select an "appropriate" approximating class. In doing so we will derive bounds on the maximum regret of the PWM rule for a large set of classes of infinite VC dimension.

The second setting we consider is one where the class of feasible allocations may have large VC dimension relative to the sample size. This could arise, for example, if the planner has many covariates on which to base assignment. As is shown in Kitagawa and Tetenov (2018), when the constraints placed on assignment are too flexible relative to the sample size available, the EWM rule may suffer from overfitting, which can result in inflated values of regret. By the same mechanism that allows PWM to select an appropriate approximating class in our first application, we can use PWM in order to select amongst simpler subclasses in this setting as well, in a way that improves the performance of the allocation rule in finite samples. We illustrate PWM's ability to reduce regret in a simulation study where the policy maker has many covariates on which to base treatment assignment, but does not know how many to use when performing best-subset selection.

The PWM rule is heavily inspired by the literature on model selection in classification: see for example the seminal work of Vapnik and Chervonenkis (1974), as well as Györfi et al. (1996), Koltchinskii (2001), Bartlett et al. (2002), Scott and Nowak (2002), Boucheron et al. (2005), Bartlett (2008), Koltchinskii (2008) among many others. The theoretical contribution of our paper is to modify and extend some of these tools to the setting of treatment choice. As pointed out in Kitagawa and Tetenov (2018), there are substantive differences between classification and treatment choice: observed outcomes are real-valued in the setting of treatment choice, and only one of the potential outcomes is observed for any given individual. When we say that we extend these tools, we mean that we prove

results for settings where the data available to the policy maker is quasi-experimental. As we will see, in such a setting the policy maker's objective function contains an estimated quantity, which is not an issue that arises in the classification problem. In deciding which tools to extend, we have attempted to strike a balance between ease of use for practitioners, theoretical appeal, and performance in simulations. The connection between classification and treatment choice has been explored in various fields, including machine learning, under the label of *policy learning* (see Beygelzimer and Langford, 2009; Kallus, 2016; Swaminathan and Joachims, 2015; Zadrozny, 2003, among others), and in epidemiology under the label of *individualized treatment rules* (examples include Qian and Murphy, 2011; Zhao et al., 2012). Kitagawa and Tetenov (2018) and Athey and Wager (2017) provide a discussion on the link between these various literatures.

The remainder of the paper is organized as follows. In Section 2.2, we setup the notation and formally define the problem that the policy maker (i.e. social planner) is attempting to solve. In Section 2.3, we introduce the PWM rule and present general results about its maximum regret. In Section 2.4, we perform a small simulation study to highlight PWM's ability to reduce regret when performing best-subset selection. In Section 2.5 we derive bounds on maximum regret of the PWM rule when the planner is constrained to what we call *monotone* allocations, and then illustrate these in an application to the JTPA study. Section 2.6 concludes.

2.2. Setup

Let Y_i denote the observed outcome of a unit i, and let D_i be a binary variable which denotes the treatment received by unit i. Let $Y_i(1)$ denote the *potential* outcome of unit i under treatment 1 (which we will refer to as "the treatment"), and let $Y_i(0)$ denote the potential outcome of unit i under treatment 0 (which we will refer to as "the control"). The

observed outcome for each unit is related to their potential outcomes through the expression:

$$Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i) \ . \tag{2.1}$$

Let $X_i \in \mathcal{X} \subset \mathbb{R}^{d_x}$ denote a vector of observed covariates for unit i. Let Q denote the distribution of $(Y_i(0), Y_i(1), D_i, X_i)$, then we assume that the planner observes a size n random sample

$$(Y_i, D_i, X_i)_{i=1}^n \sim P^n \ ,$$

where P is jointly determined by Q, and the expression in (2.1). Throughout the paper we will assume unconfoundedness, i.e.

Assumption 2.2.1. *(Unconfoundedness) The distribution Q satisfies:*

$$\Big((Y(1), Y(0)) \perp D\Big) \Big| X \ .$$

This assumption asserts that, once we condition on the observable covariates, the treatment is exogenous. This assumption will hold in a randomized controlled trial (RCT), which is our primary application of interest, since the treatment is exogenous by construction. This assumption is sometimes also made (possibly tenuously) in observational studies; it is a key identifying assumption when using matching or regression estimators in policy evaluation settings with observational data (Imbens, 2004, provides a review of these techniques, and discusses the validity of Assumption 2.2.1 in economic applications).

The planner's goal is to optimally assign the treatment to the population. The objective function we consider is utilitarian welfare, which is defined by the average of the individual outcomes in the population:

$$E_Q[Y(1)\mathbf{1}\{X \in G\} + Y(0)\mathbf{1}\{X \notin G\}] \ ,$$

where $G \subset \mathcal{X}$ represents the set of covariate values of the individuals assigned to treatment. The planner is tasked with choosing a *treatment allocation* $G \subset \mathcal{X}$ using the empirical data. Using Assumption 2.2.1, we can rewrite the welfare criterion as:

$$E_Q[Y(0)] + E_P\Big[\Big(\frac{YD}{e(X)} - \frac{Y(1-D)}{1-e(X)}\Big)\mathbf{1}\{X \in G\}\Big] ,$$

where $e(X) = E_P[D|X]$ is the propensity score. Since the first term of this expression does not depend on G, we define the planner's objective function given a choice of treatment allocation G as:

$$W(G) := E_P\Big[\Big(\frac{YD}{e(X)} - \frac{Y(1-D)}{1-e(X)}\Big)\mathbf{1}\{X \in G\}\Big] .$$

Let $\mathcal{G}$ be the class of all feasible treatment allocations. Here, we consider the possibility that the planner may be restricted in what type of allocations she can (or wants to) consider. These restrictions may arise from legal, ethical, or political considerations, or could arise as natural constraints from the economic model. Consider the following three examples of $\mathcal{G}$:

Example 2.2.1. $\mathcal{G}$ could be the set of all measurable subsets of $\mathcal{X}$. This is the largest possible class of admissible allocations. It is straightforward to show that the optimal allocation in this case is as follows: define

$$\tau(x) := E_Q[Y(1) - Y(0)|X = x] ,$$

then the optimal allocation is given by

$$G^*_{FB} := \{x \in \mathcal{X} : \tau(x) \geqslant 0\} ,$$

which assigns an individual with covariate x to treatment or control depending on whether the conditional average treatment effect at x is non-negative.

Example 2.2.2. Suppose $\mathcal{X} \subset \mathbb{R}$, and consider the class of *threshold allocations*:

$$\mathcal{G} = \{G : G = (-\infty, x] \cap \mathcal{X} \text{ or } G = [x, \infty) \cap \mathcal{X}, \text{ for } x \in \mathcal{X}\} .$$

Such a class $\mathcal{G}$ would be reasonable, for example, when assigning scholarships to students: suppose the only covariate available to the planner is a student's GPA, then it may be school policy that only threshold-type rules are to be considered.

Example 2.2.3. Let $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \subset \mathbb{R}^2$, and consider the class of *monotone allocations*:

$$\mathcal{G} = \left\{G : G = \{(x_1, x_2) \in \mathcal{X} |\, x_2 \geqslant f(x_1) \text{ for } f : \mathcal{X}_1 \to \mathcal{X}_2 \text{ increasing}\}\right\} .$$

As an example, consider again the setting of assigning scholarships to students (Example 2.2.2), but now suppose that the covariates available to the planner are parental income (x_1) and a student's GPA (x_2). The allocation rules considered in $\mathcal{G}$ are such that the GPA requirement for scholarship eligibility increases with parental income. In fact, even if the planner is *not* exogenously constrained to such allocations, this type of shape restriction could arise naturally from an economic model. Suppose, for instance, that the outcome of interest depends only on a student's innate "ability" (which is unobservable) and on whether or not the student receives the scholarship. Furthermore, suppose that the planner can only use information on GPA and parental income to assign scholarships, which have a per-unit cost. Under some modeling assumptions (outlined in Appendix A.3) on the outcome equation, and the relationship between the distributions of ability, GPA, and parental income, it can be shown that the optimal allocation is in $\mathcal{G}$. See Appendix A.3 for details.

Given a feasible class $\mathcal{G}$, we denote the highest attainable welfare by:

$$W^*_{\mathcal{G}} := \sup_{G \in \mathcal{G}} W(G) \ .$$

A *decision rule* is a function $\hat{G}$ from the observed data $\{(Y_i, D_i, X_i)\}_{i=1}^n$ into the set of admissible allocations $\mathcal{G}$. We call the rule that we develop and study in this paper the *Penalized Welfare Maximization* (or PWM) rule. As in much of the literature that follows the work of Manski (2004), we assume that the planner is interested in rules $\hat{G}$ that, on average, are close to the highest attainable welfare. To that end, the criterion by which we evaluate a decision rule is given by what we call maximum $\mathcal{G}$*-regret*:

$$\sup_P E_{P^n}[W^*_{\mathcal{G}} - W(\hat{G})] \ .$$

We note that, in contrast to many papers on statistical treatment rules which employ maximum-regret criteria, this notion of regret is defined relative to the optimum attained in $\mathcal{G}$, which is not necessarily the first-best unrestricted optimum (see Example 2.2.1). Kitagawa and Tetenov (2018) and Athey and Wager (2017) are recent papers which also focus on the $\mathcal{G}$-regret criterion.

2.3. Results

In this section, we present the main results of our paper. In Section 2.3.1, we review the properties of the empirical welfare maximization (EWM) rule of Kitagawa and Tetenov (2018), which will motivate the PWM rule and serve as an important building block in its construction. In Section 2.3.2, we define the penalized welfare maximization rule and present bounds on its maximum $\mathcal{G}$-regret for general penalties. In Section 2.3.3 we illustrate these results by applying them to some specific penalties. In Section 2.3.4 we present results for a

modification of the PWM rule for quasi-experimental settings where the propensity score is not known and must be estimated.

2.3.1. Empirical Welfare Maximization: a Review and Some Motivation

The idea behind the EWM rule is to solve a sample analog of the population welfare maximization problem:

$$\hat{G}_{EWM} \in \arg\max_{G \in \mathcal{G}} W_n(G) \ ,$$

where

$$W_n(G) := \frac{1}{n}\sum_{i=1}^{n} \tau_i \mathbf{1}\{X_i \in G\} := \frac{1}{n}\sum_{i=1}^{n} \Big[\Big(\frac{Y_i D_i}{e(X_i)} - \frac{Y_i(1-D_i)}{1-e(X_i)}\Big)\mathbf{1}\{X_i \in G\}\Big] \ . \tag{2.2}$$

In general this problem could be computationally challenging. However, Kitagawa and Tetenov (2018) show that solving such a problem is practically feasible for many applications by formulating it as a Mixed Integer Linear Program (MILP): see Appendix B.3 for details. Note that to solve this optimization problem, the planner must know the propensity score $e(\cdot)$. This assumption is reasonable if the data comes from a randomized experiment, but clearly could not be made in a setting where the planner is using observational data. Kitagawa and Tetenov (2018) derive results for a modified version of the EWM rule where the propensity score is estimated, which we will review in Section 2.3.4.

To derive their non-asymptotic bounds on the maximum $\mathcal{G}$-regret of the EWM rule, Kitagawa and Tetenov (2018) make the following additional assumptions, which we will also require for our results:

Assumption 2.3.1. *(Bounded Outcomes and Strict Overlap) The set of distributions $\mathcal{P}(M, \kappa)$ has the following properties:*

- *There exists some* $M < \infty$ *such that the support of the outcome variable* Y *is contained in* $[-\frac{M}{2}, \frac{M}{2}]$.
- *There exists some* $\kappa \in (0, 0.5)$ *such that* $e(x) \in [\kappa, 1-\kappa]$ *for all* x.

The first assumption asserts that the outcome is bounded. Since the implementation of the EWM rule or the PWM rule does not require that the planner knows M, and the existence of *some* bound on outcomes of interest to economics seems tenable (the assumption holds, for instance, if the outcome variable is binary), we view this assumption as mild. The second assumption ensures overlap in the covariate distributions, and is standard when imposing unconfoundedness. In an RCT, this assumption can be made to hold by design, but may be violated in settings with observational data.

In order to derive their results, Kitagawa and Tetenov (2018) also make the following assumption, which we will *not* require:

Assumption 2.3.2. *(Finite VC Dimension)*[1] *:* $\mathcal{G}$ *has finite VC dimension* $V < \infty$.

Such an assumption may or may not be restrictive depending on the application in question. Consider Example 2.2.2, the class of threshold allocations on $\mathbb{R}$. This class has VC dimension 2, thus Assumption 2.3.2 holds. On the other hand, it can be shown that the class of monotone allocations on $[0,1]^2$ that was introduced in Example 2.2.3 has infinite VC dimension (see Györfi et al. (1996)).

Given Assumptions 2.3.1 and 2.3.2, Kitagawa and Tetenov (2018) derive the following non-asymptotic upper bound on the maximum $\mathcal{G}$-regret of the EWM rule:

$$\sup_{P \in \mathcal{P}(M,\kappa)} E_{P^n}[W^*_{\mathcal{G}} - W(\hat{G}_{EWM})] \leqslant C\frac{M}{\kappa}\sqrt{\frac{V}{n}}\ , \tag{2.3}$$

[1]Their results can also be extended to settings where the class of treatment allocations has sufficiently small bracketing entropy (as in Tsybakov, 2004), or Hamming entropy (as in Athey and Wager, 2017). We will also not require these types of assumptions.

for some universal constant C. Moreover, when X has sufficiently "large" support, they derive the following *lower* bound: for *any* decision rule $\hat{G}$,

$$\sup_{P \in \mathcal{P}(M,\kappa)} E_{P^n}[W^*_{\mathcal{G}} - W(\hat{G})] \geqslant RM\sqrt{\frac{V-1}{n}}\,, \tag{2.4}$$

for R a universal constant and for all sufficiently large n. This shows that the rate of convergence of maximum $\mathcal{G}$-regret implied by (2.3) is the best possible, i.e. that no other decision rule could achieve a faster rate without imposing additional assumptions.

Remark 2.3.1. Theorem 2.2 in Kitagawa and Tetenov (2018), which establishes (2.4), has another interesting implication: if X has "large" support and we do not impose additional restrictions on the set of distributions $\mathcal{P}(M,\kappa)$, then it is *impossible* to derive a uniform rate of convergence of maximum $\mathcal{G}$-regret for *any* rule, for classes $\mathcal{G}$ of infinite VC dimension. This is in line with the results derived in Stoye (2009), where he shows that in a setting with a continuous covariate, and for any sample size, flipping a coin to assign individuals is minimax-regret optimal despite this rule not even being point-wise consistent. Since we will be interested in classes $\mathcal{G}$ of infinite VC dimension, we will revisit this problem later in Section 2.3. ■

Remark 2.3.2. As pointed out in Kitagawa and Tetenov (2018), the EWM rule is not invariant to positive affine transformations of the outcomes, and thus the researcher could manipulate the treatment rule in settings where they have leeway in how to code the outcome variable. To deal with this issue they suggest solving a demeaned version of the welfare maximization problem. In Appendix A.3 we discuss the demeaned version of EWM and repeat the exercises of Sections 2.4 and 2.5 using a demeaned version of EWM and PWM. ■

2.3.2. Penalized Welfare Maximization: General Results

We now consider a setting where the class $\mathcal{G}$ of admissible rules is "large", but can be "approximated" by a sequence of less complex subclasses $\mathcal{G}_k$:[2]

$$\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{G}_3 \subseteq \cdots \subseteq \mathcal{G}_k \subseteq \cdots \subseteq \mathcal{G} \ .$$

Let $\hat{G}_{n,k}$ be the EWM rule in the class $\mathcal{G}_k$. Then we can decompose the $\mathcal{G}$-regret of the rule $\hat{G}_{n,k}$ as follows:

$$E_{P^n}[W^*_{\mathcal{G}} - W(\hat{G}_{n,k})] = E_{P^n}[W^*_{\mathcal{G}_k} - W(\hat{G}_{n,k})] + W^*_{\mathcal{G}} - W^*_{\mathcal{G}_k} \ .$$

Given this decomposition, we call

$$E_{P^n}[W^*_{\mathcal{G}_k} - W(\hat{G}_{n,k})] \ ,$$

the *estimation* error of the rule $\hat{G}_{n,k}$ in the class $\mathcal{G}_k$, and we call

$$W^*_{\mathcal{G}} - W^*_{\mathcal{G}_k} \ ,$$

the *approximation* error (or bias) of the class $\mathcal{G}_k$. Note that since the classes $\{\mathcal{G}_k\}_k$ are nested, the estimation error (respectively approximation bias) is non-decreasing (resp. non-increasing) with respect to k. If one has sharp uniform bounds on these errors, then an appropriate choice of k would be one that minimizes the sum of these bounds. In Theorem 2.3.1, we derive an oracle inequality which shows that PWM selects such a k, in a data-driven fashion. We use this feature of PWM to derive bounds on maximum regret in two settings of empirical interest.

[2]As can be seen from the proofs, the results we present below remain valid even if the sequence $\{\mathcal{G}_k\}_k$ is not nested.

The first setting we consider is one where $\mathcal{G}$ has infinite VC dimension (consider Examples 2.2.1 and 2.2.3). In this setting, performing EWM on the whole class $\mathcal{G}$ may be undesirable (for example, the regret may not converge to zero). Instead, we apply EWM to an approximating class $\mathcal{G}_k$, and we allow the complexity of the approximating class to grow as the sample size increases. We present examples of relevant approximating classes in Examples 2.3.2 and 2.3.3 below. In Corollary 2.3.1 we establish a bound on maximum regret in this setting.

The second setting that we consider is one where the class $\mathcal{G}$ has finite but large VC dimension relative to the sample size. This situation can arise, for instance, in applications where the planner has a large set of covariates on which to base treatment, and where the feasible allocations are threshold allocations (see Example 2.3.1 below). The bound on regret given by (2.3) increases with the VC dimension V of $\mathcal{G}$, so that EWM tends to "overfit"' the data when V is large relative to the sample size. In such a situation, it may be beneficial to perform EWM in a class $\mathcal{G}'$ of smaller VC dimension, resulting in a smaller bound on the estimation error

$$E_{P^n}[W^*_{\mathcal{G}'} - W(\hat{G}_{EWM})] \ .$$

However, this will only be useful if it is also the case that

$$W^*_{\mathcal{G}} - W^*_{\mathcal{G}'} \ ,$$

is small. Hence we face the same tradeoff between estimation and approximation error that was noted above. In Corollary 2.3.2 we specialize Theorem 2.3.1 to a finite collection of approximating classes, and then in Corollary 2.3.3 establish a bound on maximum regret for the PWM rule which shows that it behaves as if we knew the correct class $\mathcal{G}'$ to use ex-ante, in the special case where the optimal allocation $G^* \in \mathcal{G}'$. We then apply these results to

select the number of covariates over which to perform best-subset selection with threshold allocations (see Example 2.3.1 below).

We consider the following assumption on our sequence of classes, which we call a *sieve* of $\mathcal{G}$:

Assumption 2.3.3. *The sequence of classes*

$$\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{G}_3 \subseteq \cdots \subseteq \mathcal{G}_k \subseteq \cdots \subseteq \mathcal{G}$$

is such that each class $\mathcal{G}_k$ has VC dimension V_k, which is finite.[3]

We illustrate this with some examples:

Example 2.3.1. Recall the class of threshold allocations introduced in Example 2.2.2. Let $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \subset \mathbb{R}^2$, and define $\mathcal{G}_X^1$ to be the threshold allocations on $\mathcal{X}_1$ and $\mathcal{G}_X^2$ to be the threshold allocations on $\mathcal{X}_2$. We can now define the set of *two-dimensional* threshold allocations on $\mathcal{X}$:

$$\mathcal{G} = \{G \subset \mathcal{X} : G = G_1 \times G_2,\ G_1 \in \mathcal{G}_X^1 \text{ and } G_2 \in \mathcal{G}_X^2\} .$$

To make this concrete, suppose that covariates X_1 and X_2 respectively denote age and income. Then $\mathcal{G}$ contains (for instance) allocations of the type: "receive treatment if age is above x_1 and income is below x_2" for some x_1 and x_2.

With K available covariates, it is straightforward to extend this definition to the class of K-dimensional threshold allocations. For large K, the VC dimension of $\mathcal{G}$ can become large relative to the sample size, and we may want to base treatment only on a smaller subset

[3]Kitagawa and Tetenov (2018) additionally assume that their class $\mathcal{G}$ is countable so as to avoid potential measurability concerns. We instead choose not to address these concerns explicitly, as is done in most of the literature on classification. See Van der Vaart and Wellner (1996) for a discussion of possible resolutions to this issue.

of the covariates. This is a variant of the best-subset selection problem, which has been recently studied in the classification context by Chen and Lee (2016). However, the question still remains as to *how many* covariates should be considered (that is, the size of the subset). An interesting sieve sequence for $\mathcal{G}$ is given by the following: let $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$ be defined as

$$\mathcal{G}_1 = \{\varnothing, \mathcal{X}\} \ , \ \mathcal{G}_2 = (\mathcal{G}_X^1 \otimes \mathcal{X}_2) \cup (\mathcal{X}_1 \otimes \mathcal{G}_X^2) \ , \ \mathcal{G}_3 = \mathcal{G} \ ,$$

where

$$\mathcal{G}_X^1 \otimes \mathcal{X}_2 := \{G \times \mathcal{X}_2 : G \in \mathcal{G}_X^1\}, \ \ \mathcal{X}_1 \otimes \mathcal{G}_X^2 := \{\mathcal{X}_1 \times G : G \in \mathcal{G}_X^2\} \ .$$

The sequence $\{\mathcal{G}_k\}_{k=1}^3$ corresponds to the sequence of threshold allocations that use zero, one and two covariates respectively (that each class $\mathcal{G}_k$ has finite VC dimension follows from the fact a class of threshold allocations in one dimension has finite VC dimension, and that unions of classes of finite VC dimension have finite VC dimension, see Dudley (1999))[4]. As we will illustrate below, PWM will determine in a data-driven way the number of covariates to use for treatment assignment. We will revisit this example in the simulation study of Section 2.4.

Example 2.3.2. Recall the class of monotone allocations introduced in Example 2.2.3. Suppose that $\mathcal{X} = [0,1]^2$, so that $\mathcal{G}$ has infinite VC dimension (see Györfi et al. (1996) for a proof of this fact). We will construct a useful sieve for $\mathcal{G}$, where we approximate sets in $\mathcal{G}$ with sets that feature monotone, piecewise-linear boundaries. We proceed in three steps.

[4]Note that in this example, it is actually the case that $\mathcal{G}_2$ and $\mathcal{G}_3$ have the same VC dimension. This will not be the case when we move to settings in higher dimensions.

First define, for T an integer and $0 \leqslant j \leqslant T$, the following function $\psi_{T,j} : [0,1] \to [0,1]$:

$$\psi_{T,j}(x) = \begin{cases} 1 - |Tx - j|, & x \in [\frac{j-1}{T}, \frac{j+1}{T}] \cap [0,1] \\ 0, & \text{otherwise .} \end{cases}$$

The function $\psi_{T,j}(\cdot)$ is simply a triangular kernel whose base shifts with j and is scaled by T. For example, $\psi_{4,1}(\cdot)$ is a triangular kernel with base $[0, 0.5]$, and $\psi_{8,1}(\cdot)$ is a triangular kernel with base $[0, 0.25]$. Next, using these functions, we define the following classes $\mathcal{S}_k$:

$$\mathcal{S}_k = \left\{ G : G = \{x = (x_1, x_2) \in \mathcal{X} | \sum_{j=0}^{T} \theta_j \psi_{T,j}(x_1) + x_2 \geqslant 0\} \text{ for } \theta_j \in \mathbb{R}, \forall\, 0 \leqslant j \leqslant T \right\},$$

where $T = 2^{k-1}$. These $\mathcal{S}_k$ are a special case of what Kitagawa and Tetenov (2018) call *generalized eligibility scores*, which, as shown in Dudley (1999), have VC dimension $T + 2$. The intuition behind the class $\mathcal{S}_k$ is that it divides the covariate space into treatment and control such that the boundary is a piecewise linear curve. Note that by construction it is the case that $\mathcal{S}_{k-1} \subset \mathcal{S}_k$ for every k. Finally, to construct our approximating class $\mathcal{G}_k$, we will modify the class $\mathcal{S}_k$ such that we ensure that the resulting treatment allocations are monotone.

For T an integer, let D_T be the following $T \times (T+1)$ *differentiation matrix*:

$$D_T := \begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix}.$$

Then $\mathcal{G}_k$ is defined as follows:

$$\mathcal{G}_k = \left\{ G : G \in \mathcal{S}_k \text{ and } D_T \Theta_T \geqslant 0 \ , \ \Theta_T = [\theta_0 \cdots \theta_T]' \right\} \ ,$$

for $T = 2^{k-1}$. Note that the purpose of the constraint $D_T \Theta_T \geqslant 0$ is to ensure that $\theta_k - \theta_{k-1} \geqslant 0$ for all k, which is what imposes monotonicity on the allocations. This construction, which we borrow from Beresteanu (2004), is useful as it imposes monotonicity through a *linear* constraint, which is ideal for our implementation of this sequence in Section 2.5. Proposition 2.5.1 provides a uniform rate at which $W^*_{\mathcal{G}_k} \to W^*_{\mathcal{G}}$ under some additional regularity conditions, and Corollary 2.5.1 derives the corresponding bound on maximum $\mathcal{G}$-regret of the PWM rule. It is important to mention that, under the regularity conditions we will impose, the class of monotone allocations is an example of a class for which bounds on maximum $\mathcal{G}$-regret exist for EWM, despite this class having infinite VC dimension (see Proposition 2.5.2). We will compare the bounds we derive for PWM to these bounds in the discussion following Corollary 2.5.1. In Section 2.5, we study the use of this sequence of approximating classes in an application to the JTPA study.

Example 2.3.3. Suppose the planner faces no restrictions on treatment assignment, so that $\mathcal{G}$ is the class of all measurable subsets of $\mathcal{X}$. Recall from Example 2.2.1 that that the optimal allocation in this case is given by G^*_{FB}. In this setting it may seem natural to employ the *plug-in* decision rule:

$$\hat{G}_{plug-in} := \{ x : \hat{\tau}(x) \geqslant 0 \} \ ,$$

where $\hat{\tau}(\cdot)$ is a non-parametric estimate of $\tau(\cdot)$. Under Assumption 2.2.1 many non-parametric estimates of $\tau(\cdot)$ are well understood. The Penalized Welfare Maximization Rule could provide an interesting alternative to plug-in rules in this setting by considering a sequence of classes made of *decision trees*. Decision trees are popular rules in classification because

of their natural interpretability. Intuitively, a decision tree recursively partitions the covariate space in such a way that the resulting decision rule can be understood as a series of "yes-or-no" questions involving the covariates. Using decision trees for the estimation of causal effects has recently become a popular idea in econometrics. Although we do not explore decision trees extensively in this paper, in Appendix A.3 we explain how we could accommodate them in our framework and relate them to recent work on the use of decision trees for treatment assignment, as presented in Kallus (2016) and Athey and Wager (2017). We also provide a preliminary comparison to plug-in decision rules.

Given a sieve $\{\mathcal{G}_k\}_k$, let

$$\hat{G}_{n,k} := \arg\max_{G \in \mathcal{G}_k} W_n(G) \ ,$$

be the EWM rule in the class $\mathcal{G}_k$. Our goal is to select the appropriate class k^* in which to perform EWM. We do this by selecting the class k^* in the following way: for each class $\mathcal{G}_k$, suppose we had some (potentially data dependent) measure $C_n(k)$ of the amount of "overfitting" that results from using the rule $\hat{G}_{n,k}$ (we will be more precise about the nature of $C_n(k)$ in a moment). Given such a measure $C_n(k)$, let $\{t_k\}_{k=1}^{\infty}$ be an increasing sequence of real numbers, and define the following penalized objective function:

$$R_{n,k}(G) := W_n(G) - C_n(k) - \sqrt{\frac{t_k}{n}} \ . \tag{2.5}$$

Then the *penalized welfare maximization* rule $\hat{G}_n$ is defined as follows:

$$\hat{G}_n := \hat{G}_{n,\hat{k}^*} \ ,$$

where

$$\hat{k}^* := \arg\max_k R_{n,k}(\hat{G}_{n,k}) \ .$$

In words, the PWM rule selects an allocation which maximizes a penalized version of the empirical welfare, with the penalty for allocations in $\mathcal{G}_k$ given by the term $C_n(k)$ (plus the auxiliary term $\sqrt{t_k/n}$).

Remark 2.3.3. Note that the PWM objective function $R_{n,k}(\cdot)$ includes the term: $\sqrt{t_k/n}$. This component of the objective is a technical device that is used to ensure that the classes get penalized at a sufficiently fast rate as k increases. The dependence of the penalty term on the sequence $\{t_k\}_k$ is somewhat undesirable, as it implies that the size of the penalty term for a given class depends on the location of the class in the sieve, as well as the specific choice of the sequence $\{t_k\}_k$. Ideally, we would like the penalty term to be completely determined by the class. This technical device seems−however−unavoidable, and similar terms are pervasive throughout the literature on model selection in classification: see Koltchinskii (2001), Bartlett et al. (2002), Boucheron et al. (2005), Koltchinskii (2008). We make three additional comments regarding this term. First, our results hold for *any* increasing sequence $\{t_k\}_{k=1}^{\infty}$, and the choice is reflected explicitly in the bounds that we derive. Second, if one is only interested in using PWM in settings where the sequence of classes is finite, then we will show in Corollary 2.3.2 that the $\sqrt{t_k/n}$ term is not required. Third, from our simulation results, PWM performs well for the choice $t_k = k$, and its performance is essentially unaffected by this term. For simplicity, and unless otherwise specified, we will present all of our results with this specific choice of $t_k = k$. ■

Remark 2.3.4. As noted by Kitagawa and Tetenov (2018), given a sieve $\{\mathcal{G}_k\}_k$, one can use their results to derive uniform (w.r.t $\mathcal{P}(M, \kappa)$) bounds on the estimation error. If one has in addition uniform bounds on the approximation bias, then one can consider the decision rule $\hat{G}_{n,k(n)}$, where $k(n)$ is constructed to minimize the sum of these bounds. However, the merit of such an approach would depend on obtaining "good" computable

bounds for the estimation and approximation error, which may be difficult to do in practice. For instance, the uniform bounds on the estimation error implied by the results of Kitagawa and Tetenov (2018) depend on the VC dimension of the classes $\{\mathcal{G}_k\}_k$ which may be hard to compute or bound in practice. Furthermore, such a deterministic choice of $k(n)$ may lead to suboptimal rates if the true DGP satisfies additional regularity conditions which may be unknown to the econometrician; for instance, if the true DGP belongs to a much smaller class $\mathcal{P}'(M,\kappa) \subseteq \mathcal{P}(M,\kappa)$, over which the approximation bias (uniformly) decays at a much faster rate than on $\mathcal{P}(M,\kappa)$, then the original choice of $k(n)$ will be suboptimal. Given these challenges, PWM displays two advantages. First, As shown in Theorem 2.3.1 and Corollary 2.3.1, PWM will perform−in a data-driven way−the optimal tradeoff between the approximation and estimation error, without relying on explicit bounds for these quantities. Second, PWM will select the subclass $\hat{k}$−over which to perform EWM−in a way that adapts to additional "regularities" that may be satisfied by the true DGP (see Corollary 2.3.1 below).

■

Before stating our main results about the $\mathcal{G}$-regret of the PWM rule, we first list some high-level assumptions on the penalty term $C_n(k)$. In Section 2.3.3, we will provide some specific examples of penalties that satisfy these assumptions.

Assumption 2.3.4. *There exist positive constants c_0 and c_1 such that $C_n(k)$ satisfies the following tail inequality for every n, k, and for every $\epsilon > 0$:*

$$\sup_{P \in \mathcal{P}(M,\kappa)} P^n(W_n(\hat{G}_{n,k}) - W(\hat{G}_{n,k}) - C_n(k) > \epsilon) \leqslant c_1 e^{-2c_0 n\epsilon^2} .$$

We provide some intuition for this assumption. Given an EWM rule $\hat{G}_{n,k}$, the value of the empirical welfare is given by $W_n(\hat{G}_{n,k})$. From the perspective of $\mathcal{G}$-regret, what we would really like to know is the value of population welfare $W(\hat{G}_{n,k})$. Although the latter

quantity is unknown, if we could define the (infeasible) penalty $C_n(k)$ as $W_n(\hat{G}_{n,k}) - W(\hat{G}_{n,k})$, then the penalized objective $W_n(\hat{G}_{n,k}) - C_n(k)$ would be exactly equal to $W(\hat{G}_{n,k})$. Since implementing such a $C_n(k)$ is impossible, our assumption requires for our feasible penalty to be a good (empirical) upper bound on $W_n(\hat{G}_{n,k}) - W(\hat{G}_{n,k})$. We are now ready to state our main workhorse result: an *oracle inequality* that chracterizes the $\mathcal{G}$-regret of the PWM rule.

Theorem 2.3.1. *Suppose that Assumptions 2.2.1, 2.3.1, 2.3.3 and 2.3.4 hold, and set $t_k = k$ in (2.5). Then there exist constants Δ and c_0 such that for every $P \in \mathcal{P}(M, \kappa)$:*

$$E_{P^n}[W^*_{\mathcal{G}} - W(\hat{G}_n)] \leqslant \inf_k \left[E_{P^n}[C_n(k)] + \left(W^*_{\mathcal{G}} - W^*_{\mathcal{G}_k}\right) + \sqrt{\frac{k}{n}} \right] + \sqrt{\frac{\log(\Delta e)}{2c_0 n}} \ .$$

Theorem 2.3.1 forms the basis of all the results we present in Sections 2.3.2 and 2.3.3. It says that, at least from the perspective of *pointwise* (as opposed to maximum) $\mathcal{G}$-regret, the PWM rule is able to balance the tradeoff between $E_{P^n}[C_n(k)]$ and the approximation error, at the cost of adding two additional terms that are $O(1/\sqrt{n})$. The relative importance of these terms is hard to quantify at this level of generality, and we will attempt to shed some light on them, for specific penalties, in Section 2.3.3. Note that this result does not *quite* accomplish our initial goal of balancing the estimation and approximation error along our sieve: it is possible to choose a $C_n(k)$ that satisfies Assumption 2.3.4 for which $E_{P^n}[C_n(k)]$ is too large a bound for the estimation error. For this reason, we also impose the requirement that any penalty we consider should have the following additional property:

Assumption 2.3.5. *There exists a positive constant C_1 such that, for every n, $C_n(k)$ satisfies*

$$\sup_{P \in \mathcal{P}(M,\kappa)} E_{P^n}[C_n(k)] \leqslant C_1 \sqrt{\frac{V_k}{n}} \ ,$$

where V_k is the VC dimension of $\mathcal{G}_k$.

This assumption ensures that $E_{P^n}[C_n(k)]$ is comparable to the estimation error for EWM derived in (2.3), which was shown to be rate-optimal in (2.4).

The next result we present is a bound on maximum regret for our first setting of interest: choosing the appropriate approximating class when $\mathcal{G}$ has infinite VC dimension. Note that, as discussed in Remark 2.3.1, a bound on maximum regret may not exist unless we impose some additional regularity conditions on the family of DGPs under consideration. Hence we make the additional assumption that we restrict ourselves to a set of distributions $\mathcal{P}_r$ for which there exists a uniform bound on the approximation error. Note however that we do not assume that the rate of decay of the approximation bias is necessarily known to the econometrician, thus illustrating the "oracle" nature of our results.:

Assumption 2.3.6. *Let $\mathcal{P}_r$ be a set of distributions such that*

$$\sup_{P \in \mathcal{P}_r} W^*_{\mathcal{G}} - W^*_{\mathcal{G}_k} = O(\gamma_k) \ ,$$

$$\sup_{P \in \mathcal{P}_r \cap \mathcal{P}(M,\kappa)} E_{P^n}[C_n(k)] = O(\zeta(k,n)) \ ,$$

for a sequence $\gamma_k \to 0$, and $\zeta(k,n)$ non-decreasing in k, $\zeta(k,n) \to 0$ as $n \to \infty$.

The first assumption asserts that we have a uniform bound on the approximation error. As we pointed out in Remark 2.3.1, an assumption of this type is necessary to derive a bound on maximum regret when the class $\mathcal{G}$ has infinite VC dimension. The second assumption is made to highlight the following possibility: although Assumption 2.3.5 guarantees that we can satisfy this restriction with $\zeta(k,n) = \sqrt{V_k/n}$, it is possible that, once we have imposed that P must lie in $\mathcal{P}_r$, an even *tighter* bound may exist on $C_n(k)$. We make this point to emphasize that PWM will balance the tradeoff between the estimation and approximation

error according to the *tighest* possible bounds on $E_P[C_n(K)]$ and $W^*_{\mathcal{G}} - W^*_{\mathcal{G}_k}$, regardless of whether or not we know these bounds for a given application.

Corollary 2.3.1. *Under Assumptions 2.2.1, 2.3.1, 2.3.3, 2.3.4, and 2.3.6, we have that*

$$\sup_{P \in \mathcal{P}_r \cap \mathcal{P}(M,\kappa)} E_{P^n}[W^*_{\mathcal{G}} - W(\hat{G}_n)] \leqslant \inf_k \left[O(\zeta(k,n)) + O(\gamma_k) + \sqrt{\frac{k}{n}} \right] + \sqrt{\frac{\log(\Delta e)}{2c_0 n}} \ .$$

As mentioned in Remark 2.3.4, if $\{\zeta(k,n)\}_{k,n}$ and $\{\gamma_k\}_k$ were known, then we could achieve such a result with a deterministic sequence $k(n)$. The strength of the PWM rule then is that we achieve the *same* behavior for any class $\mathcal{G}$ and approximating sequence $\{\mathcal{G}_k\}_k$ without having to know these quantities in practice. We will illustrate this result in our application Section 2.5, in the setting of Example 2.3.2.

The second Corollary we present specializes Theorem 2.3.1 to our second setting of interest: the appropriate selection of a subclass when the VC-dimension of $\mathcal{G}$ is finite and large (or comparable) in magnitude to the sample size (for example, when selecting amongst many covariates when performing best-subset selection). The result highlights two features of PWM. First, it shows that by balancing the trade-off between the approximation and estimation error, PWM can potentially lead to a reduction in regret (relative to EWM) for values of the sample size that are comparable in magnitude to the VC-dimension of $\mathcal{G}$. Second, it illustrates how our bound changes when the sieve is finite and we drop the auxiliary $\sqrt{k/n}$ component of our penalty.

Corollary 2.3.2. *Suppose that Assumptions 2.2.1, 2.3.1, 2.3.4, 2.3.5, and 2.3.3 hold, and that $\mathcal{G}_K = \mathcal{G}$ for some finite K. Furthemore, suppose that in our definition of the penalty*

we omit the term $\sqrt{k/n}$*. Then we have that*

$$E_{P^n}[W^*_{\mathcal{G}} - W(\hat{G}_n)] \leqslant \inf_{1 \leqslant k \leqslant K} \left[C_1 \sqrt{\frac{V_k}{n}} + \left(W^*_{\mathcal{G}} - W^*_{\mathcal{G}_k} \right) \right] + \sqrt{\frac{\log(K c_1 e)}{2 c_0 n}} \ .$$

Note that if the above bound is minimized at $k = K$, then the approximation error $W^*_{\mathcal{G}} - W^*_{\mathcal{G}_k}$ is zero and the resulting bound is comparable to the one derived in (2.3), with one additional term. In Section 2.3.3 we argue that for specific choices of the penalty term $C_n(k)$ this additional term is of smaller order than the $\sqrt{V_k/n}$ component of the bound.

Our final corollary of Section 2.3.2 considers the particular setting in which the constrained optimum $W^*_{\mathcal{G}}$ over the class $\mathcal{G}$ is *achieved* in $\mathcal{G}_{k_0}$, for some k_0, but that this class is unknown to the econometrician. The result shows that the resulting upper bound on maximum regret for PWM is as if we had performed EWM in the appropriate class $\mathcal{G}_{k_0}$.

Corollary 2.3.3. *Suppose that Assumptions 2.2.1, 2.3.1, 2.3.3, 2.3.4, and 2.3.5 hold, and let* $\mathcal{P}_k \subset \mathcal{P}(M, \kappa)$ *be the set of distributions such that* $G^* \in \mathcal{G}_k$*, then*

$$\sup_{P \in \mathcal{P}_k} E_{P^n}[W^*_{\mathcal{G}} - W(\hat{G}_n)] \leqslant C_1 \sqrt{\frac{V_k}{n}} + \sqrt{\frac{k}{n}} + \sqrt{\frac{\log(\Delta e)}{2 c_0 n}} \ .$$

Furthermore, if $\{\mathcal{G}_k\}_{k=1}^K$ *is finite, and we do not include the* $\sqrt{k/n}$ *term as discussed in Remark 2.3.3, then we have that:*

$$\sup_{P \in \mathcal{P}_k} E_{P^n}[W^*_{\mathcal{G}} - W(\hat{G}_n)] \leqslant C_1 \sqrt{\frac{V_{k*}}{n}} + \sqrt{\frac{\log(K c_1 e)}{2 c_0 n}} \ ,$$

where c_0*,* c_1 *are as in Assumption 2.3.4.*

2.3.3. Penalized Welfare Maximization: Some Examples of Penalties

This section serves two purposes. First, it illustrates the results of Section 2.3.2 with two concrete choices for the penalty $C_n(k)$. Second, the results help quantify the size of the auxiliary term in the bound of Theorem 2.3.1 for these penalties, so as to address the concerns presented in the discussion following Theorem 2.3.1. The first penalty we present, the Rademacher penalty, is theoretically elegant but computationally burdensome. The second penalty we present, the holdout penalty, is very intuitive and much more tractable in applications. However, the holdout penalty involves a sample-splitting procedure that some may find unappealing. Both of the penalties share the property that they do *not* require the practitioner to know the VC dimensions V_k of the approximating classes, which we feel is important to make the method broadly applicable.

2.3.3.1. The Rademacher Penalty. The first penalty we present is very attractive from a theoretical perspective, but is computationally burdensome. Let $S_n := \{(Y_i, D_i, X_i)\}_{i=1}^n$ be the observed data. Then the *Rademacher* penalty is given by

$$C_n(k) = E_\sigma\Big[\sup_{G \in \mathcal{G}_k} \frac{2}{n} \sum_{i=1}^n \sigma_i \tau_i \mathbf{1}\{X_i \in G\} \mid S_n \Big] ,$$

where τ_i is defined as in equation (2.2), and $\{\sigma_1, ..., \sigma_n\}$ are a sequence of i.i.d Rademacher variables, i.e. they take on the values $\{-1, 1\}$, each with probability half.

To clarify the origin of this penalty, recall that $C_n(k)$ must be a good upper bound on $W_n(\hat{G}_{n,k}) - W(\hat{G}_{n,k})$, which is the requirement of Assumption 2.3.4. Bounding such quantities is common in the study of empirical processes, and the usual first step is to use what is known as *symmetrization*, which gives the following bound:

$$E_{P^n}[\sup_{G \in \mathcal{G}} W_n(G) - W(G)] \leqslant E_{P^n}\Big[E_\sigma\big[\sup_{G \in \mathcal{G}} \frac{2}{n} \sum_{i=1}^n \sigma_i \tau_i \mathbf{1}\{X_i \in G\} \mid S_n \big] \Big] .$$

It is thus this inequality that inspires the definition of $C_n(k)$. The concept of Rademacher complexity[5] is pervasive throughout the statistical learning literature (see for example Koltchinskii (2001), Bartlett and Mendelson (2002), and Bartlett et al. (2002)). Intuitively, it measures a notion of complexity that is finer than that of VC dimension, and is at the same time computable from the data at hand. Furthermore, unlike the holdout penalty introduced in the next subsection, it allows both the objective function and the penalty to be estimated with all of the data.

Our first task is to prove that the conditions of Assumptions 2.3.4 and 2.3.5 hold for the Rademacher penalty:

Lemma 2.3.1. *Consider Assumptions 2.2.1, 2.3.1, 2.3.3. Let $C_n(k)$ be the Rademacher penalty as defined above. Then we have that*

$$P^n(W_n(\hat{G}_{n,k}) - W(\hat{G}_{n,k}) - C_n(k) > \epsilon) \leqslant \exp\Big(-2\Big(\frac{\kappa}{3M}\Big)^2 n\epsilon^2\Big),$$

and

$$E_{P^n}[C_n(k)] \leqslant C\frac{M}{\kappa}\sqrt{\frac{V_k}{n}}\ ,$$

where C is the same universal constant that appears in equation (2.3).

We are thus able to refine Theorem 2.3.1 to the case of the Rademacher penalty.

Proposition 2.3.1. *Consider Assumptions 2.2.1, 2.3.1, 2.3.3. Let $C_n(k)$ be the Rademacher penalty as defined above. Then we have that for every $P \in \mathcal{P}(M, \kappa)$:*

$$E_{P^n}[W^*_{\mathcal{G}} - W(\hat{G}_n)] \leqslant \inf_k \Big[E_{P^n}[C_n(k)] + \big(W^*_{\mathcal{G}} - W^*_{\mathcal{G}_k}\big) + \sqrt{\frac{k}{n}}\Big] + g(M,\kappa)\frac{M}{\kappa}\sqrt{\frac{1}{n}}\ ,$$

[5]Note that the definition of Rademacher complexity is slightly different than the definition of our penalty. Here we follow Bartlett et al. (2002) and do not include the absolute value in our definition of the penalty.

with $E_{P^n}[C_n(k)] \leqslant C\frac{M}{\kappa}\sqrt{\frac{V_k}{n}}$*, where* C *is the same universal constant as that in equation (2.3) and*

$$g(M,\kappa) := 6\sqrt{\log\left(\frac{3\sqrt{e}}{\sqrt{2}}\frac{M}{\kappa}\right)} \ .$$

Remark 2.3.5. We can now revisit the discussion following Theorem 2.3.1, about quantifying the size of the constants in the auxiliary term of the bound. In Appendix A.3 we perform a back-of-the-envelope calculation that provides insight into the size of $g(M,\kappa)$, and compares it to the size of the universal constant C derived in Kitagawa and Tetenov (2018). ■

Despite this penalty being theoretically appealing, implementing it in practical applications is problematic. The standard approach suggested in the statistical learning literature is to compute $C_n(k)$ by simulation: first, we repeatedly draw samples of $\{\sigma_i\}_{i=1}^n$, then we solve the problem

$$\max_{G\in\mathcal{G}_k} \frac{2}{n}\sum_{i=1}^{n} \sigma_i\tau_i \mathbf{1}\{X_i \in G\} \ ,$$

for each draw, and then average the result. Unfortunately, the optimization problem to be solved in the second step is computationally demanding for most classes $\mathcal{G}_k$ of interest, so that repeatedly solving it for multiple draws of $\{\sigma_i\}_{i=1}^n$ is impractical. Moreover, this procedure must be repeated for *each* class $\mathcal{G}_k$, which makes it even more prohibitive.

In the next section, we present a penalty that is not only conceptually very simple, but easy to implement as well.

2.3.3.2. The Holdout Penalty. The second penalty we introduce is motivated by the following idea: First fix some number $\ell \in (0,1)$ such that $m := n(1-\ell)$ (for expositional clarity suppose that m is an integer)[6], and let $r := n - m$. Given our original sample $S_n = \{(Y_i, D_i, X_i)\}_{i=1}^n$, let $S_n^E := \{(Y_i, D_i, X_i)\}_{i=1}^m$ denote what we call the *estimating* sample,

and let $S_n^T := \{(Y_i, D_i, X_i)\}_{i=m+1}^n$ denote the *testing* sample. Now, using S_n^E, compute $\hat{G}_{m,k}$ for each k. It seems intuitive that we could get a sense of the efficacy of $\hat{G}_{m,k}$ by applying this rule to the subsample S_n^T and computing the empirical welfare $W_r(\hat{G}_{m,k})$. We could then select the class k that results in the highest empirical welfare $W_r(\hat{G}_{m,k})$.

It turns out this idea can be formalized in our framework by treating it as a PWM-rule on the estimating sample, with the following penalty: for each EWM rule $\hat{G}_{m,k}$ estimated on S_n^E, let

$$W_m(\hat{G}_{m,k}) = \frac{1}{m} \sum_{i=1}^{m} \tau_i \mathbf{1}\{X_i \in \hat{G}_{m,k}\} \ ,$$

be the empirical welfare of the rule $\hat{G}_{m,k}$ on S_n^E and let

$$W_r(\hat{G}_{m,k}) = \frac{1}{r} \sum_{i=m+1}^{n} \tau_i \mathbf{1}\{X_i \in \hat{G}_{m,k}\} \ ,$$

be the empirical welfare of the rule $\hat{G}_{m,k}$ on S_n^T. We define the *holdout* penalty to be

$$C_m(k) := W_m(\hat{G}_{m,k}) - W_r(\hat{G}_{m,k}) \ .$$

Now, recall that the PWM rule is given by

$$\hat{G}_m = \arg\max_k \left[W_m(\hat{G}_{m,k}) - C_m(k) - \sqrt{\frac{k}{m}} \right] \ ,$$

which, given the definition of $C_m(k)$, simplifies to

$$\hat{G}_m = \arg\max_k \left[W_r(\hat{G}_{m,k}) - \sqrt{\frac{k}{m}} \right] \ .$$

Hence we see that the PWM rule with the holdout penalty reproduces the intuition presented above (with the usual addition of the $\sqrt{k/m}$ term; see Remark 2.3.3).

[6]The results would continue to hold if one were to instead define $m := \lfloor n(1-\ell) \rfloor$.

We check the conditions of Assumptions 2.3.4 and 2.3.5:

Lemma 2.3.2. *Assume Assumptions 2.2.1, 2.3.1, 2.3.3. Suppose we have a sample of size n and recall that $m = n(1-\ell)$ and $r = n - m$. Let $C_m(k)$ be the holdout penalty as defined above. Then we have that*

$$P^n(W_m(\hat{G}_{m,k}) - W(\hat{G}_{m,k}) - C_m(k) > \epsilon) \leqslant \exp\Big(-2\Big(\frac{\kappa}{M}\Big)^2 n\ell\epsilon^2\Big),$$

and

$$E_{P^n}[C_m(k)] \leqslant C\frac{M}{\kappa\sqrt{(1-\ell)}}\sqrt{\frac{V_k}{n}}\ ,$$

where C is the same universal constant that appears in equation (2.3).

With Lemma 2.3.2 established, Theorem 2.3.1 becomes:

Proposition 2.3.2. *Assume Assumptions 2.2.1, 2.3.1, 2.3.3. Suppose we have a sample of size n, and let $m = n(1-\ell)$, $r = n - m$. Let $C_m(k)$ be the holdout penalty as defined above. Then we have that for every $P \in \mathcal{P}(M,\kappa)$:*

$$E_{P^n}[W^*_{\mathcal{G}} - W(\hat{G}_m)] \leqslant \inf_k \Big[E_{P^n}[C_n(k)] + \big(W^*_{\mathcal{G}} - W^*_{\mathcal{G}_k}\big) + \sqrt{\frac{k}{n}}\Big] + g(M,\kappa,\ell)\frac{M}{\kappa\sqrt{\ell}}\sqrt{\frac{1}{n}}\ ,$$

with

$$E_{P^n}[C_n(k)] \leqslant C\frac{M}{\kappa\sqrt{(1-\ell)}}\sqrt{\frac{V_k}{n}}\ ,$$

where C is the same universal constant as that in equation (2.3) and

$$g(M,\kappa,\ell) := 2\sqrt{\log\Big(\sqrt{\frac{e}{2\ell}}\frac{M}{\kappa}\Big)}\ .$$

Remark 2.3.6. We can perform the same analysis as we did in Remark 2.3.5. In doing so we see that the difference between this result and the result in Proposition 2.3.1

is that sample-splitting introduces distortions into the constant terms through ℓ. Indeed, the tradeoff between splitting the sample into the estimating sample and testing sample is reflected in these constants. ■

As noted in Remark 2.3.6, the bound we derive for the holdout penalty is similar to what we derive for the Rademacher penalty, but with inflated constants. However, the benefit of the holdout penalty lies in the fact that it is much more practical to implement. The only remaining issue with the holdout penalty is how to split the data. Deriving some sort of data-driven procedure to choose the proportion ℓ is beyond the scope of our paper, but as a rule of thumb, we have found that it is much more important to focus on accurate estimation of the rule $\hat{G}_{m,k}$ than on the computation of $W_r(\hat{G}_{m,k})$. In other words, we recommend that the estimating sample S_n^E be a large proportion of the original sample S_n. Throughout Sections 2.4 and 2.5, we designate three quarters of the sample as the estimating sample.

2.3.4. Penalized Welfare Maximization: Estimated Propensity Score

In this section we present a modification of the PWM rule where the propensity score is not known and must be estimated from the data. This situation would arise if the planner had access to observational data instead of data from a randomized experiment. Before describing our modification of the PWM rule, we must review results about the corresponding modification of the EWM rule in Kitagawa and Tetenov (2018). The modification we consider here is what they call the *e-hybrid* EWM rule. Recall the EWM objective function as defined in equation (2.2). To define the e-hybrid EWM rule we modify this objective function by replacing τ_i with

$$\hat{\tau}_i := \Big[\frac{Y_i D_i}{\hat{e}(X_i)} - \frac{Y_i(1-D_i)}{1-\hat{e}(X_i)}\Big]\mathbf{1}\{\epsilon_n \leqslant \hat{e}(X_i) \leqslant 1-\epsilon_n\} \ ,$$

where $\hat{e}(\cdot)$ is an estimator of the propensity score, and ϵ_n is a trimming parameter such that $\epsilon_n = O(n^{-\alpha})$ for some $\alpha > 0$. The e-hybrid EWM objective function is defined as follows:

$$W_n^e(G) := \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}_i \mathbf{1}\{X_i \in G\} \ .$$

In a recent paper, Athey and Wager (2017) argue that more sophisticated estimators of the welfare objective can improve performance relative to the e-hybrid rule, and derive corresponding bounds on the maximum regret of their procedure that feature smaller constants. Modifying our method using their techniques would be an interesting direction for future work.

Since we are now estimating the propensity score, we must impose additional regularity conditions on P to guarantee a uniform rate of convergence. We make a high level assumption:

Assumption 2.3.7. *Given an estimator $\hat{e}(\cdot)$, let $\mathcal{P}_e$ be a class of data generating processes such that*

$$\sup_{P \in \mathcal{P}_e} E_{P^n}\Big[\frac{1}{n}\sum_{i=1}^{n} |\hat{\tau}_i - \tau_i|\Big] = O(\phi_n^{-1}) \ ,$$

where $\phi_n \to \infty$.

Although we do not explore low-level conditions that satisfy this assumption here, Kitagawa and Tetenov (2018) do so in their paper. To summarize their results, they show that if $\hat{e}(\cdot)$ is a local polynomial estimator, and that $e(\cdot)$ and the marginal distribution of X satisfy some smoothness conditions, then Assumption 2.3.7 is satisfied with $\phi_n = n^{-\frac{1}{n+d_x/\beta_e}}$, where β_e is a constant that determines the smoothness of $e(\cdot)$.[7]

Let $\hat{G}_{e-hybrid}$ be the solution to the e-hybrid problem in a class $\mathcal{G}$ of finite VC dimension, then Kitagawa and Tetenov (2018) derive the following bound on maximum $\mathcal{G}$-regret:

[7]To be more precise, β_e is the degree of the Holder class to which $e(\cdot)$ must belong.

$$\sup_{P\in\mathcal{P}_e\cap\mathcal{P}(M,\kappa)} E_{P^n}\Big[W^*_{\mathcal{G}} - W(\hat{G}_{e-hybrid})\Big] \leqslant O(\phi_n^{-1}\vee n^{-1/2}) \ . \tag{2.6}$$

With a non-parametric estimator of $e(\cdot)$, ϕ_n will generally be slower than $\sqrt{n}$ and hence determine the rate of convergence.

We are now ready to present the construction of the corresponding e-hybrid PWM estimator. Let $\mathcal{G}$ be an arbitrary class of allocations, and let $\{\mathcal{G}_k\}_k$ be some approximating sequence for $\mathcal{G}$. Let $\hat{G}^e_{n,k}$ be the hybrid EWM rule in the class $\mathcal{G}_k$. Let $C^e_n(k)$ be our penalty for the hybrid PWM rule. We now require that the penalty satisfies the following properties:

Assumption 2.3.8. *(Assumptions on $C^e_n(k)$)*
In addition to making assumptions about $C^e_n(k)$, we assume there exists an "infeasible penalty" $\tilde{C}_n(k)$ with the following properties:

- *There exist positive constants c_0 and c_1 such that $\tilde{C}_n(k)$ satisfies the following tail inequality for every n, k and for every $\epsilon > 0$:*

$$\sup_{P\in\mathcal{P}_e\cap\mathcal{P}(M,\kappa)} P^n(W_n(\hat{G}^e_{n,k}) - W(\hat{G}^e_{n,k}) - \tilde{C}_n(k) > \epsilon) \leqslant c_1 e^{-2c_0 n\epsilon^2}$$

- *There exists a positive constant C_1 such that, for every n, $\tilde{C}_n(k)$ satisfies*

$$\sup_{P\in\mathcal{P}_e\cap\mathcal{P}(M,\kappa)} E_{P^n}[\tilde{C}_n(k)] \leqslant C_1\sqrt{\frac{V_k}{n}} \ ,$$

  where V_k is the VC dimension of $\mathcal{G}_k$.

- *$\tilde{C}_n(k)$ and $C^e_n(k)$ are such that*

$$\sup_{P\in\mathcal{P}_e\cap\mathcal{P}(M,\kappa)} E_{P^n}\left[\sup_k\left|C^e_n(k) - \tilde{C}_n(k)\right|\right] = O(\phi_n^{-1}) \ .$$

We will provide context for these assumptions. First of all, included in the assumptions on $C_n^e(k)$ is the existence of an object $\tilde{C}_n(k)$ which we call an *infeasible penalty*. The first assumption asserts that the infeasible penalty obeys a similar tail inequality to $C_n(k)$, which was the penalty when the propensity score was known. The main difference is that $\tilde{C}_n(k)$ satisfies this assumption with respect to the *e-hybrid* EWM rule, and not the EWM rule with a known propensity. What is strange about this condition is that it is as if we were evaluating the hybrid rule through the empirical objective $W_n(\cdot)$, *which is the objective when the propensity score is known.* This is our motivation for calling $\tilde{C}_n(k)$ an infeasible penalty. Luckily, $\tilde{C}_n(k)$ is purely a theoretical device and does not serve a role in the actual implementation of PWM. We provide an example of such an infeasible penalty in the setting of the holdout penalty below.

The second assumption is the same as Assumption 2.3.5, but now with respect to the infeasible penalty $\tilde{C}_n(k)$. The third assumption simply links the true penalty $C_n^e(k)$ to the infeasible penalty $\tilde{C}_n(k)$ in such a way that both should agree asymptotically and do so at an appropriate rate.

Given this, we obtain the following analogue to Theorem 2.3.1:

Theorem 2.3.2. *Given assumptions 2.2.1, 2.3.1, 2.3.3, 2.3.7 and 2.3.8, there exist constants Δ and c_0 such that for every $P \in \mathcal{P}_e \cap \mathcal{P}(M, \kappa)$:*

$$E_{P^n}[W_{\mathcal{G}}^* - W(\hat{G}_n^e)] \leqslant \inf_k \left[E_{P^n}[\tilde{C}_n(k)] + \left(W_{\mathcal{G}}^* - W_{\mathcal{G}_k}^*\right) + \sqrt{\frac{k}{n}} \right] + O(\phi_n^{-1}) + \sqrt{\frac{\log(\Delta e)}{2c_0 n}} \ .$$

As we can see, the only difference between this bound and the bound derived in Theorem 2.3.1 is that there is an additional term of order ϕ_n^{-1}. This is also the case with the hybrid EWM estimator, as shown in (2.6).

Next, we check that the conditions in Assumption 2.3.8 are satisfied with modified versions of the holdout and Rademacher penalties. First we begin with the holdout penalty. Recall from Section 2.3.3 that the holdout method split the sample $S_n = \{(Y_i, D_i, X_i)\}_{i=1}^n$ into the estimating sample $S_n^E = \{(Y_i, D_i, X_i)\}_{i=1}^m$ of size $m = n(1-\ell)$ and the testing sample $S_n^T = \{(Y_i, D_i, X_i)\}_{i=m+1}^n$ of size $r = n - m$. The holdout penalty was then defined as

$$C_m(k) = W_m(\hat{G}_{m,k}) - W_r(\hat{G}_{m,k}) \ ,$$

where $W_m(\cdot)$ was the empirical welfare computed on S_n^E and $W_r(\cdot)$ was the empirical welfare computed on S_n^T.

To define the *hybrid holdout* penalty, let $\hat{e}^E(\cdot)$ be the propensity estimated on S_n^E, and let $\hat{e}^T(\cdot)$ be the propensity estimated on S_n^T. Define

$$W_m^e(G) := \frac{1}{m}\sum_{i=1}^m \hat{\tau}_i{}^E \mathbf{1}\{X_i \in G\} \ ,$$

where

$$\hat{\tau}_i{}^E = \Big[\frac{Y_i D_i}{\hat{e}^E(X_i)} - \frac{Y_i(1-D_i)}{1-\hat{e}^E(X_i)}\Big]\mathbf{1}\{\epsilon_n \leqslant \hat{e}^E(X_i) \leqslant 1-\epsilon_n\} \ .$$

Define $W_r^e(G)$ on the testing sample analogously. Letting $\hat{G}_{m,k}^e$ be the hybrid EWM rule computed on the estimating sample in the class $\mathcal{G}_k$, the hybrid holdout penalty is defined as:

$$C_m^e(k) := W_m^e(\hat{G}_{m,k}^e) - W_r^e(\hat{G}_{m,k}^e) \ .$$

We can now check the conditions of Assumption 2.3.8 for the hybrid holdout penalty. To do so, we must assert the existence of an infeasible penalty $\tilde{C}_m(k)$ that satisfies our assumptions. The infeasible penalty we consider is given by

$$\tilde{C}_m(k) := W_m(\hat{G}_{m,k}^e) - W_r(\hat{G}_{m,k}^e) \ ,$$

where $W_m(\cdot)$ and $W_r(\cdot)$ are defined as in Section 2.3.3, that is, they are computed *as if the propensity score were known*. We present the following lemma:

Lemma 2.3.3. *Assume Assumptions 2.2.1, 2.3.1, 2.3.3, and 2.3.7. Suppose we have a sample of size n and recall that $m = n(1-\ell)$ and $r = n-m$. Let $C_m^e(k)$ be the hybrid holdout penalty and $\tilde{C}_m(k)$ be the infeasible penalty as defined above. Then we have that*

$$P^n(W_m(\hat{G}^e_{m,k}) - W(\hat{G}^e_{m,k}) - \tilde{C}_m(k) > \epsilon) \leqslant \exp\Big(-2\Big(\frac{\kappa}{M}\Big)^2 n\ell\epsilon^2\Big),$$

$$E_{P^n}[\tilde{C}_m(k)] \leqslant C\frac{M}{\kappa\sqrt{(1-\ell)}}\sqrt{\frac{V_k}{n}}\ ,$$

and

$$\sup_{P\in\mathcal{P}_e} E_{P^n}\big[\sup_k |C_m^e(k) - \tilde{C}_m(k)|\big] = O(\phi_n^{-1})\ ,$$

where C is the same universal constant as that in equation (2.3).

We thus obtain an analogous result to Proposition 2.3.2 for PWM with the hybrid holdout penalty. Next we do the same thing for the Rademacher penalty. In fact, defining the hybrid version of the Rademacher penalty is relatively straightforward. Recall that the Rademacher penalty when the propensity score was known was defined as

$$C_n(k) = E_\sigma\Big[\sup_{G\in\mathcal{G}_k}\frac{2}{n}\sum_{i=1}^n \sigma_i\tau_i\mathbf{1}\{X_i\in G\} \mid S_n\Big]\ .$$

The *hybrid* Rademacher penalty is defined analogously:

$$C_n^e(k) = E_\sigma\Big[\sup_{G\in\mathcal{G}_k}\frac{2}{n}\sum_{i=1}^n \sigma_i\hat{\tau}_i\mathbf{1}\{X_i\in G\} \mid S_n\Big]\ .$$

To check the conditions of Assumption 2.3.8, the infeasible penalty we consider here is simply just the penalty when the propensity score is known, so that $\tilde{C}_n(k) = C_n(k)$. Hence we have the following lemma:

Lemma 2.3.4. *Assume Assumptions 2.2.1, 2.3.1, 2.3.3, and 2.3.7. Let $C_n^e(k)$ be the hybrid Rademacher penalty and $\tilde{C}_n(k)$ be the infeasible penalty as defined above. Then we have that*

$$P^n(W_n(\hat{G}^e_{n,k}) - W(\hat{G}^e_{n,k}) - \tilde{C}_n(k) > \epsilon) \leqslant \exp\Big(-2\Big(\frac{\kappa}{3M}\Big)^2 n\epsilon^2\Big),$$

$$E_{P^n}[\tilde{C}_n(k)] \leqslant C\frac{M}{\kappa}\sqrt{\frac{V_k}{n}}\ ,$$

and

$$\sup_{P\in\mathcal{P}_e} E_{P^n}\big[\sup_k |C_n^e(k) - \tilde{C}_n(k)|\big] = O(\phi_n^{-1})\ ,$$

where C is the same universal constant as that in equation (2.3).

Again, from this we obtain an analogous result to Proposition 2.3.1 for PWM with the hybrid Rademacher penalty.

2.4. A Simulation Study

In this section we perform a small simulation study to highlight the ability of the PWM rule to reduce $\mathcal{G}$-regret in an empirically relevant setting. We consider a situation where the planner has access to threshold-type allocations over five covariates, as described in Examples 2.2.2 and 2.3.1, and wishes to perform best-subset selection. The sieve sequence we consider is the same as in Example 2.3.1, where $\mathcal{G}_k$ is the set of threshold allocations on $k-1$ out of the 5 covariates. For example, $\mathcal{G}_1$ contains only the allocations $G = \varnothing$ and $G = \mathcal{X}$, which correspond to threshold allocations that use zero covariates, $\mathcal{G}_2$ contains all threshold allocations on one out of the five covariates, etc. We focus here on the setting

with five covariates for computational simplicity, but recent work by Chen and Lee (2016) suggests that solving this problem with ten or more covariates should be feasible in practice.

The problem that the planner faces is choosing how many covariates to use in the allocation: for example suppose that the distribution P is such that some of the available covariates are irrelevant for assigning treatment. Of course, the planner could perform EWM on all the covariates at once, and by the bound in equation (2.3) this is guaranteed to produce small regret in large enough samples. However, if the sample is not large, the planner may be able to achieve a reduction in regret by performing PWM. Through the lens of Corollary 2.3.3, our results say that PWM should behave *as if* we had performed EWM in the smallest class $\mathcal{G}_k$ that contains all of the relevant covariates.

To be concrete, we consider the following data generating process: Let $\mathcal{X} = [0,1]^5$, and

$$X_i = (X_{1i}, X_{2i}, ..., X_{5i}) \sim (U[0,1])^5 \ .$$

The potential outcomes for unit i are specified as:

$$Y_i(1) = 50(2X_{2i} - (1 - X_{1i})^4 - 0.5 + 0.5(X_{3i} - X_{4i})) + U_{1i} \ ,$$

$$Y_i(0) = 50(0.5(X_{3i} - X_{4i})) + U_{2i} \ ,$$

where U_1 and U_2 are distributed as $U[-20, 20]$ random variables which are independent of each other and of X. The covariates enter the potential outcomes in three different ways:

- X_{5i} is an irrelevant covariate; it does not play a role in determining potential outcomes at all.
- X_{3i} and X_{4i} affect both treatment and control equally; there will be a nonzero correlation between the observed outcome Y_i and these covariates, but they serve no purpose for treatment assignment.

- X_{1i} and X_{2i} *do* serve a purpose for assigning treatment, and both are used in the optimal threshold allocation. See Figure 2.1 below.

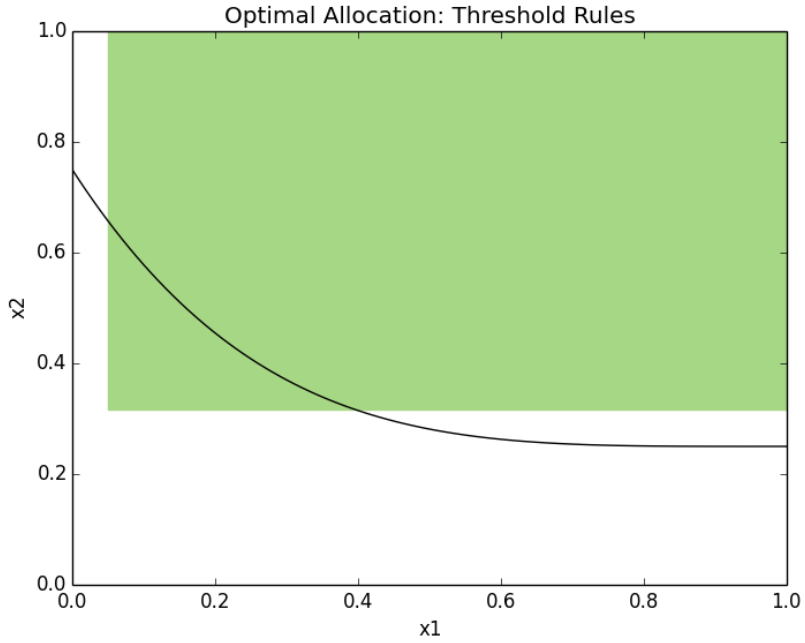

Figure 2.1. Shaded in green: the best threshold-allocation for our design.
Second-best welfare: 29.3
Traced in black: the boundary of the first-best allocation.

To implement PWM we used the holdout penalty, with 3/4 of our sample designated as the estimating sample. In Appendix B.3 we explain in detail how to implement PWM as a mixed integer linear program, and how we performed our simulations.

Our results compare the $\mathcal{G}$-regret of the PWM rule against the regret of performing EWM in $\mathcal{G}_6$ (which corresponds to the class that uses all five covariates) or performing EWM in $\mathcal{G}_3$. Recall that $\mathcal{G}_3$ is the smallest class that contains the optimal threshold allocation. In light of Corollary 2.3.3, we would hope that PWM behaves similarly to doing EWM in $\mathcal{G}_3$ directly. In Figure 2.2, we plot the regret of these rules for various sample sizes.

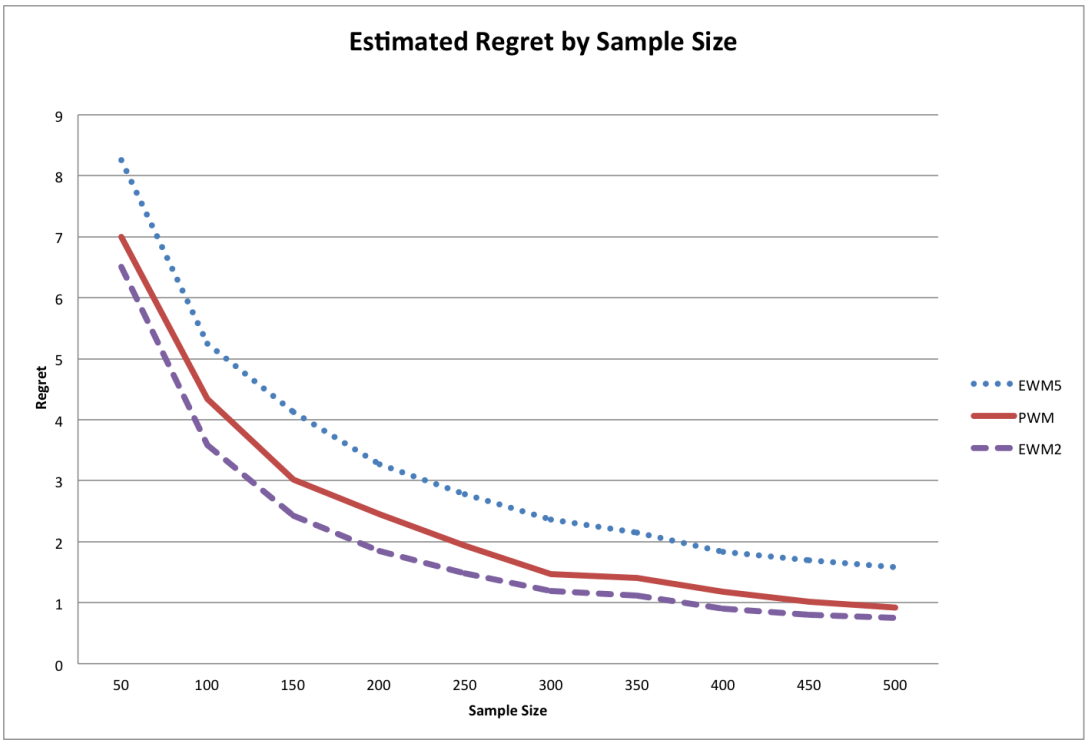

Figure 2.2. Estimated regret by sample size. Optimal (second-best) welfare: 29.3. EWM5 corresponds to $\mathcal{G}_6$ (five covariates), EWM2 corresponds to $\mathcal{G}_3$ (two covariates).

First we comment on the regret of performing EWM in $\mathcal{G}_6$ (recall that this corresponds to the set of allocations using all five covariates) vs. performing EWM in $\mathcal{G}_3$ (which corresponds to the set of allocations that use two of the five covariates). As predicted by equation (2.3), regret decreases as sample size increases. Moreover, performing EWM in $\mathcal{G}_6$ results in larger regret at every sample size: performing EWM in $\mathcal{G}_3$ results in a 6% improvement on average, across the sample sizes we consider.

Next, we comment on the performance of PWM. As we had hoped, the regret of PWM is smaller than the regret of performing EWM in $\mathcal{G}_6$ at every sample size: performing PWM results in a 4% improvement on average, across the sample sizes we consider. Moreover, the results in Figure 2.2 suggest that this gain is not just due to an improvement in very small samples, as the gap in regret seems to diminish quite slowly as sample size increases.

2.5. An Application

In this section we apply the PWM rule to experimental data from the Job Training Partnership Act (JTPA) Study. The JTPA study was a randomized controlled trial whose purpose was to measure the benefits and costs of employment and training programs. The study randomized whether applicants would be eligible to receive a collection of services provided by the JTPA related to job training, for a period of 18 months. The study collected background information about the applicants prior to the experiment, as well as data on applicants' earnings for 30 months following assignment (for a detailed description of the study, see Bloom et al. (1997)).[8]

We revisit the application setting of Kitagawa and Tetenov (2018). The outcome that we consider is total individual earnings in the 30 months following program assignment. The covariates on which we define our treatment allocations are the individual's years of education and their earnings in the year prior to the assignment. The set of allocations we consider is the set of monotone allocations defined in Example 2.2.3, but with a *non-increasing* monotone function. To be precise, let $\mathcal{X}_1$ be the covariate set of years of education, and let $\mathcal{X}_2$ be the covariate set of previous earnings, then the set of allocations we consider is given by:

$$\mathcal{G} = \left\{ G : G = \{(x_1, x_2) \in \mathcal{X} |\, x_2 \leqslant f(x_1) \text{ for } f : \mathcal{X}_1 \to \mathcal{X}_2 \text{ non-increasing}\} \right\} .$$

Let us discuss what this set of allocations means in the context of this application. This restriction imposes that, the less education you have, the more accessible is the program based on your previous earnings. For example, if an applicant with 12 years of education

[8]The sample we use is the same as that in Abadie et al. (2013), which we downloaded from ideas.repec.org/c/boc/bocode/s457801.html. We supplemented this dataset with education data from the `expbif.dta` dataset available at the W.E. Upjohn Institute website. Observations with years of education coded as '99' were dropped.

and previous earnings of $20,000 is to be accepted into the program, then an applicant with the same previous earnings and less education must also be accepted, as well as an applicant with the same level of education and less earnings. In Example 2.2.3 we discussed a situation where application-specific assumptions impose this type of constraint. In this setting, we instead argue that it is plausible that such a restriction may be exogenously imposed on the planner for political reasons; after all, it may not be politically viable to implement a job-training program where only those with high levels of education or income are accepted.

As we have previously discussed, this class of allocations will have infinite VC dimension when continuous covariates are used. Accordingly, in the results that follow, we will assume both covariates are continuous. However, note that in our application years of education is a discrete covariate. This discrepancy is not an issue for illustrating our method, and we think it is important that we make our study comparable to the one in Kitagawa and Tetenov (2018).

The approximating sequence we consider is the one described in Example 2.3.2, but now with a non-increasing monotonicity constraint. Recall that this was a sequence such that the resulting allocations partitioned the covariate space with a progressively refined, piecewise-linear, monotone boundary. Given any fixed class in this sequence, we can perform EWM in that class. For example, Figure 2.3 below illustrates the result of performing EWM on the simplest class in the approximating sequence. This class is equivalent to the class of linear treatment rules from Kitagawa and Tetenov (2018), but with an additional slope constraint.

At the other end of the spectrum, we could consider performing EWM in the most complicated class in our approximating sequence: this class corresponds to allocations that stipulate a threshold for previous income at every level of education (note that such a class

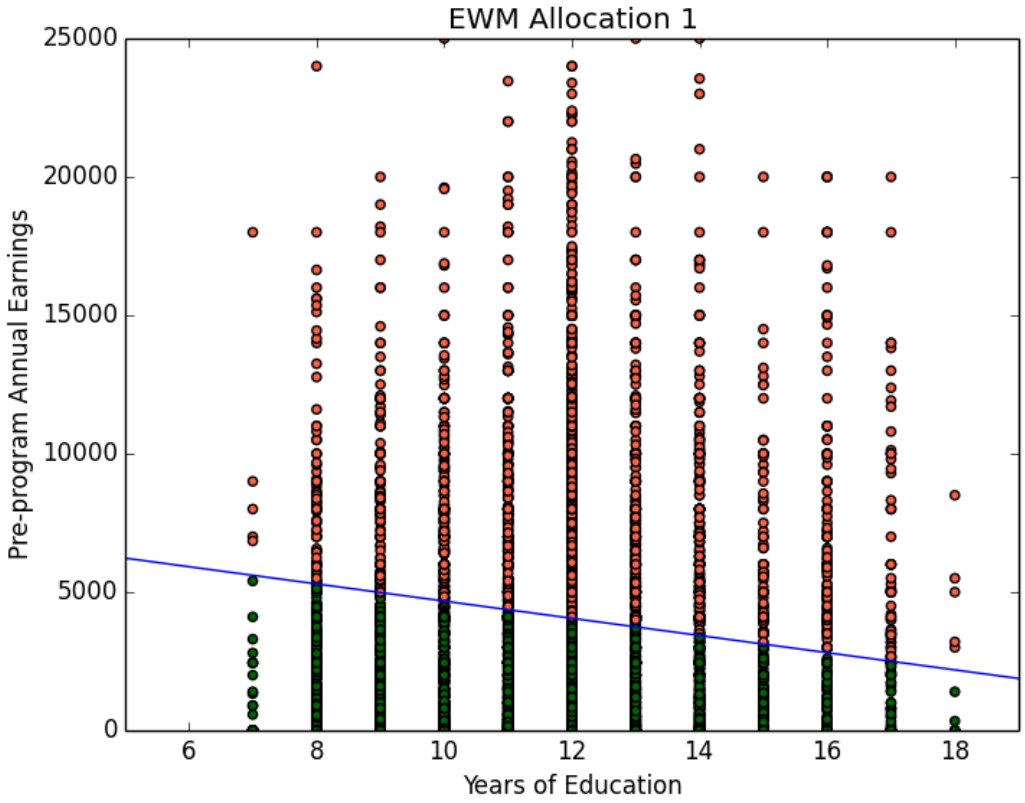

Figure 2.3. The resulting treatment allocation from performing EWM in $\mathcal{G}_1$. Each point represents a covariate pair in the sample. The region shaded in green (dark) is the prescribed treatment region, the region shaded in red (light) is the prescribed control region.

exists here because years of education is discrete). Figure 2.4 below illustrates the result of performing EWM in this class.

As we might expect, the resulting allocation in the simplest class and in the most complicated class are quite different, and given the option to choose any class from our sequence, it is not obvious which one should be chosen given the size of the experiment. Before showing the results for the PWM rule, recall from Remark 2.3.1 that, if the class $\mathcal{G}$ is has infinite VC dimension (as it would if both covariates were continuous), then we *cannot* establish a bound on maximum regret without imposing additional regularity conditions. Accordingly, we will

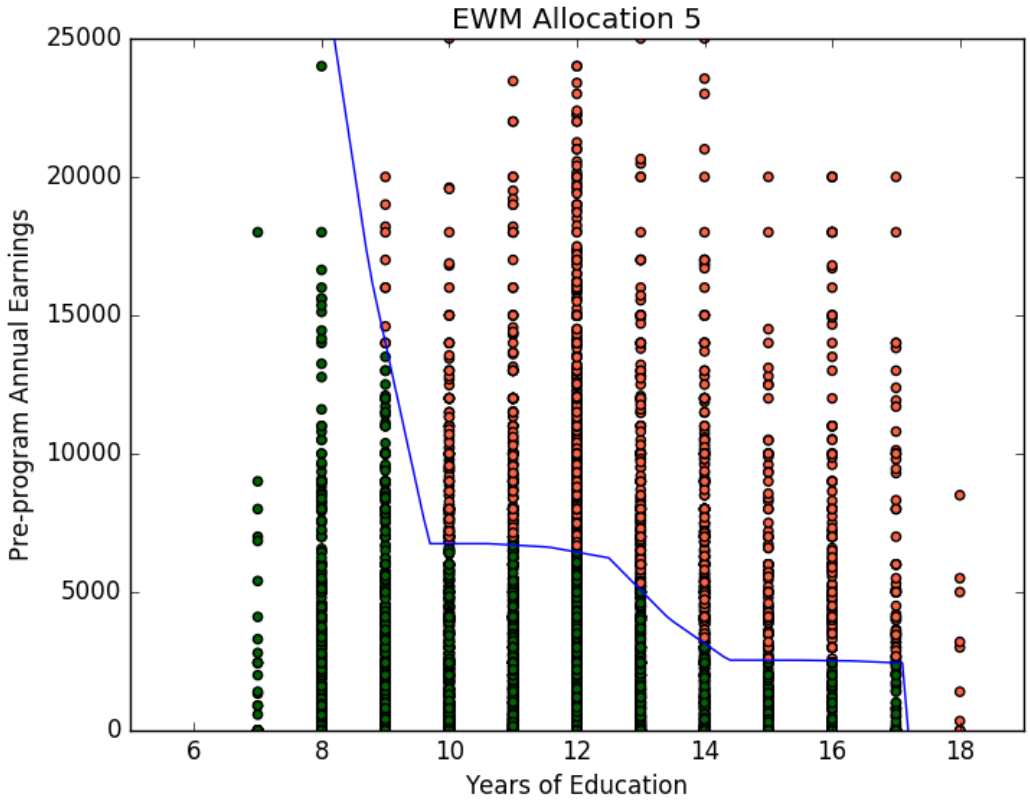

Figure 2.4. The resulting treatment allocation from performing EWM in $\mathcal{G}_5$. Each point represents a covariate pair in the sample. The region shaded in green (dark) is the prescribed treatment region, the region shaded in red (light) is the prescribed control region.

first establish a set of regularity conditions under which we derive a bound on maximum regret of the PWM rule.

We state the result for $\mathcal{X} = [0,1]^2$. We impose the following regularity condition on the distribution P:

Assumption 2.5.1. *Let $\mathcal{P}_r$ be a set of distributions such that there exists some constant A, where for every $P \in \mathcal{P}_r$, X has a density p_x with respect to Lebesgue measure on $[0,1]^2$ such that p_x is bounded above by A.*

It is worth emphasizing that we do *not* require the first best to be contained in $\mathcal{G}$, nor do we require that $W^*_{\mathcal{G}}$ even be attained. With this regularity condition imposed, we are able to derive the following uniform bound on the approximation bias $W^*_{\mathcal{G}} - W^*_{\mathcal{G}_k}$:

Proposition 2.5.1. *Under Assumption 2.5.1, the approximation bias of the approximating sequence $\{\mathcal{G}_k\}_{k=1}^{\infty}$ from Example 2.3.2 satisfies*

$$\sup_{P \in \mathcal{P}_r} W^*_{\mathcal{G}} - W^*_{\mathcal{G}_k} \leqslant A \frac{M}{\kappa} 2^{-k} \ ,$$

To illustrate the use of Proposition 2.5.1 in our setting, we derive a bound on maximum regret for monotone allocations. Proposition 2.5.1 and Corollary 2.3.1, along with the (possibly loose) bound on V_k given in Example 2.3.2 allow us to conclude that:

Corollary 2.5.1. *Let $C_n(k)$ be the Rademacher or holdout penalty. Under Assumptions 2.2.1, 2.3.1, 2.3.3, and 2.5.1, we have that*

$$\sup_{P \in \mathcal{P}_r \cap \mathcal{P}(M,\kappa)} E_{P^n}[W^*_{\mathcal{G}} - W(\hat{G}_n)] = O\big(n^{-\frac{1}{3}}\big) \ .$$

As we alluded to in the discussion of Example 2.3.2, bounds on maximum regret for EWM can be derived for the class of monotone allocations. Proposition 2.5.2 establishes such a bound by modifying the proof presented in Györfi et al. (1996) in the context of classification:

Proposition 2.5.2. *Under Assumptions 2.2.1, 2.3.1, and 2.5.1, we have that*

$$\sup_{P \in \mathcal{P}_r \cap \mathcal{P}(M,\kappa)} E_{P^n}[W^*_{\mathcal{G}} - W(\hat{G}_{EWM})] = O\big(n^{-\frac{1}{4}}\big) \ .$$

We make no claim that our bounds for PWM or EWM are sharp: for EWM, the most relevant results of which we are aware are presented in Tsybakov (2004), where he shows that

if the optimum is achieved in $\mathcal{G}$, and sufficient smoothness is imposed on the boundary of the optimal allocation, then the rate of convergence of the classification analogue of EWM is $O(n^{-1/2})$. Another relevant result from classification comes from Tsybakov and van de Geer (2005), where they develop a penalized method for classification over boundary fragments which is able to achieve a root-n rate (up to a logarithmic factor) for monotone allocations, while only assuming that the optimum is achieved. An interesting direction for future work would be to understand to what extent these techniques generalize to our setting, and also whether or not PWM is truly able to achieve a faster rate of convergence over EWM for this example under our assumptions.

In Figure 2.5, we illustrate the result of performing PWM on our sequence of classes, where we used 3/4 of our sample for estimation. In Appendix B.3 we discuss the computational details of our implementation. Note that PWM selects the allocation from the second class in our sequence, which corresponds to a piecewise-linear allocation with one allowable "kink".

2.6. Conclusion

In this paper, we introduced a new statistical decision rule for the treatment assignment problem, which we call the Penalized Welfare Maximization (PWM) rule. Our rule builds on the Empirical Welfare Maximization Rule of Kitagawa and Tetenov (2018), which is designed for situations where treatment allocation is exogenously constrained. The PWM rule is designed for settings where the policy maker may want to choose amongst a collection of such constrained classes. We established an oracle inequality for the regret of the PWM rule which shows that it is able to perform model selection over the collection of available classes. We then applied this result to two examples: the choice of the number of covariates when performing best-subset selection, and the selection of an approximating class in a sieve.

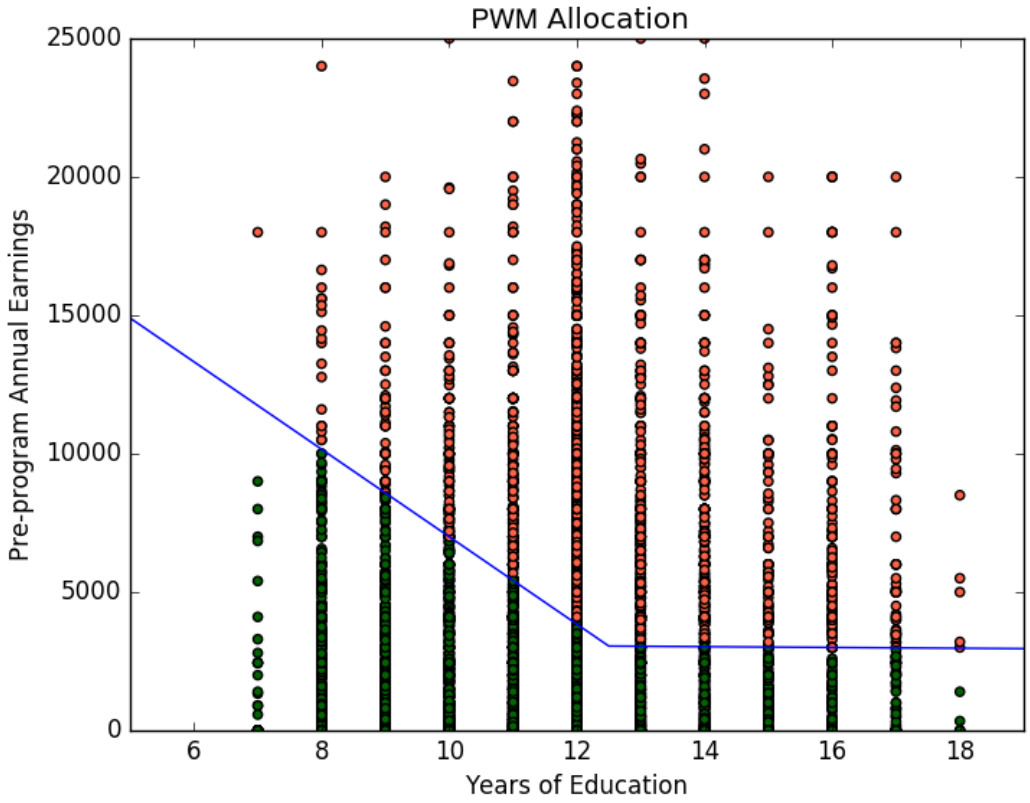

Figure 2.5. The resulting treatment allocation from performing PWM on the approximating sequence $\{\mathcal{G}_k\}_{k=1}^{5}$. Each point represents a covariate pair in the sample. The region shaded in green (dark) is the prescribed treatment region, the region shaded in red (light) is the prescribed control region.

Moving forward, we have identified some areas that we feel are worth further study. In general, implementing PWM is computationally challenging; from a practical perspective, practitioners may find it convenient to have a software package that can implement PWM in a few important examples. In particlar, decision/regression trees are becoming popular for the estimation of treatment effects, and as we illustrate in Appendix A.3 could serve as a useful approximating classes in our setting. We hope to further study the use of decision trees in the treatment assignment problem, as well as implement a software package that implements decision-tree based rules for practitioners.

CHAPTER 3

Inference with Dyadic Data: Asymptotic Behavior of the Dyadic Robust t-statistic

3.1. Introduction to Chapter 3

Over the last 25 years applied microeconomics has increasingly embraced the fact that dependence in cross-sectional data can affect inference. It has been well understood since at least the work of Moulton (1986) that failing to account for dependence in cross-sectional studies can have dire effects. In the past, researchers explicitly modeled such dependencies and used techniques such as GLS to estimate and do inference in their models. However, modern researchers are typically not satisfied with making such strong assumptions on the dependence present in the data. It is now standard practice to account for dependence by pairing standard test statistics with so-called "robust" variance estimators, analogous to the heteroskedasticity-robust variance estimator of White (1980).

In this paper we focus on inference for the regression parameters in a linear model with dyadic data. Dyadic data relates to pairs of objects; examples include data on trade between pairs of countries and data on links in a social-network setting. We will call such pairs "dyads" and the objects within them "units". Because of the paired nature of the data, dyads that share a unit in common could be correlated. In order to account for this potential dependence when conducting inference, we study the asymptotic properties of a t-statistic formed using a "robust" variance estimator known in the literature as the dyadic-robust variance estimator.

Fafchamps and Gubert (2007) were the first to propose the dyadic-robust variance estimator, under the following assumption: dyads that do *not* share a unit are uncorrelated, but otherwise the dependence between dyads is unspecified. To draw an analogy with cluster-robust inference (see Cameron and Miller, 2015, for an extensive survey), the dyadic-dependence assumption results in an "overlapping-cluster" configuration of the data, with each unit defining its own cluster. Subsequently, many applied papers in economics and political science have employed the dyadic-robust estimator under this same assumption (an incomplete list includes Aker, 2010; Baldwin and Jaimovich, 2012; Comola and Fafchamps, 2014; Echevarria and Gardeazabal, 2016; Egger and Tarlea, 2015; Leblang, 2010; Lustig and Richmond, 2017; Poznansky and Scroggs, 2016). Many empirical papers with dyadic data also make reference to the dependence assumption we describe, but then compute two-way clustered standard errors as described in Cameron et al. (2011). However, Cameron and Miller (2014) point out that this does not account for all the potential dependencies in the dyadic setting.

We present formal results under which a t-statistic that uses the dyadic-robust variance estimator is asymptotically normal. Using a central limit theorem for dependency graphs proved in Janson (1988) and careful bounding arguments, we establish a range of assumptions under which asymptotic normality holds. We then use our results to guide a simulation study of the accuracy of a normal approximation in finite samples, and discover an important setting where such an approximation is inadequate. With these insights, we propose a novel degrees of freedom correction to help alleviate the issue, and assess the performance of this correction in simulations.

Fafchamps and Gubert (2007) motivate the dyadic-robust variance estimator as an extension of the spatial HAC estimator of Conley (1999). Despite Conley's work being the

initial motivation, neither consistency of the dyadic-robust variance estimator nor asymptotic normality of the resulting t-statistic, under their maintained assumptions, follow from his results. Recently, Cameron and Miller (2014) have proposed the use of the dyadic-robust variance estimator in the analysis of trade data, and present simulation evidence to assess its performance. Both Fafchamps and Gubert (2007) and Cameron and Miller (2014) implicitly assume an asymptotic normality result for the dyadic-robust t-statistic in their analysis, but do not provide conditions under which such a result may hold. Aronow et al. (2015) prove the consistency of the dyadic-robust variance estimator for cross-sectional and panel data under more strict assumptions than those considered here, but do not attempt to study the use of this estimator for inference: although they derive the asymptotic variance of the t-statistic, they do *not* characterize its asymptotic distribution, specifically, they do not establish conditions under which the t-statistic is asymptotically normal. Our paper is the first to provide a theoretical grounding for the use of a normal approximation to the dyadic-robust t-statistic for inference in the linear model.

The remainder of the paper is organized as follows: In Section 3.2, we set up the model and the asymptotic frameworks we will study. Section 3.3 presents our results about asymptotic normality of the t-statistic. In Section 3.4 we study the finite-sample behavior of our approximation in a simulation study, and propose a degrees of freedom correction guided by our results. Section 3.5 concludes.

3.2. Setup of the Model and Asymptotic Frameworks

3.2.1. The Model

We will now formally describe the model. Consider a collection of G units indexed by $g = 1, ..., G$. The data we consider is indexed by pairs of units (g, h), which we call dyads. We do not require that each possible pair of units form a dyad. Pairs of units (g, h) for

which a dyad exists map into dyadic indices $n = 1, 2, ..., N$ through the index function $n(g, h)$, where for simplicity we make the assumption that $n(g, h) = n(h, g)$ (i.e. that we treat the dyads as non-directional), and the assumption that there are no elements of the type $n(g, g)$ (i.e. that we consider only pairs between distinct units). Given a dyadic index n, we also define the inverse correspondence ψ so that $\psi(n(g, h)) = \{g, h\}$. The model we consider is the linear model:

$$y_{n(g,h)} = \beta' \mathbf{x}_{n(g,h)} + u_{n(g,h)} \ , \tag{3.1}$$

where $\mathbf{x}_n$ is K-dimensional, with the standard conditions that $E[\mathbf{x}_n u_n] = \mathbf{0}$ and $E[\mathbf{x}_n \mathbf{x}_n'] > 0$. Our focus is on performing inference on the components of the regression parameter β.

Next, we present the dependence structure we will consider. Intuitively, we want observations that do not share a unit in common to be independent, but to allow correlation between observations otherwise. The typical assumption stated in the literature (see for example Aronow et al., 2015; Cameron and Miller, 2014) is that

$$E[u_n u_m | \mathbf{x}_n, \mathbf{x}_m] = 0, \ \text{ unless } \ \psi(n) \cap \psi(m) \neq \varnothing \ .$$

Although this assumption somewhat captures the intuition presented above, we will need to sharpen it considerably in order to prove formal results about our model. We impose the following dependence assumption on the data:

Assumption 3.2.1. *$\{(\mathbf{x}_n, u_n)\}_{n=1}^N$ are identically distributed. For any two disjoint subsets S_1, S_2 of $\{1, 2, ..., N\}$, $\{(\mathbf{x}_n, u_n)\}_{n \in S_1}$ is independent of $\{(\mathbf{x}_m, u_m)\}_{m \in S_2}$ if $\psi(n) \cap \psi(m) = \varnothing$ for every pair n, m of dyads such that $n \in S_1, m \in S_2$.*

Remark 3.2.1. Although we focus on the setting where our data is cross-sectional and the dyads are non-directional, our analysis covers settings with directional dyads, as well as panel-data with a *finite* number of time periods. For example, consider the following panel-data version of our model:

$$y_{(g,h)t} = \beta' \mathbf{x}_{(g,h)t} + \gamma_g + \gamma_h + \alpha_{gh} + u_{(g,h)t} \ ,$$

where we have explicitly indexed observations by their units, and now observations are indexed by pairs of units (g, h) as well as by time $t = 1, ..., T$. Note that we include γ_g and γ_h, which are unit-level fixed effects, as well as a dyad-level fixed effect α_{gh}. Let $\ddot{y}_{(g,h)t}$, $\ddot{\mathbf{x}}_{(g,h)t}$, and $\ddot{u}_{(g,h)t}$ denote the random variables that result from performing a *within* transformation:

$$\ddot{y}_{(g,h)t} := y_{(g,h)t} - \frac{1}{T} \sum_{s=1}^{T} y_{(g,h)s} \ ,$$

and similarly for $\mathbf{x}_{(g,h)t}$ and $\ddot{u}_{(g,h)t}$. Then the transformed model

$$\ddot{y}_{(g,h)} = \mathbf{x}_{(g,h)} \beta + \ddot{u}_{(g,h)} \ ,$$

where $\ddot{y}_{(g,h)}$ and $\ddot{u}_{(g,h)}$ are $T \times 1$ stacked vectors and $\mathbf{x}_{(g,h)}$ is a $T \times K$ stacked matrix, can be studied completely analogously to Model (3.1) above, with the assumptions $E[\mathbf{x}'_{(g,h)} \ddot{u}_{(g,h)}] = \mathbf{0}$ and $E[\mathbf{x}'_{(g,h)} \mathbf{x}_{(g,h)}] > 0$ (these assumptions are implied by standard primitive conditions on the original, untransformed model; see Wooldridge, 2010). The extension of our results to settings with growing T is more complicated and beyond the scope of this paper, as this would require additional assumptions on the nature of the dependence across time. ■

Let $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_K)'$ be the OLS estimator of β, that is

$$\hat{\beta} = \left(\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n' \right)^{-1} \sum_{n=1}^{N} \mathbf{x}_n y_n \ .$$

In this paper, we focus on the use of $\hat{\beta}$ as a means of forming a test-statistic to perform inference on β. To that end, we study the asymptotic distribution of the following root:

$$T_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{V}_{kk}}} \ ,$$

where $\hat{\beta}_k$ is the kth component of the OLS estimator of β and $\hat{V}_{kk}$ is the kkth entry of an appropriate estimator of its asymptotic variance. For a specific value β_{0k} of β_k we call the resulting statistic the *dyadic-robust t-statistic*. As mentioned in the introduction, the estimator of $\hat{V}$ we consider here is a sandwich estimator known in the literature as the dyadic-robust variance estimator:

$$\hat{V} = (\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n')^{-1} \Big(\sum_{n=1}^{N} \sum_{m=1}^{N} \mathbf{1}_{n,m} \hat{u}_n \hat{u}_m \mathbf{x}_n \mathbf{x}_m' \Big) (\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n')^{-1} \ ,$$

where $\hat{u}_n = y_n - \hat{\beta}' \mathbf{x}_n$ and $\mathbf{1}_{n,m}$ is an indicator function that equals 1 when $\psi(n) \cap \psi(m) \neq \varnothing$.

Our goal is to specify conditions under which T_k is asymptotically standard normal. As explained in the introduction, previous work on the dyadic-robust variance estimator either implicitly assumes an asymptotic normality result for T_k (Cameron and Miller, 2014; Fafchamps and Gubert, 2007), or does not explore or employ such a result (Aronow et al., 2015). The contribution of this paper is to provide general conditions under which T_k is asymptotically normal, and develop a degrees of freedom correction guided by our results.

3.2.2. A Key Condition for our Central Limit Theorem

To study the asymptotic distribution of T_k we will employ a central limit theorem for dependency graphs proved in Janson (1988). A key condition in the theorem, which we denote as Condition (2) in the appendix, plays a central role in our analysis. To simplify the exposition of our results, we introduce Condition 3.2.1 below, which is a modified, but equivalent, condition. Remark C.0.1 in the appendix establishes the equivalence of these two conditions.

For each unit g, let M_g be the number of dyads containing g. Recall that N is the total number of *dyads* in the data. Define

$$\mathcal{M}^H = \max_g M_g, \quad \mathcal{M}^L = \min_g M_g \ .$$

Note that by definition $\mathcal{M}^H \leqslant G-1$ and that $\frac{\mathcal{M}^L G}{2} \leqslant N \leqslant \frac{\mathcal{M}^H G}{2}$.

Condition 3.2.1. *Let $\mathcal{M}^H$ be as above. Let $\sigma_N^2 = Var(\sum_n \mathbf{x}_n u_n)$. Given some additional assumptions (see Theorem C.0.1), a sufficient condition for Janson's Theorem to apply in our framework is that*

$$\frac{(\frac{N}{\mathcal{M}^H})^{1/\ell}\mathcal{M}^H}{\sigma_N} \to 0 \ \ as \ \ N \to \infty \ ,$$

for some integer $\ell \geqslant 3$.

Intuitively, Janson shows that in our framework the expression above gives a bound on the higher-order cumulants of the sequence of random variables we will study, and that these cumulants vanishing is sufficient to establish asymptotic normality. A central theme of our paper is the fundamental connection Condition 3.2.1 creates between $\mathcal{M}^H$ and σ_N^2. This connection will be made more clear throughout the rest of Section 3.2.

Remark 3.2.2. It is important to emphasize that Condition 3.2.1 is simply a *sufficient* condition for asymptotic normality in our setting. Throughout Sections 3.2 and 3.3, we

use Condition 3.2.1 to motivate the assumptions we impose to ultimately prove our main asymptotic normality result (Theorem 3.3.1). In Section 3.4 we perform a simulation study that explores what can happen when Condition 3.2.1 does not hold. In Remark 3.3.2 we comment on how our results would change if we were to use alternative central limit theorems for dependency graphs. ■

3.2.3. Asymptotic Frameworks

We will now describe the two asymptotic frameworks that we consider in this paper. The first asymptotic framework we consider is one where $\mathcal{M}^H$ (and thus also $\mathcal{M}^L$) is bounded as $G \to \infty$. This framework is relevant in settings where units have few links. Model S in Figure 3.1 presents a configuration of the dyads in which we would expect such an approximation to be appropriate; note that there are 25 units (the grey nodes), but no units are contained in more than 6 dyads (the black edges). We will call this framework AF1:

Assumption 3.2.2. *(AF1)* $\mathcal{M}^H < D$ *for some constant* D *as* $G \to \infty$.

We will see in Section 3.3 that bounding $\mathcal{M}^H$ makes the analysis very clean, but AF1 may not be an appropriate framework for many settings of interest. In particular, Cameron and Miller (2014) suggest trade-data as an application where $\mathcal{M}^H$ could be very large relative to sample size. Model D in Figure 3.1 presents a configuration of the dyads such that every unit is contained in a dyad with every other unit, which is the configuration employed in all of the simulation results presented in Cameron and Miller (2014) and Aronow et al. (2015). Given such a configuration, we may not expect AF1 to provide a good approximation in this setting for large G, hence we will also study an asymptotic framework where we allow $\mathcal{M}^H \to \infty$ as $G \to \infty$. This setting is more complicated, and several subtle issues will arise.

The first issue that arises is that allowing $\mathcal{M}^H$ to grow as $G \to \infty$ now also frees $\mathcal{M}^L$ to grow as well. The most flexible possible framework would be one where we make no assumptions on $\mathcal{M}^L$ and simply require $\mathcal{M}^H \to \infty$. Unfortunately, in this case Janson's CLT cannot generally establish asymptotic normality when $\mathcal{M}^L$ is finite and fixed or grows too slowly. To illustrate, suppose that the error structure were of the form

$$u_{n(g,h)} = \alpha_g + \alpha_h + \epsilon_n \ ,$$

where α_g, α_h, ϵ_n are i.i.d. If $\mathcal{M}^L$ grows at rate $\log(G)$, $\mathcal{M}^H$ grows at rate G, and all but finitely many units form links at rate $\mathcal{M}^L$, then $Var(\sum_n \mathbf{x}_n u_n)$ would grow at rate G^2. It follows that Condition 3.2.1 would not be satisfied, so that Janson's CLT would not apply. The simulations of Section 3.4 will show that this has implications for inference when the number of dyads per unit varies wildly. In light of this, our second framework is given by:

Assumption 3.2.3. *(AF2)* $\mathcal{M}^L \geqslant cG$ *for some positive constant* c*.*

It is clear that this assumption is stronger than simply requiring that $\mathcal{M}^L \to \infty$, as it imposes an explicit rate of growth on $\mathcal{M}^L$ (and $\mathcal{M}^H$). It is possible to weaken this assumption slightly, but pinning down this rate clarifies our analysis and simplifies our notation.

Remark 3.2.3. It is crucial to note that, although this framework allows every unit to be linked together, this does *not* imply that every dyad can be correlated with every other dyad: consider the dense configuration of Model D presented in Figure 3.1. By construction, every dyad can be correlated with at most 46 other dyads, despite there being 300 dyads total. ■

Model B in Figure 3.1 highlights the kind of configuration of the dyads alluded to above that this framework may have trouble capturing: note that most of the units are only

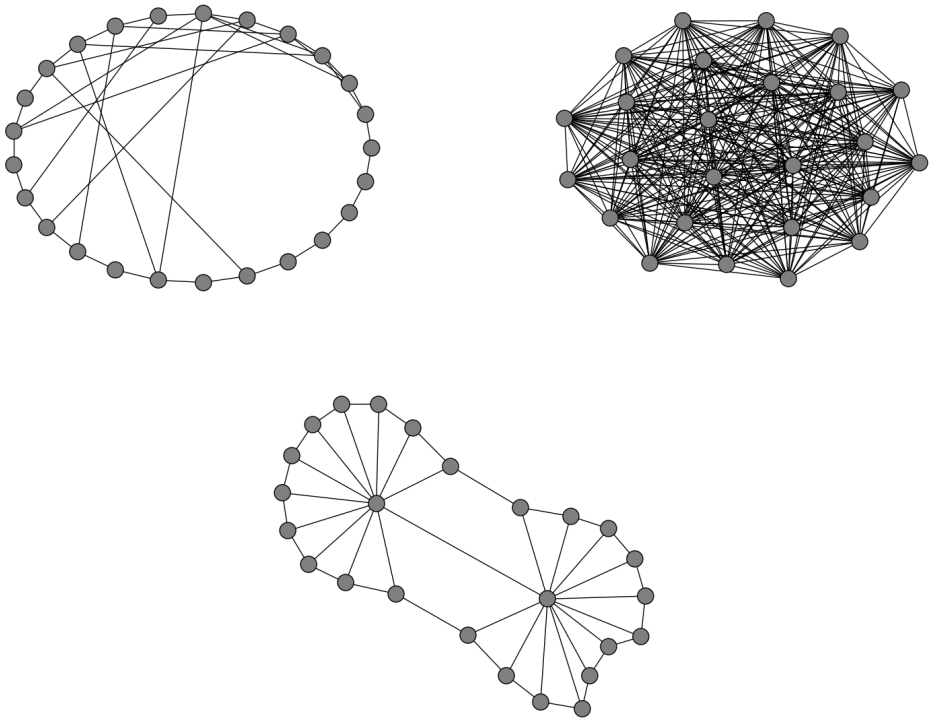

Figure 3.1. Clockwise from the top-left: Models S, D, and B with $G = 25$. Units are the grey nodes, dyads are the black edges.

contained in three dyads, but that two of the units are contained in many dyads. This type of configuration causes $\mathcal{M}^H$ to be very large relative to σ_N, which we have seen may cause Condition 3.2.1 to fail. Configurations of this type can arise in empirical settings with dyadic data: as an example, the application in Aronow et al. (2015) features a configuration in which most of the units are contained in approximately 10 dyads, but a handful of units are contained in upwards of 140 dyads. Model B will play an important role in the simulation study of Section 3.4, as well as in developing our proposed degrees of freedom adjustment.

3.2.4. The Rate of Growth of $Var(\sum_n \mathbf{x}_n u_n)$

We have already seen that the connection between the asymptotic framework and the rate of growth of $Var(\sum_n \mathbf{x}_n u_n)$ plays an important role in establishing Condition 3.2.1. In this section we will highlight another important observation concerning the rate of growth of the variance that is essential to the results presented in Section 3.3.

First consider AF1. We see immediately that imposing AF1 results in

$$Var(\sum_{n=1}^{N} \mathbf{x}_n u_n) = \sum_{n=1}^{N}\sum_{m=1}^{N} Cov(\mathbf{x}_n u_n, \mathbf{x}_m u_m)$$

growing at rate N. Hence under AF1 we will make the following assumption on the rate of growth:

Assumption 3.2.4.

$$\Omega := \lim_{G\to\infty} \frac{1}{N} Var(\sum_{n=1}^{N} \mathbf{x}_n u_n) = \lim_{G\to\infty} \frac{1}{N} \sum_{n=1}^{N}\sum_{m=1}^{N} Cov(\mathbf{x}_n u_n, \mathbf{x}_m u_m) \text{ is positive definite .}$$

We have seen in Section 3.2 that under AF2, when we allow $\mathcal{M}^H$ and $\mathcal{M}^L$ to grow as $G \to \infty$, the analysis can become more complicated. To clarify the exposition, we first consider a straightforward assumption that could be imposed under AF2:

Assumption 3.2.5.

$$\Omega := \lim_{G\to\infty} \frac{1}{NG} Var(\sum_{n=1}^{N} \mathbf{x}_n u_n) = \lim_{G\to\infty} \frac{1}{NG} \sum_{n=1}^{N}\sum_{m=1}^{N} Cov(\mathbf{x}_n u_n, \mathbf{x}_m u_m) \text{ is positive definite .}$$

This assumption is completely analogous to the standard assumption made on the variance in clustered data with strong dependence, and is implicitly the assumption made in all of the results in Aronow et al. (2015). Note that this assumption holds under an additive

common shocks error specification:

$$u_{n(g,h)} = \alpha_g + \alpha_h + \epsilon_n \ ,$$

where α_g, α_h, ϵ_n are i.i.d, which is a specification considered in the simulations of both Cameron and Miller (2014) and Aronow et al. (2015). More generally, Assumption 3.2.5 would be appropriate in applications where the dependence results in a positive correlation between dyads.

It is important to point out, however, that AF2 does not *imply* the rate of growth for $Var(\sum_n \mathbf{x}_n u_n)$ given in Assumption 3.2.5. For example, if instead the data were simply $i.i.d$, then $Var(\sum_n \mathbf{x}_n u_n)$ would grow at rate N since $Cov(\mathbf{x}_n u_n, \mathbf{x}_m u_m) = 0$ for $n \neq m$. In Section 3.4 we also consider a more interesting example: suppose we divide the units in the data into two groups, which we'll call G_A and G_B, and specify the error term as

$$u_{n(g,h)} = \begin{cases} -(\alpha_g + \alpha_h) + \epsilon_n & \text{if } g \text{ and } h \text{ belong to different groups} \\ \alpha_g + \alpha_h + \epsilon_n & \text{if } g \text{ and } h \text{ belong to the same group} \end{cases}$$

then by controlling the relative sizes of G_A and G_B, we can achieve growth rates of the form NG^r for any $r \in [0, 1]$ in Model D while still maintaining the maximal amount of dependence (see the appendix for details). Such an error structure is a stylized example of a situation where a shock to a unit could affect certain dyadic relations positively, while affecting others negatively. These examples highlight the fact that Assumptions 3.2.1 and AF2 can accommodate many plausible rates of growth of the variance. Since the goal of inference using robust-variance estimators is to be as agnostic as possible about the specific dependence structure in the data, we will consider these possibilities in our analysis. Our second assumption on the rate of growth of the variance under AF2 is given by:

Assumption 3.2.6.

$$\Omega := \lim_{G\to\infty} \frac{1}{NG^r} Var(\sum_{n=1}^{N} \mathbf{x}_n u_n) = \lim_{G\to\infty} \frac{1}{NG^r} \sum_{n=1}^{N}\sum_{m=1}^{N} Cov(\mathbf{x}_n u_n, \mathbf{x}_m u_m) \text{ is positive definite },$$

for some $r \in [0, 1]$.

Note that Assumption 3.2.5 is a special case of Assumption 3.2.6 with $r = 1$.

3.3. Asymptotic Properties of T_k

Recall that our goal is to study the asymptotic distribution of

$$T_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\widehat{V}_{kk}}} .$$

Although Fafchamps and Gubert (2007) motivate the construction of T_k by citing the work of Conley (1999), asymptotic normality under our maintained assumptions does not follow from his results. Theorem 3.3.1 contains our main result. Our development proceeds in two steps: our first set of results establish that

$$\tau_N(\hat{\beta} - \beta) \xrightarrow{d} N(0, V) ,$$

for some V to be defined later in the section. It will turn out that the rate τ_N depends on both the specific characteristics of the dependence and the asymptotic framework considered. Our second set of results establish under which additional conditions we get that

$$\tau_N^2 \hat{V} \xrightarrow{p} V .$$

By combining these two sets of results, we will see that T_k is asymptotically standard normal under a range of assumptions.

For all of the results that follow, we make the following support assumption:

Assumption 3.3.1. $\{(\mathbf{x}_n, u_n)\}_{n=1}^N$ *have uniformly bounded support for all* N*.*

Remark 3.3.1. We choose to present the results under a bounded support assumption for expositional simplicity, but this assumption can be weakened at the expense of adding extra conditions on the moments of $(\mathbf{x}, u)$, without altering the substantive conclusions of the paper. See the remarks after the proofs of Proposition 3.3.2 and Proposition 3.3.5 (Remarks C.0.2 and C.0.4, respectively) for further discussion. ■

3.3.1. Asymptotic Normality

Let us now study the asymptotic distribution of $\tau_N(\hat{\beta}-\beta)$ under both frameworks. Expanding the expression:

$$\tau_N(\hat{\beta} - \beta) = (\frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n\mathbf{x}_n')^{-1}\frac{\tau_N}{N}\sum_{n=1}^{N} \mathbf{x}_n u_n \ .$$

We show that the first term converges in probability to $E[\mathbf{x}_n\mathbf{x}_n']^{-1}$ by showing that the variance of each component of $(1/N)\sum_{n=1}^N \mathbf{x}_n\mathbf{x}_n'$ converges to zero. We show that the second term converges in distribution to a normal by Janson's Theorem. First we state the result under AF1:

Proposition 3.3.1. *Under Assumptions AF1, 3.2.1, 3.2.4, and 3.3.1,*

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V) \ ,$$

with $V = E(\mathbf{x}_n\mathbf{x}_n')^{-1}\Omega E(\mathbf{x}_n\mathbf{x}_n')^{-1}$*, and* Ω *as in Assumption 3.2.4.*

Note that under AF1 the rate τ_N is the standard $\sqrt{N}$. Intuitively, this is because AF1 imposes strong restrictions on the amount of dependence in the data.

Our asymptotic normality result under AF2 is:

Proposition 3.3.2. *Under Assumptions AF2, 3.2.1, 3.2.6 with $r > 0$ and 3.3.1,*

$$\tau_N(\hat{\beta} - \beta) \xrightarrow{d} N(0, V) \ ,$$

where $\tau_N = \sqrt{\frac{N}{G^r}}$, $V = E(\mathbf{x}_n \mathbf{x}_n')^{-1} \Omega E(\mathbf{x}_n \mathbf{x}_n')^{-1}$, and Ω is as in Assumption 3.2.6.

Note that the rate of convergence τ_N is now scaled by the growth-rate of $Var(\sum_n \mathbf{x}_n u_n)$; the smaller is r, the closer we get to the standard $\sqrt{N}$ rate. Our result explicitly excludes the case $r = 0$, this is because with $r = 0$ and all of our imposed assumptions, Condition 3.2.1 does not hold. In any case, the most likely situation in which r is exactly zero would be if the data were in fact i.i.d, and asymptotic normality then follows from many standard CLTs.

Although it may seem problematic at first that r will be unknown in practice, we will see in the following subsection that it is not necessary to know r precisely in order to do inference on β_k.

Remark 3.3.2. To prove Propositions 3.3.1 and 3.3.2 we employ a CLT for dependency graphs developed in Janson (1988). However, we could also consider using other results, such as the Berry-Esseen bounds for dependency graphs developed in, for example, Baldi and Rinott (1989), Penrose (2003), and Chen and Shao (2004). For our purposes it seems that Janson's CLT is as general (if not more general) than other available results: for example, to prove Proposition 3.3.2 by employing the results in Baldi and Rinott (1989) or Penrose (2003) would require us to assume that $r > 2/3$, and the results in Chen and Shao (2004) do not allow us to establish an asymptotic normality result under AF2 and our maintained assumptions. ■

3.3.2. Consistency of $\hat{V}$

We now turn to our second set of results: under what additional assumptions will the dyadic-robust variance estimator $\hat{V}$ converge to the asymptotic variance V? Recall that

$$\hat{V} = (\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n')^{-1} \hat{\Omega} (\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n')^{-1} ,$$

where

$$\hat{\Omega} = \Big(\sum_{n=1}^{N} \sum_{n'=1}^{N} \mathbf{1}_{n,n'} \hat{u}_n \hat{u}_{n'} \mathbf{x}_n \mathbf{x}_{n'}' \Big) .$$

To prove that $\tau_N^2 \hat{V} \xrightarrow{p} V$ we consider the "bread", $(\sum_n \mathbf{x}_n \mathbf{x}_n')^{-1}$, and the "meat", $\hat{\Omega}$, of the estimator separately. The convergence of $(1/N)\sum_n \mathbf{x}_n \mathbf{x}_n'$ to $E(\mathbf{x}_n \mathbf{x}_n')$ was proved when deriving the asymptotic distribution of $\tau_N(\hat{\beta} - \beta)$. To show that $(\tau_N/N)^2 \hat{\Omega} \xrightarrow{p} \Omega$, we show convergence in mean-square. As before, the result under AF1 is relatively straightforward:

Proposition 3.3.3. *Under Assumptions AF1, 3.2.1, 3.2.4, and 3.3.1, we have that*

$$N\hat{V} \xrightarrow{p} V .$$

The result under AF2 will again need more qualifications. First we present the result under Assumption 3.2.5:

Proposition 3.3.4. *Under Assumptions AF2, 3.2.1, 3.2.5, and 3.3.1, we have that*

$$(N/G)\hat{V} \xrightarrow{p} V .$$

Note that the rate has now slowed in accordance with the fact that the number of dependencies is growing for each unit. A special case of Proposition 3.3.4 for $\mathcal{M}^L = \mathcal{M}^H =$

$G-1$ appears in Aronow et al. (2015), whose proof generalizes to ours. Our next result generalizes Proposition 3.3.4 to the setting where instead we impose Assumption 3.2.6:

Proposition 3.3.5. *Under Assumptions AF2, 3.2.1, 3.2.6 with $r > 1/2$, and 3.3.1, we have that*

$$\tau_N^2 \hat{V} \xrightarrow{p} V \ ,$$

where τ_N is as in Proposition 3.3.2.

Note that we have restricted the rate of growth of $Var(\sum_n \mathbf{x}_n u_n)$ even more than in Proposition 3.3.2. We will provide some intuition as to why we require $r > 1/2$ in this result. In our proof we show convergence in mean-square, so heuristically we want to show that $Var((\tau^2/N^2)\hat{\Omega}) \to \mathbf{0}$. It is the case that the variance of $\hat{\Omega}$ depends not only on the covariances between observations, but on their dependence in general. Now, the larger is r, the slower the growth of τ_N, and hence the faster the convergence of τ^2/N^2 to zero. For small values of r, the convergence of τ^2/N^2 is too slow and cannot combat the growth in dependencies present in the data. This means that we cannot formally establish the consistency of $\hat{V}$ when the growth of $Var(\sum_n x_n u_n)$ is slow but the dependencies in the data are strong. We will study the implications of this via simulation in Section 3.4.

Remark 3.3.3. Note that Propositions 3.3.4 and 3.3.5 do not establish the consistency of $\hat{V}$ when the data are i.i.d. As an aside, we prove this result separately in Proposition 3.3.6 below. ■

Proposition 3.3.6. *Under Assumptions AF2, 3.3.1 and the assumption that $\{(\mathbf{x}_n, u_n)\}$ are i.i.d, we have that*

$$N\hat{V} \xrightarrow{p} V \ .$$

Remark 3.3.4. Cameron and Miller (2014) make the observation that there is no guarantee that $\hat{V}$ is positive semi-definite in finite samples. To account for this possibility they suggest the following modification: Consider the unitary decomposition $\hat{V} = U\Xi U'$ where $\Xi = diag[\lambda_1, ..., \lambda_k]$ is the diagonal matrix of eigenvalues of $\hat{V}$. Cameron and Miller (2014) suggest replacing $\hat{V}$ by

$$\hat{V}^+ = U\Xi^+ U' \ ,$$

where $\Xi^+ = diag[\lambda_1^+, ..., \lambda_k^+]$, $\lambda_j^+ := max(\lambda_j, 0)$ for all j. This modification guarantees that $\hat{V}^+$ is positive semi-definite. For the purposes of doing inference it would be even better if our estimator were positive-definite to guarantee that it is invertible. Politis (2011) suggests such a modification for a HAC estimator in a time-series setting, where $\widetilde{\lambda}_j^+ := max(\lambda_j, \epsilon_n)$ with $\epsilon_n \to 0$ at a suitable rate, and proves its consistency there. Additionally, Cameron and Miller (2014) suggest a simple finite sample correction for $\hat{V}^+$ which may help alleviate some of its finite-sample bias. ■

3.3.3. Asymptotic Normality of T_k

With all of these results in hand, we can now state the main result of the paper:

Theorem 3.3.1. *Under Assumptions 3.2.1, 3.3.1, and either:*

- *AF1 and Assumption 3.2.4,*
- *AF2 and Assumption 3.2.5,*
- *AF2 and Assumption 3.2.6 with* $r > \frac{1}{2}$,

we have that,

$$T_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{V}_{kk}}} \xrightarrow{d} N(0,1) \ .$$

Remark 3.3.5. With Theorem 3.3.1 in hand it is clear that a hypothesis test that uses the dyadic-robust t-statistic and the quantiles of a $N(0,1)$ distribution as critical values will be asymptotically valid. The same can be said for the coverage of an analogous confidence interval. ■

We want to highlight two appealing aspects of this result. The first is that the limit distribution is the *same* under both asymptotic frameworks. The second is that T_k features a type of "self-normalization" of τ_N. Specifically, although the rate of convergence of the estimator depends on which asymptotic framework we consider and which assumptions we impose on the dependence, precise knowledge of this rate is not required to construct T_k. In this respect our paper is philosophically similar to Hansen (2007), where he shows that the robust t-statistic in the linear panel model is asymptotically normal under various assumptions about the dependence across time. There is, however, one important caveat to our claim: under AF2 with Assumption 3.2.6, we only obtain the result when we impose that $r > 1/2$. This means that we have *not* established an asymptotic normality result when the configuration of the dyads is dense and the dependence features many positively and negatively correlated dyads. We treat the results under Assumption 3.2.6 as a form of robustness to small deviations from Assumption 3.2.5, which is the standard assumption imposed in the rest of the literature on strong dependence (i.e. clustering and its variations). In Section 3.4, we will use the insights we have gained from our formal analysis of the problem to conduct a simulation study under a broader range of designs than those considered previously in the literature.

Remark 3.3.6. Theorem 3.3.1 is sufficiently general to accommodate the theoretical setup considered in Aronow et al. (2015), as well as the simulation designs in Aronow et al.

(2015) and Cameron and Miller (2014). In their settings, $\mathcal{M}_H = \mathcal{M}_L = G - 1$, and $r = 1$, so that Theorem 3.3.1 applies under Assumptions AF2 and 3.2.5. ■

Remark 3.3.7. Cameron and Miller (2014) and Aronow et al. (2015) also consider the use of the dyadic-robust variance estimator in settings with M-estimators or GMM estimators. Our results can be extended to these settings under a set of "classical" regularity conditions (see Van der Vaart, 1998, Section 5.6) which include, for example, the logit or probit models considered in Cameron and Miller (2014). However, extending our results to M-estimation problems under more general regularity conditions (see for example Van der Vaart, 1998, Theorem 5.23) may require different tools, and hence different assumptions, than those considered in this paper. To this end, theoretical tools recently developed in Lee and Song (2017) look promising. ■

3.4. Simulation Evidence and a Degrees of Freedom Correction

In this section we perform a simulation-study to assess the accuracy of the normal approximation in finite samples, and the validity of the approximation for DGPs that do not satisfy the results presented in Section 3. In light of our results, we propose a novel degrees of freedom correction and study its performance via simulation as well. We use our formal results to guide the choice of designs. In an attempt to make the link between the simulations and our results as clear as possible, we consider simple designs.

The model we use throughout is a special case of what we have studied in Section 3:

$$y_n = 1 + \beta x_n + u_n \ ,$$

with x_n a scalar and $\beta = 0$. We consider two different specifications of x_n:

- When the u_n are i.i.d, $x_n \sim U[0,1]$ i.i.d, so that (x_n, u_n) are i.i.d.

- When u_n are not i.i.d, $x_{n(g,h)} = |z_g - z_h|$ where $z_g \sim U[0,1]$ i.i.d. for $g \in G$.

As pointed out in Remark 3.3.4, it is possible that $\hat{V}$ is not positive definite in finite samples, in particular when G is small. For the simulations we perform, we use the estimator $\widetilde{V}$ which is constructed by an analogous construction to $\hat{V}^+$ presented in Remark 3.3.4 but with $\tilde{\lambda}_j^+ = max(\lambda_j, \epsilon)$ for $\epsilon = 10^{-7}$. Whether or not we use $\widetilde{V}$, or $\hat{V}$ and drop those iterations where $\hat{V}$ is non-positive does not materially affect the results. In particular, for the sample sizes in which we get appropriate coverage, $\hat{V}$ and $\tilde{V}$ are always equal. Although we do not employ the finite sample corrections to $\hat{V}$ suggested in Cameron and Miller (2014), we comment on them briefly in Section 4.3 below.

3.4.1. Designs with Varying Levels of Sparsity

First we will study the normal approximation under varying levels of "sparsity" of the dyads relative to the number of units, while maintaining the standard variance Assumptions 3.2.4 and 3.2.5. We will consider three possible levels of sparsity:

- **Model D (dense)** Every possible dyad is present in the data.
- **Model S (sparse)** A design where $\mathcal{M}^H = 5$ and $\mathcal{M}^L = 2$ regardless of sample size.
- **Model B (both)** A design where $0.5\log(G) \leqslant \mathcal{M}^L \leqslant 1.5\log(G)$ and $0.5G \leqslant \mathcal{M}^H \leqslant G - 1$, such that $Var(\sum_n x_n u_n)$ grows slower than required in Proposition 3.3.2.

See the appendix for the specific construction of Models S and B. Figure 3.1 in Section 3.2 was generated using these designs for $G = 25$. Given the results presented in Section 3.3, we expect Models D and S to behave well in large samples since they satisfy the conditions of our theorem. Model B does not satisfy the conditions of our theorem since the rate of growth

of $\mathcal{M}^L$ is too slow. We saw in Section 3.2 that this has the potential to make $Var(\sum_n x_n u_n)$ grow too slowly for Janson's CLT to apply.

We will study these three models under two possible specifications of the error term. In the first specification, $u_n \sim U[-\sqrt{3}, \sqrt{3}]$ i.i.d, so that (x_n, u_n) are simply i.i.d. In the second specification we set:

$$u_{n(g,h)} = \alpha_g + \alpha_h + \epsilon_n \ ,$$

where $\alpha_g \sim U[-\sqrt{3}, \sqrt{3}]$ i.i.d for $g = 1, 2, ...G$ and $\epsilon_n \sim U[-\sqrt{3}, \sqrt{3}]$ i.i.d for $n = 1, 2, ..., N$. This is the additive common-shocks model discussed in Section 3.3.

We perform 10,000 Monte Carlo iterations and record the number of times a 95% confidence interval based on the normal approximation contains the true parameter $\beta = 0$. Table 1 presents the results under each specification. Recall that G is the number of units in the data.

		G				
	Specification	10	25	50	100	250
	i.i.d	69.8	85.1	91.3	93.2	94.4
Model D		(0.46)	(0.36)	(0.28)	(0.25)	(0.23)
	unit-level shock	63.7	86.2	92.1	93.6	94.3
		(0.48)	(0.34)	(0.27)	(0.24)	(0.23)
	i.i.d	70.0	84.8	90.1	92.8	94.5
Model S		(0.46)	(0.36)	(0.30)	(0.26)	(0.23)
	unit-level shock	66.3	82.8	90.2	92.9	94.0
		(0.47)	(0.38)	(0.30)	(0.26)	(0.24)
	i.i.d	69.6	82.4	86.9	89.3	93.1
Model B		(0.46)	(0.38)	(0.34)	(0.31)	(0.25)
	unit-level shock	65.4	81.0	84.5	86.1	89.4
		(0.48)	(0.39)	(0.36)	(0.35)	(0.31)

Table 3.1. Coverage percentages of a 95% CI for $\beta = 0$, simulation SEs in parentheses.

We note that for Models D and S, we get appropriate coverage in large samples. This is not surprising given our results, and is in line with the simulation results presented in Cameron and Miller (2014) and Aronow et al. (2015). It is interesting to note that, despite their similar performances, there is significantly less data present in the simulations for a given G under Model S than under Model D. For example, at $G = 50$ there are 1225 data points under Model D but only 86 data points under Model S. The fact that the approximations are comparable suggests that estimating the variance is much harder when the dyads are dense relative to the number of units. This makes sense seeing as the number of potential dependencies grows quadratically in Model D but is fixed in Model S.

Given that it is the number of units that determines the accuracy of the approximation, it is also clear that we require many units to get adequate coverage. In many settings of interest (for example, social networks) the number of units required should not be insurmountable. In other settings such as international trade, the number of units required could be prohibitive.

The final important observation we make is that the true coverage under Model B with unit-shocks, which was designed to fail the conditions of our theorem, is consistently worse than under either Models D and S, even when G is relatively large. In fact, even with G as large as 800 (not formally reported) we do not see coverage higher than 92% using this design. A similar phenomenon has been documented for the linear model with one-way clustering when cluster sizes differ "wildly", in MacKinnon and Webb (2017) and Carter et al. (2017). Because configurations in the spirit of Model B do arise in applications (for example, the empirical application in Aronow et al., 2015), the degrees of freedom correction we propose in the next section is specifically designed to address this issue.

3.4.2. Designs with Varying Growth Rates of $Var(\sum_n x_n u_n)$

In this section we study the accuracy of a normal approximation under Assumption 3.2.6, which allows varying rates of growth of $Var(\sum_n x_n u_n)$. We noted in Section 3.3 that even though we cannot expect to know the exact rate in practice, using T_k to test hypotheses does not require this knowledge. However, we also saw in Section 3.3 that when the model is dense, Proposition 3.3.5 does not establish consistency of $\hat{V}$ if the data is such that the growth rate of $Var(\sum_n x_n u_n)$ is too slow while still having many dependencies, which would be the case if the dependence features many positively and negatively correlated dyads.

Throughout this section we consider Models D and S. Our specification is constructed as follows: we divide the units in the data into two groups, which we call G_A and G_B, and specify the error term as

$$u_{n(g,h)} = \begin{cases} -(\alpha_g + \alpha_h) + \epsilon_n & \text{if } g \text{ and } h \text{ belong to different groups.} \\ \alpha_g + \alpha_h + \epsilon_n & \text{if } g \text{ and } h \text{ belong to the same group.} \end{cases}$$

Where $\alpha_g \sim U[-\sqrt{3}, \sqrt{3}]$ i.i.d for $g = 1, 2, ...G$ and $\epsilon_n \sim U[-\sqrt{3}, \sqrt{3}]$ i.i.d for $n = 1, 2, ..., N$.

By controlling the relative sizes of G_A and G_B, we can achieve growth rates of the form NG^r for any $r \in [0, 1]$ in Model D while still maintaining the maximal amount of dependence (see the appendix for details). Although it is clear that this design is artificial, we think it is reasonable to consider situations where shocks at the unit-level can have differing effects across dyads.

As before, we perform 10,000 Monte Carlo iterations and record the number of times a 95% confidence interval based on the normal approximation contains the true parameter $\beta = 0$. Table 2 presents the results for varying levels of r.

	r	G				
		10	25	50	100	250
	0	65.5	75.9	80.4	83.6	84.0
		(0.48)	(0.43)	(0.40)	(0.37)	(0.37)
	0.25	65.2	77.3	82.8	86.7	89.8
		(0.48)	(0.42)	(0.38)	(0.34)	(0.30)
Model D	0.5	65.2	80.1	87.3	91.5	94.7
		(0.48)	(0.40)	(0.33)	(0.28)	(0.22)
	0.75	65.3	82.6	91.1	93.9	94.4
		(0.48)	(0.38)	(0.28)	(0.24)	(0.23)
	1	63.7	82.6	92.1	93.6	94.3
		(0.48)	(0.38)	(0.27)	(0.24)	(0.23)
	0	68.2	84.1	90.3	92.8	93.8
		(0.47)	(0.37)	(0.30)	(0.26)	(0.24)
	0.25	68.6	84.1	90.1	92.9	94.2
		(0.46)	(0.37)	(0.30)	(0.26)	(0.23)
Model S	0.5	68.6	84.0	90.1	92.5	94.3
		(0.46)	(0.37)	(0.30)	(0.26)	(0.23)
	0.75	67.5	83.4	90.4	92.6	94.1
		(0.47)	(0.37)	(0.29)	(0.26)	(0.24)
	1	66.3	82.8	90.2	92.9	94.0
		(0.47)	(0.38)	(0.30)	(0.26)	(0.24)

Table 3.2. Coverage percentages of a 95% CI for $\beta = 0$. Simulation SEs in parentheses.

Unsurprisingly given our results, coverage probabilities under Model S are at the nominal level in large samples. The results under Model D are more interesting: recall that Proposition 3.3.5 established consistency only for $r > 0.5$, and indeed we see that we get appropriate coverage in large samples for $r = 0.5, 0.75$, and $r = 1$. In contrast, our simulations display poor coverage in large samples for $r = 0$ and $r = 0.25$, where Proposition 3.3.5 does not establish consistency. This suggests that $\hat{V}$ may be a poor approximation of the true asymptotic variance when there are a roughly equal number of negatively and positively correlated observations in the data. For this reason we feel that it is more appropriate to treat the

results under Assumption 3.2.6 as a form of robustness to small deviations from Assumption 3.2.5, which is the standard assumption imposed in the literature on strong dependence.

3.4.3. A Degrees of Freedom Correction

Given the simulation results in the previous sections, and the fact that configurations like Model B do arise in empirical applications, we propose a new degrees of freedom correction to help guard against the potential for under-coverage. Instead of using the critical values from a $N(0,1)$ distribution to perform inference, we propose using the critical values from a t_κ distribution where κ is given by

$$\kappa = G \cdot \left(\frac{\text{med}_g\{M_g\}}{\mathcal{M}^H} \right) ,$$

where $\text{med}_g\{M_g\}$ denotes the median of the $\{M_g\}_{g=1}^G$. This degrees of freedom adjustment is similar in spirit to those proposed for analogous inference procedures using robust variance estimators in other settings: for example, Bell and McCaffrey (2002), Donald and Lang (2007), and Imbens and Kolesar (2016) propose degrees of freedom adjustments for the heteroskedastic and clustered data settings (for a textbook treatment, see Angrist and Pischke, 2008). Although it is ad-hoc, the intuition behind our choice of κ is simple: when the configuration of the dyads satisfies our asymptotic normality assumptions, as in Models S and D, then $\text{med}_g\{M_g\}/\mathcal{M}^H$ is large, and hence the critical values derived from a t_κ distribution approach the critical values derived from a $N(0,1)$ distribution for large G. On the other hand, for configurations like Model B where most of the units are contained in a few dyads but some units are contained in many dyads, we get that $\text{med}_g\{M_g\}/\mathcal{M}^H$ is very small, which results in a down-weighting of κ and hence an enlargement of the critical value. In Table 3 we repeat our first simulation exercise, but with our degrees of freedom adjustment.

Although we still see under-coverage for very unbalanced configurations such as Model B, we do a modest job of improving coverage in this setting while maintaining proper coverage in Models S and D; hence we see that a major benefit of employing this degrees of freedom correction is that it does not require the researcher to take a stand on whether or not their configuration is "well-behaved". Moreover, this degrees of freedom correction could easily be implemented in any software package that computes dyadic standard errors and confidence intervals. Our simulations suggest that, by combining our degrees of freedom correction with the finite sample corrections for $\hat{V}$ presented in Cameron and Miller (2014), inference for most configurations of at least 150 units should behave as expected.

		G				
	Specification	10	25	50	100	250
Model D	i.i.d	73.1	86.4	91.8	93.5	94.5
		(0.44)	(0.34)	(0.27)	(0.25)	(0.23)
	unit-level shock	67.6	87.7	92.8	93.9	94.5
		(0.47)	(0.33)	(0.26)	(0.24)	(0.23)
Model S	i.i.d	76.7	87.7	91.8	93.4	94.7
		(0.43)	(0.33)	(0.27)	(0.25)	(0.22)
	unit-level shock	73.2	86.2	91.6	93.6	94.2
		(0.44)	(0.34)	(0.27)	(0.25)	(0.23)
Model B	i.i.d	74.5	88.3	92.6	94.5	95.8
		(0.46)	(0.32)	(0.26)	(0.23)	(0.2)
	unit-level shock	69.5	87.2	91.7	93.6	93.4
		(0.46)	(0.33)	(0.27)	(0.25)	(0.25)

Table 3.3. Coverage percentages of a 95% CI for $\beta = 0$, with t_κ critical vaues. Simulation SEs in parentheses.

3.5. Conclusion

In this paper we have established a range of conditions under which the dyadic-robust t-statistic is asymptotically normal. We have also seen that in situations where our theorem does not apply, using a normal approximation of T_k for inference may not be appropriate

even for reasonable sample sizes. Our analysis suggests that, when we combine our degrees of freedom correction with the finite-sample corrections to $\hat{V}$ given in Cameron and Miller (2014), inference should not be problematic for most datasets with roughly 150 units. However, if the data features a roughly equal number of positively and negatively correlated dyads, the dyadic-robust variance estimator may not provide a suitable approximation to the asymptotic variance. In our simulations this translated into confidence intervals that covered less often than the nominal coverage dictated, even in large samples. From a theoretical perspective it would be ideal to have a method of inference that is robust to this issue, but we expect that practitioners would consider such situations pathological in most settings of interest.

In our opinion the most pressing issue to explore is that, as pointed out in our simulations, the normal approximation of T_k can be very poor when we do not have a lot of units in the data. A similar problem arises in the clustered-data setting, and recent papers have studied solutions for inference with few or even finitely many clusters (see Bester et al., 2011; Cameron et al., 2008; Canay et al., 2017; Ibragimov and Müller, 2010, 2016, for recent work in this area). A very promising option has been recently proposed in Menzel (2017), who develops a bootstrap procedure for multiway/dyadic clustering that provides refinements whenever the limiting distribution is Gaussian. It would be interesting to see what other techniques could be adapted to the dyadic data setting as well.

Bibliography

Abadie, Alberto, Matthew M Chingos, and Martin R West (2013), "Endogenous stratification in randomized experiments." Technical report, National Bureau of Economic Research.

Aker, Jenny C (2010), "Information from markets near and far: Mobile phones and agricultural markets in niger." *American Economic Journal: Applied Economics*, 2, 46–59.

Aliprantis, Charalambos D and Kim C Border (1986), "Infinite dimensional analysis: a hitchhikers guide."

Angrist, Joshua D and Jörn-Steffen Pischke (2008), *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Antognini, Alessandro Baldi and Alessandra Giovagnoli (2004), "A new Ôbiased coin designÕfor the sequential allocation of two treatments." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53, 651–664.

Arlot, Sylvain, Alain Celisse, et al. (2010), "A survey of cross-validation procedures for model selection." *Statistics surveys*, 4, 40–79.

Armstrong, Timothy and Shu Shen (2015), "Inference on optimal treatment assignments."

Aronow, Peter M, Cyrus Samii, and Valentina A Assenova (2015), "Cluster–robust variance estimation for dyadic data." *Political Analysis*, 23, 564–577.

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

Athey, Susan and Guido W Imbens (2017), "The econometrics of randomized experiments." *Handbook of Economic Field Experiments*, 1, 73–140.

Athey, Susan and Stefan Wager (2017), "Efficient policy learning." *arXiv preprint arXiv:1702.02896*.

Audibert, Jean-Yves, Alexandre B Tsybakov, et al. (2007), "Fast learning rates for plug-in classifiers." *The Annals of statistics*, 35, 608–633.

Aufenanger, Tobias (2017), "Machine learning to improve experimental design." Technical report, FAU Discussion Papers in Economics.

Baldi, Pierre and Yosef Rinott (1989), "On normal approximations of distributions in terms of dependency graphs." *The Annals of Probability*, 17, 1646–1650.

Baldwin, Richard and Dany Jaimovich (2012), "Are free trade agreements contagious?" *Journal of international Economics*, 88, 1–16.

Barrios, Thomas (2014), "Optimal stratification in randomized experiments." *Manuscript, Harvard University*.

Barros, Rodrigo Coelho, Márcio Porto Basgalupp, Andre CPLF De Carvalho, and Alex A Freitas (2012), "A survey of evolutionary algorithms for decision-tree induction." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42, 291–312.

Bartlett, Peter L (2008), "Fast rates for estimation error and oracle inequalities for model selection." *Econometric Theory*, 24, 545–552.

Bartlett, Peter L, Stéphane Boucheron, and Gábor Lugosi (2002), "Model selection and error estimation." *Machine Learning*, 48, 85–113.

Bartlett, Peter L and Shahar Mendelson (2002), "Rademacher and gaussian complexities: Risk bounds and structural results." *Journal of Machine Learning Research*, 3, 463–482.

Bell, Robert M and Daniel F McCaffrey (2002), "Bias reduction in standard errors for linear regression with multi-stage samples." *Survey Methodology*, 28, 169–182.

Beresteanu, Arie (2004), "Nonparametric estimation of regression functions under restrictions on partial derivatives." Technical report, Duke University, Department of Economics.

Bertsimas, Dimitris and Jack Dunn (2017), "Optimal classification trees." *Machine Learning*, 1–44.

Bertsimas, Dimitris, Angela King, Rahul Mazumder, et al. (2016), "Best subset selection via a modern optimization lens." *The Annals of Statistics*, 44, 813–852.

Bester, C Alan, Timothy G Conley, and Christian B Hansen (2011), "Inference with dependent data using cluster covariance estimators." *Journal of Econometrics*, 165, 137–151.

Beygelzimer, Alina and John Langford (2009), "The offset tree for learning with partial labels." In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 129–138, ACM.

Bhattacharya, Debopam and Pascaline Dupas (2012), "Inferring welfare maximizing treatment assignment under budget constraints." *Journal of Econometrics*, 167, 168–196.

Bhattacharya, Rabindra Nath and Edward C Waymire (2007), *A basic course in probability theory*, volume 69. Springer.

Bloom, Howard S, Larry L Orr, Stephen H Bell, George Cave, Fred Doolittle, Winston Lin, and Johannes M Bos (1997), "The benefits and costs of jtpa title ii-a programs: Key findings from the national job training partnership act study." *Journal of human resources*, 549–576.

Blundell, Richard, Amanda Gosling, Hidehiko Ichimura, and Costas Meghir (2007), "Changes in the distribution of male and female wages accounting for employment composition using bounds." *Econometrica*, 75, 323–363.

Boucheron, Stéphane, Olivier Bousquet, and Gábor Lugosi (2005), "Theory of classification: A survey of some recent advances." *ESAIM: probability and statistics*, 9, 323–375.

Breiman, Leo, Jerome Friedman, Charles J Stone, and Richard A Olshen (1984), *Classification and regression trees*. CRC press.

Bubeck, Sébastien, Nicolo Cesa-Bianchi, et al. (2012), "Regret analysis of stochastic and nonstochastic multi-armed bandit problems." *Foundations and Trends® in Machine Learning*, 5, 1–122.

Bugni, Federico A, Ivan A Canay, and Azeem M Shaikh (2017), "Inference under covariate-adaptive randomization." *Journal of the American Statistical Association*.

Bugni, Federico A, Ivan A Canay, and Azeem M Shaikh (2018), "Inference under covariate adaptive randomization with multiple treatments."

Cameron, A Colin, Jonah B Gelbach, and Douglas L Miller (2008), "Bootstrap-based improvements for inference with clustered errors." *The Review of Economics and Statistics*, 90, 414–427.

Cameron, A Colin, Jonah B Gelbach, and Douglas L Miller (2011), "Robust inference with multiway clustering." *Journal of Business & Economic Statistics*, 29, 238–249.

Cameron, A Colin and Douglas L. Miller (2014), "Robust inference for dyadic data." *Working Paper, University of California-Davis*.

Cameron, A Colin and Douglas L Miller (2015), "A practitioner's guide to cluster-robust inference." *Journal of Human Resources*, 50, 317–372.

Canay, Ivan A, Joseph P Romano, and Azeem M Shaikh (2017), "Randomization tests under an approximate symmetry assumption." *Econometrica*, 85, 1013–1030.

Carneiro, Pedro Manuel, Sokbae Lee, and Daniel Wilhelm (2016), "Optimal data collection for randomized control trials."

Carter, Andrew V, Kevin T Schnepel, and Douglas G Steigerwald (2017), "Asymptotic behavior of at-test robust to cluster heterogeneity." *Review of Economics and Statistics*, 99, 698–709.

Cattaneo, Matias D (2010), "Efficient semiparametric estimation of multi-valued treatment effects under ignorability." *Journal of Econometrics*, 155, 138–154.

Cerf, Raphaël (1995), "An asymptotic theory for genetic algorithms." In *European Conference on Artificial Evolution*, 35–53, Springer.

Chambaz, Antoine, Mark J van der Laan, and Wenjing Zheng (2014), "Targeted covariate-adjusted response-adaptive lasso-based randomized controlled trials." *Modern Adaptive Randomized Clinical Trials: Statistical, Operational, and Regulatory Aspects*, 345–368.

Chamberlain, Gary (2011), "Bayesian aspects of treatment choice." *The Oxford Handbook of Bayesian Econometrics*, 11–39.

Chen, Le-Yu and Sokbae Lee (2016), "Best subset binary prediction." *arXiv preprint arXiv:1610.02738*.

Chen, Louis HY and Qi-Man Shao (2004), "Normal approximation under local dependence." *The Annals of Probability*, 32, 1985–2028.

Cheng, Yi, Fusheng Su, and Donald A Berry (2003), "Choosing sample size for a clinical trial using decision analysis." *Biometrika*, 90, 923–936.

Comola, Margherita and Marcel Fafchamps (2014), "Testing unilateral and bilateral link formation." *The Economic Journal*, 124, 954–976.

Conley, Timothy G (1999), "Gmm estimation with cross sectional dependence." *Journal of Econometrics*, 92, 1–45.

Cox, David Roxbee and Nancy Reid (2000), *The theory of the design of experiments*. CRC Press.

Dehejia, Rajeev H (2005), "Program evaluation as a decision problem." *Journal of Econometrics*, 125, 141–173.

Donald, Stephen G and Kevin Lang (2007), "Inference with difference-in-differences and other panel data." *The review of Economics and Statistics*, 89, 221–233.

Dudley, Richard M (1999), *Uniform central limit theorems*, volume 23. Cambridge Univ Press.

Echevarria, Jon and Javier Gardeazabal (2016), "Refugee gravitation." *Public Choice*, 169, 269–292.

Efron, Bradley (1971), "Forcing a sequential experiment to be balanced." *Biometrika*, 58, 403–417.

Egger, Peter H and Filip Tarlea (2015), "Multi-way clustering estimation of standard errors in gravity models." *Economics Letters*, 134, 144–147.

Fafchamps, Marcel and Flore Gubert (2007), "The formation of risk sharing networks." *Journal of Development Economics*, 83, 326–350.

Florios, Kostas and Spyros Skouras (2008), "Exact computation of max weighted score estimators." *Journal of Econometrics*, 146, 86–91.

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

Geer, Sara A (2000), *Empirical Processes in M-estimation*, volume 6. Cambridge university press.

Glennerster, Rachel and Kudzai Takavarasha (2013), *Running randomized evaluations: A practical guide*. Princeton University Press.

Grubinger, Thomas, Achim Zeileis, and Karl-Peter Pfeiffer (2011), "evtree: Evolutionary learning of globally optimal classification and regression trees in r." Technical report, Working Papers in Economics and Statistics.

Györfi, L, L Devroye, and G Lugosi (1996), *A probabilistic theory of pattern recognition*. Springer-Verlag.

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

Hahn, Jinyong, Keisuke Hirano, and Dean Karlan (2011), "Adaptive experimental design using the propensity score." *Journal of Business & Economic Statistics*, 29, 96–108.

Hansen, Christian B (2007), "Asymptotic properties of a robust variance matrix estimator for panel data when t is large." *Journal of Econometrics*, 141, 597–620.

Hirano, Keisuke and Jack R Porter (2009), "Asymptotics for statistical treatment rules." *Econometrica*, 77, 1683–1701.

Hu, Feifang and William F Rosenberger (2006), *The theory of response-adaptive randomization in clinical trials*, volume 525. John Wiley & Sons.

Ibragimov, Rustam and Ulrich K Müller (2010), "t-statistic based correlation and heterogeneity robust inference." *Journal of Business & Economic Statistics*, 28, 453–468.

Ibragimov, Rustam and Ulrich K Müller (2016), "Inference with few heterogeneous clusters." *Review of Economics and Statistics*, 98, 83–96.

Imbens, Guido W (2004), "Nonparametric estimation of average treatment effects under exogeneity: A review." *Review of Economics and statistics*, 86, 4–29.

Imbens, Guido W and Michal Kolesar (2016), "Robust standard errors in small samples: Some practical advice." *Review of Economics and Statistics*, 98, 701–712.

Janson, Svante (1988), "Normal convergence by higher semiinvariants with applications to sums of dependent random variables and random graphs." *The Annals of Probability*, 16, 305–312.

Kahneman, Daniel (2003), "Maps of bounded rationality: Psychology for behavioral economics." *The American economic review*, 93, 1449–1475.

Kallus, Nathan (2016), "Learning to personalize from observational data." *arXiv preprint arXiv:1608.08925*.

Kallus, Nathan (2018), "Optimal a priori balance in the design of controlled experiments." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 85–112.

Karlan, Dean and Jacob Appel (2016), *Failing in the Field: What We Can Learn When Field Research Goes Wrong*. Princeton University Press.

Karlan, Dean and Daniel H Wood (2017), "The effect of effectiveness: Donor response to aid effectiveness in a direct mail fundraising experiment." *Journal of Behavioral and Experimental Economics*, 66, 1–8.

Karlan, Dean S and Jonathan Zinman (2008), "Credit elasticities in less-developed economies: Implications for microfinance." *American Economic Review*, 98, 1040–68.

Kasy, Maximilian (2013), "Why experimenters should not randomize, and what they should do instead."

Kasy, Maximilian (2014), "Using data to inform policy." Technical report.

Kasy, Maximilian (2016), "Why experimenters might not always want to randomize, and what they could do instead." *Political Analysis*, 24, 324–338.

Kitagawa, Toru and Aleksey Tetenov (2018), "Who should be treated? empirical welfare maximization methods for treatment choice." *Econometrica*, 86, 591–616.

Kock, Anders Bredahl and Martin Thyrsgaard (2017), "Optimal sequential treatment allocation."

Koltchinskii, V (2008), "Oracle inequalities in empirical risk minimization and sparse recovery problems: Lecture notes." Technical report, Technical report, Ecole d'ete de Probabilités de Saint-Flour, 2008. 12.6.

Koltchinskii, Vladimir (2001), "Rademacher penalties and structural risk minimization." *IEEE Transactions on Information Theory*, 47, 1902–1914.

Kuznetsova, Olga M and Yevgen Tymofyeyev (2011), "Brick tunnel randomization for unequal allocation to two or more treatment groups." *Statistics in medicine*, 30, 812–824.

Leblang, David (2010), "Familiarity breeds investment: Diaspora networks and international investment." *American Political Science Review*, 104, 584–600.

Lee, Ji Hyung and Kyungchul Song (2017), "Stable limit theorems for empirical processes under conditional neighborhood dependence." *Available at SSRN: https://ssrn.com/abstract=2974393 or http://dx.doi.org/10.2139/ssrn.2974393*.

Lugosi, Gábor and Marten Wegkamp (2004), "Complexity regularization via localized random penalties." *Annals of Statistics*, 1679–1697.

Lustig, Hanno and Robert J Richmond (2017), "Gravity in fx r-squared: Understanding the factor structure in exchange rates." Technical report, National Bureau of Economic Research.

MacKinnon, James G and Matthew D Webb (2017), "Wild bootstrap inference for wildly different cluster sizes." *Journal of Applied Econometrics*, 32, 233–254.

Manski, Charles F (2004), "Statistical treatment rules for heterogeneous populations." *Econometrica*, 72, 1221–1246.

Manski, Charles F (2009), *Identification for prediction and decision*. Harvard University Press.

Mbakop, Eric and Max Tabord-Meehan (2016), "Model selection for treatment choice: Penalized welfare maximization." *arXiv preprint arXiv:1609.03167*.

Menzel, Konrad (2017), "Bootstrap with clustering in two or more dimensions." *arXiv preprint arXiv:1703.03043*.

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky (2017), "Using instrumental variables for inference about policy relevant treatment effects." Technical report, National Bureau of Economic Research.

Moulton, Brent R (1986), "Random group effects and the precision of regression estimates." *Journal of Econometrics*, 32, 385–397.

Narita, Yusuke (2018), "Toward an ethical experiment."

Penrose, Mathew (2003), *Random geometric graphs*. 5, Oxford University Press.

Politis, Dimitris N (2011), "Higher-order accurate, positive semidefinite estimation of large-sample covariance and spectral density matrices." *Econometric Theory*, 27, 703–744.

Poznansky, Michael and Matt K Scroggs (2016), "Ballots and blackmail: Coercive diplomacy and the democratic peace." *International Studies Quarterly*, sqw016.

Pukelsheim, Friedrich (2006), *Optimal design of experiments*. SIAM.

Qian, Min and Susan A Murphy (2011), "Performance guarantees for individualized treatment rules." *Annals of statistics*, 39, 1180.

Rosenberger, William F and John M Lachin (2015), *Randomization in clinical trials: theory and practice*. John Wiley & Sons.

Ryan, Elizabeth G, Christopher C Drovandi, James M McGree, and Anthony N Pettitt (2016), "A review of modern computational algorithms for bayesian optimal design." *International Statistical Review*, 84, 128–154.

Schlag, Karl H (2007), "Eleven% designing randomized experiments under minimax regret." *Un% published manuscript, European University Institute*.

Scott, Clayton and Robert Nowak (2002), "Dyadic classification trees via structural risk minimization." In *Advances in Neural Information Processing Systems*, 359–366.

Simester, Duncan I, Peng Sun, and John N Tsitsiklis (2006), "Dynamic catalog mailing policies." *Management science*, 52, 683–696.

Smith, Kirstine (1918), "On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations." *Biometrika*, 12, 1–85.

Song, Kyungchul and Zhengfei Yu (2014), "Efficient estimation of treatment effects under treatment-based sampling."

Stoye, Jörg (2009), "Minimax regret treatment choice with finite samples." *Journal of Econometrics*, 151, 70–81.

Stoye, Jörg (2012), "Minimax regret treatment choice with covariates or with limited validity of experiments." *Journal of Econometrics*, 166, 138–156.

Sverdlov, Oleksandr (2015), *Modern adaptive randomized clinical trials: statistical and practical aspects*, volume 81. CRC Press.

Swaminathan, Adith and Thorsten Joachims (2015), "Batch learning from logged bandit feedback through counterfactual risk minimization." *Journal of Machine Learning Research*, 16, 1731–1755.

Tetenov, Aleksey (2012), "Statistical treatment choice based on asymmetric minimax regret criteria." *Journal of Econometrics*, 166, 157–165.

Todd, Petra E and Kenneth I Wolpin (2003), "On the specification and estimation of the production function for cognitive achievement." *The Economic Journal*, 113, F3–F33.

Tsybakov, Alexandre B (2004), "Optimal aggregation of classifiers in statistical learning." *Annals of Statistics*, 135–166.

Tsybakov, Alexandre B and Sara A van de Geer (2005), "Square root penalty: adaptation to the margin in classification and in edge estimation." *Annals of Statistics*, 1203–1224.

Van Der Vaart, Aad (1996), "New donsker classes." *The Annals of Probability*, 24, 2128–2140.

Van der Vaart, Aad W (1998), *Asymptotic statistics*, volume 3. Cambridge university press.

Van der Vaart, Aad W and Jon A Wellner (1996), "Weak convergence." In *Weak Convergence and Empirical Processes*, 16–28, Springer.

Vapnik, Vladimir N and Alexey J Chervonenkis (1974), "Theory of pattern recognition."

Wager, Stefan and Susan Athey (2017), "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association*.

Wei, Lee-Jen (1978), "The adaptive biased coin design for sequential experiments." *The Annals of Statistics*, 92–100.

White, Halbert (1980), "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." *Econometrica: Journal of the Econometric Society*, 48, 817–838.

Wooldridge, Jeffrey M (2010), *Econometric analysis of cross section and panel data*. MIT press.

Zadrozny, Bianca (2003), "Policy mining: Learning decision policies from fixed sets of data."

Zelen, Marvin (1974), "The randomization and stratification of patients to clinical trials." *Journal of chronic diseases*, 27, 365–375.

Zhao, Yingqi, Donglin Zeng, A John Rush, and Michael R Kosorok (2012), "Estimating individualized treatment rules using outcome weighted learning." *Journal of the American Statistical Association*, 107, 1106–1118.

APPENDIX A

Appendix to Chapter 1

A.1. Proofs of Main Results in Chapter 1

The proof of Theorem 3.3.1 requires some preliminary machinery which we develop in Appendix A.2. In this section we take the following facts as given:

- We select a representative out of every equivalence class $T \in \mathcal{T}$ by defining an explicit labeling of the leaves, which we call the *canonical labeling* (Definition A.2.1).
- We endow $\mathcal{T}$ with a metric $\rho(\cdot, \cdot)$ that makes $(\mathcal{T}, \rho)$ a compact metric space (Definition A.2.2, Lemma A.2.2).
- We prove that $V(\cdot)$ is continuous in ρ (Lemma A.2.1).
- Let $\mathcal{T}^*$ be the set of minimizers of $V(\cdot)$, then it is the case given our assumptions that

$$\inf_{T^* \in \mathcal{T}^*} \rho(\widetilde{T}_m, T^*) \xrightarrow{a.s.} 0 \ ,$$

as $m \to \infty$ (note that $\rho(\cdot, \cdot)$ is measurable due to the separability of $\mathcal{T}$). Furthermore, there exists a sequence of $\sigma\{(W_i)_{i=1}^m\}/\mathcal{B}(\mathcal{T}_L)$-measurable trees $\bar{T}_m \in \mathcal{T}^*$ such that

$$\rho(\widetilde{T}_m, \bar{T}_m) \xrightarrow{a.s.} 0 \ .$$

(Lemma A.2.4)

Remark A.1.1. To simplify the exposition, we derive all our results for the subset of $\mathcal{T}_L$ which excludes trees with empty leaves. In other words, this means that we will only consider trees of depth L with exactly 2^L leaves. ■

Proof of Theorem 3.3.1

PROOF. By the derivation in the proof of Theorem 3.1 in Bugni et al. (2018), we have that

$$\sqrt{N}(\hat{\theta}(\hat{T}) - \theta) = \sum_{k=0}^{K} \left[\Omega_1(k; \hat{T}) - \Omega_0(k; \hat{T}) \right] + \sum_{k=0}^{K} \Theta_k(k; \hat{T}) \ ,$$

where

$$\Omega_a(k; T) := \frac{N(k; T)}{N_a(k; T)} \left[\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{1}\{A_i(T) = a, S_i = k\} \psi_i(a; T) \right] \ ,$$

with the following definitions:

$$\psi_i(a; T) := Y_i(a) - E[Y_i(a)|S(X)] \ ,$$

$$N(k; T) := \sum_{i=1}^{N} \mathbf{1}\{S_i = k\} \ ,$$

$$N_a(k; T) := \sum_{i=1}^{N} \mathbf{1}\{A_i(T) = a, S_i = k\} \ ,$$

and

$$\Theta_k(T) := \sqrt{N} \left(\frac{N(k; T)}{N} - p(k; T) \right) \left[E(Y(1)|S(X) = k) - E(Y(0)|S(X) = k) \right]^2 \ .$$

Note that by Assumptions 1.2.1 and 1.3.1, $\Omega_a(0;\hat{T})$ and $\Theta(0;\hat{T})$ are both $o_P(1)$, so we omit them for the rest of the analysis. To prove our result, we study the process

$$\mathbb{O}(T) := \begin{bmatrix} \Omega_0(1;T) \\ \Omega_1(1;T) \\ \Omega_0(2;T) \\ \vdots \\ \Omega_1(K;T) \\ \Theta(1;T) \\ \vdots \\ \Theta(K;T) \end{bmatrix} . \tag{A.1}$$

By Lemma A.1.1, we have that

$$\mathbb{O}(\hat{T}_m) \stackrel{d}{=} \bar{\mathbb{O}}(\bar{T}_m) + o_P(1) \ ,$$

where $\bar{\mathbb{O}}(\cdot)$ is defined in Lemma A.1.1 and $\bar{T}_m \in \mathcal{T}^*$ is defined in Lemma A.2.4 (note that we have explicitly indexed the trees by the pilot sample index m). Hence

$$\sqrt{N}(\hat{\theta}(\hat{T}_m) - \theta) \stackrel{d}{=} B'\bar{\mathbb{O}}(\bar{T}_m) + o_P(1) \ ,$$

where B is the appropriate vector of ones and negative ones to collapse $\bar{\mathbb{O}}(\bar{T})$:

$$B' = \left[-1, 1, -1, 1, \ldots, 1, 1, 1, \ldots, 1\right] \ .$$

Now, we study $\bar{\mathbb{O}}(\bar{T}_m)$ conditional on the sigma algebra generated by all of the pilot data: $\sigma\{(W_j)_{j=1}^{\infty}\}$. Note that $\bar{T}_m$ is a measurable function of the pilot data and that all other sources of randomness in $\bar{\mathbb{O}}(\bar{T}_m)$ are independent of the pilot data, so that we can "treat"

$\bar{T}_m$ as a deterministic sequence after conditioning (see Remark A.1.2). Fix a subsequence $\bar{T}_{m_j}$ of $\bar{T}_m$. By Lemma A.2.4, $\bar{T}_m \in \mathcal{T}^*$ which is a compact set, so that $\bar{T}_{m_j}$ contains a convergent (sub)subsequence:

$$\bar{T}_{m_{j_\ell}} \to \bar{T}^* \ ,$$

where $\bar{T}^*$ is in $\mathcal{T}^*$ and convergence is with respect to the metric we define in Appendix A.2. Now by repeating many of the arguments of Lemma A.1.1,

$$\bar{\mathbb{O}}(\bar{T}_{m_{j_\ell}}) = \bar{\mathbb{O}}(\bar{T}^*) + o_P(1) \ ,$$

conditional on the pilot data. By the partial sum arguments in Lemma C.1. of Bugni et al. (2018),

$$\bar{\mathbb{O}}(\bar{T}^*) \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_1(\bar{T}^*) & 0 \\ 0 & \Sigma_2(\bar{T}^*) \end{pmatrix}\right)$$

conditional on the pilot data, where $\Sigma_1(\bar{T}^*)$ and $\Sigma_2(\bar{T}^*)$ are such that

$$B'\bar{\mathbb{O}}(\bar{T}^*) \xrightarrow{d} N(0, V^*) \ ,$$

which follows from the fact that, by definition, every $T \in \mathcal{T}^*$ is a minimizer of our variance. Hence we have that

$$B'\bar{\mathbb{O}}(\bar{T}_{m_{j_\ell}}) \xrightarrow{d} N(0, V^*) \ ,$$

conditional on the pilot data, and so since every subsequence of $\bar{T}_m$ contains a sub-sub sequence that converges to the same value, we conclude that

$$B'\bar{\mathbb{O}}(\bar{T}_m) \xrightarrow{d} N(0, V^*) \ ,$$

conditional on the pilot data. By the Dominated Convergence Theorem we get that this convergence holds unconditionally as well. It thus follows that

$$\sqrt{N}(\hat{\theta}(\hat{T}) - \theta) \xrightarrow{d} N(0, V^*) \ ,$$

as desired. ■

Lemma A.1.1. *Given the Assumptions required for Theorem 3.3.1,*

$$\mathbb{O}(\hat{T}) \stackrel{d}{=} \bar{\mathbb{O}}(\bar{T}) + o_P(1) \ ,$$

where $\mathbb{O}(\cdot)$ *is defined in the proof of Theorem 3.3.1 and* $\bar{\mathbb{O}}(\cdot)$ *is defined in the proof of this result.*

Proof. By a slight modification of the argument in Lemma C1 in Bugni et al. (2018), we have that

$$\mathbb{O}(\hat{T}) \stackrel{d}{=} \widetilde{\mathbb{O}}(\hat{T}) \ ,$$

where

$$\widetilde{\mathbb{O}}(T) := \begin{bmatrix} \widetilde{\Omega}_0(1;T) \\ \widetilde{\Omega}_1(1;T) \\ \widetilde{\Omega}_0(2;T) \\ \vdots \\ \Theta(1;T) \\ \vdots \\ \Theta(K;T) \end{bmatrix} \ , \tag{A.2}$$

with

$$\widetilde{\Omega}_a(k;T) = \frac{N(k;T)}{N_a(k;T)} \left[\frac{1}{\sqrt{N}} \sum_{i=N(\hat{F}(k;T)+\hat{F}_a(k;T))+1}^{N(\hat{F}(k;T)+\hat{F}_{a+1}(k;T))} G_a^k(U_{i,(a)}(k);T) \right] ,$$

with the following definitions: $\{U_{i,(a)}(k)\}_{i=1}^N$ are i.i.d $U[0,1]$ random variables generated independently of everything else, and independently across pairs (a,k), $G_a^k(\cdot\,;T)$ is the inverse CDF of the distribution of $\psi(a;T)|S(X)=k$, $\hat{F}(k;T) := \frac{1}{N}\sum_{i=1}^N \mathbf{1}\{S_i < k\}$, and $\hat{F}_a(k;T) := \frac{1}{N}\sum_{i=1}^N \mathbf{1}\{S_i = k, A_i < a\}$. Note that here it is important that we argue that this is true for $\hat{T}$ and *not* just pointwise in $T \in \mathcal{T}$: to do this we repeat the argument in Bugni et al. (2018) for each T and then argue by conditioning on the pilot data.

Let us focus on the term in brackets. Fix some a and k for the time being, and let

$$\mathcal{G} := \{G_a^k(\cdot\,;T) : T \in \mathcal{T}\}$$

be the class of all the inverse CDFs defined above, then the empirical process $\eta_N : [0,1]\times\mathcal{G} \to \mathbb{R}$ defined by

$$\eta_N(u,f) := \frac{1}{\sqrt{N}} \sum_{i=1}^{\lfloor Nu \rfloor} f(U_i) ,$$

is known as the *sequential empirical process* (see Van der Vaart and Wellner (1996)) (note that by construction $E[f(U_i)] = 0$). By Theorem 2.12.1 in Van der Vaart and Wellner (1996), η_N converges in distribution to a tight limit in $\ell^\infty([0,1]\times\mathcal{G})$ if $\mathcal{G}$ is Donsker, which follows by Lemma A.1.4. It follows that η_N is asymptotically equicontinuous in the natural (pseudo) metric

$$d\left((u,f),(v,g)\right) = |u-v| + \rho_P(f,g) ,$$

where ρ_P is the variance pseudometric. Note that since $U_i \sim U[0,1]$ and $E[f(U_i)] = 0$ for all $f \in \mathcal{G}$, ρ_P is equal to the L^2 norm $||\cdot||$. Define $F(k;T) := P(S(X) < k)$ and $F_a(k;T) := \sum_{j<a} p(k;T)\pi_j(k)$, where $\pi_0(k) := 1-\pi(k)$, $\pi_1(k) := \pi$, then it follows by

Lemmas A.1.2, and A.1.5 that:

$$|\hat{F}_a(k;\hat{T}) - F_a(k;\bar{T})| \xrightarrow{p} 0 \ ,$$

$$|\hat{F}(k;\hat{T}) - F(k;\bar{T})| \xrightarrow{p} 0 \ ,$$

$$||G_a^k(\cdot\,;\hat{T}) - G_a^k(\cdot\,;\bar{T})|| \xrightarrow{p} 0 \ ,$$

where $\bar{T} \in \mathcal{T}^*$ as defined in Lemma A.2.4. Hence we have by asymptotic equicontinuity that

$$\eta_N\left(\hat{F}(k;\hat{T}) + \hat{F}_a(k;\hat{T}), G_a^k(\cdot\,;\hat{T})\right) = \eta_N\left(F(k;\bar{T}) + F_a(k;\bar{T}), G_a^k(\cdot\,;\bar{T})\right) + o_P(1) \ .$$

By Lemma A.1.3,

$$\frac{N(k;\hat{T})}{N_a(k;\hat{T})} = \frac{1}{\pi(k;\bar{T})} + o_P(1) \ .$$

Using the above two expressions, it can be shown that

$$\widetilde{\Omega}_a(k;\hat{T}) = \bar{\Omega}_a(k;\bar{T}) + o_P(1) \ ,$$

where

$$\bar{\Omega}_a(k;T) := \frac{1}{\pi(k;T)}\left[\frac{1}{\sqrt{N}}\sum_{i=\lfloor N(F(k;T)+F_a(k;T))\rfloor+1}^{\lfloor N(F(k;T)+F_{a+1}(k;T))\rfloor} G_a^k(U_{i,(a)}(k);T)\right] \ .$$

Now we turn our attention to $\Theta(k;T)$. To show that

$$\Theta(k;\hat{T}) = \Theta(k;\bar{T}) + o_P(1) \ ,$$

we consider the following expansion:

$$\sqrt{N}\left(\frac{N(k;T)}{N} - p(k;T)\right) = \sqrt{N}\left(\frac{n(k;T)}{n}\frac{n}{N} - \frac{n(k;T)}{n}\right) + \frac{\sqrt{N}}{\sqrt{n}}\sqrt{n}\left(\frac{n(k;T)}{n} - p(k;T)\right) \ ,$$

where we recall that $N(k) = n(k)$ for $k > 0$. The result then follows by Assumption 1.3.1, Lemma A.2.4 and standard empirical process results for

$$\sqrt{n}\left(\frac{n(k;T)}{n} - p(k;T)\right) ,$$

since the class of indicators $\{\mathbf{1}\{S(X) = k\} : S \in \mathcal{S}\}$ is Donsker for each k (since the partitions are rectangles and hence for a fixed k we get a VC class). Finally, let

$$\bar{\mathbb{O}}(T) := \begin{bmatrix} \bar{\Omega}_0(1;T) \\ \bar{\Omega}_1(1;T) \\ \bar{\Omega}_0(2;T) \\ \vdots \\ \Theta(1;T) \\ \vdots \\ \Theta(K;T) \end{bmatrix} , \tag{A.3}$$

then we have shown that

$$\mathbb{O}(\hat{T}) \stackrel{d}{=} \bar{\mathbb{O}}(\bar{T}) + o_P(1),$$

as desired. ■

Remark A.1.2. We treated various objects as "fixed" by conditioning on the sigma algebra generated by the pilot data. These arguments can be made more formal by employing the following *substitution property* of conditional expectations (see Bhattacharya and Waymire (2007)):

Let W, V be random maps into $(S_1, \mathcal{S}_1)$ and $(S_2, \mathcal{S}_2)$, respectively. Let κ be a measurable function on $(S_1 \times S_2, \mathcal{S}_1 \times \mathcal{S}_2)$. If W is $\mathcal{H}$-measurable, and $\sigma(V)$ and $\mathcal{H}$ are independent, and

$E|\kappa(W,V)| < \infty$, then

$$E[\kappa(W,V)|\mathcal{H}] = h(W) \ ,$$

where $h(w) := E[\kappa(w,V)]$. ■

Proof of Theorem 1.3.2

PROOF. Adapting the derivation in Theorem 3.3 of Bugni et al. (2018), and using the same techniques developed in the proof of Theorem 3.3.1 of this paper, it can be shown that

$$\hat{V}(\hat{T}) \stackrel{d}{=} V(\bar{T}) + o_P(1) \ .$$

By definition, $\bar{T} \in \mathcal{T}^*$ so that the result follows. ■

Proof of Proposition 1.3.2

PROOF. By definition,

$$\frac{n_1(k)}{n} = \frac{\lfloor n(k)\pi(k) \rfloor}{n} \ .$$

We bound the floor function from above and below:

$$\pi(k)\frac{n(k)}{n} \leqslant \frac{n_1(k)}{n} \leqslant \pi(k)\frac{n(k)}{n} + \frac{1}{n} \ .$$

We consider the lower bound (the upper bound proceeds identically). It suffices to show that

$$\sup_{T \in \mathcal{T}} \left| \frac{n(k;T)}{n} - p(k;T) \right| \xrightarrow{p} 0 \ .$$

Since the partitions are rectangles, for a fixed k we get a VC class and hence by the Glivenko-Cantelli theorem the result follows. ■

Proof of Proposition 1.3.1

PROOF. First note that, for a given realization of the data, there exists an optimal choice of π for every $S \in \mathcal{S}_L$ by continuity of $\widetilde{V}_m(T)$ in π (which we'll call $\pi^*(S)$), so our task is to choose $(S, \pi^*(S))$ to minimize $\widetilde{V}_m(T)$. Given this, note that for a given realization of the data, the empirical objective $\widetilde{V}_m(T)$ can take on only finitely many values, and hence a minimizer $\widetilde{T}$ exists. Re-write the population-level variance $V(T)$ as follows:

$$V(T) = E[\nu_T(X)] \ ,$$

where

$$\nu_T(x) = \left[\frac{\sigma^2_{1,S}(x)}{\pi(S(x))} - \frac{\sigma^2_{0,S}(x)}{1 - \pi(S(x))} + (\theta_S(x) - \theta)^2 \right] \ ,$$

$$\sigma^2_{a,S}(x) = Var(Y(a)|S(X) = S(x)) \ ,$$

$$\theta_S(x) = E[Y(1) - Y(0)|S(X) = S(x)] \ .$$

Write $\widetilde{V}_m(T)$ as

$$\widetilde{V}_m(T) = \frac{1}{m} \sum_{i=1}^{m} \hat{\nu}_T(X_i) \ ,$$

with

$$\hat{\nu}_T(x) = \left[\frac{\hat{\sigma}^2_{1,S}(x)}{\pi(S(x))} - \frac{\hat{\sigma}^2_{0,S}(x)}{1 - \pi(S(x))} + (\hat{\theta}_S(x) - \hat{\theta})^2 \right] \ ,$$

where the hats in the definition of $\hat{\nu}$ simply denote empirical analogs. For the sake of the proof we also introduce the following intermediate quantity:

$$V_m(T) = \frac{1}{m} \sum_{i=1}^{m} \nu_T(X_i) \ .$$

Now, let T^* be any minimizer of $V(T)$ (which exists by Lemma A.2.4), then

$$\begin{aligned} V(\widetilde{T}) - V(T^*) &= V(\widetilde{T}) - \widetilde{V}_m(\widetilde{T}) + \widetilde{V}_m(\widetilde{T}) - V(T^*) \\ &\leqslant V(\widetilde{T}) - \widetilde{V}_m(\widetilde{T}) + \widetilde{V}_m(T^*) - V(T^*) \\ &\leqslant 2 \sup_{T \in \mathcal{T}} |\widetilde{V}_m(T) - V(T)| \ . \end{aligned}$$

So if we can show

$$\sup_{T \in \mathcal{T}} |\widetilde{V}_m(T) - V(T)| \xrightarrow{a.s} 0 \ ,$$

then we are done.

To that end, by the triangle inequality:

$$\sup_{T \in \mathcal{T}} |\widetilde{V}_m(T) - V(T)| \leqslant \sup_{T \in \mathcal{T}} |\widetilde{V}_m(T) - V_m(T)| + \sup_{T \in \mathcal{T}} |V_m(T) - V(T)| \ ,$$

so we study each of these in turn. Let us look at the second term on the right hand side. This converges almost surely to zero by the Glivenko-Cantelli theorem, since the class of functions $\{\nu_T(\cdot) : T \in \mathcal{T}\}$ is Glivenko-Cantelli (this can be seen by the fact that $\nu_T(\cdot)$ can be constructed through appropriate sums, products, differences and quotients of various types of VC-subgraph functions, and by invoking Assumption 1.2.2 to avoid potential degeneracies through division). Hence it remains to show that the first term converges a.s. to zero.

Re-writing:

$$\widetilde{V}_m(T) = \sum_{k=1}^{K} \left[\left(\frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{S(X_i) = k\} \right) \left(\frac{\hat{\sigma}^2_{1,S}(k)}{\pi(k)} - \frac{\hat{\sigma}^2_{0,S}(k)}{1 - \pi(k)} + (\hat{\theta}_S(k) - \hat{\theta})^2 \right) \right] \ ,$$

and

$$V_m(T) = \sum_{k=1}^{K} \left[\left(\frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{S(X_i) = k\} \right) \left(\frac{\sigma^2_{1,S}(k)}{\pi(k)} - \frac{\sigma^2_{0,S}(k)}{1 - \pi(k)} + (\theta_S(k) - \theta)^2 \right) \right] \ ,$$

where, through an abuse of notation, we define $\sigma^2_{a,S}(k) := Var(Y(a)|S(X) = k)$ etc. By the triangle inequality it suffices to consider each difference for each $k \in [K]$ individually. Moreover, since the expression $\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{S(X_i) = k\}$ is bounded, we can factor it out and ignore it in what follows. It can be shown by repeated applications of the triangle inequality, Assumption 1.2.2, the Glivenko-Cantelli Theorem and the following expression for conditional expectation:

$$E[Y|S(X) = k] = \frac{E[Y\mathbf{1}\{S(X) = k\}]}{P(S(X) = k)} \ ,$$

that

$$\sup_{T\in\mathcal{T}} \left| \left(\frac{\hat{\sigma}^2_{1,S}(k)}{\pi(k)} - \frac{\hat{\sigma}^2_{0,S}(k)}{1-\pi(k)} + (\hat{\theta}_S(k) - \hat{\theta})^2 \right) - \left(\frac{\sigma^2_{1,S}(k)}{\pi(k)} - \frac{\sigma^2_{0,S}(k)}{1-\pi(k)} + (\theta_S(k) - \theta)^2 \right) \right| \xrightarrow{a.s} 0 \ .$$

Hence, we see that our result follows. ■

Proof of Proposition 1.3.3

Proof. For simplicity of exposition suppose that $V^*_1 > V^*_2 > ... > V^*_{\bar{L}}$. It suffices to show that

$$\left| \widetilde{V}^{(1)}(\widetilde{T}^{(2)}_L) - V^*_L \right| \xrightarrow{a.s} 0 \ ,$$

for each L, and similarly with 1 and 2 reversed. Then we he have that

$$\widetilde{V}^{CV}_L \xrightarrow{a.s} V^*_L \ ,$$

and hence

$$\hat{L} \overset{a.s}{=} \bar{L} \ ,$$

for m sufficiently large. To that end, by the triangle inequality

$$\left| \widetilde{V}^{(1)}(\widetilde{T}^{(2)}_L) - V^*_L \right| \leqslant \left| \widetilde{V}^{(1)}(\widetilde{T}^{(2)}_L) - \widetilde{V}^{(2)}(\widetilde{T}^{(2)}_L) \right| + \left| \widetilde{V}^{(2)}(\widetilde{T}^{(2)}_L) - V^*_L \right| \ .$$

Consider the second term on the RHS, applying the triangle inequality again,

$$\left|\widetilde{V}^{(2)}(\widetilde{T}_L^{(2)}) - V_L^*\right| \leqslant \left|\widetilde{V}^{(2)}(\widetilde{T}_L^{(2)}) - V(\widetilde{T}_L^{(2)})\right| + \left|V(\widetilde{T}_L^{(2)}) - V_L^*\right| \ ,$$

and both of these terms converge to zero a.s. by the arguments made in the proof of Proposition 1.3.1. Next we consider the first term on the RHS, this is bounded above by

$$\sup_T \left|\widetilde{V}^{(1)}(T) - \widetilde{V}^{(2)}(T)\right| \ ,$$

and another application of the triangle inequality yields

$$\sup_T \left|\widetilde{V}^{(1)}(T) - \widetilde{V}^{(2)}(T)\right| \leqslant \sup_T \left|\widetilde{V}^{(1)}(T) - V(T)\right| + \sup_T \left|\widetilde{V}^{(2)}(T) - V(T)\right| \ ,$$

with both terms converging to 0 a.s. by the arguments made in the proof of Proposition 1.3.1. ■

Lemma A.1.2. *Let $\hat{F}$, $\hat{F}_a$, F and F_a be defined as in the proof of Theorem 3.3.1. Given the Assumptions of Theorem 3.3.1, we have that, for $k = 1, ..., K$,*

$$|\hat{F}_a(k;\hat{T}) - F_a(k;\bar{T})| \xrightarrow{p} 0 \ ,$$

and

$$|\hat{F}(k;\hat{T}) - F(k;\bar{T})| \xrightarrow{p} 0 \ .$$

Proof. We prove the first statement for $a = 1$, as the rest of the results follow similarly. We want to show that

$$\left|\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{S_i(\hat{T}) = k, A_i(\hat{T}) = 0\} - (1 - \pi(k;\bar{T}))p(k;\bar{T})\right| \xrightarrow{p} 0 \ .$$

By the triangle inequality, we bound this above by

$$\left| \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{S_i(\hat{T}) = k, A_i(\hat{T}) = 0\} - (1 - \pi(k; \hat{T}))p(k; \hat{T}) \right| +$$

$$+ \left| (1 - \pi(k; \hat{T}))p(k; \hat{T}) - (1 - \pi(k; \bar{T}))p(k; \bar{T}) \right| \ .$$

The first line of the above expression converges to zero by Assumption 1.3.5. Next consider the second line: by Lemma A.2.4, we have that $|p(k; \hat{T}) - p(k; \bar{T})| \xrightarrow{p} 0$ and $|\pi(k; \hat{T}) - \pi(k; \bar{T})| \xrightarrow{p} 0$ (recall that $\hat{T}$ is simply $\widetilde{T}$ with and extra stratum appended for $k = 0$), and hence the second line converges to zero. ■

Lemma A.1.3. *Given the Assumptions of Theorem 3.3.1, we have that, for $k = 1, ..., K$,*

$$\frac{N(k; \hat{T})}{N_a(k; \hat{T})} = \frac{1}{\pi(k; \bar{T})} + o_P(1) \ .$$

PROOF. This follows from Assumption 1.3.5, the Glivenko-Cantelli Theorem, and the fact that $\pi(k; \bar{T})p(k; \bar{T})$ and $\frac{1}{p(k; \bar{T})}$ are $O_p(1)$. ■

Lemma A.1.4. *Given Assumption 1.2.1, the class of functions $\mathcal{G}$ defined as*

$$\mathcal{G} := \{G_a^k(\cdot\,; T) : T \in \mathcal{T}\} \ ,$$

for a given a and k is a Donsker class.

PROOF. This follows from the discussion of classes of monotone uniformly bounded functions in Van Der Vaart (1996). ■

Lemma A.1.5. *Given the Assumptions of Theorem 3.3.1, we have that, for $k = 1, ..., K$,*

$$||G_a^k(\cdot\,; \hat{T}) - G_a^k(\cdot\,; \bar{T})|| \xrightarrow{p} 0 \ .$$

PROOF. We show this for the case where $Y(a)$ is continuous. We proceed by showing convergence pointwise $a.s.$ by invoking Lemma A.5.2, and then using the dominated convergence theorem. It thus remains to show that

$$|Z_a^k(t;\hat{T}) - Z_a^k(t;\bar{T})| \xrightarrow{a.s.} 0 \ ,$$

where $Z_a^k(\cdot\,;T)$ is the CDF of the distribution of $(Y(a) - E[Y(a)|S(X)])\,\big|S(X) = k$. To that end, fix some ω in the sample space such that

$$\rho(\widetilde{T}(\omega), \bar{T}(\omega)) \to 0 \ ,$$

(note that by Lemma A.2.4 this convergence holds almost surely in ω, and recall that $\hat{T}$ is simply $\widetilde{T}$ with an extra stratum appended to $k = 0$). To emphasize the fact that we are now studying a deterministic sequence of trees, let $T_m^{(1)} = \hat{T}(\omega)$, $T_m^{(2)} = \bar{T}(\omega)$, where have have also explicitly indexed the trees by the pilot sample size. Then our goal is to show that:

$$|Z_a^k(t;T_m^{(1)}) - Z_a^k(t;T_m^{(2)})| \to 0 \ .$$

Re-writing, this difference is equal to:

$$\left| \frac{E[\mathbf{1}\{Y(a) \leqslant t + E(Y(a)|S_m^{(1)}(X) = k)\}\mathbf{1}\{S_m^{(1)}(X) = k\}]}{P(S_m^{(1)}(X) = k)} - \frac{E[\mathbf{1}\{Y(a) \leqslant t + E(Y(a)|S_m^{(2)}(X) = k)\}\mathbf{1}\{S_m^{(2)}(X) = k\}]}{P(S_m^{(2)}(X) = k)} \right| \ , \tag{A.4}$$

(where the randomness is with respect to the distribution of $(Y(a), X)$). By the triangle inequality, Assumption 1.2.2 and a little bit of algebra, this is less than or equal to

$$\begin{aligned}(\text{A.5}) \quad \frac{1}{\delta} \Big| & E[\mathbf{1}\{Y(a) \leqslant t + E(Y(a)|S_m^{(1)}(X) = k)\}\mathbf{1}\{S_m^{(1)}(X) = k\}] - \\ & E[\mathbf{1}\{Y(a) \leqslant t + E(Y(a)|S_m^{(2)}(X) = k)\}\mathbf{1}\{S_m^{(2)}(X) = k\}]\Big| + \\ & \frac{1}{\delta^2} \left|P(S_m^{(1)}(X) = k) - P(S_m^{(2)}(X) = k)\right| \ .\end{aligned}$$

The third line of the expression in (A.5) goes to zero by Lemma A.2.4. It remains to show that the rest goes to zero. Again by the triangle inequality, the first two lines of (A.5) are less than or equal to

$$\begin{aligned}(\text{A.6}) \quad \frac{1}{\delta} \Bigg(\Big| & E[\mathbf{1}\{Y(a) \leqslant t + E(Y|S_m^{(1)}(X) = k)\}\mathbf{1}\{S_m^{(1)}(X) = k\}] - \\ & E[\mathbf{1}\{Y(a) \leqslant t + E(Y(a)|S_m^{(1)}(X) = k)\}\mathbf{1}\{S_m^{(2)}(X) = k\}]\Big| + \\ & \Big| E[\mathbf{1}\{Y(a) \leqslant t + E(Y(a)|S_m^{(1)}(X) = k)\}\mathbf{1}\{S_m^{(2)}(X) = k\}] - \\ & E[\mathbf{1}\{Y(a) \leqslant t + E(Y(a)|S_m^{(2)}(X) = k)\}\mathbf{1}\{S_m^{(2)}(X) = k\}]\Big| \Bigg) \ .\end{aligned}$$

The first two lines of (A.6) are bounded above by

$$\frac{1}{\delta} \left(E \left| \mathbf{1}\{S_m^{(1)}(X) = k\} - \mathbf{1}\{S_m^{(2)}(X) = k\} \right| \right) \ ,$$

(where we recall here that this expectation is with respect to the distribution of X). This bound converges to zero by Lemma A.2.4 and the definition of the metric ρ_2 on $\mathcal{S}_L$. The last two lines of (A.6) are bounded above by

$$\frac{1}{\delta} \left(E \left| \mathbf{1}\{Y(a) \leqslant t + E(Y(a)|S_m^{(1)}(X) = k)\} - \mathbf{1}\{Y(a) \leqslant t + E(Y(a)|S_m^{(2)}(X) = k)\} \right| \right) \ .$$

By similar arguments to what we have shown above, this also converges to zero, and hence we're done. ■

A.2. A Theory of Convergence for Stratification Trees

Remark A.2.1. For the remainder of this section, suppose X is continuously distributed. Modifying the results to include discrete covariates with finite support is straightforward. Also recall that as discussed in Remark A.1.1, to simplify the exposition we derive our results for the subset of $\mathcal{T}_L$ which excludes trees with empty leaves. ■

We will define a metric ρ on the space $\mathcal{T}_L$ and study its properties. To define ρ, we write it as a product metric between a metric ρ_1 on $\mathcal{S}_L$, which we define below, and ρ_2 the Euclidean metric on $[0,1]^K$. Recall from Remark 1.2.4 that any permutation of the elements in $[K]$ simply results in a re-labeling of the partition induced by $S(\cdot)$. For this reason we explicitly define the labeling of a tree partition that we will use, which we call the *canonical labeling*:

Definition A.2.1. *(The Canonical Labeling)*

- *Given a tree partition* $\{\Gamma_D, \Gamma_U\}$ *of depth one, we assign a label of* 1 *to* Γ_D *and a label of* 2 *to* Γ_U *(recall by Remark A.1.1 that both of these are nonempty).*
- *Given a tree partition* $\{\Gamma_D^{(L-1)}, \Gamma_U^{(L-1)}\}$ *of depth* $L > 1$*, we label* $\Gamma_D^{(L-1)}$ *as a tree partition of depth* $L-1$ *using the labels* $\{1, 2, ..., K/2\}$*, and use the remaining labels* $\{K/2+1, ..., K\}$ *to label* $\Gamma_U^{(L-1)}$ *as a tree partition of depth* $L-1$ *(recall by Remark A.1.1 that each of these subtrees hase exactly* 2^{L-1} *leaves).*
- *If it is ever the case that a tree partition of depth* L *can be constructed in two different ways, we specify the partition unambiguously as follows: if the partition can be written as* $\{\Gamma_D^{(L-1)}, \Gamma_U^{(L-1)}\}$ *with cut* (j, γ) *and* $\{\Gamma_D^{\prime(L-1)}, \Gamma_U^{\prime(L-1)}\}$ *with cut* (j', γ')*,*

then we select whichever of these has the smallest pair (j, γ) where our ordering is lexicographic. If the cuts (j, γ) are equal then we continue this recursively on the subtrees, beginning with the left subtree, until a distinction can be made.

In words, the canonical labeling labels the leaves from "left-to-right" when the tree is depicted in a tree representation (and the third bullet point is used to break ties whenever multiple such representations are possible). All of our previous examples have been canonically labeled (see Examples 1.2.1, 1.2.2). From now on, given some $S \in \mathcal{S}_L$, we will use the the version of S that has been canonically labeled. Let P_X be the measure induced by the distribution of X on $\mathcal{X}$. We are now ready to define our metric $\rho_1(\cdot, \cdot)$ on $\mathcal{S}_L$ as follows:

Definition A.2.2. *For $S_1, S_2 \in \mathcal{S}_L$,*

$$\rho_1(S_1, S_2) := \sum_{k=1}^{2^L} P_X(S_1^{-1}(k) \Delta S_2^{-1}(k)) \ .$$

That ρ_1 is a metric follows from the properties of symmetric differences. We show under appropriate assumptions that $(\mathcal{S}, \rho_1)$ is a complete metric space in Lemma A.2.2, and that $(\mathcal{S}, \rho_1)$ is totally bounded in Lemma A.2.3. Hence $(\mathcal{S}, \rho_1)$ is a compact metric space under appropriate assumptions. Combined with the fact that $([0,1]^{2^L}, \rho_2)$ is a compact metric space, it follows that $(\mathcal{T}, \rho)$ is a compact metric space.

Next we show that $V(\cdot)$ is continuous in our new metric.

Lemma A.2.1. *Given Assumption 1.2.1, $V(\cdot)$ is a continuous function in ρ.*

Proof. We want to show that for a sequence $T_n \to T$, we have $V(T_n) \to V(T)$. By definition, $T_n \to T$ implies $S_n \to S$ and $\pi_n \to \pi$ where $T_n = (S_n, \pi_n)$, $T = (S, \pi)$. By the

properties of symmetric differences,

$$|P(S_n(X) = k) - P(S(X) = k)| \leqslant P_X(S_n^{-1}(k)\Delta S^{-1}(k)) \ ,$$

and hence $P(S_n(X) = k) \to P(S(X) = k)$. It remains to show that $E[f(Y(a))|S_n(X) = k] \to E[f(Y(a))|S(X) = k]$ for $f(\cdot)$ a continuous function. Re-writing:

$$E[f(Y(a))|S_n(X) = k] = \frac{E[f(Y(a))\mathbf{1}\{S_n(X) = k\}]}{P(S_n(X) = k)} \ .$$

The denominator converges by the above inequality, and the numerator converges by the above inequality combined with the boundedness of $f(Y)$. ■

Lemma A.2.2. *Given Assumptions 1.2.1 and 1.2.2, $(\mathcal{S}, \rho_1)$ is a complete metric space.*

Proof. We proceed by induction on the depth of the tree in the following fashion: Let $\Gamma_n = \times_{j=1}^d [a_{jn}, b_{jn}]$ be a Cauchy sequence w.r.t ρ_1 of *depth 0* tree partitions (i.e. simply rectangles). Suppose for the time being that we have shown that $\{a_{jn}\}_n$ and $\{b_{jn}\}_n$ are both convergent as sequences in $\mathbb{R}$, so that $\{\Gamma_n\}_n$ converges to a depth zero decision tree given by $\Gamma = \times_{j=1}^d [\lim a_{jn}, \lim b_{jn}]$.

Now for the induction step, suppose it is the case that a Cauchy sequence of depth $(L-1)$ tree partitions $\{S_n^{(L-1)}\}_n$ on $\Gamma_n = \times_{j=1}^d [a_{jn}, b_{jn}]$ converges to a depth $(L-1)$ tree partition $S^{(L-1)}$ on $\Gamma = \times_{j=1}^d [\lim a_{jn}, \lim b_{jn}]$. Consider a Cauchy sequence of depth L tree partitions $\{S_n^L\}_n$ on Γ_n, and consider the corresponding subtrees $\{S_{D;n}^{(L-1)}\}_n$ on $\Gamma_{D;n}(j_n, \gamma_n)$ and $\{S_{U;n}^{(L-1)}\}_n$ on $\Gamma_{U;n}(j_n, \gamma_n)$ for some j_n and γ_n. By the definition of ρ_1, it is immediate that $\{S_{D;n}^{(L-1)}\}_n$ and $\{S_{U;n}^{(L-1)}\}_n$ are Cauchy, and so by the induction hypothesis each of these converges to some tree $S_D^{(L-1)}$ and $S_U^{(L-1)}$ on $\Gamma_D(\lim j_n, \lim \gamma_n)$ and $\Gamma_U(\lim j_n, \lim \gamma_n)$ respectively. But

then the resulting collection $\{S_D^{(L-1)}, S_U^{(L-1)}\}$ describes a limit of the original sequence $\{S_n^L\}_n$ and so we're done.

It remains to show that our conclusion holds for the base case. Our goal is to show that for a sequence of cubes $\Gamma_n = \times_{j=1}^d [a_{jn}, b_{jn}]$ which is Cauchy, that the corresponding sequences $\{a_{jn}\}$ and $\{b_{jn}\}$ are both Cauchy as sequences in $\mathbb{R}$. First note that it suffices to treat $P_X(\cdot)$ as Lebesgue measure λ on $[0,1]^d$, since by Assumption 1.2.1, for any measurable set A,

$$P_X(A) = \int_A f_X d\lambda \geqslant c\lambda(A) \ ,$$

for some $c > 0$. Moreover to show each sequence $\{a_{jn}\}_n$ $\{b_{jn}\}_n$ is Cauchy, it suffices to argue this for $d = 1$, since we can argue for $d > 1$ by repeating the argument on the projection onto each axis. So let $d = 1$ and consider a sequence of intervals $\{[a_n, b_n]\}_n$ which is Cauchy (w.r.t to the metric induced by Lebesgue measure), then

$$\lambda([a_n, b_n] \Delta [a_{n'}, b_{n'}]) = |b_{n'} - b_n| + |a_{n'} - a_n|,$$

and hence it follows that the sequences $\{a_n\}_n$ and $\{b_n\}_n$ are Cauchy as sequences in $\mathbb{R}$, and thus convergent. It follows that $\{[a_n, b_n]\}_n$ converges to $[\lim a_n, \lim b_n]$. ■

Lemma A.2.3. *Given Assumption 1.2.1 $(\mathcal{S}_L, \rho_1)$ is a totally bounded metric space.*

Proof. Given any measurable set A, we have by Assumption 1.2.1 that

$$P_X(A) = \int_A f_X d\lambda \leqslant C\lambda(A) \ ,$$

where λ is Lebesgue measure, for some constant $C > 0$. The result now follows immediately by constructing the following ϵ-cover: at each depth L, consider the set of all trees that can

be constructed from the set of splits $\{\frac{\epsilon}{C(2^L-1)}, \frac{2\epsilon}{C(2^L-1)}, ..., 1\}$. By construction any tree in $\mathcal{S}_L$ is at most ϵ away from some tree in this set. ■

Lemma A.2.4. *Given Assumptions 1.2.1, 1.2.2, 1.3.1, and 1.3.2. Then the set $\mathcal{T}^*$ of maximizers of $V(\cdot)$ exists, and*

$$\inf_{T^* \in \mathcal{T}^*} \rho(\widetilde{T}_m, T^*) \xrightarrow{a.s.} 0 \ ,$$

where measurability of $\rho(\cdot, \cdot)$ is guaranteed by the separability of $\mathcal{T}$. Furthermore, there exists a sequence of $\sigma\{(W_i)_{i=1}^m\}/\mathcal{B}(\mathcal{T}_L)$-measurable trees $\bar{T}_m \in \mathcal{T}^*$ such that

$$\rho(\widetilde{T}_m, \bar{T}_m) \xrightarrow{a.s.} 0 \ .$$

Proof. First note that, since $(\mathcal{T}, \rho)$ is a compact metric space and $V(\cdot)$ is continuous, we have that $\mathcal{T}^*$ exists and is itself compact. Fix an $\epsilon > 0$, and let

$$\mathcal{T}_\epsilon := \{T \in \mathcal{T} : \inf_{T^* \in \mathcal{T}^*} \rho(T, T^*) > \epsilon\} \ ,$$

then it is the case that

$$\inf_{T \in \mathcal{T}_\epsilon} V(T) > V^* \ .$$

To see why, suppose not and consider a sequence $T_m \in \mathcal{T}_\epsilon$ such that $V(T_m) \to V^*$. Now by the compactness of $\mathcal{T}$, there exists a convergent subsequence $\{T_{m_\ell}\}$ of $\{T_m\}$, i.e. $T_{m_\ell} \to T'$ for some $T' \in \mathcal{T}$. By continuity, it is the case that $V(T_{m_\ell}) \to V(T')$ and by assumption we have that $V(T_{m_\ell}) \to V^*$, so we see that $T' \in \mathcal{T}^*$ but this is a contradiction.

Hence, for every $\epsilon > 0$, there exists some $\eta > 0$ such that

$$V(T) > V^* + \eta \ ,$$

for every $T \in \mathcal{T}_\epsilon$. Let ω be any point in the sample space for which we have that $V(\widetilde{T}_m(\omega)) \to V^*$, then it must be the case that $\tilde{T}_m(\omega) \notin \mathcal{T}_\epsilon$ for m sufficiently large, and hence

$$\inf_{T^* \in \mathcal{T}^*} \rho(\widetilde{T}_m, T^*) \xrightarrow{a.s.} 0 \ .$$

To make our final conclusion, it suffices to note that $\rho(\cdot, \cdot)$ is itself a continuous function and so by the compactness of $\mathcal{T}^*$, there exists some sequence of trees $\bar{T}$ such that

$$\inf_{T^* \in \mathcal{T}^*} \rho(\widetilde{T}_m, T^*) = \rho(\widetilde{T}_m, \bar{T}_m) \ .$$

Furthermore, by the continuity of ρ, the measurability of $\widetilde{T}$, and the compactness of $\mathcal{T}^*$, we can ensure the measurability of the $\bar{T}_m$, by invoking a measurable selection theorem (see Theorem 18.19 in Aliprantis and Border (1986)). ■

A.3. Supplementary Results for Chapter 1

A.3.1. Supplementary Example

In this section we present a result which complements the discussion in the introduction on how stratification can reduce the variance of the difference-in-means estimator. Using the notation from Section 1.2.2, let $\{Y_i(1), Y_i(0), X_i\}_{i=1}^n$ be i.i.d and let Y be the observed outcome. Let $S : \mathcal{X} \to [K]$ be a stratification function. Consider treatments $\{A_i\}_{i=1}^n$ which are assigned via stratified block randomization using S, with a target proportion of 0.5 in each stratum (see Example 1.2.5 for a definition). Finally, let

$$\hat{\theta} = \frac{1}{n_1} \sum_{i=1}^n Y_i A_i - \frac{1}{n - n_1} \sum_{i=1}^n Y_i (1 - A_i) \ ,$$

where $n_1 = \sum_{i=1}^{n} \mathbf{1}\{A_i = 1\}$. It can be shown using Theorem 4.1 of Bugni et al. (2017) that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V) \ ,$$

with $V = V_Y - V_S$, where V_Y does not depend on S and

$$V_S := E\left[(E[Y(1)|S(X)] + E[Y(0)|S(X)])^2\right] \ .$$

In contrast, if treatment is assigned without any stratification, then

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V') \ ,$$

with $V' = V_Y - E[Y(1)+Y(0)]^2$. It follows by Jensen's inequality that $V_S > E[Y(1)+Y(0)]^2$ as long as $E[Y(1)+Y(0)|S(X) = k]$ is not constant for all k. Hence we see that stratification lowers the asymptotic variance of the difference in means estimator as long as the outcomes are related to the covariates as described above.

A.3.2. Alternative Asymptotic Framework

In this section we present some supplementary results about the asymptotic behavior of $\hat{\theta}(\hat{T})$. We consider an asymptotic framework where the pilot study can be large relative to the total sample size:

Assumption A.3.1. *We consider the following asymptotic framework:*

$$\frac{m}{N} = \lambda + o\left(\frac{1}{\sqrt{N}}\right) \ ,$$

where $N = m + n$, for some $\lambda \in [0, 1]$ as $m, n \to \infty$.

To prove an analogous result to Theorem 3.3.1 in this setting, we impose one additional assumption:

Assumption A.3.2. *The pilot-experiment data $\{W_i\}_{i=1}^m$ was generated through a simple randomized experiment without stratification.*

In contrast, in our original asymptotic framework we made no assumptions about how the pilot experiment was performed, except to prove Proposition 1.3.1. As explained in Remark 1.2.8, if the pilot experiment were stratified, we may want to incorporate this information into the specification of $\hat{T}$. In this case Assumption A.3.2 could be weakened in various ways at the cost of making the expression for the variance in Theorem A.3.1 slightly more complicated.

We now obtain the following result about the ATE estimator $\hat{\theta}(\hat{T})$:

Theorem A.3.1. *Given Assumptions 1.2.1, 1.2.2, 1.2.3, A.3.1, 1.3.2, 1.3.4, 1.3.5, A.3.2, and 1.3.6, we have that*

$$\sqrt{N}(\hat{\theta}(\hat{T}) - \theta) \xrightarrow{d} N(0, V_\lambda^*) \ ,$$

where

$$V_\lambda^* = \lambda V_0 + (1-\lambda)V^* \ ,$$

and

$$V_0 = \frac{\sigma_0^2(0)}{1-\pi_0} + \frac{\sigma_1^2(0)}{\pi_0} \ .$$

Hence we see that in this asymptotic framework the pooled estimator $\hat{\theta}(\hat{T})$ has an asymptotic variance which is a weighted combination of the optimal variance and the variance in the pilot experiment, with weights which correspond to their relative sizes.

We now explain how to modify the proofs of Lemma A.1.1 and 3.3.1 to prove this result. In comparison to the proof of Lemma A.1.1 we now have an extra component which corresponds to the pilot stratum, but the proof continues to hold with that stratum left untouched. For Theorem 3.3.1, we modify the argument as follows. Let $R_p(\bar{T})$ denote the components of $\bar{\mathbb{O}}(\bar{T})$ which correspond to the pilot (where it is implicit that we have augmented $\bar{T}$ to include an extra stratum at $k = 0$ for the pilot data), let $R_m(\bar{T})$ denote the components of $\bar{\mathbb{O}}(\bar{T})$ which correspond to the main study, and let (C_p, C_m) be the corresponding re-arrangement of B such that $(C_p, C_m)'(R_p, R_m) = B'\bar{\mathbb{O}}$. Then we claim that

$$P(C_p' R_p(\bar{T}) \leqslant t_p, C_m' R_m(\bar{T}) \leqslant t_m) \to P(\zeta_p \leqslant t_p, \zeta_m \leqslant t_m) \ ,$$

where t_p, t_m are arbitrary real numbers and (ζ_p, ζ_m) are independent mean zero normals, independent of everything else, with variances such that $var(\zeta_p) + var(\zeta_m) = V_\lambda^*$. To see this consider the following derivation, where $\sigma\{(W_j)_{i=1}^\infty\}$ is the sigma algebra generated by the pilot data:

$$\begin{aligned}
P(C_p' R_p(\bar{T}) \leqslant t_p, C_m' R_m(\bar{T}) \leqslant t_m) &= E\left[P(C_p' R_p(\bar{T}) \leqslant t_p, C_m' R_m(\bar{T}) \leqslant t_m) | \sigma\{(W_j)_{i=1}^\infty\})\right] \\
&= E\left[P(C_m' R_m(\bar{T}) \leqslant t_m | \sigma\{(W_j)_{i=1}^\infty\}) \mathbf{1}\{C_p' R_p(\bar{T}) \leqslant t_p\}\right] \\
&= E[(P(C_m' R_m(\bar{T}) \leqslant t_m | \sigma\{(W_j)_{i=1}^\infty\}) \\
&\qquad - P(\zeta_m \in A_m)) \mathbf{1}\{C_p' R_p(\bar{T}) \leqslant t_p\}] \\
&\qquad + E\left[P(\zeta_m \in A_m) \mathbf{1}\{R_p(\bar{T}) \in A_p\}\right] \\
&= E[(P(C_m' R_m(\bar{T}) \leqslant t_m | \sigma\{(W_j)_{i=1}^\infty\}) \\
&\qquad - P(\zeta_m \in A_m)) \mathbf{1}\{C_p' R_p(\bar{T}) \leqslant t_p\}] \\
&\qquad + P(\zeta_m \in A_m) P(C_p' R_p(\bar{T}) \leqslant t_p) \ ,
\end{aligned}$$

Where the first equality comes from the law of iterated expectations, and the second equality follows from the fact that $R_p(\bar{T})$ is non-stochastic once we condition on $\sigma\{(W_j)_{i=1}^{\infty}\}$. By a standard multivariate CLT, $P(C_p' R_p(\bar{T}) \leqslant t_p) \to P(\zeta_p \leqslant t_p)$, and by the proof of Theorem 3.3.1

$$\left|P(C_m' R_m(\bar{T}) \leqslant t_m | \sigma\{(W_j)_{i=1}^{\infty}\}) - P(\zeta_m \in A_m)\right| = o_p(1) \ ,$$

and so the result follows.

A.3.3. Details on the Multiple Treatment Case

In this section we present formal results for the setting with multiple treatments. Recall from Section 1.3.2 that here we are interested in the vector of ATEs

$$\theta = (\theta_a : a \in \mathcal{A}) \ ,$$

where $\theta_a = E[Y(a) - Y(0)]$. We also generalized the concept of a stratification tree to accommodate multiple treatments, and extended our estimator $\hat{\theta}$ accordingly.

Given a matrix norm $||\cdot||$, our goal is to choose $T \in \mathcal{T}_L$ to minimize $||\mathbb{V}(T)||$ as defined in Section 1.3.2. Define $V(T) := ||\mathbb{V}(T)||$ and let V^* be the minimum of this objective function. Consider the following extensions of Assumptions 1.2.1, 1.2.2, 1.3.2, 1.3.4, and 1.3.5 to multiple treatments:

Assumption A.3.3. *Q satisfies the following properties:*

- *$Y(a) \in [-M, M]$ for some $M < \infty$, for $a \in \mathcal{A}_0$, where the marginal distributions of each $Y(a)$ are either continuous or discrete with finite support.*
- *$X \in \mathcal{X} = \times_{j=1}^{d} [b_j, c_j]$, for some $\{b_j, c_j\}_{j=1}^{d}$ finite.*

- $X = (X_C, X_D)$, *where* $X_C \in \mathbb{R}^{d_1}$ *for some* $d_1 \in \{0, 1, 2, ..., d\}$ *is continuously distributed with a bounded, strictly positive density.* $X_D \in \mathbb{R}^{d-d_1}$ *is discretely distributed with finite support.*

Assumption A.3.4. *Constrain the set of stratification trees* $\mathcal{T}_L$ *such that, for some fixed* $\nu > 0$, $\pi_a(k) \in [\nu, 1-\nu]$ *for all* T.

Assumption A.3.5. *The estimator* $\widetilde{T}$ *is a* $\sigma\{(W_i)_{i=1}^m\}/\mathcal{B}(\mathcal{T}_L)$ *measurable function of the pilot data and satisfies*

$$|V(\widetilde{T}) - V^*| \xrightarrow{a.s} 0 \ ,$$

where

$$V^* = \inf_{T \in \mathcal{T}_L} ||\mathbb{V}(T)|| \ ,$$

as $m \to \infty$.

Assumption A.3.6. *The randomization procedure is such that, for each* $T = (S, \pi) \in \mathcal{T}$*:*

$$\left[(Y_i(0), Y_i(1), ..., Y_i(|\mathcal{A}|), X_i)_{i=1}^n \perp A^{(n)}(T)\right] \Big| S^{(n)} \ .$$

Assumption A.3.7. *The randomization procedure is such that*

$$\sup_{T \in \mathcal{T}} \left| \frac{n_a(k;T)}{n} - \pi_a(k) p(k;T) \right| \xrightarrow{p} 0 \ ,$$

for each $k \in [K]$. *Where*

$$n_a(k;T) = \sum_{i=1}^n \mathbf{1}\{A_i(T) = a, S_i = k\} \ .$$

We also require the following uniqueness assumption:

Assumption A.3.8. *The minimizer* T^* *of* $V(T)$ *over* $\mathcal{T}_L$ *is unique.*

This assumption is quite strong: in general, we are not aware of any conditions that guarantee the uniqueness of the minimum of $V(T)$. Clearly this assumption could be violated, for example, if all the covariates enter the response model symmetrically, since then many distinct trees could minimize $V(T)$. However, it is not clear if such examples would arise in real applications. Finding appropriate conditions under which this should be true, or weakening the result to move away from this assumption, are important considerations for future research.

We now obtain the following result:

Theorem A.3.2. *Given Assumptions A.3.3, A.3.4, 1.2.2, 1.2.3, 1.3.1, A.3.5, A.3.6, A.3.7, and A.3.8, we have that*

$$\sqrt{N}(\hat{\theta}(\hat{T}) - \theta) \xrightarrow{d} N(\mathbf{0}, \mathbb{V}^*) \ ,$$

where $\mathbb{V}^* = \mathbb{V}(T^*)$*, as* $m, n \to \infty$*.*

Note that, since we are now imposing Assumption A.3.8, Assumption 1.3.6 is no longer required. The proof proceeds identically to the proof of Theorem 3.3.1: we simply add the necessary components to the vector $\mathbb{O}(\cdot)$ to accommodate the multiple treatments and follow the derivation in Theorem 3.1 of Bugni et al. (2018) accordingly. We also skip the final conditioning/subsequence step by invoking Assumption A.3.8.

To show that minimizing the empirical variance still satisfies Assumption 1.3.2, the argument proceeds component-wise in a manner similar to the proof of Proposition 1.3.1. Essentially the argument proceeds as follows: let $\nu_T(X)$ and $\hat{\nu}_T(X)$ be the matrix-valued analogues to those described in the proof of Proposition 1.3.1, and suppose we want to show,

for example, that

$$\sup_T |V_n(T) - V(T)| \xrightarrow{a.s} 0 \ .$$

It follows by the reverse triangle inequality that it suffices to show

$$\sup_T \left|\left| \frac{1}{m} \sum_{i=1}^{m} \nu_T(X_i) - E[\nu_T(X)] \right|\right| \xrightarrow{a.s} \mathbf{0} \ ,$$

which follows by applying the Glivenko-Cantelli Theorem component-wise.

A.4. Computational Details/Supplementary Simulation Details for Chapter 1

A.4.1. Computational Details

In this section we describe our strategy for computing stratification trees. We are interested in solving the following empirical minimization problem:

$$\widetilde{T}^{EM} \in \arg\min_{T \in \mathcal{T}_L} \widetilde{V}(T) \ ,$$

where

$$\widetilde{V}(T) := \sum_{k=1}^{K} \frac{m(k;T)}{m} \left[\left(\hat{E}[Y(1) - Y(0) | S(X) = k] - \hat{E}[Y(1) - Y(0)] \right)^2 + \left(\frac{\hat{\sigma}_0^2(k)}{1 - \pi(k)} + \frac{\hat{\sigma}_1^2(k)}{\pi(k)} \right) \right] \ ,$$

with

$$\hat{E}[Y(1) - Y(0) | S(X) = k] := \frac{1}{m_1(k;T)} \sum_{j=1}^{m} Y_j A_j \mathbf{1}\{S(X_j) = k\} - \frac{1}{m_0(k;T)} \sum_{j=1}^{m} Y_j (1 - A_j) \mathbf{1}\{S(X_j) = k\} \ ,$$

$$\hat{E}[Y(1) - Y(0)] := \frac{1}{m} \sum_{j=1}^{m} Y_j A_j - \frac{1}{m} \sum_{j=1}^{m} Y_j (1 - A_j) \ ,$$

$$\hat{\sigma}_a^2(k) := \hat{E}[Y(a)^2 | S(X) = k] - \hat{E}[Y(a) | S(X) = k]^2 \ .$$

Finding a globally optimal tree amounts to a discrete optimization problem in a large state space. Because of this, the most common approaches to fit decision trees in statistics and machine learning are greedy: they begin by searching for a single partitioning of the data

which minimizes the objective, and once this is found, the processes is repeated recursively on each of the new partitions (Breiman et al. (1984), and Friedman et al. (2001) provide a summary of these types of approaches). However, recent advances in optimization research provide techniques which make searching for globally optimal solutions feasible in our setting.

A very promising method is proposed in Bertsimas and Dunn (2017), where they describe how to encode decision tree restrictions as mixed integer linear constraints. In the standard classification tree setting, the misclassification objective can be formulated to be linear as well, and hence computing an optimal classification tree can be computed as the solution to a Mixed Integer Linear Program (MILP), which modern solvers can handle very effectively (see Florios and Skouras (2008), Chen and Lee (2016), Mbakop and Tabord-Meehan (2016), Kitagawa and Tetenov (2018), Mogstad et al. (2017) for some other applications of MILPs in econometrics). Unfortunately, to our knowledge the objective function we consider cannot be formulated as a linear or quadratic objective, and so specialized solvers such as BARON would be required to solve the resulting program. Instead, we implement an evolutionary algorithm (EA) to perform a stochastic search for a global optimum. See Barros et al. (2012) for a survey on the use of EAs to fit decision trees.

The algorithm we propose is based on the procedure described in the `evtree` package description given in Grubinger et al. (2011). In words, a "population" of candidate trees is randomly generated, which we will call the "parents". Next, for each parent in the population we select one of five functions at random and apply it to the parent (these are called the *variation operators*, as described below), which produces a new tree which we call its "child". We then evaluate the objective function for all of the trees (the parents and the children). Proceeding in parent-child pairs, we keep whichever of the two produces a smaller value for the objective. The resulting list of winners then becomes the new population of parents,

and the entire procedure repeats iteratively until the top 5% of trees with respect to the objective are within a given tolerance of each other for at least 50 iterations. The best tree is then returned. If the algorithm does not terminate after 2000 iterations, then the best tree is returned. We describe each of these steps in more detail below.

Although we do note prove that this algorithm converges to a global minimum, it is shown in Cerf (1995) that similar algorithms will converge to a global minimum in probability, as the number of iterations goes to infinity. In practice, our algorithm converges to the global minimum in simple verified examples, and consistently achieves a lower minimum than a greedy search. Moreover, it reliably converges to the same minimum in repeated runs (that is, with different starting populations) for all of the examples we consider in the paper.

Optimal Strata Proportions: Recall that for a given stratum, the optimal proportion is given by

$$\pi^* = \frac{\sigma_1}{\sigma_0 + \sigma_1} \ ,$$

where σ_0 and σ_1 are the within-stratum standard deviations for treatments 0 and 1. In practice, if $\pi^* < 0.1$ then we assign a proportion of 0.1, and if $\pi^* > 0.9$ then we assign a proportion of 0.9 (hence we choose an overlap parameter of size $\nu = 0.1$, as required in Assumption 1.2.2).

Population Generation: We generate a user-defined number of depth 1 stratification trees (typically between 500 and 1000). For each tree, a covariate and a split point is selected at random, and then the optimal proportions are computed for the resulting strata.

Variation Operators:

- *Split*: Takes a tree and returns a new tree that has had one branch split into two new leaves. The operator begins by walking down the tree at random until it finds a leaf. If the leaf is at a depth smaller than L, then a random (valid) split occurs.

Otherwise, the procedure restarts and the algorithm attempts to walk down the tree again, for a maximum of three attempts. If it does not find a suitable leaf, a *minor tree mutation* (see below) is performed. The optimal proportions are computed for the resulting strata.

- *Prune*: Takes a tree and returns a new tree that has had two leaves pruned into one leaf. The operator begins by walking down the tree at random until it finds a node whose children are leaves, and destroys those leaves. The optimal proportions are computed for the resulting strata.
- *Minor Tree Mutation*: Takes a tree and returns a new tree where the splitting value of some internal node is perturbed in such a way that the tree structure is not destroyed. To select the node, it walks down the tree a random number of steps, at random. The optimal proportions are computed for the resulting strata.
- *Major Tree Mutation*: Takes a tree and returns a new tree where the splitting value and covariate value of some internal node are randomly modified. To select the node, it walks down the tree a random number of steps, at random. This modification may result in a partition which no longer obeys a tree structure. If this is the case, the procedure restarts and repeats the algorithm for a maximum of three attempts. If it does not produce a valid tree after three attempts, it destroys any subtrees that violate the tree structure in the final attempt and returns the result. The optimal proportions are computed for the resulting strata.
- *Crossover*: Takes a tree and returns a new tree which is the result of a "crossover". The new tree is produced by selecting a second tree from the population at random, and replacing a subtree of the original tree with a subtree from this randomly selected candidate. The subtrees are selected by walking down both trees at random.

This may result in a partition which no longer obeys a tree structure, in which case it destroys any subtrees that violate the tree structure. The optimal proportions are computed for the resulting strata.

Selection: For each parent-child pair (call these T_p and T_c) we evaluate $\widetilde{V}(T_p)$ and $\widetilde{V}(T_c)$ and then keep whichever tree has the lower value. If it is the case that for a given T any stratum has less than two observations per treatment, we set $\widetilde{V}(T) = \infty$ (this acts as a rough proxy for the minimum cell size parameter δ, as specified in Assumption 1.2.2).

A.4.2. Supplementary Simulation Details

In this section we provide additional details on our implementation of the simulation study.

For each design we compute the ATE numerically. For Model 1 we find $ATE_1 = 0.1257$, for Model 2 we find $ATE_2 = 0.0862$ and for Model 3 we find $ATE_3 = 0.121$. To compute the optimal infeasible trees, we use an auxiliary sample of size $30,000$. The infeasible trees we compute are depicted in Figures A.1, A.2 and A.3 below.

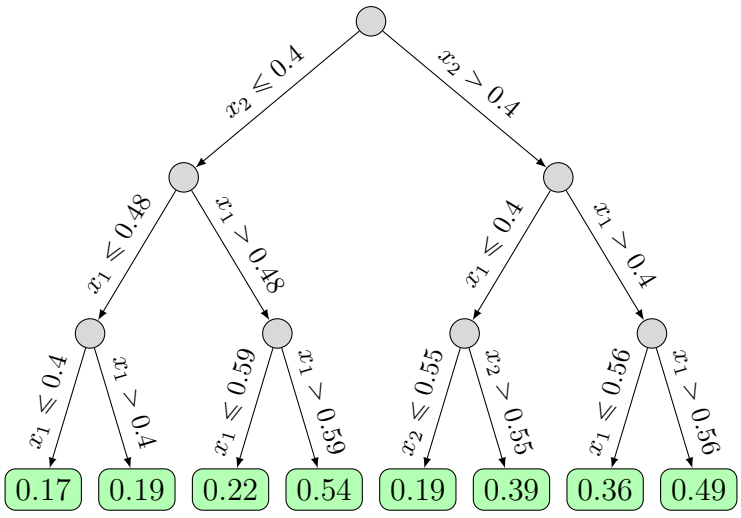

Figure A.1. Optimal Infeasible Tree for Model 1

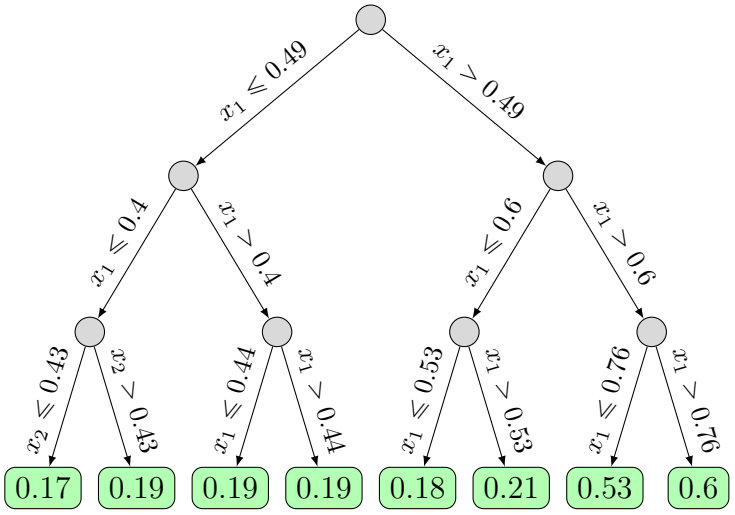

Figure A.2. Optimal Infeasible Tree for Model 2

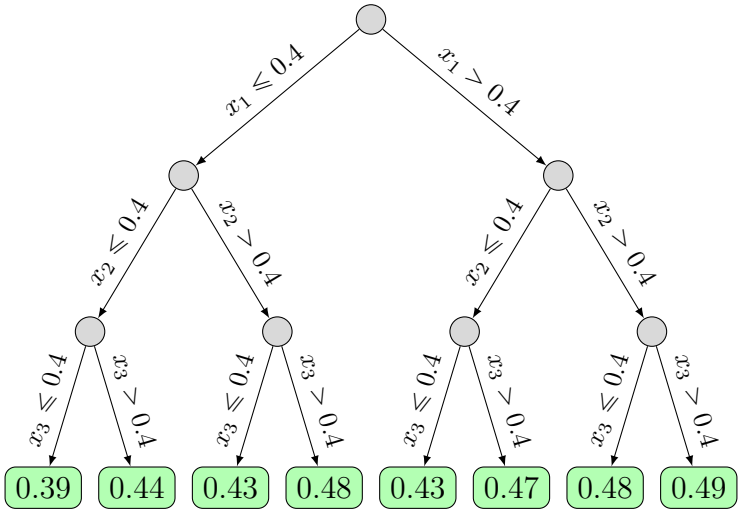

Figure A.3. Optimal Infeasible Tree for Model 3

A.4.3. Application-based Simulation

In this section we repeat the simulation exercise of Section 1.4 using an application-based simulation design, in order to assess the gains from stratification in our application. To generate the data, we draw observations from the entire dataset with replacement, and impute the missing potential outcome for each observation using nearest-neighbour matching

on the Euclidean distance between covariates. We perform the simulations with a sample size of $30,000$, which corresponds approximately to the total number of observations in the dataset. In order to reproduce the empirical setting, we conduct the experiment in two waves, with sample sizes of $12,000$ and $18,000$ in each wave, respectively. In all cases, when we stratify we consider a maximum of 4 strata, which corresponds to the number of strata in Figure 1.6, and use SBR to perform assignment. We compare the following stratification methods using the same criteria as in Section 1.4:

- No Stratification: Here we assign treatment to half the sample, with no stratification.
- Fixed Stratification: Here we use the stratification from Figure 1.6, and assign treatment to half the sample in each stratum.
- Stratification Tree: Here we perform the experiment in two waves. In the first wave, we assign individuals to treatment using the Fixed stratification, and then use this data to estimate a stratification tree. In the second wave we use the estimated tree to assign treatment.
- Cross-Validated Tree: Here we perform the experiment in two waves. In the first wave, we assign individuals to treatment using the Fixed stratification, and then use this data to estimate a stratification tree with depth selected via cross-validation. In the second wave we use the cross-validated tree to assign treatment.
- Infeasible Optimal Tree: Here we estimate an infeasible "optimal" tree by using a large auxiliary sample (see Figure A.4). In the first wave, we assign individuals to treatment using the Fixed stratification. In the second wave, we assign individuals to treatment using the infeasible tree.

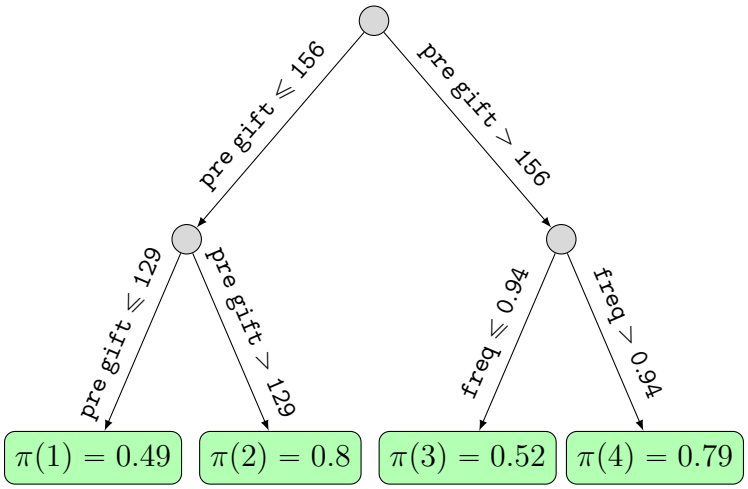

Figure A.4. Infeasible Optimal Tree for App.-based Simulation

When constructing the augmented tree $\hat{T}$, we incorporate the stratifications from both waves in accordance with Remark 1.2.8. We perform 6000 Monte Carlo iterations. Table A.1 presents the simulation results.

	Criteria			
Stratification Method	Coverage	%ΔLength	Power	%ΔRMSE
No Stratification	93.7	0.0	51.9	0.0
Fixed	93.9	-0.6	52.4	-1.6
Strat.Tree	93.0	0.3	52.2	1.1
Strat. Tree (CV)	93.8	-1.9	53.9	-3.0
Infeasible Tree	94.8	-5.9	58.1	-7.7

Table A.1. Simulation Results for Application-Based Simulation

We see in Table A.1 that the overall gains from stratification are small. The Stratification Tree performs slightly worse than no stratification, which agrees with the fact that the cross-validation procedure returned a tree of depth one in Section 1.5. However, despite the fact that the DGP is relatively complex and the gains from stratification are small, the cross-validated Stratification Tree nevertheless performs fairly well.

A.5. Auxiliary Lemmas for Chapter 1

Lemma A.5.1. *Let $\{A_n\}_n$, $\{B_n\}_n$ be sequences of continuous random variables such that*

$$|A_n - B_n| \xrightarrow{a.s.} 0 \ .$$

Furthermore, suppose that the sequences of their respective CDFs $\{F_n(t)\}_n$ $\{G_n(t)\}_n$ are both equicontinuous families at t. Then we have that

$$|F_n(t) - G_n(t)| \to 0 \ .$$

Proof. Fix some $\epsilon > 0$, and choose a $\delta > 0$ such that, for $|t' - t| < \delta$, $|G_n(t) - G_n(t')| < \epsilon$. Furthermore, choose N such that for $n \geqslant N$, $|A_n - B_n| < \delta$ a.s.. Then for $n \geqslant N$:

$$F_n(t) = P(A_n \leqslant t) \leqslant P(B_n \leqslant t + \delta) + P(|A_n - B_n| > \delta) \leqslant G_n(t) + \epsilon \ ,$$

and similarly

$$G_n(t) \leqslant F_n(t) + \epsilon \ .$$

We thus have that $|G_n(t) - F_n(t)| < \epsilon$ as desired. ■

Lemma A.5.2. *Let $\{F_n(t)\}_n$ and $\{G_n(t)\}_n$ be sequences of (absolutely) continuous CDFs with bounded support $[-M, M]$, such that*

$$|F_n(t) - G_n(t)| \to 0 \ ,$$

for all t. Let $\{F_n^{-1}\}_n$ and $\{G_n^{-1}\}_n$ be the corresponding sequences of quantile functions, and suppose that each of these form an equicontinuous family for every $p \in (0, 1)$. Then we have that

$$|F_n^{-1}(p) - G_n^{-1}(p)| \to 0 \ .$$

PROOF. Let V be a random variable that is uniformly distributed on $[-2M, 2M]$, and let $\Gamma(\cdot)$ be the CDF of V. Then it is the case that

$$|F_n(V) - G_n(V)| \xrightarrow{a.s} 0 \ .$$

By the uniform continuity of Γ and the equicontinuity properties of $\{F_n^{-1}\}_n$ and $\{G_n^{-1}\}_n$, we have that $\{P(F_n(V) \leqslant \cdot)\}_n$ and $\{P(G_n(V) \leqslant \cdot)\}_n$ are equicontinuous families for $p \in (0,1)$. It thus follows by Lemma A.5.1 that

$$|P(F_n(V) \leqslant p) - P(G_n(V) \leqslant p)| \to 0 \ .$$

By the properties of quantile functions we have that $|\Gamma(F_n^{-1}(p))) - \Gamma(G_n^{-1}(p))| \to 0$. Hence by the uniform continuity of Γ^{-1}, we can conclude that

$$|\Gamma^{-1}(\Gamma(F_n^{-1}(p))) - \Gamma^{-1}(\Gamma(G_n^{-1}(p)))| = |F_n^{-1}(p) - G_n^{-1}(p)| \to 0 \ ,$$

as desired. ■

Our final lemma completes the discussion in Remark 1.3.3. It shows that, as long as the family of quantile functions defined in Assumption 1.3.6 are continuous, and vary "continuously" in $S \in \mathcal{S}_L$, then Assumption 1.3.6 holds.

Lemma A.5.3. *Let $(\mathbb{D}, d)$ be a compact metric space. Let $\mathcal{F}$ be some class of functions*

$$\mathcal{F} = \{f_d : (0,1) \to \mathbb{R}\}_{d \in \mathbb{D}}$$

such that $f_d(\cdot)$ is continuous and bounded for every $d \in \mathbb{D}$. Define $g : \mathbb{D} \to L^\infty(0,1)$ by $g(d) = f_d(\cdot)$, and suppose that g is continuous. Then we have that, for every $x_0 \in (0,1)$, $\{f_d(\cdot, d)\}_{d \in \mathbb{D}}$ is an equicontinuous family at x_0.

PROOF. By construction, $g(\mathbb{D}) = \mathcal{F}$, and so by the continuity of g and the compactness of $\mathbb{D}$, $\mathcal{F}$ is compact. Let $\epsilon > 0$ and fix some $x_0 \in (0,1)$. Let $\mathcal{F}_{\epsilon/3} = \{f_{d_k}(\cdot)\}_{k=1}^{K}$ be a finite $\epsilon/3$ cover for $\mathcal{F}$. By continuity, there exists a $\delta > 0$ such that if $|x - x_0| < \delta$, $|f_{d_k}(x) - f_{d_k}(x_0)| < \epsilon/3$ for every $k = 1, ..., K$. By the triangle inequality, for any d:

$$|f_d(x) - f_d(x_0)| \leqslant |f_d(x) - f_{d_k}(x)| + |f_{d_k}(x) - f_{d_k}(x_0)| + |f_{d_k}(x_0) - f_d(x_0)| \ ,$$

for all $k = 1, ..., K$. It thus follows that, for $|x - x_0| < \delta$, and by virtue of the fact that $\mathcal{F}_{\epsilon/3}$ is an open cover for $\mathcal{F}$,

$$|f_d(x) - f_d(x_0)| < \epsilon \ ,$$

and hence $\{f_d(\cdot)\}_{d\in\mathbb{D}}$ is an equicontinuous family at x_0. ■

APPENDIX B

Appendix to Chapter 2

B.1. Proofs of Main Results in Chapter 2

Recall that the planner's objective function is given by

$$W(G) = E_P\left[\left(\frac{YD}{e(X)} - \frac{Y(1-D)}{1-e(X)}\right)\cdot \mathbf{1}\{X \in G\}\right] . \tag{B.1}$$

To each treatment allocation $G \in \mathcal{G}$ we associate a function $f_G : \mathbb{R}\times\mathcal{X}\times\{0,1\}\to\mathbb{R}$ defined by:

$$f_G(Z) = f_G(Y,X,D) = \left(\frac{YD}{e(X)} - \frac{Y(1-D)}{1-e(X)}\right)\cdot \mathbf{1}\{X \in G\} ,$$

where $Z = (Y,X,D)$. Let $\mathcal{F} := \{f_G : G \in \mathcal{G}\}$ denote the corresponding set of functions associated to decision rules in $\mathcal{G}$. By (B.1), any optimal allocation in $\mathcal{G}$ solves

$$G^* \in \arg\max_{G\in\mathcal{G}} E_P\left[\left(\frac{YD}{e(X)} - \frac{Y(1-D)}{1-e(X)}\right)\cdot \mathbf{1}\{X \in G\}\right] .$$

Equivalently, functions associated to optimal allocations solve

$$f^* \in \arg\max_{f\in\mathcal{F}} E_P f(Z) .$$

By an abuse of notation, for $G \in \mathcal{G}$, we set

$$W(f_G) = E_P f_G(Z) .$$

Given an approximating sequence $\{\mathcal{G}_k\}_k$ of classes of treatment allocations, let $\{\mathcal{F}_k\}_k$ denote the sequence of associated classes of functions.

The following lemma, whose proof is given in Kitagawa and Tetenov (2018) (Lemma A.1), establishes the relevant link between the classes of sets $\{\mathcal{G}_k\}_k$ and the classes of functions $\{\mathcal{F}_k\}_k$. It shows that if a class $\mathcal{G}$ has finite VC dimension, then the associated class $\mathcal{F}$ is a VC-subgraph class with dimension bounded above by that of $\mathcal{G}$.

Lemma B.1.1. *Let $\mathcal{G}$ be a VC-class of subsets of $\mathcal{X}$ with finite VC dimension V. Let g be a function from $\mathcal{Z} := \mathbb{R} \times \mathcal{X} \times \{0,1\}$ to $\mathbb{R}$. Then the set of functions $\mathcal{F}$ defined by*

$$\mathcal{F} = \{g(z) \cdot \mathbf{1}\{x \in G\} : G \in \mathcal{G}\}$$

is a VC-subgraph class with dimension at most V.

For each $k \geqslant 1$, let $\hat{f}_{n,k}$ be a maximizer of the empirical welfare over the class $\mathcal{F}_k$; that is:

$$\hat{f}_{n,k} = \arg\max_{f \in \mathcal{F}_k} W_n(f) \ ,$$

and for $f \in \mathcal{F}_k$, define the complexity-penalized estimate of welfare by

$$R_{n,k}(f) = W_n(f) - C_n(k) - \sqrt{\frac{k}{n}} \ .$$

The PWM rule $\hat{f}_{n,\hat{k}}$ is then chosen such that

$$\hat{k} = \arg\max_{k \geqslant 1} R_{n,k}(\hat{f}_{n,k}) \ .$$

In what follows, we set $\hat{f}_n := \hat{f}_{n,\hat{k}}$ and $R_n(\hat{f}_n) := R_{n,\hat{k}}(\hat{f}_{n,\hat{k}})$.

To bound the regret, we decompose it as follows

$$W^*_{\mathcal{F}} - W(\hat{f}_n) = \Big(W^*_{\mathcal{F}} - R_n(\hat{f}_n)\Big) + \Big(R_n(\hat{f}_n) - W(\hat{f}_n)\Big)\,. \tag{B.2}$$

The following lemma yields (under Assumption 2.3.4) a subgaussian tail bound for the second term on the right hand side of the preceding equality.

Lemma B.1.2. *Given Assumption 2.3.4, there exists a positive constant Δ (that does not depend on n) such that:*

$$P(R_n(\hat{f}_n) - W(\hat{f}_n) > \epsilon) \leqslant \Delta e^{-2c_o n\epsilon^2}$$

for every n.

Proof. First note that:

$$P(R_n(\hat{f}_n) - W(\hat{f}_n) > \epsilon) \leqslant P\Big(\sup_k \big(R_{n,k}(\hat{f}_{n,k}) - W(\hat{f}_{n,k})\big) > \epsilon\Big)\ ,$$

then by the union bound:

$$P\Big(\sup_k \big(R_{n,k}(\hat{f}_{n,k}) - W(\hat{f}_{n,k})\big) > \epsilon\Big) \leqslant \sum_k P(R_{n,k}(\hat{f}_{n,k}) - W(\hat{f}_{n,k}) > \epsilon)\ .$$

Now by definition of $R_{n,k}$, we have

$$\sum_k P(R_{n,k}(\hat{f}_{n,k}) - W(\hat{f}_{n,k}) > \epsilon) = \sum_k P\big(W_n(\hat{f}_{n,k}) - C_n(k) - W(\hat{f}_{n,k}) > \epsilon + \sqrt{\frac{k}{n}}\big)\ .$$

By Assumption 2.3.4,

$$\sum_k P(W_n(\hat{f}_{n,k}) - W(\hat{f}_{n,k}) - C_n(k) > \epsilon + \sqrt{\frac{k}{n}}) \leqslant \sum_k c_1 e^{-2c_o n(\epsilon + \sqrt{\frac{k}{n}})^2} \leqslant e^{-2c_o n\epsilon^2} \sum_k c_1 e^{-2kc_o}\ .$$

By setting

$$\Delta := \sum_k c_1 e^{-2kc_o} < \infty \ , \tag{B.3}$$

the result follows. ■

Proof of Theorem 2.3.1. We follow the general strategy from Bartlett et al. (2002). For every k, we have

$$W^*_{\mathcal{F}} - W(\hat{f}_n) = \left(W^*_{\mathcal{F}} - W^*_{\mathcal{F}_k}\right) + \left(W^*_{\mathcal{F}_k} - W(\hat{f}_n)\right) \ . \tag{B.4}$$

We first consider the second term in (B.4), and expand it as follows

$$W^*_{\mathcal{F}_k} - W(\hat{f}_n) = W^*_{\mathcal{F}_k} - R_n(\hat{f}_n) + R_n(\hat{f}_n) - W(\hat{f}_n) \ . \tag{B.5}$$

By the definition of R_n, the first term of expression (B.5) is bounded by

$$W^*_{\mathcal{F}_k} - R_n(\hat{f}_n) \leqslant W^*_{\mathcal{F}_k} - W_n(\hat{f}_{n,k}) + C_n(k) + \sqrt{\frac{k}{n}} \ .$$

Fix $\delta > 0$, and choose some $f^*_k \in \mathcal{F}_k$ such that $W(f^*_k) + \delta \geqslant W^*_{\mathcal{F}_k}$.[9] . We have

$$W^*_{\mathcal{F}_k} - W_n(\hat{f}_{n,k}) + C_n(k) + \sqrt{\frac{k}{n}} \leqslant W(f^*_k) + \delta - W_n(f^*_k) + C_n(k) + \sqrt{\frac{k}{n}} \ .$$

Taking expectations of both sides and letting δ converge to 0 yields

$$E[W^*_{\mathcal{F}_k} - R_n(\hat{f}_n)] \leqslant E[C_n(k)] + \sqrt{\frac{k}{n}} \ .$$

[9] If the welfare criterion achieves its maximum on $\mathcal{F}_k$, then f^*_k can be set equal to any maximizer. In general however such an optimum may not exist, and thus we must choose f^*_k will to be an "almost maximizer" of the welfare criterion on $\mathcal{F}_k$.

By Lemma B.1.2 and a standard integration argument (see for instance problem 12.1 in Györfi et al. (1996)), the second term on the right hand side of (B.5) is bounded by

$$E[R_n(\hat{f}_n) - W(\hat{f}_n)] \leqslant \sqrt{\frac{\log(\Delta e)}{2c_o n}} \ .$$

Combining these bounds yields

$$E[W^*_{\mathcal{F}} - W(\hat{f}_n)] \leqslant E[C_n(k)] + W^*_{\mathcal{F}} - W^*_{\mathcal{F}_k} + \sqrt{\frac{\log(\Delta e)}{2c_o n}} + \sqrt{\frac{k}{n}} \ ,$$

for every k, and our result follows. ■

Proof of Lemma 2.3.1. We first establish the inequality

$$P(W_n(\hat{f}_{n,k}) - W(\hat{f}_{n,k}) - C_n(k) > \epsilon) \leqslant \exp\left(-2n\epsilon^2(\frac{\kappa}{3M})^2\right) \ . \tag{B.6}$$

By two standard symmetrization arguments, we get

$$E\big[\sup_{f\in\mathcal{F}_k} W_n(f) - W(f)\big] \leqslant 2E\big[\sup_{f\in\mathcal{F}_k} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)\big] = E\big[C_n(k)\big] \ , \tag{B.7}$$

where we recall that $C_n(k) = E\left[2\sup_{f\in\mathcal{F}_k} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)|Z_1, Z_2, \cdots, Z_n\right]$ and $\{\sigma_i\}_{i=1}^n$ is an $i.i.d$ sequence of Rademacher random variables independent from the data $\{Z_i\}_{i=1}^n$. Note that

$$P(W_n(\hat{f}_{n,k}) - W(\hat{f}_{n,k}) - C_n(k) > \epsilon) \leqslant P(\sup_{f\in\mathcal{F}_k} ((W_n(f) - W(f)) - C_n(k) > \epsilon) \ ,$$

and set $M_{n,k} := \sup_{f\in\mathcal{F}_k} (W_n(f) - W(f)) - C_n(k)$. Combining the preceding inequality with (B.7) yields

$$P(W_n(\hat{f}_{n,k}) - W(\hat{f}_{n,k}) - C_n(k) > \epsilon) \leqslant P\left(M_{n,k} - EM_{n,k} > \epsilon\right) \ .$$

To control the deviations of $M_{n,k}$ from its mean, we use McDiamird's inequality (note that $M_{n,k}$ satisfies the bounded difference property with increments bounded by $\frac{3M}{n\kappa}$) which yields the inequality

$$P\left(M_{n,k} - EM_{n,k} > \epsilon\right) \leqslant \exp\left(-2n\epsilon^2(\frac{\kappa}{3M})^2\right) \ ,$$

from which our result follows.

The second inequality (where C is a universal constant)

$$E[C_n(k)] \leqslant C\frac{M}{\kappa}\sqrt{\frac{V_k}{n}} \ ,$$

follows from a chaining argument and a control on the universal entropy of VC subgraph classes (see for instance the proof of Lemma A.4 in Kitagawa and Tetenov (2018)), along with Lemma B.1.1. ■

Proof of Lemma 2.3.2. Let us assume for notational simplicity that the quantity $m = n(1-\ell)$ is an integer. We first establish the inequality

$$P(W_m(\hat{f}_{m,k}) - W(\hat{f}_{m,k}) - C_m(k) > \epsilon) \leqslant \exp\left(-2n\ell\epsilon^2(\frac{\kappa}{M})^2\right) \ . \tag{B.8}$$

By the definition of $C_m(k)$, we have

$$P(W(\hat{f}_{m,k}) - W(\hat{f}_{m,k}) - C_m(k) > \epsilon) = P(W_r(\hat{f}_{m,k}) - W(\hat{f}_{m,k}) > \epsilon) \ .$$

Now, working conditionally on $\{Z_i\}_{i=1}^m$, we get by Hoeffding's inequality that

$$P(W_r(\hat{f}_{m,k}) - W(\hat{f}_{m,k}) > \epsilon | \{Z_i\}_{i=1}^m) \leqslant \exp\left(-2n\ell\epsilon^2(\frac{\kappa}{M})^2\right) \ .$$

Since the right hand side of the preceding inequality is non random, the inequality holds unconditionally as well.

We now establish the inequality

$$E[C_m(k)] \leqslant C\frac{M}{\kappa\sqrt{(1-\ell)}}\sqrt{\frac{V_k}{n}} \ .$$

By the definition of $C_m(k)$, we have

$$E[C_m(k)] = E[W_m(\hat{f}_{m,k}) - W_r(\hat{f}_{m,k})] = E[W_m(\hat{f}_{m,k}) - W(\hat{f}_{m,k}) + W(\hat{f}_{m,k}) - W_r(\hat{f}_{m,k})] \ .$$

Note that by the law of iterated expectations, we have

$$E[W(\hat{f}_{m,k}) - W_r(\hat{f}_{m,k})] = 0 \ ,$$

and by Lemma A.4 in Kitagawa and Tetenov (2018) combined with Lemma B.1.1 there exists some universal constant C such that:

$$E[W_m(\hat{f}_{m,k}) - W(\hat{f}_{m,k})] \leqslant C\frac{M}{\kappa}\sqrt{\frac{V_k}{m}} \ .$$

Since $m = (1-\ell)n$, the result follows. ■

Proof of Propositions 2.3.1 and 2.3.2. From the inequality

$$\frac{e^{-x}}{(1-e^{-x})} \leqslant \frac{1}{x} \ ,$$

and from (B.3) and (B.6), we derive that

$$\Delta \leqslant 1/2\left(\frac{3M}{\kappa}\right)^2 \ .$$

Similarly, we derive from (B.3) and (B.8) that

$$\Delta \leqslant 1/(2l)\left(\frac{M}{\kappa}\right)^2 \ .$$

The results then follow by substituting these into the inequalities of Theorem 2.3.1. ■

Proof of Theorem 2.3.2. Our strategy here is to proceed analogously to the proof of Theorem 2.3.1 with some additional machinery. For every k, we have that:

$$W^*_{\mathcal{F}} - W(\tilde{f}^e_n) = \left(W^*_{\mathcal{F}} - W^*_{\mathcal{F}_k}\right) + \left(W^*_{\mathcal{F}_k} - W(\hat{f}^e_n)\right). \tag{B.9}$$

Adding and subtracting $R^e_n(\hat{f}^e_n)$ to the last term yields

$$W^*_{\mathcal{F}_k} - W(\hat{f}^e_n) = \left(W^*_{\mathcal{F}_k} - R^e_n(\hat{f}^e_n)\right) + \left(R^e_n(\hat{f}^e_n) - W(\hat{f}^e_n)\right). \tag{B.10}$$

Let $f^*_k := \arg\max_{f\in\mathcal{F}_k} W(f)$, (if the supremum is not achieved, apply the argument to a δ-maximizer of the welfare, and let δ tend to zero). Now consider the first term on the right hand side of (B.10). Expanding yet again gives

$$W^*_{\mathcal{F}_k} - R^e_n(\hat{f}^e_n) = W^*_{\mathcal{F}_k} - W_n(f^*_k) + W_n(f^*_k) - R^e_n(\hat{f}^e_n)\ . \tag{B.11}$$

From the definition of R^e_n, we have

$$W_n(f^*_k) - R^e_n(\hat{f}^e_n) \leqslant W_n(f^*_k) - W^e_n(f^*_k) + C^e_n(k) + \sqrt{\frac{k}{n}} \leqslant \frac{1}{n}\sum_{i=1}^{n} |\hat{\tau}_i - \tau_i| + C^e_n(k) + \sqrt{\frac{k}{n}}\ .$$

Hence, considering the above inequality and taking expectations in (B.11) yields

$$E[W^*_{\mathcal{F}_k} - R^e_n(\hat{f}^e_n))] \leqslant E\Big[\frac{1}{n}\sum_{i=1}^{n} |\hat{\tau}_i - \tau_i|\Big] + E[C^e_n(k)] + \sqrt{\frac{k}{n}}\ ,$$

and thus by Assumption 2.3.7

$$E[W^*_{\mathcal{F}_k} - R^e_n(\hat{f}^e_n))] \leqslant O(\phi_n^{-1}) + E[C^e_n(k)] + \sqrt{\frac{k}{n}}\ . \tag{B.12}$$

We now consider the second term on the right hand side of (B.10). Let $\hat{k}$ be the class k such that

$$\hat{f}_n^e = \hat{f}_{n,\hat{k}}^e \ .$$

Note that $\hat{k}$ is random. We have

$$R_n^e(\hat{f}_n^e) - W(\hat{f}_n^e) = W_n^e(\hat{f}_{n,\hat{k}}^e) - C_n^e(\hat{k}) - \sqrt{\frac{\hat{k}}{n}} - W(\hat{f}_{n,\hat{k}}^e) \ .$$

By adding and subtracting $W_n(\hat{f}_{n,\hat{k}}^e)$ and the function $\tilde{C}_n(\hat{k})$, we get

$$W_n^e(\hat{f}_{n,\hat{k}}^e) - C_n^e(\hat{k}) - \sqrt{\frac{k}{n}} - W(\hat{f}_{n,\hat{k}}^e) =$$

(B.13)

$$\Big(W_n^e(\hat{f}_{n,\hat{k}}^e) - W_n(\hat{f}_{n,\hat{k}}^e)\Big) + \Big(\tilde{C}_n(\hat{k}) - C_n^e(\hat{k})\Big) + \left(W_n(\hat{f}_{n,\hat{k}}^e) - W(\hat{f}_{n,\hat{k}}^e) - \tilde{C}_n(\hat{k}) - \sqrt{\frac{\hat{k}}{n}}\right) .$$

Note again that

$$\sup_k \big(W_n^e(\hat{f}_{n,k}^e) - W_n(\hat{f}_{n,k}^e)\big) \leqslant \frac{1}{n}\sum_{i=1}^n |\hat{\tau}_i - \tau_i| \ ,$$

and so by Assumptions 2.3.7 and 2.3.8, the first two terms of (B.13) are of order $O(\phi_n^{-1})$ in expectation. By the first part of Assumption 2.3.8, and an argument similar to the one used in the proof of Lemma B.1.2, it can be shown that

$$E\Big[\sup_k \big(W_n(\hat{f}_{n,k}^e) - W(\hat{f}_{n,k}^e) - \tilde{C}_n(k) - \sqrt{\frac{k}{n}}\big)\Big] \leqslant \sqrt{\frac{\log(\Delta e)}{2c_0 n}} \ ,$$

where Δ and c_o are the same constants that appear in B.1.2. We thus get

$$E[R_n^e(\hat{f}_n^e) - W(\hat{f}_n^e)] \leqslant O(\phi_n^{-1}) + \sqrt{\frac{\log(\Delta e)}{2m}} \ . \tag{B.14}$$

Now combining (B.12) and (B.14), we conclude that

$$E[W^*_{\mathcal{F}_k} - W(\hat{f}^e_n)] \leqslant O(\phi_n^{-1}) + E[C^e_n(k)] + \sqrt{\frac{k}{n}} + \sqrt{\frac{\log(\Delta e)}{2m}} \ .$$

Finally, by Assumption 2.3.8, we get

$$E[W^*_{\mathcal{F}} - W(\hat{f}^e_n)] \leqslant O(\phi_n^{-1}) + E[\tilde{C}_n(k)] + W^*_{\mathcal{F}} - W^*_{\mathcal{F}_k} + \sqrt{\frac{k}{n}} + \sqrt{\frac{\log(\Delta e)}{2m}} \ ,$$

for all k, and hence the result follows. ■

Proof of Lemma 2.3.3 and 2.3.4. In what follows, we verify that the third condition of Assumption 2.3.8 is satisfied for the holdout and Rademacher penalties with estimated propensity scores, as the first two conditions follow from previous arguments. If we let

$$\tilde{C}_n(k) = E_\sigma \left[2 \sup_{f \in \mathcal{F}_k} \frac{1}{n} \sum_{i=1}^n \sigma_i f(Z_i) | Z_1, Z_2, \cdots, Z_n \right] \ ,$$

which is the infeasible Rademacher penalty that depends on the unknown propensity score, then it can be shown that

$$|\tilde{C}_n(k) - C^e_n(k)| \leqslant E_\sigma \left[\frac{2}{n} \sum_{i=1}^n |\hat{\tau}_i - \tau_i| \Big| Z_1, Z_2, \cdots, Z_n \right] \ .$$

Since the right hand side does not depend on k, we conclude that

$$E \sup_{k \geqslant 1} \left| \tilde{C}_n(k) - C^e_n(k) \right| \leqslant 2E \sum_{i=1}^n |\hat{\tau}_i - \tau_i| = O(\phi_n^{-1}) \ ,$$

by Assumption 2.3.7. In the case of the holdout penalty, we can set

$$\tilde{C}_m(k) = W_m(\hat{f}^e_{m,k}) - W_r(\hat{f}^e_{m,k}) \ .$$

Note that since the propensity score is unknown, the empirical welfare criteria W_m and W_r are infeasible. It can easily be shown that for this choice of $\tilde{C}_m(k)$, we have

$$|\tilde{C}_m(k) - C_m^e(k)| \leqslant \frac{1}{m}\sum_{i=1}^{m} |\hat{\tau}_i{}^E - \tau_i| + \frac{1}{r}\sum_{i=m+1}^{n} |\hat{\tau}_i{}^T - \tau_i| \ ,$$

which yields

$$E \sup_{k \geqslant 1} \left| \tilde{C}_m(k) - C_m^e(k) \right| = O(\phi_n^{-1}) \ .$$

■

Next, we prove Proposition 2.5.1.

Let $\mathcal{G}$ be the set of monotone allocations. Let π_k denote the partition of $[0,1]$ formed by the points $x_i = i/2^k$, $i = 0, \cdots, 2^k$. By definition, for each $G \in \mathcal{G}$, there is an associated function $b_G : [0,1] \to [0,1]$ which determines the boundary of the allocation region, that is, such that $G = \{(x_1, x_2) \in \mathcal{X} : x_2 \leqslant b_G(x_1)\}$. Let $\{\mathcal{G}_k\}_k$ be the approximating sequence defined in Example 2.3.2, and define $G^* \in \mathcal{G}$ to be a set such that $W(G^*) = W_{\mathcal{G}}^*$ (if no such G^* exists, the argument proceeds by considering an "almost maximizer").

Proof of Proposition 2.5.1. Fix some $P \in \mathcal{P}_r$, where $\mathcal{P}_r$ is as defined in Assumption 2.5.1. By definition,

$$W_{\mathcal{G}}^* - W_{\mathcal{G}_k}^* \leqslant W(G^*) - W(\tilde{G}_k) \ ,$$

where $\tilde{G}_k \in \mathcal{G}_k$ is the allocation such that $b_{\tilde{G}_k}(\cdot)$ is the linear interpolation of b_{G^*} on the partition π_k. We can re-write this as

$$\begin{aligned} W(G^*) - W(\tilde{G}_k) &= E\left[\left(\frac{YD}{e(X)} - \frac{Y(1-D)}{1-e(X)}\right) \cdot \left(\mathbf{1}\{X \in G^*\} - \mathbf{1}\{X \in \tilde{G}_k\}\right)\right] \\ &\leqslant \frac{M}{\kappa} P(G^* \Delta \tilde{G}_k) \ , \end{aligned} \tag{B.15}$$

where Δ denotes the symmetric difference operator, $A\Delta B := A\backslash B \cup B\backslash A$. By Assumption 2.5.1, X has density p_X with respect to Lebesgue measure on $[0,1]^2$ such that p_x is bounded by some constant A, so that

$$P(G^*\Delta\tilde{G}_k) \leqslant A\int_0^1 |b_{G^*}(x) - b_{\tilde{G}_k}(x)|dx \ .$$

We thus conclude that if $b_{\tilde{G}_k}$ is a good L^1-approximation of b_{G^*}, then the welfare difference $W(G^*) - W(\tilde{G}_k)$ is small. To that end, it remains to bound the approximation bias. Let

$$M_i = [x_{i-1}, x_i] \times [b_{G^*}(x_{i-1}), b_{G^*}(x_i)] \ ,$$

for $i = 1, \ldots, 2^k$. It follows from the monotonicity of b_{G^*} that the graphs of the restrictions of $b_{G^*}(\cdot)$ and $b_{\tilde{G}_k}(\cdot)$ to $[x_{i-1}, x_i]$ are contained in M_i. Hence we have that

$$\int_0^1 |b_{G^*}(x) - b_{\tilde{G}_k}(x)|dx \leqslant \sum_{i=1}^{2^k} \text{area}(M_i) \ .$$

Now note that

$$\sum_{i=1}^{2^k} \text{area}(M_i) = \sum_{i=1}^{2^k} |b_{G^*}(x_i) - b_{G^*}(x_{i-1})| \cdot |x_i - x_{i-1}| = \frac{1}{2^k}\sum_{i=1}^{2^k} |b_{G^*}(x_i) - b_{G^*}(x_{i-1})| \ .$$

By monotonicity, it is the case that

$$\frac{1}{2^k}\sum_{i=1}^{2^k} |b_{G^*}(x_i) - b_{G^*}(x_{i-1})| \leqslant \frac{1}{2^k} \ ,$$

since by definition $b_{G^*} : [0,1] \to [0,1]$. We thus obtain that

$$W^*_{\mathcal{G}} - W^*_{\mathcal{G}_k} \leqslant A\frac{M}{\kappa}2^{-k} \ ,$$

as desired. ■

Next, we prove Proposition 2.5.2. Define

$$N_{\mathcal{G}}(x_1, ..., x_n) = |\{\{x_1, x_2, ..., x_n\} \cap G : G \in \mathcal{G}\}| \ ,$$

then we present the following lemma, which is proved in Györfi et al. (1996):

Lemma B.1.3. *Let $\mathcal{G}$ be the set of monotone allocations. If X has a bounded density with respect to Lebesgue measure on $[0,1]^2$, then*

$$E[N_{\mathcal{G}}(X_1, ..., X_n)] \leqslant e^{\alpha\sqrt{n}} \ .$$

for some constant α.

Proof. See Theorem 13.13 and the discussion following the proof in Györfi et al. (1996).

■

Proof of Proposition 2.5.2. By Corollary 3.4 in Geer (2000), we have that

$$P\left(\sup_{f\in\mathcal{F}} |W_n(f) - W(f)| > \epsilon\right) \leqslant 4P\left(\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)\right| > \frac{\epsilon}{4}\right) \ ,$$

for $\epsilon \geqslant \sqrt{\frac{8(M/\kappa)^2}{n}}$, where σ_i are Rademacher random variables (this follows from two symmetrizations). Write $f(Z)$ as

$$f_G(Z) = g(Z)\mathbf{1}\{X \in G\} \ ,$$

where $g(Z) = \left(\frac{YD}{e(X)} - \frac{Y(1-D)}{1-e(X)}\right)$. Conditioning on $\{Z_i = z_i = (y_i, x_i, d_i)\}_{i=1}^n$, and applying the union bound, we get that

$$P\left(\sup_{G\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i g(Z_i)\mathbf{1}\{X_i \in G\}\right| > \frac{\epsilon}{4}\middle|\{Z_i = z_i\}_{i=1}^n\right) \leqslant$$
$$N_{\mathcal{G}}(x_1, ..., x_n)\sup_{G\in\mathcal{G}} P\left(\left|\frac{1}{n}\sum_{i=1}^n \sigma_i g(Z_i)\mathbf{1}\{X_i \in G\}\right| > \frac{\epsilon}{4}\middle|\{Z_i = z_i\}_{i=1}^n\right) . \tag{B.16}$$

By Hoeffding's inequality,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n \sigma_i g(Z_i)\mathbf{1}\{X_i \in G\}\right| > \frac{\epsilon}{4}\middle|\{Z_i = z_i\}_{i=1}^n\right) \leqslant 2e^{-n\epsilon^2/c} ,$$

where $c = (4M/\kappa)^2$. Taking expectations, we can conclude that

$$P\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)\right| > \frac{\epsilon}{4}\right) \leqslant 2E[N_{\mathcal{G}}(X_1, ..., X_n)]e^{-n\epsilon^2/c} .$$

Using Lemma B.1.3, we get that

$$P\left(\sup_{f\in\mathcal{F}}|W_n(f) - W(f)| > \epsilon\right) \leqslant 8e^{\alpha\sqrt{n}}e^{-n\epsilon^2/c} ,$$

for $\epsilon \geqslant \sqrt{\frac{8(M/\kappa)^2}{n}}$. Let $\epsilon^*(n) = \sqrt{\frac{8(M/\kappa)^2}{n}}$, then the result follows by a slight modification of the integration argument presented in Problem 12.1 of Györfi et al. (1996) (split the integral of the tail probability as follows: $\int_0^\infty = \int_0^{\epsilon^*(n)} + \int_{\epsilon^*(n)}^u + \int_u^\infty$, bound the first integral by $\epsilon^*(n)$, the second by u, and the third by our tail inequality, and proceed analogously). ■

B.2. Supplementary Results for Chapter 2

B.2.1. Supplement to Example 2.2.3

We work through the claim of Example 2.2.3 in detail. Suppose the outcomes of interest to the planner are described by

$$Y(k) = g(k, A) - \mathbf{1}\{k = 1\}c \ ,$$

where A is an unobserved measure of a student's ability, and c is the per-unit cost of the scholarship to the planner. Let

$$h(a) := g(1, a) - g(0, a) \ .$$

Suppose the planner has two covariates $X = (Z, T)$, on which to base treatment, where Z is parental income and T is a student's GPA. Define

$$\tau(t, z) := E[h(A)|Z = z, T = t] = \int h(a) dF_{A|Z,T}(a|z, t) \ ,$$

to be the average treatment effect (ignoring costs) conditional on $Z = z$, $T = t$ (note that if we consider Assumption 2.3.1 then $h(A)$ has finite support, which guarantees the existence of τ). The unrestricted optimal allocation is given by

$$G^*_{FB} := \{(z, t) : \tau(z, t) \geqslant c\} \ .$$

We claimed in Example 2.2.3 that some plausible assumptions about $h(\cdot)$ and (A, T, Z) could give rise to an optimal allocation which is *monotone*, as defined in Example 2.2.3.

First, the planner makes the following assumption on $h(\cdot)$:

Assumption B.2.1. *$h(a)$ is increasing in a.*

This assumption asserts that the function g has *increasing differences*, which is a common assumption made in economics when doing comparative statics analysis. Intuitively, it says that higher ability students will realize a larger difference in outcomes if they receive the scholarship than lower ability students.

Next, the planner makes the following assumptions about the conditional distribution of $(A|Z,T)$:

Assumption B.2.2. *(FOSD of A in (Z,T))*

- $F_{A|Z,T}(\cdot|z,t) \geq^{FOSD} F_{A|Z,T}(\cdot|z,t')$ *for* $t \geqslant t'$
- $F_{A|Z,T}(\cdot|z,t) \geq^{FOSD} F_{A|Z,T}(\cdot|z',t)$ *for* $z \leqslant z'$

Stochastic-dominance assumptions of this type have been employed by, for example, Blundell et al. (2007) in the study of wage distributions. Intuitively, Assumption B.2.2 asserts that, given a fixed level of parental income, a higher GPA is an indication of higher innate ability, and that given a fixed GPA, lower levels of parental income are an indication of higher innate ability. An assumption of this type could come out of a production function for cognitive achievement, for example as studied in Todd and Wolpin (2003).

Given these assumptions, we can show that $\tau(z,t) \geqslant \tau(z,t')$ if $t \geqslant t'$, and $\tau(z,t) \geqslant \tau(z',t)$ if $z \leqslant z'$. This follows by the fact that, for an increasing function $f(\cdot)$ and two distributions G_1 and G_2, such that G_1 first order stochastically dominates G_2, it is the case that

$$\int f dG_1 \geqslant \int f dG_2 \ .$$

This establishes that the first best allocation is indeed monotone.

B.2.2. Supplement to Example 2.2.1

We elaborate on the example introduced in Example 2.2.1. We construct an approximating sequence that results in what Scott and Nowak (2002) call a *dyadic decision tree*. From now on assume it is the case that $\mathcal{X} = [0,1]^d$. First, we define a *sequential dyadic partition* (SDP). Let $\{R_1, R_2, ..., R_k\}$ be a partition of the the covariate space where each R_i is a hyper-rectangle with sides parallel to the co-ordinate axes. Given a cell R_i, let $R_i^{(1,j)}$ and $R_i^{(2,j)}$ be the hyper-rectangles formed by splitting R_i at its midpoint along the co-ordinate j. A SDP is defined recursively as follows:

- The trivial partition $\{[0,1]^d\}$ is a SDP
- If $\{R_1, R_2, ..., R_k\}$ is a SDP, then so is

$$\{R_1, ..., R_{i-1}, R_i^{(1,j)}, R_i^{(2,j)}, R_{i+1}, ..., R_k\} \ ,$$

  where $1 \leqslant i \leqslant k$ and $1 \leqslant j \leqslant d$.

In words, a SDP is formed by recusively splitting a hyper-cube at its midpoint on some coordinate. A *dyadic decision tree* (DDT) with k splits is a SDP with k partitions, paired with a $\{0,1\}$ label for each hyper-rectangle in the SDP. Given a DDT T_k with k splits, let $G(T_k)$ be the set of covariate points in $\mathcal{X}$ such that those covariates are labeled with a 1 in T_k. Our approximating class is defined as follows:

$$\mathcal{G}_k = \{G \subset \mathcal{X} : G = G(T_k) \text{ for some DDT } T_k \text{ with } k \text{ splits}\} \ .$$

It follows by results in Scott and Nowak (2002) that $\mathcal{G}_k$ has finite VC dimension. Given this approximating sequence, the PWM procedure can be applied to choose the appropriate DDT. In other words, our method could be used to choose the appropriate *depth* of a decision

tree. Kallus (2016) develops Optimal Personalization Trees, which solve a similar problem for a given class $\mathcal{G}_k$, i.e. for trees of a fixed depth. Athey and Wager (2017) use decision trees with a fixed depth as a primary motivating example for their method, and derive a bound on the Hamming entropy of the class of fixed-depth decision trees without the dyadic-split assumption we present here.

We expect that under appropriate regularity conditions we could derive bounds on the maximum regret of this version of PWM with respect to the *unrestricted* optimum. The first question one might ask is how the bounds on maximum regret of PWM with this approximating sequence would compare to the bounds on maximum regret that exist for plug-in rules. As discussed in Kitagawa and Tetenov (2018), if the plug-in rule is implemented with appropriate local-polynomial estimators, and smoothness conditions on the *regression functions* $E(Y(d)|X = x)$ are imposed, a bound on maximum regret can be derived. On the other hand, as explained in Audibert et al. (2007) in the context of classification, although results for plug-in rules typically require assumptions on smoothness of the regression functions, the analogues to our approach in classification typically require regularity conditions on the *boundary* of the decision set G^*_{FB}. In this sense, a comparison of the regularity conditions for plug-in rules and PWM-type rules would suggest that they are complementary approaches.

B.2.3. Supplement to Remark 2.3.2

The demeaned EWM rule is defined as follows: Let $Y_i^{dm} := Y_i - E_n[Y_i]$, then the demeaned EWM rule solves the following problem:

$$\max_{G \in \mathcal{G}} E_n \left[\frac{Y_i^{dm} D_i}{e(X_i)} \mathbf{1}\{X_i \in G\} + \frac{Y_i^{dm}(1 - D_i)}{1 - e(X_i)} \mathbf{1}\{X_i \in G\} \right] .$$

Analogues of the results presented in Sections 2.4 and 2.5 are available from the authors upon request.

B.2.4. Supplement to Remark 2.3.5

In this subsection we provide some simple calculations that justify the comments made in Remark 2.3.5. Consider first the Rademacher penalty, then Proposition 2.3.2 shows that

$$E_{P^n}[W^*_{\mathcal{G}} - W(\hat{G}_n)] \leqslant \inf_k \Big[C\frac{M}{\kappa}\sqrt{\frac{V_k}{n}} + \big(W^*_{\mathcal{G}} - W^*_{\mathcal{G}_k}\big) + \sqrt{\frac{k}{n}} \Big] + g(M,\kappa)\frac{M}{\kappa}\sqrt{\frac{1}{n}} \ ,$$

where C is the universal constant derived in the bound of EWM in Kitagawa and Tetenov (2018) and g is defined as

$$g(M,\kappa) := 6\sqrt{\log\Big(\frac{3\sqrt{e}}{\sqrt{2}}\frac{M}{\kappa}\Big)} \ .$$

Our first task is to quantify the size of C. By the proof of Lemma A.4. in Kitagawa and Tetenov (2018), we can see that the constant C depends on a universal constant K derived in Theorem 2.6.7 of Van der Vaart and Wellner (1996), which establishes a bound on the covering numbers of a VC subgraph class. Inspection of the proof in Van der Vaart and Wellner (1996) allows us to conclude that a suitable K is given by $K = 3\sqrt{e}/8$. Plugging this in to the expression for C derived in Kitagawa and Tetenov (2018) allows us to conclude that a suitable C is given by $C = 36.17$. Turning to $g(M,\kappa)$, we can calculate that in order for it to surpass C by an order of magnitude, we would need M/κ to be about as large as 10^{120}. This give us a sense of the relative sizes of the terms in our bound.

B.3. Computational Details for Chapter 2

In this section we provide details on how we perform the computations of Sections 2.4 and 2.5. All of our work is implemented in Python 2.7 paired with Gurobi. To clarity the

exposition, we begin with Section **??**, which is more straightforward, then proceed to Section 2.4.

B.3.1. Application Details

We will now describe how we compute each $\hat{G}_{n,k}$ to solve PWM over monotone allocations. Recall the definition of $\psi_{T,j}(x)$ as defined in Example 2.3.2. We modify this definition to accommodate the fact that our covariates do not lie in the unit interval. In particular, we restrict ourselves to levels of education that lie in the interval $[5, 20]$, which leads to the following modification.

$$\psi_{T,j}(x) = \begin{cases} 1 - |\frac{T}{15}(x-5) - j|, & x \in \left[\frac{j-1}{T/15} + 5, \frac{j+1}{T/15} + 5\right] \cap [5, 20] \\ 0, & \text{otherwise .} \end{cases}$$

Let $\Theta_T = \begin{bmatrix} \theta_0 & \theta_1 & \dots & \theta_T \end{bmatrix}'$ and let $\bar{\Theta}_T = \begin{bmatrix} -1 & \theta_0 & \theta_1 & \dots & \theta_T \end{bmatrix}'$. Let our two dimensional covariate be denoted as $x = (x^{(1)}, x^{(2)})$ where $x^{(1)}$ is level of education and $x^{(2)}$ is previous earnings. Let

$$\Psi_T(x) = \begin{bmatrix} x^{(2)} & \psi_{T,0}(x^{(1)}) & \dots & \psi_{T,T}(x^{(1)}) \end{bmatrix}' .$$

To compute $\hat{G}_{n,k}$ we solve the following mixed integer linear program (MILP), which modifies the MILP described in Kitagawa and Tetenov (2018) for "Single Linear Index Rules":

$$\begin{aligned}
\max_{\substack{\theta_0,\theta_1,\ldots,\theta_T,\\ z_1,\ldots,z_n}} \quad & \sum_{i=1}^{n} \tau_i \cdot z_i \\
\text{subject to} \quad & \frac{\bar{\Theta}_T' \Psi_T(x_i)}{c_{iT}} < z_i \leqslant \frac{\bar{\Theta}_T' \Psi_T(x_i)}{c_{iT}} + 1, \ i = 1, \ldots, n \\
& z_i \in \{0, 1\}, \ i = 1, \ldots, n \\
& D_T \Theta_T \geqslant 0
\end{aligned}$$

where $T = 2^{k-1}$, τ_i is as defined in equation (2.2), c_T is an appropriate constant (to be discussed in the following sentence), and D_T is the differentiation matrix as defined in Example 2.3.2. c_{iT} is a constant chosen such that $c_{iT} > \sup_{\Theta_T} |\bar{\Theta}_T' \Psi_T(x_i)|$, which allows us to formulate a set of what are known as "big-M" constraints. To implement such a constraint it must necessarily be the case that Θ_T is bounded, so in order to implement PWM we also include an implicit (very large) bound on the possible treatment allocations.[10]

The first two sets of constraints impose that the treatment allocation result in a piecewise linear boundary, the third set of constraints impose that this boundary is monotone. The strength of this formulation is that it imposes monotonicity via a *linear* constraint, which allows us to solve the problem as a MILP.

[10]Big-M constraints have the potential to cause numerical instabilities when solving MILPs that are poorly formulated. We found that it was important to ensure that the covariates are scaled to within the same order of magnitude and that the `IntFeasTol` and `FeasibilityTol` parameters in Gurobi were set to their smallest possible values.

B.3.2. Simulation Details

We describe a MILP to compute each $\hat{G}_{n,k}$ over threshold allocations on d covariates. Define x to be a $(d+1)$-dimensional vector where $x = (1, x^{(1)}, x^{(2)}, ..., x^{(d)})$, with the last d components denoting the d covariates, and suppose $x \in [0,1]^{d+1}$, which is the case in the simulation design. We define the threshold β_k on covariate $x^{(k)}$ to be a $(d+1)$-dimensional vector such that the first component is in $[-1, 1]$, all other components other than the $(k+1)$st are zero, and the $(k+1)$st component is one of $\{1, -1\}$. Let $A = \{1, 2, ..., d\}$ index the dimension of the threshold. We modify the MILP described in Kitagawa and Tetenov (2018) for "Multiple Linear Index Rules":

$$\max_{\substack{\{\beta_a\}_{a\in A},\\ \{z_1^a,\ldots,z_n^a\}_{a\in A},z_1^*,\ldots,z_n^*}} \quad \sum_{i=1}^{n} \tau_i \cdot z_i^*$$

$$\text{subject to} \quad \frac{x_i'\beta_a}{c} < z_i^a \leqslant \frac{x_i'\beta_a}{c} + 1,\ i = 1,\ldots,n,\ a \in A$$

$$1 - |A| + \sum_{a\in A} z_i^a \leqslant z_i^* \leqslant \frac{1}{|A|}\sum_{a\in A} z_i^a,\ i = 1,\ldots,n$$

$$\beta_a^{(1)} \in [-1,1],\ a \in A$$

$$\beta_a^{(j)} = 0,\ j > 1, j \neq a+1,\ a \in A$$

$$\sum_{a\in A} e_a = k$$

$$-e_a \leqslant \beta_a^{(1)} \leqslant e_a,\ a \in A$$

$$\beta_a^{(a+1)} = y_a^{(1)} - y_a^{(2)},\ a \in A$$

$$y_a^{(1)} + y_a^{(2)} = e_a,\ a \in A$$

$$\{z_i^a\}_{a\in A}, z_i^* \in \{0,1\},\ i = 1,\ldots,n$$

$$\{e_a\}_{a\in A} \in \{0,1\},\ a \in A$$

$$\{y_a^{(1)}\}_{a\in A}, \{y_a^{(2)}\}_{a\in A} \in \{0,1\},\ a \in A$$

The constraints serve the following roles: the first two constraints enforce the assignment of observations to treatment, the next two constraints enforce part of the structure of the threshold allocation, the fifth constraint specifies that only k thresholds can be used, and the three subsequent constraints enforce this. Again we require an appropriately chosen constant c to implement a set of big-M constraints, but in this case the choice is straightforward:

$c = d + 2$ will suffice since this guarantees that $c > x_i'\beta_a$ for any possible x_i and β_a, by construction.

Remark B.3.1. In practice, the solution of this MILP could be further optimized using the improvements developed in Bertsimas et al. (2016) and Chen and Lee (2016). ■

APPENDIX C

Appendix to Chapter 3

C.0.1. Proofs of the main Results

To simplify the exposition, we first present the proofs in the special case of

$$y_n = \beta x_n + u_n \ ,$$

where x_n is a scalar. We then explain how to extend the proofs to the case where $\mathbf{x}_n$ is a vector (see Remarks C.0.3 and C.0.5).

A comment about the general strategy: to prove convergence in probability, we prove convergence in mean-square by finding an appropriate bound on the variance that converges to zero. To prove convergence in distribution to a normal, we use the following central limit theorem for dependency graphs proved in Janson (1988). First we give the definition of a dependency graph for a family of random variables:

Defininition C.0.1. A graph Γ is a dependency graph for a family of random variables if the following two conditions are satisfied:

- There exists a one-to-one correspondence between the random variables and the vertices of the graph.
- If $\mathcal{V}_1$, $\mathcal{V}_2$ are two disjoint sets of vertices of Γ such that no edge of Γ had one endpoint in $\mathcal{V}_1$ and the other in $\mathcal{V}_2$, then the corresponding sets of random variables are independent.

Now we state the theorem:

Theorem C.0.1. (Janson 1988 Theorem 2) *Suppose that, for each N, $\{X_{Ni}\}_{i=1}^{N}$ is a family of bounded random variables; $|X_{Ni}| \leqslant A_N$ a.s. Suppose further that Γ_N is a dependency graph for this family and let D_N be the maximal degree of Γ_N (unless Γ_N has no edges at all, in which case we set $D_N = 1$). Let $S_N = \sum_i X_{Ni}$ and $\sigma_N^2 = Var(S_N)$. If there exists and integer $\ell \geqslant 3$ such that*

$$L_N := \frac{(\frac{N}{D_N})^{1/\ell} D_N A_N}{\sigma_N} \to 0 \ \ as \ \ N \to \infty \ , \tag{C.1}$$

then

$$\frac{S_N - E[S_N]}{\sigma_n} \xrightarrow{d} N(0,1) \ \ as \ \ N \to \infty \ .$$

Remark C.0.1. As will be made clear in the proof of Proposition 3.3.2, we use Theorem C.0.1 by definining an appropriate dependency graph for which $D_N = 2(\mathcal{M}^H - 1)$, which establishes the equivalence between Condition 3.2.1 and Equation C.1 for our purposes. Also note that although Janson's theorem applies to an *array* of random variables, in the sense that for a given i, X_{Ni} is allowed to change as N grows, we do not use this feature for our results. ■

Because the proofs under AF1 amount to special cases of the proofs under AF2, we prove Propositions 3.3.2 and 3.3.5 before proving Propositions 3.3.1 and 3.3.3 and 3.3.4.

Proof of Proposition 3.3.2

Proof. We have

$$\tau_N(\hat{\beta} - \beta) = (\frac{1}{N}\sum_{n=1}^{N} x_n^2)^{-1} \frac{\tau_N}{N} \sum_{n=1}^{N} x_n u_n \ .$$

First let's prove $((1/N)\sum_n x_n^2)^{-1} \xrightarrow{p} E[x_n^2]^{-1}$. By the continuous mapping theorem it is enough to show that $(1/N)\sum_n x_n^2 \xrightarrow{p} E[x^2]$. The expectation of $(1/N)\sum_n x_n^2$ is $E[x_n^2]$, so it

suffices to show that

$$Var(\frac{1}{N}\sum_{n=1}^{N} x_n^2) \to 0 \ .$$

Expanding the variance gives

$$Var(\frac{1}{N}\sum_{n=1}^{N} x_n^2) = \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N} Cov(x_n^2, x_m^2) \ .$$

For a fixed $n = n(g,h)$, we have at most $2\mathcal{M}^H - 1$ terms in the inner sum such that the covariance is nonzero. By Assumption 3.3.1, $Cov(x_n^2, x_m^2)$ is uniformly bounded. Hence the sum over n is of order $O(N\mathcal{M}^H)$. Thus under AF2 we have that

$$Var(\frac{1}{N}\sum_{n=1}^{N} x_n^2) = O\big(\frac{\mathcal{M}^H}{N}\big) = O\big(\frac{1}{G}\big) = o(1) \ .$$

Next we'll prove that

$$\frac{\tau_N}{N}\sum_{n=1}^{N} x_n u_n \xrightarrow{d} N(0,\Omega) \ .$$

We apply Janson's theorem to the family of random variables $\{x_n u_n\}_{n=1}^N$. A dependency graph $\Gamma_N = (\mathcal{V}, E)$ for this family is the graph with vertex-set $\mathcal{V} = \{x_n u_n\}_{n=1}^N$, and edge set

$$E = \{\{x_n u_n, x_m u_m\} : x_n u_n, x_m u_m \in V \ \text{ and } \ \psi(n) \cap \psi(m) \neq \varnothing\} \ .$$

That Γ_N is a dependency graph for $\{x_n u_n\}_{n=1}^N$ follows immediately from Assumption 3.2.1. The maximal degree D_N of Γ_N is $2(\mathcal{M}^H - 1)$ by definition and, by Assumption 3.3.1, $|x_n u_n| < A$ for all N for some finite constant A. It remains to check Condition (2) of Janson's theorem. Let $\Omega_N = Var(\sum_n x_n u_n)$, then:

$$L_N = \frac{(\frac{N}{2(\mathcal{M}^H-1)})^{1/\ell} 2(\mathcal{M}^H - 1)A}{\sqrt{\Omega_N}} = \frac{(\frac{N}{2(\mathcal{M}^H-1)})^{1/\ell} 2(\mathcal{M}^H - 1)A}{\sqrt{NG^r}} \cdot \big(\frac{1}{NG^r}\Omega_N\big)^{-1/2} \ .$$

Call the first term in the product R_1 and the second term R_2. R_2 converges to $\Omega^{-1/2}$ by Assumption 3.2.6. To evaluate the limit of R_1, we re-write everything in terms of the rates dictated by AF2, which gives us

$$R_1 = O\Big(\frac{G^{1/\ell}}{G^{r/2}}\Big) \ .$$

Choose ℓ sufficiently large such that $\frac{1}{\ell} - \frac{r}{2} < 0$, which is possible since $r > 0$. It then follows that $R_1 \to 0$. Now that we have established Condition (2) of Janson's theorem, we have that

$$\frac{\sum_{n=1}^{N} x_n u_n}{\sqrt{\Omega_N}} \xrightarrow{d} N(0,1) \ .$$

Re-writing:

$$\frac{\sum_{n=1}^{N} x_n u_n}{\sqrt{\Omega_N}} = \big(\frac{1}{\sqrt{NG^r}} \sum_{n=1}^{N} x_n u_n\big) \cdot (\frac{1}{NG^r}\Omega_N)^{-1/2} = \frac{\tau_N}{N} \sum_{n=1}^{N} x_n u_n \cdot (\frac{1}{NG^r}\Omega_N)^{-1/2}$$

It thus follows by Assumption 3.2.6 that

$$\frac{\tau_N}{N} \sum_{n=1}^{N} x_n u_n \xrightarrow{d} N(0,\Omega) \ .$$

Applying Slutsky's Theorem to

$$\tau_N(\hat{\beta} - \beta) = (\frac{1}{N} \sum_{n=1}^{N} x_n^2)^{-1} \frac{\tau_N}{N} \sum_{n=1}^{N} x_n u_n \ ,$$

we can conclude that

$$\tau_N(\hat{\beta} - \beta) \xrightarrow{d} N(0, V) \ ,$$

as desired. ■

Remark C.0.2. As noted after the statement of Assumption 3.3.1, we could weaken the uniform boundedness assumption by using Janson's theorem with a Lindeberg-type condition

(see Remark 3 in Janson (1988)). For example, if we instead assume that $x_n u_n$ has bounded $2+\delta$ moments, then the result could be proved under the additional assumption that $r > 2/(2+\delta)$, which agrees with our result as we take δ to infinity. ■

Remark C.0.3. The general case is proved as follows: to show the convergence of $(1/N)\sum_n \mathbf{x}_n \mathbf{x}_n'$ we repeat the argument above but component-wise. To show the convergence of $(\tau_N/N)\sum_n \mathbf{x}_n u_n$ we use Janson's Theorem paired with the Cramer-Wold device. ■

Proof of Proposition 3.3.1

PROOF. This proof follows by a similar argument to the proof of Proposition 3.3.2. We will sketch it here. Expanding:

$$\sqrt{N}(\hat{\beta}-\beta) = (\frac{1}{N}\sum_{n=1}^{N} x_n^2)^{-1}\frac{1}{\sqrt{N}}\sum_{n=1}^{N} x_n u_n \ .$$

The first term of the product converges to $E(x^2)^{-1}$ by the same argument as above. The second term of the product converges to a normal by an application of Janson's theorem C.0.1 where we note that now under AF1 the maximal degree $D_N = 2(\mathcal{M}^H - 1)$ of the dependency graph is bounded. Hence by Janson's theorem we have that

$$\frac{\sum_{n=1}^{N} x_n u_n}{\Omega_N} \xrightarrow{d} N(0,1) \ .$$

By a similar calculation to the one done in the proof of Proposition 3.3.2, we get that

$$\frac{1}{\sqrt{N}}\sum_{n=1}^{N} x_n u_n \xrightarrow{d} N(0,\Omega) \ ,$$

and thus Proposition 3.3.1 follows by an application of Slutsky's theorem. ■

Proof of Proposition 3.3.5

PROOF. We follow the general strategy of Aronow et al. (2015). First we introduce some notation for the proof. Given a dyadic index n, define $\tilde{n} = \psi(n)$. Recall that

$$\hat{V} = (\sum_{n=1}^{N} x_n^2)^{-1} \hat{\Omega} (\sum_{n=1}^{N} x_n^2)^{-1} \ ,$$

where

$$\hat{\Omega} = \sum_{n=1}^{N} \sum_{m=1}^{N} \mathbf{1}_{n,m} \hat{u}_n \hat{u}_m x_n x_m \ .$$

We proved in Proposition 3.3.2 that $\left((1/N)\sum_{n=1}^{N} x_n^2\right)^{-1} \xrightarrow{p} E(x^2)^{-1}$. So it remains to show that

$$\frac{\tau_N^2}{N^2} \hat{\Omega} \xrightarrow{p} \Omega \ .$$

Re-writing $\hat{u}_n = u_n - (\hat{\beta} - \beta) x_n$ and expanding gives:

$$\frac{\tau_N^2}{N^2} \hat{\Omega} = \frac{\tau_N^2}{N^2} \Big(R_1 + R_2 + R_3 + R_4 \Big) \ ,$$

where

$$R_1 = \sum_{n=1}^{N} \sum_{m=1}^{N} \mathbf{1}_{n,m} x_n x_m u_n u_m \ ,$$

$$R_2 = -\sum_{n=1}^{N} \sum_{m=1}^{N} \mathbf{1}_{n,m} x_n x_m^2 u_n (\hat{\beta} - \beta) \ ,$$

$$R_3 = -\sum_{n=1}^{N} \sum_{m=1}^{N} \mathbf{1}_{n,m} x_n^2 x_m u_m (\hat{\beta} - \beta) \ , \text{ and}$$

$$R_4 = \sum_{n=1}^{N} \sum_{m=1}^{N} \mathbf{1}_{n,m} x_n^2 x_m^2 (\hat{\beta} - \beta)^2 \ .$$

We will show that $(\tau_N^2/N^2) R_1 \xrightarrow{p} \Omega$ while the rest converge in probability to zero. As usual, we show convergence in mean-square. For the first term, it is the case by definition that

$$\lim_{G \to \infty} E\Big[\frac{\tau_N^2}{N^2} R_1 \Big] = \Omega \ ,$$

so it suffices to show that

$$\lim_{G\to\infty} Var\Big(\frac{\tau_N^2}{N^2}R_1\Big) = 0 \ .$$

Expanding:

$$Var\Big(\frac{\tau_N^2}{N^2}R_1\Big) = \frac{\tau_N^4}{N^4}\Big(\sum_i\sum_j\sum_k\sum_l Cov(\mathbf{1}_{i,j}x_ix_ju_iu_j, \mathbf{1}_{k,l}x_kx_lu_ku_l)\Big) \ .$$

By Assumption 3.3.1, the summands are uniformly bounded, so in order to get a suitable bound on the sum we need to understand under what conditions

$$Cov(\mathbf{1}_{i,j}x_ix_ju_iu_j, \mathbf{1}_{k,l}x_kx_lu_ku_l) \neq 0 \ .$$

First it is clear that we must have

$$\tilde{i} \cap \tilde{j} \neq \varnothing \text{ and } \tilde{k} \cap \tilde{l} \neq \varnothing \ . \tag{C.2}$$

Given (2) holds, expand the covariance:

$$Cov(x_ix_ju_iu_j, x_kx_lu_ku_l) = E[x_ix_ju_iu_jx_kx_lu_ku_l] - E[x_ix_ju_iu_j]E[x_kx_lu_ku_l] \ ,$$

then we see that we must also have that

$$\tilde{i} \cap \tilde{k} \neq \varnothing \text{ or } \tilde{i} \cap \tilde{l} \neq \varnothing \text{ or } \tilde{j} \cap \tilde{k} \neq \varnothing \text{ or } \tilde{j} \cap \tilde{l} \neq \varnothing \ . \tag{C.3}$$

Let S be the set of tuples $(i,j,k,l) \in N^4$ such that conditions (2) and (3) hold, then the cardinality of S, denoted $|S|$, is an upper bound on the number of nonzero terms in the sum. Our goal is to find an upper-bound on $|S|$.

Fix an index i, then there are $O(\mathcal{M}^H)$ indices j such that $\tilde{i} \cap \tilde{j} \neq \varnothing$. Now, for a fixed i and j such that $\tilde{i} \cap \tilde{j} \neq \varnothing$, there are $O(\mathcal{M}^H)$ possible indices k such that $\tilde{i} \cap \tilde{k} \neq \varnothing$ or

$\tilde{j} \cap \tilde{k} \neq \varnothing$. For a fixed i, j and k such that the above hold, there are $O(\mathcal{M}^H)$ possible indices l such that $\tilde{k} \cap \tilde{l} \neq \varnothing$.

Similarly, for a fixed i and j such that $\tilde{i} \cap \tilde{j} \neq \varnothing$, there are $O(\mathcal{M}^H)$ possible indices l such that $\tilde{i} \cap \tilde{l} \neq \varnothing$ or $\tilde{j} \cap \tilde{l} \neq \varnothing$. For a fixed i, j and l such that the above hold, there are $O(\mathcal{M}^H)$ possible indices k such that $\tilde{k} \cap \tilde{l} \neq \varnothing$.

Thus there are $N \cdot O(\mathcal{M}^H) \cdot O(\mathcal{M}^H) \cdot O(\mathcal{M}^H) = O(N(\mathcal{M}^H)^3)$ possible indices i, j, k, l such that $(i, j, k, l) \in S$. Re-writing using the rates dictated by AF2 gives us that $|S| = O(G^5)$ and that

$$\frac{\tau_N^4}{N^4} \leqslant K \frac{1}{G^{4+2r}} \ ,$$

for some positive constant K. Therefore we can conclude that

$$Var\big(\frac{\tau_N^2}{N^2} R_1\big) \leqslant \frac{1}{G^{4+2r}} O(G^5) = o(1)$$

for $r > \frac{1}{2}$. Thus we have shown that $(\tau_N^2/N^2)R_1 \xrightarrow{p} \Omega$. Next, we must show that the remaining terms converge to 0 in probability. All three terms follow by similar arguments so we will only present the argument for R_2. We wish to show that

$$\frac{\tau_N^2}{N^2}\Big(\sum_{n=1}^{N} \sum_{m=1}^{N} \mathbf{1}_{n,m} x_n x_m^2 u_n (\hat{\beta} - \beta)\Big) \xrightarrow{p} 0 \ .$$

We know from Proposition 3.3.2 that $\tau_N^{1-\epsilon}(\hat{\beta} - \beta) = o_P(1)$ for any $\epsilon > 0$, so it suffices to show that

$$\frac{\tau_N^{1+\epsilon}}{N^2}\Big(\sum_{n=1}^{N} \sum_{m=1}^{N} \mathbf{1}_{n,m} x_n x_m^2 u_n\Big) = O_P(1) \ ,$$

for some $\epsilon > 0$. Note that from Assumption 3.3.1 and the definition of τ,

$$E\Big[\frac{\tau_N^{1+\epsilon}}{N^2}\Big(\sum_{n=1}^{N} \sum_{m=1}^{N} \mathbf{1}_{n,m} x_n x_m^2 u_n\Big)\Big] \to 0 \ ,$$

for ϵ sufficiently small, and that

$$Var\Big(\frac{\tau_N^{1+\epsilon}}{N^2}\Big(\sum_{n=1}^{N}\sum_{m=1}^{N}\mathbf{1}_{n,m}x_n x_m^2 u_n\Big)\Big) \to 0 \ ,$$

which can be shown by similar arguments to what we have done above. Similarly, the third and fourth terms also converge to zero in probability, and hence we have that

$$\frac{\tau_N^2}{N^2}\hat{\Omega} \xrightarrow{p} \Omega \ ,$$

and ultimately that

$$\tau_N^2 \hat{V} \xrightarrow{p} V$$

as desired. ■

Remark C.0.4. Note that this proof relied on the bounded support Assumption 3.3.1 to ensure that the covariance terms in the summands were uniformly bounded. In general we would need to make assumptions on these covariances if we were to weaken the support assumption used here. ■

Remark C.0.5. The general case is proved as follows: to show the convergence of R_1 to Ω we repeat the argument above but component-wise. To show the convergence of R_2, R_3 and R_4 to zero we modify the argument slightly. Consider

$$R_2 = \sum_{n=1}^{N}\sum_{m=1}^{N}\mathbf{1}_{n,m}\mathbf{x}_n\mathbf{x}_m'(\hat{\beta}-\beta)'\mathbf{x}_n u_n \ .$$

To show $R_2 \xrightarrow{p} \mathbf{0}$ we will show that $||R_2|| \xrightarrow{p} 0$ where $||\cdot||$ is the Frobenius norm. Using the triangle inequality, the matrix Schwartz Inequality, and the definition of the Frobenius norm

we get that

$$||R_2|| \leqslant \sum_{n=1}^{N} \sum_{m=1}^{N} \mathbf{1}_{n,m} ||\mathbf{x}_n||^2 ||\mathbf{x}_m|| \cdot ||\hat{\beta} - \beta|| \cdot |u_n| \ .$$

The result then follows from the arguments presented above. ■

Proof of Proposition 3.3.3

PROOF. Again we follow the same strategy as the proof of Proposition 3.3.5. Now when getting a bound on the variance of R_1, we have that, for each fixed i, there are only finitely many j, k, l such that $(i, j, k, l) \in S$. Therefore there are $O(N)$ nonzero terms in the sum. It then follows that the variance of $(1/N)R_1$ converges to zero and the proof goes through in the same manner as before. ■

Proof of Proposition 3.3.4

PROOF. This is just a special case of Proposition 3.3.5. ■

Proof of Proposition 3.3.6

PROOF. We follow the same strategy as the proof of Proposition 3.3.5. Now when getting a bound on the variance of R_1, we have that the terms in the sum are nonzero if and only if $i = j = k = l$, so that there are N nonzero terms in the sum. It then follows that the variance of $(1/N)R_1$ converges to zero and the proof goes through in the same manner as before. ■

C.0.2. Simulation Design Details

In this section we provide some details about the construction of our designs not mentioned in the main body. All of the simulations in the paper were performed using numpy in Python 2.7.

Construction of Model S: Dyads were included in Model S according to the following rule:

- (g, h) is included if $|g - h| = 1$
- $(1, G)$ is included
- $(g, 2g)$ is included if $g \leqslant \left\lfloor \frac{G}{2} \right\rfloor$
- $(g, 3g)$ is included if $g \leqslant \left\lfloor \frac{G}{3} \right\rfloor$

Construction of Model B: Dyads were included in Model B according to the following rule:

- (g, h) is included if $|g - h| = 1$ and $g, h < G - 1$
- $(1, G - 2)$ is included
- If $G = 100$, $G = 250$, or $G = 800$, (g, h) is included if $|g - h| = 2$ and $g, h < G - 1$
- If $G = 100$ or $G = 250$, or $G = 800$, $(1, G - 3)$ and $(2, G - 2)$ are included
- If $G = 800$, (g, h) is included if $|g - h| \leqslant 4$ and $g, h < G - 1$
- If $G = 800$, $(1, G - 4)$, $(1, G - 5)$, $(2, G - 3)$, and $(2, G - 4)$ are included
- $(g, G - 1)$ is included for all $g \leqslant \left\lfloor \frac{G}{2} \right\rfloor$, and (g, G) is included for all $g > \left\lfloor \frac{G}{2} \right\rfloor$

Construction of G_A and G_B: Recall that in Section 4.2 we constructed our design by partitioning the units into two groups G_A and G_B and then specifying the error term as

$$u_{n(g,h)} = \begin{cases} -(\alpha_g + \alpha_h) + \epsilon_n & \text{if } g \text{ and } h \text{ belong to different groups.} \\ \alpha_g + \alpha_h + \epsilon_n & \text{if } g \text{ and } h \text{ belong to the same group.} \end{cases}$$

Where $\alpha_g \sim U[-\sqrt{3}, \sqrt{3}]$ i.i.d for $g = 1, 2, ...G$ and $\epsilon_n \sim U[-\sqrt{3}, \sqrt{3}]$ i.i.d for $n = 1, 2, ..., N$. In order to achieve a rate of growth of NG^r for $Var(\sum_n \mathbf{x}_n u_n)$, we construct G_A and G_B as follows:

$$G_A = \left\{ g \in G : g \leqslant \left\lfloor \frac{G - G^s}{2} \right\rfloor \right\},$$

$$G_B = \left\{ g \in G : g > \left\lfloor \frac{G - G^s}{2} \right\rfloor \right\},$$

where $s = \frac{1+r}{2}$. Expanding $Var(\sum_n \mathbf{x}_n u_n)$ shows that it indeed grows at rate NG^r.