Stratification Trees for Adaptive Randomization in Randomized Controlled Trials

Max Tabord-Meehan
Department of Economics
University of Chicago
maxtm@uchicago.edu

18th November 2020

Abstract

This paper 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). 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. By using the first wave to estimate a stratification tree, we simultaneously select which covariates to use for stratification, how to stratify over these covariates, and the assignment probabilities within these strata. Our main result shows that using this randomization procedure with an appropriate estimator results in an asymptotic variance which is minimal in the class of stratification trees. Moreover, our results are able to accommodate a large class of assignment mechanisms within strata, including stratified block randomization. In a simulation study, we find that our method, paired with an appropriate cross-validation procedure, can improve on ad-hoc choices of stratification. We conclude by applying our method to the study in Karlan and Wood (2017), where we estimate stratification trees using the first wave of their experiment.

KEYWORDS: randomized experiments; decision trees; adaptive randomization
JEL classification codes: C14, C21, C93

1 Introduction

This paper 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 D.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 2.1, 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 insights from supervised machine-learning 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 "optimal" (to be made precise later) 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 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, as we explain in Section 3.1, using partitions based on decision trees gives us theoretical and computational tractability. Third, as we explain in Section 3.2, 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; Gyorfi 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 2.5 is a popular method of randomization.

Although our main focus is on increasing efficiency, 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. To these ends, we also present results for targeting subgroup-specific effects, as well as results for multiple treatments.

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) (see Silvey, 2013, for a textbook treatment), maintains a parametric relationship for the outcomes with respect to the covariates, and targets efficient estimation of the model parameters. In contrast, our paper follows a recent literature which instead maintains a non-parametric model of potential outcomes, and targets efficient estimation of treatment effects. This recent literature can be broadly divided into "one-stage" procedures, which do not use prior data on all experimental outcomes 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 prior data on the experimental outcomes to determine how to randomize. For example, they may use 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 typically 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 work 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).

Three papers which propose response-adaptive randomization methods in a framework similar to ours are Hahn et al. (2011), Chambaz et al. (2014) and Bai (2019) (see also Viviano, 2020, for related work in the presence of network interference). Hahn et al. (2011) develop a procedure which uses the information from a first-wave experiment 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 an optimal propensity score, where the propensity score is constrained through 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. Moreover, the results in both Hahn et al. (2011) and Chambaz et al. (2014) assume that assignment was performed completely independently across individuals in a given wave. In contrast, we reiterate that our results will accommodate a large class of stratified randomization schemes. Bai (2019) derives the MSE-optimal blocking of an experimental sample for the difference-in-means estimator, given a fixed assignment proportion, and shows that this blocking takes the form of a "matched-pairs" style design. He then proposes procedures which use information from a first-wave experiment to approximate the optimal blocking in a second-wave experiment. He also shows that it is possible to combine his procedure with the one proposed in this paper, by implementing his optimal blocking *within* each stratum produced by our method. Importantly, he shows that the resulting combined procedure has an asymptotic variance which is no greater, and typically strictly smaller, than using our procedure alone.

The paper proceeds as follows: In Section 2, we provide a motivating discussion, set up the notation, and formally define the set of randomization procedures we consider. In Section 3, we present the formal results underlying the method as well as several relevant extensions. In Section 4, we perform a simulation study to assess the performance of our method in finite samples. In Section 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 and perform an application-based simulation. Section 6 concludes.

2 Preliminaries

In this section we discuss some preliminary concepts and definitions. Section 2.1 presents a series of simplified examples which we use to motivate our procedure. Section 2.2 establishes some notation and provides the definition of a *stratification tree*, as well as our notion of a *randomization procedure*.

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 (whenever $n_1 = n\pi$ exactly), 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 a standard equal 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 $\{(Y_j, A_j)\}_{j=1}^m$ which allowed us to estimate σ_0^2 and σ_1^2, 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 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. ■

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

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

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 takes the form of 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 $\{(Y_j, A_j, X_j)\}_{j=1}^m$, 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 of our estimator $\hat{\theta}_C$. 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.

2.2 Notation and Definitions

In this section we establish our notation 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 Appendix D), 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 $(Y_i(1), Y_i(0), X_i)$. Throughout the paper we assume that all of our observations are generated by i.i.d draws from Q. We restrict Q as follows:

Assumption 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 2.2. The support assumptions imposed on $(Y(1), Y(0), X)$ in Assumption 2.1 are used frequently throughout the proofs of our results. However, they may be stronger than is desirable in some applications. For example, our assumption that X be supported on a rectangle may fail in certain practical examples (see for example the set of covariates considered in Section 5). Nevertheless, the simulation results presented in Sections 4 and 5 suggest that these assumptions could be reasonably weakened. Moreover, the user does not need to specify a choice of M to implement the procedure. ■

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

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

An experiment on an i.i.d sample $\{(Y_i(1), Y_i(0), X_i)\}_{i=1}^{n}$ 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* which generates $\{A_i\}_{i=1}^{n}$. 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 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 2.1. Figure 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?".

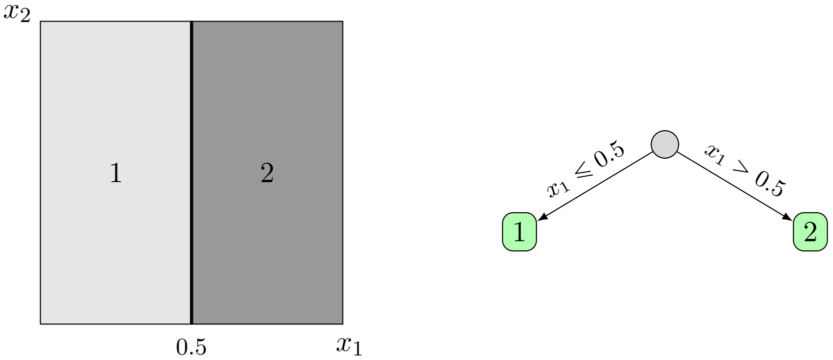

Figure 1: Two representations of a tree partition of depth 1 on $[0, 1]^2$.
Graphical representation (left), tree representation (right).

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

Definition 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*

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

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

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

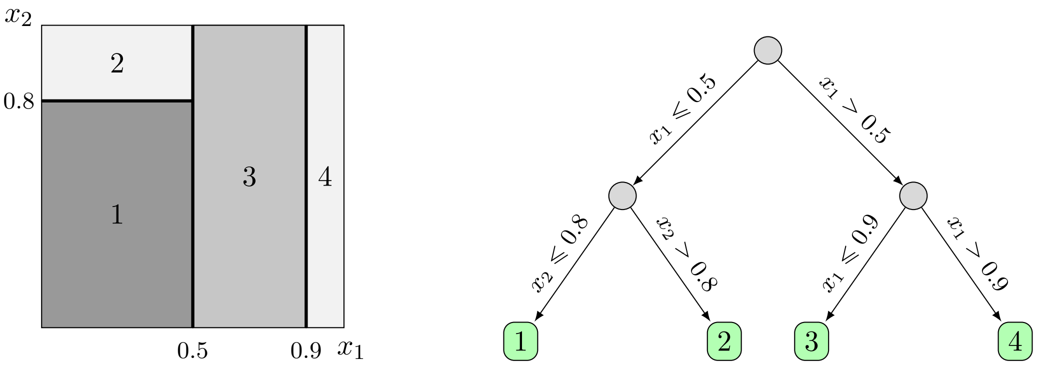

Figure 2: Two representations of a tree partition of depth 2 on $[0,1]^2$.
Graphical representation (left), tree representation (right).

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 3.2 and Appendix E, 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 employ an effective search heuristic known as an evolutionary algorithm. Third, the recursive aspect of tree partitions makes the targeting of subgroup-specific effects convenient, as we show in Section 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 2.3. 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 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 2.3. Figure 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.

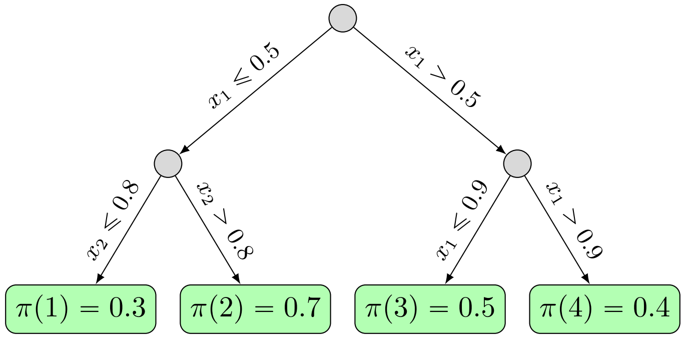

Figure 3: Representation of a Stratification Tree of Depth 2

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

Assumption 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 2.4. 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 2.5. 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 3.2. ■

For technical reasons relating to the potential non-measurability of our estimator, we will impose one additional restriction on $\mathcal{T}_L$.

Assumption 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. By an abuse of notation, we continue to denote the set of stratification trees we will consider as $\mathcal{T}_L$.*

Remark 2.6. We emphasize that this assumption is *only* used as a sufficient condition to guarantee measurability, in order to invoke Fubini's theorem in the proof of Theorem 3.1. Note that, in practice, restricting the set of stratification trees to those constructed from a finite grid satisfies Assumption 2.3. However, our results will also apply more generally. ■

Recall that we are interested in randomization procedures that stratify on baseline covariates and then assign treatment status independently across strata. For each $T \in \mathcal{T}_L$, and given an i.i.d sample $\{(Y_i(0), Y_i(1), X_i)\}_{i=1}^n$ 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 $T = (S, \pi) \in \mathcal{T}_L$. For

[1] Here "closed" is with respect to an appropriate topology on $\mathcal{T}_L$, see Appendix B for details. It is possible that Assumption 2.3 could be weakened

$T = (S, \pi)$, 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. We impose two assumptions on the randomization procedure $A^{(n)}(T)$.

First, we require the following exogeneity assumption:

Assumption 2.4. *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. Next, let $p(k;T) := P(S_i = k)$ be the population proportions of each stratum, then we also require that the randomization procedure satisfy the following "consistency" property:

Assumption 2.5. *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 2.4 and 2.5, we do not require any additional assumptions about how assignment is performed within strata. Examples 2.4 and 2.5 illustrate two randomization schemes which satisfy these assumptions and are popular in economics. Bugni et al. (2018) make similar assumptions for a *fixed* stratification and show that they are satisfied for a wide range of assignment procedures, including procedures often considered in the literature on clinical trials: see for example Efron (1971), Wei (1978), Antognini and Giovagnoli (2004), and Kuznetsova and Tymofyeyev (2011). In Proposition 2.1 below, we verify that Assumptions 2.4 and 2.5 hold for stratified block randomization (see Example 2.5), which is a common assignment procedure in economic applications.

Example 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 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 is a popular method of assignment in economics (for example, every RCT published in the Quarterly Journal of Economics in 2017 used SBR).

We conclude this section by showing that Assumptions 2.4 and 2.5 are satisfied by SBR:

Proposition 2.1. *Suppose randomization is performed through SBR (see Example 2.5), then Assumptions 2.4 and 2.5 are satisfied.*

3 Results

In this section we formally define our proposed procedure and present results about its asymptotic behavior. Section 3.1 sets up the problem and presents the main results about the asymptotic normality of our estimator. Section 3.2 considers several extensions: a cross-validation procedure to select the depth L of the stratification tree, asymptotic results for a "pooled" estimator of the ATE, and extensions for the targeting of subgroup specific effects.

3.1 Main Results

In this section we describe our procedure, and present our main formal results. Recall our discussion at the end of Section 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 $T \in \mathcal{T}_L$, let $\{(Y_i, A_i, X_i)\}_{i=1}^n$ be an experimental sample generated from a randomized experiment with randomization procedure $A^{(n)}(T)$. 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) establish asymptotic normality for a *fixed* $T = (S, \pi) \in \mathcal{T}_L$:

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

where

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

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 takes the form of the semi-parametric efficiency bound of Hahn (1998) amongst all 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 $\{(Y_j, A_j, X_j)\}_{j=1}^m$ (indexed by j) to estimate some "optimal" tree $\hat{T}$ which is designed to minimize $V(T)$. In the second stage, we perform a randomized experiment using stratified randomization with $A^{(n)}(\hat{T})$ to obtain second-wave data $\{(Y_i, A_i, X_i)\}_{i=1}^n$ (indexed by i). Finally, to analyze the results of the experiment, we consider both the "unpooled" estimator $\hat{\theta}(\hat{T})$ defined above, which uses only the second-wave data to estimate the ATE, as well as a "pooling" estimation strategy, which use both waves of data to construct an ATE estimator (see Section 3.2).

We now present the main theoretical properties of our method. In particular, we establish conditions under which the estimator $\hat{\theta}(\hat{T})$ constructed using the second wave of data is asymptotically normal, with minimal variance in the class of estimators defined above. Additionally, we provide a consistent estimator of the asymptotic variance of our estimator, and establish a form of "robustness" of our estimator to potential inconsistency of $\hat{T}$. Note that 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 3.2.

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. As in the paragraph above, denote the pilot data as $\{W_j\}_{j=1}^m := \{(Y_j, X_j, A_j)\}_{j=1}^m$. Given this pilot sample, we require the following high-level consistency property for our estimator $\hat{T}$:

Assumption 3.1. *The estimator $\hat{T}_m$ is a $\sigma\{(W_j)_{j=1}^m\}/\mathcal{B}(\mathcal{T}_L)$ measurable function of the pilot data*[2] *and satisfies*

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

where

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

as $m \to \infty$.

Note that Assumption 3.1 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 $\hat{T}$ converges to any *fixed* tree. Moreover, Assumption 3.1 imposes no explicit restrictions on how $\hat{T}$ is constructed, or even on the nature of the pilot data itself. In Proposition 3.1 below, we establish sufficient conditions on the pilot data under which an appropriate $\hat{T}$ can be constructed by solving the following empirical minimization problem:

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

where $\widetilde{V}(T)$ is an empirical analog of $V(T)$ (defined in Appendix E) constructed using the pilot data. In Section 3.2, we consider an alternative construction of $\hat{T}$ which uses cross-validation to select the depth of the tree. In general, computing $\hat{T}^{EM}$ involves solving a complicated discrete optimization problem. In Appendix E we describe an evolutionary algorithm which effectively performs a stochastic search for the global minimizer of the empirical minimization problem.

We verify Assumption 3.1 for $\hat{T}^{EM}$ when the pilot data comes from a RCT performed using simple random assignment:

Proposition 3.1. *Suppose the pilot data come from a RCT performed using simple random assignment. Under Assumptions 2.1, 2.2, and 2.3, Assumption 3.1 is satisfied for $\hat{T}^{EM}$.*

To prove our asymptotic normality result 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 3.2. *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 all $S^{-1}(k)$ nonempty. Then we assume that $\mathcal{G}$ forms a pointwise equicontinuous family.*

Remark 3.1. To our knowledge this assumption is non-standard. In Lemma C.4 we show that a sufficient condition for Assumption 3.2 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 $S \in \mathcal{S}_L$. ■

[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.

We now state the first main result of the paper: an optimality result for the unpooled estimator $\hat{\theta}(\hat{T})$. In Remark 3.2 we comment on some of the technical challenges that arise in the proof of the result.

Theorem 3.1. *Given Assumptions 2.1, 2.2, 2.3, 2.4, 2.5, 3.1, and 3.2, we have that*

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

as $m, n \to \infty$.

Remark 3.2. Here we comment on some of the technical challenges that arise in proving Theorem 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 $\hat{T}$ 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 2.3, 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 B.1) which we use in our definitions.

Next, because Assumptions 2.4 and 2.5 impose so little on the dependence structure of the randomization procedure, it is not clear how to apply standard central limit theorems. 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.2 for details). We emphasize that we do not require any assumptions on the convergence rate of $\hat{T}$ to the set of optimal trees when establishing this result. ■

Next we construct a consistent estimator for the variance V^*. Let

$$\hat{V}_H = \sum_{k=1}^{K} \frac{n(k)}{n} \left(\hat{\beta}(k) - \hat{\theta}\right)^2 \ ,$$

and let

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

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

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

We obtain the following consistency result:

Theorem 3.2. *Given Assumptions 2.1, 2.2, 2.3, 2.4, 2.5, 3.1, and 3.2, then*

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

where

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

as $m, n \to \infty$.

We finish this section by presenting a result about the limiting behavior of $\hat{\theta}(\hat{T})$ when $\hat{T}$ is not necessarily itself consistent in the sense of Assumption 3.1:

Proposition 3.2. *Let $\widetilde{T}_m$ be any sequence of trees constructed from the pilot data. Let $H_n(t;T)$ be the cdf of $\sqrt{n}(\hat{\theta}(T) - \theta)$, and let $\Phi(t;T)$ be the cdf of a $N(0, V(T))$ random variable. Given Assumptions 2.1, 2.2, 2.3, 2.4, 2.5, and 3.2, we have that*

$$\sup_{t \in \mathbb{R}} |H_n(t; \widetilde{T}_m) - \Phi(t; \widetilde{T}_m)| \xrightarrow{a.s} 0 \ ,$$

as $m, n \to \infty$.

We conclude from Proposition 3.2 that, regardless of whether or not $\hat{T}$ is consistent for an optimal tree, we may use a normal approximation of $\sqrt{n}(\hat{\theta}(\hat{T}) - \theta)$ to conduct valid inference. Indeed, we will see in the simulations of Section 4 that even in situations where $\hat{T}$ is a very poor estimate of an optimal tree, the coverage of a confidence interval constructed using our estimator is not affected.

3.2 Extensions

In this section we present some extensions to the main results. First we present a version of $\hat{T}$ whose depth is selected by cross-validation. Second, we describe a method to combine estimates of the ATE from both waves of data, and establish properties of the resulting "pooled" estimator. Finally, we explain how to accommodate the targeting of subgroup-specific effects.

3.2.1 Cross-validation to select L

In this subsection we describe a method to select the depth L via cross-validation. We focus on selecting a depth L such that the optimal tree can be well estimated using the pilot data, since in practice this seems to be the binding constraint. The tradeoff of choosing L in the first-stage estimation problem 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 $\hat{T}_L$ be the stratification tree estimated from class $\mathcal{T}_L$, then by Assumption 3.1, we have that

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

for each $L \leqslant \bar{L}$. Despite the fact that $\hat{T}_L$ 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 $\{W_j\}_{j=1}^m$ be the pilot data, and for simplicity suppose m is even. Split the pilot sample into two halves and denote these by $\mathcal{D}_1 := \{W_j\}_{j=1}^{m/2}$ and $\mathcal{D}_2 := \{W_j\}_{j=m/2+1}^m$, respectively. Now for each L, let $\hat{T}_L^{(1)}$ and $\hat{T}_L^{(2)}$ 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:

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

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:

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

with

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

where in the event of a tie we choose the smallest such L. Hence $\hat{T}^{CV}$ is chosen to be the stratification tree whose depth minimizes the cross-validation criterion $\widetilde{V}_L^{CV}$. If each $\hat{T}_L$ is estimated by minimizing the empirical variance over $\mathcal{T}_L$, as described in Section 3.1, then we can show that the cross-validated estimator satisfies the consistency property of Assumption 3.1:

Proposition 3.3. *Suppose the pilot data come from a RCT performed using simple random assignment. Under Assumptions 2.1, 2.2, and 2.3, Assumption 3.1 is satisfied for $\hat{T}^{CV} = \hat{T}^{EM}_{\hat{L}}$ in the set $\mathcal{T}_{\bar{L}}$, that is,*

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

as $m \to \infty$.

In light of Proposition 3.3 we see that all of our previous results continue to hold while using $\hat{T}^{CV}$ as our stratification tree. However, Proposition 3.3 *does not* help us conclude that $\hat{T}^{CV}$ should perform any better than $\hat{T}_{\bar{L}}$ 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 $\hat{T}^{CV}$ via simulation in Section 4, and note that it does indeed seem to protect against overfitting in practice. In Section 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).

Remark 3.3. 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. Here we focus on 2-fold cross validation because of the computational difficulties we face in solving the optimization problem to compute $\hat{T}^{EM}$. ■

3.2.2 A pooling estimator of the ATE

In this subsection we study an estimator which allows us to "pool" data from both datasets when estimating the ATE. Pooling may be particularly useful in formal two-stage randomized experiments where the first wave sample-size is large relative to the total sample-size (for example, in the application we consider in Section 5).

Let $\hat{\theta}_1$ be an estimator of the ATE constructed from the pilot data, and let $\hat{\theta}(\hat{T})$ be the estimator defined in Section 3.1. We impose the following high level assumption on the asymptotic behavior of $\hat{\theta}_1$:

Assumption 3.3. *$\hat{\theta}_1$ is an asymptotically normal estimator for the ATE:*

$$\sqrt{m}(\hat{\theta}_1 - \theta) \xrightarrow{d} N(0, V_1) \ ,$$

as $m \to \infty$. Moreover, V_1 can be consistently estimated.

Assumption 3.3 holds for a variety of standard estimators under various assignment schemes: see for example the results in Bugni et al. (2017), Bugni et al. (2018), and Bai et al. (2019). We also impose the following assumption on the relative rates of growth of the pilot and main sample.

Assumption 3.4. *Let m be the pilot data sample size, n the main data sample size, and $N = m+n$. We assume that*

$$\frac{m}{N} \to \lambda \ ,$$

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

We propose the following sample-size weighted estimator:

$$\hat{\theta}_{SW} = \widehat{\lambda}\hat{\theta}_1 + (1 - \widehat{\lambda})\hat{\theta}(\hat{T}) \ ,$$

where $\hat{\lambda} = m/N$.

Theorem 3.3 derives the limiting distribution of this estimator:

Theorem 3.3. *Given Assumptions 2.1, 2.2, 2.3, 2.4, 2.5, 3.1, 3.2, 3.3, and 3.4, we have that*

$$\sqrt{N}(\hat{\theta}_{SW} - \theta) \xrightarrow{d} N(0, \lambda V_1 + (1 - \lambda)V^*) \ ,$$

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

In words, we see that the pooled estimator $\hat{\theta}_{SW}$ 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.

3.2.3 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 4 for an example).

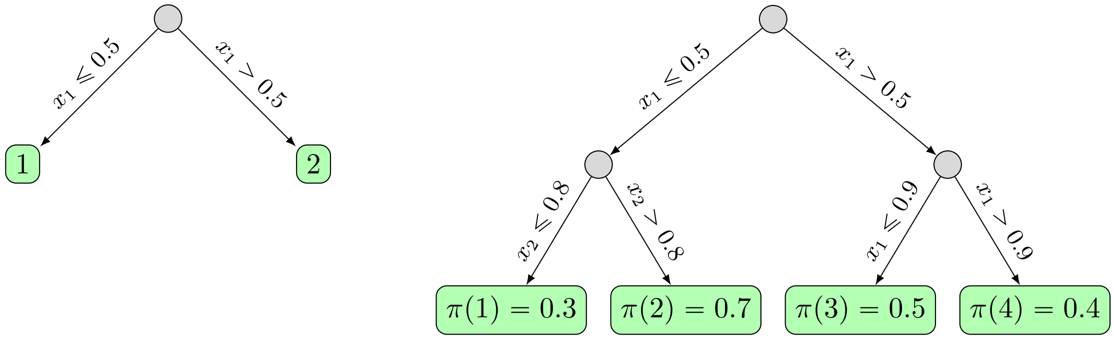

Figure 4: On the left: a tree S' whose nodes represent the subgroups of interest. On the right: an extension $T \in \mathcal{T}_2(S')$. Here $\mathcal{K}_1(T) = \{1, 2\}, \mathcal{K}_2(T) = \{3, 4\}$

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 straightforward to show using the recursive structure of stratification trees that choosing T as a solution to the following problem:

$$\min_{T \in \mathcal{T}_L(S')} V(T) \ ,$$

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.

In Section 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 $\hat{T}$ 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.

4 Simulations

In this section we analyze the finite sample behaviour of our method via a simulation study, and in particular analyze the performance of the cross-validation procedure presented in Section 3.2. We consider three DGPs in the spirit of the designs considered in Athey and Imbens (2016). We emphasize that although these designs are artificial, they highlight several interesting qualitative patterns. In Section 5, we repeat this exercise using an application-based design. For all three designs in this section, 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: $d = 2$, $\kappa_0(x) = 0.2$, $\nu_0(x) = 5$,

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

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

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

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

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

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

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

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

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 E for details). In each simulation we test five different methods of stratification. 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 E 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. 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 E. The estimator we use throughout is the sample-size weighted estimator described in Section 3.2.

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 $3,000$ Monte Carlo iterations. Table 1 presents the simulation results for Model 1.

In Table 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

Sample Size		Stratification Method	Criteria			
Pilot	Main		Coverage	%ΔLength	Power	%ΔRMSE
100	4900	No Stratification	94.5	0.0	77.2	0.0
		Ad-Hoc	94.9	-7.0	82.9	-10.4
		Strat. Tree	94.6	-0.1	78.2	-1.4
		CV Tree	95.1	-5.1	81.9	-7.7
		Optimal Tree	94.2	-18.6	91.5	-19.5
500	4500	No Stratification	94.1	0.0	77.7	0.0
		Ad-Hoc	94.0	-7.0	82.5	-6.3
		Strat. Tree	93.5	-13.5	88.0	-12.9
		CV Tree	93.9	-13.0	87.5	-13.5
		Optimal Tree	94.8	-17.0	90.8	-18.3
1500	3500	No Stratification	95.0	0.0	76.3	0.0
		Ad-Hoc	94.5	-7.0	82.6	-7.9
		Strat. Tree	93.7	-12.0	86.8	-11.9
		CV Tree	94.4	-11.6	86.7	-11.9
		Optimal Tree	94.3	-12.9	87.9	-12.1

Table 1: Simulation Results for Model 1

medium-sized pilot study (sample size 500), we see that both the stratification tree and the CV tree outperform ad-hoc stratification. Finally, when using a large pilot study (sample size 1500), we see that all three trees (strat, CV, and optimal) perform similarly to each other, and that there is a drop in performance relative to the medium-sized pilot. This is the behaviour we should have expected given the asymptotic results presented in Section 3. Next we study the results for Model 2, presented in Table 2:

In Table 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. For the large pilot, the qualitative results are similar to what we saw in Table 1. Finally, we study the results of Model 3, presented in Table 3.

In Table 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 as well as to the optimal tree.

Overall, we conclude that our proposed cross-validation procedure does a good job of protecting against overfitting. However, we would caution against using our method with small pilots.

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 behavioral 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

Sample Size		Stratification Method	Criteria			
Pilot	Main		Coverage	%ΔLength	Power	%ΔRMSE
100	4900	No Stratification	95.0	0.0	46.3	0.0
		Ad-Hoc	94.3	-1.9	48.8	-1.0
		Strat. Tree	94.5	7.0	41.6	9.5
		CV Tree	94.4	-7.8	53.8	-7.2
		Optimal Tree	93.8	-19.2	63.9	-17.5
500	4500	No Stratification	94.7	0.0	47.1	0.0
		Ad-Hoc	93.9	-1.8	48.0	-1.9
		Strat. Tree	93.6	-12.8	57.6	-10.2
		CV Tree	94.3	-14.0	58.9	-14.3
		Optimal Tree	94.1	-17.5	63.3	-15.8
1500	3500	No Stratification	94.1	0.0	47.8	0.0
		Ad-Hoc	93.8	-1.8	49.5	-0.5
		Strat. Tree	94.0	-12.4	59.1	-13.0
		CV Tree	94.2	-12.1	58.8	-12.7
		Optimal Tree	94.2	-13.3	59.1	-14.1

Table 2: Simulation Results for Model 2

Sample Size		Stratification Method	Criteria			
Pilot	Main		Coverage	%ΔLength	Power	%ΔRMSE
100	4900	No Stratification	95.0	0.0	30.5	0.0
		Ad-Hoc	94.9	-2.2	32.0	-1.6
		Strat. Tree	95.3	16.3	24.7	15.8
		CV Tree	94.6	1.0	30.9	3.1
		Optimal Tree	94.6	-7.3	35.4	-5.8
500	4500	No Stratification	94.8	0.0	31.0	0.0
		Ad-Hoc	94.9	-2.2	32.7	-2.3
		Strat. Tree	95.0	-2.1	31.8	-2.3
		CV Tree	94.9	-1.8	31.2	-2.7
		Optimal Tree	94.4	-6.7	33.9	-5.3
1500	3500	No Stratification	95.1	0.0	29.4	0.0
		Ad-Hoc	94.9	-2.2	31.8	-0.5
		Strat. Tree	95.4	-4.1	31.9	-2.9
		CV Tree	95.5	-3.6	32.2	-3.8
		Optimal Tree	95.9	-5.2	33.4	-5.1

Table 3: Simulation Results for Model 3

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 5 depicts the stratification used in Karlan and Wood (2017)[3].

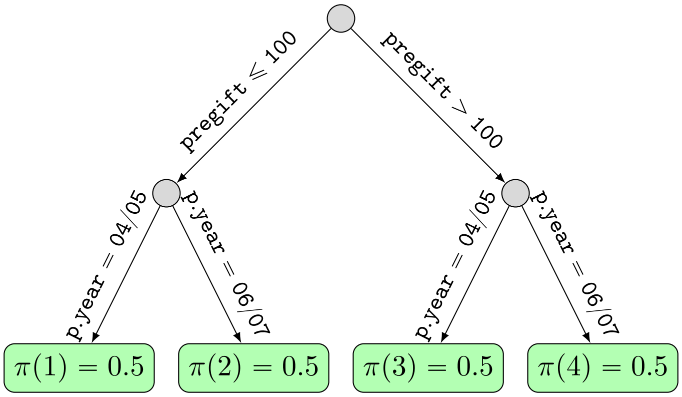

Figure 5: Stratification used in Karlan and Wood (2017)

We estimate two different stratification trees using data from the first wave of the experiment (with a sample size of 10, 869), 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 Section 3. 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 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 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 (we set the upper bound as roughly the 97th percentile in the dataset, although in practice this should be set using historical data).

Figure 6 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.

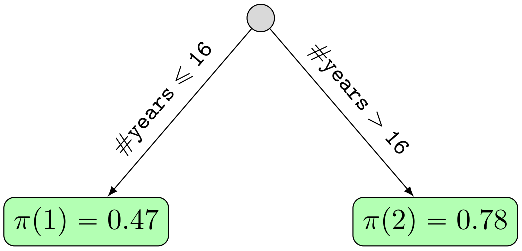

Figure 6: Unrestricted Stratification Tree estimated from Karlan and Wood (2017) data

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 7 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.

[3] Although Karlan and Wood (2017) claim to use a different stratification in the second-wave experiment, their exact implementation is not clear from the available data. 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.

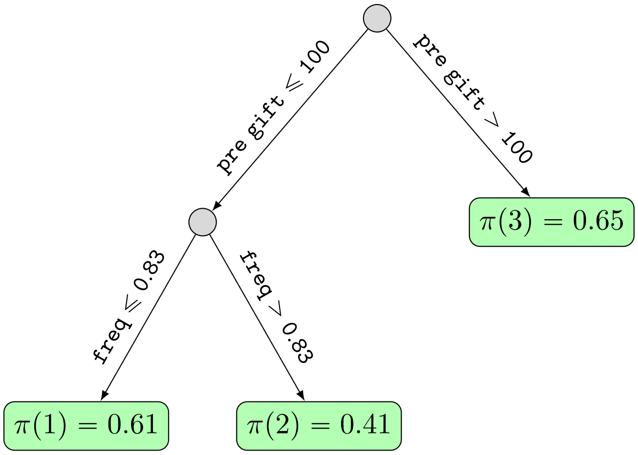

Figure 7: Restricted Stratification Tree estimated from Karlan and Wood (2017) data

These results are not necessarily surprising given the nature of the experiment: with very high probability, a recipient of either mailer is likely to make no donation at all, and hence we might expect limited heterogeneity in the potential outcomes with respect to our observable characteristics. This suggests 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 large sample, this may suggest that there is relatively little heterogeneity in the outcome with respect to the observable characteristics.

To further assess the potential gains from stratification in this application, we repeat the simulation exercise of Section 4 with an application-based simulation design. 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 the (scaled) covariates. We perform the simulations with a sample size of $30,000$, which corresponds approximately to the total number of observations in the dataset. 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 5, and use SBR to perform assignment. We compare the following stratification methods using the same criteria as in Section 4:

- No Stratification: Here we assign treatment to half the sample, with no stratification.
- Fixed Stratification: Here we use the stratification from Figure 5, 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 Appendix E). 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.

We perform 6000 Monte Carlo iterations. Table 4 presents the simulation results. We see in Table 4 that the overall gains from our procedure are small, which as we explained above may not be surprising given the nature of the experiment. 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 Figure 6. As was the case in the simulations of Section 4, the cross-validated stratification tree protects against overfitting, and seems to perform fairly well relative to the other feasible methods presented. To put these (modest) gains in perspective, the fixed stratification design would require 500 additional observations to match the performance of our cross-validated tree, and the no-stratification design would require 1000 additional observations.

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.

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.

A Proofs of Main Results

The proof of Theorem 3.1 requires some preliminary machinery which we develop in Appendix B. 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 B.1).
- We endow $\mathcal{T}$ with a metric $\rho(\cdot,\cdot)$ that makes $(\mathcal{T},\rho)$ a compact metric space (Definition B.2, Lemma B.2).
- We prove that $V(\cdot)$ is continuous in ρ (Lemma B.1).
- Let $\mathcal{T}^*$ be the set of minimizers of $V(\cdot)$, then this set is compact (in the topology induced by ρ), and it is the case given our assumptions that

  $$\inf_{T^* \in \mathcal{T}^*} \rho(\hat{T}_m, T^*) \xrightarrow{a.s.} 0 \ ,$$

  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

  $$\rho(\hat{T}_m, \bar{T}_m) \xrightarrow{a.s.} 0 \ .$$

  (Lemma B.4)

Remark A.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 3.1

Proof. Let $E_1[\cdot]$ and $E_2[\cdot]$ denote the expectations with respect to the first wave and second wave data, respectively. By Lemmas B.4 and A.1, we obtain immediately that

$$E_2[\mathbf{1}\{\sqrt{n}(\hat{\theta}(\hat{T}_m) - \theta) \leqslant t\}] \xrightarrow{a.s} \Phi^*(t) \ ,$$

where $\Phi^*(t)$ is the CDF of a $N(0, V^*)$ random variable. By the dominated convergence theorem, we get that

$$E_1[E_2[\mathbf{1}\{\sqrt{n}(\hat{\theta}(\hat{T}_m) - \theta) \leqslant t\}]] \to \Phi^*(t) \ .$$

Finally, by Fubini's theorem,

$$P(\sqrt{n}(\hat{\theta}(\hat{T}_m) - \theta) \leqslant t) = E[\mathbf{1}\{\sqrt{n}(\hat{\theta}(\hat{T}_m) - \theta) \leqslant t\}] = E_1[E_2[\mathbf{1}\{\sqrt{n}(\hat{\theta}(\hat{T}_m) - \theta) \leqslant t\}]] \to \Phi^*(t) \ ,$$

as desired. ■

Lemma A.1. *Let $\{T_m^{(1)}\}_m$ be a sequence of trees such that there exists a sequence $\{T_m^{(2)}\}_m$ where $\rho(T_m^{(1)}, T_m^{(2)}) \to 0$, and $T_m^{(2)} \in \mathcal{T}^*$ for all m. Given the Assumptions required for Theorem 3.1,*

$$\sqrt{n}(\hat{\theta}(T_m^{(1)}) - \theta) \xrightarrow{d} N(0, V^*) \ .$$

Proof. By the derivation in the proof of Theorem 3.1 in Bugni et al. (2018), we have that

$$\sqrt{n}(\hat{\theta}(T_m^{(1)}) - \theta) = \sum_{k=1}^{K} \left[\Omega_1(k; T_m^{(1)}) - \Omega_0(k; T_m^{(1)})\right] + \sum_{k=1}^{K} \Theta_k(k; T_m^{(1)}) \ ,$$

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

$$\Theta_k(T) := \sqrt{n}\left(\frac{n(k;T)}{n} - p(k;T)\right)\left[E(Y(1)|S(X) = k) - E(Y(0)|S(X) = k)\right]^2 \ .$$

To prove our result, we study the process

$$\mathbb{O}(T) = [\Omega_0(1;T) \ \ \Omega_1(1;T) \ \ \Omega_0(2;T) \ \ldots \ \Omega_1(K;T) \ \ \Theta(1;T) \ \ldots \ \Theta(K;T)]' \ .$$

By Lemma A.2, we have that

$$\mathbb{O}(T_m^{(1)}) \overset{d}{=} \bar{\mathbb{O}}(T_m^{(2)}) + o_P(1) \ ,$$

where $\bar{\mathbb{O}}(\cdot)$ is defined in Lemma A.2. Hence

$$\sqrt{n}(\hat{\theta}(T_m^{(1)}) - \theta) \overset{d}{=} O_n(T_m^{(2)}) + o_P(1) \ ,$$

where

$$O_n(T_m^{(2)}) = B'\bar{\mathbb{O}}(T_m^{(2)}) \ ,$$

and B is the appropriate vector of ones and negative ones to collapse $\bar{\mathbb{O}}(T)$:

$$B' = [-1, 1, -1, 1, \ldots, 1, 1, 1, \ldots, 1] \ .$$

It remains to show that $O_n(T_m^{(2)}) \xrightarrow{d} N(0, V^*)$, and then the result will follow. To that end, fix a strictly increasing indexing $(n_1, m_1) < ... < (n_\ell, m_\ell) < ...$ (where the inequality is to be interpreted componentwise). By the compactness of $\mathcal{T}^*$, $\{T_{m_\ell}^{(2)}\}$ contains a convergent subsequence (which by an abuse of notation we continue to index by m_ℓ, with corresponding index n_ℓ), so that:

$$T_{m_\ell}^{(2)} \to T^* \ ,$$

for some $T^* \in \mathcal{T}^*$. By the asymptotic equicontinuity of $\bar{\mathbb{O}}(\cdot)$ established in Lemma A.2, we have that

$$\bar{\mathbb{O}}(T_{m_\ell}^{(2)}) = \bar{\mathbb{O}}(T^*) + o_P(1) \ ,$$

and by the partial sum arguments in Lemma C.1. of Bugni et al. (2018), it follows that

$$O_{n_\ell}(T^*) \xrightarrow{d} N(0, V^*) \ ,$$

since T^* is an optimal tree. Hence we have that

$$O_{n_\ell}(T^{(2)}_{m_\ell}) \xrightarrow{d} N(0, V^*) \ .$$

By Lemma C.1 (applied to the CDFs), we get that

$$O_n(T^{(2)}_m) \xrightarrow{d} N(0, V^*) \ ,$$

as $m, n \to \infty$, and so the result follows. ■

Lemma A.2. *Let $\{T^{(1)}_m\}_m$ be a sequence of trees such that there exists a sequence $\{T^{(2)}_m\}_m$ where $\rho(T^{(1)}_m, T^{(2)}_m) \to 0$, and $T^{(2)}_m \in \mathcal{T}^*$ for all m. Given the Assumptions required for Theorem 3.1,*

$$\mathbb{O}(T^{(1)}_m) \overset{d}{=} \bar{\mathbb{O}}(T^{(2)}_m) + o_P(1) \ ,$$

as $n \to \infty$, where $\mathbb{O}(\cdot)$ is defined in the proof of Lemma A.1 and $\bar{\mathbb{O}}(\cdot)$ is defined in the proof of this result.

Proof. By the argument in Lemma C1 in Bugni et al. (2018), we have that

$$\mathbb{O}(T) \overset{d}{=} \widetilde{\mathbb{O}}(T) \ ,$$

where

$$\widetilde{\mathbb{O}}(T) = \Big[\widetilde{\Omega}_0(1;T) \ \ \widetilde{\Omega}_1(1;T) \ \ \widetilde{\Omega}_0(2;T) \ \dots \ \widetilde{\Omega}_1(K;T) \ \ \Theta(1;T) \ \dots \ \Theta(K;T)\Big]' \ .$$

with

$$\widetilde{\Omega}_a(k;T) = \frac{n(k;T)}{n_a(k;T)}\left[\frac{1}{\sqrt{n}}\sum_{i=n(\hat{F}(k;T)+\hat{F}_a(k;T))+1}^{n(\hat{F}(k;T)+\hat{F}_{a+1}(k;T))} G^k_a(U_{i,(a)}(k);T)\right] \ ,$$

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), $G^k_a(\cdot\,;T)$ 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 $\hat{F}_a(k;T) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{S_i = k, A_i < a\}$.

Let us focus on the term in brackets. Fix some a and k for the time being, and let

$$\mathcal{G} := \{G^k_a(\cdot\,;T) : T \in \mathcal{T}\}$$

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.5. 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.3, and A.6 that:

$$|\hat{F}_a(k;T^{(1)}_m) - F_a(k;T^{(2)}_m)| \xrightarrow{p} 0 \ ,$$

$$|\hat{F}(k;T_m^{(1)}) - F(k;T_m^{(2)})| \xrightarrow{p} 0 \ ,$$

$$||G_a^k(\cdot\ ;T_m^{(1)}) - G_a^k(\cdot\ ;T_m^{(2)})|| \to 0 \ ,$$

as $m \to \infty$. Hence we have by asymptotic equicontinuity that

$$\eta_n\left(\hat{F}(k;T_m^{(1)}) + \hat{F}_a(k;T_m^{(1)}), G_a^k(\cdot\ ;T_m^{(1)})\right) = \eta_n\left(F(k;T_m^{(2)}) + F_a(k;T_m^{(2)}), G_a^k(\cdot\ ;T_m^{(2)})\right) + o_P(1) \ .$$

By Lemma A.4,

$$\frac{n(k;T_m^{(1)})}{n_a(k;T_m^{(1)})} = \frac{1}{\pi(k;T_m^{(2)})} + o_P(1) \ .$$

Using the above two expressions, it can be shown that

$$\tilde{\Omega}_a(k;T_m^{(1)}) = \bar{\Omega}_a(k;T_m^{(2)}) + o_P(1) \ ,$$

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 $\Theta(k;T)$. By standard empirical process results for

$$\sqrt{n}\left(\frac{n(k;T)}{n} - p(k;T)\right) \ ,$$

it can be shown that

$$\Theta(k;T_m^{(1)}) = \Theta(k;T_m^{(2)}) + o_P(1) \ ,$$

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

$$\bar{\mathbb{O}}(T) = \left[\bar{\Omega}_0(1;T)\ \ \bar{\Omega}_1(1;T)\ \ \bar{\Omega}_0(2;T)\ \ \ldots\ \ \bar{\Omega}_1(K;T)\ \ \Theta(1;T)\ \ \ldots\ \ \Theta(K;T)\right]' \ .$$

then we have shown that

$$\mathbb{O}(T_m^{(1)}) \overset{d}{=} \bar{\mathbb{O}}(T_m^{(2)}) + o_P(1),$$

as desired. ■

Proof of Theorem 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 3.1 of this paper, it can be shown that

$$\hat{V}(\hat{T}) \overset{d}{=} V(\bar{T}) + o_P(1) \ .$$

By definition, $\bar{T} \in \mathcal{T}^*$ so that the result follows. ■

Proof of Proposition 2.1

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 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 B.4), then

$$\begin{aligned} V(\hat{T}) - V(T^*) &= V(\hat{T}) - \widetilde{V}_m(\hat{T}) + \widetilde{V}_m(\hat{T}) - V(T^*) \\ &\leqslant V(\hat{T}) - \widetilde{V}_m(\hat{T}) + \widetilde{V}_m(T^*) - V(T^*) \\ &\leqslant 2 \sup_{T \in \mathcal{T}} |\widetilde{V}_m(T) - V(T)| \; . \end{aligned}$$

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 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 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 3.2

Proof. It suffices to show that for any deterministic sequence of trees $\{T_m\}$, we have that

$$\chi_n(T_m) := \sup_t |H_n(t, T_m) - \Phi(t; T_m)| \to 0 \ .$$

Fix a strictly increasing indexing $(n_1, m_1) < ... < (n_\ell, m_\ell) < ...$ (where the inequality is to be interpreted componentwise). By the compactness of $\mathcal{T}_L$, we have that $\{T_{m_\ell}\}$ has a convergent subsequence (which by an abuse of notation we continue to index by m_ℓ and n_ℓ as in the proof of Lemma A.1) such that $T_{m_\ell} \to T'$ for some $T' \in \mathcal{T}_L$. By identical arguments to those used in the proofs of Lemmas A.1 and A.2 combined with Polya's theorem, it is the case that

$$\sup_t |H_{n_\ell}(t; T_{m_\ell}) - \Phi(t, T')| \to 0 \ .$$

By the continuity of $V(\cdot)$ we get that $V(T_{m_\ell}) \to V(T')$, and hence

$$|\Phi(t; T_{m_\ell}) - \Phi(t; T')| \to 0 \ ,$$

for every t. By the continuity of $\Phi(t; T')$ we get by an argument identical to the proof of Polya's theorem that

$$\sup_t |\Phi(t; T_{m_\ell}) - \Phi(t; T')| \to 0 \ .$$

It then follows by the triangle inequality that

$$\chi_{n_\ell}(T_{m_\ell}) \to 0 \ .$$

By Lemma C.1 it follows that $\chi_n(T_m)$ itself converges to zero, and hence we are done. ■

Proof of Proposition 3.3

Proof. For simplicity of exposition suppose that $V_1^* > V_2^* > ... > V_{\bar{L}}^*$. It suffices to show that

$$\left|\tilde{V}^{(1)}(\hat{T}_L^{(2)}) - V_L^*\right| \xrightarrow{a.s} 0 \ ,$$

for each L, and similarly with 1 and 2 reversed. Then we he have that

$$\tilde{V}_L^{CV} \xrightarrow{a.s} V_L^* \ ,$$

and hence

$$\hat{L} \overset{a.s}{=} \bar{L} \ ,$$

for m sufficiently large. To that end, by the triangle inequality

$$\left|\tilde{V}^{(1)}(\hat{T}_L^{(2)}) - V_L^*\right| \leqslant \left|\tilde{V}^{(1)}(\hat{T}_L^{(2)}) - \tilde{V}^{(2)}(\hat{T}_L^{(2)})\right| + \left|\tilde{V}^{(2)}(\hat{T}_L^{(2)}) - V_L^*\right| \ .$$

Consider the second term on the RHS, applying the triangle inequality again,

$$\left|\tilde{V}^{(2)}(\hat{T}_L^{(2)}) - V_L^*\right| \leqslant \left|\tilde{V}^{(2)}(\hat{T}_L^{(2)}) - V(\hat{T}_L^{(2)})\right| + \left|V(\hat{T}_L^{(2)}) - V_L^*\right| \ ,$$

and both of these terms converge to zero a.s. by the arguments made in the proof of Proposition 3.1. Next we consider the first term on the RHS, this is bounded above by

$$\sup_T \left|\tilde{V}^{(1)}(T) - \tilde{V}^{(2)}(T)\right| \ ,$$

and another application of the triangle inequality yields

$$\sup_T \left|\tilde{V}^{(1)}(T) - \tilde{V}^{(2)}(T)\right| \leqslant \sup_T \left|\tilde{V}^{(1)}(T) - V(T)\right| + \sup_T \left|\tilde{V}^{(2)}(T) - V(T)\right| \ ,$$

with both terms converging to 0 a.s. by the arguments made in the proof of Proposition 3.1. ■

Proof of Theorem 3.3

Proof. Let $t_1, t_2 \in \mathbb{R}$ be arbitrary, then we will to show that

$$P\left(\sqrt{m}(\hat{\theta}_1 - \theta) \leqslant t_1, \sqrt{n}(\hat{\theta}(\hat{T}) - \theta) \leqslant t_2\right) \to \Phi_1(t_1)\Phi^*(t_2) \ ,$$

where $\Phi_1(\cdot)$ is the CDF of a $N(0, V_1)$ random variable, and $\Phi^*(\cdot)$ is the CDF of a $N(0, V^*)$ random variable. The result will then follow by Assumption 3.4 and Slutsky's theorem. As in the proof of Theorem 3.1, let $E_1[\cdot]$ and $E_2[\cdot]$ denote the expectations with respect to the first and second wave data, respectively, then by Fubini's theorem,

$$P\left(\sqrt{m}(\hat{\theta}_1 - \theta) \leqslant t_1, \sqrt{n}(\hat{\theta}(\hat{T}) - \theta) \leqslant t_2\right) = E_1\left[E_2\left[\mathbf{1}\{\sqrt{m}(\hat{\theta}_1 - \theta) \leqslant t_1\}\mathbf{1}\{\sqrt{n}(\hat{\theta}(\hat{T}) - \theta) \leqslant t_2\}\right]\right] \ .$$

Adding and subtracting $P(\sqrt{m}(\hat{\theta}_1 - \theta) \leqslant t_1)\Phi^*(t_2)$ gives, after some additional algebra,

$$E_1\left[E_2\left[\mathbf{1}\{\sqrt{m}(\hat{\theta}_1 - \theta) \leqslant t_1\}\mathbf{1}\{\sqrt{n}(\hat{\theta}(\hat{T}) - \theta) \leqslant t_2\}\right]\right] = E_1\left[\left(E_2\left[\mathbf{1}\{\sqrt{n}(\hat{\theta}(\hat{T}) - \theta) \leqslant t_2\}\right] - \Phi^*(t_2)\right)\mathbf{1}\{\sqrt{m}(\hat{\theta}_1 - \theta) \leqslant t_1\}\right] + \\ + P(\sqrt{m}(\hat{\theta}_1 - \theta) \leqslant t_1)\Phi^*(t_2) \ .$$

By Assumption 3.3, we have that

$$P(\sqrt{m}(\hat{\theta}_1 - \theta) \leqslant t_1)\Phi^*(t_2) \to \Phi_1(t_1)\Phi^*(t_2) \ .$$

It remains to show that

$$E_1\left[\left(E_2\left[\mathbf{1}\{\sqrt{n}(\hat{\theta}(\hat{T}) - \theta) \leqslant t_2\}\right] - \Phi^*(t_2)\right)\mathbf{1}\{\sqrt{m}(\hat{\theta}_1 - \theta) \leqslant t_1\}\right] \to 0 \ .$$

By the triangle inequality,

$$\left|E_1\left[\left(E_2\left[\mathbf{1}\{\sqrt{n}(\hat{\theta}(\hat{T}) - \theta) \leqslant t_2\}\right] - \Phi^*(t_2)\right)\mathbf{1}\{\sqrt{m}(\hat{\theta}_1 - \theta) \leqslant t_1\}\right]\right| \leqslant E_1\left|E_2\left[\mathbf{1}\{\sqrt{n}(\hat{\theta}(\hat{T}) - \theta) \leqslant t_2\}\right] - \Phi^*(t_2)\right| \ .$$

By the argument used in the proof of Theorem 3.1,

$$\left|E_2\left[\mathbf{1}\{\sqrt{n}(\hat{\theta}(\hat{T}) - \theta) \leqslant t_2\}\right] - \Phi^*(t_2)\right| \xrightarrow{a.s} 0 \ .$$

Hence our result follows by applying Dominated Convergence. ■

Lemma A.3. *Let $\hat{F}$, $\hat{F}_a$, F and F_a be defined as in the proof of Lemma A.2. Let $T_m^{(1)}$, $T_m^{(2)}$ be defined as in the statement of Lemma A.2. Given the Assumptions of Theorem 3.1, we have that, for $k = 1, ..., K$,*

$$|\hat{F}_a(k; T_m^{(1)}) - F_a(k; T_m^{(2)})| \xrightarrow{p} 0 \ ,$$

and

$$|\hat{F}(k; T_m^{(1)}) - F(k; T_m^{(2)})| \xrightarrow{p} 0 \ .$$

Proof. We prove the first statement for $a = 1$, and the rest of the results follow similarly. We want to show that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{S_i(T_m^{(1)}) = k, A_i(T_m^{(1)}) = 0\} - (1 - \pi(k; T_m^{(2)}))p(k; T_m^{(2)})\right| \xrightarrow{p} 0 \ .$$

By the triangle inequality, we bound this above by

$$\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{S_i(T_m^{(1)})=k,A_i(T_m^{(1)})=0\}-(1-\pi(k;T_m^{(1)}))p(k;T_m^{(1)})\right|+$$

$$+\left|(1-\pi(k;T_m^{(1)}))p(k;T_m^{(1)})-(1-\pi(k;T_m^{(2)}))p(k;T_m^{(2)})\right| \ .$$

The first line of the above expression converges to zero by Assumption 2.5. Next consider the second line: by assumption, we have that $|p(k;T_m^{(1)})-p(k;T_m^{(2)})|\to 0$ and $|\pi(k;T_m^{(1)})-\pi(k;T_m^{(2)})|\to 0$ and hence the second line converges to zero. ■

Lemma A.4. *Let $T_m^{(1)}$, $T_m^{(2)}$ be defined as in the statement of Lemma A.2. Given the Assumptions of Theorem 3.1, we have that, for $k=1,...,K$,*

$$\frac{n(k;T_m^{(1)})}{n_a(k;T_m^{(1)})}=\frac{1}{\pi(k;T_m^{(2)})}+o_P(1) \ .$$

Proof. This follows from Assumption 2.5, the Glivenko-Cantelli Theorem, and the fact that $\pi(k;T_m^{(2)})p(k;T_m^{(2)})$ and $\frac{1}{p(k;T_m^{(2)})}$ are bounded. ■

Lemma A.5. *Given Assumption 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.6. *Let $T_m^{(1)}$, $T_m^{(2)}$ be defined as in the statement of Lemma A.2. Given the Assumptions of Theorem 3.1, we have that, for $k=1,...,K$,*

$$||G_a^k(\cdot\,;T_m^{(1)})-G_a^k(\cdot\,;T_m^{(2)})||\to 0 \ .$$

Proof. We show this for the case where $Y(a)$ is continuous. We proceed by showing convergence pointwise by invoking Lemma C.3, and then using the dominated convergence theorem. It thus remains to show that

$$|Z_a^k(t;T_m^{(1)})-Z_a^k(t;T_m^{(2)})|\to 0 \ ,$$

where $Z_a^k(\cdot\,;T)$ is the CDF of the distribution of $(Y(a)-E[Y(a)|S(X)])\,|S(X)=k$. Re-writing, we have that

$$Z_a^k(t;T)=\frac{E[\mathbf{1}\{Y(a)\leqslant t+E(Y(a)|S(X)=k)\}\mathbf{1}\{S(X)=k\}]}{P(S(X)=k)} \ ,$$

Hence by the triangle inequality, Assumption 2.2 and a little bit of algebra, we get that

$$|Z_a^k(t;T_m^{(1)})-Z_a^k(t;T_m^{(2)})|\leqslant\frac{1}{\delta}\left|R_{m1}-R_{m2}\right|+\frac{1}{\delta^2}\left|R_{m3}\right| \ ,$$

where

$$R_{mj}=E[\mathbf{1}\{Y(a)\leqslant t+E(Y(a)|S_m^{(j)}(X)=k)\}\mathbf{1}\{S_m^{(j)}(X)=k\}] \ \text{ for } j=1,2 \ ,$$

$$R_{m3} = P(S_m^{(1)}(X) = k) - P(S_m^{(2)}(X) = k) \ .$$

$|R_{m3}|$ goes to zero by assumption. It remains to show that $|R_{m1} - R_{m2}|$ converges to zero. Again by the triangle inequality,

$$|R_{m1} - R_{m2}| \leqslant |R_{m1} - R_{m4}| + |R_{m4} - R_{m2}| \ ,$$

where

$$R_{m4} = E[\mathbf{1}\{Y(a) \leqslant t + E(Y(a)|S_m^{(1)}(X) = k)\}\mathbf{1}\{S_m^{(2)}(X) = k\}] \ .$$

By another application of the triangle inequality,

$$|R_{m1} - R_{m4}| \leqslant E\left|\mathbf{1}\{S_m^{(1)}(X) = k\} - \mathbf{1}\{S_m^{(2)}(X) = k\}\right| \ ,$$

and this bound converges to zero by assumption. Finally,

$$|R_{m4} - R_{m2}| \leqslant E\left|\mathbf{1}\{Y(a) \leqslant t + E(Y(a)|S_m^{(1)}(X) = k)\} - \mathbf{1}\{Y(a) \leqslant t + E(Y(a)|S_m^{(2)}(X) = k)\}\right| \ .$$

By similar arguments to what we have used above, it can be shown that

$$|E(Y(a)|S_m^{(1)}(X) = k) - E(Y(a)|S_m^{(2)}(X) = k)| \to 0 \ ,$$

and hence it can be shown that $|R_{m4} - R_{m2}|$ also converges to zero. ■

Acknowledgments

I am grateful for advice and encouragement from Ivan Canay, Joel Horowitz, and Chuck Manski. I would also like to thank Eric Auerbach, Lori Beaman, Stephane Bonhomme, Ivan Fernandez-Val, Hidehiko Ichimura, Sasha Indarte, Seema Jayachandran, Vishal Kamat, Dean Karlan, Cynthia Kinnan, Dennis Kristensen, Ryan Lee, Eric Mbakop, Matt Masten, Denis Nekipelov, Sam Norris, Susan Ou, Azeem Shaikh, Mikkel Solvsten, Imran Rasul, Alex Torgovitsky, Chris Udry, Takuya Ura, Andreas Wachter, Joachim Winter, and seminar participants at many institutions for helpful comments and discussions. This research was supported in part through the computational resources and staff contributions provided for the Quest high performance computing facility at Northwestern University, and the Acropolis computing cluster at the University of Chicago.

Stratification Method	Criteria			
	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 4: Simulation Results for Application-Based Simulation

References

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

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.

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.*

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

Bai, Yuehao (2019), "Optimality of matched-pair designs in randomized controlled trials." *Available at SSRN 3483834.*

Bai, Yuehao, Azeem Shaikh, and Joseph P Romano (2019), "Inference in experiments with matched pairs." *University of Chicago, Becker Friedman Institute for Economics Working Paper.*

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.

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

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

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."

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

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.

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

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

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

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

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.

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.

Gyorfi, L Devroye L, Gabor Lugosi, and L Devroye (1996), "A probabilistic theory of pattern recognition."

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.

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

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

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 (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.

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.

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*.

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.

Narita, Yusuke (2018), "Toward an ethical experiment."

Pukelsheim, Friedrich (2006), *Optimal design of experiments*. SIAM.

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.

Silvey, Samuel (2013), *Optimal design: an introduction to the theory for parameter estimation*, volume 1. Springer Science & Business Media.

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."

Sverdlov, Oleksandr (2015), *Modern adaptive randomized clinical trials: statistical and practical aspects*, volume 81. CRC Press.

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.

Viviano, Davide (2020), "Experimental design under network interference." *arXiv preprint arXiv:2003.08421*.

Wei, Lee-Jen (1978), "The adaptive biased coin design for sequential experiments." *The Annals of Statistics*, 92–100.

Zelen, Marvin (1974), "The randomization and stratification of patients to clinical trials." *Journal of chronic diseases*, 27, 365–375.

B A Theory of Convergence for Stratification Trees (for online publication)

Remark B.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, 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 2.3 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 B.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 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 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 $\{\Gamma_D^{'(L-1)}, \Gamma_U^{'(L-1)}\}$ 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 2.1, 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 B.2. *For $S_1, S_2 \in \mathcal{S}_L$,*

$$\rho_1(S_1, S_2) := \sum_{k=1}^{2^L} P_X(S_1^{-1}(k) \Delta S_2^{-1}(k)) \ .$$

Where $A\Delta B := A\backslash B \cup B\backslash A$ denotes the symmetric difference of A and B. That ρ_1 is a metric follows from the properties of symmetric differences and Assumption 2.1. We show under appropriate assumptions that $(\mathcal{S}, \rho_1)$ is a complete metric space in Lemma B.2, and that $(\mathcal{S}, \rho_1)$ is totally bounded in Lemma B.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 B.1. *Given Assumption 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 B.2. *Given Assumptions 2.1 and 2.2, $(\mathcal{S}, \rho_1)$ is a complete metric space.*

Proof. Let $\{S_n\}_n$ be a Cauchy sequence in $\mathcal{S}_L$. It follows by the definition of ρ_1 that for each k, $\{S_n^{-1}(k)\}_n$ is a sequence of d-dimensional cubes which is itself Cauchy in the metric $P_X(\cdot\Delta\cdot)$. Fix a k and consider the resulting sequence of cubes $\Gamma_n = \times_{j=1}^d [a_{jn}, b_{jn}]$ for $n = 1, 2, ...$, we will show that this sequence converges to some cube $\Gamma = \times_{j=1}^d [a_j, b_j]$, where $a_j = \lim_n a_{jn}$, $b_j = \lim_n b_{jn}$. The resulting partition formed by all of these limit cubes will be our limit of $\{S_n\}_n$.

To that end, we will show that for a Cauchy sequence of cubes $\{\Gamma_n\}_n$, the corresponding sequences $\{a_{jn}\}$ and $\{b_{jn}\}$ are all Cauchy as sequences in $\mathbb{R}$. First note that if $\{\Gamma_n\}_n$ is Cauchy with respect to the metric induced by P_X, then it is Cauchy with respect to the metric induced by Lebesgue measure λ on $[0,1]^d$, since by Assumption 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 by Assumptions 2.1 and 2.2, it follows that if $\{\Gamma_n\}_n$ is Cauchy w.r.t to the metric induced by λ, then each sequence of intervals $\{[a_{jn}, b_{jn}]\}_n$ for $j = 1..., d$ is Cauchy w.r.t to the metric induced by Lebesgue measure on $[0,1]$ (which we denote by λ_1). By the properties of symmetric differences, when $[a_{jn}, b_{jn}] \cap [a_{jn'}, b_{jn'}] \neq \varnothing$ for $n \neq n'$,

$$\lambda_1([a_{jn}, b_{jn}]\Delta[a_{jn'}, b_{jn'}]) = |b_{jn'} - b_{jn}| + |a_{jn'} - a_{jn}|,$$

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 $[\lim a_n, \lim b_n]$, and hence that Γ_n converges to Γ, as desired. ■

Lemma B.3. *Given Assumption 2.1 $(\mathcal{S}_L, \rho_1)$ is a totally bounded metric space.*

Proof. Given any measurable set A, we have by Assumption 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 $\{\frac{\epsilon}{C(2^{2L})}, \frac{2\epsilon}{C(2^{2L})}, ..., 1\}$. By construction any tree in $\mathcal{S}_L$ is at most ϵ away from some tree in this set. ■

Lemma B.4. *Given Assumptions 2.1, 2.2, and 3.1. Then the set $\mathcal{T}^*$ of maximizers of $V(\cdot)$ exists, and*

$$\inf_{T^* \in \mathcal{T}^*} \rho(\widetilde{T}_m, T^*) \xrightarrow{a.s.} 0 \ ,$$

as $m \to \infty$, 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 $\bar{T}_m$ 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)). ■

C Auxiliary Lemmas (for online publication)

Lemma C.1. *Let $\{x_{n,m}\}$ be a doubly-indexed sequence of real numbers. If for any strictly increasing indexing $(n_1, m_1) < (n_2, m_2) < ... < (n_\ell, m_\ell) < ...$ (where the inequality is to be interpreted componentwise) the sequence $\{x_{n_\ell, m_\ell}\}$ contains a convergent subsequence which converges to x, then $x_{n,m} \to x$ as $n, m \to \infty$.*

Proof. Suppose not, then there exists some $\epsilon > 0$ such that for any $M \in \mathbb{N}$, we can find $n', m' > M$ such that $|x_{n',m'} - x| > \epsilon$. We use this fact to construct the following sequence: first pick $n_1, m_1 > 1$ such that $|x_{n_1,m_1} - x| > \epsilon$. Next pick $n_2, m_2 > \max(n_1, m_1)$ such that $|x_{n_2,m_2} - x| > \epsilon$. Continue to pick $n_{\ell+1}, m_{\ell+1} > \max(n_\ell, m_\ell)$ such that $|x_{n_{\ell+1},m_{\ell+1}} - x| > \epsilon$. The resulting sequence $\{x_{n_\ell,m_\ell}\}$ satisfies to conditions of the lemma but contains no subsequence converging to x. Hence the result follows by contradiction. ■

Lemma C.2. *Let $\{A_n\}_n$, $\{B_n\}_n$ be sequences of continuous random variables such that*

$$|A_n - B_n| \xrightarrow{p} 0 \ .$$

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 $|t' - t| < \delta$, $|G_n(t) - G_n(t')| < \epsilon/2$. Furthermore, choose N such that for $n \geqslant N$, $P(|A_n - B_n| > \delta) < \epsilon/2$. 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 C.3. *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 C.2 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 3.1. It shows that, as long as the family of quantile functions defined in Assumption 3.2 are continuous, and vary "continuously" in $S \in \mathcal{S}_L$, then Assumption 3.2 holds.

Lemma C.4. *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 $x_0 \in (0,1)$, $\{f_d(\cdot)\}_{d \in \mathbb{D}}$ 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. ∎

D Supplementary Results (for online publication)

D.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 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 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.

D.2 Extension to the Case of 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, outer$(v) := vv'$, and ι_M is a vector of ones of length M. We note that this variance matrix takes the form of the semi-parametric efficiency bound derived in Cattaneo (2010) for the discretization implied by $S(\cdot)$.

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 section 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.

Consider the following extensions of Assumptions 2.1, 2.2, 3.1, 2.4, and 2.5 to multiple treatments:

Assumption D.1. *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 D.2. *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 D.3. *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 D.4. *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 D.5. *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

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

Consider also the following potentially strong uniqueness assumption:

Assumption D.6. *The minimizer T^* of $V(T)$ over $\mathcal{T}_L$ is unique.*

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 $V(T)$. 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.

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 3.1 continue to hold in the multiple treatment setting as well. For example:

Theorem D.1. *Given Assumptions D.1, D.2, 2.2, 2.3, D.3, D.4, D.5, and D.6, 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 D.6, Assumption 3.2 is no longer required. The proof proceeds identically to the proof of Theorem 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 D.6.

E Computational Details and Supplementary Simulation Details (for online publication)

E.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

$$\hat{E}[Y(1) - Y(0)|S(X) = k] := \frac{1}{m_1(k;T)} \sum_{j=1}^{m} Y_j A_j \mathbf{1}\{S(X_j) = k\} - \frac{1}{m_0(k;T)} \sum_{j=1}^{m} Y_j (1 - A_j) \mathbf{1}\{S(X_j) = k\} \ ,$$

$$\hat{E}[Y(1) - Y(0)] := \frac{1}{m_1} \sum_{j=1}^{m} Y_j A_j - \frac{1}{m_0} \sum_{j=1}^{m} Y_j (1 - A_j) \ ,$$

$$\hat{\sigma}_a^2(k) := \hat{E}[Y(a)^2|S(X) = k] - \hat{E}[Y(a)|S(X) = k]^2 \ .$$

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 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 2.2).

E.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 8, 9 and 10 below.

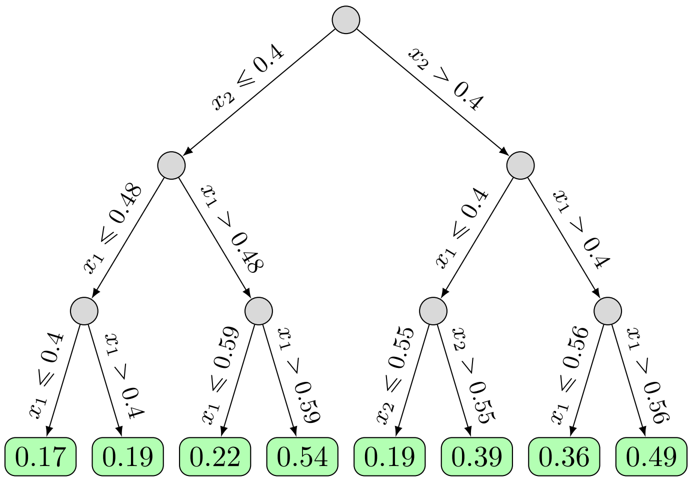

Figure 8: Optimal Infeasible Tree for Model 1

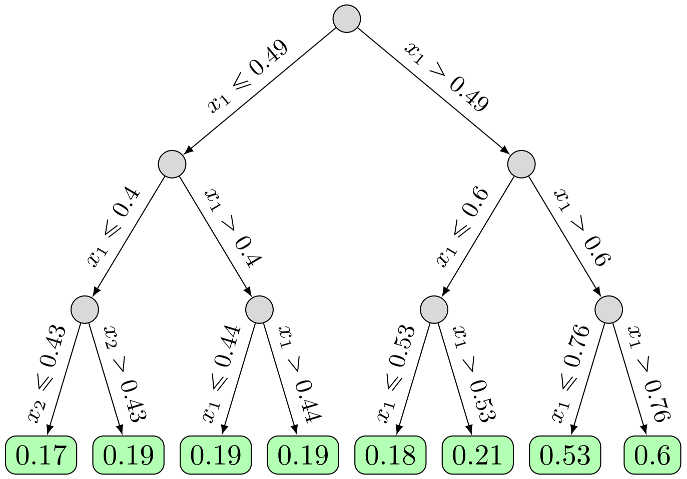

Figure 9: Optimal Infeasible Tree for Model 2

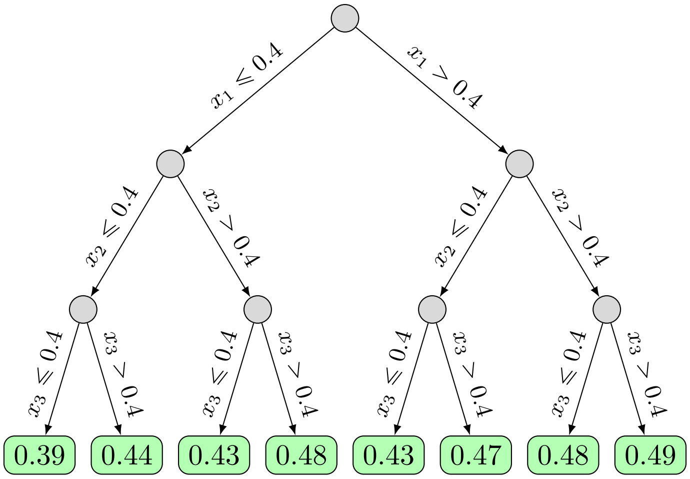

Figure 10: Optimal Infeasible Tree for Model 3

For the application-based design, the ATE is computed to be 0.61. The infeasible tree we computed is depicted in Figure 11.

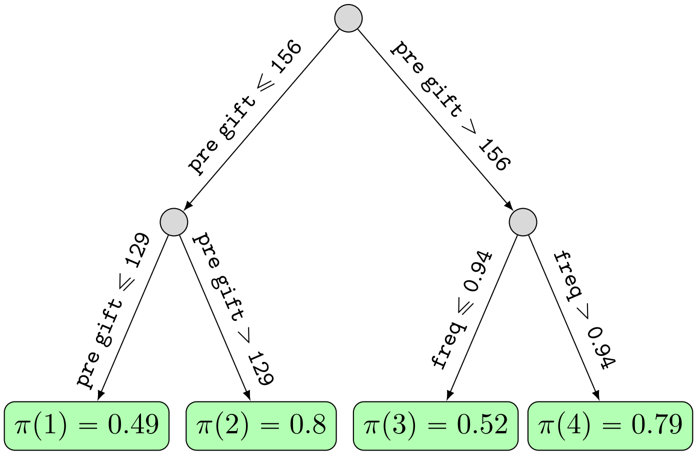

Figure 11: Infeasible Optimal Tree for App.-based Simulation