Statistical Inference for Heterogeneous Treatment Effects Discovered by Generic Machine Learning in Randomized Experiments*

Kosuke Imai† Michael Lingzhi Li‡

April 23, 2024

Abstract

Researchers are increasingly turning to machine learning (ML) algorithms to investigate causal heterogeneity in randomized experiments. Despite their promise, ML algorithms may fail to accurately ascertain heterogeneous treatment effects under practical settings with many covariates and small sample size. In addition, the quantification of estimation uncertainty remains a challenge. We develop a general approach to statistical inference for heterogeneous treatment effects discovered by a generic ML algorithm. We apply the Neyman's repeated sampling framework to a common setting, in which researchers use an ML algorithm to estimate the conditional average treatment effect and then divide the sample into several groups based on the magnitude of the estimated effects. We show how to estimate the average treatment effect within each of these groups, and construct a valid confidence interval. In addition, we develop nonparametric tests of treatment effect homogeneity across groups, and rank-consistency of within-group average treatment effects. The validity of our methodology does not rely on the properties of ML algorithms because it is solely based on the randomization of treatment assignment and random sampling of units. Finally, we generalize our methodology to the cross-fitting procedure by accounting for the additional uncertainty induced by the random splitting of data.

Key Words: causal inference, causal heterogeneity, conditional average treatment effect, cross-fitting, multiple testing, randomization inference, sample splitting

*We thank the Sloan foundation (Economics Program; 2020–13946) for partial support. The proposed methodology is implemented through an open-source R package, evalITR, which is freely available for download at the Comprehensive R Archive Network (CRAN; `https://CRAN.R-project.org/package=evalITR`). We also thank Kevin Du, Lucas Janson, and Yash Nair for their feedback on an earlier version of this paper.

†Professor, Department of Government and Department of Statistics, Harvard University, Cambridge, MA 02138. Phone: 617–384–6778, Email: Imai@Harvard.Edu, URL: https://imai.fas.harvard.edu

‡Operation Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139. mlli@mit.edu

1 Introduction

A growing number of researchers are turning to machine learning (ML) algorithms to uncover causal heterogeneity in randomized experiments. ML algorithms are appealing because in many applications the structure of heterogeneous treatment effects is unknown. Despite their promise, however, relatively little theoretical properties have been established for many of these algorithms. In addition, the choice of tuning parameter values remains to be often difficult and consequential in practice. As a result, ML algorithms may fail to ascertain heterogeneous treatment effects under common settings with many covariates and small sample size. Furthermore, one major challenge is the quantification of statistical uncertainty when estimating heterogeneous treatment effects using ML algorithms.

In this paper, we develop a general approach to statistical inference for heterogeneous treatment effects estimated through the application of a generic ML algorithm to experimental data. We apply the Neyman (1923)'s repeated sampling framework to a common setting, in which researchers use ML algorithms to estimate the conditional average treatment effect (CATE) given pre-treatment covariates and then divide the sample into several groups based on the magnitude of these estimated effects. We show how to obtain a consistent estimate of the average treatment effect within each of these groups — the sorted group average treatment effect (GATES; Chernozhukov *et al.* (2023)) — and construct an asymptotically valid confidence interval.

We also propose two nonparametric tests of treatment effect heterogeneity that are of interest to applied researchers. First, we test whether there exists any treatment effect heterogeneity across groups. Second, we develop a statistical test of the rank-consistency of GATES. If an ML algorithm produces a reasonable scoring rule, the rank ordering of GATES based on their magnitude should be monotonic. To accommodate the use of various ML algorithms, we make no assumption about their properties. Specifically, ML algorithms do not have to be consistent or unbiased. This is possible because the validity of our confidence intervals and nonparametric tests solely depends on the randomization of treatment assignment and random sampling of units. Thus, our approach imposes only a minimal set of assumptions on the underlying data generating process.

We first consider the settings, in which an external data set is used to estimate the CATE. For example, researchers may apply an ML algorithm to an observational dataset. Alternatively, an experimental dataset may be split into the training and validation data sets where an ML algorithm is first applied to the training data to estimate the CATE, and the validation data is then used to estimate the GATES. Here, we treat the estimated CATE function as fixed and do not account for the uncertainty that arises from its estimation.

To incorporate this additional source of uncertainty, we further generalize our methodology to the cross-fitting procedure, which randomly splits the data into multiple folds. Each fold is used as the validation data to estimate the GATES while the remaining folds serve as the corresponding training data to estimate the CATE. After repeating this for each fold, we aggregate the GATES estimates to the entire sample. Unlike the sample-splitting case where we condition on the split, we account for additional uncertainty induced by the randomness of its cross-fitting procedure. This directly addresses the fact that when the sample size is small the GATES estimate may vary considerably due to the random splitting of data.

Related Literature. The proposed methodology builds on the existing literature about statistical inference for heterogeneous treatment effects. In an early work, Crump *et al.* (2008) propose nonparametric tests of treatment effect heterogeneity. The authors rely on the consistency of sieve methods under the assumption that heterogeneous treatment effects are a smooth function of covariates. In contrast, our methodology does not require the consistent estimation of the CATE by ML algorithms. Moreover, while Crump *et al.* assume the continuous differentiability of the CATE, we only require its continuity.

Ding *et al.* (2016) propose an alternative approach based on Fisher's randomization test. Similar to our proposed methodology, this test neither requires modeling assumptions nor imposes restrictive assumptions on data generating process. In fact, their test yields conservative p-values without asymptotic approximation whereas other approaches including ours are only valid in large samples. The authors, however, test restrictive sharp null hypotheses. For example, Ding *et al.* (2016) consider a null hypothesis that the individual treatment effect is constant within each group and the effect only varies across groups. In contrast, we focus on the null hypotheses about average treatment effects that may vary within and across groups under the Neyman's repeated sampling framework. While our tests are valid only asymptotically, our simulation studies show that they perform reasonably well in small samples. In addition, Ding *et al.* (2019) use the Neyman's repeated sampling framework to explore treatment effect heterogeneity like we do, but rely entirely on the linear regression and does not allow for the use of more flexible ML algorithms.

More recently, Chernozhukov *et al.* (2023) study the settings that are identical to the ones considered in this paper. Similar to our methodology, the authors do not impose strong assumptions on the properties of ML algorithms that are used to estimate the CATE. However, to incorporate the additional uncertainty of the cross-fitting procedure, Chernozhukov *et al.* (2023) propose to repeat the procedure many times and aggregate the resulting p-values. We avoid such a computationally intensive procedure and instead use the Neyman's repeated sampling framework to conduct valid statistical inference under cross-fitting. In simulation studies reported elsewhere (Imai and Li, 2023a), we show that our confidence intervals are less conservative than those proposed by Chernozhukov *et al.* (2023) in finite samples.

Other researchers also have considered GATES and related quantities. For example, Yadlowsky *et al.* (2021) establish the asymptotic properties for a related general class of metrics that summarize the effect of treatment prioritization rules. In addition to the different focus, the authors assume that a treatment prioritization rule of interest is fixed and do not consider the uncertainty that arises from its estimation. Dwivedi *et al.* (2020) also estimate the GATES to explore treatment effect heterogeneity and develop calibration methods. However, they do not derive the asymptotic distribution of GATES and hence stop short of providing formal statistical methods.

Finally, Imai and Li (2023b) show how to evaluate an individualized treatment rule derived from the application of a generic ML algorithm in general settings including the one based on cross-fitting. We build on this work and derive the asymptotic properties of the GATES estimator. Imai and Li (2023c) further extends the methodology proposed in this paper and develop uniform asymptotic confidence bands. This allows researchers to choose, with a statistical guarantee, a group of individuals who are predicted to benefit from or be harmed by the treatment, using the estimated CATE based on a generic ML algorithm. They do not, however, consider the estimation uncertainty of the CATE.

2 The Proposed Methodology

We start by developing our methodology in a setting where the conditional average treatment effect (CATE) function is estimated using a separate data set, but is considered fixed when estimating the sorted group average treatment effect (GATES) and conducting statistical tests. For instance, the estimated CATE might come from an external, possibly observational, dataset. An alternative is *sample splitting*, where the sample is divided randomly into training and evaluation sets. The training data is used for CATE estimation via a machine learning algorithm, and the evaluation data for GATES estimation. In this section, we do not account for the uncertainty in estimating the CATE. In Section 3, we extend our methodology to cross-fitting, incorporating this estimation uncertainty.

2.1 Setup

Suppose that we have an independently and identically distributed (i.i.d.) sample of n units from a super-population $\mathcal{P}$. Let T_i represent the treatment assignment indicator variable, which is equal to 1 if unit i is assigned to the treatment condition and is equal to 0 otherwise, i.e., $T_i \in \mathcal{T} = \{0, 1\}$. For each unit, we observe the outcome variable $Y_i \in \mathcal{Y}$ and a vector of pre-treatment covariates, $\boldsymbol{X}_i \in \mathcal{X}$, where $\mathcal{Y}$ and $\mathcal{X}$ represent the support of the outcome variable and that of the pre-treatment covariates, respectively.

We require the standard causal inference assumptions of consistency and no interference between units, denoting the potential outcome for unit i under the treatment condition $T_i = t$ as $Y_i(t)$ for $t = 0, 1$ (e.g., Neyman, 1923; Holland, 1986; Rubin, 1990). The observed outcome is given by $Y_i = Y_i(T_i)$. For notational simplicity, we assume that the treatment assignment is completely randomized with exactly n_1 units assigned to the treatment condition though the extensions to other experimental designs and unconfounded observational designs are possible. We formally state these assumptions below.

Assumption 1 (No Interference between Units) *The potential outcomes for unit i do not depend on the treatment status of other units. That is, for all $t_1, t_2, \ldots, t_n \in \{0, 1\}$, we have, $Y_i(T_1 = t_1, T_2 = t_2, \ldots, T_n = t_n) = Y_i(T_i = t_i)$.*

Assumption 2 (Random Sampling of Units) *Each of n units, represented by a three-tuple consisting of two potential outcomes and pre-treatment covariates, is assumed to be independently sampled from a super-population $\mathcal{P}$, i.e.,*

$$(Y_i(1), Y_i(0), \boldsymbol{X}_i) \overset{\text{i.i.d.}}{\sim} \mathcal{P}$$

Assumption 3 (Complete Randomization) *For any $\boldsymbol{t} \in \{0, 1\}^n$ such that $\sum_{i=1}^n \boldsymbol{t}_i = n_1$, the treatment assignment probability is given by,*

$$\Pr(\boldsymbol{T} = \boldsymbol{t} \mid \{Y_{i'}(1), Y_{i'}(0), \boldsymbol{X}_{i'}\}_{i'=1}^n) = \frac{1}{\binom{n_1}{n}}.$$

Suppose that a researcher applies an ML algorithm to a training dataset and estimate the CATE. As noted earlier, this training dataset can be obtained through the sample splitting or it may be an external dataset. The CATE is defined as,

$$\tau(\boldsymbol{x}) \;=\; \mathbb{E}(Y_i(1) - Y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x}),$$

for any $\boldsymbol{x} \in \mathcal{X}$. The ML algorithm produces the following scoring rule,

$$s : \mathcal{X} \longrightarrow \mathcal{S} \subset \mathbb{R} \tag{1}$$

where a greater score indicates a higher priority to receive the treatment. Without loss of generality, we assume that the scoring rule is bijective, i.e., $s(\boldsymbol{x}) \neq s(\boldsymbol{x}')$ for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ with $\boldsymbol{x} \neq \boldsymbol{x}'$. Note that one can always redefine $\mathcal{X}$ to satisfy this assumption.

As noted earlier, we assume almost nothing about the properties of this scoring rule derived by the ML algorithm. In particular, the scoring rule does not have to be a consistent estimate of CATE. In fact, the scoring rule need not even be an estimate of CATE so long as it satisfies the definition given in Equation (1).

2.2 Estimation and Inference

Given the setup introduced above, we first consider the estimation and inference for the sorted group average treatment effect (GATES), which is a common quantity of interest in applied research and is studied by Chernozhukov *et al.* (2023). The idea is that researchers sort units into a total of K groups based on the quantile of the scoring rule, and then estimate the average treatment effect within each group. For simplicity, we assume that the number of treated and control units, i.e., n_1 and n_0, are multiples of K. The formal definition of GATES is given by,

$$\tau_k \;=\; \mathbb{E}(Y_i(1) - Y_i(0) \mid c_{k-1}(s) < s(\boldsymbol{X}_i) \leq c_k(s)) \tag{2}$$

for $k = 1, 2, \ldots, K$ where c_k represents the cutoff between the $(k-1)$th and kth groups and is defined as,

$$c_k(s) \;=\; \inf\{c \in \mathbb{R} \mid \Pr(s(\boldsymbol{X}_i) \leq c) \geq k/K\},$$

for $k = 1, 2, \ldots, K$, with $c_0 = -\infty$. Equivalently, GATES can be seen as a special case of the rank-weighted average treatment effect (RATE) with $\alpha(u) = \mathbf{1}\{\frac{k-1}{K} < u \leq \frac{k}{K}\}$ (Yadlowsky *et al.*, 2021).

Thus, units that belong to the Kth group, for example, represent those who are likely to have the greatest treatment effect according to the ML algorithm whereas those in the first group are likely to have the least treatment effect. However, we do not assume that the GATES is monotonic, i.e., $\tau_k \leq \tau_{k'}$ for all $k < k'$. This is important because we want to impose as little restriction on the underlying scoring rule as possible. Indeed, if the scoring rule is not a good estimate of CATE, such an assumption may be violated. To address this problem, we later develop a statistical test of this monotonicity assumption.

We consider the following estimator of GATES using the experimental data,

$$\hat{\tau}_k \;=\; \frac{K}{n_1}\sum_{i=1}^{n} Y_i T_i \hat{f}_k(\boldsymbol{X}_i) - \frac{K}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)\hat{f}_k(\boldsymbol{X}_i), \tag{3}$$

for $k = 1, 2, \ldots, K$ where $\hat{f}_k(\boldsymbol{X}_i) = \mathbf{1}\{s(\boldsymbol{X}_i) > \hat{c}_{k-1}(s)\} - \mathbf{1}\{s(\boldsymbol{X}_i) > \hat{c}_k(s)\}$, and $\hat{c}_k(s) \;=\; \inf\{c \in \mathbb{R} : \sum_{i=1}^{n} \mathbf{1}\{s(\boldsymbol{X}_i) \leq c\} \geq nk/K\}$ is the estimated cutoff. First, we derive the bias bound and exact variance of the GATES estimator.

Theorem 1 (Bias Bound and Exact Variance of the GATES Estimator) *Under Assumptions 1–3, the bias of the proposed estimator of GATES given in Equation* (3) *can be bounded as follows,*

$$\begin{aligned}
&\mathbb{P}(|\mathbb{E}\{\hat{\tau}_k - \tau_k \mid \hat{c}_k(s), \hat{c}_{k-1}(s)\}| \geq \epsilon) \\
\leq\;& 1 - B\left(\frac{k}{K} + \gamma_k(\epsilon), \frac{nk}{K}, n - \frac{nk}{K} + 1\right) + B\left(\frac{k}{K} - \gamma_k(\epsilon), \frac{nk}{K}, n - \frac{nk}{K} + 1\right) \\
& - B\left(\frac{k-1}{K} + \gamma_{k-1}(\epsilon), \frac{n(k-1)}{K}, n - \frac{n(k-1)}{K} + 1\right)
\end{aligned}$$

$$+B\left(\frac{k-1}{K}-\gamma_{k-1}(\epsilon), \frac{n(k-1)}{K}, n-\frac{n(k-1)}{K}+1\right),$$

for any given constant $\epsilon > 0$ *where* $B(\epsilon, \alpha, \beta)$ *is the incomplete beta function (if* $\alpha \leq 0$ *and* $\beta > 0$*, we set* $B(\epsilon, \alpha, \beta) := H(\epsilon)$ *for all* ϵ *where* $H(\epsilon)$ *is the Heaviside step function), and*

$$\gamma_k(\epsilon) \;=\; \frac{\epsilon}{K \max_{c \in [c_k(s)-\epsilon,\; c_k(s)+\epsilon]} \mathbb{E}(Y_i(1) - Y_i(0) \mid s(\boldsymbol{X}_i) = c)}.$$

The variance of the estimator is given by,

$$\mathbb{V}(\hat{\tau}_k) \;=\; K^2\left(\frac{\mathbb{E}(S_{k1}^2)}{n_1}+\frac{\mathbb{E}(S_{k0}^2)}{n_0}\right)+\frac{(n-K)\kappa_{k11}}{n-1}-\kappa_{k1}^2,$$

where $S_{kt}^2 = \sum_{i=1}^n (Y_{ki}(t) - \overline{Y_k(t)})^2/(n-1)$*,* $\kappa_{kt} = \mathbb{E}(Y_i(1) - Y_i(0) \mid \hat{f}_k(\boldsymbol{X}_i) = t)$*, and* $\kappa_{ktt} = \mathbb{E}[(Y_i(1) - Y_i(0))(Y_j(1) - Y_j(0)) \mid \hat{f}_k(\boldsymbol{X}_i) = \hat{f}_k(\boldsymbol{X}_j) = t]$ *for* $i \neq j$ *with* $Y_{ki}(t) = \hat{f}_k(\boldsymbol{X}_i)Y_i(t)$*, and* $\overline{Y_k(t)} = \sum_{i=1}^n Y_{ki}(t)/n$*, for* $t = 0, 1$*.*

Proof is given in Supplementary Appendix S1.

When compared to the standard variance estimator, there are additional two terms. These terms result from the fact that the cutoff points are estimated, yielding a cross-unit correlation in terms of $\hat{f}_k(\boldsymbol{X}_i)Y_i(t)$. Since exactly n/K data points are taken to have $\hat{f}_k(\boldsymbol{X}_i) = 1$, the value of this function is generally negatively correlated across units, i.e., $\mathrm{Corr}(\hat{f}_k(\boldsymbol{X}_i), \hat{f}_k(\boldsymbol{X}_j)) < 0$.

The variance can be consistently estimated by replacing each unknown parameter with its sample analogue:

$$\begin{aligned}
\widehat{\mathbb{E}(S_{kt}^2)} &= \frac{1}{n_t-1}\sum_{i=1}^n \mathbf{1}\{T_i = t\}(Y_{ki} - \overline{Y_{kt}})^2,\\
\hat{\kappa}_{kt} &= \frac{\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}T_iY_i}{\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}T_i} - \frac{\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}(1-T_i)Y_i}{\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}(1-T_i)},\\
\hat{\kappa}_{ktt} &= \frac{[\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}T_iY_i]^2 - \sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}T_iY_i^2}{[\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}T_i]^2 - \sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}T_i}\\
&\quad + \frac{[\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}(1-T_i)Y_i]^2 - \sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}(1-T_i)Y_i^2}{[\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}(1-T_i)]^2 - \sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}(1-T_i)}\\
&\quad - 2\frac{[\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}(1-T_i)Y_i][\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}T_iY_i]}{[\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}(1-T_i)][\sum_{i=1}^n \mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i) = t\}T_i]}.
\end{aligned}$$

for $t = 0, 1$ where $Y_{ki} = \hat{f}_k(\boldsymbol{X}_i)Y_i$ and $\overline{Y}_{kt} = \sum_{i=1}^n \mathbf{1}\{T_i = t\}Y_{ki}/n_t$. The expression of $\hat{\kappa}_{ktt}$ above enables the calculation in $O(n)$ rather than $O(n^2)$ time. The details of the derivation is given in Appendix S2.

We can further derive the asymptotic sampling distribution of the GATE estimator by requiring the following continuity assumption and moment conditions:

Assumption 4 (Continuity of CATE at the Thresholds) *Let $F(c) = \Pr(s(\boldsymbol{X}_i) \leq c)$ represent the cumulative distribution function of the scoring rule and define its pseudo-inverse $F^{-1}(p) = \inf\{c : F(c) \geq p\}$ for $p \in [0, 1]$. The CATE function $\mathbb{E}(Y_i(1) - Y_i(0) \mid s(\boldsymbol{X}_i) = F^{-1}(p))$ is assumed to be left-continuous with bounded variation on any interval $(\theta, 1 - \theta)$ with $\theta > 0$, and continuous in p at $p = 1/K, \ldots, (K - 1)/K$.*

Assumption 5 (Moment Conditions) *For each $t = 0, 1$, we have*

1. $\mathbb{V}(Y_i(t)) > 0$;
2. $\mathbb{E}(Y_i(t)^3) < \infty$.

Assumption 4 is similar to the assumption commonly used in the literature that the CATE is continuous in the covariates $\boldsymbol{X}_i$ (e.g., Künzel *et al.*, 2018; Wager and Athey, 2018), but we only require continuity at the thresholds, $1/K, \cdots, (K - 1)/K$ and bounded variation everywhere else. We will show in Proposition 1 below that Assumption 4 is among the weakest assumptions necessary for our asymptotic results. In particular, this assumption requires that the scoring rule cannot be discontinuous at the thresholds unless the CATE is constant in the scoring rule, i.e. $\mathbb{E}(Y_i(1) - Y_i(0) \mid s(\boldsymbol{X}_i) = F^{-1}(p)) = \mathbb{E}(Y_i(1) - Y_i(0))$ for all p.

We now present the asymptotic sampling distribution of GATES estimator.

Theorem 2 (Asymptotic Sampling Distribution of GATES Estimator) *Under Assumptions 1–5, we have,*

$$\frac{\hat{\tau}_k - \tau_k}{\sqrt{\mathbb{V}(\hat{\tau}_k)}} \xrightarrow{d} N(0, 1)$$

for $k = 1, \ldots, K$ where $\mathbb{V}(\hat{\tau}_k)$ is given in Theorem 1.

Proof is given in Supplementary Appendix S3. We emphasize that Theorem 2 does not impose a strong assumption about the properties of the ML algorithm used to generate the scoring rule s.

In fact, the continuity of the CATE at the thresholds (Assumption 4) is among the weakest assumptions that can ensure the validity of Theorem 2. To see this, consider an alternative assumption that there exists a threshold at which CATE is bounded but discontinuous, slightly relaxing Assumption 4. The following proposition shows that this assumption is not sufficient for Theorem 2.

Proposition 1 (Insufficiency of Bounded Variation) *Suppose Assumptions 1–3 and 5 hold. Further assume that the there exists a threshold k/K, such that $\mathbb{E}(Y_i(1) - Y_i(0) \mid s(\boldsymbol{X}_i) = F^{-1}(p))$, is discontinuous (but bounded) at $p = k/K$. Then, there exist a scoring rule s and a population $\mathcal{P}$ such that as $n \to \infty$ with $0 < n_1/n < 1$ staying constant, we have,*

$$\mathbb{E}\left(\frac{\hat{\tau}_k - \tau_k}{\sqrt{\mathbb{V}(\hat{\tau}_k)}}\right) \not\to 0.$$

Proof is given in Supplementary Appendix S4. Proposition 1 shows that if the CATE is mildly discontinuous at a threshold, then we cannot sufficiently control the bias in estimating the boundary points, $c_k(s)$. Under this scenario, the bias decays at the rate of $n^{-1/2}$, which is not fast enough for the application of the central limit theorem.

2.3 Nonparametric Test of Treatment Effect Heterogeneity

In many applications, heterogeneous treatment effects are imprecisely estimated. Researchers may wish to know whether the treatment effect heterogeneity discovered by ML algorithms represents signal rather than noise. In addition, checking the statistical significance of each GATES suffers from multiple testing problems. To address these challenges, we develop a nonparametric test of treatment effect heterogeneity. In particular, we consider the following null hypothesis that all GATEs are equal to one another,

$$H_0: \ \tau_1 = \tau_2 = \cdots = \tau_K. \tag{4}$$

This null hypothesis is equivalent to $\tau_k = \tau$ for any k where $\tau = \mathbb{E}(Y_i(1) - Y_i(0))$ represents the overall average treatment effect (ATE). Thus, we consider the following test statistic,

$$\hat{\boldsymbol{\tau}} = (\hat{\tau}_1 - \hat{\tau}, \cdots, \hat{\tau}_K - \hat{\tau})^\top,$$

where

$$\hat{\tau} = \frac{1}{n_1}\sum_{i=1}^n Y_i T_i - \frac{1}{n_0}\sum_{i=1}^n Y_i(1 - T_i).$$

To derive the asymptotic reference distribution of this test statistic,

Imai and Li (2023b) derive the bias bound and the exact variance of this PAPE estimator. Leveraging those results, the following theorem shows that we can utilize a χ^2 distribution as an asymptotic approximation to the reference distribution when testing treatment effect heterogeneity.

Theorem 3 (Nonparametric Test of Treatment Effect Heterogeneity) *Suppose Assumptions 1–5 hold. Under H_0 defined in Equation (4) and against the alternative $H_1 : \mathbb{R}^K \setminus H_0$, as $n \to \infty$ with $0 < n_1/n < 1$ stays constant, we have,*

$$\hat{\boldsymbol{\tau}}^\top \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\tau}} \xrightarrow{d} \chi^2_K$$

where the entries of the covariance matrix $\boldsymbol{\Sigma}$ are defined as follows,

$$\begin{aligned}
\boldsymbol{\Sigma}_{kk} &= K^2 \Bigg[\frac{\mathbb{E}(S^{*2}_{k1})}{n_1} + \frac{\mathbb{E}(S^{*2}_{k0})}{n_0} \\
&\quad + \frac{1}{K^3}\left\{(K-2)\left(\frac{n-K}{n-1}\kappa_{kk11} - \kappa^2_{k1}\right) - \frac{2n(K-1)}{(n-1)}\kappa_{kk01} + 2\kappa_{k1}\kappa_{k0}\right\}\Bigg], \\
\boldsymbol{\Sigma}_{kk'} &= K^2 \left\{\frac{\mathbb{E}(S^{*2}_{kk'1})}{n_1} + \frac{\mathbb{E}(S^{*2}_{kk'0})}{n_0}\right\} + \frac{1}{K}\Bigg\{(K-2)\left(\kappa_{kk'11} - \kappa_{k1}\kappa_{k'1}\right) \\
&\quad - \frac{Kn - n - 1}{n-1}\left(\kappa_{kk'10} + \kappa_{kk'01}\right) + \kappa_{k1}\kappa_{k'0} + \kappa_{k0}\kappa_{k'1}\Bigg\},
\end{aligned}$$

for $k, k' \in \{1, \cdots K\}$ and $k \neq k'$ where $S^{*2}_{kt} = \sum_{i=1}^n (Y^*_{ki}(t) - \overline{Y^*_k(t)})^2/(n-1)$, $S^{*2}_{kk't} = \sum_{i=1}^n (Y^*_{ki}(t) - \overline{Y^*_k(t)})(Y^*_{k'i}(t) - \overline{Y^*_{k'}(t)})/(n-1)$, $\kappa_{kt} = \mathbb{E}(Y_i(1) - Y_i(0) \mid \hat{f}_k(\boldsymbol{X}_i) = t)$, and $\kappa_{kk'ts} = \mathbb{E}[(Y_i(1) - Y_i(0))(Y_j(1) - Y_j(0)) \mid \hat{f}_k(\boldsymbol{X}_i) = t, \hat{f}_{k'}(\boldsymbol{X}_i) = s]$ for $i \neq j$ with $Y^*_{ki}(t) = (\hat{f}_k(\boldsymbol{X}_i) - 1/K)Y_i(t)$, and $\overline{Y^*_k(t)} = \sum_{i=1}^n Y^*_{ki}(t)/n$, for $t = 0, 1$.

Proof is given in Supplementary Appendix S5. Similar to Theorem 1, there is an additional third term in the variance beyond the two standard terms, induced by the fact that $\hat{f}_k(\boldsymbol{X}_i)$ is negatively correlated across units. In practice, we replace the entries of $\boldsymbol{\Sigma}$ with their sample analogues, which result in a consistent estimator $\widehat{\boldsymbol{\Sigma}}$. By Slutsky's Lemma, the asymptotic distribution is not affected by this substitution.

2.4 Nonparametric Test of Rank-Consistent Treatment Effect Heterogeneity

To evaluate the quality of the scoring rule produced by an ML algorithm, we can test whether or not the rank of estimated GATES is consistent with that of the true GATES. The relevant null hypothesis is given by,

$$H_0^* : \tau_1 \leq \tau_2 \leq \cdots \leq \tau_K. \tag{5}$$

Unlike the null hypothesis for treatment effect heterogeneity given in Equation (4), this is a composite null hypothesis.

To characterize the sampling distribution under this null hypothesis H_0^*, we consider the following optimization problem,

$$\boldsymbol{\mu}^*(\boldsymbol{x}) \;=\; \operatorname*{argmin}_{\boldsymbol{\mu}} \|\boldsymbol{\mu} - \boldsymbol{x}\|_2^2 \quad \text{subject to } \mu_1 \leq \mu_2 \leq \cdots \leq \mu_K,$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K)^\top$ and $\boldsymbol{x} \in \mathbb{R}^K$. If $\boldsymbol{x} \sim N(0, \boldsymbol{\Sigma})$, the following test statistic has a mixture of appropriately weighted χ^2 distribution with K degrees of freedom, called chi-bar-squared distribution (Shapiro, 1988),

$$(\boldsymbol{x} - \boldsymbol{\mu}^*(\boldsymbol{x}))^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}^*(\boldsymbol{x})) \;\sim\; \bar{\chi}_K^2.$$

Using this fact, the next theorem derives a nonparametric test of rank-consistent treatment effect heterogeneity that is asymptotically uniformly most powerful.

THEOREM 4 (NONPARAMETRIC TEST OF RANK-CONSISTENT TREATMENT EFFECT HETEROGENEITY) *Suppose that Assumptions 1–5 hold. Then, as $n \to \infty$ and $0 < n_1/n < 1$ stays constant, an asymptotically uniformly most powerful test* *of size* α *for the null hypothesis* H_0^* *defined in Equation* (5) *against the alternative* $H_1^* : \mathbb{R}^K \setminus H_0^*$ *has the following critical region,*

$$\{\hat{\boldsymbol{\tau}} \in \mathbb{R}^K \mid (\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}}))^\top \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}})) > C_\alpha\},$$

for some constant C_α *that only depends on* α*. The expression of* $\boldsymbol{\Sigma}$ *is given in Theorem 3. Under* H_0^* *and as* $n \to \infty$*, we have,*

$$(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}^*(\hat{\boldsymbol{\tau}}))^\top \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}^*(\hat{\boldsymbol{\tau}})) \xrightarrow{d} \bar{\chi}_K^2.$$

Proof is given in Supplementary Appendix S6. In practice, we use Monte Carlo simulations to approximately compute the critical values.

While our test is the asymptotically most powerful test of its type, it is likely to be conservative as we control the critical value based on the worst-case scenario among all the distributions consistent with the null hypothesis. In the literature on statistical tests of moment inequalities, scholars have developed subsampling and moment selection techniques that can improve their statistical power (see e.g., Andrews and Guggenberger, 2009; Andrews and Soares, 2010; Canay, 2010; Chernozhukov *et al.*, 2019). Canay *et al.* (2023) provides an up-to-date review.

3 Generalization to Cross-Fitting

In this section, we generalize our methodology to *cross-fitting*, in which researchers use the same experimental data to first generate the scoring rule using an ML algorithm and then estimate the GATES based on the resulting scoring rule. In comparison with *sample splitting* discussed in Section 2 where they are done on separate samples, cross-fitting could potentially be much more efficient. The key challenge, however, is the incorporation of additional uncertainty due to the random splitting of the data. We show how to overcome this under the Neyman's repeated sampling framework.

3.1 Estimation and Inference

Under cross-fitting, we randomly divide the experimental data into $L \geq 2$ folds of equal size $m = n/L$ where for the sake of simplicity we assume n is a multiple of

L, and each fold contains m_1 treated units with m_0 control units, i.e, $m = m_0 + m_1$. We maintain Assumptions 1–3 introduced in Section 2.1. Then, for each $\ell = 1, 2, \ldots, L$, we use the ℓth fold as a validation dataset $\mathcal{Z}_\ell = \{\boldsymbol{X}_i^{(\ell)}, T_i^{(\ell)}, Y_i^{(\ell)}\}_{i=1}^m$ to conduct statistical tests and estimate the GATES. We use the remaining folds, $\mathcal{Z}_{-\ell} = \{\boldsymbol{X}_i^{(-\ell)}, T_i^{(-\ell)}, Y_i^{(-\ell)}\}_{i=1}^{n-m}$, as the training dataset to estimate the scoring rule with an ML algorithm.

Suppose that we define a generic ML algorithm as a deterministic map from the space of training data $\boldsymbol{\mathcal{Z}}_{\text{train}}$ to the space of scoring rules $\mathcal{S}$:

$$F : \boldsymbol{\mathcal{Z}}_{\text{train}} \to \mathcal{S}.$$

Then, for a given training data set $\mathcal{Z}_{\text{train}}$ of size $n - m$, the estimated scoring rule is given by,

$$\hat{s}_{\mathcal{Z}_{\text{train}}^{n-m}} \;=\; F(\mathcal{Z}_{\text{train}}^{n-m}). \tag{6}$$

We now generalize the definition of the GATES to the cross-fitting case,

$$\tau_k(F, n-m) \;=\; \mathbb{E}[\mathbb{E}\{Y_i(1) - Y_i(0) \mid c_{k-1}(\hat{s}_{\mathcal{Z}_{\text{train}}^{n-m}}) \leq \hat{s}_{\mathcal{Z}_{\text{train}}^{n-m}}(\boldsymbol{X}_i) \leq c_k(\hat{s}_{\mathcal{Z}_{\text{train}}^{n-m}})\}], \tag{7}$$

where the inner expectation is taken over the distribution of $\{\boldsymbol{X}_i, Y_i(0), Y_i(1)\}$ among the units who belong to the kth group, and the outer expectation is taken over all possible training sets of size $n - m$ from $\mathcal{Z}_{\text{train}}^{n-m}$ the population $\mathcal{P}$.

This generalized GATES is not a function of fixed scoring rule. Rather, it is a function of ML algorithm F itself (as well as the sample size of training data, $n-m$). Intuitively, it represents the average of GATES based on all observations that score between $(k-1)/K \times 100$th percentile and $k/K \times 100$th percentile under the ML algorithm F across all possible training datasets of size $n - m$. Alternatively, the cross-fitted GATE can be seen as a weighted average of GATEs that are specific to scoring rules where weights are determined by the training data and the particular ML algorithm.

Algorithm 1 Estimation of the Sorted Group Average Treatment Effects (GATES) under Cross-fitting

Input: Data $\mathcal{Z} = \{\boldsymbol{X}_i, T_i, Y_i\}_{i=1}^n$, Machine learning algorithm F, Estimator $\hat{\tau}_k$, Number of folds L

Output: Estimated GATES $\{\hat{\tau}_k(F, n-m)\}_{k=1}^K$

Split the data $\mathcal{Z}$ into L random subsets of equal size $\{\mathcal{Z}_1, \cdots, \mathcal{Z}_L\}$
Set $m \leftarrow n/L$ and $\ell \leftarrow 1$
while $\ell \leq L$ **do**
  $\mathcal{Z}_{-\ell} = \{\mathcal{Z}_1, \cdots, \mathcal{Z}_{\ell-1}, \mathcal{Z}_{\ell+1}, \cdots, \mathcal{Z}_L\}$ ▷ Create the training dataset
  $\hat{s}_{-\ell} = F(\mathcal{Z}_{-\ell})$ ▷ Estimate the scoring rule s by applying F to $\mathcal{Z}_{-\ell}$
  $\hat{\tau}_k^\ell = \hat{\tau}_k(\mathcal{Z}_\ell)$ for each $k = 1, 2, \ldots, K$ ▷ Calculate the GATES estimator using $\mathcal{Z}_\ell$
  $\ell \leftarrow \ell + 1$
end while
return $\hat{\tau}_k(F, n-m) = \frac{1}{L}\sum_{\ell=1}^L \hat{\tau}_k^\ell$ for each $k = 1, 2, \ldots, K$

We describe estimation and inference for $\tau_k(F, n-m)$. For each fold ℓ, we first estimate a scoring rule s by applying an ML algorithm F to the training data $\mathcal{Z}_{-\ell}$,

$$\hat{s}_\ell = F(\mathcal{Z}_{-\ell}). \tag{8}$$

We then estimate the GATES based on the validation data $\mathcal{Z}_\ell$, using the following estimator that is analogous to the one defined in Equation (3),

$$\begin{aligned}&\hat{\tau}_k^\ell(F, n-m)\\ = \; & K\left[\frac{1}{m_1}\sum_{i=1}^m Y_i^{(\ell)} T_i^{(\ell)} \hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)}) + \frac{1}{m_0}\sum_{i=1}^m Y_i^{(\ell)}(1-T_i^{(\ell)})\left\{1 - \hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)})\right\} - \frac{1}{m_0}\sum_{i=1}^m Y_i^{(\ell)}(1-T_i^{(\ell)})\right],\end{aligned}$$

where $\hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)}) = \mathbf{1}\{\hat{s}_\ell(\boldsymbol{X}_i^{(\ell)}) \geq \hat{c}_{k-1}^\ell(\hat{s}_\ell)\} - \mathbf{1}\{\hat{s}_\ell(\boldsymbol{X}_i^{(\ell)}) \geq \hat{c}_k^\ell(\hat{s}_\ell)\}$, and $\hat{c}_k^\ell(\hat{s}_\ell) = \inf\{c \in \mathbb{R} : \sum_{i=1}^m \mathbf{1}\{\hat{s}_\ell(\boldsymbol{X}_i^{(\ell)}) > c\} \leq mk/K\}$ represents the estimated cutoff in the ℓth subsample. Repeating this for each fold and averaging the results gives us the final GATES estimator,

$$\hat{\tau}_k(F, n-m) = \frac{1}{L}\sum_{\ell=1}^L \hat{\tau}_k^\ell \tag{9}$$

for $k = 1, 2, \ldots, K$. Algorithm 1 summarizes this estimation procedure.

We extend our bias and variance results under sample splitting (Theorem 1) to the cross-fitting case by incorporating the additional randomness induced by the cross-fitting procedure.

Theorem 5 (Bias Bound and Exact Variance of the GATES Estimator under Cross-fitting) *Under Assumptions 1–3, the bias of the proposed GATES estimator given in Equation* (9) *can be bounded as follows,*

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{P}\left(\left|\mathbb{E}\{\hat{\tau}_k(F,n-m)-\tau_k(F,n-m)\mid \hat{c}_k(\hat{s}_{\mathcal{Z}^{n-m}_{train}}),\hat{c}_{k-1}(\hat{s}_{\mathcal{Z}^{n-m}_{train}})\}\right|\geq\epsilon \;\middle|\; \mathcal{Z}^{n-m}_{train}\right)\right]\\
\leq\;& 1-B\left(\frac{k}{K}+\gamma_k(\epsilon),\frac{nk}{K},n-\frac{nk}{K}+1\right)+B\left(\frac{k}{K}-\gamma_k(\epsilon),\frac{nk}{K},n-\frac{nk}{K}+1\right)\\
&-B\left(\frac{k-1}{K}+\gamma_{k-1}(\epsilon),\frac{n(k-1)}{K},n-\frac{n(k-1)}{K}+1\right)\\
&+B\left(\frac{k-1}{K}-\gamma_{k-1}(\epsilon),\frac{n(k-1)}{K},n-\frac{n(k-1)}{K}+1\right),
\end{aligned}$$

for any given constant $\epsilon>0$ where $B(\epsilon,\alpha,\beta)$ is the incomplete beta function (if $\alpha\leq 0$ and $\beta>0$, we set $B(\epsilon,\alpha,\beta):=H(\epsilon)$ for all ϵ where $H(\epsilon)$ is the Heaviside step function), and

$$\gamma_k(\epsilon) \;=\; \frac{\epsilon}{K\mathbb{E}\{\max_{c\in[c_k(\hat{s}_{\mathcal{Z}^{n-m}_{train}}(\boldsymbol{X}_i))-\epsilon,\ c_k(\hat{s}_{\mathcal{Z}^{n-m}_{train}}(\boldsymbol{X}_i))+\epsilon]}\mathbb{E}(Y_i(1)-Y_i(0)\mid \hat{s}_{\mathcal{Z}^{n-m}_{train}}(\boldsymbol{X}_i)=c)\}}.$$

The variance of the estimator is given by,

$$\begin{aligned}
&\mathbb{V}(\hat{\tau}_k(F,n-m))\\
&=K^2\left(\frac{\mathbb{E}(S^2_{Fk1})}{m_1}+\frac{\mathbb{E}(S^2_{Fk0})}{m_0}\right)+\frac{(n-K)\mathbb{E}_\ell(\kappa^\ell_{k11})}{n-1}-\mathbb{E}_\ell[(\kappa^\ell_{k1})^2]+\mathbb{V}\left(\kappa^\ell_{k1}\right)-\frac{L-1}{L}\mathbb{E}(S^2_{Fk}),
\end{aligned}$$

where $S^2_{Fkt}=\sum_{i=1}^m(Y^\ell_{ki}(t)-\overline{Y^\ell_k(t)})^2/(m-1)$, $S^2_{Fk}=\sum_{\ell=1}^L(\hat{\tau}^{(\ell)}_k-\hat{\tau}_k(F,n-m))^2/(L-1)$, $\kappa^\ell_{kt}=\mathbb{E}(Y_i(1)-Y_i(0)\mid \hat{f}^\ell_k(\boldsymbol{X}_i)=t)$, and $\kappa^\ell_{ktt}=\mathbb{E}[(Y_i(1)-Y_i(0))(Y_j(1)-Y_j(0))\mid \hat{f}^\ell_k(\boldsymbol{X}_i)=\hat{f}^\ell_k(\boldsymbol{X}_j)=t]$ for $i\neq j$ with $Y^\ell_{ki}(t)=\hat{f}^\ell_k(\boldsymbol{X}^{(\ell)}_i)Y^{(\ell)}_i(t)$, and $\overline{Y^\ell_k(t)}=\sum_{i=1}^m Y^\ell_{ki}(t)/n$, for $t=0,1$.

Proof is given in Supplementary Appendix S7. When compared to Theorem 1, although the bias bound is of a similar form, the variance expression implies two additional terms. The first additional term, $\mathbb{V}\left(\kappa^\ell_{k1}\right)$, accounts for the variation across training data sets. The second negative term, $-(L-1)\mathbb{E}(S^2_{Fk})/L$, represents the efficiency gain of the cross-fitting procedure. As expected, when $L=1$, the expression reduces to the sample splitting case (see Theorem 1).

The estimation of $\mathbb{E}(S^2_{Fkt})$, $\mathbb{E}\{(\kappa^\ell_{kt})^2\}$, $\mathbb{E}\{(\kappa^\ell_{ktt})\}$ and $\mathbb{V}(\kappa^\ell_{kt})$ is straightforward and based on their sample analogues:

$$\widehat{\mathbb{E}(S^2_{Fkt})} \;=\; \frac{1}{(m-1)L}\sum_{\ell=1}^{L}\sum_{i=1}^{m}\mathbf{1}\{T^{(\ell)}_i=t\}(Y^\ell_{ki}-\overline{Y^\ell_{kt}})^2,$$

$$\mathbb{E}\widehat{\left\{(\kappa_{kt}^{\ell})^2\right\}} \;=\; \frac{1}{L}\sum_{\ell=1}^{L}\left(\hat{\kappa}_{kt}^{\ell}\right)^2, \quad \widehat{\mathbb{V}(\kappa_{kt}^{\ell})} \;=\; \frac{1}{L-1}\sum_{\ell=1}^{L}(\hat{\kappa}_{kt}^{\ell}-\overline{\hat{\kappa}_{kt}^{\ell}})^2,$$

where $Y_{ki}^{\ell}=\hat{f}_k^{\ell}(\boldsymbol{X}_i)Y_i^{(\ell)}$, $\overline{Y_{kt}^{\ell}}=\sum_{i=1}^{m}\mathbf{1}\{T_i=t\}Y_{ki}^{(\ell)}/m$, $\overline{\hat{\kappa}_{kt}^{\ell}}=\sum_{\ell=1}^{L}\hat{\kappa}_{kt}^{\ell}/L$ and

$$\begin{aligned}
\hat{\kappa}_{kt}^{\ell} =& \frac{\sum_{i=1}^{m}\mathbf{1}\{\hat{f}_k^{\ell}(\boldsymbol{X}_i^{(\ell)})=t\}T_i^{(\ell)}Y_i^{(\ell)}}{\sum_{i=1}^{m}\mathbf{1}\{\hat{f}_k^{\ell}(\boldsymbol{X}_i^{(\ell)})=t\}T_i^{(\ell)}} - \frac{\sum_{i=1}^{m}\mathbf{1}\{\hat{f}_k^{\ell}(\boldsymbol{X}_i^{(\ell)})=t\}(1-T_i^{(\ell)})Y_i^{(\ell)}}{\sum_{i=1}^{m}\mathbf{1}\{\hat{f}_k^{\ell}(\boldsymbol{X}_i^{(\ell)})=t\}(1-T_i^{(\ell)})}, \\
\hat{\kappa}_{ktt}^{\ell} =& \frac{[\sum_{i=1}^{m}\mathbf{1}\{\hat{f}_k^{\ell}(\boldsymbol{X}_i^{(\ell)})=t\}T_i^{(\ell)}Y_i^{(\ell)}]^2-\sum_{i=1}^{m}\mathbf{1}\{\hat{f}_k^{\ell}(\boldsymbol{X}_i^{(\ell)})=t\}T_i^{(\ell)}(Y_i^{(\ell)})^2}{[\sum_{i=1}^{m}\mathbf{1}\{\hat{f}_k^{\ell}(\boldsymbol{X}_i^{(\ell)})=t\}T_i^{(\ell)}]^2-\sum_{i=1}^{m}\mathbf{1}\{\hat{f}_k^{\ell}(\boldsymbol{X}_i^{(\ell)})=t\}T_i} \\
&-\frac{[\sum_{i=1}^{m}\mathbf{1}\{\hat{f}_k^{\ell}(\boldsymbol{X}_i^{(\ell)})=t\}(1-T_i^{(\ell)})Y_i^{(\ell)}]^2-\sum_{i=1}^{m}\mathbf{1}\{\hat{f}_k^{\ell}(\boldsymbol{X}_i^{(\ell)})=t\}(1-T_i^{(\ell)})(Y_i^{(\ell)})^2}{[\sum_{i=1}^{m}\mathbf{1}\{\hat{f}_k^{\ell}(\boldsymbol{X}_i^{(\ell)})=t\}(1-T_i^{(\ell)})]^2-\sum_{i=1}^{m}\mathbf{1}\{\hat{f}_k^{\ell}(\boldsymbol{X}_i^{(\ell)})=t\}(1-T_i^{(\ell)})},
\end{aligned}$$

In contrast, the estimation of $\mathbb{E}(S_{Fk}^2)$ requires care. In particular, although it is tempting to estimate $\mathbb{E}(S_{Fk}^2)$ using a realization of S_{Fk}^2, this estimate is highly variable especially when L is small. As a result, it often yields a negative overall variance estimate. We address this problem by applying Lemma 1 from Nadeau and Bengio (2000) to $\hat{\tau}_k(F,n-m)$, which gives,

$$\mathbb{V}(\hat{\tau}_k(F,n-m)) \geq \; \mathbb{E}(S_{Fk}^2).$$

Since Theorem 5 implies:

$$\mathbb{V}(\hat{\tau}_k(F,n-m)) \leq \; K^2\left(\frac{\mathbb{E}(S_{Fk1}^2)}{m_1}+\frac{\mathbb{E}(S_{Fk0}^2)}{m_0}\right)+\frac{(n-K)\mathbb{E}_{\ell}[\kappa_{k11}^{\ell}]}{n-1}-\mathbb{E}_{\ell}[(\kappa_{k1}^{\ell})^2]+\mathbb{V}\left(\kappa_{k1}^{\ell}\right),$$

this inequality suggests the following estimator of $\mathbb{E}(S_{Fk}^2)$,

$$\widehat{\mathbb{E}(S_{Fk}^2)} \;=\; \min\left(S_{Fk}^2, K^2\left(\frac{\widehat{\mathbb{E}(S_{Fk1}^2)}}{m_1}+\frac{\widehat{\mathbb{E}(S_{Fk0}^2)}}{m_0}\right)+\frac{(n-K)\mathbb{E}_{\ell}[\kappa_{k11}^{\ell}]}{n-1}-\mathbb{E}_{\ell}[(\kappa_{k1}^{\ell})^2]+\widehat{\mathbb{V}\left(\kappa_{k1}^{\ell}\right)}\right). \tag{10}$$

Although this yields a conservative estimate of $\mathbb{V}(\hat{\tau}_k(F,n-m))$ in finite samples, the bias appears to be relatively small in practice (see Section 4). In Appendix S8, we show that the estimator is consistent as L goes to infinity and sufficiently large m.

To establish the asymptotic sampling distribution of our cross-fitting GATES estimator, we first extend our CATE continuity condition (Assumption 4) by assuming that each CATE given a training data set is continuous and the average CATE (over all possible training data sets) is bounded.

Assumption 6 (Continuity of CATE at the Thresholds under Cross-Fitting) *Let* $F_{\mathcal{Z}^{n-m}_{train}}(c) = \Pr(\hat{s}_{\mathcal{Z}^{n-m}_{train}}(\boldsymbol{X}_i) \leq c)$ *represent the cumulative distribution function of the scoring rule under training set* $\mathcal{Z}^{n-m}_{train}$ *and define its pseudo-inverse as* $F^{-1}_{\mathcal{Z}^{n-m}_{train}}(p) = \inf\{c : F_{\mathcal{Z}^{n-m}_{train}}(c) \geq p\}$ *for* $p \in [0,1]$*. Then, for all but asymptotically measure-zero set of possible training sets* $\mathcal{Z}^{n-m}_{train}$ *of size* $n-m$*, the CATE function* $\tau_{\mathcal{Z}^{n-m}_{train}}(p) = \mathbb{E}(Y_i(1) - Y_i(0) \mid \hat{s}_{\mathcal{Z}^{n-m}_{train}}(\boldsymbol{X}_i) = F^{-1}_{\mathcal{Z}^{n-m}_{train}}(p))$ *is left-continuous with bounded variation on any interval* $(\epsilon, 1-\epsilon)$ *with* $0 < \epsilon < 1/2$*, and continuous in* p *at* $p = 1/K, \ldots, (K-1)/K$*. Furthermore, we assume* $\lim_{n\to\infty} \mathbb{E}_{\mathcal{Z}^{n-m}_{train}}[\max_{p\in[0,1]} \tau_{\mathcal{Z}^{n-m}_{train}}(p)] < \infty$*.*

In addition, we require the ML algorithm F to be stable.

Assumption 7 (ML Algorithm Stability) *Let* $\mathcal{Z}^n_{train}$ *be a training dataset of size* n *and* $\hat{s}_{\mathcal{Z}^n_{train}} = F(\mathcal{Z}^n_{train})$ *represent the estimated scoring rule that results from the application of an ML algorithm* F *to the training dataset. Then, as* $m \to \infty$ *(with* L *fixed), for any* a, b *with* $a < b$*:*

$$\|\mathbb{E}[Y_i(1) - Y_i(0) \mid a \leq \hat{s}_{\mathcal{Z}^n_{train}}(\boldsymbol{X}_i) \leq b]\|_2 = o\left(m^{-1}\right).$$

The expectation is taken over the distribution of $\{\boldsymbol{X}_i, Y_i(0), Y_i(1)\}$ among those units in the population $\mathcal{P}$ who belong to the group defined by the conditioning set. The outer norm is computed across the random sampling of training data set of size n from the same population. Assumption 7 implies that as the size of training data approaches infinity, L_2 norm of the resulting scoring rule $\hat{s}_{\mathcal{Z}^n_{\text{train}}}$ stabilizes sufficiently quickly at a rate faster than $O(m^{-1})$. The required rate is consistent with the asymptotic conditions needed for other related cross-validation settings (e.g., Austern and Zhou, 2020). Importantly, we do not assume that the ML algorithm converges to the true CATE.

Finally, the next theorem established the asymptotic distribution of GATES estimator under cross-fitting.

Theorem 6 (Asymptotic Sampling Distribution of GATES Estimator under Cross-Fitting) *Suppose* L *is fixed. Then, under Assumptions 1–3, 5–7, we have, as* m *goes to infinity,*

$$\frac{\hat{\tau}_k(F, n-m) - \tau_k(F, n-m)}{\sqrt{\mathbb{V}(\hat{\tau}_k(F, n-m))}} \xrightarrow{d} N(0,1)$$

where the expression of $\mathbb{V}(\hat{\tau}_k(F, n-m))$ *is given in Theorem 5.*

Proof is given in Supplementary Appendix S9, and is similar to the proof of Theorem 2.

3.2 Nonparametric Tests of Treatment Effect Heterogeneity

We now extend the nonparametric tests of treatment effect heterogeneity and its rank-consistency introduced in Sections 2.3 and 2.4 to the cross-fitting setting. Similar to Chernozhukov *et al.* (2023), we account for the additional uncertainty due to random splitting. Unlike their method, however, the proposed tests do not require a computationally intensive resampling procedure.

Our first null hypothesis of interest is that the GATES are all equal to the ATE,

$$H_{F0}: \ \tau_1(F, n-m) \ = \ \tau_2(F, n-m) \ = \ \cdots \ = \ \tau_K(F, n-m). \tag{11}$$

This null hypothesis depends on the ML algorithm F whereas the null hypothesis given in Equation (4) depends on the (fixed) scoring rule.

The following theorem generalizes the result of Theorem 3 to cross-fitting.

Theorem 7 (Nonparametric Test of Treatment Effect Heterogeneity Under Cross-fitting) *Suppose L is fixed. Then, under Assumptions 1–3, 5–7, and the null hypothesis H_{F0} defined in Equation* (11) *and against the alternative $H_{F1}: \mathbb{R}^K \setminus H_{F0}$, as $m \to \infty$, and $0 < m_1/m < 1$ stays constant, we have,*

$$\hat{\boldsymbol{\tau}}_F^\top \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\tau}}_F \ \xrightarrow{d} \ \chi^2_K$$

where $\hat{\boldsymbol{\tau}}_F = (\hat{\tau}_1(F, n-m) - \hat{\tau}, \cdots, \hat{\tau}_K(F, n-m) - \hat{\tau})$, and $\boldsymbol{\Sigma}$ is defined as for $k, k' \in \{1, \cdots K\}$:

$$\begin{aligned}
\boldsymbol{\Sigma}_{kk} \ &= K^2\left(\frac{\mathbb{E}(S^{*2}_{Fk1})}{m_1} + \frac{\mathbb{E}(S^{*2}_{Fk0})}{m_0}\right) - \frac{L-1}{L}\mathbb{E}(S^2_{Fk}) + \mathbb{V}\left(\kappa^\ell_{k1}\right) \\
&\quad + \frac{1}{K}\mathbb{E}_\ell\left\{(K-2)\left(\frac{n-K}{n-1}\kappa^\ell_{kk11} - (\kappa^\ell_{k1})^2\right) - \frac{2n(K-1)}{(n-1)}\kappa^\ell_{kk01} + 2\kappa^\ell_{k1}\kappa^\ell_{k0}\right\} \\
\boldsymbol{\Sigma}_{kk'} \ &= K^2\left(\frac{\mathbb{E}(S^{*2}_{Fkk'1})}{m_1} + \frac{\mathbb{E}(S^{*2}_{Fkk'0})}{m_0}\right) - \frac{L-1}{L}\mathbb{E}(S^2_{Fkk'}) + \mathrm{Cov}\left(\kappa^\ell_{k1}, \kappa^\ell_{k'1}\right) \\
&\quad + \ \frac{1}{K}\mathbb{E}_\ell\left\{(K-2)\left(\kappa^\ell_{kk'11} - \kappa^\ell_{k1}\kappa^\ell_{k'1}\right) - \frac{Kn-n-1}{n-1}\left(\kappa^\ell_{kk'10} + \kappa^\ell_{kk'01}\right) + \kappa^\ell_{k1}\kappa^\ell_{k'0} + \kappa^\ell_{k0}\kappa^\ell_{k'1}\right\}
\end{aligned}$$

where $S^{*2}_{Fkt} = \sum_{i=1}^m (Y^{*\ell}_{ki}(t) - \overline{Y^{*\ell}_k(t)})^2/(m-1)$, $S^{*2}_{Fkk't} = \sum_{i=1}^m (Y^{*\ell}_{ki}(t) - \overline{Y^{*\ell}_k(t)})(Y^\ell_{k'i}(t) - \overline{Y^{*\ell}_{k'}(t)})/(m-1)$, $S^2_{Fkk'} = \sum_{\ell=1}^L (\hat{\tau}^\ell_k(F, n-m) - \hat{\tau}_k(F, n-m))(\hat{\tau}^\ell_{k'}(F, n-m) - \hat{\tau}_{k'}(F, n-$

$m))/(L-1)$, $\kappa^{\ell}_{kt} = \mathbb{E}(Y_i(1)-Y_i(0) \mid \hat{f}^{\ell}_k(\boldsymbol{X}_i) = t)$ and $\kappa^{\ell}_{kk'ts} = \mathbb{E}[(Y_i(1)-Y_i(0))(Y_j(1)-Y_j(0)) \mid \hat{f}^{\ell}_k(\boldsymbol{X}_i) = t, \hat{f}^{\ell}_{k'}(\boldsymbol{X}_i) = s]$ with $Y^{*\ell}_{ki}(t) = (\hat{f}^{\ell}_k(\boldsymbol{X}^{(\ell)}_i) - 1/K)Y^{(\ell)}_i(t)$, and $\overline{Y^{*\ell}_k(t)} = \sum_{i=1}^m Y^{*\ell}_{ki}(t)/m$, for $t = 0, 1$.

Proof is given in Supplementary Appendix S10. Compared to Theorem 3, the only difference appears in the expression of the covariance matrix $\boldsymbol{\Sigma}$ due to the efficiency gains of the cross-validation procedure. Similar to Theorem 5, the estimation of $\mathbb{E}(S^2_{Fkk'})$ for $k = k'$ requires care, and we utilize the consistent estimator as identified in Equation (10). If the resulting covariance matrix estimate is not positive definite, we find the nearest positive definite matrix in the L_2 norm by utilizing the algorithm of Higham (2002).

Finally, we extend the nonparametric test of rank-consistent treatment effect heterogeneity (Theorem 4) to cross-fitting. The null hypothesis is given by,

$$H^*_{F0} : \tau_1(F, n-m) \le \tau_2(F, n-m) \le \cdots \le \tau_K(F, n-m). \tag{12}$$

Now, we present the result.

Theorem 8 (Nonparametric Test of Rank-Consistent Treatment Effect Heterogeneity Under Cross-Fitting) *Suppose L is fixed. Then, under Assumptions 1–3, 5–7, as $m \to \infty$ and $0 < m_1/m < 1$ stays constant, the uniformly most powerful test of size α for the null hypothesis H^*_{F0} defined in Equation (12) against the alternative $H^*_{F1} : \mathbb{R}^K \setminus H^*_{F0}$ has the following critical region,*

$$\{\hat{\boldsymbol{\tau}}_F \in \mathbb{R}^K \mid (\hat{\boldsymbol{\tau}}_F - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}}_F))^\top \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\tau}}_F - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}}_F)) > C_\alpha\},$$

for some constant C_α that only depends on α. Furthermore, under H_{F0} and as $n \to \infty$, we have,

$$(\hat{\boldsymbol{\tau}}_F - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}}_F))^\top \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\tau}}_F - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}}_F)) \xrightarrow{d} \bar{\chi}^2_K,$$

where $\hat{\boldsymbol{\tau}}_F$ and $\boldsymbol{\Sigma}$ are defined in Theorem 7.

Proof directly follows from the fact by Theorem 7, $\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\tau}}_F$ is asymptotically normally distributed with variance-covariance matrix $\boldsymbol{I}$, which is an identity matrix of size $K \times K$. Then, following the same steps as those in Supplementary Appendix S6 immediately establishes the result.

4 A Simulation Study

We undertake a simulation study to examine the finite sample performance of the proposed methodology. We consider both sample-splitting and cross-fitting cases. For the estimation of GATES, we evaluate the bias and variance of the proposed estimators as well as the coverage of their confidence intervals. For hypothesis tests, we examine the actual power and size of the proposed tests. We show that the proposed methodology performs well even when the sample size is as small as 100.

4.1 The Setup

We utilize the data generating process from the 2016 Atlantic Causal Inference Conference (ACIC) Competition. We briefly describe its simulation setting here and refer interested readers to Dorie *et al.* (2019) for additional details. The focus of this competition was the inference of average treatment effect in observational studies. There are a total of 58 pre-treatment covariates $\boldsymbol{X}$, including 3 categorical, 5 binary, 27 count data, and 13 continuous variables. The data were taken from a real-world study with the sample size $n = 4802$.

In our simulation, we assume that the empirical distribution of these covariates represent the population $\mathcal{P}$ and obtain each simulation sample via bootstrap. We consider small and moderate sample sizes: $n = 100$, 500, and 2500. For the sample-splitting case, the models are pre-trained on the original dataset from the 2016 ACIC data challenge, and the sample size n refers to the number of testing samples. For the cross-validation case, n refers to the total dataset size, which we then conduct 5-fold cross-validation, $L = 5$. One important change from the original competition is that instead of utilizing a propensity model to determine T, we assume that the treatment assignment is completely randomized, i.e., $\Pr(T_i = 1) = 1/2$, and the treatment and control groups are of equal size, i.e., $n_1 = n_0 = n/2$.

To generate the outcome variable, we use one of the settings from the competition, which is based on the generalized additive model with polynomial basis functions. The model represents a setting, in which there exists a substantial amount of treatment effect heterogeneity. The formula for this outcome model is reproduced here:

$$
\begin{aligned}
\mathbb{E}(Y_i(t) \mid \boldsymbol{X}_i) = {} & 1.60 + 0.53 \times x_{29} - 3.80 \times x_{29}(x_{29} - 0.98)(x_{29} + 0.86) - 0.32 \times \mathbf{1}\{x_{17} > 0\} \\
& + 0.21 \times \mathbf{1}\{x_{42} > 0\} - 0.63 \times x_{27} + 4.68 \times \mathbf{1}\{x_{27} < -0.61\} - 0.39 \times (x_{27} + 0.91)\mathbf{1}\{x_{27} < -0.91\} \\
& + 0.75 \times \mathbf{1}\{x_{30} \leq 0\} - 1.22 \times \mathbf{1}\{x_{54} \leq 0\} + 0.11 \times x_{37}\mathbf{1}\{x_4 \leq 0\} - 0.71 \times \mathbf{1}\{x_{17} \leq 0, t = 0\} \\
& - 1.82 \times \mathbf{1}\{x_{42} \leq 0, t = 1\} + 0.28 \times \mathbf{1}\{x_{30} \leq 0, t = 0\} \\
& + \{0.58 \times x_{29} - 9.42 \times x_{29}(x_{29} - 0.67)(x_{29} + 0.34)\} \times \mathbf{1}\{t = 1\} \\
& + (0.44 \times x_{27} - 4.87 \times \mathbf{1}\{x_{27} < -0.80\}) \times \mathbf{1}\{t = 0\} - 2.54 \times \mathbf{1}\{t = 0, x_{54} \leq 0\}.
\end{aligned}
$$

Throughout, we set $K = 5$ so that observations are sorted into five groups based on the magnitude of the CATE. For the case of sample-splitting, we can directly compute the true values of GATES using the outcome model and evaluate each quantity based on the entire original data set. For the cross-validation case, however, the exact calculation of GATES true values would require averaging over all combinations of training data sets from the original data set. Since this is computationally prohibitive, we obtain their approximate true values by independently sampling 10,000 training data sets. For each training dataset, we train an ML algorithm F using 5-fold cross-validation. Then, we use the sample mean of each estimated causal quantity across the 10,000 simulated data sets as our approximate truth.

We evaluate Bayesian Additive Regression Trees (BART) (see Chipman *et al.*, 2010; Hill, 2011; Hahn *et al.*, 2020) and Causal Forest (Wager and Athey, 2018), and LASSO (Tibshirani, 1996). The number of trees were tuned through the 5-fold cross validation for both algorithms. For implementation, we use R 3.6.3 with the following packages: bartMachine (version 1.2.6) for BART, grf (version 2.0.1) for Causal Forest, and glmnet (version 4.1-2) for LASSO. The number of trees was tuned through 5-

		$n_{\text{test}} = 100$			$n_{\text{test}} = 500$			$n_{\text{test}} = 2500$		
Estimator	truth	bias	s.d.	coverage	bias	s.d.	coverage	bias	s.d.	coverage
Causal Forest										
$\hat{\tau}_1$	2.164	0.034	2.486	93.8%	0.041	1.071	95.0%	0.007	0.467	96.0%
$\hat{\tau}_2$	4.001	0.011	2.551	93.7	−0.060	1.183	94.4	−0.002	0.510	95.3
$\hat{\tau}_3$	4.583	−0.018	2.209	94.0	−0.003	0.956	96.4	0.020	0.421	95.8
$\hat{\tau}_4$	4.931	−0.077	2.500	94.6	0.001	1.138	94.3	0.003	0.517	95.6
$\hat{\tau}_5$	5.728	−0.058	3.332	96.0	−0.010	1.499	95.0	−0.009	0.661	95.2
BART										
$\hat{\tau}_1$	2.092	0.016	3.188	94.0%	−0.014	1.402	96.2%	0.009	0.626	95.8%
$\hat{\tau}_2$	3.913	0.127	2.918	95.1	0.028	1.388	94.0	−0.003	0.618	95.3
$\hat{\tau}_3$	4.478	−0.077	2.218	94.3	−0.041	0.968	95.0	−0.001	0.425	95.1
$\hat{\tau}_4$	5.042	−0.154	2.366	94.2	0.014	1.106	95.8	0.015	0.495	95.4
$\hat{\tau}_5$	5.881	−0.019	2.510	94.7	−0.019	1.104	94.4	−0.000	0.489	95.0
LASSO										
$\hat{\tau}_1$	3.243	0.028	2.507	94.1%	0.049	1.119	95.1%	0.003	0.769	95.1%
$\hat{\tau}_2$	3.817	−0.012	1.848	93.6	−0.013	0.834	94.5	−0.000	0.684	95.4
$\hat{\tau}_3$	4.318	−0.013	2.095	94.2	−0.002	0.930	94.5	0.010	0.516	95.0
$\hat{\tau}_4$	4.788	−0.041	2.475	94.0	−0.015	1.101	94.6	−0.001	0.480	94.6
$\hat{\tau}_5$	5.241	−0.046	3.921	94.4	0.021	1.739	95.1	0.002	0.505	95.3

Table 1: The Finite Sample Performance of the GATES Estimators under Sample-splitting. The table presents the estimated bias and standard deviation of the GATES estimators as well as the empirical coverage of their 95% confidence intervals for Causal Forest, BART, and LASSO. The machine learning algorithms are trained on the original dataset from the 2016 ACIC data challenge.

fold cross-validation for BART and Causal Forest. The regularization parameter was tuned similarly for LASSO.

4.2 Finite-Sample Performance of the Proposed Estimators

Table 1 presents the results for the estimation of GATES in the sample-splitting case. According to this simulation setup, Causal Forest and BART appear to identify treatment effect heterogeneity better than LASSO. For example, for BART, the largest and smallest GATES are 5.89 and 2.09, respectively. In contrast, the gap between the corresponding quantities is much smaller for the LASSO, roughly equaling 2 points.

For each sample size, we conducted 1,000 simulation trials. For all three algorithms, the estimated biases of the proposed GATES estimators are negligibly small, accounting for less than 5% of their estimated standard deviation in almost all cases.

Estimator	$n = 100$ truth	bias	s.d.	coverage	$n = 500$ truth	bias	s.d.	coverage	$n = 2500$ truth	bias	s.d.	coverage
Causal Forest												
$\hat{\tau}_1$	3.976	−0.053	2.971	94.0%	2.900	−0.007	1.572	95.6%	2.210	−0.007	0.594	97.7%
$\hat{\tau}_2$	4.173	−0.061	2.584	95.9	4.112	−0.038	1.075	98.2	4.057	0.011	0.541	98.6
$\hat{\tau}_3$	4.286	−0.012	2.560	96.7	4.510	−0.054	1.058	97.7	4.545	0.019	0.465	98.1
$\hat{\tau}_4$	4.400	−0.119	2.865	97.4	4.799	0.066	1.149	97.9	4.951	−0.009	0.509	98.6
$\hat{\tau}_5$	4.569	0.140	3.447	94.1	5.086	0.001	1.620	96.0	5.643	−0.006	0.620	98.3
LASSO												
$\hat{\tau}_1$	4.191	−0.125	3.196	97.6%	4.017	−0.025	1.488	96.0%	3.752	−0.004	0.669	96.0%
$\hat{\tau}_2$	4.205	0.036	2.281	97.5	4.137	−0.069	1.027	97.9	4.028	−0.019	0.590	98.9
$\hat{\tau}_3$	4.268	−0.126	2.354	96.6	4.291	−0.019	1.000	97.9	4.323	0.037	0.488	97.5
$\hat{\tau}_4$	4.334	−0.003	2.536	96.8	4.430	0.035	1.174	96.8	4.571	0.033	0.642	97.2
$\hat{\tau}_5$	4.406	0.111	3.615	96.2	4.530	0.047	1.811	95.0	4.732	0.022	0.697	95.3

Table 2: The Finite Sample Performance of the GATES Estimators under Cross-fitting. The table presents the estimated bias and standard deviation of the proposed GATES estimators as well as the empirical coverage of their 95% confidence intervals for Causal Forest and LASSO.

The bias also generally decreases as the sample size grows. We also find that the empirical coverage of the confidence intervals is close to the theoretical 95% value even when the sample size is as small as $n = 100$.

We obtain similar findings for the cross-fitting case. Table 2 shows the results for Causal Forest and LASSO. Unfortunately, BART is too computationally intensive to include for this simulation. For the results of Causal Forest and LASSO, we utilize 1,000 trials as before. As seen in the sample-splitting case, the estimated biases of the proposed GATES estimators are relatively small even when $n = 100$ and becomes negligible when $n = 500$.

Recall that under the 5-fold cross-fitting, for example, $n = 500$ implies the evaluation sample of size 100 for each fold. And, yet, combining the five folds leads to a much lower standard deviation than the sample-splitting case with the $n = 100$ case in Table 1. The results are similar when comparing the $n = 2500$ cross-fitting case with the $n = 500$ sample-splitting case. Indeed, in some cases, the reduction in standard deviation is more than 50 percent. This experimentally demonstrates the efficiency gain from using a cross-fitting approach. We further find that although Theorem 5 implies that the proposed variance estimate is conservative, the results show only the slight overcoverage of the confidence intervals. In Imai and Li (2023a) we show that the methodology proposed in Chernozhukov *et al.* (2023) leads to more significant overcoverage of the confidence intervals.

4.3 Finite-Sample Performance of the Proposed Hypothesis Tests

We next examine the finite sample performance of the proposed hypothesis tests. Due to the aforementioned computational intensity of BART, we focus on Causal Forest and LASSO. For each simulated data set, we conduct hypothesis tests of two null hypotheses of interest: treatment effect homogeneity (see Equations (4) and (11) for sample-splitting and cross-fitting, respectively) and rank-consistency of the GATES (see Equations (5) and (12) for sample-splitting and cross-fitting cases, respectively).

According to the true values shown in Tables 1 and 2, the null hypothesis of treatment effect homogeneity is false while the rank-consistency null hypothesis is correct. This suggests that the proposed test should reject the former hypothesis more frequently as the sample size increases whereas it should reject the latter hypothesis no more frequently than the specified size of the test, which we set to 5% throughout.

We first consider the sample-splitting setting based on 500 simulation trials. Table 3 presents the rejection rate and median p-value for each scenario across different training and testing data sizes, denoted by n_{train} and n_{test}, respectively. We find that for Causal Forest, the training data of size 400 and the test data of size 2000 are required to reject the null hypothesis of treatment effect homogeneity with a high probability. This highlights the difficulty of identifying treatment effect heterogeneity in randomized experiments. For the hypothesis test of the rank-consistency of GATES, we find that if trained with a small sample ($n_{\text{train}} = 100$), Causal Forest might falsely reject the null hypothesis but this false rejection rate is less than the size of the test regardless of the size of the test data. This reflects the conservative

	$n_{\text{test}} = 100$		$n_{\text{test}} = 500$		$n_{\text{test}} = 2500$	
	rejection rate	median p-value	rejection rate	median p-value	rejection rate	median p-value
Causal Forest						
H_0: *Treatment effect homogeneity*						
$n_{\text{train}} = 100$	5.2%	0.504	7.4%	0.529	19.6%	0.361
$n_{\text{train}} = 400$	9.0	0.459	22.0	0.254	74.4	0.002
$n_{\text{train}} = 2000$	13.0	0.367	40.4	0.092	96.0	0.000
H_0^*: *Rank consistency of GATES*						
$n_{\text{train}} = 100$	4.0%	0.583	2.2%	0.624	2.2%	0.704
$n_{\text{train}} = 400$	2.8	0.687	0.2	0.820	0.2	0.907
$n_{\text{train}} = 2000$	1.2	0.707	0.2	0.852	0.0	0.967
LASSO						
H_0: *Treatment effect homogeneity*						
$n_{\text{train}} = 100$	5.8%	0.496	5.2%	0.544	9.6%	0.516
$n_{\text{train}} = 400$	7.0	0.557	4.0	0.578	10.4	0.468
$n_{\text{train}} = 2000$	6.2	0.489	9.4	0.519	26.2	0.249
H_0^*: *Rank consistency of GATES*						
$n_{\text{train}} = 100$	4.6%	0.525	3.0%	0.584	5.4%	0.596
$n_{\text{train}} = 400$	6.0	0.494	1.8	0.600	2.4	0.687
$n_{\text{train}} = 2000$	3.2	0.608	1.4	0.698	1.2	0.851

Table 3: The Finite Sample Performance of the Hypothesis Tests for Treatment Effect Homogeneity and Rank-consistency of GATES under Sample-splitting. The results are based on Causal Forest and LASSO. The table presents the percent of 500 simulation trials where each null hypothesis is rejected using the 5% test size. In addition, the median p-value across all trials is also shown. The results are presented for different training data sizes n_{train} and different test data sizes n_{test}.

nature of our test as discussed at the end of Section 2.

We obtain similar findings for LASSO, where small training data leads to low rejection rates for the treatment effect homogeneity hypothesis and some false rejection of the rank consistency hypothesis. As before, the false rejection rates are approximately 5% or lower (the small number of simulations induce some noise). Interestingly, the proposed test is much less powerful for LASSO than for Causal Forest. Even when the size of training data is 2,000 and the test data size is 2,500, the rejection rate is only slightly above 25%. This is consistent with the finding in Section 4.2 that LASSO discovers less treatment effect heterogeneity than Causal Forest.

We also examine the performance of our hypothesis tests under cross-fitting, again

	$n = 100$		$n = 500$		$n = 2500$	
	rejection rate	median p-value	rejection rate	median p-value	rejection rate	median p-value
Causal Forest						
Homogeneous Treatment Effects	1.4%	0.790	4.6%	0.712	51.4%	0.041
Consistent Treatment Effects	1.4%	0.702	0.8%	0.845	0.0%	0.976
LASSO						
Homogeneous Treatment Effects	0.6%	0.880	1.8%	0.850	9.0%	0.664
Consistent Treatment Effects	1.0%	0.722	0.6%	0.769	0.2%	0.889

Table 4: The Finite Sample Performance of the Hypothesis Tests for Treatment Effect Homogeneity and Rank-consistency of GATES under Cross-fitting. The results are based on Causal Forest and LASSO. The table presents the percent of 500 simulation trials where each null hypothesis is rejected using the 5% test size and also the median p-value across all trials.

using 500 simulation trials. Table 4 presents the rejection rate and median p-value across different sample sizes. We use $L = 5$ fold cross-fitting for all simulations. Note that the $n = 500$ case under cross-fitting is analogous in the size of training and testing data to the ($n_{\text{train}} = 400, n_{\text{test}} = 100$) case for sample splitting. Similarly, the $n = 2500$ case under cross-fitting corresponds to the ($n_{\text{train}} = 2,000, n_{\text{test}} = 500$) case under sample-splitting.

For both Causal Forest and LASSO, the rejection rate of the homogeneous treatment effect hypothesis is lower in the $n = 500$ case compared with the corresponding sample-splitting case, reflecting the additional uncertainty due to the sampling of training data (under sample-splitting, the scoring rule is regarded as fixed). However, when the sample size is $n = 2,500$, for both algorithms the rejection rate of homogeneous treatment effects is higher under cross-fitting than sample-splitting, demonstrating that the efficiency gain of cross-fitting outweigh its additional sampling uncertainty. For the hypothesis test of rank-consistency, we find that the rejection rate under cross-fitting is significantly lower than the nominal test size for all cases.

5 An Empirical Application

To demonstrate the applicability of the proposed framework, we utilize the experimental data from the male sub-sample of the National Supported Work Demonstration (NSW) (LaLonde, 1986; Dehejia and Wahba, 1999). NSW was a temporary employment program to help disadvantaged workers by providing them with work experience and counseling in a sheltered environment. Specifically, qualified applicants were randomly assigned to the treatment and control groups, where the workers in the treatment group were given a guaranteed job for 9 to 18 months. The primary outcome of interest is the annualized earnings in 1978, 36 months after the program. The data contains a total of $n = 722$ observations, with $n_1 = 297$ participants assigned to the treatment group and $n_0 = 425$ participants in the control group. There are 7 available pre-treatment covariates $\boldsymbol{X}$ that records the demographics and pre-treatment earnings of the participants.

We evaluate Causal Forest, BART, and LASSO under the two settings considered in this paper. For sample-splitting, we randomly select 67% of the data (484 observations) to serve as a training dataset. We use the remaining 238 samples to estimate the GATES and conduct the proposed hypothesis tests. For cross-fitting, we first randomly split the data into 3 folds, i.e., $L = 3$. We use each fold once as a testing set, while the remaining two folds are the training set. The number of trees was tuned through 5-fold cross-validation for BART and Causal Forest within each training dataset. The regularization parameter was tuned similarly for LASSO.

We focus on the quintile GATES ($K = 5$). Table 5 presents the results (reported in 1,000 US dollars) under the sample-splitting and cross-fitting settings. We find that Causal Forest is able to produce statistically significantly positive GATES for the highest quintile group ($\hat{\tau}_5$) under both sample-splitting and cross-fitting. Thus, unlike the other two algorithms, Causal Forest can identify a 20% subset that benefits

	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$	$\hat{\tau}_4$	$\hat{\tau}_5$
Sample-splitting					
Causal Forest	3.40	0.13	−0.85	−1.91	7.21
	[−1.29, 3.40]	[−5.37, 5.63]	[−5.22, 3.52]	[−5.16, 1.34]	[1.22, 13.19]
BART	2.90	−0.73	−0.02	3.25	2.57
	[−2.25, 8.06]	[−5.05, 3.58]	[−3.47, 3.43]	[−1.53, 8.03]	[−3.82, 8.97]
LASSO	1.86	2.62	−2.07	1.39	4.17
	[−3.59, 7.30]	[−1.69, 6.93]	[−5.39, 1.26]	[−2.95, 5.73]	[−2.30, 10.65]
Cross-fitting					
Causal Forest	−3.72	1.05	5.32	−2.64	4.55
	[−6.52, −0.93]	[−2.28, 4.37]	[2.63, 8.01]	[−5.07, −0.22]	[1.14, 7.96]
BART	0.40	−0.15	−0.40	2.52	2.19
	[−3.79, 4.59]	[−2.54, 2.23]	[−3.37, 2.56]	[−0.99, 6.03]	[−0.73, 5.11]
LASSO	0.65	0.45	−2.88	1.32	5.02
	[−3.65, 4.94]	[−3.28, 4.18]	[−5.38, −0.38]	[−1.83, 4.48]	[−0.14, 10.18]

Table 5: The Estimated GATES and their 95% Confidence Intervals based on Causal Forest, BART, and LASSO under Sample-splitting and Cross-fitting. The estimated GATES based on quintiles are reported in 1,000 US dollars. Sample-splitting is done using 67% of the sample as the training data and 33% of the sample as the evaluation data. For cross-fitting, we use 3 folds of equal size.

significantly from the temporary employment program.

Two additional observations are worth noting. First, the confidence intervals are generally narrower in the cross-fitting case compared to the sample-splitting case. This finding is consistent with the fact that cross-fitting is more efficient than sample-splitting. Second, the three algorithms failed to produce any statistically significant positive GATES for the remaining groups. This may be because there are few additional workers who benefit from the program. Alternatively, it is also possible that such workers exist but the algorithms are unable to identify them.

To formally evaluate the statistical significance of several GATES estimates, we must account for the potential multiple testing problem. Thus, we apply the proposed hypothesis tests to evaluate the null hypotheses of treatment effect homogeneity and rank-consistency of the GATES. Table 6 presents the resulting values of test statistics and p-values. We find that under sample-splitting, only Causal Forest is able to reject the null hypothesis of treatment effect homogeneity at the 10% level. However, under

	Causal Forest		BART		LASSO	
	stat	p-value	stat	p-value	stat	p-value
Sample-splitting						
Homogeneous Treatment Effects	9.78	0.082	2.76	0.737	5.26	0.362
Rank-consistent Treatment Effects	3.07	0.323	1.13	0.657	3.14	0.302
Cross-fitting						
Homogeneous Treatment Effects	30.29	0.000	2.32	0.803	10.79	0.056
Rank-consistent Treatment Effects	0.06	0.691	0.04	0.885	0.45	0.711

Table 6: The Results of the Proposed Hypothesis Tests under Sample-splitting and Cross-fitting Using Causal Forest, BART, and LASSO. The values of test statistics and p-values are presented. We test the null hypotheses of treatment effect homogeneity and rank-consistency of the GATES.

cross-fitting, both Causal Forest and LASSO algorithms can reject the null hypothesis at the 10% level, with Causal Forest being able to reject the hypothesis at even the 0.1% level. In contrast, BART fails to reject the treatment effect homogeneity hypothesis under both sample-splitting and cross-fitting. The results with Causal Forest suggest that the identification of a statistically significant GATES estimate for one subgroup under cross-fitting is able to grant enough power to reject the null hypothesis that the average treatment effects are homogeneous across all subgroups. Finally, we find that all three algorithms fail to reject the null hypothesis of the rank-consistency of GATEs. Thus, under our conservative tests, there is no strong statistical evidence that these algorithms are producing unreliable GATES.

6 Concluding Remarks

Many randomized experiments have a limited sample size and the resulting treatment effect estimates are often small and noisy. Yet, applied researchers often use machine learning algorithms to estimate heterogeneous treatment effects. Therefore, it is important to statistically distinguish signal from noise. We have developed the framework that does not impose a strong assumption on machine learning algorithms and hence is applicable to a wide range of situations. The proposed methodology allows researchers to construct confidence intervals on the estimated average treatment effects within a group identified by any machine learning algorithm. We also show how to conduct formal hypothesis tests about heterogeneous treatment effects. Our method solely relies upon the randomization of treatment assignment and the random sampling of units, and hence yields reliable statistical inference even when the sample size is relatively small and machine learning algorithms are not performing well.

References

Andrews, D. W. and Guggenberger, P. (2009). Validity of subsampling and "plug-in asymptotic" inference for parameters defined by moment inequalities. *Econometric Theory* **25**, 3, 669–709.

Andrews, D. W. and Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica* **78**, 1, 119–157.

Austern, M. and Zhou, W. (2020). Asymptotics of cross-validation. *arXiv preprint arXiv:2001.11111* .

Bhattacharya, P. K. (1974). Convergence of sample paths of normalized sums of induced order statistics. *The Annals of Statistics* 1034–1039.

Canay, I., Illanes, G., and Velez, A. (2023). A user's guide to inference in models defined by moment inequalities. Tech. rep., National Bureau of Economic Research.

Canay, I. A. (2010). El inference for partially identified models: Large deviations optimality and bootstrap validity. *Journal of Econometrics* **156**, 2, 408–425.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2019). Inference on causal and structural parameters using many moment inequalities. *The Review of Economic Studies* **86**, 5, 1867–1900.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernandez-Val, I. (2023). Fisher-schultz lecture: Generic machine learning inference on heterogeneous treatment effects in randomized experiments. Tech. rep., arXiv:1712.04802.

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

Cramér, H. and Wold, H. (1936). Some theorems on distribution functions. *Journal of the London Mathematical Society* **1**, 4, 290–294.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics* **90**, 3, 389–405.

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

Ding, P., Feller, A., and Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **78**, 3, 655–671.

Ding, P., Feller, A., and Miratrix, L. (2019). Decomposing treatment effect variation. *Journal of the American Statistical Association* **114**, 525, 304–317.

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science* **34**, 1, 43–68.

Dwivedi, R., Tan, Y. S., Park, B., Wei, M., Horgan, K., Madigan, D., and Yu, B. (2020). Stable discovery of interpretable subgroups via calibration in causal studies. *International Statistical Review* **88**, S135–S178.

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

Higham, N. J. (2002). Computing the nearest correlation matrix—a problem from finance. *IMA journal of Numerical Analysis* **22**, 3, 329–343.

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

Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association* **81**, 945–960.

Imai, K. and Li, M. L. (2023a). Comment on "generic machine learning inference on heterogeneous treatment effects in randomized experiments.". Tech. rep., Working Paper.

Imai, K. and Li, M. L. (2023b). Experimental evaluation of individualized treatment rules. *Journal of the American Statistical Association* **118**, 541, 242–256.

Imai, K. and Li, M. L. (2023c). Statistical performance guarantee for selecting those predicted to benefit most from treatment. Tech. rep., arXiv preprint 2310.07973.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2018). Meta-learners for estimating heterogeneous treatment effects using machine learning. Tech. rep., arXiv:1706.03461.

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

Nadeau, C. and Bengio, Y. (2000). Inference for the generalization error. In *Advances in neural information processing systems*, 307–313.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. (translated in 1990). *Statistical Science* **5**, 465–480.

Rubin, D. B. (1990). Comments on "On the application of probability theory to agricultural experiments. Essay on principles. Section 9" by J. Splawa-Neyman translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. *Statistical Science* **5**, 472–480.

Shapiro, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. *International Statistical Review/Revue Internationale de Statistique* **56**, 1, 49–62.

Shorack, G. R. (1972). Functions of order statistics. *The Annals of Mathematical Statistics* **43**, 2, 412–427.

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

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

Wellner, J. A. (1977). A glivenko-cantelli theorem and strong laws of large numbers for functions of order statistics. *The Annals of Statistics* **5**, 3, 473–480.

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., and Wager, S. (2021). Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv preprint arXiv:2111.07966* .

Supplementary Appendix

S1 Proof of Theorem 1

We first rewrite the expectation of the proposed estimator in Equation (3) as,

$$\mathbb{E}(\hat{\tau}_k) \;=\; K\mathbb{E}\left\{Y_i\left(f^*(\boldsymbol{X}_i, \hat{c}_k(s))\right) - Y_i\left(f^*(\boldsymbol{X}_i, \hat{c}_{k-1}(s))\right)\right\},$$

where $f^*(\boldsymbol{X}_i, c) = \mathbf{1}\{s(\boldsymbol{X}_i) < c\}$. Similarly, we can also write the estimand in Equation (2) as,

$$\tau_k \;=\; K\mathbb{E}\left\{Y_i\left(f^*(\boldsymbol{X}_i, c_k(s))\right) - Y_i\left(f^*(\boldsymbol{X}_i, c_{k-1}(s))\right)\right\}.$$

Now, define $F(c) = \mathbb{P}(s(\boldsymbol{X}_i) \leq c)$. Without loss of generality, assume $\hat{c}_k(s) > c_k(s)$ and $\hat{c}_{k-1}(s) > c_{k-1}(s)$. If this is not the case, we simply switch the upper and lower limits of the integrals in the proof below. Then, the bias of the estimator is given by,

$$\begin{aligned}
&\frac{|\mathbb{E}(\hat{\tau}_k) - \tau_k|}{K} \\
\leq\;& \left|\mathbb{E}\left\{Y_i\left(f^*(\boldsymbol{X}_i, \hat{c}_k(s))\right) - Y_i\left(f(\boldsymbol{X}_i, c_k(s))\right)\right\}\right| + \left|\mathbb{E}\left\{Y_i\left(f^*(\boldsymbol{X}_i, \hat{c}_{k-1}(s))\right) - Y_i\left(f^*(\boldsymbol{X}_i, c_{k-1}(s))\right)\right\}\right| \\
=\;& \left|\mathbb{E}_{\hat{c}_k(s)}\left[\int_{c_k(s)}^{\hat{c}_k(s)} \mathbb{E}(\tau_i \mid s(\boldsymbol{X}_i) = c)\mathrm{d}F(c)\right]\right| + \left|\mathbb{E}_{\hat{c}_{k-1}(s)}\left[\int_{c_{k-1}(s)}^{\hat{c}_{k-1}(s)} \mathbb{E}(\tau_i \mid s(\boldsymbol{X}_i) = c)\mathrm{d}F(c)\right]\right| \\
=\;& \left|\mathbb{E}_{F(\hat{c}_k(s))}\left[\int_{F(c_k(s))}^{F(\hat{c}_k(s))} \mathbb{E}(\tau_i \mid s(\boldsymbol{X}_i) = F^{-1}(x))\mathrm{d}x\right]\right| \\
&+ \left|\mathbb{E}_{F(\hat{c}_{k-1}(s))}\left[\int_{F(c_{k-1}(s))}^{F(\hat{c}_{k-1}(s))} \mathbb{E}(\tau_i \mid s(\boldsymbol{X}_i) = F^{-1}(x))\mathrm{d}x\right]\right| \\
\leq\;& \mathbb{E}_{F(\hat{c}_k(s))}\left[\left|F(\hat{c}_k(s)) - \frac{k}{K}\right| \times \max_{c \in [c_k(s), \hat{c}_k(s)]} |\mathbb{E}(\tau_i \mid s(\boldsymbol{X}_i) = c)|\right] \\
&+ \mathbb{E}_{F(\hat{c}_{k-1}(s))}\left[\left|F(\hat{c}_{k-1}(s)) - \frac{k-1}{K}\right| \times \max_{c \in [c_{k-1}(s), \hat{c}_{k-1}(s)]} |\mathbb{E}(\tau_i \mid s(\boldsymbol{X}_i) = c)|\right]
\end{aligned}$$

By the definition of $\hat{c}_k(s)$, $F(\hat{c}_k(s))$ is the nk/Kth order statistic of n independent uniform random variables, and thus follows the Beta distribution with the shape and scale parameters equal to nk/K and $n - nk/K + 1$, respectively. For the special case where $k - 1 = 0$, we define the 0th order statistic of n uniform random variables to be 0, and by extension also define the "beta distribution" with shape parameter ≤ 0 to be $H(x)$ where $H(x)$ is the Heaviside step function. Therefore, we have,

$$\mathbb{P}\left(|F(\hat{c}_k(s)) - \frac{k}{K}| > \epsilon\right) \;=\; 1 - B\left(\frac{k}{K} + \epsilon, \frac{nk}{K}, n - \frac{nk}{K} + 1\right) + B\left(\frac{k}{K} - \epsilon, \frac{nk}{K}, n - \frac{nk}{K} + 1\right),$$

where $B(\epsilon, \alpha, \beta) = \int_0^\epsilon t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t$ is the incomplete beta function. Similarly, we have

$$\begin{aligned}
\mathbb{P}(|F(\hat{c}_{k-1}(s)) - \frac{k-1}{K}| > \epsilon) \;=\;& 1 - B\left(\frac{k-1}{K} + \epsilon, \frac{n(k-1)}{K}, n - \frac{n(k-1)}{K} + 1\right) \\
&+ B\left(\frac{(k-1)}{K} - \epsilon, \frac{n(k-1)}{K}, n - \frac{n(k-1)}{K} + 1\right).
\end{aligned}$$

Combining the above results yields the desired bias bound expression.

To derive the exact variance, we first apply the law of total variance to Equation (3),

$$\begin{aligned}
\mathbb{V}(\hat{\tau}_k) &= \mathbb{V}\left[\mathbb{E}\left\{K\left(\frac{1}{n_1}\sum_{i=1}^n \hat{f}_k(\boldsymbol{X}_i)T_iY_i(1) - \frac{1}{n_0}\sum_{i=1}^n \hat{f}_k(\boldsymbol{X}_i)(1-T_i)Y_i(0)\right) \,\middle|\, \boldsymbol{X}, \{Y_i(1), Y_i(0)\}_{i=1}^n\right\}\right] \\
&\quad + \mathbb{E}\left[\mathbb{V}\left\{K\left(\frac{1}{n_1}\sum_{i=1}^n \hat{f}_k(\boldsymbol{X}_i)T_iY_i(1) - \frac{1}{n_0}\sum_{i=1}^n \hat{f}_k(\boldsymbol{X}_i)(1-T_i)Y_i(0)\right) \,\middle|\, \boldsymbol{X}, \{Y_i(1), Y_i(0)\}_{i=1}^n\right\}\right] \\
&= K^2\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^n \{Y_{ki}(1) - Y_{ki}(0)\}\right) \\
&\quad + K^2\mathbb{E}\left[\mathbb{V}\left\{\frac{1}{n_1}\sum_{i=1}^n \hat{f}_k(\boldsymbol{X}_i)T_iY_i(1) - \frac{1}{n_0}\sum_{i=1}^n \hat{f}_k(\boldsymbol{X}_i)(1-T_i)Y_i(0) \,\middle|\, \boldsymbol{X}, \{Y_i(1), Y_i(0)\}_{i=1}^n\right\}\right].
\end{aligned} \tag{S1}$$

Applying the standard result from Neyman's finite sample variance analysis to the second term shows that this term is equal to,

$$K^2\mathbb{E}\left\{\frac{1}{n}\left(\frac{n_0}{n_1}S_{k1}^2 + \frac{n_1}{n_0}S_{k0}^2 + 2S_{k01}\right)\right\}. \tag{S2}$$

where $S_{k01} = \sum_{i=1}^n (Y_{ki}(0) - \overline{Y_k(0)})(Y_{ki}(1) - \overline{Y_k(1)})/(n-1)$. Since $Y_{ki}(t)$ and $Y_{kj}(t)$ are correlated, we apply Lemma 1 of Nadeau and Bengio (2000) to the first term, yielding,

$$\begin{aligned}
&\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^n \{Y_{ki}(1) - Y_{ki}(0)\}\right) \\
&= \mathrm{Cov}(Y_{ki}(1) - Y_{ki}(0), Y_{kj}(1) - Y_{kj}(0)) + \frac{1}{n}\mathbb{E}(S_{k1}^2 + S_{k0}^2 - 2S_{k01}),
\end{aligned} \tag{S3}$$

for $i \neq j$ where

$$\begin{aligned}
&\mathrm{Cov}(Y_{ki}(1) - Y_{ki}(0), Y_{kj}(1) - Y_{kj}(0)) \\
&= \mathrm{Cov}\left(\hat{f}_k(\boldsymbol{X}_i)\tau_i, \hat{f}_k(\boldsymbol{X}_j)\tau_j\right) \\
&= \Pr(\hat{f}_k(\boldsymbol{X}_i) = \hat{f}_k(\boldsymbol{X}_j) = 1)\mathbb{E}[\tau_i\tau_j \mid \hat{f}_k(\boldsymbol{X}_i) = \hat{f}_k(\boldsymbol{X}_j) = 1] - \Pr(\hat{f}_k(\boldsymbol{X}_i) = 1)^2\mathbb{E}[\tau_i \mid \hat{f}_k(\boldsymbol{X}_i) = 1]^2 \\
&= \frac{n-K}{K^2(n-1)}\mathbb{E}[\tau_i\tau_j \mid \hat{f}_k(\boldsymbol{X}_i) = \hat{f}_k(\boldsymbol{X}_j) = 1] - \frac{1}{K^2}\mathbb{E}[\tau_i \mid \hat{f}_k(\boldsymbol{X}_i) = 1]^2 \\
&= \frac{(n-K)\kappa_{k11}}{K^2(n-1)} - \frac{\kappa_{k1}^2}{K^2}
\end{aligned}$$

Substituting Equations (S2) and (S3) into Equation S1, we obtain the desired variance expression. □

S2 Derivation of $\hat{\kappa}_{ktt}$

We first rewrite κ_{ktt} as:

$$\kappa_{ktt} = \sum_{u,v\in\{0,1\}} (-1)^{u+v}\mathbb{E}[Y_i(u)Y_j(v) \mid \hat{f}_k(\boldsymbol{X}_i) = \hat{f}_k(\boldsymbol{X}_j) = t].$$

We can estimate each conditional expectation term inside of the summation using its sample analogue:

$$\frac{\sum_{i=1}^{n}\sum_{j\neq i}\mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i)=\hat{f}_k(\boldsymbol{X}_j)=t\}\{1-u+(2u-1)T_i\}\{1-v+(2v-1)T_j\}Y_iY_j}{\sum_{i=1}^{n}\sum_{j\neq i}\mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i)=\hat{f}_k(\boldsymbol{X}_j)=t\}\{1-u+(2u-1)T_i\}\{1-v+(2v-1)T_j\}}$$

We can further simplify the computation by rewriting the numerator as:

$$\left[\sum_{i=1}^{n}\mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i)=t\}\{1-u+(2u-1)T_i\}Y_i\right]\left[\sum_{i=1}^{n}\mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i)=t\}\{1-v+(2v-1)T_i\}Y_i\right]$$
$$-\sum_{i=1}^{n}\mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i)=t\}\{1-u+(2u-1)T_i\}\{1-v+(2v-1)T_i\}Y_i^2.$$

Similarly, we can rewrite the denominator as follows:

$$\left[\sum_{i=1}^{n}\mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i)=t\}(1-u+(2u-1)T_i)\right]\left[\sum_{i=1}^{n}\mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i)=t\}\{1-v+(2v-1)T_i\}\right]$$
$$-\sum_{i=1}^{n}\mathbf{1}\{\hat{f}_k(\boldsymbol{X}_i)=t\}\{1-u+(2u-1)T_i\}\{1-v+(2v-1)T_i\}.$$

Putting these terms together, we obtain the expression of $\hat{\kappa}_{ktt}$ given in Section 2.2.

S3 Proof of Theorem 2

Given a tuple of n samples $\{Y_i, T_i, \boldsymbol{X}_i\}_{i=1}^n$, we first reorder the sample to $(Y_{[i,n]}, T_{[i,n]}, \boldsymbol{X}_{[i,n]})$ based on the magnitude of the scoring rule, such that

$$s(\boldsymbol{X}_{[1,n]}) \leq s(\boldsymbol{X}_{[2,n]}) \leq \cdots \leq s(\boldsymbol{X}_{[n,n]})$$

Then, the proposed GATES estimator can be rewritten as

$$\hat{\tau}_k = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{\frac{(k-1)n}{K} < i \leq \frac{kn}{K}\right\}U_{[i,n]} \tag{S4}$$

where

$$U_{[i,n]} := KY_{[i,n]}\left(\frac{T_{[i,n]}}{q} - \frac{1-T_{[i,n]}}{1-q}\right), \tag{S5}$$

where $q = n_1/n$. Now, we prove the following two lemmas.

Lemma S1 *Let $(X_1, Y_1), (X_2, Y_2), \cdots$ be a sequence of random vectors. For each $n \geq 1$, $(X_1, Y_1), \cdots, (X_n, Y_n)$ possesses a joint distribution. Let $\boldsymbol{Z}_n = ((X_1, Y_1), \cdots, (X_n, Y_n))$ and $\boldsymbol{X}_n = (X_1, \cdots, X_n)$, and let $W_n(\boldsymbol{Z}_n)$ and $S_n(\boldsymbol{X}_n)$ be measurable vector-valued functions of $\boldsymbol{Z}_n$ and $\boldsymbol{X}_n$ respectively. Suppose $S_n(\boldsymbol{X}_n)$ converges in distribution to F_S and the conditional distribution $W_n(\boldsymbol{Z}_n) \mid \boldsymbol{X}_n$ converges in distribution to F_W in probability, where F_W does not depend on $\boldsymbol{X}_n$. Then, we have that:*

$$(W_n(\boldsymbol{Z}_n), S_n(\boldsymbol{X}_n)) \to F_W F_S$$

Proof The characteristic function of the joint distribution of $(W_n(\boldsymbol{Z}_n), S_n(\boldsymbol{X}_n))$ can be written as:

$$\varphi_{W_nS_n}(t_1, t_2) = \mathbb{E}[\exp\{i(t_1 W_n(\boldsymbol{Z}_n) + t_2 S_n(\boldsymbol{X}_n))\}]$$

Let $W \sim F_W$ and $S \sim F_S$. Then the characteristic function of $F_W F_S$ can be written as:

$$\varphi_{WS}(t_1, t_2) = \mathbb{E}[\exp\{i(t_1 W)\}]\mathbb{E}[\exp\{i(t_2 S)\}]$$

We then have:

$$\begin{aligned}
&|\varphi_{W_nS_n}(t_1, t_2) - \varphi_{WS}(t_1, t_2)| \\
=& |\mathbb{E}[\exp\{i(t_1 W_n(\boldsymbol{Z}_n) + t_2 S_n(\boldsymbol{X}_n))\}] - \mathbb{E}[\exp\{i(t_1 W)\}]\mathbb{E}[\exp\{i(t_2 S)\}]| \\
\leq& |\mathbb{E}[\mathbb{E}[\exp\{i(t_1 W_n(\boldsymbol{Z}_n) + t_2 S_n(\boldsymbol{X}_n))\} \mid \boldsymbol{X}_n]] - \mathbb{E}[\exp\{i(t_1 W)\}]\mathbb{E}[\exp\{i(t_2 S_n(\boldsymbol{X}_n))\}]| \\
&+ |\mathbb{E}[\exp\{i(t_1 W)\}]\mathbb{E}[\exp\{i(t_2 S_n(\boldsymbol{X}_n))\}] - \mathbb{E}[\exp\{i(t_1 W)\}]\mathbb{E}[\exp\{i(t_2 S)\}]| \\
\leq& \mathbb{E}[|\mathbb{E}[\exp(i t_1 W_n(\boldsymbol{Z}_n)) \mid \boldsymbol{X}_n] - \mathbb{E}[\exp(i t_1 W)]|] + |\mathbb{E}[\exp\{i(t_2 S_n(\boldsymbol{X}_n))\}] - \mathbb{E}[\exp\{i(t_2 S)\}]|,
\end{aligned}$$

where the last inequality follows from the fact that all characteristic functions satisfy $|\varphi| \leq 1$. This expression converges to zero in probability due to the convergence of $S_n(\boldsymbol{X}_n)$ and $W_n(\boldsymbol{Z}_n) \mid \boldsymbol{X}_n$ respectively. Therefore, we have:

$$(W_n(\boldsymbol{Z}_n), S_n(\boldsymbol{X}_n)) \to F_W F_S$$

□

Lemma S2 $\lim\limits_{n\to\infty} \mathbb{E}(\hat{\tau}_k) - \tau_k = O\left(n^{-1}\right)$

Proof We bound the bias of $\mathbb{E}(\hat{\tau}_k)$ by appealing to Theorem 1 of Imai and Li (2023b), which implies,

$$\begin{aligned}
|\mathbb{E}(\hat{\tau}_k) - \tau_k| \leq & \left| K\mathbb{E}\left[\int_{F(c_k(s))}^{F(\hat{c}_k(s))} \mathbb{E}(Y_i(1) - Y_i(0) \mid s(\boldsymbol{X}_i) = F^{-1}(x))\mathrm{d}x\right]\right| \\
& + \left| K\mathbb{E}\left[\int_{F(c_{k-1}(s))}^{F(\hat{c}_{k-1}(s))} \mathbb{E}(Y_i(1) - Y_i(0) \mid s(\boldsymbol{X}_i) = F^{-1}(x))\mathrm{d}x\right]\right|. \qquad \text{(S6)}
\end{aligned}$$

By the definition of $\hat{c}_k(s)$, $F(\hat{c}_k(s))$ is the nk/Kth order statistic of n independent uniform random variables, and therefore, follows the Beta distribution with the shape and scale parameters equal to nk/K and $n - nk/K + 1$, respectively.

Now, by Assumption 4, we can compute the first-order Taylor expansion of $\int_a^x \mathbb{E}(Y_i(1) - Y_i(0) \mid s(\boldsymbol{X}_i) = F^{-1}(x))\mathrm{d}x$:

$$\begin{aligned}
|\mathbb{E}(\hat{\tau}_k) - \tau_k| \leq & |K\mathbb{E}\left[a_0\{F(\hat{c}_k(s)) - F(c_k(s))\} + o(F(\hat{c}_k(s)) - F(c_k(s)))\right]| \\
& + |K\mathbb{E}\left[a_1\{F(\hat{c}_{k-1}(s)) - F(c_{k-1}(s))\} + o(F(\hat{c}_{k-1}(s)) - F(c_{k-1}(s)))\right]| \\
=& |Ka_0| \left|\frac{nk}{K(n+1)} - \frac{k}{K}\right| + |Ka_1| \left|\frac{n(k-1)}{K(n+1)} - \frac{k-1}{K}\right| + o\left(n^{-1}\right) \\
=& O\left(n^{-1}\right).
\end{aligned}$$

□

Now, using these two lemmas, we prove the main result. Letting $u(s) = \mathbb{E}[U_i \mid s(\boldsymbol{X}_i) = s]$, we decompose $\hat{\tau}_k$ into two parts,

$$\hat{\tau}_k = \underbrace{\frac{1}{n} \sum_{\frac{(k-1)n}{K} < i \leq \frac{kn}{K}} U_{[i,n]} - u(s(\boldsymbol{X}_{[i,n]}))}_{\hat{\tau}_k^{(1)}} + \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left\{ \frac{(k-1)n}{K} < i \leq \frac{kn}{K} \right\} u(s(\boldsymbol{X}_{[i,n]}))}_{\hat{\tau}_k^{(2)}}. \tag{S7}$$

Consider the first term. By the general theory of induced order statistics presented in Bhattacharya (1974), $U_{[i,n]} - u(s(\boldsymbol{X}_{[i,n]}))$ for $i = 1, \cdots, n$ are independent of one another conditional on $\boldsymbol{X}_n = (\boldsymbol{X}_{[1,n]}, \cdots, \boldsymbol{X}_{[n,n]})$. Define the random variables $Z_{[i,n]}$ as distributed according to the joint conditional distribution $U_{[i,n]} - u(s(\boldsymbol{X}_{[i,n]})) \mid \boldsymbol{X}_n$. Then, we have

$$\hat{\tau}_k^{(1)} = \frac{1}{n} \sum_{\frac{(k-1)n}{K} < i \leq \frac{kn}{K}} Z_{[i,n]},$$

where $Z_{[i,n]}$ are conditionally independent and $\mathbb{E}[Z_{[i,n]}] = 0$ by construction. Therefore, by Assumption 5, we can utilize the Berry-Esseen Theorem. Define:

$$\sigma_1^2(n) = \frac{1}{n} \sum_{\frac{(k-1)n}{K} < i \leq \frac{kn}{K}} \mathbb{V}(Z_{[i,n]})$$

$$\rho_1(n) = \frac{1}{n} \sum_{\frac{(k-1)n}{K} < i \leq \frac{kn}{K}} \mathbb{E}(|Z_{[i,n]}|^3)$$

Then the Berry-Esseen Theorem states that for $W \sim N(0,1)$, we have:

$$d\left(\frac{\sqrt{n}\hat{\tau}_k^{(1)}}{\sqrt{\sigma_1^2(n)}}, W\right) \leq \frac{C_0}{\sqrt{n}} \left(\sigma_1^2(n)\right)^{-3/2} \rho_1(n)$$

where $d(\cdot,\cdot)$ is the Kolmogorov distance. Now define the asymptotic variance and third moment by:

$$\sigma_1^2 = \lim_{n\to\infty} \sigma_1^2(n)$$

$$\rho_1 = \lim_{n\to\infty} \rho_1(n)$$

Both quantities exist by the strong law of large numbers for functions of order statistics (see Theorem 4 of Wellner (1977)). Specifically, by the strong law, σ_1^2 and ρ_1 does not depend on $\boldsymbol{X}_n$ for all but at most a measure zero set of $\boldsymbol{X}_n$. Therefore, the Berry-Esseen theorem implies that:

$$\sqrt{n}\hat{\tau}_k^{(1)} \mid \boldsymbol{X}_n \xrightarrow{d} N(0, \sigma_1^2) \text{ with probability 1} \tag{S8}$$

Next, consider the second term of Equation (S7). To prove the convergence of this summation of a function of order statistics, we utilize Theorem 1 and Example 1 from Shorack (1972), which we restate in our notation below:

THEOREM S1 (SHORACK (1972)) *Consider an independently and identically distributed random sample $X_1, \cdots, X_n$ of size n from a cumulative distribution function F, and a function of bounded variation g such that $\mathbb{E}[g(X)^3] < \infty$. Define:*

$$T_n = \frac{1}{n} \sum_{i=1}^{n} J\left(\frac{i}{n}\right) g(X_{[i,n]})$$

where $X_{[i,n]}$ is the ith order statistics of the sample, and J is a function that is continuous except at a finite number of points at which $g(F^{-1})$ is continuous. Suppose that there exists $\delta > 0$ such that:

$$|J(t)| \leq M(t(1-t))^{-\frac{1}{6}+\delta} \quad \forall 0 < t < 1$$

Then, we have:

$$\sqrt{n}(T_n - \mathbb{E}[T_n]) \xrightarrow{d} N(0, \sigma^2)$$

where $\sigma^2 = \lim_{n\to\infty} n\mathbb{V}(T_n) < \infty$.

Now, set $X_i = s(\boldsymbol{X}_i)$, $g(\cdot) = u(\cdot)$, and $J(t) = \mathbf{1}\{(k-1)n/K < tn \leq kn/K\}$. Then, we have $T_n = \hat{\tau}_k^2$. Assumption 5 guarantees $\mathbb{E}[g(X)^3] < \infty$. The function $J(t)$ is discontinuous only at the quantile points $t = k/K$ and $t = \frac{k-1}{K}$, and Assumption 4 guarantees the continuity of $g(F^{-1})$ at those points. The function J clearly satisfies the bounding condition with $\delta = 1/6$ and $M = 1$. Therefore, define the asymptotic variance as $\sigma_2^2 = \lim_{n\to\infty} n\mathbb{V}(\hat{\tau}_k^{(2)})$, and we can utilize Theorem S1 to show the following convergence:

$$\sqrt{n}(\hat{\tau}_k^{(2)} - \mathbb{E}(\hat{\tau}_k^{(2)})) \xrightarrow{d} N(0, \sigma_2^2) \tag{S9}$$

Now, we aim to combine the results given in Equations (S8) and (S9). Using Lemma S2, we can replace τ_k with $\mathbb{E}(\hat{\tau}_k)$ by adding a small bias term. Then, we have

$$\begin{aligned}\sqrt{n}(\hat{\tau}_k - \tau_k) &= \sqrt{n}(\hat{\tau}_k^{(1)} - \mathbb{E}(\hat{\tau}_k^{(1)})) + \sqrt{n}(\hat{\tau}_k^{(2)} - \mathbb{E}(\hat{\tau}_k^{(2)})) + O\left(n^{-1/2}\right) \\ &\xrightarrow{d} N(0, \sigma_1^2 + \sigma_2^2) \end{aligned}\tag{S10}$$

where the last line follows from the application of Lemma S1 to the convergence results given in Equations (S8) and (S9).

Equivalently, we can write Equation (S10) as,

$$\sqrt{n}\frac{\hat{\tau}_k - \tau_k}{\sqrt{\sigma_1^2 + \sigma_2^2}} \xrightarrow{d} N(0, 1)$$

Now, note that by the law of total variance, we have that

$$\begin{aligned} n\mathbb{V}(\hat{\tau}_k) &= n\mathbb{E}[\mathbb{V}(\hat{\tau}_k^{(1)} + \hat{\tau}_k^{(2)} \mid \boldsymbol{X}_n)] + n\mathbb{V}[\mathbb{E}(\hat{\tau}_k^{(1)} + \hat{\tau}_k^{(2)} \mid \boldsymbol{X}_n)] \\ &= n\mathbb{E}[\mathbb{V}(\hat{\tau}_k^{(1)} \mid \boldsymbol{X}_n)] + n\mathbb{V}(\hat{\tau}_k^{(2)}) \\ &\to \sigma_1^2 + \sigma_2^2 \end{aligned}\tag{S11}$$

Therefore, by Slutsky's lemma, we have that:

$$\frac{\hat{\tau}_k - \tau_k}{\sqrt{\mathbb{V}(\hat{\tau}_k)}} \xrightarrow{d} N(0, 1)$$

□

S4 Proof of Proposition 1

We prove this proposition by finding an example that satisfies it. Define $t(x) = \mathbb{E}(Y_i(1) - Y_i(0) \mid s(\boldsymbol{X}_i) = F^{-1}(x))$. Then, consider a scoring function s and a population such that:

$$t(x) = \begin{cases} 2 & x \geq F(c_k(s)) \\ 1 & x < F(c_k(s)) \end{cases}$$

Note that $t(x)$ is bounded everywhere but has a discontinuity. By definition of $\hat{c}_k(s)$, $F(\hat{c}_k(s))$ follows the Beta distribution with the shape and scale parameters equal to nk/K and $n - nk/K + 1$, respectively. Therefore, we have the following normal approximation:

$$\sqrt{n+1}\left(F(\hat{c}_k(s)) - \frac{nk}{K(n+1)}\right) \xrightarrow{d} N\left(0, \frac{k}{K}\left(1 - \frac{k}{K}\right)\right)$$

In particular, as $n \to \infty$, $F(\hat{c}_k(s))$ is distributed approximately symmetric around $F(c_k(s)) = \frac{k}{K}$ with an error of $O(n^{-1})$ and has a standard deviation of $O(n^{-1/2})$. Thus, we have,

$$\begin{aligned} \mathbb{E}(\hat{\tau}_k) - \tau_k &= K\mathbb{E}\left[\int_{F(c_k(s))}^{F(\hat{c}_k(s))} f(x)\mathrm{d}x\right] + K\mathbb{E}\left[\int_{F(c_{k-1}(s))}^{F(\hat{c}_{k-1}(s))} f(x)\mathrm{d}x\right] \\ &= (2-1)O\left(n^{-1/2}\right) + (1-1)O\left(n^{-1/2}\right) + O\left(n^{-1}\right) \\ &= O\left(n^{-1/2}\right) \end{aligned}$$

We can now conclude $\sqrt{n}(\mathbb{E}(\hat{\tau}_k) - \tau_k) \not\to 0$. □

S5 Proof of Theorem 3

We wish to prove that for $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1, \cdots, \hat{\tau}_K)$, $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_K)$, and $\boldsymbol{\Sigma}_n = \mathbb{V}(\hat{\boldsymbol{\tau}})$, we have:

$$\boldsymbol{\Sigma}_n^{-1/2}(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}) \xrightarrow{d} N(0, \boldsymbol{I})$$

where $\boldsymbol{I}$ is the $K \times K$ identity matrix.

By Equation (S10) in the proof of Theorem 2, for all $k = 1, \cdots, k$ we have,

$$\sqrt{n}(\hat{\tau}_k - \tau_k) \xrightarrow{d} N(0, \sigma_k^2),$$

where $\sigma_k^2 = \lim_{n\to\infty} n\mathbb{V}(\hat{\tau}_k)$. To prove the multi-dimensional result, we utilize the Cramer-Wold device, which we restate below:

Theorem S2 (Cramér and Wold (1936)) *Let $\boldsymbol{X}_n = (X_{n1}, \cdots, X_{nk})$ and $\boldsymbol{X} = (X_1, \cdots, X_k)$ be k-dimensional random vectors. Then $\boldsymbol{X}_n \to \boldsymbol{X}$ if and only if for all $(t_1, \cdots, t_k) \in \mathbb{R}^k$, we have:*

$$\sum_{i=1}^k t_i X_{ni} \xrightarrow{d} \sum_{i=1}^k t_i X_i$$

Now, consider $\boldsymbol{t} = (t_1, \cdots, t_K) \in \mathbb{R}^K$ and $\hat{\tau}_{\boldsymbol{t}} = \sum_{k=1}^K t_k \hat{\tau}_k$. Then, we can write $\hat{\tau}_{\boldsymbol{t}}$ as:

$$\hat{\tau}_{\boldsymbol{t}} = \frac{1}{n}\sum_{i=1}^n \left(\sum_{k=1}^K t_k \mathbf{1}\left\{\frac{(k-1)n}{K} < i \leq \frac{kn}{K}\right\}\right) U_{[i,n]}$$

where $U_{[i,n]}$ is defined in Equation (S5). We use the same proof strategy as the one used to prove Theorem 1. We define $u(s) = \mathbb{E}[U_i \mid s(\boldsymbol{X}_i) = s]$ and write:

$$\begin{aligned}\hat{\tau}_{\boldsymbol{t}} \;&=\; \underbrace{\frac{1}{n}\sum_{k=1}^{K} t_k \sum_{\frac{(k-1)n}{K} < i \leq \frac{kn}{K}} U_{[i,n]} - u(s(\boldsymbol{X}_{[i,n]}))}_{\hat{\tau}_{\boldsymbol{t}}^{(1)}} \\ &\quad + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K} t_k \mathbf{1}\left\{\frac{(k-1)n}{K} < i \leq \frac{kn}{K}\right\}\right) u(s(\boldsymbol{X}_{[i,n]}))}_{\hat{\tau}_{\boldsymbol{t}}^{(2)}}\end{aligned}$$

Using Lindberg's Central Limit Theorem for the conditional distribution of $\hat{\tau}_{\boldsymbol{t}}^{(1)}$ given $\boldsymbol{X}_n$ and applying Theorem S1 to $\hat{\tau}_{\boldsymbol{t}}^{(2)}$ yield,

$$\begin{aligned}\sqrt{n}(\hat{\tau}_{\boldsymbol{t}}^{(1)} - \mathbb{E}(\hat{\tau}_{\boldsymbol{t}}^{(1)})) \mid \boldsymbol{X}_n \;&\xrightarrow{d}\; N(0, \sigma_1^2) \\ \sqrt{n}(\hat{\tau}_{\boldsymbol{t}}^{(2)} - \mathbb{E}(\hat{\tau}_{\boldsymbol{t}}^{(2)})) \;&\xrightarrow{d}\; N(0, \sigma_2^2)\end{aligned}$$

where $\sigma_1^2 = \lim_{n\to\infty} n\mathbb{V}(\hat{\tau}_{\boldsymbol{t}}^{(1)} \mid \boldsymbol{X}_n)$ and $\sigma_2^2 = \lim_{n\to\infty} n\mathbb{V}(\hat{\tau}_{\boldsymbol{t}}^{(2)})$. This implies,

$$\sqrt{n}(\hat{\tau}_{\boldsymbol{t}} - \tau_{\boldsymbol{t}}) \;=\; \sqrt{n}\left(\hat{\tau}_{\boldsymbol{t}}^{(1)} - \mathbb{E}[\hat{\tau}_{\boldsymbol{t}}^{(1)}]\right) + \sqrt{n}\left(\hat{\tau}_{\boldsymbol{t}}^{(2)} - \mathbb{E}[\hat{\tau}_{\boldsymbol{t}}^{(2)}]\right) + O\left(n^{-1/2}\right) \xrightarrow{d} N(0, \sigma_1^2 + \sigma_2^2)$$

where the result follows from Lemma S1. Since we can easily show $n\mathbb{V}(\hat{\tau}_t) \to \sigma_1^2 + \sigma_2^2$, we have: $\sqrt{n}(\hat{\tau}_{\boldsymbol{t}} - \tau_{\boldsymbol{t}}) \to N(0, \lim_{n\to\infty} n\mathbb{V}(\hat{\tau}_{\boldsymbol{t}}))$. Therefore, by the Cramer-Wold device (Theorem S2), we have $\sqrt{n}(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}) \to N(0, \lim_{n\to\infty} n\boldsymbol{\Sigma}_n)$. Finally, Slutsky's Lemma implies the desired result,

$$\boldsymbol{\Sigma}_n^{-1/2}(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}) \to N(0, \boldsymbol{I}).$$

To derive the expression for the covariance matrix $\boldsymbol{\Sigma}_{kk'}$, we utilize the same approach as the one used in the proof of Theorem 1. We first apply the law of total covariance to obtain:

$$\begin{aligned}&\mathrm{Cov}(\hat{\tau}_k, \hat{\tau}_{k'}) \\ &= K^2\,\mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^{n}\{Y_{ki}^*(1) - Y_{ki}^*(0)\}, \frac{1}{n}\sum_{i=1}^{n}\{Y_{k'i}^*(1) - Y_{k'i}^*(0)\}\right) \\ &\quad + K^2\mathbb{E}\left[\mathrm{Cov}\left\{\frac{1}{n_1}\sum_{i=1}^{n}\left(\hat{f}_k(\boldsymbol{X}_i) - \frac{1}{K}\right)T_iY_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}\left(\hat{f}_k(\boldsymbol{X}_i) - \frac{1}{K}\right)(1-T_i)Y_i(0),\right.\right. \\ &\qquad\left.\left.\frac{1}{n_1}\sum_{i=1}^{n}\left(\hat{f}_{k'}(\boldsymbol{X}_i) - \frac{1}{K}\right)T_iY_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}\left(\hat{f}_{k'}(\boldsymbol{X}_i) - \frac{1}{K}\right)(1-T_i)Y_i(0)\;\middle|\;\boldsymbol{X}, \{Y_i(1), Y_i(0)\}_{i=1}^{n}\right\}\right].\end{aligned} \tag{S12}$$

Applying Neyman's finite sample variance analysis to the second term shows that this term is equal to:

$$K^2\mathbb{E}\left\{\frac{1}{n}\left(\frac{n_0}{n_1}S_{kk'1}^{*2} + \frac{n_1}{n_0}S_{kk'0}^{*2} + 2S_{kk'01}^{*}\right)\right\}, \tag{S13}$$

where $S^*_{kk'01} = \sum_{i=1}^n (Y^*_{ki}(0) - \overline{Y^*_k(0)})(Y^*_{k'i}(1) - \overline{Y^*_{k'}(1)})/(n-1)$. Since $Y^*_{ki}(t)$ and $Y^*_{k'j}(t)$ are correlated, we have:

$$\begin{aligned} &\operatorname{Cov}\left(\frac{1}{n}\sum_{i=1}^n (Y^*_{ki}(1) - Y^*_{ki}(0)), \frac{1}{n}\sum_{i=1}^n (Y^*_{k'i}(1) - Y^*_{k'i}(0))\right) \\ &= \operatorname{Cov}(Y^*_{ki}(1) - Y^*_{ki}(0), Y^*_{k'j}(1) - Y^*_{k'j}(0)) + \frac{1}{n}\mathbb{E}(S^{*2}_{kk'1} + S^{*2}_{kk'0} - 2S^*_{kk'01}). \end{aligned} \tag{S14}$$

For $k \neq k'$, we have:

$$\begin{aligned} &\operatorname{Cov}(Y^*_{ki}(1) - Y_{ki}(0), Y^*_{k'j}(1) - Y_{k'j}(0)) \\ =& \operatorname{Cov}\left(\left(\hat{f}_k(\boldsymbol{X}_i) - \frac{1}{K}\right)\tau_i, \left(\hat{f}_{k'}(\boldsymbol{X}_j) - \frac{1}{K}\right)\tau_j\right) \\ =& \left(1 - \frac{2}{K}\right)\operatorname{Cov}(\hat{f}_k(\boldsymbol{X}_i)\tau_i, \hat{f}_{k'}(\boldsymbol{X}_j)\tau_j) - \frac{1}{K}\operatorname{Cov}((1 - \hat{f}_k(\boldsymbol{X}_i))\tau_i, \hat{f}_{k'}(\boldsymbol{X}_j)\tau_j) \\ &- \frac{1}{K}\operatorname{Cov}((1 - \hat{f}_{k'}(\boldsymbol{X}_i))\tau_i, \hat{f}_k(\boldsymbol{X}_j)\tau_j) \\ =& \left(1 - \frac{2}{K}\right)\left(\frac{1}{K^2}\kappa_{kk'11} - \frac{1}{K^2}\kappa_{k1}\kappa_{k'1}\right) \\ &- \frac{1}{K}\left\{\frac{n/K(n - n/K - 1)}{n(n-1)}\kappa_{kk'01} + \frac{n/K(n - n/K - 1)}{n(n-1)}\kappa_{kk'10} - \frac{1}{K^2}\kappa_{k1}\kappa_{k'0} - \frac{1}{K^2}\kappa_{k0}\kappa_{k'1}\right\} \\ =& \frac{1}{K^3}\left\{(K-2)(\kappa_{kk'11} - \kappa_{k1}\kappa_{k'1}) - \frac{Kn - n - 1}{n-1}(\kappa_{kk'10} + \kappa_{kk'01}) + \kappa_{k1}\kappa_{k'0} + \kappa_{k0}\kappa_{k'1}\right\}. \end{aligned}$$

For $k = k'$, we have

$$\begin{aligned} &\operatorname{Cov}(Y^*_{ki}(1) - Y^*_{ki}(0), Y^*_{kj}(1) - Y^*_{kj}(0)) \\ =& \operatorname{Cov}\left(\left(\hat{f}_k(\boldsymbol{X}_i) - \frac{1}{K}\right)\tau_i, \left(\hat{f}_k(\boldsymbol{X}_j) - \frac{1}{K}\right)\tau_j\right) \\ =& \left(1 - \frac{2}{K}\right)\operatorname{Cov}(\hat{f}_k(\boldsymbol{X}_i)\tau_i, \hat{f}_k(\boldsymbol{X}_j)\tau_j) - \frac{2}{K}\operatorname{Cov}((1 - \hat{f}_k(\boldsymbol{X}_i))\tau_i, \hat{f}_k(\boldsymbol{X}_j)\tau_j) \\ =& \left(1 - \frac{2}{K}\right)\left\{\frac{n-K}{K^2(n-1)}\kappa_{k11} - \frac{1}{K^2}\kappa^2_{k1}\right\} - \frac{2}{K}\left\{\frac{n(K-1)}{K^2(n-1)}\kappa_{k01} - \frac{1}{K^2}\kappa_{k1}\kappa_{k0}\right\} \\ =& \frac{1}{K^3}\left\{(K-2)\left(\frac{n-K}{n-1}\kappa_{k11} - \kappa^2_{k1}\right) - \frac{2n(K-1)}{(n-1)}\kappa_{k01} + 2\kappa_{k1}\kappa_{k0}\right\}. \end{aligned}$$

Substituting Equations (S13) and (S14) into Equation S12, we obtain the desired covariance expression.

□

S6 Proof of Theorem 4

The proof of Theorem 3 above establishes that $\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\tau}}$ is asymptotically normally distributed with the identity variance matrix $\boldsymbol{I}$. For simplicity, throughout this proof, we will assume that $\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\tau}}$ is exactly normally distributed with unknown mean $\boldsymbol{\theta} = (\tau_1 - \tau, \cdots, \tau_K - \tau)$, i.e., $\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\tau}} \sim N(\boldsymbol{\theta}, \boldsymbol{I})$.

Let the likelihood of the data $\hat{\boldsymbol{\tau}}$ under the null and alternative hypotheses as $L_{\hat{\boldsymbol{\tau}}}(H^C_0)$ and $L_{\hat{\boldsymbol{\tau}}}(H^C_1)$. Under the asymptotic normal assumption, the likelihood ratio is given by:

$$\frac{L_{\hat{\boldsymbol{\tau}}}(H_0^C)}{L_{\hat{\boldsymbol{\tau}}}(H_1^C)} = \begin{cases} \exp\left\{(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}_1(\hat{\boldsymbol{\tau}}))^\top \boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}_1(\hat{\boldsymbol{\tau}}))\right\} & \boldsymbol{\theta} \in \Theta_0 \\ \exp\left\{-(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}}))^\top \boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}}))\right\} & \boldsymbol{\theta} \in \Theta_1 \end{cases}$$

Where $\boldsymbol{\mu}_i(\hat{\boldsymbol{\tau}})$ are the optimal mean vectors given data $\hat{\boldsymbol{\tau}}$ for region j of the hypothesis test, and is the solution to the following optimization problems for $j \in \{0, 1\}$:

$$\boldsymbol{\mu}_j(\hat{\boldsymbol{\tau}}) = \underset{\boldsymbol{\mu} \in \Theta_j}{\operatorname{argmin}} \|\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}\|^2$$

We can identify the optimal means $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_0)$ for each region of the hypothesis test through this optimization problem because the multivariate normal distribution is spherical and symmetric.

We use $(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}}))^\top \boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}}))$ as our test statistic. Note that when $\hat{\boldsymbol{\tau}} \in \Theta_0$, the statistic is always 0, so the null hypothesis is never rejected and thus we are consistent. Given that we have a composite test, we are interested in finding the uniformly most powerful test. This requires calculating the size of a test α, as a function of the critical value $C(\alpha)$:

$$\alpha = \sup_{\boldsymbol{\theta} \in \Theta_0} \Pr((\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}}))^\top \boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}})) > C(\alpha) \mid \boldsymbol{\theta})$$

Since the supremum must occur at the boundary $\partial\boldsymbol{\Theta}_0$ of the polytope $\boldsymbol{\Theta}_0$ the set Θ_0, the probability of exceeding $C(\alpha)$ is maximized when the solid angle of the $\boldsymbol{\Theta}_0$ region is minimized. By considering the shape of the polytope $\boldsymbol{\Theta}_0$, we recognize that the boundary points, which minimize the solid angle, are precisely those on the boundary when all constraints are active:

$$\alpha = \sup_t \Pr((\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}}))^\top \boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}})) > C(\alpha) \mid \tau_1 - \tau = \cdots = \tau_K - \tau = t).$$

We now note that we have translational invariance on this boundary, i.e., all points along $\tau_1 - \tau = \cdots = \tau_K - \tau$ have the same probability, yielding,

$$\alpha = \Pr((\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}}))^\top \boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}})) > C(\alpha) \mid \tau_1 - \tau = \cdots = \tau_K - \tau = 0)$$

Therefore, to identify the value of α, we just need the CDF of the statistic $(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}}))^\top \boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}_0(\hat{\boldsymbol{\tau}}))$ when $\hat{\tau} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. This can be easily estimated using Monte Carlo simulation. □

S7 Proof of Theorem 5

The derivation of bias is essentially identical to that given in Supplementary Appendix S1 and thus is omitted. To derive the variance, we first introduce the following useful lemma, adapted from Nadeau and Bengio (2000).

Lemma S3

$$\begin{aligned} \mathbb{E}(S_{Fk}^2) &= \mathbb{V}(\hat{\tau}_k^\ell) - \operatorname{Cov}(\hat{\tau}_k^\ell, \hat{\tau}_k^{\ell'}), \\ \mathbb{V}(\hat{\tau}_k(F, n-m)) &= \frac{\mathbb{V}(\hat{\tau}_k^\ell)}{L} + \frac{L-1}{L} \operatorname{Cov}(\hat{\tau}_k^\ell, \hat{\tau}_k^{\ell'}). \end{aligned}$$

where $\ell \neq \ell'$.

The lemma implies,

$$\mathbb{V}(\hat{\tau}_k(F, n-m)) \ = \ \mathbb{V}(\hat{\tau}_k^\ell) - \frac{L-1}{L}\mathbb{E}(S_{Fk}^2).$$

We then follow the same process of derivation as in Appendix S1 for the first term. The only difference occurs in the derivation of the covariance term:

$$\begin{aligned}
&\text{Cov}(Y_{ki}^\ell(1) - Y_{ki}^\ell(0), Y_{kj}^\ell(1) - Y_{kj}^\ell(0)) \\
=&\mathbb{E}_\ell\left[\text{Cov}_{\boldsymbol{X},Y}\left(\hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)})\tau_i, \hat{f}_k^\ell(\boldsymbol{X}_j^{(\ell)})\tau_j\right)\right] + \text{Cov}_\ell\left[\mathbb{E}_{\boldsymbol{X},Y}[\hat{f}_k^\ell(\boldsymbol{X}_i^{(\ell)})\tau_i], \mathbb{E}_{\boldsymbol{X},Y}[\hat{f}_k^\ell(\boldsymbol{X}_j^{(\ell)})\tau_j]\right] \\
=&\mathbb{E}_\ell\left[\frac{(n-K)\kappa_{k11}^\ell}{K^2(n-1)} - \frac{1}{K^2}(\kappa_{k1}^\ell)^2\right] + \frac{1}{K^2}\mathbb{V}_\ell\left(\kappa_{k1}^\ell\right).
\end{aligned}$$

□

S8 Proof of Consistency of $\widehat{\mathbb{E}(S_{Fk}^2)}$

We show that $\widehat{\mathbb{E}(S_{Fk}^2)}$ is consistent as L approaches infinity under the assumption that the fourth moments $\mathbb{E}(Y_i(t)^4) < \infty$ for $t = 0, 1$ and a sufficiently large value of m. Theorem 5 implies,

$$\begin{aligned}
\mathbb{V}(\hat{\tau}_k(F, n-m)) = \ & K^2\left(\frac{\mathbb{E}(S_{Fk1}^2)}{m_1} + \frac{\mathbb{E}(S_{Fk0}^2)}{m_0}\right) \\
& + \mathbb{E}_\ell\left[\frac{(n-K)\kappa_{k11}^\ell}{K^2(n-1)} - \frac{1}{K^2}(\kappa_{k1}^\ell)^2\right] + \mathbb{V}\left(\kappa_{k1}^\ell\right) - \frac{L-1}{L}\mathbb{E}(S_{Fk}^2),
\end{aligned}$$

Now, define:

$$\mathbb{V}(\hat{\tau}_k(\widehat{F, n-m})) = K^2\left(\frac{\widehat{\mathbb{E}(S_{Fk1}^2)}}{m_1} + \frac{\widehat{\mathbb{E}(S_{Fk0}^2)}}{m_0}\right) + \frac{(n-K)\widehat{\mathbb{E}[\kappa_{k11}^\ell]}}{K^2(n-1)} - \frac{1}{K^2}(\widehat{\mathbb{E}[\kappa_{k1}^\ell]})^2 + \widehat{\mathbb{V}\left(\kappa_{k1}^\ell\right)}$$

By construction, we have that as $m \to \infty$:

$$\frac{\mathbb{V}(\hat{\tau}_k(\widehat{F, n-m}))}{\mathbb{V}(\hat{\tau}_k(F, n-m)) + \frac{L-1}{L}\mathbb{E}(S_{Fk}^2)} \xrightarrow{p} 1$$

Applying Lemma 1 from Nadeau and Bengio (2000) to $\hat{\tau}_k(F, n-m)$ gives

$$\mathbb{V}(\hat{\tau}_k(F, n-m)) \geq \ \mathbb{E}(S_{Fk}^2).$$

Therefore, we have:

$$\lim_{m\to\infty} \frac{\mathbb{V}(\hat{\tau}_k(\widehat{F, n-m}))}{\mathbb{E}(S_{Fk}^2) + \frac{L-1}{L}\mathbb{E}(S_{Fk}^2)} \geq 1$$

Note that we can write $\widehat{\mathbb{E}(S_{Fk}^2)}$ as:

$$\widehat{\mathbb{E}(S_{Fk}^2)} \ = \ \min\left(S_{Fk}^2, \mathbb{V}(\hat{\tau}_k(\widehat{F, n-m}))\right).$$

By definition of S^2_{Fk}, if the fourth moments of $Y_i(t)$ exist, we have $\mathbb{V}(S^2_{Fk}) = O(L^{-1})$, and thus as $L \to \infty$:

$$\frac{S^2_{Fk}}{\mathbb{E}(S^2_{Fk})} \xrightarrow{p} 1$$

Let $\epsilon > 0$. There exists L_0 such that for all $L > L_0$, $|\frac{S^2_{Fk}}{\mathbb{E}(S^2_{Fk})} - 1| < \epsilon$. Similarly, there exists m_0 such that for all $\mathbb{V}(\hat{\tau}_k(\widehat{F, n-m})) > (1-\epsilon)\mathbb{E}(S^2_{Fk})\ \forall m > m_0$. Therefore, for all $m > m_0$, we have that:

$$\lim_{L\to\infty} \frac{\widehat{\mathbb{E}(S^2_{Fk})}}{\mathbb{E}(S^2_{Fk})} = \lim_{L\to\infty} \frac{\min\left(S^2_{Fk}, \mathbb{V}(\hat{\tau}_k(\widehat{F, n-m}))\right)}{\mathbb{E}(S^2_{Fk})} = \lim_{L\to\infty} \frac{S^2_{Fk}}{\mathbb{E}(S^2_{Fk})} \xrightarrow{p} 1$$

□

S9 Proof of Theorem 6

We first shows the bias is small.

LEMMA S4 $\lim_{n\to\infty} |\hat{\tau}_k(F, n-m) - \tau_k(F, n-m)| = O\left(m^{-1}\right)$

Proof

$$\begin{aligned} |\hat{\tau}_k(F, n-m) - \tau_k(F, n-m)| &\leq \frac{1}{L}\sum_{\ell=1}^{L} |\mathbb{E}(\hat{\tau}^\ell_k(F, n-m)) - \tau^\ell_k(F, n-m)| \\ &= \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{E}_{\mathcal{Z}^{-\ell}}\left[O\left(m^{-1}\right)\right] \\ &= O\left(m^{-1}\right). \end{aligned}$$

The first equality follows because the estimator for each fold $\hat{\tau}^\ell_k(F, n-m)$ is equivalent to the non-cross-fitting estimator under m samples and so Lemma S2 is applicable. The second equality follows from Assumption 6. □

We first write:

$$\hat{\tau}_k(F, n-m) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\left\{\frac{(k-1)m}{K} < i \leq \frac{km}{K}\right\} U_{[i,m]}$$

where $U_{[i,m]} \in \mathbb{R}$ is defined as,

$$U_{[i,m]} := \frac{1}{L}\sum_{\ell=1}^{L} K\hat{f}^\ell_k(\boldsymbol{X}^{(\ell)}_{[i,m]})Y^{(\ell)}_{[i,m]}\left(\frac{T^{(\ell)}_{[i,m]}}{q} - \frac{1-T^{(\ell)}_{[i,m]}}{1-q}\right).$$

and $(Y^{(\ell)}_{[i,m]}, T^{(\ell)}_{[i,m]}, \boldsymbol{X}^{(\ell)}_{[i,m]})$ are ordered separately for each split ℓ such that:

$$s^\ell(\boldsymbol{X}^{(\ell)}_{[i,m]}) \leq s^\ell(\boldsymbol{X}^{(\ell)}_{[i,m]}) \leq \cdots \leq s^\ell(\boldsymbol{X}^{(\ell)}_{[i,m]})$$

Now by Assumption 7, there exists a fixed scoring rule $s(\boldsymbol{X})$ and corresponding treatment rule $f_k(\boldsymbol{X}_i) = \mathbf{1}\{s(\boldsymbol{X}_i) > c_{k-1}(s)\} - \mathbf{1}\{s(\boldsymbol{X}_i) > c_k(s)\}$ such that we can write:

$$U_{[i,m]} = \tilde{U}_{[i,m]} + \epsilon_{[i,m]}$$

$$\tilde{U}_{[i,m]} := \frac{1}{L}\sum_{\ell=1}^{L} K f_k(\boldsymbol{X}_{[i,m]}^{(\ell)})Y_{[i,m]}^{(\ell)}\left(\frac{T_{[i,m]}^{(\ell)}}{q} - \frac{1-T_{[i,m]}^{(\ell)}}{1-q}\right)$$
$$\tilde{\tau}_k(F, n-m) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\left\{\frac{(k-1)m}{K} < i \leq \frac{km}{K}\right\}\tilde{U}_{[i,m]}$$

where

$$\begin{aligned}
\mathbb{E}[\epsilon_{[i,m]}] &= \mathbb{E}\left[\frac{1}{L}\sum_{l=1}^{L} K(f_k(\boldsymbol{X}_{[i,m]}^{(\ell)}) - \hat{f}_k^{\ell}(\boldsymbol{X}_{[i,m]}^{(\ell)}))Y_{[i,m]}^{(\ell)}\left(\frac{T_{[i,m]}^{(\ell)}}{q} - \frac{1-T_{[i,m]}^{(\ell)}}{1-q}\right)\right] \\
&\leq \sqrt{\mathbb{E}\left[\left(\frac{1}{L}\sum_{l=1}^{L} K(f_k(\boldsymbol{X}_{[i,m]}^{(\ell)}) - \hat{f}_k^{\ell}(\boldsymbol{X}_{[i,m]}^{(\ell)}))Y_{[i,m]}^{(\ell)}\left(\frac{T_{[i,m]}^{(\ell)}}{q} - \frac{1-T_{[i,m]}^{(\ell)}}{1-q}\right)\right)^2\right]} \\
&= o\left(m^{-1/2}\right)
\end{aligned}$$

Then, we can apply the proof of Theorem 2 on $\tilde{U}_{[i,m]}$ as f_k is fixed, which gives:

$$\frac{\tilde{\tau}_k(F, n-m) - \mathbb{E}[\tilde{\tau}_k(F, n-m)]}{\sqrt{\mathbb{V}(\tilde{\tau}_k(F, n-m))}} \to N(0,1)$$

Since $\mathbb{V}(\tilde{\tau}_k(F, n-m)) = \mathbb{V}(\hat{\tau}_k(F, n-m)) + o(m^{-1})$ and $\hat{\tau}_k(F, n-m) = \tilde{\tau}_k(F, n-m) + o(m^{-1/2})$, we have:

$$\frac{\hat{\tau}_k(F, n-m) - \tau_k(F, n-m)}{\sqrt{\mathbb{V}(\hat{\tau}_k(F, n-m))}} \to N(0,1)$$

□

S10 Proof of Theorem 7

The proof follows identically to the proof of Theorem 3 by applying the Cramer-Wold Device in Theorem S2 to the sequence $\sum_{k=1}^{K} t_k \hat{\tau}_k(F, n-m)$.