# Training and Testing with Multiple Splits: A Central Limit Theorem for Split-Sample Estimators

Bruno Fava*

November 25, 2025

[Click here for the latest version]

**Abstract**

As predictive algorithms grow in popularity, using the same dataset to both train and test a new model has become routine across research, policy, and industry. Sample-splitting attains valid inference on model properties by using separate subsamples to estimate the model and to evaluate it. However, this approach has two drawbacks, since each task uses only part of the data, and different splits can lead to widely different estimates. Averaging across multiple splits, I develop an inference approach that uses more data for training, uses the entire sample for testing, and improves reproducibility. I address the statistical dependence from reusing observations across splits by proving a new central limit theorem for a large class of split-sample estimators under arguably mild and general conditions. Importantly, I make no restrictions on model complexity or convergence rates. I show that confidence intervals based on the normal approximation are valid for many applications, but may undercover in important cases of interest, such as comparing the performance between two models. I develop a new inference approach for such cases, explicitly accounting for the dependence across splits. Moreover, I provide a measure of reproducibility for p-values obtained from split-sample estimators. Finally, I apply my results to two important problems in development and public economics: predicting poverty and learning heterogeneous treatment effects in randomized experiments. I show that my inference approach with repeated cross-fitting achieves better power than existing alternatives, often enough to reveal statistical significance that would otherwise be missed.

*Department of Economics, Northwestern University. Contact: brunofava@u.northwestern.edu.
I am incredibly grateful to Federico Bugni, Ivan Canay, Joel Horowitz, and Dean Karlan for their unparalleled advising. I thank Eric Auerbach, Denis Chetverikov, Federico Crippa, Georgy Egorov, Annie Liang, and Amilcar Velez for helpful discussions. All errors are my own. This research was supported in part through the computational resources and staff contributions provided for the Quest high performance computing facility at Northwestern University which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology.

# 1 Introduction

As predictive algorithms transform empirical economics, policy, and industry, it is now routine to use the same dataset to train and evaluate a new model. For example, when training a machine learning algorithm to predict treatment effects, create a targeted policy rule, or automate consumer credit scoring, it is essential to evaluate the quality of the predictions and assess whether implementation would generate disparate impact across demographic groups. However, standard methods for training and evaluating typically waste part of the data by splitting the sample into training and testing sets. I develop a new inference approach that uses the entire sample for both tasks by combining multiple splits, improving statistical power and reproducibility. I provide valid confidence intervals under weak conditions for model properties such as accuracy and fairness calculated using cross-fitting, repeated sample-splitting or repeated cross-fitting.

Specifically, I study a setting in which an analyst (a researcher, policymaker, or industry practitioner) wishes to use the same dataset to both:

(i) train a new model, and

(ii) evaluate some of its properties, such as a measure of accuracy or fairness.

For example, consider a government using machine learning (ML) to target recipients of a poverty alleviation program. Step (i) consists of training a model to predict, for example, families at higher risk of falling below the poverty line, while step (ii) consists of constructing a confidence interval for the out-of-sample mean squared error or rate of correct classifications.

Using the same observations for both steps (i) and (ii) creates a form of statistical dependence that makes inference challenging. For example, standard central limit theorems (CLTs) assume independence, which is violated in this setting. This difficulty is often overcome by randomly splitting the sample into two, one part to train the model (training sample), and the other to evaluate its properties (evaluation sample). Since each task is conducted with separate data, such statistical dependence is not generated, and one can use standard approaches to inference. This procedure, however, has three drawbacks: it uses only part of the data for training the model, only part of the data for evaluating its properties, and different random splits can lead to widely different estimates and potentially affect statistical significance.

My main contribution is an inference approach that averages estimates across multiple sample splits, improving upon a standard 50/50 split by using more data for training, using twice as much data for evaluation, and improving reproducibility. In empirical applications and Monte Carlo experiments, I show that these improvements

often reveal statistical significance that would otherwise be missed. The main challenge of using multiple splits is a new form of statistical dependence due to reusing observations in both training and evaluation roles across different splits. I address this challenge by proving a new CLT for a large class of split-sample estimators under weak conditions. My CLT builds on previous literature in two key dimensions: (i) it applies to a large class of estimators and split-sample procedures, and (ii) it imposes no restrictions on model complexity or rates of convergence and stability, only requiring that the estimated model converges to an arbitrary limit at any rate. This generality is crucial for accommodating popular ML algorithms, such as random forests or neural networks. Moreover, I characterize when confidence intervals based on the normal approximation are valid, showing they may fail in important cases such as comparing the performance of two models or learning features of heterogeneous treatment effects. I develop a new inference approach for such cases that explicitly accounts for the dependence across splits, leveraging my CLT. Finally, I develop a reproducibility measure for p-values from split-sample procedures, quantifying whether the number of repetitions is sufficiently large to ensure reproducible inference.

To illustrate the technical challenges and empirical implications of my results, consider the problem of predicting poverty as a simple running example, which is one of my applications. Accurate out-of-sample poverty prediction is central to Development Economics for understanding poverty dynamics and designing targeted interventions. I focus on assessing predictive accuracy as the natural starting point, though my framework applies more broadly. Consider a sample $D = (Y_i, X_i)_{i=1}^n$ of $n$ households, where $X_i$ are covariates and $Y_i$ is a binary indicator equal to one if household $i$ is below the poverty line 13 years after the covariates were measured. The goal is to use the sample to (i) train a model $\hat{\eta}(x)$ to predict poverty by estimating $P(Y = 1|X = x)$, for example using a machine learning algorithm, and (ii) evaluate its accuracy, for example by estimating and calculating a confidence interval (CI) for the out-of-sample mean squared error (MSE)

$$\theta_{\hat{\eta}} = \mathbb{E}\left[\left(Y_{new} - \hat{\eta}(X_{new})\right)^2 |D\right] = \int \left(y - \hat{\eta}(x)\right)^2 dP(y, x),$$

where $(Y_{new}, X_{new})$ is an out-of-sample observation drawn from the same distribution as the sample. Note that $\theta_{\hat{\eta}}$ is data-dependent, and is thus different from targeting a parameter $\theta_{\eta_0}$ for some fixed $\eta_0$. In the policy prediction example, the researcher is not interested in the out-of-sample accuracy of an ideal but unknown model $\eta_0$. Instead, they are interested in the accuracy of the actually estimated model $\hat{\eta}$.

In this context, confidence intervals are often constructed using sample-splitting. If the entire sample is used for both tasks, standard CLTs do not apply to the average

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\eta}(X_i))^2,$$

since the summands are not independent. For example, $Y_1 - \hat{\eta}(X_1)$ and $Y_2 - \hat{\eta}(X_2)$ are dependent since $\hat{\eta}$ is estimated with both $(Y_1, X_1)$ and $(Y_2, X_2)$. A standard approach to handle this dependence is to impose complexity restrictions on how $\hat{\eta}$ is estimated, such as Donsker conditions. These restrictions hold for simple procedures like ordinary least squares, but fail for complex machine learning algorithms frequently used in applied problems (Chernozhukov et al., 2018). Sample-splitting avoids this dependence without strong assumptions: randomly split $\{1, \ldots, n\}$ into sets $\mathsf{s}_1$ and $\mathsf{s}_2$ of size for example $n/2$, use data in $\mathsf{s}_1$ to estimate $\hat{\eta}_1$, and data in $\mathsf{s}_2$ to calculate the average

$$\hat{\theta}_{\hat{\eta}_1} = \frac{1}{n/2} \sum_{i \in \mathsf{s}_2} (Y_i - \hat{\eta}_1(X_i))^2. \tag{1.1}$$

Since the summands in $\hat{\theta}_{\hat{\eta}_1}$ are independent conditional on $\mathsf{s}_1$, standard CLTs apply, and the normal approximation gives a valid CI for $\theta_{\hat{\eta}_1}$. However, this procedure uses only half of the data for each task, and different random splits can lead to widely different estimates and potentially different conclusions about statistical significance.

Using multiple splits can improve upon these drawbacks but introduces a new challenge. Consider, for example, two-fold cross-fitting, where the roles of samples $\mathsf{s}_1$ and $\mathsf{s}_2$ are reversed and the final estimator averages the split-specific estimates. That is, estimate $\hat{\eta}_1$ using $\mathsf{s}_1$ and $\hat{\eta}_2$ using $\mathsf{s}_2$, then calculate the final estimator

$$\hat{\theta}_{\hat{\eta}} = \frac{1}{n} \left[ \sum_{i \in \mathsf{s}_1} (Y_i - \hat{\eta}_2(X_i))^2 + \sum_{i \in \mathsf{s}_2} (Y_i - \hat{\eta}_1(X_i))^2 \right], \tag{1.2}$$

where $\hat{\eta} = (\hat{\eta}_1, \hat{\eta}_2)$. The estimand in this case is the MSE of an average model, as discussed in the next paragraph. While this estimator averages over all $n$ observations, standard CLTs do not apply due to a different form of statistical dependence: the first sum is not independent of the second since both use the entire dataset. My first main contribution is a central limit theorem for a large class of estimators that includes $\hat{\theta}_{\hat{\eta}}$, which I use to construct valid CIs. In addition to using the entire sample for evaluation in (1.2), which reduces the variance of the asymptotic distribution compared to that of (1.1), more data can be used for training by increasing the number of folds. With 3 folds, for example, three models are trained, each using two-thirds of the data, with the remaining third used to evaluate the MSE. Finally, reproducibility is improved by repeating the splitting process multiple times and averaging the estimators over repetitions.

I show that $\sqrt{n}(\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}})$ is asymptotically normal under weak conditions, targeting its out-of-sample expectation

$$\theta_{\hat{\eta}} = \mathbb{E}\left[ \frac{1}{2} \left( Y_{new} - \hat{\eta}_1(X_{new}) \right)^2 + \frac{1}{2} \left( Y_{new} - \hat{\eta}_2(X_{new}) \right)^2 | D \right].$$

In the example above, $\theta_{\hat{\eta}}$ is mathematically equivalent to the MSE of the average model, that is, $\theta_{\hat{\eta}} = \theta_{\bar{\eta}}$, where

$$\theta_{\bar{\eta}} = \mathbb{E}\left[(Y_{new} - \bar{\eta}(X_{new}))^2 \,|D\right], \qquad \bar{\eta}(x) = \frac{1}{2}\hat{\eta}_1(x) + \frac{1}{2}\hat{\eta}_2(x).$$

This happens anytime the outcome is binary, and holds for the MSE, mean absolute deviation, among others, including when averaging over multiple folds and repetitions. In the poverty prediction example, this means that a researcher or policymaker can use model $\bar{\eta}$ for out-of-sample predictions, which will have MSE $\theta_{\hat{\eta}}$. For continuous outcomes, the researcher has two options. The first is to use a model $\tilde{\eta}(x)$ that predicts a value in $(\hat{\eta}_1(x), \hat{\eta}_2(x))$ at random. This model has an out-of-sample MSE equal to $\theta_{\hat{\eta}}$. Alternatively, one could still use $\bar{\eta}$, which has the guarantee to perform better or equal than $\tilde{\eta}$ in terms of out-of-sample accuracy due to a risk-contraction property (for details, see Appendix A).

I make three main contributions. First, I prove a new central limit theorem for a large class of split-sample estimators under mild conditions. Specifically, I make no restrictions on the complexity of the models $\hat{\eta}$, or on their rates of convergence or algorithmic stability. For sample-average estimators, my CLT follows under a standard moments condition and assuming that $\hat{\eta}$ converges to an arbitrary limit, at any rate. I show that the normal approximation yields a valid CI in many applications, but may fail to do so in important cases of interest, such as comparing the performance between two models or some instances when $\hat{\eta}$ converges to a constant. My second contribution builds on the CLT to develop a new inference approach that covers such cases, explicitly accounting for the dependence across splits. I focus on the case of comparing the performance between two models, and discuss how the arguments apply more broadly to other cases. Finally, I develop a reproducibility measure for p-values obtained from split-sample estimators. It addresses a common concern: another researcher using the same dataset, but different splits, may reach a different conclusion about statistical significance. For a given (large) number of repetitions of sample-splitting/cross-fitting, my measure quantifies p-value reproducibility, assessing whether the number of repetitions is sufficiently large to ensure reproducible inference.

Other contributions include a central limit theorem for split-sample empirical processes, which I use to prove my main central limit theorem, and may be of independent interest. I also apply this CLT to develop a new *ensemble* method for learning features of heterogeneous treatment effects in randomized experiments, following the framework of Chernozhukov et al. (2025b). The ensemble method improves on previous alternatives by using the entire sample for evaluation, more data for training, and combining multiple machine learning predictors, potentially improving power and avoiding issues of multiple hypothesis testing.

I apply my inference approaches to two important problems in development and public economics: predicting poverty and learning heterogeneous treatment effects in randomized experiments. In the first application, using a panel from Ghana (Osei et al., 2022) and Monte Carlo experiments, I show that repeated cross-fitting outperforms previous alternative approaches in detecting predictive power for being below the poverty line 13 years ahead. In the second application, I revisit the experiment of Karlan and List (2007) on charitable giving and conduct Monte Carlo simulations, and show that my ensemble method achieves improved power for detecting heterogeneous treatment effects compared to previous alternatives.

The rest of the paper is structured as follows. Section 1.1 summarizes related work, and Section 2 establishes the setup and notation. Section 3 establishes a central limit theorem for split-sample Z-estimators, and Section 4 develops inference using the normal approximation and for comparing two models. I introduce my measure of reproducibility in Section 5. Finally, I implement my inference approaches in two empirical applications: predicting poverty in Ghana in Section 6 and heterogeneous treatment effects in charitable giving in Section 7. Section 8 concludes. Proofs are delayed to Appendix B.

## 1.1 Related Work

I contribute to the literature on inference using multiple splits of the data. The literature on risk estimation via cross-validation provides related results establishing asymptotic normality for sample-average estimators based on multiple splits. Like my approach, these CLTs target data-dependent parameters, but rely on different types of assumptions and focus on the particular case of sample-averages. Dudoit and van der Laan (2005) consider estimators that average over all possible splits or cross-splits of the sample, assume bounded loss function and require $\hat{\eta}$ to be loss consistent for a risk minimizing function, whereas I assume that $\hat{\eta}$ converges to an arbitrary limit. Austern and Zhou (2020) and Bayle et al. (2020) provide CLTs under rate assumptions on the algorithmic stability of $\hat{\eta}$. Bayle et al. (2020) provide two CLTs using estimators based on a single repetition of cross-fitting, one relying on rate conditions for algorithmic stability, and the second requires a "conditional variance convergence" assumption that they verify using rates for loss stability. My result does not require verifying loss stability conditions, which may not be satisfied in some high-dimensional settings (Bates et al., 2024), and my result allows for any ML algorithm as long as $\hat{\eta}$ converges to an arbitrary limit at any rate, in the sense established in Section 3. LeDell et al. (2015) provide a CLT for the particular case of estimating the area under the curve (AUC) measure via cross-validation, and Andrews et al. (2022) derive confidence intervals for the different but related problem of learning the transfer error of a model across domains. Moreover, I document that asymptotic normality of

split-sample estimators does not immediately lead to valid inference for the important problem of comparing the performance between two models, and I construct a new inference approach for this case that explicitly incorporates the dependence across splits.

A different class of related results show asymptotic normality using cross-fitting when targeting parameters that are not data-dependent. These approaches require stronger conditions on $\hat{\eta}$ that may not hold in general for nonparametric models with more than a handful of covariates, such as requiring $\hat{\eta}$ to converge in probability at some specified rate (Luedtke and Van Der Laan, 2016; Belloni et al., 2017; Chernozhukov et al., 2018; Benkeser et al., 2020; Imai and Li, 2025). Leveraging the data-dependent parameter of interest, my CLT (Theorem 3.1) requires no complexity restrictions, and assumes $\hat{\eta}$ converges in probability to any limit at any rate.

In the context of learning features of heterogeneous treatment effects in randomized trials, Chernozhukov et al. (2025b) proposed taking the median of estimators, confidence intervals and p-values across splits, similarly focusing on a data-dependent parameter, without relying on complexity or rate assumptions. Wager (2024) proposed a modified, sequential approach based on Luedtke and Van Der Laan (2016), and Chernozhukov et al. (2025a) suggested taking the median over repetitions of the sequential approach. In the same framework, Imai and Li (2025) developed inference using cross-fitting, relying on rate assumptions. My results build on this literature in four main dimensions, relying on the mild assumption that the trained models converge to any limit, at any rate. First, my estimator uses all observations on the role of evaluation sample, leading to a smaller variance of its asymptotic distribution. Second, my approach does not exhibit a tradeoff between training and evaluation sample sizes, allowing for more data to be used to train the models. Third, I provide inference for an interpretable estimand under no rate assumptions on the trained models, while Chernozhukov et al. (2025b) require a rate of concentration condition for coverage of their median estimand, which requires for example the training sample to be large relative to the evaluation sample. Finally, I introduce a new *ensemble* method that combines predictions from multiple ML algorithms, potentially improving statistical power for detecting HTE, and avoiding issues of multiple hypothesis testing.

The literature on learning features of heterogeneous treatment effects with multiple splits is a subset of a broader literature on aggregating potentially dependent p-values (Rüger, 1978; Rüschendorf, 1982; Meng, 1994; Meinshausen et al., 2009; Gasparin et al., 2025). These approaches similarly apply to data-dependent parameters under weak conditions, and typically target size control under the worst data generating process, thus being conservative in general. My confidence intervals are asymptotically exact and improve statistical power.

Finally, my work is complementary to Ritzwoller and Romano (2023). They provide a stopping algorithm for determining how many times to repeat sample-splitting

to ensure reproducibility of averages over split-sample statistics, for example, the average point estimate. I take the number of repetitions as given and focus on inference, providing a measure of reproducibility for p-values calculated using multiple splits. While Ritzwoller and Romano (2023) uses an asymptotic framework that takes the sample size fixed and assumes a small threshold for the variability of the average split-sample statistic, my framework uses a growing sample size and number of repetitions.

# 2 Setup

I consider a setup in which an analyst (a researcher, policymaker, or industry practitioner) wishes to use a dataset to both (i) train a new model and (ii) evaluate some of its properties. This is typically the case when one wants to train a new model to automate or assist decision-making, for example using a machine learning algorithm. Since these algorithms, despite their potential, may perform poorly in practice or have disparate performance across groups, it is often important to assess their accuracy and fairness. I use the term fairness as in the algorithmic fairness literature (Chouldechova and Roth, 2018; Cowgill and Tucker, 2020; Barocas and Selbst, 2016) and provide example measures below. I state the analyst's goals, discuss the parameter of interest with examples, and introduce the split-sample procedures I study.

The first goal of the analyst is to train a model $\hat{\eta} \in H$ using an algorithm and a dataset $D = (W_i)_{i=1}^n$, where each $W_i \in \mathcal{W}$ is an iid draw from a distribution $P$. I use *train* to denote the fitting/estimation of $\hat{\eta}$ using $D$, *training algorithm* (or just *algorithm*) for the procedure that maps $D$ to $\hat{\eta}$, and *estimated model* (or just *model*) for the realized function $\hat{\eta}$. For example, one can use the Random Forests algorithm to train a new model $\hat{\eta}$. The sets $H$ and $\mathcal{W}$ are in principle unconstrained, and $H$ depends on the choice of training algorithm. Typically, the analyst will estimate $\hat{\eta}$ by minimizing some loss function. My setup, however, is agnostic to the choice of training algorithm, and all results hold for any algorithm as long as $\hat{\eta}$ converges to an arbitrary limit at any rate, in the sense defined in Section 3.

The second goal of the analyst is to use $D$ to evaluate some performance property of $\hat{\eta}$, denoted $\theta_{\hat{\eta}}$. Specifically, the analyst wishes to construct a confidence interval $\widehat{\mathrm{CI}}_\alpha$ for $\theta_{\hat{\eta}}$ such that, for $\alpha \in (0, 1)$,

$$\liminf_{n\to\infty} P\left(\theta_{\hat{\eta}} \in \widehat{\mathrm{CI}}_\alpha\right) \geqslant 1 - \alpha, \tag{2.1}$$

where the probability accounts for the randomness in both $\hat{\eta}$ and $\widehat{\mathrm{CI}}_\alpha$. $\theta_{\hat{\eta}}$ can be, for example, a measure of accuracy or fairness of $\hat{\eta}$. The parameter of interest, $\theta_{\hat{\eta}}$, depends on the data through the estimated model $\hat{\eta}$. This differs from the standard

semiparametric literature, where the parameter of interest takes the form of $\theta_{\eta_0}$ for some nuisance function $\eta_0$. In the applications I consider, the object of interest is $\theta_{\hat{\eta}}$ since the analyst/policymaker is interested in the accuracy or fairness of the specific estimated model $\hat{\eta}$ they can actually implement. This is different from evaluating the performance of an ideal but unknown model $\eta_0$.

I provide three examples of such parameters of interest, and then discuss related cases in the literature where the parameter of interest is data-dependent.

**Example 1** (Mean Squared Error)**.** *The individual observations are $W = (Y, X)$, where $Y \in \mathbb{R}$ is an outcome and $X \in \mathcal{X} \subseteq \mathbb{R}^d$ is a set of covariates with $d \geqslant 1$. $\hat{\eta} : \mathcal{X} \to \mathbb{R}$ is a function that predicts $Y$ from $X$. In the poverty prediction example discussed in Section 1 and developed in Section 6, $Y$ is a binary indicator for whether a household is below the poverty line, and $\hat{\eta}(x)$ is an estimate of $P(Y = 1|X = x)$. The mean squared error (MSE) of model $\hat{\eta}$ is*

$$\theta_{\hat{\eta}} = \int (y - \hat{\eta}(x))^2 \, dP(y, x).$$

*A related estimand, also covered by my framework, is the difference in MSE between two groups, which is a measure of fairness (Auerbach et al., 2024; Liang et al., 2024; Liu and Molinari, 2024). Let $W = (Y, X, G)$, where $G \in \{a, b\}$ indicates group membership (e.g., racial groups). Then,*

$$\theta'_{\hat{\eta}} = \int (y - \hat{\eta}(x))^2 \, dP_{Y,X|G=a}(y, x) - \int (y - \hat{\eta}(x))^2 \, dP_{Y,X|G=b}(y, x) \tag{2.2}$$

*quantifies how much better $\hat{\eta}$ performs for one group relative to the other, where $P_{Y,X|G=g}$ is the conditional distribution of $(Y, X)$ given $G = g$.* □

**Example 2** (Classification Rate - Binary Classifiers)**.** *The individual observations are $W = (Y, X)$, where $Y$ is a binary outcome and $X \in \mathcal{X} \subseteq \mathbb{R}^d$ is a set of covariates, for some $d \geqslant 1$. $\hat{\eta} : \mathcal{X} \to \{0, 1\}$ is a function that predicts whether $Y = 1$ or $Y = 0$. The correct classification rate of model $\hat{\eta}$ is*

$$\theta_{\hat{\eta}} = \int \mathbb{I}\left(y = \hat{\eta}(x)\right) dP(y, x).$$

*$\theta_{\hat{\eta}}$ is a measure of accuracy, corresponding to the probability that $\hat{\eta}$ classifies an observation correctly.*

*Similar to* (2.2)*, the difference in classification rate between two groups is a measure of fairness.* □

**Example 3** (Classification Rate - Probabilistic Classifiers)**.** *The previous example can be generalized to accommodate probabilistic classifiers* $\hat{\eta} : \mathcal{X} \to [0,1]$*, with* $\hat{\eta}(X)$ *being the estimated probability that* $Y = 1$ *given* $X$*. The correct classification rate is given by*

$$\theta_{\hat{\eta}} = \int \left[\hat{\eta}(x)\mathbb{I}\left(y=1\right) + (1-\hat{\eta}(x))\mathbb{I}\left(y=0\right)\right] dP(y,x).$$

*This is equivalent to the probability (taking* $\hat{\eta}$ *fixed) that* $a_{\hat{\eta}}(X) = Y$*, where* $a_{\hat{\eta}}(X) = 1$ *with probability* $\hat{\eta}(X)$*, independent of* $D$*. A measure of fairness can be defined similar to* (2.2)*.* □

There are several examples in the literature where the parameter of interest takes the form of a data-dependent $\theta_{\hat{\eta}}$. This occurs anytime the hypothesis of interest is selected only after the data has been used (Dawid, 1994). An important case is the approach of Chernozhukov et al. (2025b) to inference on features of heterogeneous effects in randomized trials, which I revisit in Section 7. Other examples include evaluating the impacts of data-driven algorithms in policy applications (Potash et al., 2015; Kuzmanovic et al., 2024), measuring welfare gains generated from data-driven rules (Kitagawa and Tetenov, 2018; Ida et al., 2024), and the "inference on winners" framework of Andrews et al. (2024).

My setup also applies to some cases where the parameter of interest is not data dependent, but is estimated using split-sample techniques. For example, in Fava (2025) I develop an approach to inference on points of the distribution of treatment effects. Although the parameter of interest, $\theta$, is not data dependent, I incorporate covariate-adjustment terms $\hat{\eta}$ that yield bounds $\theta_{\hat{\eta},L} \leqslant \theta \leqslant \theta_{\hat{\eta},U}$. Inference on $\theta$ can then be derived from the asymptotic distribution of split-sample estimators $(\hat{\theta}_{\hat{\eta},L}, \hat{\theta}_{\hat{\eta},U})$, centered around the bounds $(\theta_{\hat{\eta},L}, \theta_{\hat{\eta},U})$. Other examples where $\theta_{\hat{\eta}}$ is informative about a parameter $\theta$ include learning the mean outcome under an optimal treatment regime (Shi et al., 2020; Fischer-Abaigar et al., 2024), and averages of intersection bounds (Ji et al., 2024; Semenova, 2025). Another type of application is when $\theta = \theta_{\hat{\eta}}$ does not depend on $\hat{\eta}$, yet estimating $\theta_{\hat{\eta}}$ leads to some better properties. This is the case of adding a covariate-adjustment term for learning the average treatment effect in a randomized trial, as I discuss in Appendix I.

I consider four split-sample procedures for attaining the analyst's goals: 1) sample-splitting, 2) cross-fitting, 3) repeated sample-splitting, and 4) repeated cross-fitting. First, I introduce some notation. Let $[n] = \{1, \ldots, n\}$ and the dataset $D = (W_i)_{i\in[n]}$ be an iid sample of $W \sim P$. I denote the training algorithm by $\mathcal{A} : \mathcal{W}^m \to H$, a function that takes a sample of size $m$ and returns a value $\eta \in H$. The dependence on $m$ is suppressed in the notation of $\mathcal{A}$. For any subsample $\mathsf{s} \subset [n]$, $D_{\mathsf{s}} = \{W_i\}_{i\in\mathsf{s}}$.

Sample-splitting consists of taking a random subsample $\mathsf{s} \subseteq [n]$ of size $b$, using its complement $\tilde{\mathsf{s}} = [n] \setminus \mathsf{s}$ to train the model $\hat{\eta}_{\tilde{\mathsf{s}}} = \mathcal{A}(D_{\tilde{\mathsf{s}}})$, and calculating $\widehat{\mathrm{CI}}_\alpha$ from $\mathsf{s}$ for

the parameter $\theta_{\hat{\eta}_{\tilde{\mathsf{s}}}}$. Cross-fitting consists of partitioning $[n]$ into $K$ roughly equal-sized subsets (*folds*) $(\mathsf{s}_k)_{k=1}^K$, at random. For $k = 1, \ldots, K$, train a model $\hat{\eta}_{\tilde{\mathsf{s}}_k} = \mathcal{A}(D_{\tilde{\mathsf{s}}_k})$, that is, using all observations except those in fold $k$. Each model $\hat{\eta}_{\tilde{\mathsf{s}}_k}$ is trained from $n(K-1)/K$ observations when $n$ is a multiple of $K$. In Section 3, I discuss different ways to aggregate the $K$ models into an estimand $\theta_{\hat{\eta}}$, where $\hat{\eta} = (\hat{\eta}_{\tilde{\mathsf{s}}_k})_{k=1}^K$, and the construction of a confidence interval $\widehat{\mathrm{CI}}_\alpha$ for $\theta_{\hat{\eta}}$. I consider $K$ fixed as $n \to \infty$.

Repeated sample-splitting and cross-fitting consist of repeating the procedures above $M$ times. That is, for repeated sample-splitting, take $M$ independent, random subsamples of $[n]$ of size $b$, $(\mathsf{s}_{m,1})_{m=1}^M$, and train $M$ models $(\hat{\eta}_{\tilde{\mathsf{s}}_{m,1}})_{m=1}^M$. For repeated cross-fitting, take $M$ independent, random partitions of $[n]$ into $K$ roughly equal-sized folds, $\mathcal{R} = (r_m)_{m=1}^M$, where each $r_m = (\mathsf{s}_{m,k})_{k=1}^K$ forms a partition of $[n]$. For each subsample $\mathsf{s}_{m,k}$, train a model $\hat{\eta}_{\tilde{\mathsf{s}}_{m,k}}$ using all observations except the ones in $\mathsf{s}_{m,k}$, giving a total of $MK$ models. I discuss different ways to aggregate the multiple splits in Section 3.

I give a unified notation to the four split-sample procedures described above. Let $\mathcal{R} = (r_m)_{m=1}^M$ denote a collection of $M$ random splits of the sample, where each split can be either:

- Sample-splitting: $K = 1$ and $r_m = (\mathsf{s}_{m,1})$ with $\mathsf{s}_{m,1} \subset [n]$ of size $b$, or
- $K$-fold cross-fitting: $K > 1$ and $r_m = (\mathsf{s}_{m,k})_{k=1}^K$ forms a partition of $[n]$.

I use $K = 1$ to denote sample-splitting for convenience. $K = 1$ means that $r_m$ consists of one subsample, of size $b < n$ chosen by the researcher. For cross-fitting, I assume folds are equal-sized if $n$ is a multiple of $K$, and have sizes $\lfloor n/K \rfloor$ and $\lceil n/K \rceil$ otherwise, and define $b = n/K$. Define $\pi = \lim_{n\to\infty} b/n$, $\pi \in (0, 1)$. With this notation, $K = 1$ denotes sample-splitting and $K > 1$ denotes cross-fitting. I allow $M$ to grow as $n$ increases, and denote

$$\bar{M} = \lim_{n\to\infty} M \in \mathbb{N} \cup \{+\infty\}.$$

This notation unifies the four split-sample procedures described previously, as shown in Table 1. I use the term *multiple splits* to denote any of the three procedures that

Table 1: Classification of Split-Sample Procedures

| Limit number of repetitions ($\bar{M}$) | Number of folds ($K$) | |
|---|---|---|
| | 1 | $> 1$ |
| 1 | Sample-splitting | Cross-fitting |
| $> 1$ | Repeated sample-splitting | Repeated cross-fitting |

use more than one split ($\bar{M} > 1$ and/or $K > 1$). In all cases, I assume that the splits are taken at random uniformly over all possible splits or cross-splits. Although the number of possible splits is finite for any given $n$, I consider that the $M$ repetitions are taken independently, with repetition. This assumption reflects common practice, as the computationally feasible number of repetitions is usually much smaller than the total number of possible splits, so that the probability of taking two identical splits is negligible.

I compare the four split-sample procedures in terms of statistical power, modeling power, and reproducibility properties in Section 3.2.

# 3 CLT for Split-Sample Z-Estimators

I prove a central limit theorem for split-sample Z-estimators, defined as zeroes of empirical moment equations. Z-estimators are a large class of estimators which include averages, linear regressions, and most M-estimators, since the parameter value that maximizes some objective function is the same that sets its partial derivatives to zero. This CLT can be used off-the-shelf in many applications, including the poverty prediction application in Section 6. First, in Section 3.1, I define split-sample Z-estimators and Z-estimands, introduce the assumptions used, and state the CLT. Finally, in Section 3.2, I compare the four split-sample procedures (sample-splitting, cross-fitting, repeated sample-splitting, and cross-fitting).

I provide a more accessible exposition for the particular case of sample average estimators, such as the MSE (Example 1), in Appendix D. I prove a new CLT for split-sample empirical processes in Appendix E, which I use to prove my CLT for Z-estimators and may be of independent interest.

## 3.1 Main Result

Since Z-estimators can be nonlinear, unlike the mean squared error (Example 1), different approaches to aggregating multiple splits lead to different estimators and estimands. I discuss three such approaches. Let $\|\cdot\|$ be the Euclidean norm, $\psi_{\theta,\eta} : \mathcal{W} \to \mathbb{R}^d$ be measurable functions for $\theta \in \Theta \subseteq \mathbb{R}^d$ and $\eta \in H$ ($H$ is defined as in Section 2), and $d \geqslant 1$. For $\eta \in H$, let $\Psi_\eta(\theta) = P\psi_{\theta,\eta}$, $\hat{\Psi}_{\mathsf{s},\eta}(\theta) = |\mathsf{s}|^{-1} \sum_{i \in \mathsf{s}} \psi_{\theta,\eta}(W_i)$, and $\dot{\Psi}_\eta$ be the Jacobian matrix of $\Psi_\eta(\theta_\eta)$. As in Section 2, let $\mathcal{R}$ denote a collection of splits with $M$ repetitions and $K$ folds, and let $\hat{\eta} = \hat{\eta}_{\mathcal{R}} = \left(\hat{\eta}_{\mathsf{s}_{m,k}}\right)_{m \in [M], k \in [K]}$.

The first type of estimand is an average across split-specific estimands:

$$\theta^{(1)}_{\hat{\eta}} = \frac{1}{MK} \sum_{r \in \mathcal{R}} \sum_{\mathsf{s} \in r} \theta^{(1)}_{\hat{\eta}_{\mathsf{s}}}, \tag{3.1}$$

where $\theta_\eta^{(1)}$ for $\eta \in H$ is the unique solution for $\theta$ in $\Psi_\eta(\theta) = 0$, i.e.,

$$\Psi_\eta(\theta_\eta^{(1)}) = 0.$$

(3.1) consists of solving the moment condition $\Psi_{\hat{\eta}_{\tilde{\mathsf{s}}}}(\theta) = 0$ for each split $\tilde{\mathsf{s}}$, and averaging over the split-specific estimands. The Z-estimator for (3.1) is

$$\hat{\theta}_{\hat{\eta}}^{(1)} = \frac{1}{MK} \sum_{r \in \mathcal{R}} \sum_{\mathsf{s} \in r} \hat{\theta}_{\hat{\eta}_{\tilde{\mathsf{s}}}}^{(1)}, \tag{3.2}$$

where $\hat{\theta}_{\hat{\eta}_{\tilde{\mathsf{s}}}}^{(1)} \in \arg\min_{\theta \in \Theta} \left\| \hat{\Psi}_{\mathsf{s}, \hat{\eta}_{\tilde{\mathsf{s}}}}(\theta) \right\|$. This approach is analogous to the DML1 estimator in Chernozhukov et al. (2018).

The second type of estimand solves the average of the moment conditions. That is, $\theta_{\hat{\eta}}^{(2)}$ uniquely solves

$$\frac{1}{MK} \sum_{r \in \mathcal{R}} \sum_{\mathsf{s} \in r} \Psi_{\hat{\eta}_{\tilde{\mathsf{s}}}} \left( \theta_{\hat{\eta}}^{(2)} \right) = 0. \tag{3.3}$$

The associated Z-estimator is given by

$$\hat{\theta}_{\hat{\eta}}^{(2)} \in \arg\min_{\theta \in \Theta} \left\| \frac{1}{MK} \sum_{r \in \mathcal{R}} \sum_{\mathsf{s} \in r} \hat{\Psi}_{\mathsf{s}, \hat{\eta}_{\tilde{\mathsf{s}}}}(\theta) \right\|. \tag{3.4}$$

This approach is analogous to the DML2 estimator in Chernozhukov et al. (2018).

Finally, the third type of estimand is a hybrid of the previous two approaches. It solves the moment condition at each cross-split of the sample, and averages across repetitions. That is,

$$\theta_{\hat{\eta}}^{(3)} = \frac{1}{M} \sum_{r \in \mathcal{R}} \theta_{\hat{\eta}_r}^{(2)}, \tag{3.5}$$

where $\theta_{\hat{\eta}_r}^{(2)}$ uniquely solves

$$\frac{1}{K} \sum_{\mathsf{s} \in r} \Psi_{\hat{\eta}_{\tilde{\mathsf{s}}}} \left( \theta_{\hat{\eta}_r}^{(2)} \right) = 0.$$

The associated Z-estimator is given by

$$\hat{\theta}_{\hat{\eta}}^{(3)} = \frac{1}{M} \sum_{r \in \mathcal{R}} \hat{\theta}_{\hat{\eta}_r}^{(3)}, \tag{3.6}$$

where

$$\hat{\theta}_{\hat{\eta}_r}^{(3)} \in \arg\min_{\theta \in \Theta} \left\| \frac{1}{K} \sum_{\mathsf{s} \in r} \hat{\Psi}_{\mathsf{s}, \hat{\eta}_{\tilde{\mathsf{s}}}}(\theta) \right\|. \tag{3.7}$$

In this approach, each $\hat{\theta}^{(3)}_{\hat{\eta}_r}$ uses the whole sample both for calculating $\hat{\eta}_r$ and the average in (3.7), and the final estimator $\hat{\theta}^{(3)}_{\hat{\eta}}$ is the average of the cross-fitting estimators across repetitions. Note that $\theta^{(1)}_{\hat{\eta}} = \theta^{(3)}_{\hat{\eta}}$ if $K = 1$, $\theta^{(2)}_{\hat{\eta}} = \theta^{(3)}_{\hat{\eta}}$ if $M = 1$, and $\theta^{(1)}_{\hat{\eta}} = \theta^{(2)}_{\hat{\eta}} = \theta^{(3)}_{\hat{\eta}}$ if $K = M = 1$. The estimators are not assumed to be unique, but I assume the estimands and the limit of the estimators to be unique.

For a concrete example, I consider below the particular case of sample-averages, as in the example of calculating the MSE for poverty prediction (Example 1).

**Example 4** (Split-sample averages)**.**
*Let $\psi_{\theta,\eta}(w) = f_\eta(w) - \theta$ for some known $f_\eta$. In this case, the three estimators coincide:*

$$\hat{\theta}^{(j)}_{\hat{\eta}} = \frac{1}{MK} \sum_{r \in \mathcal{R}} \sum_{\mathsf{s} \in r} \frac{1}{|\mathsf{s}|} f_{\hat{\eta}_{\mathsf{s}}}(W_i)$$

*for any $j$, and the estimand is*

$$\theta^{(j)}_{\hat{\eta}} = \frac{1}{MK} \sum_{r \in \mathcal{R}} \sum_{\mathsf{s} \in r} \int f_{\hat{\eta}_{\mathsf{s}}}(w) dP(w).$$

□

$\hat{\theta}^{(2)}_{\hat{\eta}}$ can be interpreted as the value of $\theta$ that solves the moment condition for a randomized function that takes value across $\left(\hat{\eta}_{\tilde{\mathsf{s}}_{m,k}}\right)_{m \in [M], k \in [K]}$ uniformly at random. That is, (3.3) is equivalent to

$$\int \psi_{\hat{\theta}^{(2)}_{\hat{\eta}}, \hat{\eta}_\xi}(w) dP(w, \xi) = 0,$$

where $dP(w, \xi) = dP(w)dP(\xi)$ and $\xi$ takes value in $(\tilde{\mathsf{s}}_{m,k})_{m \in [M], k \in [K]}$ uniformly at random. If, for example, each $\hat{\eta}_{\mathsf{s}}$ is a probabilistic classifier as in Example 3, $\hat{\theta}^{(2)}_{\hat{\eta}}$ can be interpreted as solving the moment condition for a randomized rule $\bar{\eta}(x)$ that predicts a positive classification with probability $(MK)^{-1} \sum_{r \in \mathcal{R}} \sum_{\mathsf{s} \in r} \hat{\eta}_{\mathsf{s}}(x)$.

I provide a CLT for the three estimators $(\hat{\theta}^{(1)}_{\hat{\eta}}, \hat{\theta}^{(2)}_{\hat{\eta}}, \hat{\theta}^{(3)}_{\hat{\eta}})$. Below, I establish my main regularity conditions.

**Assumption 3.1.** *For some $\Theta' \subseteq \Theta$, the following conditions hold:*

(i) *For some $\delta > 0$,*

$$\sup_{P \in \mathcal{P}, \eta \in H, \theta \in \Theta'} \mathbb{E}_P \left[ |\psi_{\theta,\eta}(W)|^{2+\delta} \right] < \infty;$$

(ii) *There exists* $\eta_P^* \in H$ *such that for* $\tilde{\eta} = \mathcal{A}(D)$, $W \perp D$, *and every* $\theta \in \Theta'$,

$$\left|\psi_{\theta,\tilde{\eta}}(W) - \psi_{\theta,\eta_P^*}(W)\right| \xrightarrow{P} 0$$

*uniformly in* $P \in \mathcal{P}$.

□

Assumption 3.1(i) is a standard moments condition for CLTs. Assumption 3.1(ii) is a mild stability condition on $\tilde{\eta}$. Importantly, $\tilde{\eta}$ is allowed to converge at any rate and to any limit $\eta_P^*$, which may depend on $P$. It holds, for example, if

$$\psi_{\theta,\tilde{\eta}}(w) \xrightarrow{P} \psi_{\theta,\eta_P^*}(w)$$

pointwise for every $w \in \mathcal{W}$ and $\theta \in \Theta'$. This condition is more interpretable but stronger than required (see Assumption E.2 in Appendix E). Assumption 3.1(ii) differs from the typical approach in the double machine learning literature where faster convergence rates (often $n^{-1/4}$) are required for nuisance functions (e.g., Chernozhukov et al., 2018). The key difference between the two approaches is that I target a different, data-dependent parameter.

My CLT relies on the additional technical regularity conditions Assumption B.1, which I delay to Appendix B.1. This assumption adapts standard conditions for consistency and asymptotic normality of Z-estimators to the context of split-sample estimators (e.g., Van der Vaart, 2000; van der Vaart and Wellner, 2023). This is a weak assumption that holds in many settings, and it mostly concerns the choice of $\psi_{\theta,\eta}$. First, it assumes that the classes $\mathcal{F}_\eta = \{\psi_{\theta,\eta,j} : \theta \in \Theta'\}$ are Donsker, which restricts complexity along $\theta \in \Theta'$ but does not restrict the complexity of $H$. Second, it requires $\hat{\theta}_{\hat{\eta}}^{(j)}$ to nearly solve the moment conditions, and $\theta_{\hat{\eta}}^{(j)}$ to be unique and well-separated zeroes of the population moment conditions. Finally, it assumes that $\Psi_\eta$ is differentiable in $\theta$ for $\eta \in H$, and the Jacobian is continuous in $\eta$ around $\eta_P^*$. Assumption B.1 holds, for example, in the case of sample averages (Example 4), or the "fraction in poverty by tercile" estimator in the poverty prediction application in Section 6.

Theorem 3.1 is the first main result of this paper.

**Theorem 3.1.** *(CLT for split-sample Z-estimators)*
*Let Assumptions 3.1 and B.1 hold. Then, for* $j \in \{1, 2, 3\}$,

$$\sqrt{n}\left(\hat{\theta}_{\hat{\eta}}^{(j)} - \theta_{\hat{\eta}}^{(j)}\right) \rightsquigarrow \mathcal{N}\left(0, V_{\eta_P^*}\right)$$

*uniformly in* $P \in \mathcal{P}$, *where*

$$V_{\eta_P^*} = V_{\bar{M},K}\dot{\Psi}_{\eta_P^*}^{-1}\left(P\psi_{\theta_{\eta_P^*},\eta_P^*}\psi_{\theta_{\eta_P^*},\eta_P^*}^T\right)\left(\dot{\Psi}_{\eta_P^*}^{-1}\right)^T,$$

*and*

$$V_{\bar{M},K} = \begin{cases} \bar{M}^{-1}\left(\pi^{-1} + \bar{M} - 1\right), & \text{if } K = 1 \text{ and } \bar{M} < \infty \\ 1, & \text{otherwise.} \end{cases}$$

□

The limiting variance $V_{\eta_P^*}$ is the product of two terms, the scalar $V_{\bar{M},K}$ and a positive semidefinite matrix. The choice of split-sample procedure only affects $V_{\eta_P^*}$ through $V_{\bar{M},K}$, which acts as a variance-inflating term since $V_{\bar{M},K} \geqslant 1$. When using a single split ($K = 1, \bar{M} = 1$), the asymptotic variance is inflated by $V_{\bar{M},K} = \pi^{-1}$, where $\pi$ is the fraction of the sample used to evaluate $\hat{\theta}_{\hat{\eta}}^{(j)}$ (as opposed to training $\hat{\eta}$). This occurs because $\hat{\theta}_{\hat{\eta}}^{(j)}$ is calculated from only $b = \pi n$ observations. When using repeated sample-splitting ($K = 1$ and $\bar{M} > 1$), $V_{\bar{M},K} = \bar{M}^{-1}\pi^{-1} + \bar{M}^{-1}(\bar{M} - 1)$ is an average of $\pi^{-1}$ and 1 with weights proportional to 1 and $\bar{M} - 1$. This occurs since each observation is picked a different number of times across splits for calculating $\hat{\theta}_{\hat{\eta}}^{(j)}$. A larger number of repetitions leads to more balance in how often each observation is selected, and $V_{\bar{M},K}$ decreases with larger $\bar{M}$. In fact, if $\bar{M} = \infty$, there is perfect balance in large samples and $V_{\bar{M},K} = 1$. When using cross-fitting ($K > 1$), all observations are used an equal amount of times, and $V_{\bar{M},K} = 1$. For intuition on this result, consider the particular case of sample averages (Example 4). In this case,

$$\hat{\theta}_{\hat{\eta}} = \frac{1}{\bar{M}} \sum_{r \in \mathcal{R}} \frac{1}{K} \sum_{\mathsf{s} \in r} \frac{1}{b} \sum_{i \in \mathsf{s}} f_{\hat{\eta}_{\mathsf{s}}}(W_i)$$

is the same for $j = 1, 2, 3$. If $\bar{M} = K = 1$, $\hat{\theta}_{\hat{\eta}}$ averages over $b = \pi n$ observations. If $\bar{M} > 1$ and $K = 1$, different observations are picked by splits $\mathsf{s}$ a different, random amount of times, and larger $\bar{M}$ leads to more balance. If $K > 1$, $\frac{1}{K}\sum_{\mathsf{s} \in r} \frac{1}{b} \sum_{i \in \mathsf{s}} f_{\hat{\eta}_{\mathsf{s}}}(W_i)$ is an average over all observations, the entire sample is used equally, and the variance-inflation term is minimum. Hence, the asymptotic variance is minimized using cross-fitting with any number of folds $K > 1$ and repetitions $\bar{M}$.

Theorem 3.1 appears to be new. The literature on risk estimation via cross-validation provides related results establishing asymptotic normality for sample average estimators based on multiple splits. Like my approach, these CLTs target data-dependent parameters, though they rely on different types of assumptions, and focus on the specific case of sample-averages. Dudoit and van der Laan (2005) consider estimators that average over all possible splits or cross-splits of the sample, assume bounded loss function and requires $\hat{\eta}$ to be loss consistent for a risk minimizing function, whereas I assume $\hat{\eta}$ converges to any limit. (Austern and Zhou, 2020; Bayle et al., 2020) provide CLTs under rate assumptions on the algorithmic stability of $\hat{\eta}$. Bayle et al. (2020) provides two CLTs using estimators based on a single repetition

of cross-fitting, one relying on rate condition for algorithmic stability, and the second requires a "conditional variance convergence" assumption that they verify using rates for loss stability. My result does not require verifying a loss stability condition, which may not be satisfied in some high-dimensional settings (Bates et al., 2024), and my result allows for any ML algorithm as long as Assumption D.1(ii) holds. LeDell et al. (2015) provides a CLT for the particular case of estimating the area under the curve (AUC) measure via cross-validation.

A different class of related results are CLTs with cross-fitting for parameters that are not data-dependent. These approaches require stronger conditions on $\hat{\eta}$, such as requiring $\hat{\eta}$ to converge in probability at some specified rate, typically $n^{-1/4}$ (Luedtke and Van Der Laan, 2016; Chernozhukov et al., 2018; Benkeser et al., 2020; Imai and Li, 2025). Theorem 3.1 requires no complexity restrictions, and assumes $\hat{\eta}$ converges in probability to any limit at any rate.

A central limit theorem for the class of split-sample Z-estimators appears to be new. The characterization of the asymptotic variance, specifically how the variance-inflating term $V_{\bar{M},K}$ depends on the number of splits $\bar{M}$ when $K = 1$, also appears to be new. The proof uses a new CLT for split-sample empirical stated in Appendix E, which also appears to be new and may be of independent interest.

**Remark 3.1.** *In the double machine learning context, which targets a different parameter $\theta_{\eta_0}$ and uses $M = 1$, simulation evidence (Chernozhukov et al., 2018) and theoretical results (Velez, 2024) suggest using DML2 over DML1. It is unclear whether similar arguments hold for comparing $\hat{\theta}^{(1)}_{\hat{\eta}}$ and $\hat{\theta}^{(2)}_{\hat{\eta}}$, and how they compare with $\hat{\theta}^{(3)}_{\hat{\eta}}$. Exploring theoretical and empirical properties of the three methods is an interesting direction for future research.* □

## 3.2 Comparison of Split-Sample Procedures

I compare the four split-sample procedures (sample-splitting, cross-fitting, repeated sample-splitting, and repeated cross-fitting) in terms of statistical power, modeling power, reproducibility, and computation time.

Cross-fitting and repeated cross-fitting, as well as repeated sample-splitting with $\bar{M} = \infty$, all exhibit the highest statistical power since they all minimize the variance of the asymptotic distribution in Theorem 3.1. Repeated sample-splitting with $1 < \bar{M} < \infty$ comes second, and sample-splitting yields the largest variance.

I say that an estimator has better modeling power than another if the models in $\left(\hat{\eta}_{\tilde{\mathsf{s}}_{m,k}}\right)_{m\in[M],k\in[K]}$ are trained using larger datasets. Using more data for training typically leads to models with smaller expected loss, as I formalize in Appendix C. For sample-splitting or repeated sample-splitting, modeling power increases by picking a smaller $b$ (and $\pi$), so that more data is used to train each $\hat{\eta}_{\tilde{\mathsf{s}}_{m,k}}$. However, if $\bar{M} < \infty$,

a smaller $b$ leads to smaller statistical power, since fewer data are used as evaluation sample at each split. When using cross-fitting, modeling power increases with $K$, since $b = n/K$. In this case, the returns to increasing $K$ are diminishing. For example, if $K = 2$, $\hat{\eta}_{\mathfrak{s}_{m,k}}$ is calculated with 50% of the sample, and this fraction raises to 90% with $K = 10$. If $K = 20$, however, the fraction only raises by another 5%. Although a large value of $K$ or small value of $\pi$ (when $K = 1$) lead to better modelling power, my asymptotic framework takes these quantities as fixed. This means that the quality of the asymptotic approximation may be poor if $K$ is large (or $\pi$ small) relative to the sample size. For example, my asymptotic framework does not accommodate for leave-one-out cross-fitting, that is, $K = n$.

I formalize the fact that increasing $M$ leads to better reproducibility properties in Section 5. For example, as $M$ increases, it becomes more likely that two researchers using the same dataset but different random splits will reach the same conclusion about statistical significance of $\theta_{\hat{\eta}}$. Although I make no formal comparison between the cases $K = 1$ and $K > 1$ in terms of reproducibility, I note that Ritzwoller and Romano (2023) documented the difference in variance between repeated sample-splitting and cross-fitting in an earlier draft.[1] Comparing cross-fitting with $M$ repetitions to sample-splitting with $KM$ repetitions, they argued that in principle it is possible that split-sample estimators have smaller variance conditional on data when $K = 1$ instead of $K > 1$, but show empirical evidence that cross-fitting typically leads to better reproducibility.

Table 2 summarizes the comparison of the four procedures.

Table 2: Comparison of Split-Sample Procedures

| Procedure | Statistical Power | Modeling Power | Reproducibility | Computation Time |
|---|---|---|---|---|
| Sample-splitting | Low | Low | Low | Low |
| Cross-fitting | High | High | Medium | Medium |
| Repeated sample-splitting | Med/High* | Med/High* | High** | High |
| Repeated cross-fitting | High | High | High** | High |

*High if $\bar{M} = \infty$, medium if $\bar{M} < \infty$.

**Whether repeated sample-splitting or cross-fitting dominates depends on application.

Modeling power considers the trade-off with statistical power: for sample-splitting and repeated sample-splitting with $\bar{M} < \infty$, increasing modeling power requires decreasing statistical power. Computation time and reproducibility columns compare repeated cross-fitting with $M$ repetitions to repeated sample-splitting with $KM$ repetitions.

[1]This discussion appears in the second version at `https://arxiv.org/abs/2311.14204` (dated December 9, 2023).

The choices of $M$, $K$, and $\pi$ (when $K = 1$) involve tradeoffs. Statistical power is maximized when $K > 1$ or $\bar{M} = \infty$ (Appendix D), and the reproducibility properties improve with larger $M$ and are ambiguously affected by $K$, despite empirical evidence that $K > 1$ usually leads to better properties (Ritzwoller and Romano, 2023). For $K = 1$ and $\bar{M} < \infty$, there is a tradeoff between statistical and modelling powers, unlike with cross-fitting. A larger $M$ is always beneficial in terms of reproducibility (and statistical power when $K = 1$), but this comes at the cost of higher computation time. Hence, I recommend choosing $M$ as large as computationally convenient, and $K > 1$ but small, since that provides valid asymptotic inference, maximum statistical power, and likely better reproducibility properties. In Section 5, I provide a measure to assess whether a given $M$ is sufficiently large to ensure reproducibility of p-values calculated from split-sample Z-estimators.

# 4 Inference on Split-Sample Estimands

The CLT in Theorem 3.1 can be directly applied to conduct inference on many split-sample estimands. However, confidence intervals based on the normal approximation may fail to cover $\theta_{\hat{\eta}}$ at the nominal level in some important cases of interest. First, in Section 4.1, I consider inference when the normal approximation is asymptotically exact, and discuss why this approximation may not be precise in some contexts. Then, in Section 4.2, I propose a new approach for the particular cases of inference on comparisons between models, which explicitly accounts for the dependence across splits.

I discuss in Section 4.1 that a typical case when the normal approximation CI may have coverage probability smaller than nominal is when the variance of a moment function is allowed to be zero in the limit. I provide a general method for inference that covers this case in Appendix F, by exploring the faster-than-$\sqrt{n}$ convergence rate of the empirical moment functions and introducing a tuning parameter. I also discuss in Sections 4.1 and 4.2 that although Section 4.2 considers the specific case of comparing two models, the arguments developed in that section apply more broadly, covering other cases such as the Generic ML approach of Chernozhukov et al. (2025b) (see Appendix B.5.5).

## 4.1 Inference from Normal Approximation

Consider the problem of conducting inference on $h(\theta_{\hat{\eta}})$, where $\theta_{\hat{\eta}}$ is any of the split-sample Z-estimands in Section 3, and $h$ is any scalar differentiable function with row-vector of partial derivatives $\dot{h}(\theta_{\eta_P^*}) \neq 0$. This encompasses many cases of interest, for example when $h(\theta_{\hat{\eta}})$ is a subset of the vector $\theta_{\hat{\eta}}$ or a linear combination of its

entries, as in the application of Section 6. An application of Theorem 3.1 and the delta-method yields

$$\sqrt{n}\left(h(\hat{\theta}_{\hat{\eta}}) - h(\theta_{\hat{\eta}})\right) \rightsquigarrow \mathcal{N}(0, \sigma^2_{\eta^*_P}), \tag{4.1}$$

where $\hat{\theta}_{\hat{\eta}}$ is a Z-estimator as in (3.4), and

$$\sigma^2_{\eta^*_P} = \dot{h}(\theta_{\eta^*_P}) V^*_{\eta^*_P} \dot{h}(\theta_{\eta^*_P})^T.$$

If $\sigma^2_{\eta^*_P} > 0$, one can calculate the plug-in estimator

$$\hat{\sigma}^2_{\hat{\eta}} = \dot{h}(\hat{\theta}_{\hat{\eta}}) \hat{V}_{\hat{\eta}} \dot{h}(\hat{\theta}_{\hat{\eta}})^T, \tag{4.2}$$

where $\hat{V}_{\hat{\eta}}$ is given in (B.11) in Appendix B.2, and the confidence interval

$$\left[h(\hat{\theta}_{\hat{\eta}}) - n^{-1/2} z_{1-\alpha/2} \hat{\sigma}_{\hat{\eta}}, h(\hat{\theta}_{\hat{\eta}}) + n^{-1/2} z_{1-\alpha/2} \hat{\sigma}_{\hat{\eta}}\right] \tag{4.3}$$

contains $h(\theta_{\hat{\eta}})$ with probability approaching $1-\alpha$, where $z_\alpha$ is the $\alpha$-th quantile of the standard normal distribution.

**Theorem 4.1.** *(Asymptotic Exactness of Normal Approximation CI)*
*Let the conditions of Theorem 3.1 hold, $V_{\eta^*_P}$ be positive definite, assume there exists an estimator $\widehat{\dot{\Psi}}_{\hat{\eta}}$ such that*

$$\left\| \widehat{\dot{\Psi}}_{\hat{\eta}} - \dot{\Psi}_{\eta^*_P} \right\| \xrightarrow{P} 0$$

*uniformly in $P \in \mathcal{P}$, and that $\inf_{P\in\mathcal{P}} \left\| \dot{h}(\theta_{\eta^*_P}) \right\| > 0$. Then, for any sequence $(P_n)_{n\geqslant 1} \subseteq \mathcal{P}$,*

$$P_n\left(h(\theta_{\hat{\eta}}) \in \left[h(\hat{\theta}_{\hat{\eta}}) - n^{-1/2} z_{1-\alpha/2} \hat{\sigma}_{\hat{\eta}}, h(\hat{\theta}_{\hat{\eta}}) + n^{-1/2} z_{1-\alpha/2} \hat{\sigma}_{\hat{\eta}}\right]\right) \to 1-\alpha.$$

□

Theorem 4.1 assumes the existence of a consistent estimator of $\dot{\Psi}_{\eta^*_P}$. If $\psi_{\theta,\eta}(w)$ is differentiable in $\theta$, this assumption is satisfied by the plug-in estimator defined in (B.10) in Appendix B.2 under a uniform integrability condition on this derivative. Otherwise, consistent estimators of $\dot{\Psi}_{\eta^*_P}$ can typically be constructed on a case-by-case basis (Hansen, 2022). Note that the probability in Theorem 4.1 is taken over both the random estimand $h(\theta_{\hat{\eta}})$ and the CI.

Theorem 4.1 implies that (4.3) contains $h(\theta_{\hat{\eta}})$ with probability approaching $1-\alpha$ in many settings. However, in some cases, (4.3) may not cover $h(\theta_{\hat{\eta}})$ with nominal probability, as illustrated in the two examples below.

**Example 5.** *Consider a dataset with covariates $X$, outcome $Y \in \mathbb{R}$, and moment function $\psi_{\theta,\eta}(y,x) = y\eta(x) - \theta$, so*

$$\hat{\theta}_{\hat{\eta}} = \frac{1}{MK} \sum_{r \in \mathcal{R}} \frac{1}{n} \sum_{\mathsf{s} \in r} \sum_{i \in \mathsf{s}} Y_i \hat{\eta}_{\tilde{\mathsf{s}}}(X_i).$$

*The limit variance in* (4.1) *is*

$$\sigma^2_{\eta^*_P} = \mathrm{Var}_P \left[ Y \eta^*_P(X) \right].$$

*If $\sigma^2_{\eta^*_P} > 0$,* (4.3) *contains $h(\theta_{\hat{\eta}})$ with probability approaching $1 - \alpha$. However, if $\eta^*_P(x) = 0$ for all $x$, $\sigma^2_{\eta^*_P} = 0$,*

$$\sqrt{n} \left( h(\hat{\theta}_{\hat{\eta}}) - h(\theta_{\hat{\eta}}) \right) \xrightarrow{P} 0,$$

*and* (4.3) *may not contain $h(\theta_{\hat{\eta}})$ with nominal probability.* □

**Example 6.** *Consider a dataset with covariates $X$, outcome $Y \in \mathbb{R}$, and moment function*

$$\psi_{\theta,\eta}(y,x) = \begin{pmatrix} y - \theta_0 - \theta_1 \eta(x) \\ (y - \theta_0 - \theta_1 \eta(x)) \eta(x) \end{pmatrix},$$

*that is, for each subsample $\mathsf{s}$, $\hat{\theta}_{\tilde{\mathsf{s}}}$ is the OLS estimator for $(\theta_{0,\tilde{\mathsf{s}}}, \theta_{1,\tilde{\mathsf{s}}})$ in the regression*

$$Y_i = \theta_{0,\tilde{\mathsf{s}}} + \theta_{1,\tilde{\mathsf{s}}} \hat{\eta}_{\tilde{\mathsf{s}}}(X_i) + \varepsilon_i$$

*using observations $i \in \mathsf{s}$. Focusing on the slope coefficient, the final estimator can be, for example,*

$$\hat{\theta}^{(1)}_{1,\hat{\eta}} = \frac{1}{MK} \sum_{r \in \mathcal{R}} \sum_{\mathsf{s} \in r} \hat{\theta}_{1,\tilde{\mathsf{s}}}.$$

*If $E[Z^T Z]$ is positive definite, where $Z = (1, \eta^*_P(X))^T$, the conditions of Theorem 3.1 are met, and* (4.3) *contains $\theta^{(1)}_{1,\hat{\eta}}$ with probability approaching $1-\alpha$. However, if $\eta^*_P(x)$ is constant in $x$, $E[Z^T Z]$ is not invertible, the conditions of Theorem 3.1 are not met, and* (4.3) *may contain $h(\theta_{\hat{\eta}})$ with probability below the nominal level.* □

Examples 5 and 6 have two features in common: the normal approximation CI may undercover $\theta_{\hat{\eta}}$ only when $\eta^*_P(x)$ is constant, and one of the empirical moment equations evaluated at the true parameter converges to zero at a rate faster than $n^{-1/2}$:

$$\min_{j \in \{1,\ldots,d\}} \left| \frac{1}{\sqrt{n}} \sum_{i \in \mathsf{s}} \psi_{\theta_{\hat{\eta}_{\tilde{\mathsf{s}}}}, \hat{\eta}_{\tilde{\mathsf{s}}}, j}(W_i) \right| \xrightarrow{P} 0, \tag{4.4}$$

where $\psi_{\theta,\eta,j}$ is the $j$-th entry of the vector $\psi_{\theta,\eta}$. In Section 4.2, I develop an approach that can be used to test whether $\eta_P^*(x)$ is constant, and although I focus on the particular case of comparing the performance between two models, the arguments apply more broadly and could be used to provide a valid CI for the problems in Examples 5 and 6 under the same conditions of Theorem 4.1. In Appendix F, I establish a general approach to inference on $\theta_{\hat{\eta}}$ that allows (4.4) to happen. The approach explores the faster-than-$\sqrt{n}$ convergence rate to provide an asymptotically uniformly valid CI by introducing a tuning parameter.

## 4.2 Inference on Model Comparisons

In several applications, the goal is not only to create a new model $\hat{\eta}$ and assess some property $\theta_{\hat{\eta}}$, but to compare such properties between two models. For example, if $\theta_{\hat{\eta}}$ is a measure of accuracy such as the mean squared error (Example 1), one might want to infer if $\hat{\eta}$ has better performance than a baseline model that predicts the sample mean of $Y$ for all observations. This is the case in the application of Section 6, where the goal is to assess whether a random forest model has predictive power for poverty, that is, whether it achieves smaller MSE than using the sample average. Alternatively, one might want to compare the performance of using different machine learning algorithms, such as training $\hat{\eta}$ with neural networks versus random forests. I show that the CLTs of the previous sections give a valid inference approach when both models do not have similar performances in large samples. However, if the models have similar performance, the asymptotic distribution of the difference in performance is degenerate at the $\sqrt{n}$ rate, and CIs based on the asymptotic approximation may fail to control size. In this section, I build on the CLT of Section 3 to develop an inference method that is valid for both cases. Although this section focuses on the particular case of comparing two models, I discuss in the end of Section 4.1 that the arguments developed in this section apply more broadly.

The setting is as follows. $\hat{\theta}_{\hat{\eta}}$ denotes any of the estimators $(\hat{\theta}^{(1)}_{\hat{\eta}}, \hat{\theta}^{(2)}_{\hat{\eta}}, \hat{\theta}^{(3)}_{\hat{\eta}})$ of Section 3, assumed to be a scalar ($d = 1$) (alternatively, one could consider a scalar transformation $h(\hat{\theta}_{\hat{\eta}})$ as in Section 4.1). I refer to the parameter $\theta_{\hat{\eta}}$ (defined analogously) as a *performance* measure for expositional convenience, though the results apply more generally. I focus on comparing $\theta_{\hat{\eta}}$ to the performance $\theta_{\hat{b}}$ of a baseline model $\hat{b} \in H$ computed using the entire sample, that is, without forms of sample-splitting. $\hat{b}$ is assumed to come from a parametric model, and it can be, for example, the sample average $\hat{b}(x) = n^{-1}\sum_{i=1}^{n} Y_i$ in Examples 1 and 3. Following the notation of Section 3, $\theta_{\hat{b}}$ is the unique solution for $\theta$ in $\Psi_{\hat{b}}(\theta) = 0$, i.e.,

$$\Psi_{\hat{b}}(\theta_{\hat{b}}) = 0.$$

Similarly, the estimator $\hat{\theta}_{\hat{b}}$ is a (near) zero of the empirical moment condition,

$$\hat{\theta}_{\hat{b}} \in \arg\min_{\theta\in\Theta} \left\| \frac{1}{n}\sum_{i=1}^{n} \psi_{\theta,\hat{b}}(W_i) \right\|.$$

In Appendix G, I discuss how to extend the current setting for comparing $\theta_{\hat{\eta}}$ to the performance of another model $\hat{\eta}'$ computed with the same split-sample approach as $\hat{\eta}$. Let

$$\mathcal{S} = (\mathsf{s}_{m,k})_{m\in[M],k\in[K]}$$

be a collection of $MK$ splits of the sample, that is, a vectorization of $\mathcal{R}$ defined in Section 2. Notice that each $\mathsf{s} \in \mathcal{S}$ is associated with a model $\hat{\eta}_{\mathsf{s}}$, as in (3.1).

To see the challenge of conducting inference based on $\hat{\theta}_{\hat{\eta}} - \hat{\theta}_{\hat{b}}$, consider a simplified setting where each $\hat{\theta}_{\hat{\eta}_{\mathsf{s}}}$ (as in (3.2)) is a sample average, that is, $\psi_{\theta,\eta}(w) = f_\eta(w) - \theta$ for some $f_\eta$ and $\hat{\theta}_{\hat{\eta}_{\mathsf{s}}} = |\mathsf{s}|^{-1}\sum_{i\in\mathsf{s}} f_{\hat{\eta}_{\mathsf{s}}}(W_i)$. The CLT in Theorem 3.1 gives

$$\sqrt{n}\left(\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(f_{\eta_P^*}(W_i) - Pf_{\eta_P^*}\right) + o_P(1),$$

and the normal approximation gives an asymptotically valid CI for $\theta_{\hat{\eta}}$. Similarly, if $\hat{b}$ converges to some model $b_P \in H$,

$$\sqrt{n}\left(\hat{\theta}_{\hat{b}} - \theta_{\hat{b}}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(f_{b_P}(W_i) - Pf_{b_P}\right) + o_P(1),$$

and these two results can be combined to construct a CI for $\theta_{\hat{\eta}} - \theta_{\hat{b}}$ based on a normal approximation. However, if the baseline model $b_P$ is the same as $\eta_P^*$, both estimators have the same limit, the difference

$$\sqrt{n}\left[\left(\hat{\theta}_{\hat{\eta}} - \hat{\theta}_{\hat{b}}\right) - \left(\theta_{\hat{\eta}} - \theta_{\hat{b}}\right)\right] = o_P(1) \tag{4.5}$$

has a degenerate limit in probability, and the CLT of Section 3 does not inform how to compute a CI for $\theta_{\hat{\eta}} - \theta_{\hat{b}}$.

First, I develop a test for whether any of the models $\hat{\eta}_{\mathsf{s}}$ perform better than $\hat{b}$, then show how this test can be used to construct a CI for $\theta_{\hat{\eta}} - \theta_{\hat{b}}$. Both results build on my CLT for Z-estimators.

### 4.2.1 A Multivariate One-sided Test for Model Differences

From (4.5), the asymptotic distribution of $\hat{\theta}_{\hat{\eta}} - \hat{\theta}_{\hat{b}}$ centered around the parameter of interest $\theta_{\hat{\eta}} - \theta_{\hat{b}}$ is degenerate at the $n^{-1/2}$ rate if $b_P = \eta_P^*$. Yet, for each split $\mathsf{s} \in \mathcal{S}$,

$$\begin{aligned}
&\sqrt{n}\left[\left(\hat{\theta}_{\hat{\eta}_{\tilde{\mathsf{s}}}} - \hat{\theta}_{\hat{b}}\right) - \left(\theta_{\hat{\eta}_{\tilde{\mathsf{s}}}} - \theta_{\hat{b}}\right)\right] \\
&= \frac{\sqrt{n}}{|\mathsf{s}|}\sum_{i\in\mathsf{s}}\left(f_{\eta_P^*}(W_i) - Pf_{\eta_P^*}\right) - \frac{\sqrt{n}}{n}\sum_{i=1}^{n}\left(f_{\eta_P^*}(W_i) - Pf_{\eta_P^*}\right) + o_P(1) \qquad (4.6)
\end{aligned}$$

has a non-degenerate limit since the first average does not include observations $i \in \tilde{\mathsf{s}}$. I explore this fact to construct a test of whether any model $\hat{\eta}_{\tilde{\mathsf{s}}}$ has better performance than $\hat{b}$, then develop a CI for $\theta_{\hat{\eta}} - \theta_{\hat{b}}$ in the following subsection.

Consider the hypothesis test

$$\begin{cases} H_{0,\hat{\eta}} : \theta_{\hat{\eta}_{\tilde{\mathsf{s}}}} - \theta_{\hat{b}} \geqslant 0 & \text{for all } \mathsf{s} \in \mathcal{S}, \\ H_{A,\hat{\eta}} : \theta_{\hat{\eta}_{\tilde{\mathsf{s}}}} - \theta_{\hat{b}} < 0 & \text{for some } \mathsf{s} \in \mathcal{S}. \end{cases} \qquad (4.7)$$

If $\theta_\eta$ is a measure of performance such as the mean squared error, having $\theta_{\hat{\eta}_{\tilde{\mathsf{s}}}} - \theta_{\hat{b}} < 0$ means that $\hat{\eta}_{\tilde{\mathsf{s}}}$ performs better than $\hat{b}$. The hypotheses $H_{0,\hat{\eta}}$ and $H_{A,\hat{\eta}}$ depend on $\hat{\eta}$ due to the data-dependent parameter of interest $\theta_{\hat{\eta}}$. Testing such hypotheses is analogous to constructing a confidence interval for a data-dependent parameter as in (2.1). Let

$$\delta_{\hat{\eta}} = \left(\theta_{\hat{\eta}_{\tilde{\mathsf{s}}}} - \theta_{\hat{b}}\right)_{\mathsf{s}\in\mathcal{S}},$$

and similarly define

$$\hat{\delta}_{\hat{\eta}} = \left(\hat{\theta}_{\hat{\eta}_{\tilde{\mathsf{s}}}} - \hat{\theta}_{\hat{b}}\right)_{\mathsf{s}\in\mathcal{S}}.$$

An application of Theorem 3.1 gives

$$\sqrt{n}\left(\hat{\delta}_{\hat{\eta}} - \delta_{\hat{\eta}}\right) \rightsquigarrow \mathcal{N}\left(0, \Sigma\right),$$

for some nonzero $\Sigma$ that can be consistently estimated with $\hat{\Sigma}$ (see equation B.12 in Appendix B.2). Since splits reuse observations, the off-diagonal terms of $\Sigma$ explicitly incorporate the dependence across splits.

Denote by $\hat{\sigma}_{\mathsf{s}}^2$ the entry of the main diagonal of $\hat{\Sigma}$ associated with $\mathsf{s} \in \mathcal{S}$, that is, with the term $\hat{\theta}_{\hat{\eta}_{\tilde{\mathsf{s}}}} - \hat{\theta}_{\hat{b}}$. I propose computing the test-statistic

$$T(\hat{\delta}_{\hat{\eta}}, n^{-1}\hat{\Sigma}) = \sum_{\mathsf{s}\in\mathcal{S}}\left(\min\left\{\sqrt{n}\frac{\hat{\theta}_{\hat{\eta}_{\tilde{\mathsf{s}}}} - \hat{\theta}_{\hat{b}}}{\hat{\sigma}_{\mathsf{s}}}, 0\right\}\right)^2.$$

This type of test statistic has been considered for example in Chernozhukov et al. (2007); Romano and Shaikh (2008); Andrews and Guggenberger (2009); Romano and Shaikh (2010) in the context of moment inequalities. Critical values $\hat{c}_{1-\alpha}$ can be computed via Monte Carlo: simulate $Z \sim \mathcal{N}(0, \hat{\Sigma})$ and estimate $\hat{c}_{1-\alpha}$ as the $1-\alpha$ quantile of $T(Z, n^{-1}\hat{\Sigma})$. I note that, alternatively, one could use the likelihood ratio test statistic.

Asymptotic exactness of this test under the least favorable null follows from similar conditions to Theorem 3.1, established below.

**Assumption 4.1.** *Assumptions 3.1 and B.2 hold with scalar $\hat{\theta}_{\hat{\eta}}$ ($d = 1$). Additionally,*

(i) $V_{\eta_P^*} > 0$*;*

(ii) *For some $b_P \in H$,*

$$\sqrt{n}\left(\hat{\theta}_{\hat{b}} - \theta_{\hat{b}}\right) - \sqrt{n}\left(\hat{\theta}_{b_P} - \theta_{b_P}\right) \xrightarrow{P} 0$$

*and*

$$\sqrt{n}\left(\theta_{\hat{b}} - \theta_{b_P}\right) \xrightarrow{P} 0$$

*uniformly in $P \in \mathcal{P}$.*

□

Assumption B.2 consists of more technical conditions, which are delayed to the appendix for ease of exposition. For example, they extend the Z-estimator assumptions on $\eta_P^*$ to $b_P$. Assumption 4.1(i) requires the limiting variance of $\sqrt{n}(\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}})$ to be positive, and Assumption 4.1(ii) defines the requirements on the baseline (parametric) estimator $\hat{b}$. It holds, for example, if $\hat{b}$ belongs to a Donsker class with probability approaching one, which typically happens for parametric models such as the sample average $\hat{b}(x) = n^{-1}\sum_{i=1}^n Y_i$.

**Theorem 4.2.** *(Size control of multivariate one-sided test for model differences) Let Assumption 4.1 hold. Then, for any $\bar{c}_2 > 0$,*

$$\limsup_{n\to\infty} \sup_{P \in \left\{P \in \mathcal{P} : P(\delta_{\hat{\eta}} \geqslant 0) > \bar{c}_2\right\}} P\left(T(\hat{\delta}_{\hat{\eta}}, n^{-1}\hat{\Sigma}) > \hat{c}_{1-\alpha} \;\middle|\; \delta_{\hat{\eta}} \geqslant 0\right) = \alpha.$$

*For any sequence $(P_n)_{n\geqslant 1} \subseteq \mathcal{P}$ with $\lim_{n\to\infty} P_n(\delta_{\hat{\eta}} = 0) > 0$,*

$$\lim_{n\to\infty} P_n\left(T(\hat{\delta}_{\hat{\eta}}, n^{-1}\hat{\Sigma}) > \hat{c}_{1-\alpha} \;\middle|\; \delta_{\hat{\eta}} = 0\right) = \alpha.$$

□

Theorem 4.2 appears to be new. It establishes size control: the probability of rejecting the null hypothesis, conditional on it being true, does not exceed $\alpha$ in large samples. Note that the probabilities in Theorem 4.2 are not random objects, they integrate over the distribution of the data conditional on the events $\delta_{\hat{\eta}} \geqslant 0$ or $\delta_{\hat{\eta}} = 0$. Alternative approaches for testing across multiple splits of the sample typically aggregate p-values or confidence intervals computed separately for each split, without accounting for the dependence structure across splits (see, e.g., Chernozhukov et al., 2025b; Gasparin et al., 2025). For example, Chernozhukov et al. (2025b) propose aggregating the median of p-values or CIs across splits. Because these methods do not incorporate the correlation across splits, they are conservative in most data-generating processes, as they guard against the worst-case dependence structure. In contrast, my approach explicitly accounts for the dependence across splits, which enables the test to achieve exactness under the least favorable null $\delta_{\hat{\eta}} = 0$ in a uniform sense across DGPs. The proof is made possible by the decomposition in (4.6), which follows from the new CLT in Section 3.

The result above requires the probability of the conditioning event to be bounded away from zero using the constant $\bar{c}_2 > 0$. This could lead to an apparent uniformity issue for sequences of DGPs $(P_n)_{n\geqslant 1}$ with $P_n(\delta_{\hat{\eta}} \geqslant 0) \to 0$, for example. For such sequences, the probability of rejecting the null conditional on the null being true could be greater than $\alpha$. This is not, however, an issue for empirical practice: for such sequences the probability of being under the null itself converges to zero. Incorrectly rejecting the null is not a concern when the probability of the null being true is zero.

### 4.2.2 A Confidence Interval for the Average Performance

I construct a new confidence interval for $\theta_{\hat{\eta}} - \theta_{\hat{b}}$ based on two insights from the previous subsections. The first is that a CI based on the normal approximation using Theorem 3.1 is asymptotically exact if $\theta_{\hat{\eta}} - \theta_{\hat{b}}$ converges in probability to a value different from zero, since in this case the terms in (4.5) do not cancel out. The second insight is that the case $\theta_{\hat{\eta}} - \theta_{\hat{b}} \xrightarrow{P} 0$ is closely connected with the null hypothesis of the one-sided test developed in the previous subsection. Hence, my CI consists of using the normal approximation if the one-sided test is rejected, and an extended CI in case it is not.

Define the normal approximation CI

$$\widehat{\mathrm{CI}}_{\alpha,\mathcal{N}} = \left[ \left(\hat{\theta}_{\hat{\eta}} - \hat{\theta}_{\hat{b}}\right) - z_{1-\alpha/2}\frac{\hat{\sigma}_{\hat{\delta}}}{\sqrt{n}}, \left(\hat{\theta}_{\hat{\eta}} - \hat{\theta}_{\hat{b}}\right) + z_{1-\alpha/2}\frac{\hat{\sigma}_{\hat{\delta}}}{\sqrt{n}} \right],$$

where $\hat{\sigma}_{\hat{\delta}}$ is a standard error for $\hat{\theta}_{\hat{\eta}} - \hat{\theta}_{\hat{b}}$ (see equation B.13 in Appendix B.2), and an extended CI

$$\widehat{\mathrm{CI}}_{\alpha,\mathrm{ext}} = \mathrm{Conv}\left(\widehat{\mathrm{CI}}_{\alpha,\mathcal{N}} \cup \{0\}\right),$$

where $\text{Conv}(\cdot)$ denotes the convex hull, that is, $\widehat{\text{CI}}_{\alpha,\text{ext}}$ has all the elements in $\widehat{\text{CI}}_{\alpha,\mathcal{N}}$, 0, and all elements in between. The final CI is given by

$$\widehat{\text{CI}}_\alpha = \begin{cases} \widehat{\text{CI}}_{\alpha,\mathcal{N}}, \text{ if } T(\hat{\delta}_{\hat{\eta}}, \hat{\Sigma}) > \hat{c}_{1-\alpha} \\ \widehat{\text{CI}}_{\alpha,\text{ext}}, \text{ otherwise.} \end{cases}$$

$\widehat{\text{CI}}_\alpha$ is based on a pre-test, using different inference approaches depending on whether the one-sided test is rejected or not. This construction is motivated by the following facts, which are formalized in Theorems B.1, 4.3 and 4.4. If $\theta_{\hat{\eta}} - \theta_{\hat{b}}$ converges in probability to a negative value, $P(T(\hat{\delta}_{\hat{\eta}}, \hat{\Sigma}) > \hat{c}_{1-\alpha}) \to 1$, and $\widehat{\text{CI}}_{\alpha,\mathcal{N}}$ is used, which is asymptotically exact. If $\theta_{\hat{\eta}} - \theta_{\hat{b}}$ converges in probability to a positive value, $P(T(\hat{\delta}_{\hat{\eta}}, \hat{\Sigma}) > \hat{c}_{1-\alpha}) \to 0$, $\widehat{\text{CI}}_{\alpha,\mathcal{N}}$ is asymptotically exact but $\widehat{\text{CI}}_{\alpha,\text{ext}}$ is used, which is valid since it is wider than $\widehat{\text{CI}}_{\alpha,\mathcal{N}}$, although conservative. This asymmetric construction is a choice, which reflects the motivating problem of this section of learning whether the new model $\hat{\eta}$ performs better (instead of worse) than the baseline model $\hat{b}$. Finally, if $\theta_{\hat{\eta}} - \theta_{\hat{b}} \xrightarrow{P} 0$, intuitively $P(T(\hat{\delta}_{\hat{\eta}}, \hat{\Sigma}) > \hat{c}_{1-\alpha})$ should be close to $\alpha$ given Theorem 4.2. If that happens, $\widehat{\text{CI}}_{\alpha,\text{ext}}$ covers $\theta_{\hat{\eta}} - \theta_{\hat{b}}$ with high probability since it includes 0, the limit of $\theta_{\hat{\eta}} - \theta_{\hat{b}}$. However, this guarantee depends on additional conditions as I discuss next, since $P(\delta_{\hat{\eta}} \geqslant 0)$ may not converge to one even if $\delta_{\hat{\eta}} \xrightarrow{P} 0$.

First, I show that $\widehat{\text{CI}}_\alpha$ is valid pointwise in $P \in \mathcal{P}$, assuming that if $\eta_P^* = b_P$, then the parametric model is well-specified in the sense that it minimizes the error $\theta_\eta$ in $\eta$, that is, $\theta_\eta \geqslant \theta_{b_P}$ for all $\eta \in H$. Then, I establish conditions under which $\widehat{\text{CI}}_\alpha$ is valid uniformly in $P \in \mathcal{P}$.

**Theorem 4.3.** *(Pointwise Asymptotic Validity of $\widehat{\text{CI}}_\alpha$)*
*Let Assumption 4.1 hold. Then, for any $P \in \mathcal{P}$ such that either*

*(i)* $\theta_{\eta_P^*} \neq \theta_{b_P}$, *or*

*(ii)* $\theta_{b_P} \leqslant \inf_{\eta \in H} \theta_\eta$,

$$\liminf_{n\to\infty} P\left((\theta_{\hat{\eta}} - \theta_{\hat{b}}) \in \widehat{\text{CI}}_\alpha\right) \geqslant 1 - \alpha.$$

□

Further, I show that $\widehat{\text{CI}}_\alpha$ is asymptotically valid for most sequences of $\theta_{\hat{\eta}} - \theta_{\hat{b}}$, and discuss why it may fail for specific sequences. Then, I establish that the additional condition Assumption 4.2 is sufficient for $\widehat{\text{CI}}_\alpha$ to be asymptotically uniformly valid in $P \in \mathcal{P}$. Later, I propose a modification to $\widehat{\text{CI}}_\alpha$ that gives uniform validity under only Assumption 4.1.

**Assumption 4.2.** *For any sequence* $(P_n)_{n\geqslant 1} \subseteq \mathcal{P}$ *such that* $\theta_{\eta^*_{P_n}} - \theta_{b_{P_n}} \to 0$,

$$\sqrt{n}\,(\theta_{\hat{\eta}} - \theta_{\hat{b}}) \xrightarrow{P_n} 0.$$

□

**Theorem 4.4.** *(Uniform Asymptotic Validity of* $\widehat{\mathrm{CI}}_\alpha$*)*
*Let Assumption 4.1 hold. For any* $\bar{c}_3 > 0$ *and* $\bar{c}_4 > 0$, *define*

$$\mathcal{P}_{\bar{c}_3,\bar{c}_4} = \left\{ P \in \mathcal{P} : P\Big((\theta_{\hat{\eta}} - \theta_{\hat{b}}) \geqslant 0 \vee (\theta_{\hat{\eta}} - \theta_{\hat{b}}) \leqslant \bar{c}_3\Big) > \bar{c}_4 \right\}.$$

*Then,*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}_{\bar{c}_3,\bar{c}_4}} P\left((\theta_{\hat{\eta}} - \theta_{\hat{b}}) \in \widehat{\mathrm{CI}}_\alpha \;\middle|\; (\theta_{\hat{\eta}} - \theta_{\hat{b}}) \geqslant 0 \vee (\theta_{\hat{\eta}} - \theta_{\hat{b}}) \leqslant \bar{c}_3\right) = 1 - \alpha.$$

*Moreover, if Assumption 4.2 holds,*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} P\left((\theta_{\hat{\eta}} - \theta_{\hat{b}}) \in \widehat{\mathrm{CI}}_\alpha\right) = 1 - \alpha.$$

□

Under Assumption 4.1, $\widehat{\mathrm{CI}}_\alpha$ covers $\theta_{\hat{\eta}} - \theta_{\hat{b}}$ when this difference is positive or "sufficiently" negative, with $\mathcal{P}_{\bar{c}_3,\bar{c}_4}$ only requiring this event to happen with positive probability. If $\theta_{\hat{\eta}} - \theta_{\hat{b}}$ converges to any negative value, coverage is asymptotically exact (Theorem B.1). If it converges to a positive value, similarly, the normal approximation CI is exact, and the extended $\widehat{\mathrm{CI}}_{\alpha,\mathrm{ext}}$ is conservative. Failure of coverage may happen only if $\theta_{\hat{\eta}} - \theta_{\hat{b}} \xrightarrow{P} 0^-$, that is, it converges to zero "from the left". For such sequences, the components of $\delta_{\hat{\eta}}$, $\theta_{\hat{\eta}_\mathbb{S}} - \theta_{\hat{b}}$, may be enough negative so that the one-sided test rejects the null with high probability, but since they converge to zero, the terms in (4.5) cancel out, and the normal approximation CI may undercover. Importantly, $\widehat{\mathrm{CI}}_\alpha$ is valid in the case of interest $\theta_{\hat{\eta}} - \theta_{\hat{b}} \geqslant 0$, that is, when $\hat{\eta}$ performs equally or worse than the baseline model $\hat{b}$. This is the case, for example, when the parametric model is well-specified, as in Theorem 4.3, since $\sqrt{n}\,(\theta_{\hat{b}} - \theta_{b_P}) \xrightarrow{P} 0$ from Assumption 4.1. Hence, $\widehat{\mathrm{CI}}_\alpha$ may overstate the advantage of $\hat{\eta}$ when it slightly outperforms $\hat{b}$, but not when their performances are equal or when $\hat{\eta}$ performs worse.

Assumption 4.2 rules out the problematic sequences by ensuring that if $\theta_{\hat{\eta}} - \theta_{\hat{b}} \xrightarrow{P} 0^-$, $\theta_{\hat{\eta}}$ is close enough to $\theta_{\hat{b}}$ in large samples so that the one-sided test does not reject with probability higher than $\alpha$. It is motivated by the fact that machine learning algorithms typically penalize deviations from the mean. If there is little signal to be learned by $\hat{\eta}$, that is, $\theta_{\eta^*_{P_n}}$ is close to $\theta_{b_{P_n}}$, it may be reasonable to expect that

regularization will make the estimates $\hat{\eta}$ closer to $\hat{b}$ than to $\eta^*_{P_n}$. For example, in the case of estimating a linear model with the Lasso, if the true coefficients are very small, penalization leads to estimated coefficients exactly equal to 0 with high probability (Zhao and Yu, 2006; Zhang and Huang, 2008; Wüthrich and Zhu, 2023). However, this assumption may not lead to a good approximation for the behavior of DGPs where $\theta_{\eta^*_{P_n}}$ is sufficiently distant from $\theta_{b_{P_n}}$ and $\hat{\eta}$ is estimated with no or little regularization.

Next, I provide an alternative, more conservative CI that gives uniform coverage without relying on Assumption 4.2. It deals with sequences with $\theta_{\hat{\eta}} - \theta_{\hat{b}} \xrightarrow{P} 0^-$ by modifying $\widehat{\mathrm{CI}}_\alpha$ to be more conservative in the one-sided test. For any $\bar{c}_5 > 0$, consider the modified version of the test in (4.7):

$$\begin{cases} H_{0,\hat{\eta}} : \theta_{\hat{\eta}_{\mathsf{s}}} - \theta_{\hat{b}} \geqslant -\bar{c}_5 & \text{for all } \mathsf{s} \in \mathcal{S}, \\ H_{A,\hat{\eta}} : \theta_{\hat{\eta}_{\mathsf{s}}} - \theta_{\hat{b}} < -\bar{c}_5 & \text{for some } \mathsf{s} \in \mathcal{S}. \end{cases}$$

$\bar{c}_5$ represents a degree of slackness on how large $-(\theta_{\hat{\eta}_{\mathsf{s}}} - \theta_{\hat{b}})$ has to be to reject the null hypothesis. The final CI is given by

$$\widehat{\mathrm{CI}}'_\alpha = \begin{cases} \widehat{\mathrm{CI}}_{\alpha,\mathcal{N}}, \text{ if } T(\hat{\delta}_{\hat{\eta}} + \bar{c}_5, \hat{\Sigma}) > \hat{c}_{1-\alpha} \\ \widehat{\mathrm{CI}}_{\alpha,\mathrm{ext}}, \text{ otherwise,} \end{cases}$$

and the critical value $\hat{c}_{1-\alpha}$ is the same as before. A large $\bar{c}_5$ gives more robustness in finite samples in the sense that

$$P\left((\theta_{\hat{\eta}} - \theta_{\hat{b}}) \in \widehat{\mathrm{CI}}'_\alpha\right)$$

is (weakly) increasing in $\bar{c}_5$. On the other hand, a large $\bar{c}_5$ leads to less power. Importantly, this approach is not necessary if the goal is to test the null $H_{0,\hat{\eta}} : \theta_{\hat{\eta}} - \theta_{\hat{b}} = 0$, since this cased is covered by Theorem 4.4. The modified confidence interval $\widehat{\mathrm{CI}}'_\alpha$ is intended for researchers who may want to be careful not to overestimate the magnitude of $\theta_{\hat{\eta}} - \theta_{\hat{b}}$ when it is small but negative.

**Theorem 4.5.** *(Uniform Asymptotic Validity of $\widehat{\mathrm{CI}}'_\alpha$)*
*Let Assumption 4.1 hold and fix any $\bar{c}_5 > 0$. Then,*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} P\left((\theta_{\hat{\eta}} - \theta_{\hat{b}}) \in \widehat{\mathrm{CI}}'_\alpha\right) \geqslant 1 - \alpha.$$

□

# 5 Reproducibility

The split-sample estimators and estimands defined in Section 3 depend not only on the algorithm used to estimate $\hat{\eta}$, but also on the specific splits of the sample $\mathcal{R}$. In applications, this may lead to the undesirable phenomenon that different researchers with the same dataset, using different random splits $\mathcal{R}$, may reach different conclusions in terms of statistical significance. Intuitively, by averaging over multiple splits, this phenomenon becomes less likely. In this section, I first formalize this intuition by establishing basic reproducibility properties of split-sample Z-estimators. Then, I develop a measure that quantifies the reproducibility of p-values from hypothesis tests based on Z-estimators for a given number of repetitions $M$.

## 5.1 Basic Reproducibility Properties

I establish two basic reproducibility properties for the three versions of split-sample Z-estimators defined in Section 3. The two results formalize the notion that, for fixed $n$, choosing to use a larger number of repetitions $M$ improves reproducibility of the estimators. The results exploit the fact that $\hat{\theta}^{(1)}_{\hat{\eta}}$ and $\hat{\theta}^{(3)}_{\hat{\eta}}$ are averages over $M$ independent repetitions $r \in \mathcal{R}$. For the second estimator, I use the fact that, conditional on the data $D$, $\hat{\theta}^{(2)}_{\hat{\eta}}$ is still a Z-estimator where the "observations" are the splits $r \in \mathcal{R}$ and the target parameter is the value of $\theta$ that solves the moment condition averaged over all possible splits. This allows me to explore large $M$ properties of $\hat{\theta}^{(2)}_{\hat{\eta}}$ using arguments similar to those applied to Z-estimators (e.g., Theorem 5.9 in Van der Vaart, 2000). For $\hat{\theta}^{(2)}_{\hat{\eta}}$, I require an additional technical condition which I delay to Appendix B.3.1. This assumption is analogous to standard conditions for proving consistency of Z-estimators, and holds, for example, if $\Theta'$ is bounded, $\psi_{\theta,\eta}$ is Lipschitz in $\theta$ with a Lipschitz constant that does not depend on $\eta$ or $w$, and if the solution to the moment condition averaged over all possible splits is unique.

The first reproducibility property is that, for fixed $n$, the variance of the Z-estimators conditional on the data converge to zero as $M$ grows. Conditional on the data, the estimators vary only due to the random partitioning. This approximates the behavior of the estimators when the number of repetitions $M$ is chosen to be large. This guarantees that two researchers with the same dataset and different splits will calculate estimators that are arbitrarily close to each other with high probability for large enough $M$.

**Proposition 5.1.** *Let 3.1 hold, $\pi, K$ be arbitrary, and $j \in \{1, 2, 3\}$. Additionally, let B.3 hold if $j = 2$. Then,*

$$\mathrm{Var}_P\left[\hat{\theta}^{(j)}_{\hat{\eta}} \mid D\right] \xrightarrow{P} 0$$

*as $M \to \infty$ with $n$ fixed.* □

For the estimators $\hat{\theta}^{(1)}_{\hat{\eta}}$ and $\hat{\theta}^{(3)}_{\hat{\eta}}$, I show that the conditional variance is strictly decreasing in $M$. This establishes a stronger property than the asymptotic result in Proposition 5.1: not only does reproducibility improve as $M \to \infty$, but every increase in $M$ strictly reduces variance and thus improves reproducibility.

**Proposition 5.2.** *Let 3.1 hold, $n$ be fixed, $M, \pi, K$ be arbitrary, and $j \in \{1, 3\}$. Then, if*

$$\text{Var}_P\left[\hat{\theta}^{(j)}_{\hat{\eta}} \mid D\right] > 0,$$

$\text{Var}_P\left[\hat{\theta}^{(j)}_{\hat{\eta}} \mid D\right]$ *is strictly decreasing in $M$.* □

## 5.2 A Reproducibility Measure

I propose a reproducibility measure for p-values from hypothesis tests based on transformations of split-sample Z-estimators. Specifically, I study reproducibility of the p-value for testing $H_{0,\hat{\eta}} : h(\theta_{\hat{\eta}}) = \tau$ versus $H_{A,\hat{\eta}} : h(\theta_{\hat{\eta}}) \neq \tau$ (and its one-sided versions) for $h : \Theta \to \mathbb{R}$ differentiable. The hypotheses $H_{0,\hat{\eta}}$ and $H_{A,\hat{\eta}}$ depend on $\hat{\eta}$ since the parameter of interest, $\theta_{\hat{\eta}}$, depends on $\hat{\eta}$. Testing this hypothesis is analogous to constructing a CI for $\theta_{\hat{\eta}}$: in fact, inverting this test for all values of $\tau$ at significance level $\alpha$ gives the confidence interval of Section 4.1.

I begin by defining the reproducibility measure, then describe the asymptotic framework I use and the technical challenges involved. Finally, I establish the limit distribution of the difference of t-statistics constructed from different random splits, and apply this result to construct the reproducibility measure. As in Section 3, I consider $M$ repetitions of sample-splitting with $K$ folds ($K = 1$ denotes repeated sample-splitting).

The goal of this section is to construct a measure $\hat{\delta}(\beta)$, for $\beta \in (0, 0.5)$, that satisfies

$$P\left(p_2 > p_1 + \hat{\delta}(\beta) \,\middle|\, D\right) = \beta + o_P(1),$$

where $p_1$ and $p_2$ are p-values for $H_{0,\hat{\eta}}$ calculated with separate, independent splits. This measure provides the following guarantee: if a researcher calculates a p-value $p_1$ using one set of random splits, then a second researcher using the same dataset, but different splits, will obtain a p-value exceeding $p_1 + \hat{\delta}(\beta)$ with probability approximately $\beta$. This allows researcher 1 to assess whether their result would remain statistically significant without the computational cost of re-running the analysis. For example, if $p_1 < 0.05$ but $p_1 + \hat{\delta}(\beta) > 0.05$ for some small $\beta$, the researcher may need to increase $M$ to guarantee reproducibility of their finding.

I consider an asymptotic regime where both the number of repetitions $M$ and the sample size $n$ grow to infinity, which is the main technical challenge for proving

validity of my reproducibility measure. An alternative framework is to consider the data $D$ fixed, let $M \to \infty$, and treat each repetition as an independent observation. Although this alternative regime facilitates statistical analysis, it provides asymptotic guarantees only when $M$ is large relative to $n$. In practice, choosing $M$ much larger than $n$ is often computationally intractable. My asymptotic framework better reflects much of empirical practice by allowing $M$ to grow slower than $n$, so that $M$ can be, for instance, a small fraction of $n$. The proofs of my results under this asymptotic regime rely on the CLT of Section 3.

I focus on the estimator $\hat{\theta}_{\hat{\eta}} = \hat{\theta}^{(2)}_{\hat{\eta}}$ from Section 3, and similar results can be extended to $\hat{\theta}^{(1)}_{\hat{\eta}}$ and $\hat{\theta}^{(3)}_{\hat{\eta}}$ using similar techniques. The $\hat{\theta}^{(2)}_{\hat{\eta}}$ case is much more challenging because, unlike $\hat{\theta}^{(1)}_{\hat{\eta}}$ and $\hat{\theta}^{(3)}_{\hat{\eta}}$, $\hat{\theta}^{(2)}_{\hat{\eta}}$ is not an average of $M$ independent terms conditional on the data.

The setting follows Section 3. Additionally, let $\mathcal{R}_1$ and $\mathcal{R}_2$ be independent collections of $M$ splits of the data with $K$ folds (uniformly at random). Let $\hat{\eta}_1$ and $\hat{\eta}_2$ be calculated with $\mathcal{R}_1$ and $\mathcal{R}_2$ respectively, which leads to analogous definitions of $\hat{\theta}_{\hat{\eta}_j}$, $\theta_{\hat{\eta}_j}$, and $\hat{\sigma}_{\hat{\eta}_j}$ for $j = 1, 2$. Under the null hypothesis and the conditions of Theorem 3.1, the t-statistic

$$t_{\hat{\eta}_j} = \frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_j}) - \tau)}{\hat{\sigma}_{\hat{\eta}_j}} \rightsquigarrow \mathcal{N}(0, 1),$$

where $\hat{\sigma}_{\hat{\eta}_j}$ is given as in (4.2), $\dot{h}(\theta)$ is a row vector with the partial derivatives of $h(\theta)$ evaluated at $\theta$, and $\hat{V}_{\hat{\eta}}$ is a plug-in estimator for $V_{\eta^*_P}$ defined in Appendix B.1. Based on this result, one can calculate p-values

$$p^{\pm}_j = 2\Phi\left(-\left|\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_j}) - \tau)}{\hat{\sigma}_{\hat{\eta}_j}}\right|\right),$$

$$p^{+}_j = \Phi\left(\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_j}) - \tau)}{\hat{\sigma}_{\hat{\eta}_j}}\right), \quad p^{-}_j = \Phi\left(-\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_j}) - \tau)}{\hat{\sigma}_{\hat{\eta}_j}}\right),$$

for $H_{0,\hat{\eta}_j} : h(\theta_{\hat{\eta}_j}) = \tau$ versus $H_{A,\hat{\eta}} : h(\theta_{\hat{\eta}}) \neq \tau$ and its one-sided versions.

The asymptotic regime assumes $M^{-1} n \sigma^2_D = O_P(1)$, where $\sigma^2_D$ (defined in (B.17) in the appendix) reflects the variance of $t_{\hat{\eta}_1}$ conditional on the data. Since $h(\hat{\theta}_{\hat{\eta}})$ and $\hat{\sigma}_{\hat{\eta}}$ converge to non-random quantities $h(\theta_{\eta^*_P})$ and $\sigma_{\eta^*_P}$ respectively, $\sigma^2_D \xrightarrow{P} 0$. Hence, the asymptotic regime requires $M \to \infty$ at a rate slower than $n$. The rate of convergence of $\sigma^2_D$ depends on the rate at which $\tilde{\eta} = \mathcal{A}(D)$ converges to $\eta^*_P$, and may be slow especially when $\tilde{\eta}$ is estimated nonparametrically. In Theorem 5.3, I show that a safe guideline for achieving the reproducibility guarantees established below is to choose $M$ of comparable magnitude to $n$.

I characterize below a central limit theorem for the difference of t-statistics constructed using different splits, which is the main ingredient for deriving my reproducibility measure in Theorem 5.2. Both results rely on the fairly technical Assumption B.4, stated in Appendix B.3.2. The key condition is a Donsker-type requirement on $\{\Psi_{\hat{\eta}_{\mathsf{s}}} : \mathsf{s} \subseteq [n]\}$ and $\psi_{\theta,\eta,i}\psi_{\theta,\eta,j}$. This condition holds, for example, if $\Theta'$ and $\psi_{\theta,\eta}$ are bounded and the cross products of the entries of $\psi_{\theta,\eta}$ are Lipschitz. Importantly, Assumption B.4 does not restrict the complexity of $\hat{\eta}$, it only restricts the complexity of the function classes over $\theta \in \Theta'$, and not over $\eta \in H$.

**Theorem 5.1.** *(Reproducibility of t-statistics based on Z-estimators)*
*Let Assumptions 3.1 and B.4 hold. Then, for any $\tau \in \mathbb{R}$,*

$$\left(\frac{\sqrt{n}\hat{\sigma}_D}{\sqrt{M}}\right)^{-1}\left(\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1})-\tau)}{\hat{\sigma}_{\hat{\eta}_1}} - \frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_2})-\tau)}{\hat{\sigma}_{\hat{\eta}_2}}\right) \rightsquigarrow \mathcal{N}(0,1)$$

*conditional on $D$ with probability approaching one.* □

I introduce my reproducibility measure for each of the three tests (two-sided and both one-sided tests), where $\Phi$ is the standard normal cdf, and formalize their guarantees in Theorem 5.2.

$$\hat{\delta}^{\pm}(\beta) = 2\Phi\left(-\left|\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1})-\tau)}{\hat{\sigma}_{\hat{\eta}_1}}\right| - \frac{\sqrt{n}\hat{\sigma}_D}{\sqrt{M}}\Phi^{-1}(\beta/2)\right) - 2\Phi\left(-\left|\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1})-\tau)}{\hat{\sigma}_{\hat{\eta}_1}}\right|\right),$$

$$\hat{\delta}^{+}(\beta) = \Phi\left(\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1})-\tau)}{\hat{\sigma}_{\hat{\eta}_1}} - \frac{\sqrt{n}\hat{\sigma}_D}{\sqrt{M}}\Phi^{-1}(\beta)\right) - \Phi\left(\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1})-\tau)}{\hat{\sigma}_{\hat{\eta}_1}}\right),$$

$$\hat{\delta}^{-}(\beta) = \Phi\left(-\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1})-\tau)}{\hat{\sigma}_{\hat{\eta}_1}} - \frac{\sqrt{n}\hat{\sigma}_D}{\sqrt{M}}\Phi^{-1}(\beta)\right) - \Phi\left(-\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1})-\tau)}{\hat{\sigma}_{\hat{\eta}_1}}\right).$$

**Theorem 5.2.** *(Reproducibility of p-values based on Z-estimators)*
*Let Assumptions 3.1 and B.4 hold, and $\tau \in \mathbb{R}$. For any $\beta \in (0, 0.5)$ and*

$$(p_j, \hat{\delta}(\beta)) \in \left\{(p_j^+, \hat{\delta}^+(\beta)), (p_j^-, \hat{\delta}^-(\beta)), (p_j^{\pm}, \hat{\delta}^{\pm}(\beta))\right\},$$

*it follows that*

$$P\left(p_2 > p_1 + \hat{\delta}(\beta) \,\middle|\, D\right) \leqslant \beta + o_P(1), \tag{5.1}$$

*with equality if $p_j \in \{p_j^+, p_j^-\}$.* □

Theorem 5.2 gives a novel measure of reproducibility for p-values based on split-sample Z-estimators. The guarantee of reproducibility in (5.1) is inspired by the definition of $(\xi, \beta)$-reproducibility in Ritzwoller and Romano (2023). They provide an algorithm for deciding how many repetitions $M$ of the sample-splitting procedure are necessary to guarantee reproducibility of the average across split-sample statistics. This covers, for example, the estimators $\hat{\theta}^{(1)}_{\hat{\eta}}$ and $\hat{\theta}^{(3)}_{\hat{\eta}}$. My approach complements theirs by focusing on reproducibility of inference, examining p-value rather than average statistics. My results hold for $\hat{\theta}^{(2)}_{\hat{\eta}}$, and the arguments can easily be extended to $\hat{\theta}^{(1)}_{\hat{\eta}}$ and $\hat{\theta}^{(3)}_{\hat{\eta}}$. Ritzwoller and Romano (2023)'s procedure takes as input the desired level of reproducibility, and outputs the required number of repetitions $M$ that guarantees such reproducibility. My approach takes $M$ as input (assumed "large"), and outputs a measure of how much reproducibility is guaranteed by such $M$. The asymptotic regimes also differ: Ritzwoller and Romano (2023) takes the data as fixed and considers that the desired threshold for the variability of the average split-sample statistic is small, while my framework considers $n$ and $M$ large.

The result in Theorem 5.2 relies on choosing $M$ such that $M^{-1}n\sigma_D^2 = O_P(1)$. In practice, it may be hard to choose $M$ that satisfies this condition since the rate at which $\sigma_D^2 \xrightarrow{P} 0$ is in general unknown. I show that if $M$ grows too fast, i.e., if $M^{-1}n\sigma_D^2 \xrightarrow{P} 0$, the distribution in Theorem 5.1 collapses and the guarantees in Theorem 5.2 hold conservatively. This gives a safe guideline for empirical implementation: choose $M$ to be at least a small fraction of $n$, such as $M = 0.1n$, and the guarantee in Theorem 5.2 will hold conservatively.

**Theorem 5.3.** *(Reproducibility under $M^{-1}n\sigma_D^2 \xrightarrow{P} 0$)*
*Let Assumptions 3.1 and B.4 hold, replacing B.4(v) with $M^{-1}n\sigma_D^2 \xrightarrow{P} 0$. Then, for any $\tau \in \mathbb{R}$,*

$$\left(\frac{\sqrt{n}\hat{\sigma}_D}{\sqrt{M}}\right)^{-1}\left(\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1}) - \tau)}{\hat{\sigma}_{\hat{\eta}_1}} - \frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_2}) - \tau)}{\hat{\sigma}_{\hat{\eta}_2}}\right) \xrightarrow{P} 0.$$

*For*

$$(p_j, \hat{\delta}(\beta)) \in \left\{(p_j^+, \hat{\delta}^+(\beta)), (p_j^-, \hat{\delta}^-(\beta)), (p_j^\pm, \hat{\delta}^\pm(\beta))\right\},$$

*and $\beta \in (0, 0.5)$,*

$$P\left(p_2 > p_1 + \hat{\delta}(\beta) \,\middle|\, D\right) \xrightarrow{P} 0.$$

□

# 6 Application 1: Poverty Prediction in Ghana

Understanding the drivers of poverty is at the root of much of Development Economics. For research, being able to better predict poverty dynamics is of first-order importance to both form hypotheses and then validate theories that explain poverty and poverty dynamics. For policy, accurate predictions of current or future poverty could enable better targeting of interventions (ideally then combined with causal inference on policies and interventions).

Using a sample of 319 households in urban Accra from the ISSER-Northwestern-Yale Long Term Ghana Socioeconomic Panel Survey (GSPS) (Osei et al., 2022), I examine how well I can predict which households will be below the poverty line 13 years ahead. The outcome of interest is an indicator for whether a household is below the poverty line in the fourth wave of GSPS (2022/2023), and I use covariates measured in wave 1 (2009/2010), that is, 13 years before. Of the 319 households, 22 were below the poverty line in wave 4 (around 7%). I use predictive covariates including household demographics, parental education, religion, political and traditional leadership experience, asset holdings, and financial indicators (see Appendix B.4 for details). Although I focus on the binary indicator of below the poverty line, the approach applies more broadly and could use other outcomes such as level of consumption or assets.

I estimate two quantities: the mean squared error (MSE) and the fraction in poverty by tercile of predicted probability of being below the poverty line. In both cases, I use repeated cross-fitting with $K = 3$ and $M = 200$, and fit random forest models using the `R` package `ranger` implemented through `mlr3`. Let $i \in \{1, \ldots, 319\}$, $Y_i$ denote the indicator of whether household $i$ is below the poverty line in wave 4 of GSPS and $X_i$ the set of covariates measured in wave 1. The estimated MSE is given by

$$\hat{\theta}_{\hat{\eta},\text{MSE}} = \frac{1}{M} \sum_{r \in \mathcal{R}} \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - \hat{\eta}_{\tilde{\mathsf{s}}}(X_i) \right)^2 .$$

For $j \in \{1, 2\}$, let $\hat{t}_{j,\tilde{\mathsf{s}}}$ be the first and second terciles of $(\hat{\eta}_{\tilde{\mathsf{s}}}(X_i))_{i=1}^n$, that is,

$$\hat{t}_{j,\tilde{\mathsf{s}}} = \inf \left\{ t : \frac{1}{|\mathsf{s}|} \sum_{i \in \mathsf{s}} \mathbf{1}\{\hat{\eta}_{\tilde{\mathsf{s}}}(X_i) \leqslant t\} \geqslant \frac{j}{3} \right\},$$

and let $\hat{t}_{0,\tilde{\mathsf{s}}} = -\infty$, $\hat{t}_{3,\tilde{\mathsf{s}}} = \infty$. For $j \in \{1, 2, 3\}$, the fraction in poverty in tercile $j$ of predicted probability of being below the poverty line is given by

$$\hat{\theta}_{\hat{\eta},\text{Frac}j} = \frac{1}{MK} \sum_{r \in \mathcal{R}} \sum_{\mathsf{s} \in r} \frac{\sum_{i \in \mathsf{s}} Y_i \mathbb{I}\left( \hat{t}_{j-1,\tilde{\mathsf{s}}} < \hat{\eta}_{\tilde{\mathsf{s}}}(X_i) \leqslant \hat{t}_{j,\tilde{\mathsf{s}}} \right)}{\sum_{i \in \mathsf{s}} \mathbb{I}\left( \hat{t}_{j-1,\tilde{\mathsf{s}}} < \hat{\eta}_{\tilde{\mathsf{s}}}(X_i) \leqslant \hat{t}_{j,\tilde{\mathsf{s}}} \right)}.$$

I show in Appendix B.4 that $\hat{\theta}_{\hat{\eta},\text{Frac}j}$ is a Z-estimator.

I also compare the MSE of the models estimated with random forests to the MSE of using the sample average, as described in Section 4.2. In particular, I report p-values for the test of Section 4.2.1.

I calculate the MSE estimators and the one-sided test both in the real data and in two Monte Carlo designs, described in Appendix B.4. The data generating processes are designed to be similar to the original dataset, preserving the empirical marginals and rank-based dependence structure of the observed data. In the first design, denoted *Correlated*, the outcome $Y$ is correlated to the covariates $X$. In the second design, denoted *Uncorrelated*, the outcome is independent of the covariates. I run around 5,000 Monte Carlo iterations for each of the three designs – real data, "correlated" and "uncorrelated" simulated data –, drawing 200 new random splits of the sample at each Monte Carlo iteration. For the real data, the only source of randomness are the 200 splits, while for the simulation designs I draw a new dataset at each iteration (with 200 splits for each dataset). For each simulated dataset and split, I also calculate the difference between top and bottom terciles, $\hat{\theta}_{\hat{\eta},\text{Frac3}} - \hat{\theta}_{\hat{\eta},\text{Frac1}}$.

I compare the estimates and p-values of using repeated cross-fitting (RCF) with three alternatives. The first is the standard "twice the median" (TTM) rule (Rüger, 1978; Gasparin et al., 2025; Chernozhukov et al., 2025b): calculate the p-value (for difference in MSE or "top minus bottom" estimator) separately for each fold, that is, using a third of the data, take the median of the 600 p-values (200 repetitions, 3 folds) and multiply it by 2. The second is the Sequential Aggregation (Seq) approach of Luedtke and Van Der Laan (2016) and Wager (2024): train a random forest using only fold 1, compute the t-statistic using fold 2, then train a random forest using folds 1 and 2 and compute the t-statistic in fold 3. The p-value for each repetition of cross-fitting uses as final t-statistic $\sqrt{2}$ times the average of the two t-statistics. Finally, the final p-value for each Monte Carlo iteration is twice the median over the 200 p-values coming from the 200 repetitions, similar to Chernozhukov et al. (2025a). The third method is standard sample-splitting (SS): train a random forest using two thirds of the data, calculate p-value in the excluded third, not aggregating across repetitions.

Figure 1 presents the p-values for whether random forest MSE is lower than sample average MSE, and accuracy ($1 - MSE$) point estimates across Monte Carlo iterations for the two simulation designs as well as for the real data. In the uncorrelated design, all methods exhibit similar accuracy on average, with sample-splitting having larger variance since it does not aggregate across multiple splits. All methods are conservative: the p-values concentrate around 1. For sample-splitting, this happens since the sample average is the best predictor of $Y$ in this design, and the random forests are noisy estimates that have larger MSE. The other methods are conservative for the same reason, and TTM and Seq are more conservative since they take twice

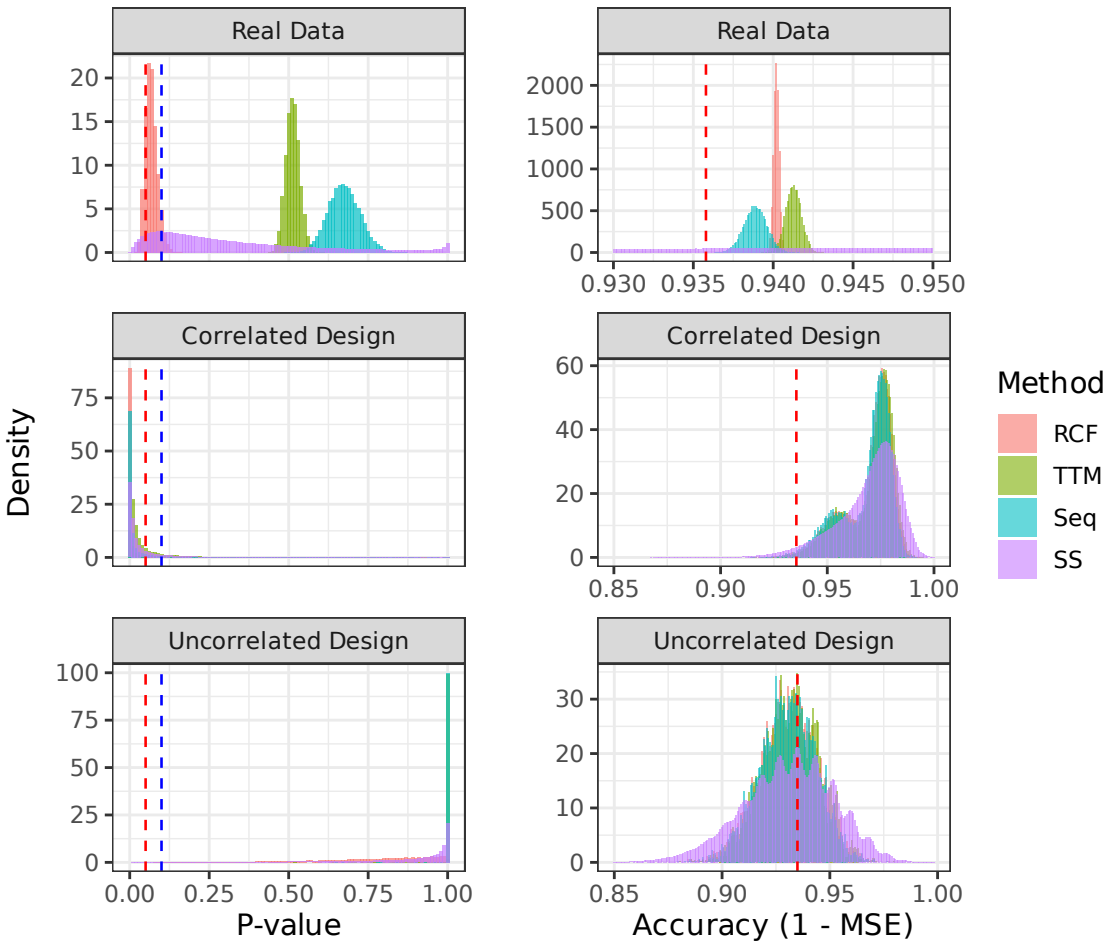


Figure 1: Accuracy Comparison Across Methods and Datasets

Notes: Left panels show distribution across Monte Carlo iterations of p-values for testing whether random forest MSE is lower than sample average MSE. Vertical red and blue lines are respectively 0.05 and 0.10. Right panels show distribution of accuracy $(1 - MSE)$ of the random forest, and vertical red lines are the accuracy of the sample average. Rows show results for real data (top), simulations from correlated design (middle), and simulations from uncorrelated design (bottom). Methods: RCF (repeated cross-fitting), TTM (twice-the-median), Seq (sequential aggregation), SS (standard sample-splitting). Top-right panel excludes sample-splitting observations higher than 0.95 or smaller than 0.93 to improve visualization. The number of iterations for the real dataset, correlated design and uncorrelated design are, respectively, 11841, 28335, and 7485. SS uses the same number of iterations, multiplied by 600 (200 repetitions, 3 folds).

the median p-value, which guards against the worst DGP. For the correlated design, all methods correctly give small p-values, with RCF being more concentrated around

zero.

In the real dataset, Seq often has the smallest accuracy, and TTM the highest, while RCF stands in between. Seq has smaller accuracy since one the two models that it averages over is trained with only a third of the data. RCF and TTM, on the other hand, always use two thirds of the data for training. The only difference between the two numbers is that RCF averages model performances over 200 repetitions while TTM takes the median. Hence, the higher accuracy of TTM reflects the distribution of model accuracies being left-skewed. Only RCF manages to consistently reject the null, concluding that poverty can be predicted from the observed covariates using a random forest model. TTM and Seq are more conservative, with Seq having larger p-values than TTM due to its lower accuracy.

A comparison similar to Figure 1 for the top minus bottom estimator $\hat{\theta}_{\hat{\eta},\text{Frac3}} - \hat{\theta}_{\hat{\eta},\text{Frac1}}$ is presented in Figure 5.

Table 3: Poverty Prediction by Tercile in Real Dataset

| Method | Variable | Estimate | 95% CI | p-value |
|---|---|---|---|---|
| RCF | Bottom tercile | 0.046 | [0.007, 0.085] | |
| | Top tercile | 0.122 | [0.059, 0.184] | |
| | Top minus bottom | 0.076 | [0.002, 0.150] | 0.023 |
| TTM | Bottom tercile | 0.056 | [−0.021, 0.133] | |
| | Top tercile | 0.114 | [0.006, 0.223] | |
| | Top minus bottom | 0.083 | [−0.052, 0.208] | 0.228 |
| Seq | Top minus bottom | – | – | 0.150 |

Notes: Estimates of fraction below poverty line by tercile of predicted probability of being below the poverty line. Fraction below poverty line in entire sample is around 7%. Bottom tercile corresponds to $\hat{\theta}_{\hat{\eta},\text{Frac1}}$ and top tercile to $\hat{\theta}_{\hat{\eta},\text{Frac3}}$. RCF, TTM and Seq correspond respectively to repeated cross-fitting, twice-the-median, and sequential aggregation. All estimates aggregate over 7,104,600 splits.

Table 3 shows the point estimates and CIs for the estimators $\hat{\theta}_{\hat{\eta},\text{Frac1}}$, $\hat{\theta}_{\hat{\eta},\text{Frac3}}$, and their difference using RCF and TTM in the real dataset, as well as p-values for testing whether the difference between top and bottom groups is positive (that is, top tercile has a larger fraction below the poverty line than the bottom tercile). These final estimates aggregate over all the 7,104,600 splits displayed in Figure 1, averaging for RCF and taking the median for TTM. I do not display the point estimates for Seq since Wager (2024) focuses on testing, but the p-value indicates that the difference between top and bottom groups is not significant. Table 3 shows that the difference between top and bottom terciles is statistically significant only for RCF.

# 7 Application 2: Heterogeneous Treatment Effects in Charitable Giving

There has been growing interest in the literature for learning features of heterogeneous treatment effects using machine learning (Chernozhukov et al., 2025b; Wager, 2024; Imai and Li, 2025; for applications, see, e.g., Bryan et al., 2024; Athey et al., 2025; Johnson et al., 2023). I revisit the Generic Machine Learning framework of Chernozhukov et al. (2025b) (henceforth CDDF), and propose a new *ensemble* estimator that uses the entire sample for calculating confidence intervals, more data for training machine learning algorithms, and aggregates predictions over multiple ML predictors into an ensemble. I first revisit CDDF's approach, and second introduce my ensemble estimator. Theoretical properties are delayed to Appendix B.5. Finally, I compare my estimator to the approaches of CDDF and of Wager (2024) in a Monte Carlo design and in an empirical application using data from Karlan and List (2007). The simulation exercise shows gains in power using the ensemble method, and the ensemble approach is the only to detect statistically significant treatment effect heterogeneity in the empirical application.

## 7.1 The Generic ML Approach of Chernozhukov et al. (2025b)

CDDF proposed a method for learning features of treatment effect heterogeneity in randomized trials. In this section, I focus on their Sorted Group Average Treatment Effects (GATES) estimand. This approach consists of using a machine learning (ML) algorithm and pre-treatment covariates to find groups of individuals with larger and smaller average treatment effects (ATEs). If such groups exist, this means that treatment effect is heterogeneous and that this heterogeneity can be explained at least in part by observable characteristics. Moreover, one can explore how these groups differ in terms of these characteristics. They call this last step Classification Analysis (CLAN), and although I focus on GATES to simplify exposition, my results also hold for CLAN.

First, I define some notation. Let $D = (Y_i, T_i, X_i)_{i=1}^n$ denote the data, where $Y$ is a scalar outcome, $T$ is the treatment assignment indicator, and $X$ is a vector of pre-treatment covariates. I assume that $(Y_i, T_i, X_i)$ are drawn i.i.d. from a distribution $P \in \mathcal{P}$. Let $\mathcal{A}$ denote an ML algorithm, a function that takes a dataset as input, and outputs an estimate of the Conditional Average Treatment Effect (CATE) function,

$$\eta_P(x) = \mathbb{E}_P\left[Y(1) - Y(0) | X = x\right].$$

For example, $\mathcal{A}$ could be Causal Forests (Wager and Athey, 2018), or based on

Random Forests, Neural Networks, or Gradient Boosting.[2] For any subsample $\mathsf{s} \subseteq \{1, \ldots, n\}$, let $D_{\mathsf{s}} = \{Y_i, T_i, X_i\}_{i \in s}$, $\tilde{\mathsf{s}} = \{1, \ldots, n\} \setminus \mathsf{s}$, and $\hat{\eta}_{\tilde{\mathsf{s}}} = \mathcal{A}(D_{\tilde{\mathsf{s}}})$, that is, $\hat{\eta}_{\tilde{\mathsf{s}}}$ is the model trained with algorithm $\mathcal{A}$ using the subsample $D_{\tilde{\mathsf{s}}}$.

The procedure is given as follows. First, take $M$ random subsets of $\{1, \ldots, n\}$ of size $\pi n$. For each $m = 1, \ldots, M$, denote the subsample by $\mathsf{s}_m$, where $\mathsf{s}_m \subseteq \{1, \ldots, n\}$ and $|\mathsf{s}_m| = \pi n$. For each repetition $m$, call $\mathsf{s}_m$ the main sample, and $\tilde{\mathsf{s}}_m = \{1, \ldots, n\} \setminus \mathsf{s}_m$ the auxiliary sample. For $m = 1, \ldots, M$, train the model

$$\hat{\eta}_{\tilde{\mathsf{s}}_m} = \mathcal{A}(D_{\tilde{\mathsf{s}}_m}) \tag{7.1}$$

using data from the auxiliary sample. In the main sample, calculate predicted individual treatment effects (ITEs) $\hat{\tau}_i = \hat{\eta}_{\tilde{\mathsf{s}}_m}(X_i)$. Sort $(\hat{\tau}_i)_{i \in \mathsf{s}}$ into $J$ quantile groups $G_1, \ldots, G_J$, where

$$G_j = \{i \in \{1, \ldots, n\} : \hat{\tau}_i \in I_j\}, \tag{7.2}$$

with $I_j = [\hat{d}_{j-1}, \hat{d}_j)$, $-\infty = \hat{d}_0 < \hat{d}_1 < \cdots < \hat{d}_J = \infty$, and $(\hat{d}_j)_{j=0}^J$ are calculated such that the number of observations in $(G_j)_{j=1}^J$ is balanced or nearly balanced. For example, with $J = 4$, $(G_j)_{j=1}^J$ is a partition of the sample into quartiles of $(\hat{\tau}_i)_{i \in s}$. Calculate the split-specific GATES estimator by running the weighted regression

$$Y_i = \alpha Z_i + \sum_{j=1}^{J} \gamma_j^{(m)} \left[T_i - p(X_i)\right] \mathbb{I}\left(i \in G_j\right) + \varepsilon_i, \qquad i \in \mathsf{s}_m, \tag{7.3}$$

with weights $\omega_i = \{p(X_i)\left[1 - p(X_i)\right]\}^{-1}$, where $p(x) = P(T = 1 | X = x)$ is the (known) propensity score. These weights guarantee correct identification of ATEs when the propensity score is not constant, that is, it ensures

$$\gamma_j^{(m)} = \mathbb{E}_P\left[Y_i(1) - Y_i(0) | i \in G_j\right].$$

Denote the estimates by $(\hat{\gamma}_j^{(m)})_{j=1}^J$. A frequent parameter of interest is

$$\delta^{(m)} = \gamma_J^{(m)} - \gamma_1^{(m)},$$

the difference in ATEs between the top and bottom groups of predicted ITEs. This parameter can be estimated with the analogue

$$\hat{\delta}^{(m)} = \hat{\gamma}_J^{(m)} - \hat{\gamma}_1^{(m)},$$

[2]For example, one could use any of these three algorithms to estimate separately the functions $\mathbb{E}_P\left[Y(1) | X = x\right]$ and $\mathbb{E}_P\left[Y(0) | X = x\right]$, and use the difference of the two estimated functions as an estimate of the CATE.

and a CI can be calculated as usual,

$$(L^{(m)}, U^{(m)}) = (\hat{\delta}^{(m)} - z_{1-\alpha/2}\hat{\sigma}^{(m)}/\sqrt{\pi n}, \hat{\delta}^{(m)} + z_{1-\alpha/2}\hat{\sigma}^{(m)}/\sqrt{\pi n}), \tag{7.4}$$

where $\hat{\sigma}^{(m)}/\sqrt{\pi n}$ is a heteroscedasticity-robust standard error for $\hat{\delta}^{(m)}$ calculated as usual from the OLS regression (7.3), and $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. Finally, the final estimators and CIs are given by

$$\hat{\delta} = \text{Med}(\hat{\delta}^{(m)})$$

and

$$(L, U) = \left(\text{Med}(L^{(m)}), \text{Med}(U^{(m)})\right),$$

where Med denotes the median across repetitions $m$. Conditions for the validity of this CI are established in Theorem 4.3 of CDDF.

This approach carries a tradeoff that's not present in my method, and it considers a single ML algorithm $\mathcal{A}$. The tradeoff regards the choice of $\pi$: a larger $\pi$ means more data is used to estimate the regression (7.3), leading to narrower CIs in (7.4); but fewer data are used to train the ML model in (7.1), likely yielding a worse estimate of the CATE. Moreover, regularity condition R3 in CDDF requires $\pi$ to be relatively small to guarantee that the CI $[L, U]$ covers the median of $\delta^{(m)}$ across all possible splits. My ensemble approach presented next avoids this tradeoff since it uses the entire sample for estimation and a larger sample for training. The ensemble estimator also incorporates more than one ML algorithm, which is important if one does not want to commit beforehand to any specific algorithm. Although CDDF's approach can be repeated with different algorithms, that comes with potential issues of multiple hypothesis testing.

In the next subsection I propose a new GATES estimator that (i) uses the entire sample to calculate $(\hat{\gamma}_j)_{j=1}^J$ in (7.3), and (ii) combines predictions from multiple ML algorithms to form an *ensemble*, eliminating the need for algorithm selection.

## 7.2 An Ensemble Estimator

Before defining my ensemble estimator, I introduce some additional notation. Theoretical properties are delayed to Appendix B.5. Let $A$ denote the number of machine learning algorithms that will be used for predicting ITEs. For $a = 1, \ldots, A$, let $\mathcal{A}_a$ denote an ML algorithm, that is, a function that takes a dataset as input, and outputs an estimate of the CATE. For example, one could choose $\mathcal{A}_1$ to use Random Forests, $\mathcal{A}_2$ Neural Nets, and $\mathcal{A}_3$ Gradient Boosting. For $\mathsf{s} \subseteq [1, \ldots, n]$ and $a = 1, \ldots, A$, let

$$\hat{\eta}_{\mathsf{s},a} = \mathcal{A}_a(D_{\mathsf{s}}),$$

that is, $\hat{\eta}_{\mathsf{s},a}$ is the model trained with algorithm $\mathcal{A}_a$ using the subsample $D_{\mathsf{s}}$.

The ensemble approach is summarized in Algorithm 1. The first difference is that instead of splitting the sample into two sets, I split it into $K$ roughly equal-sized folds $(\mathsf{s}_k)_{k=1}^K$, again repeating the process $M$ times. I calculate $A$ predicted ITEs for each individual using the $A$ ML algorithms, trained using all folds except the one that contains observation $i$. I denote the predicted ITEs by $\hat{\tau}_{i,a} = \hat{\eta}_{\tilde{\mathsf{s}}_{k(i)}}(X_i)$, where $k(i)$ is such that $i \in \mathsf{s}_{k(i)}$. Then, to calibrate the weights for combining the multiple ML predictions into one, I split the sample again into $L$ different folds, for each repetition $m = 1, \ldots, M$. Let $\{\mathsf{s}'_\ell\}_{\ell=1}^L$ denote the $L$ folds ($m$ is not incorporated in the notation to simplify exposition). For $\ell = 1, \ldots, L$, estimate the weighted regression

$$Y_i = \alpha_1 + \sum_{a=1}^{A} \beta_a(\hat{\tau}_{i,a} - \bar{\tau}_a)\left[T_i - p(X_i)\right] + \alpha_2 Z_i + \varepsilon_i, \qquad i \in \tilde{\mathsf{s}}'_\ell, \tag{7.5}$$

with weights $\omega_i = \{p(X_i)\left[1 - p(X_i)\right]\}^{-1}$. In (7.5), $\bar{\tau}_a = \frac{1}{n - |\mathsf{s}'_\ell|} \sum_{i \notin \mathsf{s}'_\ell} \hat{\tau}_{i,a}$, $p(X_i)$ is the propensity score, and $Z_i$ is a vector of functions of $X_i$, for example $Z_i = (X_{1,i}, p(X_i))'$, where $X_{1,i}$ is a subset of $X_i$. The role of $Z_i$ is only reducing noise in estimation, so this term can be omitted if desired. Denote the estimates of $(\beta_{\ell,a})_{a=1}^A$ by $(\hat{\beta}_{\ell,a})_{a=1}^A$. The final predicted ITE is then given by

$$\hat{\tau}_i = \sum_{a=1}^{A} \hat{\beta}_{\ell,a} \hat{\tau}_{i,a}, i \in \mathsf{s}'_\ell.$$

Repeating this process for $\ell = 1, \ldots, L$ gives $\hat{\tau}_i$ for every observation. I sort $(\hat{\tau}_i)_{i \in \mathsf{s}}$ into groups separately by fold. That is, for $k = 1, \ldots, K$,

$$G_{j,k} = \left\{i \in \mathsf{s}_k : \hat{\tau}_i \in I_{j,k}\right\},$$

with $I_{j,k} = [\hat{d}_{j-1,k}, \hat{d}_{j,k})$, $-\infty = \hat{d}_{0,k} < \hat{d}_{1,k} < \cdots < \hat{d}_{J,k} = \infty$, and $(\hat{d}_{j,k})_{j=0}^J$ are calculated such that the number of observations in $(G_{j,k})_{j=1}^J$ is balanced or nearly balanced. Finally, the split-specific GATES estimator uses the whole sample, defining

$$G_j = \bigcup_{k=1}^{K} G_{j,k}, \tag{7.6}$$

and running the weighted regression

$$Y_i = \alpha Z_i + \sum_{j=1}^{J} \gamma_j^{(m)} \left[T_i - p(X_i)\right] \mathbb{I}\left(i \in G_j\right) + \varepsilon_i, \qquad i \in \{1, \ldots, n\}, \tag{7.7}$$

**Algorithm 1** Ensemble Method for GATES

**Input:** Dataset $D = (Y_i, T_i, X_i)_{i=1}^n$, ML algorithms $(\mathcal{A}_a)_{a=1}^A$, repetitions $M$, number of folds $K$ (training) and $L$ (calibration), number of groups $J$.
**Output:** GATES estimates $(\hat{\gamma}_j)_{j=1}^J$ and standard errors $(\hat{\sigma}_j)_{j=1}^J$

**for** $m = 1, \ldots, M$ **do**
  **Train ML models:** Split $D$ into $K$ folds $(\mathsf{s}_k)_{k=1}^K$
  **for** $k = 1, \ldots, K$ and $a = 1, \ldots, A$ **do**
    Train $\hat{\eta}_{\bar{\mathsf{s}}_k,a} = \mathcal{A}_a(D_{\bar{\mathsf{s}}_k})$; compute $\hat{\tau}_{i,a} = \hat{\eta}_{\bar{\mathsf{s}}_k,a}(X_i)$ for $i \in \mathsf{s}_k$
  **end for**
  **Calibrate ensemble:** Split $D$ into $L$ different folds $(\mathsf{s}'_\ell)_{\ell=1}^L$
  **for** $\ell = 1, \ldots, L$ **do**
    Estimate $(\hat{\beta}_{\ell,a})_{a=1}^A$ using $D_{\bar{\mathsf{s}}'_\ell}$ as in (7.5)
    Compute $\hat{\tau}_i = \sum_{a=1}^A \hat{\beta}_{\ell,a}\hat{\tau}_{i,a}$ for $i \in \mathsf{s}'_\ell$
  **end for**
  **Compute GATES:** Sort $(\hat{\tau}_i)_{i=1}^n$ into $(G_j)_{j=1}^J$ as in (7.6)
  Estimate $(\hat{\gamma}_j^{(m)}, \hat{\sigma}_j^{(m)})_{j=1}^J$ with (7.7)
**end for**
Compute: $(\hat{\gamma}_j)_{j=1}^J = \frac{1}{M}\sum_{m=1}^M (\hat{\gamma}_j^{(m)})_{j=1}^J$, $(\hat{\sigma}_j)_{j=1}^J = \frac{1}{M}\sum_{m=1}^M (\hat{\sigma}_j^{(m)})_{j=1}^J$
**return** $(\hat{\gamma}_j, \hat{\sigma}_j)_{j=1}^J$

with weights $\omega_i = \{p(X_i)\,[1 - p(X_i)]\}^{-1}$.

(7.5) is very close to the Best Linear Predictor (BLP) regression of CDDF, except that it uses the $A$ predicted ITEs instead of just one. The intuition behind (7.5) is that $(\beta_a)_{a=1}^A$ are the best linear predictor coefficients of a regression where the true CATE $\eta_P(X_i)$ is the response variable, and $(\hat{\tau}_{i,a})_{a=1}^A$ are the independent variables (see Theorem 3.1 of CDDF). Hence, $\sum_{a=1}^A \beta_a\hat{\tau}_{i,a}$ is the best linear approximation of $\eta_P(X_i)$ given $(\hat{\tau}_{i,a})_{a=1}^A$.

The final estimator averages over repetitions,

$$\hat{\delta}_{\hat{\eta}} = \hat{\delta} = \frac{1}{M}\sum_{m=1}^M \hat{\delta}^{(m)},$$

where, as before, $\hat{\delta}^{(m)} = \hat{\gamma}_J^{(m)} - \hat{\gamma}_1^{(m)}$, with $\hat{\gamma}_J^{(m)}$ and $\hat{\gamma}_1^{(m)}$ being the estimates from (7.7). The final standard error is

$$\hat{\sigma}_{\hat{\eta}} = \hat{\sigma} = \frac{1}{M}\sum_{m=1}^M \frac{\hat{\sigma}^{(m)}}{\sqrt{n}}, \tag{7.8}$$

where $\hat{\sigma}^{(m)}/\sqrt{n}$ is a heteroscedasticity-robust standard error for $\hat{\delta}^{(m)}$ calculated as

usual from the OLS regression (7.7). The parameter of interest is

$$\delta_{\hat{\eta}} = \delta = \frac{1}{M} \sum_{m=1}^{M} \gamma_J^{(m)} - \gamma_1^{(m)},$$

where $\gamma_J^{(m)}$ and $\gamma_1^{(m)}$ are defined in (7.7).

## 7.3 Application to Charitable Giving and Monte Carlo Experiments

I compare my new ensemble approach to two alternative methods in an empirical application and in Monte Carlo experiments. I revisit Karlan and List (2007), which sent fundraising letters to prior donors of a liberal nonprofit organization in the United States, randomizing the match ratio offered (1:1, 2:1, or 3:1) versus no match for a control group. I pool all match treatments into a single treatment group, focusing on the binary treatment of receiving any match offer versus none. The outcome of interest is the amount donated in dollars. The predictive covariates I use include individual donation history (frequency, recency, amount), gender, state-level political variables (Bush vote share, count of court cases in which the organization was either a party to or filed a brief), and zip code-level demographics and economics (race, age, household size, income, homeownership, education, urbanization) (see Appendix B.5 for details). I focus on the subset of 6,419 donors who donated within the last two months, as they were more responsive to the solicitation and the smaller sample facilitates computation of the Monte Carlo experiments.

I compare the ensemble with CDDF's approach, described in Section 7.1, and the sequential aggregation approach of Luedtke and Van Der Laan (2016), Wager (2024), and Chernozhukov et al. (2025a). Sequential aggregation (Seq) consists of splitting the sample into $K$ folds, for $k = 2, \ldots, K$ train an ML model using folds 1 through $k-1$, and compute GATES in the $K$-th fold. The final estimator is the average over the $K-1$ estimates, and the p-value uses the final t-statistic equal to $\sqrt{K-1}$ times the average of the fold-specific t-statistics. This approach uses more data for calculating GATES and p-values ($n(K-1)/K$ observations), but trains some ML models using fewer data (the first model uses $n/K$ observations). I aggregate the final estimates and p-values taking the median over $M$ repetitions as in Chernozhukov et al. (2025a).

I compute the three approaches across four designs: (i) using the real data (real), (ii) using the real data but shuffling the treatment assignment indicator at random (so there is no treatment effect heterogeneity) (real-shuffled), (iii) drawing from a DGP where treatment effect is partially predictable using covariates (mc-hte), (iv)

drawing from a DGP where treatment effect heterogeneity is independent of covariates (mc-nohte). The two DGPs are meant to be similar to the real data, preserving the marginal distributions of covariates and rank-correlation structure, as described in Appendix B.5. Across all methods and datasets, at each Monte Carlo iteration I use 100 repetitions of sample-splitting, take random samples (without replacement) of sizes $n = 500, 1000, 2000, 6419$ (entire dataset), and compare the number of folds $K = 2, 3, 5, 10$ (for CDDF, the ML is trained with $n(K-1)/K$ observations and GATES calculated in the remaining sample). For Ensemble, I draw at random between 1 and 4 ML algorithms among 10 popular algorithms available in `R`'s `mlr3verse`: XGBoost (`xgboost`), Random Forest (`ranger`), Neural Networks (`nnet`), Elastic Net (`glmnet`), k-Nearest Neighbors (`kknn`), Linear Regression (`lm`), Decision Trees (`rpart`), Fast Nearest Neighbors (`fnn`), Multivariate Adaptive Regression Splines (`earth`), and Gradient Boosting (`gbm`). For CDDF and Seq, I draw one of the same ten algorithms at random, for each Monte Carlo iteration. I show the number of iterations used for each specification in Table 4 in the appendix.

Figure 2 shows the gains in power of using the ensemble method in the real dataset. It displays boxplots of one-sided p-values for testing whether the top tercile of predicted treatment effects has a larger ATE than the bottom tercile. A small p-value means rejecting the null hypothesis of no detectable treatment effect heterogeneity. With $n = 6419$ (the entire dataset), Ensemble with 4 algorithms detects treatment effect heterogeneity at the 10% level in more than 75% of the iterations. Seq and CDDF give p-values above 10% in most iterations. None of the methods are powered enough to reject the null consistently with $n = 2000$.

Figure 3 is similar to Figure 2, except that it uses the synthetic DGP where there is no detectable heterogeneity. It shows that all methods correctly fail to reject the null in most iterations. Similar figures for designs real-shuffled and mc-hte are presented in Appendix B.5.

Figure 4 shows the rejection probabilities at the 5% significance level, that is, the percentage of iterations with p-value below 5%. For the two datasets with no detectable heterogeneity, real-shuffled and mc-nohte, all methods are conservative when $K = 2$ or $K = 3$, they yield rejection probabilities below the nominal level. In the real-shuffle design with $n = 6419$ and $K = 5$ or $K = 10$, the ensemble methods reject the null with probability slightly higher than nominal, but smaller than 10%. With $n = 2000$, only Ensemble 4 rejects the null with probability higher than nominal with $K \geqslant 5$ in the real-shuffled design. In the real dataset, CDDF almost never detects HTE, and Seq detects in less than 20% of iterations with $K = 10$ and $n = 6419$. The ensemble methods have higher power especially in the specifications using the entire dataset. For example, Ensemble 2 detects heterogeneity in around 50% of iterations with $K = 3$ folds. In the synthetic dataset where there is detectable heterogeneity, mc-hte, as well as in the real data, Ensemble 2 and 4 have higher power across all

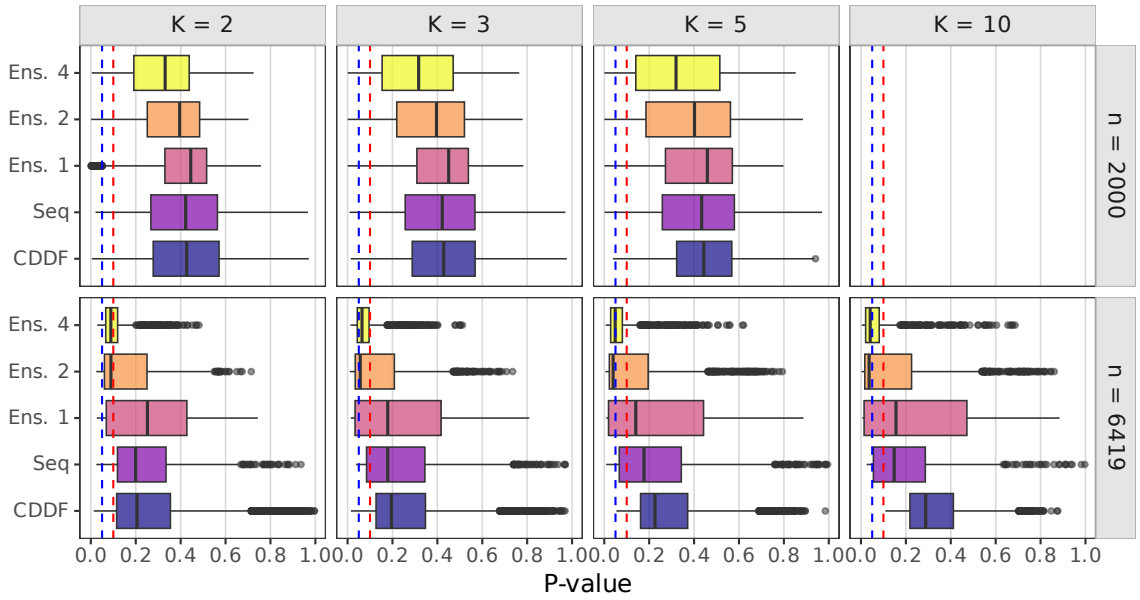


Figure 2: Distribution of p-values for Top - Bottom GATES Groups – Real Dataset

Notes: Distribution of one-sided p-values for testing whether the top tercile has a larger ATE than the bottom tercile across Monte Carlo iterations using the real dataset. Rows show different sample sizes ($n = 2000, 6419$), columns show different numbers of folds ($K = 2, 3, 5, 10$). Each box represents the distribution across Monte Carlo iterations with 100 repetitions of sample-splitting per iteration. Sources of randomness are the subsample when $n = 2000$, which ML algorithms are used, and how the data are split. Red dashed line at 0.1, blue dashed line at 0.05. Specifications with $K = 10, n = 2000$ are excluded.

specifications.

As I discuss in Appendix B.5, the rejection probability under the null of no detectable heterogeneity could in principle be above the nominal level when using the normal approximation CI. In Appendix B.5.5, I propose an alternative CI that controls size under the null, at the expense of being more conservative and requiring more computational time. However, I note that extensive simulation experiments, including but not limited to the design of Figure 4, suggest that Ensemble 4 is conservative for relatively small values of $K$. Hence, my recommendation for empirical practice is to use the normal approximation CI with Ensemble 4 and $K = 3$.

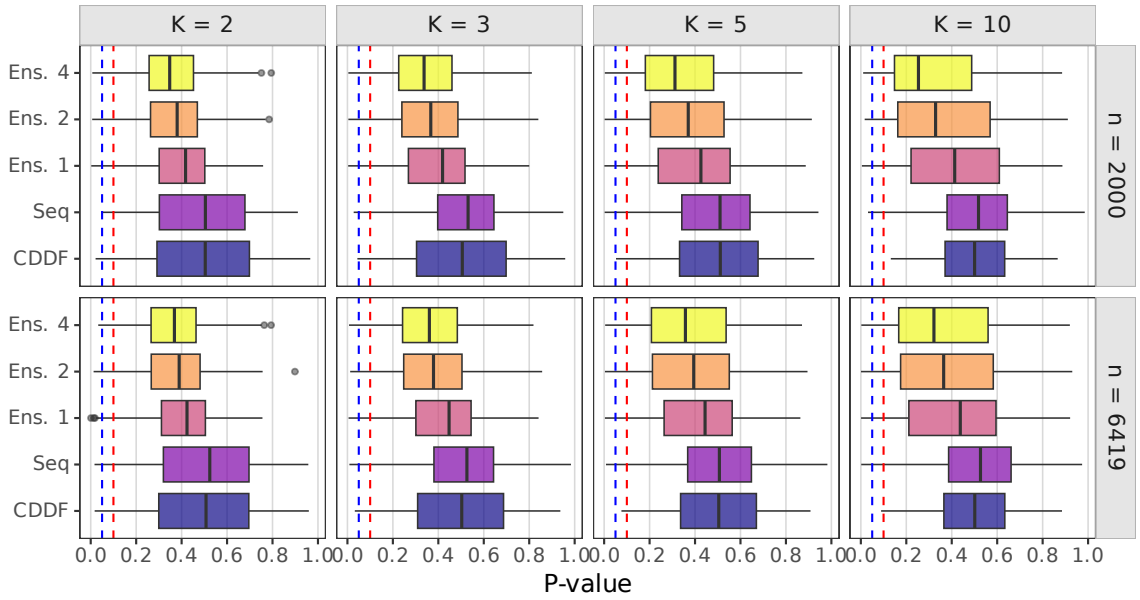


Figure 3: Distribution of p-values for Top - Bottom GATES Groups – Synthetic DGP with no Heterogeneity

Notes: Distribution of one-sided p-values for testing whether the top tercile has a larger ATE than the bottom tercile across Monte Carlo iterations using the real dataset. Rows show different sample sizes ($n = 2000, 6419$), columns show different numbers of folds ($K = 2, 3, 5, 10$). Ens. 1, Ens. 2, and Ens. 4 represent the Ensemble method using respectively 1, 2, and 4 algorithms. Each box represents the distribution across Monte Carlo iterations with 100 repetitions of sample-splitting per iteration. Boxplots show the median (center line), interquartile range (box), and whiskers extending to 1.5 times the IQR, with points beyond shown as outliers. Data is generated from a synthetic DGP where there is no explainable treatment effect heterogeneity (Appendix B.5). Red dashed line at 0.1, blue dashed line at 0.05. Specifications with $K = 10, n = 2000$ are excluded.

# 8 Conclusion

As predictive algorithms become increasingly popular, using the same dataset to both train and test a new model has become routine across research, policy, and industry. I derived a new inference approach on model properties that averages across several splits of the sample, where at each split one part is used to train a model and the remaining to evaluate it. Compared to a standard 50-50 sample-splitting, my approach improves statistical and modeling power by using more data for training and evaluating, and improves reproducibility, so two researchers using different splits are more likely to reach the same conclusion about statistical significance. Although

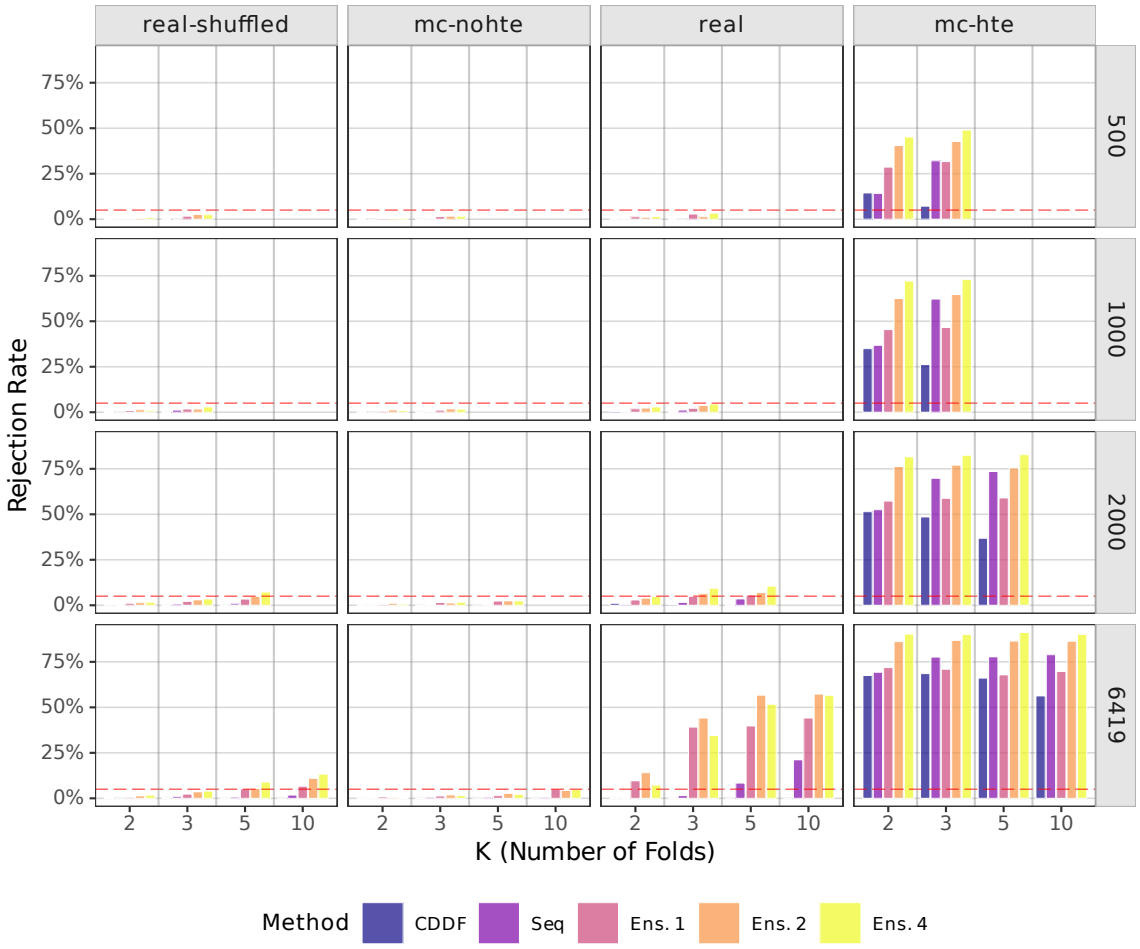


Figure 4: Rejection probabilities for Top - Bottom GATES Groups at 5% Significance Level

Notes: Percentage of Monte Carlo iterations with p-value below 5% for testing whether the top tercile has a larger ATE than the bottom tercile. Rows show different sample sizes ($n = 500, 1000, 2000, 6419$), columns show what simulation design is used. Specifications with $K \geqslant 5, n \leqslant 1000$ and $K = 10, n = 2000$ are excluded.

the practice of averaging over multiple splits is not new, the confidence intervals and establishing their validity appears to be new.

I addressed the main technical challenge, the dependence created by reusing observations across splits, by proving a central limit theorem for the large class of split-sample Z-estimators. Leveraging the data-dependent parameter of interest, my

CLT does not require restricting the complexity of the model or its convergence rate, unlike in the classic semiparametrics problem that used cross-fitting and focused on a different parameter that is not data-dependent. This generality is important as it allows the model to be learned with potentially complex machine learning algorithms, as is commonly done across research, policy, and industry.

Using the CLT, I constructed CIs based on the normal approximation that are valid in a large class of problems, and documented cases where this approximation may fail to cover the parameter of interest at nominal rate. I provided a new approach to inference for such problems, focusing on the particular case of inference when comparing the performance between two models. The approach builds on my CLT, and I discussed how the arguments can be extended to other problems. I also provided a general approach that allows the moment functions to have zero limit variance in Appendix F, by exploring the faster-than-$\sqrt{n}$ convergence of the empirical moment equations and a tuning parameter.

In Section 5, I derived a new reproducibility measure for p-values calculated with split-sample Z-estimators. This measure is especially useful when computational resources are limited, quantifying whether a given number of split-sample repetitions suffices for two researchers using different splits to reach similar conclusions about statistical significance with high probability.

Finally, I illustrated the empirical implications of my results by revisiting two important problems in development and public economics: predicting poverty and learning heterogeneous treatment effects in randomized experiments. Using a panel from Ghana (Osei et al., 2022) and Monte Carlo experiments, repeated cross-fitting performed better than previous alternatives in detecting predictive power for being below the poverty line 13 years ahead. For the heterogeneous treatment effects application, I developed a new *ensemble* method that uses the entire sample for evaluation, more data for training, and combines multiple machine learning predictors. I revisited Karlan and List (2007)'s experiment on charitable giving and conducted Monte Carlo simulations. In both cases, my ensemble method achieved improved power for detecting heterogeneous treatment effects compared to previous alternatives.

# References


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

Andrews, I., D. Fudenberg, L. Lei, A. Liang, and C. Wu (2022): "The transfer performance of economic models," *arXiv preprint arXiv:2202.04796*.

Andrews, I., T. Kitagawa, and A. McCloskey (2024): "Inference on winners," *The Quarterly Journal of Economics*, 139, 305–358.

Athey, S., N. Keleher, and J. Spiess (2025): "Machine learning who to nudge: causal vs predictive targeting in a field experiment on student financial aid renewal," *Journal of Econometrics*, 105945.

Auerbach, E., A. Liang, K. Okumura, and M. Tabord-Meehan (2024): "Testing the Fairness-Accuracy Improvability of Algorithms," *arXiv preprint arXiv:2405.04816*.

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

Barocas, S. and A. D. Selbst (2016): "Big data's disparate impact," *Calif. L. Rev.*, 104, 671.

Bates, S., T. Hastie, and R. Tibshirani (2024): "Cross-validation: what does it estimate and how well does it do it?" *Journal of the American Statistical Association*, 119, 1434–1445.

Bayle, P., A. Bayle, L. Janson, and L. Mackey (2020): "Cross-validation confidence intervals for test error," *Advances in Neural Information Processing Systems*, 33, 16339–16350.

Belloni, A., V. Chernozhukov, I. Fernandez-Val, and C. Hansen (2017): "Program evaluation and causal inference with high-dimensional data," *Econometrica*, 85, 233–298.

Benkeser, D., M. Petersen, and M. J. van der Laan (2020): "Improved small-sample estimation of nonlinear cross-validated prediction metrics," *Journal of the American Statistical Association*, 115, 1917–1932.

Bryan, G., D. Karlan, and A. Osman (2024): "Big loans to small businesses: Predicting winners and losers in an entrepreneurial lending experiment," *American Economic Review*, 114, 2825–2860.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21, C1–C68.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernández-Val (2025a): "Reply to: Comments on "Fisher–Schultz Lecture: Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, With an Application to Immunization in India"," *Econometrica*, 93, 1177–1181.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernández-Val (2025b): "Fisher–Schultz Lecture: Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, With an Application to Immunization in India," *Econometrica*, 93, 1121–1164.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and confidence regions for parameter sets in econometric models 1," *Econometrica*, 75, 1243–1284.

Chouldechova, A. and A. Roth (2018): "The Frontiers of Fairness in Machine Learning," .

Cowgill, B. and C. E. Tucker (2020): "Algorithmic fairness and economics," *Columbia Business School Research Paper*.

Davidson, J. (2021): *Stochastic Limit Theory: An Introduction for Econometricians*, Oxford University Press.

Dawid, A. (1994): "Selection paradoxes of Bayesian inference," *Lecture Notes-Monograph Series*, 211–220.

Dudoit, S. and M. J. van der Laan (2005): "Asymptotics of cross-validated risk estimation in estimator selection and performance assessment," *Statistical methodology*, 2, 131–154.

Fava, B. (2025): "Predicting the Distribution of Treatment Effects via Covariate-Adjustment, with an Application to Microcredit," .

Fischer-Abaigar, U., C. Kern, N. Barda, and F. Kreuter (2024): "Bridging the gap: Towards an expanded toolkit for AI-driven decision-making in the public sector," *Government Information Quarterly*, 41, 101976.

Gasparin, M., R. Wang, and A. Ramdas (2025): "Combining exchangeable p-values," *Proceedings of the National Academy of Sciences*, 122.

Hansen, B. (2022): *Econometrics*, Princeton University Press.

Ida, T., T. Ishihara, K. Ito, D. Kido, T. Kitagawa, S. Sakaguchi, and S. Sasaki (2024): "Dynamic targeting: Experimental evidence from energy rebate programs," Tech. rep., National Bureau of Economic Research.

Imai, K. and M. L. Li (2025): "Statistical inference for heterogeneous treatment effects discovered by generic machine learning in randomized experiments," *Journal of Business & Economic Statistics*, 43, 256–268.

Ji, W., L. Lei, and A. Spector (2024): "Model-Agnostic Covariate-Assisted Inference on Partially Identified Causal Effects," .

Johnson, M. S., D. I. Levine, and M. W. Toffel (2023): "Improving regulatory effectiveness through better targeting: Evidence from OSHA," *American Economic Journal: Applied Economics*, 15, 30–67.

Karlan, D. and J. A. List (2007): "Does price matter in charitable giving? Evidence from a large-scale natural field experiment," *American Economic Review*, 97, 1774–1793.

Kitagawa, T. and A. Tetenov (2018): "Who should be treated? empirical welfare maximization methods for treatment choice," *Econometrica*, 86, 591–616.

Kuzmanovic, M., D. Frauen, T. Hatt, and S. Feuerriegel (2024): "Causal machine learning for cost-effective allocation of development aid," in *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, 5283–5294.

LeDell, E., M. Petersen, and M. van der Laan (2015): "Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates," *Electronic journal of statistics*, 9, 1583.

Liang, A., J. Lu, X. Mu, and K. Okumura (2024): "Algorithm Design: A Fairness-Accuracy Frontier," *arXiv preprint arXiv:2112.09975*.

Liu, Y. and F. Molinari (2024): "Inference for an Algorithmic Fairness-Accuracy Frontier," *arXiv preprint arXiv:2402.08879*.

Luedtke, A. R. and M. J. Van Der Laan (2016): "Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy," *Annals of statistics*, 44, 713.

Meinshausen, N., L. Meier, and P. Bühlmann (2009): "p-Values for High-Dimensional Regression," *Journal of the American Statistical Association*, 104, 1671–1681.

Meng, X.-L. (1994): "Posterior predictive $p$-values," *The annals of statistics*, 22, 1142–1160.

Osei, R., I. Osei-Akoto, E. Aryeetey, F. Dzanku, C. Udry, and E. G. Center (2022): "ISSER-Northwestern-Yale Long Term Ghana Socioeconomic Panel Survey (GSPS)," `https://doi.org/10.7910/DVN/E5QP0F`, Harvard Dataverse, V1, UNF:6:JLtXxepgNXfzyX0ThGLDiw==.

Potash, E., J. Brew, A. Loewi, S. Majumdar, A. Reece, J. Walsh, E. Rozier, E. Jorgenson, R. Mansour, and R. Ghani (2015): "Predictive modeling for public health: Preventing childhood lead poisoning," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2039–2047.

Ritzwoller, D. M. and J. P. Romano (2023): "Reproducible aggregation of sample-split statistics," *arXiv preprint arXiv:2311.14204*.

Romano, J. P. and A. M. Shaikh (2008): "Inference for identifiable parameters in partially identified econometric models," *Journal of Statistical Planning and Inference*, 138, 2786–2807.

——— (2010): "Inference for the identified set in partially identified econometric models," *Econometrica*, 78, 169–211.

Rüger, B. (1978): "Das maximale signifikanzniveau des Tests:"Lehne H o ab, wenn k unter n gegebenen tests zur ablehnung führen"," *Metrika*, 25, 171–178.

Rüschendorf, L. (1982): "Random variables with maximum sums," *Advances in Applied Probability*, 14, 623–632.

Semenova, V. (2025): "Debiased Machine Learning of Aggregated Intersection Bounds and Other Causal Parameters," .

Shi, C., W. Lu, and R. Song (2020): "Breaking the Curse of Nonregularity with Subagging—Inference of the Mean Outcome under Optimal Treatment Regimes," *Journal of Machine Learning Research*, 21, 1–67.

Shi, X. (2015): "Model selection tests for moment inequality models," *Journal of Econometrics*, 187, 1–17.

Van der Vaart, A. W. (2000): *Asymptotic statistics*, vol. 3, Cambridge university press.

van der Vaart, A. W. and J. A. Wellner (2023): *Weak convergence and empirical processes: with applications to statistics*, Springer.

Velez, A. (2024): "On the Asymptotic Properties of Debiased Machine Learning Estimators," *arXiv preprint arXiv:2411.01864*.

Wager, S. (2024): "Sequential Validation of Treatment Heterogeneity," *arXiv preprint arXiv:2405.05534*.

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

Wüthrich, K. and Y. Zhu (2023): "Omitted variable bias of Lasso-based inference methods: A finite sample analysis," *Review of Economics and Statistics*, 105, 982–997.

Zhang, C.-H. and J. Huang (2008): "The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression," *The Annals of Statistics*, 36, 1567–1594.

Zhao, P. and B. Yu (2006): "On model selection consistency of Lasso," *The Journal of Machine Learning Research*, 7, 2541–2563.

# A Bounding the Performance of Average Model

Let $Y$ be a scalar outcome, $X$ a set of covariates, and $(\hat{\eta}_{\tilde{\mathsf{s}}})_{\mathsf{s}\in\mathcal{S}}$ be a collection of models estimated through multiple splits of the sample, where $\tilde{\mathsf{s}}$ is the complement of $\mathsf{s}$, as in Section 2. For example, $\mathcal{S}$ can be a vectorization of $\mathcal{R}$ defined in Section 2, $\mathcal{S} = (\mathsf{s}_{m,k})_{m\in[M],k\in[K]}$. Denote $\bar{\eta}(x) = \frac{1}{|\mathcal{S}|}\sum_{\mathsf{s}\in\mathcal{S}} \hat{\eta}_{\tilde{\mathsf{s}}}(x)$. If $Y$ is binary, some algebra manipulation gives the following equalities:

$$\theta_{\bar{\eta},1} = \int |y - \bar{\eta}(x)|\, dP(y,x) = \frac{1}{|\mathcal{S}|}\sum_{\mathsf{s}\in\mathcal{S}} \int |y - \hat{\eta}_{\tilde{\mathsf{s}}}(x)|\, dP(y,x) = \theta_{\hat{\eta},1},$$
$$\theta_{\bar{\eta},2} = \int (y - \bar{\eta}(x))^2\, dP(y,x) = \frac{1}{|\mathcal{S}|}\sum_{\mathsf{s}\in\mathcal{S}} \int (y - \hat{\eta}_{\tilde{\mathsf{s}}}(x))^2\, dP(y,x) = \theta_{\hat{\eta},2}.$$

Hence, one can use either $\bar{\eta}$ or a model $\tilde{\eta}(x)$ that takes value in $(\hat{\eta}_{\tilde{\mathsf{s}}}(x))_{\mathsf{s}\in\mathcal{S}}$ uniformly at random, and both will yield the same out-of-sample mean absolute deviation and mean squared error.

For the general case, if $Y$ is continuous, an application of the triangle inequality establishes a risk-contraction property for $\bar{\eta}$:

$$\theta_{\bar{\eta},1} = \int |y - \bar{\eta}(x)|\, dP(y,x) \leqslant \frac{1}{|\mathcal{S}|}\sum_{\mathsf{s}\in\mathcal{S}} \int |y - \hat{\eta}_{\tilde{\mathsf{s}}}(x)|\, dP(y,x) = \theta_{\hat{\eta},1},$$
$$\theta_{\bar{\eta},2} = \sqrt{\int (y - \bar{\eta}(x))^2\, dP(y,x)} \leqslant \frac{1}{|\mathcal{S}|}\sum_{\mathsf{s}\in\mathcal{S}} \sqrt{\int (y - \hat{\eta}_{\tilde{\mathsf{s}}}(x))^2\, dP(y,x)} = \theta_{\hat{\eta},2}.$$

Similar results hold for other distance-based functional forms where the triangle inequality applies. Although my framework does not cover the parameters $\theta_{\bar{\eta},1}$ and $\theta_{\bar{\eta},2}$, it covers $\theta_{\hat{\eta},1}$ and $\theta_{\hat{\eta},2}$, which are upper bounds on the error rate of using model $\bar{\eta}$. Hence, if one uses model $\bar{\eta}$ for out-of-sample prediction, they have the guarantee that its accuracy will be at least as large (error at least as small) as the error they can estimate, $\theta_{\hat{\eta},1}$ or $\theta_{\hat{\eta},2}$. Note that the root mean squared error estimand $\theta_{\hat{\eta},2}$ is similar although different from the one discussed in Section 1. In this case, the estimator is also covered by Section 3 and given by

$$\hat{\theta}_{\hat{\eta},2} = \frac{1}{|\mathcal{S}|}\sum_{\mathsf{s}\in\mathcal{S}} \sqrt{\frac{1}{|\mathsf{s}|}\sum_{i\in\mathsf{s}} (Y_i - \hat{\eta}_{\tilde{\mathsf{s}}}(X_i))^2}.$$

# B Proofs and Extra Definitions

The following notation is used throughout the proofs. If unspecified, $X \xrightarrow{P} Y$ denotes convergence in probability uniformly in $P \in \mathcal{P}$, that is, for every $\varepsilon > 0$,

$\sup_{P\in\mathcal{P}} P(|X-Y|>\varepsilon)\to 0$. $X_n = o_P(a_n) \iff X_n/a_n \xrightarrow{P} 0$. $X_n = O_P(a_n) \iff (\forall \varepsilon>0, \exists M>0$ and $N>0$ s.t. $n>N \implies \sup_{P\in\mathcal{P}} P\left(\left|\frac{X_n}{a_n}\right|>M\right)<\varepsilon)$. $\rightsquigarrow$ means weak convergence uniformly in $P\in\mathcal{P}$.

## B.1 Proofs and Extra Definitions of Section 3

Define

$$\Psi_{\hat{\eta}}(\theta) = \frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\Psi_{\hat{\eta}_{\mathsf{s}}}(\theta),$$

$$\hat{\Psi}_{\hat{\eta}}(\theta) = \frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\hat{\Psi}_{\mathsf{s},\hat{\eta}_{\mathsf{s}}}(\theta),$$

$$\hat{\Psi}_{\eta_P^*}(\theta) = \frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\hat{\Psi}_{\mathsf{s},\eta_P^*}(\theta),$$

$$\dot{\Psi}_{\hat{\eta}}(\theta) = \frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\dot{\Psi}_{\hat{\eta}_{\mathsf{s}}}(\theta),$$

$$\dot{\Psi}_{\hat{\eta}} = \dot{\Psi}_{\hat{\eta}}(\theta_{\hat{\eta}}),$$

where $\dot{\Psi}_{\hat{\eta}_{\mathsf{s}}}(\theta)$ is the Jacobian matrix of $\Psi_\eta(\theta)$, its derivative in $\theta$.

**Assumption B.1.** *For some $\Theta'\subseteq\Theta$, the following conditions hold:*

(i) $\left\{\theta_{\eta_P^*}\in\Theta : P\in\mathcal{P}\right\} \subseteq int(\Theta')$, *and the classes $\mathcal{F}_\eta = \{\psi_{\theta,\eta,j} : \theta\in\Theta'\}$ are P-Donsker uniformly in $P\in\mathcal{P}$ and $\eta\in H$ in the sense defined in Assumption E.1 with $T=\Theta'$, where $j=1,\dots,d$, and $\psi_{\theta,\eta,j}$ is the $j$-th coordinate of $\psi_{\theta,\eta}$;*

(ii) *The estimators $\hat{\theta}^{(1)}_{\hat{\eta}}, \hat{\theta}^{(2)}_{\hat{\eta}}, \hat{\theta}^{(3)}_{\hat{\eta}}$ satisfy*

$$\sqrt{n}\left\|\hat{\Psi}_{\mathsf{s},\hat{\eta}_{\mathsf{s}}}(\hat{\theta}^{(1)}_{\hat{\eta}_{\mathsf{s}}})\right\| \xrightarrow{P} 0 \ \ \forall \mathsf{s}\in(\mathsf{s}_{m,k})_{m\in[M],k\in[K]},$$

$$\sqrt{n}\left\|\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\hat{\Psi}_{\mathsf{s},\hat{\eta}_{\mathsf{s}}}(\hat{\theta}^{(2)}_{\hat{\eta}})\right\| \xrightarrow{P} 0,$$

$$\sqrt{n}\left\|\frac{1}{K}\sum_{\mathsf{s}\in r}\hat{\Psi}_{\mathsf{s},\hat{\eta}_{\mathsf{s}}}(\hat{\theta}^{(3)}_{\hat{\eta}_r})\right\| \xrightarrow{P} 0 \ \ \forall r\in\mathcal{R},$$

*uniformly in $P\in\mathcal{P}$;*

(iii) *For every* $\varepsilon > 0$,

$$\sup_{P\in\mathcal{P}} \sup_{\left\|\theta-\theta_{\eta_P^*}\right\|>\varepsilon} -\left\|\Psi_{\eta_P^*}(\theta)\right\| < 0 = \left\|\Psi_{\eta_P^*}(\theta_{\eta_P^*})\right\|;$$

(iv) *For* $\tilde{\eta} = \mathcal{A}(D)$,

$$\left\|\dot{\Psi}_{\tilde{\eta}} - \dot{\Psi}_{\eta_P^*}\right\| \xrightarrow{P} 0$$

*uniformly in* $P \in \mathcal{P}$*;*

(v) $\Psi_\eta$ *is differentiable at* $\theta_\eta$ *for* $\eta \in H$*, and for some* $\bar{c}_1 > 0$,

$$\inf_{P\in\mathcal{P}} \left|\det\left(\dot{\Psi}_{\eta_P^*}\right)\right| \geqslant \bar{c}_1.$$

□

Assumption B.1(i) is a Donsker condition for a subset $\Theta'$ that contains $\theta_{\eta_P^*}$ in its interior. Importantly, Assumption E.1, defined in Appendix E, does not restrict the complexity of the class of trained models $H$, and it allows $\hat{\eta}$ to be estimated with any machine learning algorithm as long as Assumption 3.1(ii) holds. It restricts the complexity of $\psi_{\theta,\eta}$ only along $\theta \in \Theta'$, and not along $\eta \in H$. Assumption B.1(i) holds, for example, if $\Theta'$ is bounded and $\psi_{\theta,\eta}$ is Lipschitz in $\theta$ with a Lipschitz constant that does not depend on $\eta$ or $w$. Assumption B.1(ii) allows for approximate Z-estimators which nearly solve the moment condition, and is immediately satisfied for exact Z-estimators, for example when

$$\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\hat{\Psi}_{\mathsf{s},\hat{\eta}_{\bar{\mathsf{s}}}}(\hat{\theta}^{(2)}_{\hat{\eta}}) = 0$$

in the case of $\hat{\theta}^{(2)}_{\hat{\eta}}$. Assumption B.1(iii) requires $\theta_{\eta_P^*}$ to be a unique and well-separated zero of $\Psi_{\eta_P^*}$, and can be replaced by the higher-level condition that $\|\hat{\theta}^{(j)}_{\hat{\eta}} - \theta^{(j)}_{\hat{\eta}}\| \xrightarrow{P} 0$ uniformly in $P \in \mathcal{P}$ for $j \in \{1,2,3\}$. Assumption B.1(iv) holds under the condition that $\dot{\Psi}_{\eta_P^*}$ is continuous in $\eta$ around $\eta_P^*$. Finally, Assumption B.1(v) requires the absolute determinant of the Jacobian to be bounded away from zero, which guarantees its invertibility in a uniform sense over $P \in \mathcal{P}$.

**Lemma B.1.** *Let Assumptions 3.1 and B.1 hold. Then, uniformly in* $P \in \mathcal{P}$*,*

$$\sup_{\theta\in\Theta'}\left\|\hat{\Psi}_{\hat{\eta}}(\theta) - \Psi_{\hat{\eta}}(\theta)\right\| \xrightarrow{P} 0 \tag{B.1}$$

$$\sup_{\theta\in\Theta'}\left\|\hat{\Psi}_{\eta_P^*}(\theta) - \Psi_{\eta_P^*}(\theta)\right\| \xrightarrow{P} 0 \tag{B.2}$$

$$\sup_{\theta\in\Theta'}\left\|\Psi_{\hat{\eta}}(\theta) - \Psi_{\eta_P^*}(\theta)\right\| \xrightarrow{P} 0 \tag{B.3}$$

*Proof of Lemma B.1.* (B.1) and (B.2) follow from asymptotic equicontinuity established in Theorem E.1. (B.3) follows from asymptotic equicontinuity of $\Psi_{\hat{\eta}}(\theta) - \Psi_{\eta_P^*}(\theta)$ (follows from Assumption E.1(v)) and pointwise in $\theta$ convergence (Assumption E.2). ■

**Lemma B.2.** *Let Assumptions 3.1 and B.1 hold. Then,*

$$\left\|\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right\| \xrightarrow{P} 0.$$

*Proof of Lemma B.2.* By Assumption B.1(iii), for any $\varepsilon > 0$, there is $\gamma > 0$ such that

$$\left\|\theta - \theta_{\eta_P^*}\right\| > \varepsilon \implies \left\|\Psi_{\eta_P^*}(\theta)\right\| > \gamma.$$

Hence,

$$\sup_{P\in\mathcal{P}} P\left(\left\|\hat{\theta}_{\hat{\eta}} - \theta_{\eta_P^*}\right\| > \varepsilon\right) \leqslant \sup_{P\in\mathcal{P}} P\left(\left\|\Psi_{\eta_P^*}(\hat{\theta}_{\hat{\eta}})\right\| > \gamma\right) \to 0,$$

since

$$\begin{aligned}\left\|\Psi_{\eta_P^*}(\hat{\theta}_{\hat{\eta}})\right\| &\leqslant \left\|\Psi_{\eta_P^*}(\hat{\theta}_{\hat{\eta}}) - \Psi_{\hat{\eta}}(\hat{\theta}_{\hat{\eta}})\right\| + \left\|\hat{\Psi}_{\hat{\eta}}(\hat{\theta}_{\hat{\eta}}) - \Psi_{\hat{\eta}}(\hat{\theta}_{\hat{\eta}})\right\| + o_P(1) \\ &= o_P(1),\end{aligned}$$

by Assumption B.1(ii), (B.3), and (B.1). This implies $\left\|\hat{\theta}_{\hat{\eta}} - \theta_{\eta_P^*}\right\| \xrightarrow{P} 0$ uniformly in $P \in \mathcal{P}$.

Similar happens for $\left\|\theta_{\hat{\eta}} - \theta_{\eta_P^*}\right\|$. For any $\varepsilon > 0$, there is $\gamma > 0$ such that

$$\sup_{P\in\mathcal{P}} P\left(\left\|\theta_{\hat{\eta}} - \theta_{\eta_P^*}\right\| > \varepsilon\right) \leqslant \sup_{P\in\mathcal{P}} P\left(\left\|\Psi_{\eta_P^*}(\theta_{\hat{\eta}})\right\| > \gamma\right) \to 0,$$

since $\Psi_{\hat{\eta}}(\theta_{\hat{\eta}}) = 0$ and

$$\left\|\Psi_{\eta_P^*}(\theta_{\hat{\eta}})\right\| = \left\|\Psi_{\eta_P^*}(\theta_{\hat{\eta}}) - \Psi_{\hat{\eta}}(\theta_{\hat{\eta}})\right\| \xrightarrow{P} 0$$

uniformly in $P \in \mathcal{P}$ by (B.3).

The result follows from the triangle inequality. ■

*Proof of Theorem 3.1.* I first show the result for the case of $\theta_{\hat{\eta}} = \theta_{\hat{\eta}}^{(2)}$ (and $\hat{\theta}_{\hat{\eta}} = \hat{\theta}_{\hat{\eta}}^{(2)}$). Differentiability of $\Psi_{\hat{\eta}}$ and Assumption B.1(iv) gives

$$\begin{aligned}\Psi_{\hat{\eta}}(\hat{\theta}_{\hat{\eta}}) - \Psi_{\hat{\eta}}\left(\theta_{\hat{\eta}}\right) &= \dot{\Psi}_{\hat{\eta}}\left(\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right) + o_P\left(\left\|\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right\|\right) \\ &= \dot{\Psi}_{\eta_P^*}\left(\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right) + o_P\left(\left\|\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right\|\right). \qquad \text{(B.4)}\end{aligned}$$

Asymptotic equicontinuity gives

$$\sqrt{n}\left(\Psi_{\hat{\eta}}(\hat{\theta}_{\hat{\eta}}) - \Psi_{\hat{\eta}}(\theta_{\hat{\eta}})\right) = -\sqrt{n}\left(\hat{\Psi}_{\hat{\eta}}(\hat{\theta}_{\hat{\eta}}) - \Psi_{\hat{\eta}}(\hat{\theta}_{\hat{\eta}})\right) + o_P(1) \tag{B.5}$$

$$= -\sqrt{n}\left(\hat{\Psi}_{\hat{\eta}}(\theta_{\hat{\eta}}) - \Psi_{\hat{\eta}}(\theta_{\hat{\eta}})\right) + o_P(1) \tag{B.6}$$

$$= -\sqrt{n}\hat{\Psi}_{\hat{\eta}}(\theta_{\hat{\eta}}) + o_P(1) \tag{B.7}$$

$$= O_P(1), \tag{B.8}$$

where (B.5) uses $\sqrt{n}\hat{\Psi}_{\hat{\eta}}(\hat{\theta}_{\hat{\eta}}) = o_P(1)$ (Assumption B.1(ii)) and $\Psi_{\hat{\eta}}(\theta_{\hat{\eta}}) = 0$, and (B.6) uses Assumption B.1(i) and Theorem E.1, and

$$\left\|\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right\| \xrightarrow{P} 0$$

uniformly in $P \in \mathcal{P}$, established in Lemma B.2. Note that Assumption 3.1(ii), used for Theorem E.1, is stronger than Assumption E.2 (see proof of Theorem D.1).

By invertibility of $\dot{\Psi}_{\eta_P^*}$,

$$\sqrt{n}\left\|\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right\| \leqslant \left\|\dot{\Psi}_{\eta_P^*}^{-1}\right\| \left\|\dot{\Psi}_{\eta_P^*}\sqrt{n}\left(\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right)\right\|.$$

Plugging (B.4) in the right-hand side gives

$$\sqrt{n}\left\|\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right\| \left\|\dot{\Psi}_{\eta_P^*}^{-1}\right\|^{-1} \leqslant \left\|\sqrt{n}\left(\Psi_{\hat{\eta}}(\hat{\theta}_{\hat{\eta}}) - \Psi_{\hat{\eta}}\left(\theta_{\hat{\eta}}\right)\right) + o_P\left(\sqrt{n}\left\|\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right\|\right)\right\|,$$

which implies

$$\sqrt{n}\left\|\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right\| \left(\left\|\dot{\Psi}_{\eta_P^*}^{-1}\right\|^{-1} + o_P(1)\right) \leqslant \left\|\sqrt{n}\left(\Psi_{\hat{\eta}}(\hat{\theta}_{\hat{\eta}}) - \Psi_{\hat{\eta}}\left(\theta_{\hat{\eta}}\right)\right)\right\| = O_P(1),$$

where the equality follows from (B.8) and Assumption B.1(v). As a consequence,

$$o_P\left(\sqrt{n}\left\|\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right\|\right) = o_P(1). \tag{B.9}$$

Finally, combining (B.4) and (B.7) gives

$$\begin{aligned}\dot{\Psi}_{\eta_P^*}\sqrt{n}\left(\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right) &= -\sqrt{n}\hat{\Psi}_{\hat{\eta}}(\theta_{\hat{\eta}}) + o_P\left(\sqrt{n}\left\|\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right\|\right) + o_P(1) \\ &= -\sqrt{n}\hat{\Psi}_{\hat{\eta}}(\theta_{\hat{\eta}}) + o_P(1).\end{aligned}$$

Hence,

$$\begin{aligned}\sqrt{n}\left(\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right) &= -\dot{\Psi}_{\eta_P^*}^{-1}\sqrt{n}\hat{\Psi}_{\hat{\eta}}(\theta_{\hat{\eta}}) + o_P(1), \\ &= -\dot{\Psi}_{\eta_P^*}^{-1}\sqrt{n}\hat{\Psi}_{\eta_P^*}(\theta_{\eta_P^*}) + o_P(1)\end{aligned}$$

by applying Theorem E.1, and the result follows for $\theta^{(2)}_{\hat{\eta}}$. Note that Assumption 3.1(ii) is stronger than Assumption E.2 (see proof of Theorem D.1).

The results for $\theta^{(1)}_{\hat{\eta}}$ and $\theta^{(3)}_{\hat{\eta}}$ follow similarly. For $\theta^{(1)}_{\hat{\eta}}$, applying the same arguments above with $K = 1$ and $M = 1$ gives

$$\sqrt{n}\left(\hat{\theta}_{\hat{\eta}_{\tilde{\mathsf{s}}}} - \theta_{\hat{\eta}_{\tilde{\mathsf{s}}}}\right) = -\dot{\Psi}^{-1}_{\eta^*_P}\sqrt{n}\hat{\Psi}_{\hat{\eta}_{\tilde{\mathsf{s}}}}(\theta_{\hat{\eta}_{\tilde{\mathsf{s}}}}) + o_P(1)$$

for any $\tilde{\mathsf{s}} \in (\tilde{\mathsf{s}}_{m,k})_{m\in[M],k\in[K]}$, and the result follows for $j = 1$ by summing over $\mathsf{s} \in (\mathsf{s}_{m,k})_{m\in[M],k\in[K]}$:

$$\sqrt{n}\left(\hat{\theta}^{(1)}_{\hat{\eta}} - \theta^{(1)}_{\hat{\eta}}\right) = -\dot{\Psi}^{-1}_{\eta^*_P}\sqrt{n}\hat{\Psi}_{\eta^*_P}(\theta_{\eta^*_P}) + o_P(1).$$

Similar holds for $j = 3$ applying the arguments above with $M = 1$ and $K > 1$ and summing over $r \in \mathcal{R}$. ■

## B.2 Proofs and Extra Definitions of Section 4

If $\psi_{\theta,\eta}$ is differentiable in $\theta$, let $\dot{\psi}_{\theta,\eta}(w)$ be the Jacobian matrix of $\psi_{\theta,\eta}(w)$, where the derivatives are taken in respect to $\theta$. In that case, $\hat{\dot{\Psi}}_{\hat{\eta}}$ can be given by

$$\hat{\dot{\Psi}}_{\hat{\eta}} = \frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\frac{1}{b}\sum_{i\in\mathsf{s}}\dot{\psi}_{\hat{\theta}_{\hat{\eta}},\hat{\eta}_{\tilde{\mathsf{s}}}}(W_i). \tag{B.10}$$

Define

$$V_{M,K} = \begin{cases} M^{-1}\left(n/b + M - 1\right), & \text{if } K = 1 \\ 1, & \text{otherwise,} \end{cases}$$

$$\hat{V}_{\hat{\eta}} = V_{M,K}\hat{\dot{\Psi}}^{-1}_{\hat{\eta}}\left(\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\frac{1}{b}\sum_{i\in\mathsf{s}}\psi_{\hat{\theta}_{\hat{\eta}},\hat{\eta}_{\tilde{\mathsf{s}}}}(W_i)\psi^T_{\hat{\theta}_{\hat{\eta}},\hat{\eta}_{\tilde{\mathsf{s}}}}(W_i)\right)\left(\hat{\dot{\Psi}}^{-1}_{\hat{\eta}}\right)^T. \tag{B.11}$$

*Proof of Theorem 4.1.* Under the conditions of the theorem, for $j \in \{1, 2, 3\}$,

$$\sqrt{n}\hat{V}^{-1/2}_{\hat{\eta}}\left(\hat{\theta}^{(j)}_{\hat{\eta}} - \theta^{(j)}_{\hat{\eta}}\right) \rightsquigarrow \mathcal{N}\left(0, I_d\right)$$

uniformly in $P \in \mathcal{P}$, where $I_d$ is the identity matrix. Consistency of the inner term to $\left(P\psi_{\theta_{\eta^*_P},\eta^*_P}\psi^T_{\theta_{\eta^*_P},\eta^*_P}\right)$ follows similarly to the proof of Theorem D.2, and the result follows from the continuous mapping theorem, Theorem 3.1 and the delta method. ■

**Assumption B.2.** *The following conditions hold:*

(i) *There exists a consistent estimator* $\hat{V}_{\hat{\eta}} \xrightarrow{P} V_{\eta_P^*}$ *uniformly in* $P \in \mathcal{P}$*;*

(ii) $\left\| \dot{\Psi}_{\hat{b}} - \dot{\Psi}_{b_P} \right\| \xrightarrow{P} 0$ *uniformly in* $P \in \mathcal{P}$*;*

(iii) $\left| \hat{\theta}_{\hat{b}} - \theta_{\hat{b}} \right| \xrightarrow{P} 0$ *uniformly in* $P \in \mathcal{P}$*;*

(iv) $\sup_{P \in \mathcal{P}} \dot{\Psi}_{b_P}^{-1} < \infty$.

□

Item Assumption B.2(i) requires $V_{\eta_P^*}$ to be consistently estimable, which can typically be verified as in Theorem 4.1. Item Assumption B.2(ii) through Item Assumption B.2(iv) adapt conditions Assumption B.1(iv) through Assumption B.1(v) to $b_P$ instead of $\eta_P^*$.

I give below a formula for $\hat{\Sigma}$ for the case of sample averages, that is, $\psi_{\theta,\eta}(w) = f_\eta(w) - \theta$. Analogous estimators can be defined for the general case using the fact that $\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}$ is asymptotically linear:

$$\sqrt{n}\left(\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right) = -\dot{\Psi}_{\eta_P^*}^{-1} \sqrt{n} \hat{\Psi}_{\eta_P^*}(\theta_{\eta_P^*}) + o_P(1),$$

from Theorem 3.1.

$$\hat{\Sigma} = \left(\hat{\Sigma}_{j,\ell}\right)_{j,\ell=1}^{MK}, \tag{B.12}$$

where for splits $\mathsf{s}_j, \mathsf{s}_\ell \in \mathcal{S}$ with complements $\tilde{\mathsf{s}}_j, \tilde{\mathsf{s}}_\ell$,

$$\begin{aligned}
\hat{\Sigma}_{j,j} &= \frac{1}{n^2} \sum_{i \in \tilde{\mathsf{s}}_j} \left( f_{\hat{b}}(W_i) - \bar{f}_{\hat{b},\tilde{\mathsf{s}}_j} \right)^2 + \frac{1}{n^2} \sum_{i \in \mathsf{s}_j} \left( \tilde{f}_j(W_i) - \bar{\tilde{f}}_j \right)^2, \\
\hat{\Sigma}_{j,\ell} &= \frac{1}{n^2} \sum_{i \in \tilde{\mathsf{s}}_j \cap \tilde{\mathsf{s}}_\ell} \left( f_{\hat{b}}(W_i) - \bar{f}_{\hat{b},\tilde{\mathsf{s}}_j \cap \tilde{\mathsf{s}}_\ell} \right)^2 \\
&\quad + \frac{1}{n^2} \sum_{i \in \tilde{\mathsf{s}}_j \cap \mathsf{s}_\ell} \left( f_{\hat{b}}(W_i) - \bar{f}_{\hat{b},\tilde{\mathsf{s}}_j \cap \mathsf{s}_\ell} \right) \left( \tilde{f}_\ell(W_i) - \bar{\tilde{f}}_{\ell,\tilde{\mathsf{s}}_j \cap \mathsf{s}_\ell} \right) \\
&\quad + \frac{1}{n^2} \sum_{i \in \mathsf{s}_j \cap \tilde{\mathsf{s}}_\ell} \left( \tilde{f}_j(W_i) - \bar{\tilde{f}}_{j,\mathsf{s}_j \cap \tilde{\mathsf{s}}_\ell} \right) \left( f_{\hat{b}}(W_i) - \bar{f}_{\hat{b},\mathsf{s}_j \cap \tilde{\mathsf{s}}_\ell} \right) \\
&\quad + \frac{1}{n^2} \sum_{i \in \mathsf{s}_j \cap \mathsf{s}_\ell} \left( \tilde{f}_j(W_i) - \bar{\tilde{f}}_{j,\mathsf{s}_j \cap \mathsf{s}_\ell} \right) \left( \tilde{f}_\ell(W_i) - \bar{\tilde{f}}_{\ell,\mathsf{s}_j \cap \mathsf{s}_\ell} \right) \quad \text{for } j \neq \ell,
\end{aligned}$$

where $\tilde{f}_j(W_i) = f_{\hat{b}}(W_i) - \frac{n}{|\mathsf{s}_j|} f_{\hat{\eta}_{\tilde{\mathsf{s}}_j}}(W_i)$, and for any set $\mathsf{s} \subseteq \{1, \ldots, n\}$, $\bar{f}_{\hat{b},\mathsf{s}} = |\mathsf{s}|^{-1} \sum_{i \in \mathsf{s}} f_{\hat{b}}(W_i)$ and $\bar{\tilde{f}}_{j,\mathsf{s}} = |\mathsf{s}|^{-1} \sum_{i \in \mathsf{s}} \tilde{f}_j(W_i)$, with $\bar{\tilde{f}}_j = \bar{\tilde{f}}_{j,\mathsf{s}_j}$.

Again, I give a standard error for the case of sample averages, and analogous estimators can be constructed for the general case following, e.g., Theorem 4.1.

$$\hat{\sigma}^2_{\hat{\delta}} = \hat{\sigma}^2_{\hat{\eta}} + \hat{\sigma}^2_{\hat{b}} - 2\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\frac{1}{|\mathsf{s}|}\sum_{i\in\mathsf{s}}\left(f_{\hat{\eta}_{\mathsf{s}}}(W_i) - \hat{\theta}_{\mathsf{s}}\right)\left(f_{\hat{\eta}_{\hat{b}}}(W_i) - \hat{\theta}_{\hat{b}}\right), \tag{B.13}$$

where $\hat{\sigma}_{\hat{\eta}}$ is defined as in (4.2) and

$$\hat{\sigma}^2_{\hat{b}} = \frac{1}{n}\sum_{i=1}^{n}\left(f_{\hat{b}}(W_i) - \hat{\theta}_{\hat{b}}\right)^2.$$

**Proposition B.1.** $\hat{\Sigma} \xrightarrow{P} \Sigma$ *uniformly in* $P \in \mathcal{P}$. □

**Proposition B.2.** $\hat{\sigma}_{\hat{\delta}} \xrightarrow{P} \sigma_{\hat{\delta}}$ *uniformly in* $P \in \mathcal{P}$. □

The two propositions above follow from a law of large numbers and Assumption 3.1(i) (assumed in Assumption 4.1).

Coverage of $\widehat{\mathrm{CI}}_\alpha$ is exact along any sequences where $\theta_{\eta^*_{P_n}} < \theta_{b_{P_n}}$ in the limit, without relying on Assumption 4.2.

**Theorem B.1.** *(Asymptotic exactness of* $\widehat{\mathrm{CI}}_\alpha$*)*
*Let Assumption 4.1 hold. Then, for any sequence* $(P_n)_{n\geqslant 1} \subseteq \mathcal{P}$ *such that* $\lim_{n\to\infty}\theta_{\eta^*_{P_n}} - \theta_{b_{P_n}} < 0$,

$$\lim_{n\to\infty} P_n\left((\theta_{\hat{\eta}} - \theta_{\hat{b}}) \in \widehat{\mathrm{CI}}_\alpha\right) = 1 - \alpha.$$

□

*Proof of Theorem B.1.* Follows from (B.16) and Proposition B.2. ■

For the proof of Theorem 4.2, define

$$\delta_{\eta^*_P} = \left(\theta_{\eta^*_P} - \theta_b\right)_{\mathsf{s}\in\mathcal{S}},$$

and

$$\hat{\delta}_{\eta^*_P} = \left(\hat{\theta}_{\eta^*_P} - \hat{\theta}_b\right)_{\mathsf{s}\in\mathcal{S}}.$$

*Proof of Theorem 4.2.* I first show the result for the case $\delta_{\hat{\eta}} = 0$. Let $C_2 > 0$, $(P_n)_{n\geqslant 1} \subseteq \mathcal{P}$ arbitrary such that $P_n(\delta_{\hat{\eta}} = 0) > C_2$. For any $\varepsilon > 0$ and $\mathsf{s} \in \mathcal{S}$, denote the event

$$E_{\mathsf{s}} = \left|\sqrt{n}\left[\left(\hat{\theta}_{\hat{\eta}_{\mathsf{s}}} - \hat{\theta}_{\hat{b}}\right) - (\theta_{\hat{\eta}_{\mathsf{s}}} - \theta_{\hat{b}})\right] - \sqrt{n}\left[\left(\hat{\theta}_{\eta^*_P} - \hat{\theta}_b\right) - \left(\theta_{\eta^*_P} - \theta_b\right)\right]\right| > \varepsilon.$$

By Theorem 3.1 and Assumption 4.1(ii),

$$P_n\left(E_{\mathsf{s}}\right) \to 0,$$

which implies

$$\begin{aligned} P_n\left(E_{\mathsf{s}}|\delta_{\hat{\eta}}=0\right) &\leqslant P_n\left(E_{\mathsf{s}}|\delta_{\hat{\eta}}=0\right) P_n(\delta_{\hat{\eta}}=0) C_2^{-1} \\ &\leqslant \left(P_n\left(E_{\mathsf{s}}|\delta_{\hat{\eta}}=0\right) P_n(\delta_{\hat{\eta}}=0) + P_n\left(E_{\mathsf{s}}|\delta_{\hat{\eta}}\neq 0\right) P_n(\delta_{\hat{\eta}}\neq 0)\right) C_2^{-1} \\ & P_n\left(E_{\mathsf{s}}\right) C_2^{-1} \to 0. \end{aligned}$$

Hence,

$$P_n\left(\left|\sqrt{n}\left(\hat{\delta}_{\hat{\eta}}-\delta_{\hat{\eta}}\right)-\sqrt{n}\left(\hat{\delta}_{\eta_P^*}-\delta_{\eta_P^*}\right)\right|>\varepsilon \;\middle|\; \delta_{\hat{\eta}}=0\right)\to 0,$$

and

$$\sqrt{n}\left(\hat{\delta}_{\hat{\eta}}-\delta_{\hat{\eta}}\right) \rightsquigarrow \mathcal{N}(0,\Sigma)$$

conditional on $\delta_{\hat{\eta}}=0$. Together with Proposition B.1 and the continuous mapping theorem, this implies

$$T(\hat{\delta}_{\hat{\eta}}, n^{-1}\hat{\Sigma}) \rightsquigarrow T(Z,\Sigma),$$

where $Z\sim\mathcal{N}(0,\Sigma)$. The result follows since the quantiles of $\mathcal{N}(0,\hat{\Sigma})$ converge to those of $\mathcal{N}(0,\Sigma)$ by the continuous mapping theorem and Proposition B.1.

Similar happens for the case $\delta_{\hat{\eta}}\geqslant 0$. The inequality comes from the fact that

$$\sqrt{n}\hat{\delta}_{\hat{\eta}} \geqslant \sqrt{n}\left(\hat{\delta}_{\hat{\eta}}-\delta_{\hat{\eta}}\right) \rightsquigarrow \mathcal{N}(0,\Sigma).$$

■

*Proof of Theorem 4.3.* Follows from Theorem 4.2, using

$$\sqrt{n}\left(\hat{\theta}_{\hat{b}}-\theta_{\hat{b}}\right)=\sqrt{n}\left(\hat{\theta}_{b_P}-\theta_{b_P}\right)+o_P(1)$$

from Assumption 4.1(ii), so that $\sqrt{n}\left(\hat{\theta}_{\hat{\eta}}-\theta_{\hat{b}}\right)\geqslant o_P(1)$ when $\theta_{\eta_P^*}=\theta_{b_P}$. ■

*Proof of Theorem 4.4.* For the first result, an argument similar to the proof of Theorem 4.2 conditional on

$$\left(\theta_{\hat{\eta}}-\theta_{\hat{b}}\right)\geqslant 0 \vee \left(\theta_{\hat{\eta}}-\theta_{\hat{b}}\right)\leqslant \bar{c}_3 \tag{B.14}$$

implies

$$\sqrt{n}\left(\hat{\delta}_{\hat{\eta}}-\delta_{\hat{\eta}}\right) \rightsquigarrow \mathcal{N}(0,\Sigma) \tag{B.15}$$

and

$$\sqrt{n}\left(\hat{\theta}_{\hat{\eta}}-\theta_{\hat{\eta}}\right)-\sqrt{n}\left(\hat{\theta}_{\hat{b}}-\theta_{\hat{b}}\right)=-\dot{\Psi}_{\eta_P^*}^{-1}\sqrt{n}\hat{\Psi}_{\eta_P^*}(\theta_{\eta_P^*})+\dot{\Psi}_{b_P}^{-1}\sqrt{n}\hat{\Psi}_{b_P}(\theta_{b_P})+o_P(1) \quad \text{(B.16)}$$

conditional on (B.14), uniformly in $P\in\mathcal{P}_{\bar{c}_3,\bar{c}_4}$. (B.16) uses Assumption 4.1(ii), Theorem 3.1, and Proposition B.2. If $(P_n)_{n\geqslant 1}\subseteq\mathcal{P}_{\bar{c}_3,\bar{c}_4}$ is such that $\lim_{n\to\infty}\theta_{\eta_{P_n}^*}-\theta_{b_{P_n}}\leqslant\bar{c}_3$, the result follows Proposition B.2 since (B.16) is asymptotically normal (nondegenerate). $\bar{c}_3<\lim_{n\to\infty}\theta_{\eta_{P_n}^*}-\theta_{b_{P_n}}<0$ is ruled out since that implies

$$P_n\left((\theta_{\hat{\eta}}-\theta_{\hat{b}})\geqslant 0\vee(\theta_{\hat{\eta}}-\theta_{\hat{b}})\leqslant\bar{c}_3\right)\to 0.$$

If $\lim_{n\to\infty}\theta_{\eta_{P_n}^*}-\theta_{b_{P_n}}\geqslant 0$, the result follows from (B.15) and Proposition B.1.

For the second result, note that (B.15) and (B.16) also hold unconditionally. For any sequence with $\lim_{n\to\infty}\theta_{\eta_{P_n}^*}-\theta_{b_{P_n}}<0$, the result follows from (B.16), and for sequences with $\lim_{n\to\infty}\theta_{\eta_{P_n}^*}-\theta_{b_{P_n}}>0$ it holds from (B.15). If $\lim_{n\to\infty}\theta_{\eta_{P_n}^*}-\theta_{b_{P_n}}=0$, Assumption 4.2 implies

$$\sqrt{n}\delta_{\hat{\eta}}\xrightarrow{P_n}0,$$

and the result follows from (B.15). ■

*Proof of Theorem 4.5.* Follows as in the proof of Theorem 4.4, except for sequences with $-\bar{c}_5\leqslant\lim_{n\to\infty}\theta_{\eta_{P_n}^*}-\theta_{b_{P_n}}\leqslant 0$, where the result follows from using (B.15) and Proposition B.1. ■

## B.3 Proofs and Extra Definitions of Section 5

### B.3.1 Proofs and Extra Definitions of Section 5.1

**Assumption B.3.** *The following conditions hold:*

(i) *For every $\varepsilon>0$,*

$$P\left(\sup_{\theta\in\Theta'}\left\|\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\hat{\Psi}_{\mathsf{s},\hat{\eta}_{\mathsf{s}}}(\theta)-\hat{\Psi}_D(\theta)\right\|>\varepsilon\;\middle|\;D\right)\xrightarrow{P}0,$$

*where*

$$\hat{\Psi}_D(\theta)=\mathbb{E}_P\left[\hat{\Psi}_{\xi,\hat{\eta}_{\tilde{\xi}}}(\theta)\mid D\right],$$

*and $\xi$ is a random subset of $[n]$ of size $b$ (as defined in Section 2);*

(ii) *For every* $\varepsilon > 0$,

$$\sup_{\left\|\theta-\hat{\theta}_D\right\|>\varepsilon} -\left\|\hat{\Psi}_D(\theta)\right\| < 0 = \left\|\hat{\Psi}_D(\hat{\theta}_D)\right\|$$

*with probability* 1, *where* $\hat{\theta}_D$ *uniquely solves* $\left\|\hat{\Psi}_D(\hat{\theta}_D)\right\| = 0$.

□

*Proof of Proposition 5.1.* For $j \in \{1, 3\}$, the result follows from a Law of Large Numbers since $\hat{\theta}^{(j)}_{\hat{\eta}}$ is an average of $M$ iid observations (conditional on data). Note that convergence in probability to a point implies convergence of the variance to zero given uniform square integrability (Assumption E.1(iv)). For $j = 2$, consistency follows from consistency of M-estimators (for example, Theorem 5.9 in Van der Vaart (2000)). ■

*Proof of Proposition 5.2.* Let $X(r) = \frac{1}{K}\sum_{\mathsf{s}\in r}\hat{\theta}^{(1)}_{\hat{\eta}_{\mathsf{s}}}$ if $j = 1$ and $X(r) = \hat{\theta}^{(2)}_{\hat{\eta}_r}$ if $j = 3$. Then,

$$\hat{\theta}^{(j)}_{\hat{\eta}} = \frac{1}{M}\sum_{r\in\mathcal{R}} X(r),$$

and $X(r) \perp X(r')$ conditional on $D$ for $r \neq r'$. It follows that

$$\mathrm{Var}_P\left[\hat{\theta}^{(j)}_{\hat{\eta}} \mid D\right] = \frac{1}{M}\mathrm{Var}_P\left[X(r) \mid D\right]$$

is strictly decreasing in $M$ as long as $\mathrm{Var}_P\left[X(r) \mid D\right] > 0$. ■

### B.3.2 Proofs and Extra Definitions of Section 5.2

I first define some objects used in the proofs.

$$g_\theta(r) = \frac{1}{K}\sum_{\mathsf{s}\in r} v_D^{-1}\Psi_{\hat{\eta}_{\mathsf{s}}}(\theta), \quad G_{\hat{\eta}}(\theta) = \frac{1}{M}\sum_{r\in\mathcal{R}} g_\theta(r), \quad G_{\bar{\eta}}(\theta) = \mathbb{E}_P\left[g_\theta(r)|D\right],$$

and $\theta_{\bar{\eta}}$ uniquely solves $G_{\bar{\eta}}(\theta_{\bar{\eta}}) = 0$. Note that $\theta_{\hat{\eta}}$ uniquely solves $G_{\hat{\eta}}(\theta_{\hat{\eta}}) = 0$.

$$\dot{G}_{\hat{\eta}} = \frac{1}{M}\sum_{r\in\mathcal{R}}\frac{1}{K}\sum_{\mathsf{s}\in r} v_D^{-1}\dot{\Psi}_{\hat{\eta}_{\mathsf{s}}}(\theta_{\hat{\eta}_{\mathsf{s}}});$$

$$\dot{G}_{\bar{\eta}} = \mathbb{E}_P\left[\frac{1}{K}\sum_{\mathsf{s}\in r} v_D^{-1}\dot{\Psi}_{\hat{\eta}_{\mathsf{s}}}(\theta_{\bar{\eta}}) \;\Big|\; D\right];$$

$$\hat{V}_G = \left(\frac{1}{MK}\sum_{r\in\mathcal{R}_1}\sum_{\mathsf{s}\in r}\hat{\Psi}_{\mathsf{s},\hat{\eta}_{\bar{\mathsf{s}}}}(\hat{\theta}_{\hat{\eta}_1})\right)\left(\frac{1}{MK}\sum_{r\in\mathcal{R}_1}\sum_{\mathsf{s}\in r}\hat{\Psi}_{\mathsf{s},\hat{\eta}_{\bar{\mathsf{s}}}}(\hat{\theta}_{\hat{\eta}_1})\right)^T;$$

$$\hat{\mathcal{V}}_{\hat{\eta}}(\theta) = \left(\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\frac{1}{b}\sum_{i\in\mathsf{s}}\psi_{\theta,\hat{\eta}_{\bar{\mathsf{s}}}}(W_i)\psi^T_{\theta,\hat{\eta}_{\bar{\mathsf{s}}}}(W_i)\right);$$

$$\hat{\mathcal{V}}_{\eta_P^*}(\theta) = \left(\frac{1}{n}\sum_{i=1}^{n}\psi_{\theta,\eta_P^*}(W_i)\psi^T_{\theta,\eta_P^*}(W_i)\right);$$

$\mathcal{V}_\eta(\theta) = P\hat{\mathcal{V}}_\eta(\theta)$. Note that

$$\sqrt{n}\left(\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\frac{1}{b}\sum_{i\in\mathsf{s}}\psi_{\theta,\eta_P^*}(W_i)\psi^T_{\theta,\eta_P^*}(W_i)\right) = \sqrt{n}\hat{\mathcal{V}}_{\theta,\eta_P^*} + o_P(1)$$

from (E.11) (the equality holds without $o_P(1)$ if $K > 1$).

$$\sigma^2_{\hat{\eta}} = V_{M,K}\dot{h}(\hat{\theta}_{\hat{\eta}})\hat{\dot{\Psi}}^{-1}_{\hat{\eta}}\mathcal{V}_{\hat{\eta}}(\hat{\theta}_{\hat{\eta}})\left(\hat{\dot{\Psi}}^{-1}_{\hat{\eta}}\right)^T\dot{h}(\hat{\theta}_{\hat{\eta}})^T;$$

$$v^2_D = \mathrm{Var}_P\left[\sigma^{-1}_{\eta_P^*}\dot{h}(\theta_{\bar{\eta}})\dot{G}^{-1}_{\bar{\eta}}\sqrt{M}G_{\hat{\eta}}(\theta_{\bar{\eta}}) \mid D\right];$$

$$\hat{v}^2_D = \hat{\sigma}^{-2}_{\hat{\eta}_1}\dot{h}(\hat{\theta}_{\hat{\eta}_1})\hat{\dot{\Psi}}^{-1}_{\hat{\eta}_1}\hat{V}_G\left(\hat{\dot{\Psi}}^{-1}_{\hat{\eta}_1}\right)^T\dot{h}(\hat{\theta}_{\hat{\eta}_1})^T;$$

$$\zeta^2_D = \mathrm{Var}_P\left[2^{-1}\sigma^{-3}_{\eta_P^*}(h(\theta_{\eta_P^*}) - \tau)V_{M,K}\dot{h}(\theta_{\eta_P^*})\dot{\Psi}^{-1}_{\eta_P^*}\sqrt{M}\mathcal{V}_{\hat{\eta}}(\hat{\theta}_{\hat{\eta}})\left(\dot{\Psi}^{-1}_{\eta_P^*}\right)^T\dot{h}(\theta_{\eta_P^*})^T \;\middle|\; D\right];$$

$$\hat{a} = (\hat{a}_1\ \cdots\ \hat{a}_d) = \dot{h}(\hat{\theta}_{\hat{\eta}_1})\hat{\dot{\Psi}}^{-1}_{\hat{\eta}_1};$$

$\hat{v}_{(i,j)}$ are the entries of $\hat{\mathcal{V}}_{\hat{\eta}_1}(\hat{\theta}_{\hat{\eta}_1})$,

$$\hat{c}_{(i,j),(i',j')} = \frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\left(\frac{1}{b}\sum_{\ell\in s}\psi_{\theta,\hat{\eta}_{\bar{\mathsf{s}}},i}(W_\ell)\psi_{\theta,\hat{\eta}_{\bar{\mathsf{s}}},j}(W_\ell) - \hat{v}_{(i,j)}\right)$$
$$\times\left(\frac{1}{b}\sum_{\ell\in s}\psi_{\theta,\hat{\eta}_{\bar{\mathsf{s}}},i'}(W_\ell)\psi_{\theta,\hat{\eta}_{\bar{\mathsf{s}}},j'}(W_\ell) - \hat{v}_{(i',j')}\right);$$

$$\hat{\zeta}^2_D = 2^{-2}\hat{\sigma}^{-6}_{\hat{\eta}_1}(h(\hat{\theta}_{\hat{\eta}_1}) - \tau)^2V^2_{M,K}\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{i'=1}^{d}\sum_{j'=1}^{d}\hat{a}_i\hat{a}_j\hat{a}_{i'}\hat{a}_{j'}\hat{c}_{(i,j),(i',j')};$$

$$\rho_{v,\zeta} = \mathrm{Cov}_P \Bigg[ \sigma_{\eta_P^*}^{-1} \dot{h}(\theta_{\bar{\eta}}) \dot{G}_{\bar{\eta}}^{-1} \sqrt{M} G_{\hat{\eta}}(\theta_{\bar{\eta}}),$$

$$2^{-1} \sigma_{\eta_P^*}^{-3} (h(\theta_{\eta_P^*}) - \tau) V_{M,K} \dot{h}(\theta_{\eta_P^*}) \dot{\Psi}_{\eta_P^*}^{-1} \sqrt{M} \mathcal{V}_{\hat{\eta}}(\hat{\theta}_{\hat{\eta}}) \left( \dot{\Psi}_{\eta_P^*}^{-1} \right)^T \dot{h}(\theta_{\eta_P^*})^T \,\Bigg|\, D \Bigg];$$

$$\hat{d}_{i,(j,\ell)} = \frac{1}{MK} \sum_{r \in \mathcal{R}} \sum_{\mathsf{s} \in r} \left( \hat{\Psi}_{\mathsf{s},\hat{\eta}_{\bar{\mathsf{s}}},i}(\hat{\theta}_{\hat{\eta}_1}) - \hat{\Psi}_{\hat{\eta}_1,i}(\hat{\theta}_{\hat{\eta}_1}) \right) \left( \frac{1}{b} \sum_{i' \in s} \psi_{\theta,\hat{\eta}_{\bar{\mathsf{s}}},j}(W_{i'}) \psi_{\theta,\hat{\eta}_{\bar{\mathsf{s}}},\ell}(W_{i'}) - \hat{v}_{(j,\ell)} \right);$$

$$\hat{\rho}_{v,\zeta} = 2^{-1} \hat{\sigma}_{\hat{\eta}_1}^{-4} (h(\hat{\theta}_{\hat{\eta}_1}) - \tau) V_{M,K} \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{\ell=1}^{d} \hat{a}_i \hat{a}_j \hat{a}_\ell \hat{d}_{i,(j,\ell)};$$

$$\sigma_D^2 = 2(v_D^2 + \zeta_D^2 + 2\rho_{v,\zeta}); \tag{B.17}$$

$$\hat{\sigma}_D^2 = 2(\hat{v}_D^2 + \hat{\zeta}_D^2 + 2\hat{\rho}_{v,\zeta}).$$

**Assumption B.4.** *The following conditions hold:*

(i) *For any $\delta_n \downarrow 0$ and $\varepsilon > 0$,*

$$P \left( \sup_{\|\theta - \theta'\| < \delta_n} \left\| \sqrt{M} \left( G_{\hat{\eta}} - G_{\bar{\eta}} \right)(\theta) - \sqrt{M} \left( G_{\hat{\eta}} - G_{\bar{\eta}} \right)(\theta') \right\| > \varepsilon \,\middle|\, D \right) \xrightarrow{P} 0$$

*uniformly in $P \in \mathcal{P}$;*

(ii) *For $(i,j) \in [d]^2$, Assumption E.1 holds with $T = \Theta'$ and*

$$\mathcal{F}_\eta = \{ \psi_{\theta,\eta,i} \psi_{\theta,\eta,j} : \theta \in \Theta' \},$$

*where $\psi_{\theta,\eta,i}$ is the $i$-th coordinate of $\psi_{\theta,\eta}$;*

(iii) *There exists an estimator $\hat{\dot{\Psi}}_{\hat{\eta}}$ such that*

$$\left\| \hat{\dot{\Psi}}_{\hat{\eta}} - \dot{\Psi}_{\eta_P^*} \right\| \xrightarrow{P} 0$$

*uniformly in $P \in \mathcal{P}$;*

(iv) *For some $\underline{v} > 0$,*

$$\sigma_{\eta_P^*}^2 = \dot{h}(\theta_{\eta_P^*}) V_{\eta_P^*} \dot{h}(\theta_{\eta_P^*})^T \geqslant \underline{v};$$

(v) $M^{-1} n \sigma_D^2 = O_P(1)$.

(vi) *Either*

$$v_D^{-1}\zeta_D \xrightarrow{P} c_1 \neq 1$$

*or*

$$2\frac{\rho_{v,\zeta}}{\zeta_D v_D} \xrightarrow{P} c_2 \neq -1.$$

□

Assumption B.4(i) is a Donsker condition on $\left\{v_D^{-1}\Psi_{\hat{\eta}_s} : s \subseteq [n]\right\}$ conditional on the data. It is similar to Assumption B.1, and can typically be verified using arguments similar to the ones used to verify Assumption E.1(vi). It holds, for example, if $\Theta'$ is bounded and $\psi_{\theta,\eta}$ is Lipschitz in $\theta$ with a Lipschitz constant that does not depend on $\eta$ or $w$ (see, e.g., Example 19.7 in Van der Vaart, 2000). Assumption B.4(ii) is a Donsker condition similar to Assumption B.1(i), but in terms of the product $\psi_{\theta,\eta,i}\psi_{\theta,\eta,j}$ instead of $\psi_{\theta,\eta,i}$. It is used to derive asymptotic normality of the standard errors $\hat{\sigma}_{\hat{\eta}_1}, \hat{\sigma}_{\hat{\eta}_2}$. If $\psi_{\theta,\eta,i}(w) \leqslant \bar{C}$ for some $\bar{C} < \infty$, that is, if the functions $\psi_{\theta,\eta,i}$ are uniformly bounded, then Assumption B.4(ii) is implied by Assumption B.1(i) (see, e.g., Example 2.10.10 in van der Vaart and Wellner, 2023). Assumption B.4(iii) assumes the existence of a consistent estimator of $\dot{\Psi}_{\eta_P^*}$. If $\psi_{\theta,\eta}(w)$ is differentiable in $\theta$, the plug-in estimator defined in (B.10) satisfies this assumption under a uniform integrability condition on this derivative. Otherwise, consistent estimators can typically be constructed on a case-by-case basis (Hansen, 2022). Assumption B.4(iv) requires the asymptotic variance of $h(\hat{\theta}_{\hat{\eta}})$ to be lower bounded. Assumption B.4(v) establishes the asymptotic regime. Finally, Assumption B.4(vi) restricts a corner case where the variance of the t-statistic $\hat{\sigma}_{\hat{\eta}_1}^{-1}\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1}) - \tau)$ is zero because of perfect negative correlation between $\hat{\sigma}_{\hat{\eta}_1}^{-1}$ and $h(\hat{\theta}_{\hat{\eta}_1})$. Note that the quantities $\rho_{v,\zeta}, \zeta_D, v_D$ can all be consistently estimated with $\hat{\rho}_{v,\zeta}, \hat{\zeta}_D, \hat{v}_D$ defined previously.

Before proving Theorem 5.1, I establish some key intermediary results.

**Lemma B.3.** *Let the conditions of Theorem 5.1 hold. Then,*

$$\sigma_D^{-1}v_D = O_P(1), \quad \sigma_D^{-1}\zeta_D = O_P(1).$$

*Proof.* I show $\sigma_D^{-1}v_D = O_P(1)$ and the second result follows analogously.

$$\sigma_D^2 = 2(v_D^2 + \zeta_D^2 + 2\rho_{v,\zeta}),$$

$$\sigma_D^{-2}v_D^2 = 2^{-1}(1 + v_D^{-2}\zeta_D^2 + 2v_D^{-2}\rho_{v,\zeta})^{-1},$$

$$
\begin{aligned}
v_D^{-2}\zeta_D^2 + 2v_D^{-2}\rho_{v,\zeta} &= v_D^{-2}\zeta_D^2 + 2v_D^{-1}\zeta_D\frac{\rho_{v,\zeta}}{\zeta_D v_D} \\
&= v_D^{-1}\zeta_D\left(v_D^{-1}\zeta_D + 2\frac{\rho_{v,\zeta}}{\zeta_D v_D}\right) \\
&= \begin{cases} O_P(1), \text{ if } v_D^{-1}\zeta_D = O_P(1), \\ o_P(1), \text{ if } v_D\zeta_D^{-1} = o_P(1), \end{cases}
\end{aligned}
$$

since $\left|\frac{\rho_{v,\zeta}}{\zeta_D v_D}\right| \leqslant 1$. Note that

$$
v_D^{-1}\zeta_D\left(v_D^{-1}\zeta_D + 2\frac{\rho_{v,\zeta}}{\zeta_D v_D}\right) \xrightarrow{P} -1 \iff v_D^{-1}\zeta_D \xrightarrow{P} 1 \wedge 2\frac{\rho_{v,\zeta}}{\zeta_D v_D} \xrightarrow{P} -1,
$$

which is ruled out by Assumption B.4(vi). ■

**Theorem B.2.** *Let Assumption B.4 hold. Then, for any $\varepsilon > 0$,*

$$
P\left(\left\|v_D^{-1}\sqrt{M}\left(\theta_{\hat{\eta}} - \theta_{\bar{\eta}}\right) - \left(-v_D^{-1}\dot{G}_{\bar{\eta}}^{-1}\sqrt{M}G_{\hat{\eta}}(\theta_{\bar{\eta}})\right)\right\| > \varepsilon \;\middle|\; D\right) \xrightarrow{P} 0
$$

*uniformly in $P \in \mathcal{P}$, and hence*

$$
\sup_{P\in\mathcal{P}} P\left(\left\|v_D^{-1}\sqrt{M}\left(\theta_{\hat{\eta}} - \theta_{\bar{\eta}}\right) - \left(-v_D^{-1}\dot{G}_{\bar{\eta}}^{-1}\sqrt{M}G_{\hat{\eta}}(\theta_{\bar{\eta}})\right)\right\| > \varepsilon\right) \to 0.
$$

*Moreover,*

$$
v_D^{-1}\dot{G}_{\bar{\eta}}^{-1}\sqrt{M}G_{\hat{\eta}}(\theta_{\bar{\eta}}) = O_P(1).
$$

□

*Proof of Theorem B.2.* For a random variable $X_M$ and a deterministic (conditional on $D$) sequence $a_M(D)$, I use $X_M = o_{P|D}(a_M(D))$ to denote

$$
P\left(\left\|\frac{X_M}{a_M(D)}\right\| > \varepsilon \;\middle|\; D\right) \xrightarrow{P} 0
$$

uniformly in $P \in \mathcal{P}$ for any $\varepsilon > 0$, and analogously define $O_{P|D}(a_M(D))$ similar to the $O_P$ notation.

By differentiability of $G$,

$$
\begin{aligned}
&v_D^{-1}\sqrt{M}\left(G_{\bar{\eta}}(\theta_{\hat{\eta}}) - G_{\bar{\eta}}(\theta_{\bar{\eta}})\right) \\
&= v_D^{-1}\sqrt{M}\dot{G}_{\bar{\eta}}\left(\theta_{\hat{\eta}} - \theta_{\bar{\eta}}\right) + o_{P|D}\left(\left\|v_D^{-1}\dot{G}_{\bar{\eta}}^{-1}\sqrt{M}\left(\theta_{\hat{\eta}} - \theta_{\bar{\eta}}\right)\right\|\right). \qquad \text{(B.18)}
\end{aligned}
$$

Further,

$$\sqrt{M}\left(G_{\bar{\eta}}(\theta_{\hat{\eta}}) - G_{\bar{\eta}}(\theta_{\bar{\eta}})\right) = -\sqrt{M}\left(G_{\hat{\eta}}(\theta_{\hat{\eta}}) - G_{\bar{\eta}}(\theta_{\hat{\eta}})\right) \tag{B.19}$$

$$= -\sqrt{M}\left(G_{\hat{\eta}}(\theta_{\bar{\eta}}) - G_{\bar{\eta}}(\theta_{\bar{\eta}})\right) + o_{P|D}(1) \tag{B.20}$$

$$= -\sqrt{M}G_{\hat{\eta}}(\theta_{\bar{\eta}}) + o_{P|D}(1) \tag{B.21}$$

$$= O_{P|D}(1). \tag{B.22}$$

(B.19) uses the definitions $G_{\bar{\eta}}(\theta_{\bar{\eta}}) = G_{\hat{\eta}}(\theta_{\hat{\eta}}) = 0$, and (B.20) uses Assumption B.4(i). (B.22) follows from the Lindeberg CLT.

Combining (B.18) and (B.21) gives

$$\begin{aligned}
&v_D^{-1}\sqrt{M}\left(\theta_{\hat{\eta}} - \theta_{\bar{\eta}}\right) \\
&= -v_D^{-1}\dot{G}_{\bar{\eta}}^{-1}\sqrt{M}G_{\hat{\eta}}(\theta_{\bar{\eta}}) + o_{P|D}\left(\left\|v_D^{-1}\dot{G}_{\bar{\eta}}^{-1}\sqrt{M}\left(\theta_{\hat{\eta}} - \theta_{\bar{\eta}}\right)\right\|\right) + o_{P|D}\left(\left\|v_D^{-1}\dot{G}_{\bar{\eta}}^{-1}\right\|\right) \\
&= -v_D^{-1}\dot{G}_{\bar{\eta}}^{-1}\sqrt{M}G_{\hat{\eta}}(\theta_{\bar{\eta}}) + o_{P|D}\left(1\right),
\end{aligned}$$

since

$$v_D^{-1}\dot{G}_{\bar{\eta}}^{-1} = \mathbb{E}_P\left[\frac{1}{K}\sum_{\mathsf{s}\in r}\dot{\Psi}_{\hat{\eta}_{\mathsf{s}}}(\theta_{\bar{\eta}}) \;\middle|\; D\right]^{-1} = \dot{\Psi}_{\eta_P^*}^{-1}(\theta_{\eta_P^*}) + o_P(1) = O_P(1)$$

by Assumption B.1(v), and an argument similar to (B.9), exploring (B.22), gives

$$\sqrt{M}\left\|\theta_{\hat{\eta}} - \theta_{\bar{\eta}}\right\| = O_{P|D}(1).$$

The second result follows since, for any events $A$ and $B$,

$$P(A|B) = o_P(1) \implies \sup_{P\in\mathcal{P}} P(A) = \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[P(A|B)\right] \to 0.$$

Finally,

$$\left\|v_D^{-1}\dot{G}_{\bar{\eta}}^{-1}\sqrt{M}G_{\hat{\eta}}(\theta_{\bar{\eta}})\right\| \leqslant \left\|v_D^{-1}\dot{G}_{\bar{\eta}}^{-1}\right\|\left\|\sqrt{M}G_{\hat{\eta}}(\theta_{\bar{\eta}})\right\| = O_P(1)O_{P|D}(1).$$

■

*Proof of Theorem 5.1.* The proof is divided into three main steps. First, I show that

$$v_D^{-1}\sqrt{M}(h(\hat{\theta}_{\hat{\eta}_1}) - h(\hat{\theta}_{\hat{\eta}_2})) = -\dot{h}(\theta_{\bar{\eta}})v_D^{-1}\dot{G}_{\bar{\eta}}^{-1}\sqrt{M}\left(G_{\hat{\eta}_1}(\theta_{\bar{\eta}}) - G_{\hat{\eta}_2}(\theta_{\bar{\eta}})\right) + o_P\left(1\right). \tag{B.23}$$

Second, I show that

$$\zeta_D^{-1}\sqrt{M}\left(\hat{\sigma}_{\hat{\eta}_1} - \hat{\sigma}_{\hat{\eta}_2}\right) \tag{B.24}$$

$$= (2\sigma_{\eta_P^*})^{-1}V_{M,K}\dot{h}(\theta_{\eta_P^*})\dot{\Psi}_{\eta_P^*}^{-1}\zeta_D^{-1}\sqrt{M}\left(\mathcal{V}_{\hat{\eta}_1}(\hat{\theta}_{\hat{\eta}_1}) - \mathcal{V}_{\hat{\eta}_2}(\hat{\theta}_{\hat{\eta}_2})\right)\left(\dot{\Psi}_{\eta_P^*}^{-1}\right)^T\dot{h}(\theta_{\eta_P^*})^T + o_P(1).$$

Finally, I combine the previous steps to reach the result.

**Step one.**

$$
\begin{aligned}
&v_D^{-1}\sqrt{M}(h(\hat{\theta}_{\hat{\eta}_1}) - h(\hat{\theta}_{\hat{\eta}_2})) \\
&= v_D^{-1}\sqrt{M}(h(\hat{\theta}_{\hat{\eta}_1}) - h(\theta_{\hat{\eta}_1})) - v_D^{-1}\sqrt{M}(h(\hat{\theta}_{\hat{\eta}_2}) - h(\theta_{\hat{\eta}_2})) + v_D^{-1}\sqrt{M}(h(\theta_{\hat{\eta}_1}) - h(\theta_{\hat{\eta}_2})) \\
&= v_D^{-1}\sqrt{M}(h(\theta_{\hat{\eta}_1}) - h(\theta_{\hat{\eta}_2})) + o_P(1),
\end{aligned}
$$

since

$$
\begin{aligned}
\sqrt{n}\left(h(\hat{\theta}_{\hat{\eta}_1}) - h(\theta_{\hat{\eta}_1})\right) &= \dot{h}(\theta_{\hat{\eta}_1})\sqrt{n}(\hat{\theta}_{\hat{\eta}_1} - \theta_{\hat{\eta}_1}) + o_P\left(\sqrt{n}\left\|\hat{\theta}_{\hat{\eta}_1} - \theta_{\hat{\eta}_1}\right\|\right) \\
&= \dot{h}(\theta_{\hat{\eta}_1})\dot{\Psi}_{\eta_P^*}^{-1}\sqrt{n}\hat{\Psi}_{\hat{\eta}_1}(\theta_{\hat{\eta}_1}) + o_P(1) \\
&= \dot{h}(\theta_{\eta_P^*})\dot{\Psi}_{\eta_P^*}^{-1}\sqrt{n}\left(\hat{\Psi}_{\eta_P^*}(\theta_{\eta_P^*}) - \Psi_{\eta_P^*}(\theta_{\eta_P^*})\right) + o_P(1),
\end{aligned}
$$

where $o_P\left(\sqrt{n}\left\|\hat{\theta}_{\hat{\eta}_1} - \theta_{\hat{\eta}_1}\right\|\right) = o_P(1)$ by Theorem 3.1, the second equality holds from Theorem 3.1, and the last equality from Theorem E.1, using the fact that $\left\|\theta_{\hat{\eta}_1} - \theta_{\eta_P^*}\right\| = o_P(1)$. Note that $v_D^{-1}\sqrt{M}/\sqrt{n} = O_P(1)$ from Lemma B.3.

By differentiability of $h$,

$$
v_D^{-1}\sqrt{M}\left(h(\theta_{\hat{\eta}_1}) - h(\theta_{\bar{\eta}})\right) = v_D^{-1}\sqrt{M}\dot{h}(\theta_{\bar{\eta}})(\theta_{\hat{\eta}_1} - \theta_{\bar{\eta}}) + o_P\left(1\right),
$$

since $v_D^{-1}\sqrt{M}\left\|\theta_{\hat{\eta}_1} - \theta_{\bar{\eta}}\right\| = O_P(1)$ from Theorem B.2. This implies

$$
v_D^{-1}\sqrt{M}\left(h(\theta_{\hat{\eta}_1}) - h(\theta_{\hat{\eta}_2})\right) = \dot{h}(\theta_{\bar{\eta}})v_D^{-1}\sqrt{M}(\theta_{\hat{\eta}_1} - \theta_{\hat{\eta}_2}) + o_P\left(1\right).
$$

Theorem B.2 gives

$$
\begin{aligned}
v_D^{-1}\sqrt{M}\left(\theta_{\hat{\eta}_1} - \theta_{\hat{\eta}_2}\right) &= v_D^{-1}\sqrt{M}\left(\theta_{\hat{\eta}_1} - \theta_{\bar{\eta}}\right) - v_D^{-1}\sqrt{M}\left(\theta_{\hat{\eta}_2} - \theta_{\bar{\eta}}\right) \\
&= -\dot{G}_{\bar{\eta}}^{-1}v_D^{-1}\sqrt{M}\left(G_{\hat{\eta}_1}(\theta_{\bar{\eta}}) - G_{\hat{\eta}_2}(\theta_{\bar{\eta}})\right) + o_P(1).
\end{aligned}
$$

(B.23) follows from combining the two previous displays.

**Step two.**

$$
\begin{aligned}
\zeta_D^{-1}\sqrt{M}(\hat{\sigma}^2_{\hat{\eta}_1} - \hat{\sigma}^2_{\hat{\eta}_2}) &= \zeta_D^{-1}\sqrt{M}(\hat{\sigma}^2_{\hat{\eta}_1} - \sigma^2_{\hat{\eta}_1}) - \zeta_D^{-1}\sqrt{M}(\hat{\sigma}^2_{\hat{\eta}_2} - \sigma^2_{\hat{\eta}_2}) + \zeta_D^{-1}\sqrt{M}(\sigma^2_{\hat{\eta}_1} - \sigma^2_{\hat{\eta}_2}) \\
&= \zeta_D^{-1}\sqrt{M}(\sigma^2_{\hat{\eta}_1} - \sigma^2_{\hat{\eta}_2}) + o_P(1),
\end{aligned}
$$

since

$$
\begin{aligned}
&\zeta_D^{-1}\sqrt{M}(\hat{\sigma}^2_{\hat{\eta}_1} - \sigma^2_{\hat{\eta}_1}) - \zeta_D^{-1}\sqrt{M}(\hat{\sigma}^2_{\hat{\eta}_2} - \sigma^2_{\hat{\eta}_2}) \\
&= \left(\frac{\sqrt{M}}{\sqrt{n}}\zeta_D^{-1}\right)\left(\sqrt{n}(\hat{\sigma}^2_{\hat{\eta}_1} - \sigma^2_{\hat{\eta}_1}) - \sqrt{n}(\hat{\sigma}^2_{\hat{\eta}_2} - \sigma^2_{\hat{\eta}_2})\right) \\
&= O_P(1)\left(\sqrt{n}(\hat{\sigma}^2_{\hat{\eta}_1} - \sigma^2_{\hat{\eta}_1}) - \sqrt{n}(\hat{\sigma}^2_{\hat{\eta}_2} - \sigma^2_{\hat{\eta}_2})\right),
\end{aligned}
$$

and

$$\begin{aligned}
&\sqrt{n}(\hat{\sigma}^2_{\hat{\eta}_1} - \sigma^2_{\hat{\eta}_1}) - \sqrt{n}(\hat{\sigma}^2_{\hat{\eta}_2} - \sigma^2_{\hat{\eta}_2}) \\
&= V_{M,K}\dot{h}(\hat{\theta}_{\hat{\eta}_1})\hat{\dot{\Psi}}^{-1}_{\hat{\eta}_1}\sqrt{n}\left(\hat{\mathcal{V}}_{\hat{\eta}_1}(\hat{\theta}_{\hat{\eta}_1}) - \mathcal{V}_{\hat{\eta}_1}(\hat{\theta}_{\hat{\eta}_1})\right)\left(\hat{\dot{\Psi}}^{-1}_{\hat{\eta}_1}\right)^T \dot{h}(\hat{\theta}_{\hat{\eta}_1})^T \\
&\quad - V_{M,K}\dot{h}(\hat{\theta}_{\hat{\eta}_2})\hat{\dot{\Psi}}^{-1}_{\hat{\eta}_2}\sqrt{n}\left(\hat{\mathcal{V}}_{\hat{\eta}_2}(\hat{\theta}_{\hat{\eta}_2}) - \mathcal{V}_{\hat{\eta}_2}(\hat{\theta}_{\hat{\eta}_2})\right)\left(\hat{\dot{\Psi}}^{-1}_{\hat{\eta}_2}\right)^T \dot{h}(\hat{\theta}_{\hat{\eta}_2})^T \\
&= V_{M,K}\dot{h}(\hat{\theta}_{\hat{\eta}_1})\hat{\dot{\Psi}}^{-1}_{\hat{\eta}_1}\sqrt{n}\left(\hat{\mathcal{V}}_{\eta^*_P}(\theta_{\bar{\eta}}) - \mathcal{V}_{\eta^*_P}(\theta_{\bar{\eta}})\right)\left(\hat{\dot{\Psi}}^{-1}_{\hat{\eta}_1}\right)^T \dot{h}(\hat{\theta}_{\hat{\eta}_1})^T \\
&\quad - V_{M,K}\dot{h}(\hat{\theta}_{\hat{\eta}_2})\hat{\dot{\Psi}}^{-1}_{\hat{\eta}_2}\sqrt{n}\left(\hat{\mathcal{V}}_{\eta^*_P}(\theta_{\bar{\eta}}) - \mathcal{V}_{\eta^*_P}(\theta_{\bar{\eta}})\right)\left(\hat{\dot{\Psi}}^{-1}_{\hat{\eta}_2}\right)^T \dot{h}(\hat{\theta}_{\hat{\eta}_2})^T + o_P(1) \\
&= o_P(1),
\end{aligned}$$

where the second equality follows from Assumption B.4(ii) and Theorem E.1, and the last equality uses $\sqrt{n}\left(\hat{\mathcal{V}}_{\eta^*_P}(\theta_{\bar{\eta}}) - \mathcal{V}_{\eta^*_P}(\theta_{\bar{\eta}})\right) = O_P(1)$.

Finally,

$$\zeta_D^{-1}\sqrt{M}(\sigma_{\hat{\eta}_1} - \sigma_{\hat{\eta}_2}) = \frac{\zeta_D^{-1}\sqrt{M}(\sigma^2_{\hat{\eta}_1} - \sigma^2_{\hat{\eta}_2})}{\sigma_{\hat{\eta}_1} + \sigma_{\hat{\eta}_2}} = (2\sigma_{\eta^*_P})^{-1}\zeta_D^{-1}\sqrt{M}(\sigma^2_{\hat{\eta}_1} - \sigma^2_{\hat{\eta}_2}) + o_P(1),$$

and

$$\begin{aligned}
&\zeta_D^{-1}\sqrt{M}(\sigma^2_{\hat{\eta}_1} - \sigma^2_{\hat{\eta}_2}) \\
&= V_{M,K}\dot{h}(\hat{\theta}_{\hat{\eta}_1})\hat{\dot{\Psi}}^{-1}_{\hat{\eta}_1}\zeta_D^{-1}\sqrt{M}\left(\mathcal{V}_{\hat{\eta}_1}(\hat{\theta}_{\hat{\eta}_1}) - \mathcal{V}_{\hat{\eta}_2}(\hat{\theta}_{\hat{\eta}_2})\right)\left(\hat{\dot{\Psi}}^{-1}_{\hat{\eta}_1}\right)^T \dot{h}(\hat{\theta}_{\hat{\eta}_1})^T + o_P(1) \\
&= V_{M,K}\dot{h}(\theta_{\eta^*_P})\dot{\Psi}^{-1}_{\eta^*_P}\zeta_D^{-1}\sqrt{M}\left(\mathcal{V}_{\hat{\eta}_1}(\hat{\theta}_{\hat{\eta}_1}) - \mathcal{V}_{\hat{\eta}_2}(\hat{\theta}_{\hat{\eta}_2})\right)\left(\dot{\Psi}^{-1}_{\eta^*_P}\right)^T \dot{h}(\theta_{\eta^*_P})^T + o_P(1),
\end{aligned}$$

using the fact that $\zeta_D^{-1}\sqrt{M}\left(\mathcal{V}_{\hat{\eta}_1}(\hat{\theta}_{\hat{\eta}_1}) - \mathcal{V}_{\hat{\eta}_2}(\hat{\theta}_{\hat{\eta}_2})\right) = O_P(1)$.

**Step three.**

$$
\begin{aligned}
&\left(\frac{\sqrt{n}\hat{\sigma}_D}{\sqrt{M}}\right)^{-1}\left(\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1})-\tau)}{\hat{\sigma}_{\hat{\eta}_1}}-\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_2})-\tau)}{\hat{\sigma}_{\hat{\eta}_2}}\right)\\
&=\hat{\sigma}_D^{-1}\sqrt{M}\left(\frac{h(\hat{\theta}_{\hat{\eta}_1})-\tau}{\hat{\sigma}_{\hat{\eta}_1}}-\frac{h(\hat{\theta}_{\hat{\eta}_2})-\tau}{\hat{\sigma}_{\hat{\eta}_2}}\right)\\
&=\hat{\sigma}_D^{-1}\sqrt{M}\frac{h(\hat{\theta}_{\hat{\eta}_1})-h(\hat{\theta}_{\hat{\eta}_2})}{\hat{\sigma}_{\hat{\eta}_1}}-\hat{\sigma}_D^{-1}\sqrt{M}(\hat{\sigma}_{\hat{\eta}_1}-\hat{\sigma}_{\hat{\eta}_2})\frac{h(\hat{\theta}_{\hat{\eta}_2})-\tau}{\hat{\sigma}_{\hat{\eta}_1}\hat{\sigma}_{\hat{\eta}_2}}\\
&=\sigma_D^{-1}\sqrt{M}\frac{h(\hat{\theta}_{\hat{\eta}_1})-h(\hat{\theta}_{\hat{\eta}_2})}{\sigma_{\eta_P^*}}-\sigma_D^{-1}\sqrt{M}(\hat{\sigma}_{\hat{\eta}_1}-\hat{\sigma}_{\hat{\eta}_2})\frac{h(\theta_{\eta_P^*})-\tau}{\sigma^2_{\eta_P^*}}+o_P(1)\\
&=-\sigma^{-1}_{\eta_P^*}\dot{h}(\theta_{\bar{\eta}})\dot{G}^{-1}_{\bar{\eta}}\sigma_D^{-1}\sqrt{M}\left(G_{\hat{\eta}_1}(\theta_{\bar{\eta}})-G_{\hat{\eta}_2}(\theta_{\bar{\eta}})\right)+o_P\left(\sigma_D^{-1}v_D\right)\\
&\quad-2^{-1}\sigma^{-3}_{\eta_P^*}(h(\theta_{\eta_P^*})-\tau)V_{M,K}\dot{h}(\theta_{\eta_P^*})\dot{\Psi}^{-1}_{\eta_P^*}\frac{\sqrt{M}}{\sigma_D}\left(\mathcal{V}_{\hat{\eta}_1}(\hat{\theta}_{\hat{\eta}_1})-\mathcal{V}_{\hat{\eta}_2}(\hat{\theta}_{\hat{\eta}_2})\right)\left(\dot{\Psi}^{-1}_{\eta_P^*}\right)^T\dot{h}(\theta_{\eta_P^*})^T\\
&\quad+o_P\left(\sigma_D^{-1}\zeta_D\right)+o_P(1)\\
&\rightsquigarrow\mathcal{N}(0,1),
\end{aligned}
$$

conditional on $D$ with probability approaching one, by Lindeberg's CLT, by definition of $\sigma_D$, and since $G_{\hat{\eta}_1}(\theta_{\bar{\eta}})\perp G_{\hat{\eta}_2}(\theta_{\bar{\eta}}),\mathcal{V}_{\hat{\eta}_2}(\hat{\theta}_{\hat{\eta}_2})$ and $\mathcal{V}_{\hat{\eta}_1}(\hat{\theta}_{\hat{\eta}_1})\perp G_{\hat{\eta}_2}(\theta_{\bar{\eta}}),\mathcal{V}_{\hat{\eta}_2}(\hat{\theta}_{\hat{\eta}_2})$ conditional on $D$. Note that $o_P\left(\sigma_D^{-1}v_D\right),o_P\left(\sigma_D^{-1}\zeta_D\right)=o_P(1)$ by Lemma B.3. ■

*Proof of Theorem 5.2.* For $(p_j,\hat{\delta}(\beta))=(p_j^+,\hat{\delta}^+(\beta))$,

$$
\begin{aligned}
&P\left(p_2^+>p_1^++\hat{\delta}^+(\beta)\,\middle|\,D\right)\\
&=P\left(\Phi\left(\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_2})-\tau)}{\hat{\sigma}_{\hat{\eta}_2}}\right)>\Phi\left(\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1})-\tau)}{\hat{\sigma}_{\hat{\eta}_1}}\right)+\hat{\delta}^+(\beta)\,\middle|\,D\right)\\
&=P\left(\Phi\left(\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_2})-\tau)}{\hat{\sigma}_{\hat{\eta}_2}}\right)>\Phi\left(\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1})-\tau)}{\hat{\sigma}_{\hat{\eta}_1}}-\frac{\sqrt{n}\hat{\sigma}_D}{\sqrt{M}}\Phi^{-1}(\beta)\right)\,\middle|\,D\right)\\
&=P\left(\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_2})-\tau)}{\hat{\sigma}_{\hat{\eta}_2}}-\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1})-\tau)}{\hat{\sigma}_{\hat{\eta}_1}}>-\frac{\sqrt{n}\hat{\sigma}_D}{\sqrt{M}}\Phi^{-1}(\beta)\,\middle|\,D\right)\\
&=P\left(\left(\frac{\sqrt{n}\hat{\sigma}_D}{\sqrt{M}}\right)^{-1}\left(\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1})-\tau)}{\hat{\sigma}_{\hat{\eta}_1}}-\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_2})-\tau)}{\hat{\sigma}_{\hat{\eta}_2}}\right)<\Phi^{-1}(\beta)\,\middle|\,D\right)\\
&=\beta+o_P(1),
\end{aligned}
$$

where the last equality follows from Theorem 5.1.

For $(p_j, \hat{\delta}(\beta)) = (p_j^-, \hat{\delta}^-(\beta))$,

$$
\begin{aligned}
&P\left(p_2^- > p_1^- + \hat{\delta}^-(\beta) \,\middle|\, D\right) \\
&= P\left(\Phi\left(-\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_2}) - \tau)}{\hat{\sigma}_{\hat{\eta}_2}}\right) > \Phi\left(-\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1}) - \tau)}{\hat{\sigma}_{\hat{\eta}_1}}\right) + \hat{\delta}^-(\beta) \,\middle|\, D\right) \\
&= P\left(\Phi\left(-\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_2}) - \tau)}{\hat{\sigma}_{\hat{\eta}_2}}\right) > \Phi\left(-\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1}) - \tau)}{\hat{\sigma}_{\hat{\eta}_1}} - \frac{\sqrt{n}\hat{\sigma}_D}{\sqrt{M}}\Phi^{-1}(\beta)\right) \,\middle|\, D\right) \\
&= P\left(\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1}) - \tau)}{\hat{\sigma}_{\hat{\eta}_1}} - \frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_2}) - \tau)}{\hat{\sigma}_{\hat{\eta}_2}} > -\frac{\sqrt{n}\hat{\sigma}_D}{\sqrt{M}}\Phi^{-1}(\beta) \,\middle|\, D\right) \\
&= P\left(\left(\frac{\sqrt{n}\hat{\sigma}_D}{\sqrt{M}}\right)^{-1}\left(\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1}) - \tau)}{\hat{\sigma}_{\hat{\eta}_1}} - \frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_2}) - \tau)}{\hat{\sigma}_{\hat{\eta}_2}}\right) > -\Phi^{-1}(\beta) \,\middle|\, D\right) \\
&= 1 - \Phi\left(-\Phi^{-1}(\beta)\right) + o_P(1) \\
&= \beta + o_P(1).
\end{aligned}
$$

For $(p_j, \hat{\delta}(\beta)) = (p_j^{\pm}, \hat{\delta}^{\pm}(\beta))$,

$$
\begin{aligned}
&P\left(p_2^{\pm} > p_1^{\pm} + \hat{\delta}^{\pm}(\beta) \,\middle|\, D\right) \\
&= P\left(2\Phi\left(-\left|\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_2}) - \tau)}{\hat{\sigma}_{\hat{\eta}_2}}\right|\right) > 2\Phi\left(-\left|\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1}) - \tau)}{\hat{\sigma}_{\hat{\eta}_1}}\right|\right) + \hat{\delta}^{\pm}(\beta) \,\middle|\, D\right) \\
&= P\left(2\Phi\left(-\left|\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_2}) - \tau)}{\hat{\sigma}_{\hat{\eta}_2}}\right|\right) > 2\Phi\left(-\left|\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1}) - \tau)}{\hat{\sigma}_{\hat{\eta}_1}}\right| - \frac{\sqrt{n}\hat{\sigma}_D}{\sqrt{M}}\Phi^{-1}(\beta/2)\right) \,\middle|\, D\right) \\
&= P\left(\left|\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1}) - \tau)}{\hat{\sigma}_{\hat{\eta}_1}}\right| - \left|\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_2}) - \tau)}{\hat{\sigma}_{\hat{\eta}_2}}\right| > -\frac{\sqrt{n}\hat{\sigma}_D}{\sqrt{M}}\Phi^{-1}(\beta/2) \,\middle|\, D\right) \\
&\leqslant P\left(\left|\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_2}) - \tau)}{\hat{\sigma}_{\hat{\eta}_2}} - \frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1}) - \tau)}{\hat{\sigma}_{\hat{\eta}_1}}\right| > -\frac{\sqrt{n}\hat{\sigma}_D}{\sqrt{M}}\Phi^{-1}(\beta/2) \,\middle|\, D\right) \\
&= 2P\left(\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_2}) - \tau)}{\hat{\sigma}_{\hat{\eta}_2}} - \frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1}) - \tau)}{\hat{\sigma}_{\hat{\eta}_1}} < \frac{\sqrt{n}\hat{\sigma}_D}{\sqrt{M}}\Phi^{-1}(\beta/2) \,\middle|\, D\right) \\
&= 2P\left(\left(\frac{\sqrt{n}\hat{\sigma}_D}{\sqrt{M}}\right)^{-1}\left(\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_2}) - \tau)}{\hat{\sigma}_{\hat{\eta}_2}} - \frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1}) - \tau)}{\hat{\sigma}_{\hat{\eta}_1}}\right) < \Phi^{-1}(\beta/2) \,\middle|\, D\right) \\
&= 2\Phi\left(\Phi^{-1}(\beta/2)\right) + o_P(1) \\
&= \beta + o_P(1).
\end{aligned}
$$

■

*Proof of Theorem 5.3.* The first result follows since, from the proof of Theorem 5.1,

$$\left(\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1}) - \tau)}{\hat{\sigma}_{\hat{\eta}_1}} - \frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_2}) - \tau)}{\hat{\sigma}_{\hat{\eta}_2}}\right) = O_P(1).$$

For $(p_j, \hat{\delta}(\beta)) = (p_j^{+}, \hat{\delta}^{+}(\beta))$, from the proof of Theorem 5.2,

$$
\begin{aligned}
&P\left(p_2^{+} > p_1^{+} + \hat{\delta}^{+}(\beta) \,\middle|\, D\right) \\
&\quad = P\left(\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1}) - \tau)}{\hat{\sigma}_{\hat{\eta}_1}} - \frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_2}) - \tau)}{\hat{\sigma}_{\hat{\eta}_2}} < \left(\frac{\sqrt{n}\hat{\sigma}_D}{\sqrt{M}}\right)\Phi^{-1}(\beta) \,\middle|\, D\right),
\end{aligned}
$$

which converges to zero since

$$\frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_1}) - \tau)}{\hat{\sigma}_{\hat{\eta}_1}} - \frac{\sqrt{n}(h(\hat{\theta}_{\hat{\eta}_2}) - \tau)}{\hat{\sigma}_{\hat{\eta}_2}} = O_P(1)$$

from Theorem 5.1, and

$$\left(\frac{\sqrt{n}\hat{\sigma}_D}{\sqrt{M}}\right)\Phi^{-1}(\beta) \xrightarrow{P} -\infty$$

since $\Phi^{-1}(\beta) < 0$. Analogous results follow for $(p_j^-, \hat{\delta}^-(\beta))$ and $(p_j^\pm, \hat{\delta}^\pm(\beta))$. ■

## B.4 Details of Section 6

### B.4.1 Covariates Description

The following variables from the Ghana Socioeconomic Panel Survey are used as predictive covariates for poverty prediction in Section 6:

**Household Demographics**

- `children`: Number of children in household
- `adults`: Number of adults in household
- `female_head`: Indicator for female household head
- `married_head`: Indicator for married household head
- `spouse_in`: Indicator for spouse living in the household

**Religion**

- `christian`: Proportion Christian
- `muslim`: Proportion Muslim
- `traditional`: Proportion traditional religion

**Political and Traditional Leadership**

- `ever_political_office`: Indicator for ever holding political office
- `today_political_office`: Indicator for currently holding political office
- `ever_traditional_office`: Indicator for ever holding traditional office
- `today_traditional_office`: Indicator for currently holding traditional office

**Parental Education**

- `father_primary`: Indicator for father completed primary education
- `father_middle`: Indicator for father completed middle school
- `father_secondary`: Indicator for father completed secondary education
- `father_tertiary`: Indicator for father completed tertiary education
- `mother_primary`: Indicator for mother completed primary education
- `mother_middle`: Indicator for mother completed middle school
- `mother_secondary`: Indicator for mother completed secondary education
- `mother_tertiary`: Indicator for mother completed tertiary education

**Asset Holdings**

- `plot_acreage`: Total land holdings in acres
- `livestock_value`: Total value of livestock
- `livestock_expenses`: Annual livestock maintenance expenses

**Financial Resources**

- `health_insurance`: Proportion of household members covered by health insurance
- `savings_home`: Amount of savings kept at home
- `d_saving_bank`: Distance to nearest bank (in km)
- `savings_bank`: Amount of savings in bank account

### B.4.2 Fraction Per Tercile as a Z-Estimator

For a given split $\mathsf{s}$, the vector

$$\left(\left(\frac{\sum_{i\in\mathsf{s}} Y_i \mathbb{I}\left(\hat{t}_{j-1,\tilde{\mathsf{s}}} < \hat{\eta}_{\tilde{\mathsf{s}}}(X_i) \leqslant \hat{t}_{j,\tilde{\mathsf{s}}}\right)}{\sum_{i\in\mathsf{s}} \mathbb{I}\left(\hat{t}_{j-1,\tilde{\mathsf{s}}} < \hat{\eta}_{\tilde{\mathsf{s}}}(X_i) \leqslant \hat{t}_{j,\tilde{\mathsf{s}}}\right)}\right)_{j=1}^{3}, (\hat{t}_{j,\tilde{\mathsf{s}}})_{j=1}^{2}\right)^{T}$$

is a Z-estimator with the moment functions

$$\psi_{(\theta,t),\eta}(y,x) = \begin{pmatrix} y\mathbb{I}\left(t_0 < \eta(x) \leqslant t_1\right) - \theta_1 \mathbb{I}\left(t_0 < \eta(x) \leqslant t_1\right) \\ y\mathbb{I}\left(t_1 < \eta(x) \leqslant t_2\right) - \theta_2 \mathbb{I}\left(t_1 < \eta(x) \leqslant t_2\right) \\ y\mathbb{I}\left(t_2 < \eta(x) \leqslant t_3\right) - \theta_3 \mathbb{I}\left(t_2 < \eta(x) \leqslant t_3\right) \\ \mathbb{I}\left(\eta(x) \leqslant t_1\right) - \frac{1}{3} \\ \mathbb{I}\left(\eta(x) \leqslant t_2\right) - \frac{2}{3} \end{pmatrix}.$$

Hence, the final estimators $\hat{\theta}_{\hat{\eta},\mathrm{Frac}j}$ are averages over split-specific estimators as in (3.2).

Note that the conditions in Theorem 3.1 are met whenever $\eta_P^*(x)$ is not flat in $x$. This condition is testable, for example using the one-sided test for the accuracy in Figure 1.

### B.4.3 Monte Carlo Designs

I simulate outcome and covariates by (i) converting each observed column to rank-based uniforms $U = \mathrm{rank}(X)/(n+1)$, (ii) Gaussianizing to $Z = \Phi^{-1}(U)$ and estimating the latent normal correlation $\Sigma^*$, (iii) drawing $Z^* \sim \mathcal{N}(0, \Sigma^*)$ and mapping back to uniforms $U^* = \Phi(Z^*)$, and (iv) inverting each margin with the empirical CDF of the corresponding variable. For the correlated design, I modify $\Sigma^*$ by multiplying by 3 the first row/column, the one corresponding to the correlation between outcome and covariates, and use as correlation matrix its nearest positive definite matrix in case the modified $\Sigma^*$ is no longer positive definite. For the uncorrelated design, I sample covariates the same way, and the outcome is sampled independently from a binomial distribution with probability 0.07.

### B.4.4 Comparison of Top-Bottom Estimates

Figure 5 compares the top minus bottom estimates across datasets and methods, similar to Figure 1.

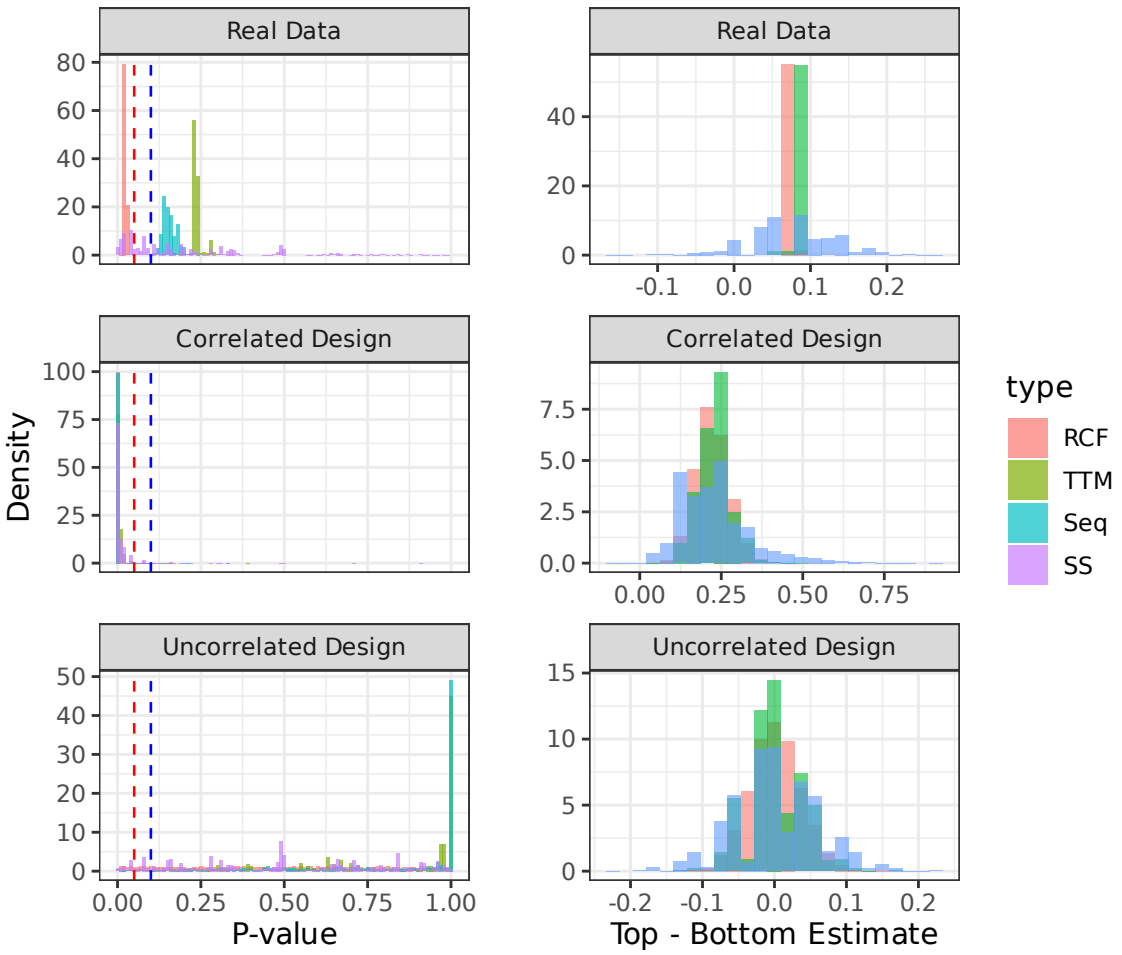


Figure 5: Comparison of Top-Bottom Estimates Across Methods and Datasets

**Notes:** Left panels show distribution across Monte Carlo iterations of p-values for testing whether the top tercile has a higher fraction below the poverty line than the bottom tercile. Vertical red and blue lines are respectively 0.05 and 0.10. Right panels show distribution of point estimates for the difference between top and bottom terciles. Rows show results for real data (top), simulations from correlated design (middle), and simulations from uncorrelated design (bottom). Methods: RCF (repeated cross-fitting), TTM (twice-the-median), Seq (sequential aggregation), SS (sample-splitting).

## B.5 Details of Section 7

### B.5.1 Covariates Description

**Donation History Variables:**

- `hpa`: Highest previous contribution
- `freq`: Number of prior donations
- `years`: Number of years since initial donation
- `mrm2`: Number of months since last donation

**Individual Demographics:**

- `female`: Female indicator

**State-Level Political Variables:**

- `cases`: Count of court cases between 2002 and 2005 in which the organization was either a party to or filed a brief
- `perbush`: State vote share for Bush
- `nonlit`: Count of incidences relevant to this organization from each state reported in 2004-5 (values range from zero to six) in the organization's monthly newsletter to donors

**Zip Code Demographics and Economics:**

- `pwhite`: Proportion white within zip code
- `pblack`: Proportion black within zip code
- `page18_39`: Proportion age 18-39 within zip code
- `ave_hh_sz`: Average household size within zip code
- `median_hhincome`: Median household income within zip code
- `powner`: Proportion house owner within zip code
- `psch_atlstba`: Proportion who finished college within zip code
- `pop_propurban`: Proportion of population urban within zip code

### B.5.2 Monte Carlo Designs

The designs are explicitly calibrated to the observed data so that simulated covariates and outcomes are distributionally aligned with the original sample.

*Treatment assignment.* I draw the treatment assignment indicator from a Bernoulli distribution with mean 0.5.

*Covariates and potential outcome under control.* Starting from the observed outcome and covariate matrix for the control sample, I form pseudo-uniforms for each column by ranking within sample and scaling, $U = \text{rank}(X)/(n+1)$. I then Gaussianize to $Z = \Phi^{-1}(U)$ and estimate the latent normal correlation $\Sigma^*$ on these $Z$ (taking the nearest positive definite matrix if needed). To generate synthetic $Y(0)$ and covariates, I draw $Z^* \sim \mathcal{N}(0, \Sigma^*)$, map to uniforms $U^* = \Phi(Z^*)$, and invert each margin via the empirical CDF of the corresponding original variable.

*Treatment effect.* From the original data, I estimate two arm-specific components as functions of treatment and covariates. The first is a logistic regression for whether $Y = 0$ (no donation), using treatment, covariates and their interactions. The second is a Poisson regression, with amount of donation as outcome and same variables in the model. For generating simulated observations, the treatment effect is zero with probability $q_0(x, y_0) - q_1(x, y_0)$ (rounded to zero or one if necessary), where

$$q_d(x, y_0) = (1 - \pi_d(x))\ \hat{P}(Y \geqslant y_0{+}1 \mid X{=}x, D{=}d),$$

with $x$ being the covariates, $y_0$ the value of potential outcome under control, $\pi_d(x)$ the probability that $Y = 0$ coming from the logit model with coefficients associated with treatment = 1 being multiplied by 4, and $\hat{P}(Y \geqslant y_0{+}1 \mid X{=}x, D{=}d)$ coming from the Poisson model with mean multiplied by 0.05. Conditional on the treatment effect being different from zero, I draw $Y(1)$ from a truncated Poisson distribution starting at $Y(0)$ with the same mean coming from the Poisson regression.

*Final outcome.* For the design where treatment effect heterogeneity is predictable, I generate the observed outcome as $Y(1)$ if treatment is 1, and $Y(0)$ otherwise. For the design where treatment effect heterogeneity is not predictable, I generate the entire dataset exactly the same way, but shuffle the treatment assignment indicator at random as the last step.

### B.5.3 Additional Figures and Table

Figure 6 displays results with the real dataset with shuffled treatment indicator (at random, so treatment effect is constant and equal to zero), and Figure 7 displays results for the synthetic DGP where there is explainable treatment effect heterogeneity. Table 4 gives the number of Monte Carlo iterations used for each specification.

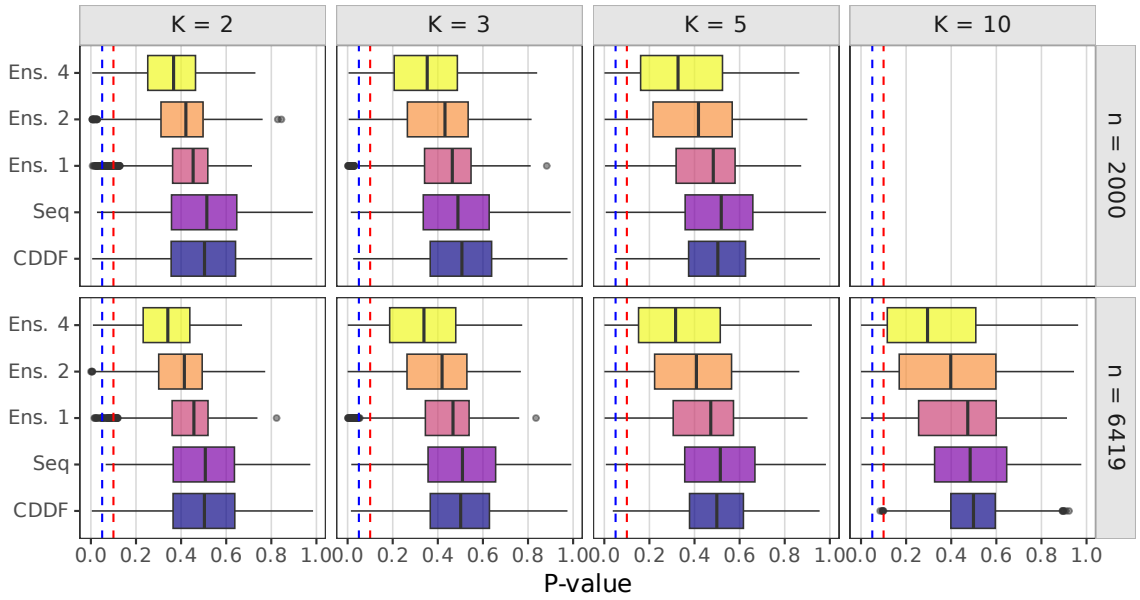


Figure 6: Distribution of p-values for Top - Bottom GATES Groups – Real Data with Shuffled Treatment Assignment

Notes: Distribution of one-sided p-values for testing whether the top tercile has a larger ATE than the bottom tercile across Monte Carlo iterations using the real dataset. Rows show different sample sizes ($n = 2000, 6419$), columns show different numbers of folds ($K = 2, 3, 5, 10$). Ens. 1, Ens. 2, and Ens. 4 represent the Ensemble method using respectively 1, 2, and 4 algorithms. Each box represents the distribution across Monte Carlo iterations with 100 repetitions of sample-splitting per iteration. Boxplots show the median (center line), interquartile range (box), and whiskers extending to 1.5 times the IQR, with points beyond shown as outliers. Sources of randomness are the subsample when $n = 2000$, which ML algorithms are used, how the data are split, and how the treatment assignment indicator is shuffled. Red dashed line at 0.1, blue dashed line at 0.05. Specifications with $K = 10, n = 2000$ are excluded.

### B.5.4 Theoretical Properties of Ensemble Approach

I establish the theoretical properties of the ensemble estimator using the CLTs proven in this paper. I show that when there is detectable heterogeneity, i.e., when the ensemble weights $(\hat{\beta}_a)_{a=1}^A$ do not converge to zero, the confidence interval based on the normal approximation is asymptotically exact. If there is no detectable heterogeneity, however, my theoretical result gives no coverage guarantee to the normal approximation CI. Extensive simulation exercises, including but not limited to those of Section 7, suggest that the normal approximation CI is actually conservative under the null hypothesis of no heterogeneity for small values of $A$ and $K$ such as $A = 4$

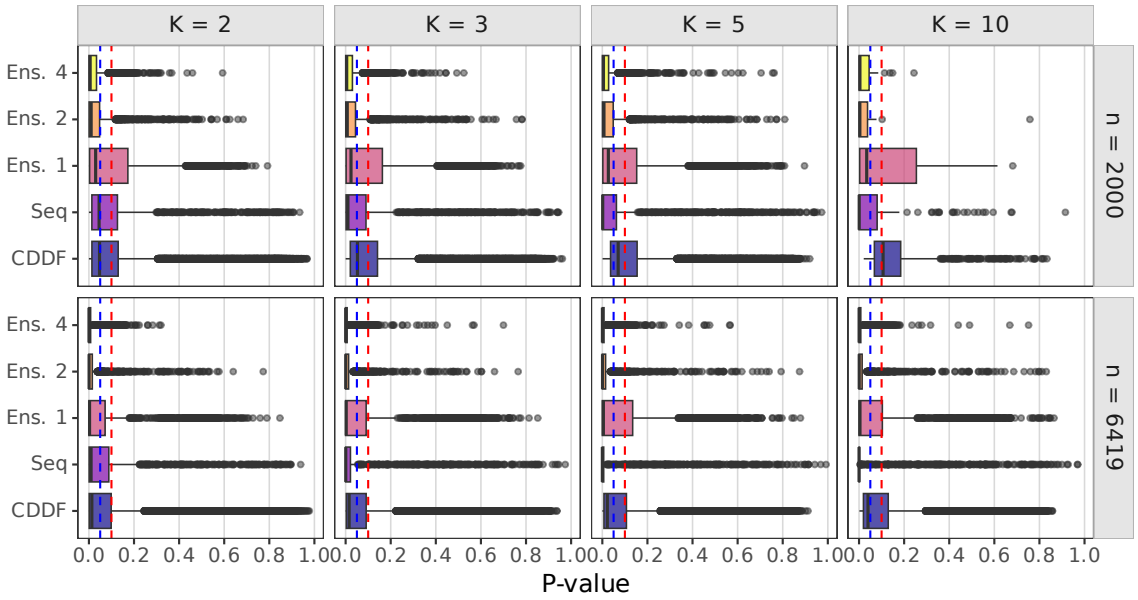


Figure 7: Distribution of p-values for Top - Bottom GATES Groups – Synthetic DGP with Heterogeneity

Notes: Distribution of one-sided p-values for testing whether the top tercile has a larger ATE than the bottom tercile across Monte Carlo iterations using the real dataset. Rows show different sample sizes ($n = 2000, 6419$), columns show different numbers of folds ($K = 2, 3, 5, 10$). Ens. 1, Ens. 2, and Ens. 4 represent the Ensemble method using respectively 1, 2, and 4 algorithms. Each box represents the distribution across Monte Carlo iterations with 100 repetitions of sample-splitting per iteration. Boxplots show the median (center line), interquartile range (box), and whiskers extending to 1.5 times the IQR, with points beyond shown as outliers. Data is generated from a synthetic DGP where there is explainable treatment effect heterogeneity (Appendix B.5). Red dashed line at 0.1, blue dashed line at 0.05. Specifications with $K = 10, n = 2000$ are excluded.

and $K = 3$. Hence, my recommendation for empirical practice is to use the normal approximation CI with no more than 4 algorithms and 5 folds. I also propose an adaptive approach using ideas developed in Section 4.2 that is valid even when there is no detectable heterogeneity, at the cost of having smaller power.

First, I introduce additional notation. Denote the set of splits

$$\mathcal{S} = (\mathsf{s}_{m,k})_{m\in[M],k\in[K]}\,,$$

and the set of model $\hat{\eta} = (\hat{\eta}_{\mathsf{s}})_{\mathsf{s}\in\mathcal{S}}$. I use $F_P(x)$ to denote the cdf of the random variable $\sum_{a=1}^{A} \beta^*_{Pa}\eta^*_{Pa}(X)$ and

$$F_P^{-1}(p) = \inf\left\{x \in \mathcal{X} : p \leqslant F_P(x)\right\}.$$

Table 4: Number of Monte Carlo Iterations by Specification

| | data | n500 | | n1000 | | n2000 | | | n6419 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Data Type | K2 | K3 | K2 | K3 | K2 | K3 | K5 | K2 | K3 | K5 | K10 |
| CDDF | Real (Shuffled) | 26,175 | 6,990 | 27,585 | 20,976 | 16,850 | 17,050 | 17,125 | 13,784 | 14,029 | 13,393 | 11,632 |
| CDDF | MC: No HTE | 26,109 | 26,320 | 26,637 | 26,211 | 16,433 | 15,805 | 18,756 | 12,384 | 12,142 | 12,376 | 11,676 |
| CDDF | Real Data | 23,324 | 4,327 | 27,026 | 12,845 | 17,204 | 17,131 | 8,579 | 13,765 | 14,005 | 13,420 | 9,417 |
| CDDF | MC: With HTE | 21,283 | 18,018 | 27,783 | 25,352 | 17,303 | 17,120 | 18,073 | 13,756 | 13,545 | 13,900 | 11,794 |
| Seq | Real (Shuffled) | 644 | 192 | 670 | 502 | 1,316 | 1,164 | 1,040 | 1,054 | 1,092 | 1,338 | 446 |
| Seq | MC: No HTE | 668 | 700 | 760 | 570 | 1,182 | 1,246 | 1,454 | 1,230 | 1,150 | 1,296 | 1,214 |
| Seq | Real Data | 1,154 | 172 | 1,342 | 650 | 1,134 | 1,380 | 568 | 1,018 | 1,004 | 1,430 | 656 |
| Seq | MC: With HTE | 508 | 464 | 696 | 622 | 1,366 | 1,368 | 1,248 | 1,302 | 1,262 | 1,320 | 1,226 |
| Ens. 1 | Real (Shuffled) | 3,177 | 901 | 3,390 | 2,638 | 1,969 | 2,084 | 1,990 | 1,493 | 1,543 | 1,580 | 1,505 |
| Ens. 1 | MC: No HTE | 3,399 | 3,368 | 3,327 | 3,335 | 2,105 | 2,057 | 2,346 | 1,492 | 1,537 | 1,560 | 1,522 |
| Ens. 1 | Real Data | 2,871 | 491 | 3,372 | 1,572 | 2,038 | 2,014 | 1,032 | 1,493 | 1,614 | 1,530 | 1,268 |
| Ens. 1 | MC: With HTE | 2,744 | 2,229 | 3,206 | 3,132 | 2,004 | 2,040 | 2,091 | 1,549 | 1,569 | 1,563 | 1,549 |
| Ens. 2 | Real (Shuffled) | 3,183 | 841 | 3,433 | 2,664 | 2,096 | 2,078 | 1,974 | 1,543 | 1,552 | 1,571 | 1,485 |
| Ens. 2 | MC: No HTE | 3,370 | 3,409 | 3,417 | 3,340 | 2,124 | 2,003 | 2,374 | 1,582 | 1,538 | 1,544 | 1,455 |
| Ens. 2 | Real Data | 2,865 | 499 | 3,269 | 1,584 | 2,160 | 1,999 | 987 | 1,574 | 1,556 | 1,561 | 1,265 |
| Ens. 2 | MC: With HTE | 2,625 | 2,226 | 3,429 | 3,170 | 2,120 | 2,052 | 2,075 | 1,589 | 1,585 | 1,588 | 1,443 |
| Ens. 4 | Real (Shuffled) | 3,261 | 868 | 3,476 | 2,512 | 2,048 | 2,072 | 2,035 | 1,581 | 1,643 | 1,511 | 1,389 |
| Ens. 4 | MC: No HTE | 3,367 | 3,405 | 3,421 | 3,410 | 2,069 | 1,996 | 2,319 | 1,524 | 1,438 | 1,525 | 1,420 |
| Ens. 4 | Real Data | 2,876 | 546 | 3,375 | 1,614 | 2,052 | 2,089 | 991 | 1,567 | 1,575 | 1,506 | 1,114 |
| Ens. 4 | MC: With HTE | 2,569 | 2,196 | 3,451 | 3,081 | 2,050 | 2,073 | 2,116 | 1,591 | 1,547 | 1,614 | 1,484 |

For some results, I focus on a set $\mathcal{P}_{hte} \subseteq \mathcal{P}$ such that $\left(F_P^{-1}(t)\right)_{P\in\mathcal{P}_{hte}}$ is equicontinuous at points $t = j/J$ for $j = 1, \ldots, J$. This is a collection of DGPs where the $J$ quantiles of the limit predicted ITE $\sum_{a=1}^{A} \beta_{Pa}^{*}\eta_{Pa}^{*}(X)$ are well-defined. This is required so that the groups defined in (7.6) are well-defined in the limit. Note that $F_P^{-1}(j/J)$ being continuous implies that the limit predictor $\sum_{a=1}^{A} \beta_{Pa}^{*}\eta_{Pa}^{*}(X)$ is not flat in $X$, so this class essentially excludes DGPs where there is no detectable heterogeneity, that is, where the true CATE $\eta_P(x)$ is flat in $x$.

My first result is that the normal approximation CI is asymptotically exact when there is detectable heterogeneity. It relies on Assumption B.5, defined in Appendix B.5.6. It is a mild but technical assumption that requires: (i) the weights $\hat{\beta}_{\ell,a}$ have finite limits, (ii) a standard moments condition, (iii) propensity scores are bounded away from 0 and 1, (iv) the variance-covariance matrix of the regressors $Z$ is positive definite, and (v) the models estimated with ML converge to any limit at any rate.

**Theorem B.3.** *Let Assumption B.5 hold, and let $\mathcal{P}_{hte} \subseteq \mathcal{P}$ be such that $\left(F_P^{-1}(t)\right)_{P\in\mathcal{P}_{hte}}$ is equicontinuous at points $t = j/J$ for $j = 1, \ldots, J$. Then, for any sequence $(P_n)_{n\geqslant 1} \subseteq \mathcal{P}_{hte}$,*

$$P_n\left(\delta_{\hat{\eta}} \in \left[\hat{\delta}_{\hat{\eta}} - z_{1-\alpha/2}\hat{\sigma}_{\hat{\eta}}, \hat{\delta}_{\hat{\eta}} + z_{1-\alpha/2}\hat{\sigma}_{\hat{\eta}}\right]\right) \to 1 - \alpha.$$

□

Although Theorem B.3 does not cover cases when there is no detectable heterogeneity, extensive simulation exercises, including but not limited to the ones of Section 7, suggest that the coverage probability is larger than $1-\alpha$ in those cases at least when $A \leqslant 4$, $K \leqslant 5$, that is, the CI Theorem B.3 is conservative. Next, I consider a test for detectable heterogeneity that can be used, for example, when $A > 4$ and/or $K > 5$. If the test rejects no detectable heterogeneity, the normal approximation CI may be used.

### B.5.5 A Test for Detectable Heterogeneity

I propose using a version of the test proposed in Section 4.2.1 for testing whether the models $\hat{\eta} = (\hat{\eta}_{\tilde{\mathsf{s}}_{m,k}})$ have explanatory power for heterogeneous treatment effects. Specifically, I first calculate the mean squared of residuals from the BLP regression

$$Y_i = \alpha_1 + \sum_{a=1}^{A} \beta_a (\hat{\eta}_{\tilde{\mathsf{s}},a}(X_i) - \bar{\tau}_{\mathsf{s},a}) \left[T_i - p(X_i)\right] + \alpha_2 Z_i + \varepsilon_i, \qquad i \in \mathsf{s} \tag{B.25}$$

with weights $\omega_i = \{p(X_i)\left[1-p(X_i)\right]\}^{-1}$, $\bar{\tau}_{\mathsf{s},a} = |s|^{-1} \sum_{i \in \mathsf{s}} \hat{\eta}_{\tilde{\mathsf{s}},a}(X_i)$, for $\mathsf{s} \in \mathcal{S}$, as in (7.5) but at the fold level. Denote it by

$$MSR_{\mathsf{s}} = \frac{1}{|s|} \sum_{i \in \mathsf{s}} \left( Y_i - \hat{\alpha}_{1,\mathsf{s}} + \sum_{a=1}^{A} \hat{\beta}_{a,\mathsf{s}} (\hat{\eta}_{\tilde{\mathsf{s}},a}(X_i) - \bar{\tau}_{\mathsf{s},a}) \left[T_i - p(X_i)\right] + \hat{\alpha}_{2,\mathsf{s}} Z_i \right)^2$$

I compare $(MSR_{\mathsf{s}})_{\mathsf{s} \in \mathcal{S}}$ with

$$MSR_b = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{\alpha}_{1,b} + \hat{\alpha}_{2,b} Z_i\right)^2,$$

where $\hat{\alpha}_{1,b}$ and $\hat{\alpha}_{2,b}$ are the estimates from the weighted least squares regression

$$Y_i = \alpha_1 + \alpha_2 Z_i + \varepsilon_i, \qquad i \in 1, \ldots, n.$$

Let $\hat{\Sigma}$ be an estimate of the asymptotic variance of $\sqrt{n}\left(MSR_{\mathsf{s}} - MSR_b\right)_{\mathsf{s} \in \mathcal{S}}$, and $\hat{\sigma}^2_{\mathsf{s}}$ are the entries of the main diagonal. I propose calculating the test-statistic

$$\hat{T} = \sum_{\mathsf{s} \in \mathcal{S}} \left( \min \left\{ \sqrt{n} \frac{MSR_{\mathsf{s}} - MSR_b}{\hat{\sigma}_{\mathsf{s}}}, 0 \right\} \right)^2.$$

I establish the validity of this test in Theorem B.4, where $\hat{c}_{1-\alpha}$ is calculated as in Section 4.2.1. The result follows from Theorem 4.2.

**Theorem B.4.** *Let Assumption B.5 hold, and let*

$$\mathcal{P}_0 = \{P \in \mathcal{P} : \exists c \in \mathbb{R}, \eta_P(x) = c\}.$$

*Then, for any sequence* $(P_n)_{n\geqslant 1} \subseteq \mathcal{P}_0$,

$$P_n\left(\hat{T} > \hat{c}_{1-\alpha}\right) \to 1-\alpha.$$

□

Denote the normal approximation CI

$$\widehat{\mathrm{CI}}_{\alpha,\mathcal{N}} = \left[\hat{\delta}_{\hat{\eta}} - z_{1-\alpha/2}\hat{\sigma}_{\hat{\eta}}, \hat{\delta}_{\hat{\eta}} + z_{1-\alpha/2}\hat{\sigma}_{\hat{\eta}}\right],$$

and the extended CI

$$\widehat{\mathrm{CI}}_{\alpha,\mathrm{ext}} = \mathrm{Conv}\left(\widehat{\mathrm{CI}}_{\alpha,\mathcal{N}} \cup \{0\}\right),$$

where Conv denotes the convex hull, that is, $\widehat{\mathrm{CI}}_{\alpha,\mathrm{ext}}$ has all the elements in $\widehat{\mathrm{CI}}_{\alpha,\mathcal{N}}$, 0, and all elements in between. For a given fixed $\bar{c}_5 \geqslant 0$, denote the final CI

$$\widehat{\mathrm{CI}}_{\alpha} = \begin{cases} \widehat{\mathrm{CI}}_{\alpha,\mathcal{N}}, \text{ if } \hat{T}' > \hat{c}_{1-\alpha} \\ \widehat{\mathrm{CI}}_{\alpha,\mathrm{ext}}, \text{ otherwise,} \end{cases}$$

Theorem B.4 implies that this CI is asymptotically valid pointwise in $P \in \mathcal{P}$ for $\bar{c}_5 = 0$, and uniformly in $P \in \mathcal{P}$ for any $\bar{c}_5 > 0$.

### B.5.6 Proofs and Extra Definitions

Define $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ as the space that contains the covariates $X \in \mathcal{X}$ for some integer $d_x > 0$. Let $Y^T = (Y_i)_{i=1}^n$. For any

$$d = (d_{j,k})_{j\in[J],k\in[K]},$$

$$\beta = (\beta_{\ell,a})_{\ell\in[L],a\in[A]},$$

and $\eta \in H$, let

$$H^T_{\eta,\beta,d} = \left(Z_i, \left[\{T_i - p(X_i)\}\mathbb{I}\left(d_{j-1,k(i)} \leqslant \sum_{a=1}^{A} \beta_{\ell(i),a}\eta_a(X_i) < d_{j,k(i)}\right)\right]_{j=1}^{J}\right)_{i=1}^{n}.$$

$$H^T_{\hat{\eta},\beta,d} = \left(Z_i, \left[\{T_i - p(X_i)\}\mathbb{I}\left(d_{j-1,k(i)} \leqslant \sum_{a=1}^{A} \beta_{\ell(i),a}\hat{\eta}_{\hat{\mathsf{s}}_{k(i)},a}(X_i) < d_{j,k(i)}\right)\right]_{j=1}^{J}\right)_{i=1}^{n}.$$

$\Omega$ is the n-by-n diagonal matrix of weights:

$$\Omega = \text{diag}(\omega_1, \ldots, \omega_n).$$

$\beta^*_{P\ell,a}$ is the coefficient of the linear projection with weights $\omega$ of $Y$ on

$$\left[\{T_i - p(X_i)\}\left(\eta^*_{Pa}(X) - \mathbb{E}_P\left[\eta^*_{Pa}(X)\right]\right)\right]_{j=1}^{J}$$

when that is well-defined, and zero otherwise. Note $\beta^*_{P\ell,a}$ is the same for all $\ell$ since the limit $\eta^*_{Pa}$ does not depend on the data. Let $F_P(x)$ be the cdf of the random variable $\sum_{a=1}^{A} \beta^*_{Pa}\eta^*_{Pa}(X)$ and

$$F_P^{-1}(p) = \inf\left\{x \in \mathcal{X} : p \leqslant F_P(x)\right\}.$$

Define

$$d^*_{Pj,k} = F_P^{-1}(j/J).$$

Similarly, $d^*_{Pj,k}$ is the same for all $k$.

$$\hat{d} = (\hat{d}_{j,k})_{j\in[J],k\in[K]},$$

$$\hat{\beta} = (\hat{\beta}_{\ell,a})_{\ell\in[L],a\in[A]}.$$

Define $\theta_{\eta,\beta,d}$ and column vector $\varepsilon_{\eta,\beta,d}$ such that

$$Y = H_{\eta,\beta,d}\theta_{\eta,\beta,d} + \varepsilon_{\eta,\beta,d}. \tag{B.26}$$

$\hat{\theta}^{(m)}_{\hat{\eta},\hat{\beta},\hat{d}} = (\hat{\alpha}, (\hat{\gamma}^J_{j=1}))^T$ are the estimates from (7.7), and $\theta^{(m)}_{\hat{\eta},\hat{\beta},\hat{d}}$ denotes $\hat{\eta}, \hat{\beta}, \hat{d}$ from the $m$-th repetition.

**Assumption B.5.** *The following conditions hold:*

(i) *For some $B = (B_\beta \times B_d) \subset \mathbb{R}^{LA} \times \mathbb{R}^{JK}$ with compact $B_\beta$,*

$$\bigcup_{P\in\mathcal{P}} (\beta^*_P, d^*_P) \subseteq B;$$

(ii) *For some $\bar{c}_6 > 0$,*

$$\sup_{P\in\mathcal{P}} \sup_{\eta\in H,(\beta,d)\in B} \mathbb{E}_P\left[\left|H^T_{\eta,\beta,d,i}\varepsilon_{\eta,\beta,d,i}\right|^{2+\bar{c}_6}\right] < \infty;$$

(iii) *For some $\bar{c}_7 > 0$, and all $x \in \mathcal{X}$,*

$$\bar{c}_7 < p(x) < 1 - \bar{c}_7;$$

(iv) $\inf_{P\in\mathcal{P}} \det\left(\mathrm{Var}_P\left[Z\right]\right) > 0$.

(v) *There exists* $(\eta^*_{Pa})_{a=1}^A$ *such that*

$$\mathbb{E}_P\left[\left|\tilde{\eta}_a(X) - \eta^*_{Pa}(X)\right| |D\right] \xrightarrow{P} 0$$

*uniformly in* $P \in \mathcal{P}$, *where* $\tilde{\eta}_a = \mathcal{A}_a(D)$ *and* $X \perp D$.

□

**Theorem B.5.** *Let Assumption B.5 hold, and* $\mathcal{P}_{hte} \subseteq \mathcal{P}$ *be such that* $\left(F_P^{-1}(t)\right)_{P\in\mathcal{P}_{hte}}$ *is equicontinuous at points* $t = j/J$ *for* $j = 1, \ldots, J$. *Then,*

$$\sqrt{n}\left(\hat{\theta}_{\hat{\eta},\hat{\beta},\hat{d}} - \theta_{\hat{\eta},\hat{\beta},\hat{d}}\right) - \sqrt{n}\,\mathbb{E}_P\left[H^T_{\eta^*_P,\beta^*_P,d^*_P}\Omega H_{\eta^*_P,\beta^*_P,d^*_P}\right]^{-1} H^T_{\eta^*_P,\beta^*_P,d^*_P}\Omega\varepsilon_{\eta^*_P,\beta^*_P,d^*_P} \xrightarrow{P} 0$$

*uniformly in* $P \in \mathcal{P}_{hte}$. □

*Proof of Theorem B.5.* First, note that equicontinuity of $\left(F_P^{-1}(t)\right)_{P\in\mathcal{P}_{hte}}$ implies the $J$ quantiles groups to be well-defined, which together with Assumption B.5 implies

$$\inf_{P\in\mathcal{P}_{hte}} \det\left(H^T_{\eta^*_P,\beta^*_P,d^*_P}\Omega H_{\eta^*_P,\beta^*_P,d^*_P}\right) > 0.$$

For each $m = 1, \ldots, M$, using (B.26) leads to the decomposition

$$\hat{\theta}^{(m)}_{\hat{\eta},\hat{\beta},\hat{d}} - \theta^{(m)}_{\hat{\eta},\hat{\beta},\hat{d}} = \left(H^T_{\hat{\eta},\hat{\beta},\hat{d}}\Omega H_{\hat{\eta},\hat{\beta},\hat{d}}\right)^{-1} H^T_{\hat{\eta},\hat{\beta},\hat{d}}\Omega\varepsilon_{\hat{\eta},\hat{\beta},\hat{d}}.$$

$$\left(n^{-1}H^T_{\hat{\eta},\hat{\beta},\hat{d}}\Omega H_{\hat{\eta},\hat{\beta},\hat{d}}\right)^{-1} \xrightarrow{P} \mathbb{E}_P\left[H^T_{\eta^*_P,\beta^*_P,d^*_P}\Omega H_{\eta^*_P,\beta^*_P,d^*_P}\right]^{-1}$$

by a uniform law of large numbers. The terms in $n^{-1/2}H^T_{\hat{\eta},\hat{\beta},\hat{d}}\Omega\varepsilon_{\hat{\eta},\hat{\beta},\hat{d}}$ are given by

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \omega_i\varepsilon_{\hat{\eta},\hat{\beta},\hat{d},i}Z_i, \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n \omega_i\varepsilon_{\hat{\eta},\hat{\beta},\hat{d},i}\left[T_i - p(X_i)\right]\mathbb{I}\left(\hat{d}_{j-1,k(i)} \leqslant \sum_{a=1}^A \beta_{\ell(i),a}\hat{\eta}_{\hat{\mathbf{s}}_{k(i)},a}(X_i) < \hat{d}_{j,k(i)}\right)\right\}_{j=1}^J\right).$$

These are split-sample empirical processes as in Theorem E.1, with functions

$$f_{(\beta,d),\eta,1}(y,h,\omega) = \omega(y - h^T\theta_{\eta,\beta,d})z$$

and

$$f_{(\beta,d),\eta,1+j}(y,h,k,\ell,\omega) = \omega(y - h^T\theta_{\eta,\beta,d})(t - p(x))\mathbb{I}\left(d_{j-1,k} \leqslant \sum_{a=1}^A \beta_{\ell,a}\hat{\eta}_{\hat{\mathbf{s}}_k,a}(X_i) < \hat{d}_{j,k}\right).$$

Step one of the proof of Theorem E.1 gives

$$\sup_{(\beta,d)\in B} \left\| n^{-1/2} H^T_{\hat{\eta},\beta,d} \Omega \varepsilon_{\hat{\eta},\beta,d} - n^{-1/2} H^T_{\eta^*_P,\beta,d} \Omega \varepsilon_{\eta^*_P,\beta,d} \right\| \xrightarrow{P} 0.$$

Together with consistency of $(\hat{\beta}, \hat{d})$ to $(\beta^*_P, d^*_P)$, which follows from a uniform law of large numbers, this gives

$$n^{-1/2} H^T_{\hat{\eta},\hat{\beta},\hat{d}} \Omega \varepsilon_{\hat{\eta},\hat{\beta},\hat{d}} = n^{-1/2} H^T_{\eta^*_P,\hat{\beta},\hat{d}} \Omega \varepsilon_{\eta^*_P,\hat{\beta},\hat{d}} + o_P(1).$$

Finally, asymptotic equicontinuity in $(\beta, d)$ gives

$$n^{-1/2} H^T_{\eta^*_P,\hat{\beta},\hat{d}} \Omega \varepsilon_{\eta^*_P,\hat{\beta},\hat{d}} = n^{-1/2} H^T_{\eta^*_P,\beta^*_P,d^*_P} \Omega \varepsilon_{\eta^*_P,\beta^*_P,d^*_P} + o_P(1).$$

Summing over $m \in M$ concludes the proof. ■

*Proof of Theorem B.3.* Follows from Theorem B.5, Lyapunov's CLT and consistency of $\hat{\sigma}_{\hat{\eta}}$, which follows by a law of large numbers. ■

*Proof of Theorem B.4.* Follows directly from Theorem 4.2, noting that

$$\mathbb{E}_P \left[ MSR_{\mathsf{s}} | \tilde{\mathsf{s}} \right] \geqslant \mathbb{E}_P \left[ MSR_b \right]$$

always holds when $\eta_{0,P}(x)$ is flat, since in that case the true coefficients $(\beta_a)$ in regression (B.25) are all zero. ■

# C Modeling Power

I formalize the notion that using a larger sample for training is desirable by the analyst by introducing the concept of modeling power. This appendix uses notation introduced in Section 2. I say that an estimator has better modeling power than another if its collection of splits has a smaller expected loss. Although my results rely on no assumptions on the training algorithm $\mathcal{A}$ other than a mild stability condition on $\mathcal{A}(D)$, in practice, $\mathcal{A}$ typically minimizes some loss function. For example, in Example 2, logistic regression minimizes log-likelihood, and neural networks minimize classification error over a class of network architectures. Let $\hat{\eta}_{\mathcal{R}} = \left( \hat{\eta}_{\tilde{\mathsf{s}}_{m,k}} \right)_{m\in[M],k\in[K]}$, $\ell_\eta(W)$ be a loss function,

$$\phi(\eta) = \int \ell_\eta(w) dP(w)$$

be the loss value of function $\eta$, and

$$\phi(\hat{\eta}_{\mathcal{R}}) = (MK)^{-1} \sum_{r\in\mathcal{R}} \sum_{\mathsf{s}\in r} \phi(\hat{\eta}_{\mathsf{s}}).$$

Note that $\phi(\hat{\eta}_{\mathcal{R}})$ is equal to the expected value of $\phi(\hat{\eta}_{\mathsf{s}})$ over $\mathsf{s} \in (\tilde{\mathsf{s}}_{m,k})_{m\in[M],k\in[K]}$ uniformly at random, which is equivalent to the loss value of using a function $\hat{\eta}$ that takes value in $\hat{\eta}_{\mathcal{R}}$ uniformly at random. The expected loss is defined as $\mathbb{E}_P[\phi(\hat{\eta}_{\mathcal{R}})]$.

The expected loss, and thus the modeling power of an estimator depends only on the sample size used to estimate the functions in $\hat{\eta}_{\mathcal{R}}$. That is because

$$\mathbb{E}_P[\phi(\hat{\eta}_{\mathcal{R}})] = (MK)^{-1}\sum_{r\in\mathcal{R}}\sum_{s\in r}\mathbb{E}_P[\phi(\hat{\eta}_{\mathsf{s}})] = \mathbb{E}_P[\phi(\hat{\eta}_{\xi})],$$

where $\xi$ is a random subset of $[n]$ of size $n-b$, with $b = n/K$ if $K > 1$, and assuming that $n$ is a multiple of $K$ for simplicity. If $\hat{\eta}_{\xi}$ is calculated with the goal of minimizing the loss $\phi(\eta)$ with respect to $\eta$, it is reasonable to assume that the expected loss $\mathbb{E}_P[\phi(\hat{\eta}_{\xi})]$ decreases with the sample size used to calculate $\hat{\eta}_{\xi}$. If that is the case, the expected loss increases with $b$, since fewer data are used to estimate each $\hat{\eta}_{\tilde{\mathsf{s}}_{m,k}}$. Hence, to increase modeling power when $K = 1$, one can pick a smaller $b$ (and $\pi$). However, if $\bar{M} < \infty$, a smaller $b$ leads to smaller statistical power, since fewer data are used as evaluation sample at each split. When using cross-fitting, modeling power increases with $K$, since $b = n/K$. In this case, the returns to increasing $K$ are diminishing. For example, if $K = 2$, $\hat{\eta}_{\tilde{\mathsf{s}}_{m,k}}$ is calculated with 50% of the sample, and this fraction raises to 90% with $K = 10$. If $K = 20$, however, the fraction only raises by another 5%. Although a large value of $K$ or small value of $\pi$ (when $K = 1$) lead to better modelling power, my asymptotic framework takes these quantities as fixed. This means that the quality of the asymptotic approximation may be poor if $K$ is large (or $\pi$ small) relative to the sample size. For example, my asymptotic framework does not accommodate for leave-one-out cross-fitting, that is, $K = n$.

# D CLT for Split-Sample Averages

I derive a CLT for split-sample estimators based on sample averages. The objective is to expose my main result in an accessible setting, and discuss the main insights of the proof. The result is generalized in Appendix E, where I derive a functional CLT uniformly over a large set of data generating processes, and in Section 3 where I prove a CLT for Z-estimators.

The notation follows Section 2. Additionally, let $f_\eta : \mathcal{W} \to \mathbb{R}$ be measurable functions for $\eta \in H$, and define

$$Pf_\eta = \int_w f_\eta(w)dP(w), \tag{D.1}$$

that is, $Pf_\eta$ is a marginal expectation that takes $\eta$ as fixed. This is typical notation in the empirical process literature.

**Example 3** (Revisited)**.** *In the probabilistic classifiers example,* $W = (Y, X)$*,* $\eta$ *is a function that predicts the probability of* $Y = 1$ *given* $X$*, and*

$$f_\eta(w) = \eta(x)\mathbb{I}(y = 1) + (1 - \eta(x))\mathbb{I}(y = 0).$$

$Pf_\eta$ *is the correct classification rate of predictor* $\eta$*.* □

In this section, I consider estimators of the form

$$\hat{\theta}_{\hat{\eta}} = \frac{1}{M}\sum_{r\in\mathcal{R}}\frac{1}{K}\sum_{\mathsf{s}\in r}\frac{1}{b}\sum_{i\in\mathsf{s}} f_{\hat{\eta}_{\mathsf{s}}}(W_i), \tag{D.2}$$

where $\mathcal{R}$ is a collection of $M$ random splits or cross-splits of the sample, $K$ is the number of folds ($K = 1$ denotes sample-splitting), $b$ is the size of each subsample $\mathsf{s}$ (either the chosen subsample size when $K = 1$ or the approximate fold size $n/K$ when $K > 1$), and $\hat{\eta} = \hat{\eta}_{\mathcal{R}} = ((\hat{\eta}_{\mathsf{s}})_{\mathsf{s}\in r})_{r\in\mathcal{R}}$. I show in Theorem D.1 that $\hat{\theta}_{\hat{\eta}}$ is $\sqrt{n}$-Gaussian when centered around its marginal expectation

$$\theta_{\hat{\eta}} = P\hat{\theta}_{\hat{\eta}} = \frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r} Pf_{\hat{\eta}_{\mathsf{s}}}.$$

In Example 3, $\theta_{\hat{\eta}}$ is the fraction of individuals correctly classified under a rule that predicts $Y = 1$ with probability

$$\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\hat{\eta}_{\mathsf{s}}(x)$$

for an individual with characteristics $X = x$.

Assumption D.1 establishes sufficient conditions for the CLT in Theorem D.1.

**Assumption D.1.** (i) *For some* $\delta > 0$*,*

$$\sup_{\eta\in H}\mathbb{E}_P\left[|f_\eta(W)|^{2+\delta}\right] < \infty.$$

(ii) *For some* $\eta^* \in H$ *and* $\tilde{\eta} = \mathcal{A}(D)$*,*

$$f_{\tilde{\eta}}(w) \xrightarrow{P} f_{\eta^*}(w)$$

*pointwise for every* $w$*.*

□

Assumption D.1(i) is a standard moments condition for CLTs, uniformly over possible values of $\eta$. Assumption D.1(ii) is a mild stability condition on $\tilde{\eta}$. Importantly, $\tilde{\eta}$ is allowed to converge at any rate and to any limit $\eta^*$. This condition is more interpretable but stronger than what I use for proving the more general CLTs in Appendix E and section 3. Assumption D.1(ii) differs from the typical approach in the double machine learning literature where faster convergence rates (often $n^{-1/4}$) are required for nuisance functions, in a context where the target parameter does not depend on the estimated model $\hat{\eta}$ (e.g., Chernozhukov et al., 2018).

**Theorem D.1.** *Let Assumption D.1 hold. Then,*

$$\sqrt{n}\left(\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right) \rightsquigarrow \mathcal{N}\left(0, V_{\bar{M},K} P\left(f_{\eta^*} - P f_{\eta^*}\right)^2\right),$$

*where*

$$V_{\bar{M},K} = \begin{cases} \bar{M}^{-1}\left(\pi^{-1} + \bar{M} - 1\right), & \text{if } K = 1 \text{ and } \bar{M} < \infty \\ 1, & \text{otherwise.} \end{cases}$$

□

Theorem D.1 can be used to construct confidence intervals with the standard error

$$\hat{\sigma}_{\hat{\eta}} = \sqrt{V_{M,K}} \frac{1}{MK} \sum_{r \in \mathcal{R}} \sum_{\mathsf{s} \in r} \hat{\sigma}_{\hat{\eta}_{\mathsf{s}}},$$

where

$$\hat{\sigma}^2_{\hat{\eta}_{\mathsf{s}}} = \frac{1}{b} \sum_{i \in \mathsf{s}} \left( f_{\hat{\eta}_{\mathsf{s}}}(W_i) - \frac{1}{b} \sum_{i \in \mathsf{s}} f_{\hat{\eta}_{\mathsf{s}}}(W_i) \right)^2$$

and

$$V_{M,K} = \begin{cases} M^{-1}\left(n/b + M - 1\right), & \text{if } K = 1 \\ 1, & \text{otherwise.} \end{cases}$$

**Theorem D.2.** *Let Assumption D.1 hold and* $P\left(f_{\eta^*} - P f_{\eta^*}\right)^2 > 0$. *Then,*

$$P\left(\theta_{\hat{\eta}} \in \left[\hat{\theta}_{\hat{\eta}} - z_{1-\alpha/2}\frac{\hat{\sigma}_{\hat{\eta}}}{\sqrt{n}}, \hat{\theta}_{\hat{\eta}} + z_{1-\alpha/2}\frac{\hat{\sigma}_{\hat{\eta}}}{\sqrt{n}}\right]\right) \to 1 - \alpha.$$

□

The proof of Theorem D.1 relies on four main insights. I show them for the case of repeated cross-fitting, assuming that $n$ is a multiple of $K$ for simplicity. I provide a more detailed proof in Appendix D.1, and a formal proof follows from the more

general Theorems 3.1 and E.1. The first insight and main argument of the proof is to show that

$$\sqrt{n}\left(\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f_{\eta^*}(W_i) - Pf_{\eta^*}\right) + o_P(1). \tag{D.3}$$

Once this is established, the result follows from Lyapunov's CLT, since $(f_{\eta^*}(W_i))_{i=1}^n$ are iid. The second insight is that an application of Markov and Hölder inequalities gives that a sufficient condition for (D.3) is that

$$\mathrm{Var}_P\left[\frac{1}{\sqrt{b}}\sum_{i\in\xi}\left[f_{\hat{\eta}_{\tilde{\xi}}}(W_i) - Pf_{\hat{\eta}_{\tilde{\xi}}}\right] - \left[f_{\eta^*}(W_i) - Pf_{\eta^*}\right]\right] \to 0, \tag{D.4}$$

where $\xi$ is a random subset of $[n]$ of size $b = n/K$ and $\tilde{\xi}$ is its complement. The third insight is that an application of the Law of Total Variance gives

$$\begin{aligned}
&\mathrm{Var}_P\left[\frac{1}{\sqrt{b}}\sum_{i\in\xi}\left[f_{\hat{\eta}_{\tilde{\xi}}}(W_i) - Pf_{\hat{\eta}_{\tilde{\xi}}}\right] - \left[f_{\eta^*}(W_i) - Pf_{\eta^*}\right]\right] \\
&= \mathbb{E}_P\left[\mathrm{Var}_P\left[\frac{1}{\sqrt{b}}\sum_{i\in\xi}\left(f_{\hat{\eta}_{\tilde{\xi}}}(W_i) - Pf_{\hat{\eta}_{\tilde{\xi}}}\right) - \left(f_{\eta^*}(W_i) - Pf_{\eta^*}\right) \,\middle|\, D_{\tilde{\xi}}\right]\right] \qquad \text{(D.5)} \\
&= \mathbb{E}_P\left[\mathrm{Var}_P\left[\left(f_{\hat{\eta}_{\tilde{\xi}}}(W) - Pf_{\hat{\eta}_{\tilde{\xi}}}\right) - \left(f_{\eta^*}(W) - Pf_{\eta^*}\right) \,\middle|\, D_{\tilde{\xi}}\right]\right]. \qquad \text{(D.6)}
\end{aligned}$$

Since the summands in (D.5) are iid conditional on $D_{\tilde{\xi}}$, (D.5) equals (D.6), which does not rely on the term $\sqrt{b}$. This is the crucial step that enables asymptotic normality without requiring an assumption on the rate at which $f_{\hat{\eta}_{\tilde{\xi}}}(W)$ converges to $f_{\eta^*}(W)$.

The final insight is that Assumption D.1 gives a sufficient condition for (D.6) to converge to zero. For any $\varepsilon > 0$,

$$\begin{aligned}
&\mathbb{E}_P\left[\mathrm{Var}_P\left[\left(f_{\hat{\eta}_{\tilde{\xi}}}(W) - Pf_{\hat{\eta}_{\tilde{\xi}}}\right) - \left(f_{\eta^*}(W) - Pf_{\eta^*}\right) \,\middle|\, D_{\tilde{\xi}}\right]\right] \\
&\leqslant \mathbb{E}_P\left[\left(f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right)^2\right] \\
&= \mathbb{E}_P\left[\left(f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right)^2 \mathbb{I}\left(\left|f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right| \leqslant \varepsilon\right)\right] \\
&\quad + \mathbb{E}_P\left[\left(f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right)^2 \mathbb{I}\left(\left|f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right| > \varepsilon\right)\right],
\end{aligned}$$

where the first term is bounded by $\varepsilon^2$. By Hölder's inequality, the second term is bounded by

$$\mathbb{E}_P\left[\left|f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right|^{2+\delta}\right]^{\frac{2}{2+\delta}} P\left(\left|f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right| > \varepsilon\right)^{\frac{\delta}{2+\delta}}.$$

The first term above is bounded by Assumption D.1(i), and the second term can be made arbitrarily small since

$$P\left(\left|f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right| > \varepsilon\right) = \mathbb{E}_P\left[P\left(\left|f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right| > \varepsilon \;\middle|\; W\right)\right]$$

converges to zero by the dominated convergence theorem, since

$$P\left(\left|f_{\hat{\eta}_{\tilde{\xi}}}(w) - f_{\eta^*}(w)\right| > \varepsilon \;\middle|\; W = w\right) = P\left(\left|f_{\hat{\eta}_{\tilde{\xi}}}(w) - f_{\eta^*}(w)\right| > \varepsilon\right) \to 0$$

from Assumption D.1(ii) and independence of $W$ and $\hat{\eta}_{\tilde{\xi}}$. The result follows since $\varepsilon$ can be made arbitrarily small.

## D.1 Proofs

*Proof of Theorem D.1.* I provide a detailed proof for the repeated cross-fitting case discussed in Appendix D, since that contains the main insights of the proof. A complete and formal proof follows from the more general Theorem E.1.

The argument consists of showing that

$$\sqrt{n}\left(\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f_{\eta^*}(W_i) - Pf_{\eta^*}\right) + o_P(1)$$

and applying Lyapunov's CLT to the first term on the right side of the equality.

Define $h(w, \eta) = [f_\eta(w) - Pf_\eta] - [f_{\eta^*}(w) - Pf_{\eta^*}]$ and note that

$$\begin{aligned}
&\sqrt{n}\left(\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}\right) - \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f_{\eta^*}(W_i) - Pf_{\eta^*}\right) \\
&= \sqrt{bK}\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\frac{1}{b}\sum_{i\in\mathsf{s}} h(W_i, \hat{\eta}_{\tilde{\mathsf{s}}}),
\end{aligned}$$

since $b = n/K$ for cross-fitting. For any $\varepsilon > 0$, it holds that

$$
\begin{aligned}
&P\left(\left|\sqrt{bK}\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\frac{1}{b}\sum_{i\in\mathsf{s}}h(W_i,\hat{\eta}_{\tilde{\mathsf{s}}})\right| > \varepsilon\right) \\
&\leqslant P\left(\frac{\sqrt{K}}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\left|\frac{1}{\sqrt{b}}\sum_{i\in\mathsf{s}}h(W_i,\hat{\eta}_{\tilde{\mathsf{s}}})\right| > \varepsilon\right) \\
&\leqslant \varepsilon^{-1}\frac{\sqrt{K}}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\mathbb{E}_P\left[\left|\frac{1}{\sqrt{b}}\sum_{i\in\mathsf{s}}h(W_i,\hat{\eta}_{\tilde{\mathsf{s}}})\right|\right] &\text{(D.7)}\\
&= \varepsilon^{-1}\sqrt{K}\,\mathbb{E}_P\left[\left|\frac{1}{\sqrt{b}}\sum_{i\in\xi}h(W_i,\hat{\eta}_{\tilde{\xi}})\right|\right] &\text{(D.8)}\\
&\leqslant \varepsilon^{-1}\sqrt{K}\,\mathbb{E}_P\left[\left|\frac{1}{\sqrt{b}}\sum_{i\in\xi}h(W_i,\hat{\eta}_{\tilde{\xi}})\right|^2\right]^{1/2} &\text{(D.9)}\\
&= \varepsilon^{-1}\sqrt{K}\,\mathrm{Var}_P\left[\frac{1}{\sqrt{b}}\sum_{i\in\xi}h(W_i,\hat{\eta}_{\tilde{\xi}})\right]^{1/2}. &\text{(D.10)}
\end{aligned}
$$

(D.7) follows from Markov's inequality. (D.8) defines $\xi$ as a random subset of $[n]$ of size $b$, and uses the fact that the expected value does not depend on how the sample is (randomly) split. (D.9) follows from Hölder's inequality. (D.10) follows since

$$
\begin{aligned}
\mathbb{E}_P\left[h(W,\hat{\eta}_{\tilde{\xi}})\right] &= \mathbb{E}_P\left[\left(f_{\hat{\eta}_{\tilde{\xi}}}(W) - Pf_{\hat{\eta}_{\tilde{\xi}}}\right) - \left(f_{\eta^*}(W) - Pf_{\eta^*}\right)\right] \\
&= \mathbb{E}_P\left[\mathbb{E}_P\left[f_{\hat{\eta}_{\tilde{\xi}}}(W) - Pf_{\hat{\eta}_{\tilde{\xi}}} \,\middle|\, D_{\tilde{\xi}}\right]\right] \\
&= 0
\end{aligned}
$$

by definition.

Since $K$ is assumed fixed, it is enough to show that

$$\begin{aligned}
&\mathrm{Var}_P\left[\frac{1}{\sqrt{b}}\sum_{i\in\xi} h(W_i,\hat{\eta}_{\tilde{\xi}})\right] \\
&= \mathrm{Var}_P\left[\frac{1}{\sqrt{b}}\sum_{i\in\xi}\left(f_{\hat{\eta}_{\tilde{\xi}}}(W_i) - Pf_{\hat{\eta}_{\tilde{\xi}}}\right) - \left(f_{\eta^*}(W_i) - Pf_{\eta^*}\right)\right] \\
&= \mathbb{E}_P\left[\mathrm{Var}_P\left[\frac{1}{\sqrt{b}}\sum_{i\in\xi}\left(f_{\hat{\eta}_{\tilde{\xi}}}(W_i) - Pf_{\hat{\eta}_{\tilde{\xi}}}\right) - \left(f_{\eta^*}(W_i) - Pf_{\eta^*}\right) \;\middle|\; D_{\tilde{\xi}}\right]\right] \qquad \text{(D.11)} \\
&= \mathbb{E}_P\left[\mathrm{Var}_P\left[\left(f_{\hat{\eta}_{\tilde{\xi}}}(W) - Pf_{\hat{\eta}_{\tilde{\xi}}}\right) - \left(f_{\eta^*}(W) - Pf_{\eta^*}\right) \;\middle|\; D_{\tilde{\xi}}\right]\right] \qquad \text{(D.12)} \\
&= \mathbb{E}_P\left[\mathrm{Var}_P\left[f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W) \;\middle|\; D_{\tilde{\xi}}\right]\right] \qquad \text{(D.13)}
\end{aligned}$$

converges to zero. (D.11) follows from the Law of Total Variance, since

$$\mathbb{E}_P\left[\left(f_{\hat{\eta}_{\tilde{\xi}}}(W_i) - Pf_{\hat{\eta}_{\tilde{\xi}}}\right) - \left(f_{\eta^*}(W_i) - Pf_{\eta^*}\right) \;\middle|\; D_{\tilde{\xi}}\right] = 0.$$

(D.12) follows since the observations are iid conditional on $D_{\tilde{\xi}}$.

To show convergence to zero of (D.13), consider the inequality

$$\begin{aligned}
\mathbb{E}_P\left[\mathrm{Var}_P\left[f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W) \;\middle|\; D_{\tilde{\xi}}\right]\right] &\leqslant \mathbb{E}_P\left[\mathbb{E}_P\left[\left(f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right)^2 \;\middle|\; D_{\tilde{\xi}}\right]\right] \\
&= \mathbb{E}_P\left[\left(f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right)^2\right].
\end{aligned}$$

For any fixed $\varepsilon > 0$,

$$\begin{aligned}
&\mathbb{E}_P\left[\left(f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right)^2\right] \\
&= \mathbb{E}_P\left[\left(f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right)^2 \mathbb{I}\left(\left|f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right| \leqslant \varepsilon\right)\right] \\
&\quad + \mathbb{E}_P\left[\left(f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right)^2 \mathbb{I}\left(\left|f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right| > \varepsilon\right)\right].
\end{aligned}$$

The first term is bounded by $\varepsilon^2$. By Hölder's inequality,

$$\begin{aligned}
&\mathbb{E}_P\left[\left(f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right)^2 \mathbb{I}\left(\left|f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right| > \varepsilon\right)\right] \\
&\leqslant \mathbb{E}_P\left[\left|f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right|^{2+\delta}\right]^{\frac{2}{2+\delta}} P\left(\left|f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right| > \varepsilon\right)^{\frac{\delta}{2+\delta}}.
\end{aligned}$$

The first term above is bounded by Assumption D.1(i), and the second term can be made arbitrarily small since

$$P\left(\left|f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right| > \varepsilon\right) = \mathbb{E}_P\left[P\left(\left|f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W)\right| > \varepsilon \;\middle|\; W\right)\right]$$

converges to zero by the dominated convergence theorem, since

$$P\left(\left|f_{\hat{\eta}_{\tilde{\xi}}}(w) - f_{\eta^*}(w)\right| > \varepsilon \;\middle|\; W = w\right) = P\left(\left|f_{\hat{\eta}_{\tilde{\xi}}}(w) - f_{\eta^*}(w)\right| > \varepsilon\right) \to 0$$

from Assumption D.1(ii) and independence of $W$ and $\hat{\eta}_{\tilde{\xi}}$. The result follows since $\varepsilon$ can be made arbitrarily small. ■

*Proof of Theorem D.2.* Note

$$\hat{\sigma}^2_{\hat{\eta}_{\tilde{\mathsf{s}}}} = \frac{1}{b}\sum_{i\in\mathsf{s}}\left(f_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W_i)^2\right) - \left(\frac{1}{b}\sum_{i\in\mathsf{s}} f_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W_i)\right)^2.$$

By a law of large numbers conditional on $\tilde{\mathsf{s}}$,

$$\frac{1}{b}\sum_{i\in\mathsf{s}} f_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W_i)^2 - \mathbb{E}_P\left[f_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W)^2 | D_{\tilde{\mathsf{s}}}\right] \xrightarrow{P} 0,$$

and similarly

$$\frac{1}{b}\sum_{i\in\mathsf{s}} f_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W_i) - \mathbb{E}_P\left[f_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W) | D_{\tilde{\mathsf{s}}}\right] \xrightarrow{P} 0.$$

Hence,

$$\hat{\sigma}^2_{\hat{\eta}_{\tilde{\mathsf{s}}}} - \left(\mathbb{E}_P\left[f_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W)^2 | D_{\tilde{\mathsf{s}}}\right] - \mathbb{E}_P\left[f_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W) | D_{\tilde{\mathsf{s}}}\right]^2\right) \xrightarrow{P} 0.$$

Fix $\varepsilon > 0$ and define $h_{\hat{\eta}_{\tilde{\mathsf{s}}}}(w) = |f_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W) - f_{\eta^*}(W)|$.

$$\begin{aligned}
&\mathbb{E}_P\left[h_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W) | D_{\tilde{\mathsf{s}}}\right] \\
&= \mathbb{E}_P\left[h_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W)\mathbb{I}\left(h_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W) \leqslant \varepsilon\right) | D_{\tilde{\mathsf{s}}}\right] + \mathbb{E}_P\left[h_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W)\mathbb{I}\left(h_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W) > \varepsilon\right) | D_{\tilde{\mathsf{s}}}\right] \\
&\leqslant \varepsilon + \mathbb{E}_P\left[h_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W)^{1+\delta} | D_{\tilde{\mathsf{s}}}\right]^{\frac{1}{1+\delta}} P\left(h_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W) > \varepsilon | D_{\tilde{\mathsf{s}}}\right)^{\frac{\delta}{1+\delta}}
\end{aligned}$$

by Hölder's inequality. The term $\mathbb{E}_P\left[h_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W)^{1+\delta} | D_{\tilde{\mathsf{s}}}\right]^{\frac{1}{1+\delta}}$ is bounded by Assumption D.1(i), and I show that $P\left(h_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W) > \varepsilon | D_{\tilde{\mathsf{s}}}\right)^{\frac{\delta}{1+\delta}}$ converges in probability to zero. In the proof of Theorem D.1, I established that

$$\mathbb{E}_P\left[P\left(\left|f_{\hat{\eta}_{\tilde{\xi}}}(w) - f_{\eta^*}(w)\right| > \varepsilon \;\middle|\; D_{\tilde{\mathsf{s}}}\right)\right] = P\left(\left|f_{\hat{\eta}_{\tilde{\xi}}}(w) - f_{\eta^*}(w)\right| > \varepsilon\right) \to 0.$$

This implies that $P\left(\left|f_{\hat{\eta}_{\tilde{\xi}}}(w) - f_{\eta^*}(w)\right| > \varepsilon \;\middle|\; D_{\tilde{\mathsf{s}}}\right) \xrightarrow{P} 0$ since $L_1$ convergence implies convergence in probability. Hence,

$$\mathbb{E}_P\left[\left|f_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W) - f^*_\eta(W)\right| | D_{\tilde{\mathsf{s}}}\right] \xrightarrow{P} 0,$$

which implies

$$\mathbb{E}_P\left[f_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W) | D_{\tilde{\mathsf{s}}}\right] - \mathbb{E}_P\left[f^*_\eta(W)\right] \xrightarrow{P} 0.$$

A similar argument gives

$$\mathbb{E}_P\left[f_{\hat{\eta}_{\tilde{\mathsf{s}}}}(W)^2 | D_{\tilde{\mathsf{s}}}\right] - \mathbb{E}_P\left[f^*_\eta(W)^2\right] \xrightarrow{P} 0.$$

Combining results implies

$$\hat{\sigma}^2_{\hat{\eta}_{\tilde{\mathsf{s}}}} \xrightarrow{P} P\left(f_{\eta^*} - P f_{\eta^*}\right)^2.$$

The result follows from Theorem D.1, since $V_{M,K}/V_{\bar{M},K} \to 1$. ■

# E CLT for Split-Sample Empirical Processes

I derive a CLT for empirical processes based on a broad class of split-sample procedures, uniformly over a large class of probability distributions. This section generalizes Appendix D, which gives a more accessible exposition focusing on the particular case of sample averages. The CLT of this section can be used to prove asymptotic normality for a large class of estimators. That is the case for Z-estimators, which I develop in Section 3. Moreover, this CLT can be used to establish asymptotic consistency of the bootstrap in several applications, following, for example, the arguments in Chapter 3.7 of van der Vaart and Wellner (2023).

The notation follows Section 2. Let $\mathcal{P}$ be a set of probability distributions, and $D = \{W_i\}_{i\in[n]}$, the dataset, be an iid sample of $W \sim P \in \mathcal{P}$. I denote the expected value under $P \in \mathcal{P}$ by $\mathbb{E}_P$, and the variance by $\mathrm{Var}_P$. Given a set $T$, let $f_{t,\eta} : \mathcal{W} \to \mathbb{R}$ be measurable functions for $t \in T$ and $\eta \in H$, with $H$ defined as in Section 2, and let $\mathcal{F}_\eta = \{f_{t,\eta} : t \in T\}$. $\hat{\eta} = \hat{\eta}_{\mathcal{R}} = \left(\hat{\eta}_{\mathsf{s}_{m,k}}\right)_{m\in[M],k\in[K]}$, $\|f\|_{Q,r} = \left(\int |f|^r dQ\right)^{1/r}$, $L_r(Q) = \|\cdot\|_{Q,r}$, and $\mathcal{Q}$ denotes all finitely discrete probability distributions. I use $|x|$ to denote cardinality when $x$ is a set and absolute value when $x$ is scalar. I denote by $N$ and $N_{[]}$ respectively the covering and bracketing numbers, as in Definitions 2.1.5 and 2.1.6 of van der Vaart and Wellner (2023). For $\mathsf{s} \subseteq [n]$, define the empirical measure

$$\mathbb{P}_{\mathsf{s}} f_{t,\eta} = \frac{1}{|\mathsf{s}|}\sum_{i\in\mathsf{s}} f_{t,\eta}(W_i),$$

the marginal expectation

$$Pf_{t,\eta} = \int_w f_{t,\eta}(w)dP(w),$$

and the empirical process

$$\mathbb{G}_{n,\hat{\eta}}(t) = \sqrt{n}\frac{1}{M}\sum_{r\in\mathcal{R}}\frac{1}{K}\sum_{\mathsf{s}\in r}\left(\mathbb{P}_{\mathsf{s}}f_{t,\hat{\eta}_{\bar{\mathsf{s}}}} - Pf_{t,\hat{\eta}_{\bar{\mathsf{s}}}}\right).$$

I establish below sufficient conditions for the CLT for split-sample empirical processes, presented in Theorem E.1.

**Assumption E.1.** *The following conditions hold:*

(i) *$T$ is totally bounded for some semimetric $\rho$;*

(ii) *For every $\eta \in H$ and $t \in T$, $f_{t,\eta}$ is measurable;*

(iii) *For all $\eta \in H$, there exists a measurable envelope function $F_\eta$; That is, $F_\eta : \mathcal{W} \to \mathbb{R}$ is such that $|f_{t,\eta}(w)| \leqslant F_\eta(w) < \infty$ for all $t \in T$ and $w \in \mathcal{W}$;*

(iv) $\lim_{B\to\infty}\sup_{P\in\mathcal{P}}\sup_{\eta\in H}\mathbb{E}_P\left[F_\eta(W)^2\mathbb{I}\left(F_\eta(W) > B\right)\right] = 0$*;*

(v) *For every $\delta_n \downarrow 0$,*

$$\sup_{P\in\mathcal{P}}\sup_{\eta\in H}\sup_{\rho(t,t')<\delta_n}\mathbb{E}_P\left[\left(f_{t,\eta}(W) - f_{t',\eta}(W)\right)^2\right] \to 0;$$

(vi) *One of the following conditions holds for all $\delta_n \downarrow 0$:*

$$\sup_{\eta\in H}\sup_{Q\in\mathcal{Q}}\int_0^{\delta_n}\sqrt{\log N\left(\varepsilon, \mathcal{F}_\eta, L_2(Q)\right)}d\varepsilon \to 0, \tag{E.1}$$

*or*

$$\sup_{P\in\mathcal{P}}\sup_{\eta\in H}\int_0^{\delta_n}\sqrt{\log N_{[]}\left(\varepsilon, \mathcal{F}_\eta, L_2(P)\right)}d\varepsilon \to 0; \tag{E.2}$$

□

**Assumption E.2.** *There exists $\eta_P^* \in H$ such that for $\tilde{\eta} = \mathcal{A}(D)$, $W \perp D$, and every $t \in T$,*

$$\mathrm{Var}_P\left[f_{t,\tilde{\eta}}(W) - f_{t,\eta_P^*}(W) \,\middle|\, D\right] \xrightarrow{P} 0$$

*uniformly in $P \in \mathcal{P}$.* □

Although technical, Assumption E.1 is a weak condition that is satisfied in many applications. Assumption E.1(i) through E.1(vi) are standard Donsker conditions in the literature of weak convergence of empirical processes (e.g., van der Vaart and Wellner, 2023), generalized for the presence of the functions $\eta \in H$. In fact, if $T = \{t\}$ and $\mathcal{P} = \{P\}$ are singletons, these conditions are implied by the "$2+\delta$" moments condition in Assumption D.1(i) (Proposition H.1). These assumptions are standard for proving functional CLTs by limiting the complexity of the sets $T$ and $\mathcal{F}_\eta$. In addition to ensuring that each set $\mathcal{F}_\eta$ is Donsker, Assumption E.1 requires that the inequalities and convergences be uniform in $\eta \in H$. Importantly, Assumption E.1(vi) does not restrict the complexity of the class $H$, and it does not imply the much stronger condition that $\bigcup_{\eta \in H} \mathcal{F}_\eta$ is Donsker. In applications, except for the restrictions on $\mathcal{P}$, Assumption E.1(i) through Assumption E.1(vi) are verifiable since they depend only on the choices of $T$ and $\mathcal{F}_\eta$, and typically do not depend on how $\eta$ is calculated. The assumptions on $\mathcal{P}$ involve the mild uniform square integrability condition Assumption E.1(iv), and the smoothness condition Assumption E.1(v).

Assumptions Assumption E.1(i) through Assumption E.1(vi) give standard conditions for a CLT when $\mathcal{R}$ consists of a single sample split. The proof for the case of multiple splits relies on the additional Assumption E.2. This is a weak stability condition that requires $\tilde{\eta}$ to converge at any rate to any function $\eta_P^*$, which is allowed to depend on $P$. If $T$ and $\mathcal{P}$ are singletons, this is implied by Assumption D.1(ii) (Proposition H.1). Note that the requirement is pointwise in $t \in T$, and it holds, for example, if $f_{t,\tilde{\eta}}(w) \xrightarrow{P} f_{t,\eta_P^*}(w)$ for almost all $w \in \mathcal{W}$.

**Theorem E.1.** *(CLT for split-sample empirical processes)*
*Let Assumptions E.1 and E.2 hold. Then, the sequence $\mathbb{G}_{n,\hat{\eta}}$ is asymptotically $\rho$-equicontinuous uniformly in $P \in \mathcal{P}$ and*

$$\sup_{t \in T} \left| \mathbb{G}_{n,\hat{\eta}}(t) - \mathbb{G}_{n,\eta_P^*}(t) \right| \xrightarrow{P} 0$$

*uniformly in $P \in \mathcal{P}$, where*

$$\mathbb{G}_{n,\eta_P^*}(t) = \sqrt{n} \frac{1}{M} \sum_{r \in \mathcal{R}} \frac{1}{K} \sum_{\mathsf{s} \in r} \left( \mathbb{P}_{\mathsf{s}} f_{t,\eta_P^*} - P f_{t,\eta_P^*} \right).$$

*For any sequence $(P_n)_{n \geqslant 1} \subseteq \mathcal{P}$ such that, for every $t, t' \in T$,*

$$\mathbb{E}_{P_n} \left[ \left( f_{t,\eta_{P_n}^*}(W) - P_n f_{t,\eta_{P_n}^*} \right) \left( f_{t',\eta_{P_n}^*}(W) - P_n f_{t',\eta_{P_n}^*} \right) \right] \to \sigma_{t,t'}, \tag{E.3}$$

*for some $\sigma_{t,t'}$,*

$$\mathbb{G}_{n,\hat{\eta}} \rightsquigarrow \mathbb{G}_{\eta^*}$$

*in $\ell^\infty(T)$, where $\mathbb{G}_{\eta^*}$ is a tight Gaussian process. Moreover, the covariance function of $\mathbb{G}_{\eta^*}$ is given by $V_{\bar{M},K}\sigma_{t,t'}$, where*

$$V_{\bar{M},K} = \begin{cases} \bar{M}^{-1}\left(\pi^{-1} + \bar{M} - 1\right), & \text{if } K = 1 \text{ and } \bar{M} < \infty \\ 1, & \text{otherwise.} \end{cases}$$

□

To the best of my knowledge, this appears to be the first central limit theorem for empirical processes that average over multiple splits of the sample. This result enables asymptotic inference for a large class of split-sample estimators. For example, combined with the functional delta method, it immediately implies asymptotic normality of Hadamard differentiable functionals of the split-sample empirical measure

$$\sqrt{n}\frac{1}{M}\sum_{r\in\mathcal{R}}\frac{1}{K}\sum_{\mathsf{s}\in r}\mathbb{P}_{\mathsf{s}}f_{t,\hat{\eta}_{\mathsf{s}}}.$$

In Section 3, I use Theorem E.1 as a building block to prove asymptotic normality of split-sample Z-estimators, a broad class that cover many if not most estimators used in practice, including the ones in Section 6.

## E.1 Proofs

**Lemma E.1.** *Let Assumptions Assumption E.1(i) through Assumption E.1(vi) hold, $(\eta_n)_{n\geqslant 1} \subseteq H$ be a deterministic sequence, $(P_n)_{n\geqslant 1} \subseteq \mathcal{P}$, and $s \subseteq [n]$ be a random (uniformly) subset of $[n]$ such that $|\mathsf{s}| \to \infty$ as $n \to \infty$. Define*

$$X_{n,s}(t) = \frac{1}{\sqrt{|\mathsf{s}|}}\sum_{i\in\mathsf{s}}\left(f_{t,\eta_n}(W_i) - P_n f_{t,\eta_n}\right)$$

*Then, the sequence $X_{n,s}$ is asymptotically $\rho$-equicontinuous.*

*Proof of Lemma E.1.*
The result follows from an application of Theorems 2.11.1 and 2.11.9 in van der Vaart and Wellner (2023), respectively for when conditions (E.1) and (E.2) hold. Their notation is adapted with $m_n = |\mathsf{s}|$, $\mathcal{F} = T$, and $Z_{ni}(t) = |\mathsf{s}|^{-1/2} f_{t,\eta_n}(W_i)$, where it is implicit in the notation that $W_i \sim P_n$ (alternatively, one could denote $W_{ni}$ instead of $W_i$). The presence of the suprema over $P \in \mathcal{P}$ and $\eta \in H$ guarantee that the conditions in those theorems hold for any sequences $(\eta_n)_{n\geqslant 1}$ and $(P_n)_{n\geqslant 1}$. ■

**Lemma E.2.** *Let Assumptions Assumption E.1(i) through Assumption E.1(vi) hold, and $s \subseteq [n]$ be a random (uniformly) subset such that $|\mathsf{s}| \to \infty$ as $n \to \infty$. Define*

$$X_{n,s,\eta}(t) = \frac{1}{\sqrt{|\mathsf{s}|}} \sum_{i \in \mathsf{s}} \left( f_{t,\eta}(W_i) - P f_{t,\eta} \right).$$

*Then, the sequence $X_{n,s,\hat{\eta}_{\tilde{\mathsf{s}}}}$ is asymptotically $\rho$-equicontinuous uniformly in $P \in \mathcal{P}$.*

*Proof of Lemma E.2.* Let $\mathcal{F}_{\eta,\delta} = \{f - g : f, g \in \mathcal{F}_\eta, \rho(f,g) < \delta\}$ and $\varepsilon > 0$.

$$\begin{aligned}
&\sup_{P \in \mathcal{P}} P\left( \|X_{n,s,\hat{\eta}_{\tilde{\mathsf{s}}}}\|_{\mathcal{F}_{\hat{\eta}_{\tilde{\mathsf{s}}},\delta}} > \varepsilon \right) \\
&= \sup_{P \in \mathcal{P}} \int_{D_s, D_{\tilde{\mathsf{s}}}} \mathbb{I}\left( \left\| X_{n,s,\hat{\eta}_{\tilde{\mathsf{s}}}(D_{\tilde{\mathsf{s}}})}(D_s) \right\|_{\mathcal{F}_{\hat{\eta}_{\tilde{\mathsf{s}}}(D_{\tilde{\mathsf{s}}}),\delta}} > \varepsilon \right) dP(D_s, D_{\tilde{\mathsf{s}}}) \qquad \text{(E.4)} \\
&= \sup_{P \in \mathcal{P}} \int_{D_{\tilde{\mathsf{s}}}} \left[ \int_{D_s} \mathbb{I}\left( \left\| X_{n,s,\hat{\eta}_{\tilde{\mathsf{s}}}(D_{\tilde{\mathsf{s}}})}(D_s) \right\|_{\mathcal{F}_{\hat{\eta}_{\tilde{\mathsf{s}}}(D_{\tilde{\mathsf{s}}}),\delta}} > \varepsilon \right) dP(D_s) \right] dP(D_{\tilde{\mathsf{s}}}) \qquad \text{(E.5)} \\
&\leqslant \sup_{P \in \mathcal{P}} \int_{D_{\tilde{\mathsf{s}}}} \sup_{\eta \in H} \left[ \int_{D_s} \mathbb{I}\left( \|X_{n,s,\eta}(D_s)\|_{\mathcal{F}_{\eta,\delta}} > \varepsilon \right) dP(D_s) \right] dP(D_{\tilde{\mathsf{s}}}) \\
&= \sup_{P \in \mathcal{P}} \sup_{\eta \in H} \left[ \int_{D_s} \mathbb{I}\left( \|X_{n,s,\eta}(D_s)\|_{\mathcal{F}_{\eta,\delta}} > \varepsilon \right) dP(D_s) \right] \\
&= \sup_{P \in \mathcal{P}} \sup_{\eta \in H} P\left( \|X_{n,s,\eta}(D_s)\|_{\mathcal{F}_{\eta,\delta}} > \varepsilon \right),
\end{aligned}$$

where (E.4) makes explicit the dependence of $X_{n,s,\hat{\eta}_{\tilde{\mathsf{s}}}}$ on the subsample $D_s$ and of $\hat{\eta}_{\tilde{\mathsf{s}}}$ on $D_{\tilde{\mathsf{s}}}$, and (E.5) uses the fact that the split is random and $D_s, D_{\tilde{\mathsf{s}}}$ are independent.

Hence, for an arbitrary $\delta_n \downarrow 0$, $\sup_{P \in \mathcal{P}} P\left( \|X_{n,s,\hat{\eta}_{\tilde{\mathsf{s}}}}\|_{\mathcal{F}_{\hat{\eta}_{\tilde{\mathsf{s}}},\delta_n}} > \varepsilon \right) \to 0$ follows from

$$\sup_{P \in \mathcal{P}} P\left( \|X_{n,s,\eta_n}(D_s)\|_{\mathcal{F}_{\eta_n,\delta_n}} > \varepsilon \right) \to 0$$

for any deterministic $(\eta_n)_{n \geqslant 1} \subseteq H$, which is established in Lemma E.1. ■

*Proof of Theorem E.1.*
The proof is divided into three main steps. First, I show that

$$\sup_{t \in T} \left| \mathbb{G}_{n,\hat{\eta}}(t) - \mathbb{G}_{n,\eta_P^*}(t) \right| \xrightarrow{P} 0$$

uniformly in $P \in \mathcal{P}$. Second, I show that $\mathbb{G}_{n,\eta_P^*}$ is asymptotically $\rho$-equicontinuous. Finally, I prove the Gaussian limit of $(\mathbb{G}_{n,\eta_P^*}(t))_{t \in T'}$ for any finite $T' \subseteq T$.

**Step one.** Let $h_t(w,\eta) = [f_{t,\eta}(w) - Pf_{t,\eta}] - \left[f_{t,\eta_P^*}(w) - Pf_{t,\eta_P^*}\right]$, $\pi_n = b/n$, and fix $\varepsilon > 0$. It follows that

$$\begin{aligned}
&\sup_{P\in\mathcal{P}} P\left(\sup_{t\in T}\left|\mathbb{G}_{n,\hat{\eta}}(t) - \mathbb{G}_{n,\eta_P^*}(t)\right| > \varepsilon\right) \\
&= \sup_{P\in\mathcal{P}} P\left(\sup_{t\in T}\left|\sqrt{n}\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\frac{1}{b}\sum_{i\in\mathsf{s}} h_t(W_i,\hat{\eta}_{\mathsf{s}})\right| > \varepsilon\right) \\
&\leqslant \sup_{P\in\mathcal{P}} P\left(\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\sup_{t\in T}\left|\frac{\sqrt{n}}{b}\sum_{i\in\mathsf{s}} h_t(W_i,\hat{\eta}_{\mathsf{s}})\right| > \varepsilon\right) \\
&\leqslant \varepsilon^{-1}\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\sup_{P\in\mathcal{P}}\mathbb{E}_P\left[\sup_{t\in T}\left|\frac{\sqrt{n}}{b}\sum_{i\in\mathsf{s}} h_t(W_i,\hat{\eta}_{\mathsf{s}})\right|\right] \qquad \text{(E.6)} \\
&= \varepsilon^{-1}\sup_{P\in\mathcal{P}}\mathbb{E}_P\left[\sup_{t\in T}\left|\frac{\sqrt{n}}{b}\sum_{i\in\xi} h_t(W_i,\hat{\eta}_{\xi})\right|\right] \qquad \text{(E.7)} \\
&= \varepsilon^{-1}\pi_n^{-1/2}\sup_{P\in\mathcal{P}}\mathbb{E}_P\left[\sup_{t\in T}\left|\frac{1}{\sqrt{b}}\sum_{i\in\xi} h_t(W_i,\hat{\eta}_{\xi})\right|\right], \qquad \text{(E.8)}
\end{aligned}$$

where (E.6) follows from Markov's inequality, and (E.7) defines $\xi$ as a random subset of $[n]$ (uniformly over all subsets).

Since $\pi_n \to \pi \in (0,1)$, (E.8) converges to zero if the term inside the expectation convergences in probability to zero uniformly in $P \in \mathcal{P}$, since it is uniformly integrable (by Assumption E.1(iv)). This follows from stochastic equicontinuity of $\frac{1}{\sqrt{b}}\sum_{i\in\xi} h_t(W_i,\hat{\eta}_{\xi})$ (as a process indexed by $t \in T$) and pointwise convergence in $t$, by applying Theorem 22.9 in Davidson (2021). Stochastic equicontinuity follows since

$$\frac{1}{\sqrt{b}}\sum_{i\in\xi} h_t(W_i,\hat{\eta}_{\xi}) = \frac{1}{\sqrt{b}}\sum_{i\in\xi}\left[f_{t,\hat{\eta}_{\xi}}(W_i) - Pf_{t,\hat{\eta}_{\xi}}\right] - \frac{1}{\sqrt{b}}\sum_{i\in\xi}\left[f_{t,\eta_P^*}(W_i) - Pf_{t,\eta_P^*}\right]$$

is a sum of two stochastically equicontinuous processes, respectively by Lemma E.2 and Lemma E.1. For pointwise convergence, I show that the variance converges to

zero, and note $h_t(w,\eta)$ is mean zero by construction. For an arbitrary $t \in T$,

$$\begin{aligned}
&\sup_{P\in\mathcal{P}} \mathrm{Var}_P\left[\frac{1}{\sqrt{b}}\sum_{i\in\xi} h_t(W_i,\hat{\eta}_{\tilde{\xi}})\right] \\
&= \sup_{P\in\mathcal{P}} \mathrm{Var}_P\left[\frac{1}{\sqrt{b}}\sum_{i\in\xi}\left(f_{t,\hat{\eta}_{\tilde{\xi}}}(W_i) - Pf_{t,\hat{\eta}_{\tilde{\xi}}}\right) - \left(f_{t,\eta_P^*}(W_i) - Pf_{t,\eta_P^*}\right)\right] \\
&= \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[\mathrm{Var}_P\left[\frac{1}{\sqrt{b}}\sum_{i\in\xi}\left(f_{t,\hat{\eta}_{\tilde{\xi}}}(W_i) - Pf_{t,\hat{\eta}_{\tilde{\xi}}}\right) - \left(f_{t,\eta_P^*}(W_i) - Pf_{t,\eta_P^*}\right) \,\middle|\, D_{\tilde{\xi}}\right]\right] \quad \text{(E.9)} \\
&= \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[\mathrm{Var}_P\left[\left(f_{t,\hat{\eta}_{\tilde{\xi}}}(W) - Pf_{t,\hat{\eta}_{\tilde{\xi}}}\right) - \left(f_{t,\eta_P^*}(W) - Pf_{t,\eta_P^*}\right) \,\middle|\, D_{\tilde{\xi}}\right]\right] \quad \text{(E.10)} \\
&= \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[\mathrm{Var}_P\left[f_{t,\hat{\eta}_{\tilde{\xi}}}(W) - f_{t,\eta_P^*}(W) \,\middle|\, D_{\tilde{\xi}}\right]\right],
\end{aligned}$$

where (E.9) uses the Law of Total Variance and the fact that $\mathbb{E}_P\left[h_t(W_i,\hat{\eta}_{\tilde{\xi}})|D_{\tilde{\xi}}\right] = 0$, and (E.10) follows since the summands are iid conditional on $D_{\tilde{\xi}}$. Finally, the last term converges to zero from Assumption E.2. Note that since $f_{t,\eta}$ are uniformly square integrable by Assumption E.1(iv), convergence in probability of the conditional variance implies its convergence in $L_1$.

**Step two.** Let $\lambda_i = (\pi_n MK)^{-1}\left|\left\{s \in \{s_{m,k}\}_{m\in[M],k\in[K]} : i \in \mathsf{s}\right\}\right|$ and note that

$$\mathbb{G}_{n,\eta_P^*}(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \lambda_i\left(f_{t,\eta_P^*}(W_i) - Pf_{t,\eta_P^*}\right).$$

Let $\lambda = (\lambda_i)_{i\in[n]}$, $\mathcal{F}_{\eta_P^*,\delta} = \left\{f - g : f,g \in \mathcal{F}_{\eta_P^*}, \rho(f,g) < \delta\right\}$, $\varepsilon > 0$, and $\delta_n \downarrow 0$.

$$\begin{aligned}
\sup_{P\in\mathcal{P}} P\left(\left\|\mathbb{G}_{n,\eta_P^*}\right\|_{\mathcal{F}_{\eta_P^*,\delta_n}} > \varepsilon\right) &= \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[P\left(\left\|\mathbb{G}_{n,\eta_P^*}\right\|_{\mathcal{F}_{\eta_P^*,\delta_n}} > \varepsilon \,\middle|\, \lambda\right)\right] \\
&\leqslant \sup_{\lambda:\lambda_i\leqslant\pi_n^{-1}} \sup_{P\in\mathcal{P}} P\left(\left\|\mathbb{G}_{n,\eta_P^*}\right\|_{\mathcal{F}_{\eta_P^*,\delta_n}} > \varepsilon \,\middle|\, \lambda\right).
\end{aligned}$$

The last term converges to zero from asymptotic equicontinuity of $\frac{1}{\sqrt{n}}\sum_{i=1}^n \lambda_{n,i}\left(f_{t,\eta_{P_n}^*}(W_i) - Pf_{t,\eta_{P_n}^*}\right)$ for arbitrary sequences $(P_n)_{n\geqslant 1} \subseteq \mathcal{P}$ and $(\lambda_{n,i})_{n\geqslant 1, i\in[n]}$ satisfying $\lambda_{n,i} \leqslant \pi_n^{-1}$ for all $n, i$. Asymptotic equicontinuity can be verified under Assumption E.1, for example,

by applying Theorem 2.11.1 (when (E.1) holds) and Theorem 2.11.9 (when (E.2) holds) of van der Vaart and Wellner (2023). Their notation is adapted with $m_n = n$, $\mathcal{F} = T$, and $Z_{ni}(t) = n^{-1/2}\lambda_{n,i} f_{t,\eta^*_{P_n}}(W_i)$, where it is implicit in the notation that $W_i \sim P_n$ (alternatively, one could denote $W_{ni}$ instead of $W_i$). For $\gamma > 0$, note that

$$\begin{aligned}
&\mathbb{E}_P\left[\sup_{t\in T}\left(n^{-1/2}\lambda_{n,i} f_{t,\eta^*_{P_n}}(W_i)\right)^2 \mathbb{I}\left(\sup_{t\in T}\left|n^{-1/2}\lambda_{n,i} f_{t,\eta^*_{P_n}}(W_i)\right| > \gamma\right)\right] \\
&\leqslant \pi_n^{-2}\,\mathbb{E}_P\left[\sup_{t\in T}\left(n^{-1/2} f_{t,\eta^*_{P_n}}(W_i)\right)^2 \mathbb{I}\left(\sup_{t\in T}\left|n^{-1/2}\pi_n^{-1} f_{t,\eta^*_{P_n}}(W_i)\right| > \gamma\right)\right],
\end{aligned}$$

and

$$\begin{aligned}
&\left(n^{-1/2}\lambda_{n,i} f_{t,\eta^*_{P_n}}(W_i) - n^{-1/2}\lambda_{n,i} f_{t',\eta^*_{P_n}}(W_i)\right)^2 \\
&\leqslant \pi_n^{-2}\left(n^{-1/2} f_{t,\eta^*_{P_n}}(W_i) - n^{-1/2} f_{t',\eta^*_{P_n}}(W_i)\right)^2,
\end{aligned}$$

for any $t, t' \in T$, $n$, and $i \in [n]$.

**Step three.** If $K > 1$, $\lambda_i = 1$ for all $i$, and the Gaussian limit follows from Lindeberg's CLT and the Cramér-Wold device, using Assumption E.1(iv).

For $K = 1$ and $\bar{M} < \infty$, let

$$M_i = \left|\left\{s \in \{s_{m,k}\}_{m\in[M],k\in[K]} : i \in \mathsf{s}\right\}\right|,$$

so $\lambda_i = (\pi_n M)^{-1} M_i$.

$$\mathrm{Var}_{P_n}\left[\mathbb{G}_{n,\eta^*_P}(t) \mid \lambda\right] = \mathbb{E}_{P_n}\left[\left(f_{t,\eta^*_{P_n}}(W) - P_n f_{t,\eta^*_{P_n}}\right)^2\right]\frac{1}{n}\sum_{i=1}^n \lambda_i^2.$$

Without loss of generality, let $M = \bar{M}$.

$$\frac{1}{n}\sum_{i=1}^n \lambda_i^2 = \frac{1}{\pi_n^2 \bar{M}^2}\sum_{j=1}^{\bar{M}} j^2 \frac{1}{n}\sum_{i=1}^n \mathbb{I}\left(M_i = j\right).$$

In Lemma H.1, I show that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{I}\left(M_i = j\right) \xrightarrow{P_n} \binom{\bar{M}}{j}\pi^j(1-\pi)^{\bar{M}-j}.$$

Hence,

$$\sum_{j=1}^{\bar{M}} j^2 \frac{1}{n}\sum_{i=1}^n \mathbb{I}\left(M_i = j\right) \xrightarrow{P_n} \sum_{j=1}^{\bar{M}} j^2 \binom{\bar{M}}{j}\pi^j(1-\pi)^{\bar{M}-j} = \pi(1-\pi)\bar{M} + (\pi\bar{M})^2,$$

since the sum in the right is the second moment of a binomial distribution with parameters $\bar{M}$ and $\pi$. Collecting the results,

$$\frac{1}{n}\sum_{i=1}^{n}\lambda_i^2 \xrightarrow{P_n} 1 + (1-\pi)\pi^{-1}M^{-1}.$$

The Gaussian limit follows from Lindeberg's CLT conditional on $\lambda$ and the dominated convergence theorem, and the Cramér-Wold device.

Finally, let $K = 1$ and $\bar{M} = \infty$. I show that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\lambda_i\left(f_{t,\eta^*_{P_n}}(W_i) - P_n f_{t,\eta^*_{P_n}}\right) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(f_{t,\eta^*_{P_n}}(W_i) - P_n f_{t,\eta^*_{P_n}}\right) \tag{E.11}$$

converges to zero in $L_2$. For the mean,

$$\begin{aligned}
&\mathbb{E}_{P_n}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\lambda_i - 1)\left(f_{t,\eta^*_{P_n}}(W_i) - P_n f_{t,\eta^*_{P_n}}\right)\right] \\
&= \sqrt{n}\,\mathbb{E}_{P_n}\left[(\lambda_1 - 1)\left(f_{t,\eta^*_{P_n}}(W_1) - P_n f_{t,\eta^*_{P_n}}\right)\right] \\
&= \sqrt{n}\,\mathbb{E}_{P_n}\Bigg[(\lambda_1 - 1)\underbrace{\mathbb{E}_{P_n}\left[f_{t,\eta^*_{P_n}}(W_1) - P_n f_{t,\eta^*_{P_n}} \;\middle|\; \lambda_1\right]}_{=0}\Bigg].
\end{aligned}$$

For the variance,

$$\begin{aligned}
&\mathrm{Var}_{P_n}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\lambda_i - 1)\left(f_{t,\eta^*_{P_n}}(W_i) - P_n f_{t,\eta^*_{P_n}}\right)\right] \\
&= \mathbb{E}_{P_n}\left[\mathrm{Var}_{P_n}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\lambda_i - 1)\left(f_{t,\eta^*_{P_n}}(W_i) - P_n f_{t,\eta^*_{P_n}}\right) \;\middle|\; \lambda\right]\right] \\
&= \mathrm{Var}_{P_n}\left[f_{t,\eta^*_{P_n}}(W_i) - P_n f_{t,\eta^*_{P_n}}\right]\mathbb{E}_{P_n}\left[\frac{1}{n}\sum_{i=1}^{n}(\lambda_i - 1)^2\right] \\
&= \mathrm{Var}_{P_n}\left[f_{t,\eta^*_{P_n}}(W_i) - P_n f_{t,\eta^*_{P_n}}\right]\mathbb{E}_{P_n}\left[(\lambda_1 - 1)^2\right],
\end{aligned}$$

where the first equality follows since $\mathbb{E}_{P_n}\left[f_{t,\eta^*_{P_n}}(W) - P_n f_{t,\eta^*_{P_n}} \;\middle|\; \lambda\right] = 0$ by the Law of Total Variance, and the second equality since the summands are iid conditional on

$\lambda$. Since $\lambda_1 - 1$ is bounded, $\mathbb{E}_{P_n}\left[(\lambda_1 - 1)^2\right] \to 0$ if $\lambda_1 \xrightarrow{P_n} 1$, which follows from

$$\begin{aligned}\lambda_1 &= (\pi_n M)^{-1} M_1 \\ &= (\pi_n)^{-1} \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}\left(1 \in s_{m,1}\right) \xrightarrow{P_n} 1\end{aligned}$$

by a law of large numbers, since $\mathbb{E}_{P_n}\left[\mathbb{I}\left(1 \in s_{m,1}\right)\right] = P_n(1 \in s_{m,1}) = \pi_n$ and splits are independent. Finally, the Gaussian limit follows from Lindeberg's CLT and the Cramér-Wold device. ■

# F Inference with Fast Converging Moments

Consider the setting of Section 4.1. The normal approximation CI (4.3) may not cover $h(\theta_{\hat{\eta}})$ with nominal probability when the variance of any moment function evaluated at the limit parameter $\theta_{\eta_P^*}$ is 0, that is,

$$\text{Var}_P\left[\psi_{\theta_{\eta_P^*}, \eta_P^*, j}(W)\right] = 0 \tag{F.1}$$

for any $j \in [1, \ldots, d]$. If that happens, either $\sigma^2_{\eta_P^*} = 0$, $\dot{\Psi}_{\eta_P^*}$ is not invertible, or both. If $\sigma^2_{\eta_P^*} = 0$, (4.1) implies that the centered estimator multiplied by $\sqrt{n}$ converges in probability to zero, and the normal approximation in (4.3) may not be accurate. Similarly, if $\dot{\Psi}_{\eta_P^*}$ is not invertible, $V^*_{\eta_P^*}$ is not well-defined, and the normal approximation may be inaccurate. In this subsection, I provide an approach to inference on $\theta_{\hat{\eta}}$ that is general in considering the class of Z-estimators in Section 3.

I explore the fact that (F.1) implies that the empirical moment equation evaluated at $\theta_{\hat{\eta}}$ converges faster than the typical $\sqrt{n}$ rate to construct a confidence interval for $\theta_{\hat{\eta}}$ that is uniformly asymptotically valid regardless of whether (F.1) happens or not. The issue discussed in this section is not important for every application. First, I discuss examples of when one may or may not comfortably assume that (F.1) does not hold. Then, I propose a confidence interval, prove its uniform asymptotic validity, and characterize its power properties. I focus on the estimator $\hat{\theta}_{\hat{\eta}} = \hat{\theta}^{(2)}_{\hat{\eta}}$ from Section 3, and the results can be extended to $\hat{\theta}^{(1)}_{\hat{\eta}}$ and $\hat{\theta}^{(3)}_{\hat{\eta}}$ using similar techniques.

## F.1 Examples

In many cases, the researcher can safely assume that (F.1) does not happen, depending on the setup and definition of $\psi_{\theta,\eta}$. In other cases, as in Section 7, (F.1) can happen under one of the main hypotheses of interest. I present examples of both cases below.

**Example 2** (Revisited)**.** *In Example 2, $W = (Y, X)$, $Y$ is binary, and $\hat{\eta} : \mathcal{X} \to \{0, 1\}$ is a predictor of $Y$ using covariates $X$. The parameter of interest is a split-sample Z-estimand with $\psi_{\theta,\eta}(y, x) = \mathbb{I}(y = \eta(x)) - \theta$:*

$$\theta_{\hat{\eta}} = \frac{1}{MK} \sum_{r \in \mathcal{R}} \sum_{\mathsf{s} \in r} \int \mathbb{I}\left(y = \hat{\eta}_{\tilde{\mathsf{s}}}(x)\right) dP(y, x).$$

*The variance*

$$\mathrm{Var}_P \left[ \psi_{\theta_{\eta_P^*}, \eta_P^*}(Y, X) \right] = P\left(Y = \eta_P^*(X)\right) \left[1 - P\left(Y = \eta_P^*(X)\right)\right]$$

*is positive unless $\eta_P^*(X)$ always predicts $Y$ correctly or always incorrectly. In practice, predictive algorithms rarely have a near perfect (or imperfect) performance, and in many cases the researcher can confidently assume $\mathrm{Var}_P[\psi_{\theta_{\eta_P^*}, \eta_P^*}(Y, X)] > 0$.* □

**Example 9.** *Consider a dataset with covariates $X$, a mean zero continuous outcome $Y \in \mathbb{R}$, and the goal of assessing whether a predictor $\hat{\eta}(X)$ has predictive power for $Y$. One way of assessing predictive power for $Y$ is by conducting inference on the covariance*

$$\theta_{\hat{\eta}} = \frac{1}{MK} \sum_{r \in \mathcal{R}} \sum_{\mathsf{s} \in r} \int y \hat{\eta}_{\tilde{\mathsf{s}}}(x) dP(y, x).$$

*$\theta_{\hat{\eta}}$ is a Z-estimand with moment function $\psi_{\theta,\eta}(y, x) = y\eta(x) - \theta$, and its limit variance is*

$$\mathrm{Var}_P \left[ \psi_{\theta_{\eta_P^*}, \eta_P^*}(Y, X) \right] = \mathrm{Var}_P \left[ Y \eta_P^*(X) \right].$$

*Let $\eta^*(x) = \mathbb{E}[Y | X = x]$ be the limit of $\hat{\eta}(x)$. If $X$ has no predictive power for $Y$, for example because $Y$ and $X$ are independent, $\eta^*(x) = 0$, and $\mathrm{Var}_P[Y \eta_P^*(X)] = 0$. Hence, the CI in (4.3) may fail to achieve nominal coverage asymptotically.* □

**Remark F.1.** *When $\mathrm{Var}_P[\psi_{\theta_{\eta_P^*}, \eta_P^*}(Y, X)] = 0$, the asymptotic distribution of $\hat{\theta}_{\hat{\eta}}$ may depend on the specific structure of how $\hat{\eta}$ is calculated. Let $Y$ be a mean zero scalar random variable, $K = 2$, $M = 1$, and $(\mathsf{s}, \tilde{\mathsf{s}})$ be a 2-fold cross-split of the data of equal sizes. Let $\psi_{\theta,\hat{\eta}_{\tilde{\mathsf{s}}}}(y) = y \bar{y}_{\tilde{\mathsf{s}}}^d - \theta$ for some odd positive $d$, where $\bar{y}_{\tilde{\mathsf{s}}} = \frac{1}{|\tilde{\mathsf{s}}|} \sum_{i \in \tilde{\mathsf{s}}} Y_i$. Then, $\hat{\theta}_{\hat{\eta}} = \frac{1}{2}\left(\bar{y}_{\mathsf{s}} \bar{y}_{\tilde{\mathsf{s}}}^d + \bar{y}_{\tilde{\mathsf{s}}} \bar{y}_{\mathsf{s}}^d\right)$ and $n^{1/2+d/2} \hat{\theta}_{\hat{\eta}} = \frac{1}{2}\left((\sqrt{n}\bar{y}_{\mathsf{s}})(\sqrt{n}\bar{y}_{\tilde{\mathsf{s}}})^d + (\sqrt{n}\bar{y}_{\tilde{\mathsf{s}}})(\sqrt{n}\bar{y}_{\mathsf{s}})^d\right)$, which follows a non-trivial distribution that depends on $d$. If, for example, $d = 3$, $(\sqrt{n}\bar{y}_{\mathsf{s}})^d$ is approximately distributed as the cube of a standard normal distribution, and $d = 5$ leads to a different distribution. Moreover, the dependence between $(\sqrt{n}\bar{y}_{\mathsf{s}})(\sqrt{n}\bar{y}_{\tilde{\mathsf{s}}})^d$ and $(\sqrt{n}\bar{y}_{\tilde{\mathsf{s}}})(\sqrt{n}\bar{y}_{\mathsf{s}})^d$ is not trivial.* □

## F.2 An Adaptive Confidence Interval

I show how to construct a confidence interval $\hat{C}_{1-\alpha}$ that satisfies

$$\lim_{n\to\infty} \inf_{P\in\mathcal{P}} P(\theta_{\hat{\eta}} \in \hat{C}_{1-\alpha}) = 1-\alpha,$$

regardless of whether (F.1) may hold or not, by introducing a tuning parameter. In Section 4.2 and Appendix B.5, I propose a different approach for the particular cases of inference on comparisons between models and in the Generic ML context of Chernozhukov et al. (2025b), which explicitly account for the dependence across splits.

I construct $\hat{C}_{1-\alpha}$ by inverting the test

$$\begin{cases} H_{0,\hat{\eta}}: & h(\theta_{\hat{\eta}}) = \tau \\ H_{A,\hat{\eta}}: & h(\theta_{\hat{\eta}}) \neq \tau, \end{cases} \tag{F.2}$$

that is, $\hat{C}_{1-\alpha}$ contains all values of $\tau$ for which the null hypothesis is not rejected at significance level $\alpha$. My approach consists of a data-driven procedure to choose one of two p-values for testing (F.2): $p_c(\tau)$ or $p_e(\tau)$. $p_c(\tau)$ is a *conservative* p-value, meant to be valid when (F.1) holds, that is, the p-value a researcher would use if they knew (F.1) were true. $p_e(\tau)$ is an *exact* p-value, coming from the normal approximation (4.3), as it achieves exact nominal coverage in large samples when (F.1) does not hold. Hence, I test (F.2) with the p-value $p_e(\tau)$ when the data suggest that the empirical moment equations are away from zero, and with $p_c(\tau)$ otherwise. The idea of using different tests based on pre-testing some condition (in this case, whether the empirical moment equations are away from zero), is similar to Shi (2015), in the context of moment inequalities. Specifically,

$$p_e(\tau) = 2\Phi\left(-\left|\frac{\sqrt{n}\left(h(\hat{\theta}_{\hat{\eta}}) - \tau\right)}{\hat{\sigma}_{\hat{\eta}}}\right|\right),$$

$$\hat{\Psi}_{\min}(\tau) = \min_{j\in[d]}\left|\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\hat{\Psi}_{\mathsf{s},\hat{\eta}_{\bar{\mathsf{s}}},j}(\tau)\right|,$$

$$\hat{\Psi}(\tau) = \left\|\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\hat{\Psi}_{\mathsf{s},\hat{\eta}_{\bar{\mathsf{s}}}}(\tau)\right\|,$$

$$a_n(\tau) = \mathbb{I}\left(\hat{\Psi}_{\min}(\tau)\hat{\Psi}(\tau) > \gamma_n\right),$$

$$p(\tau) = a_n(\tau)p_e(\tau) + \left[1 - a_n(\tau)\right]p_c(\tau),$$

$$\hat{C}_{1-\alpha} = \{\tau \in \mathbb{R} : p(\tau) \leqslant \alpha\}.$$

The idea behind $a_n$ is that $\sqrt{n}\hat{\Psi}_{\min}(\tau) \xrightarrow{P} 0$ when $\text{Var}_P[\psi_{\theta_{\eta_P^*},\eta_P^*}(W)] = 0$. The sequence $\gamma_n$ is a tuning parameter that should ideally be specified before the data analysis. The properties of $\gamma_n$ and $p_c(\tau)$ are specified in Assumption F.1 below.

**Assumption F.1.** *The following conditions hold:*

(i) $\sup_{P\in\mathcal{P}} P(p_c(\theta_{\hat{\eta}}) \leqslant \alpha) \leqslant \alpha$;

(ii) $n\gamma_n \to \gamma \in (0, \infty)$;

(iii) *The set $\mathcal{P}$ can be decomposed as $\mathcal{P} = \mathcal{P}_+ \bigcup \mathcal{P}_0$, where*

*(a) For every $\varepsilon > 0$,*

$$\sup_{P\in\mathcal{P}_+} \sup_{\left\|\theta-\theta_{\eta_P^*}\right\|>\varepsilon} -\left\|\Psi_{\eta_P^*}(\theta)\right\| < 0 = \left\|\Psi_{\eta_P^*}(\theta_{\eta_P^*})\right\|,$$

*$\Psi_\eta$ is differentiable at $\theta_\eta$ for $\eta \in H$, and for some $\bar{c}_1 > 0$,*

$$\inf_{P\in\mathcal{P}_+} \left|\det\left(\dot{\Psi}_{\eta_P^*}\right)\right| \geqslant \bar{c}_1;$$

*(b) $\sup_{P\in\mathcal{P}_0} \left\|\Psi_{\eta_P^*}(\theta_{\eta_P^*})\right\| = 0$, $\theta_{\hat{\eta}} \xrightarrow{P} \theta_{\eta_P^*}$ for some $\theta_{\eta_P^*} \in \Theta'$ uniformly in $P \in \mathcal{P}_0$, and $\sup_{P\in\mathcal{P}_0} \min_{j\in[d]} \text{Var}_P\left[\psi_{\theta_{\eta_P^*},\eta_P^*,j}(W)\right] = 0$.*

□

Assumption F.1(i) requires the p-value $p_c(\tau)$ to be valid, even if conservative, including when $\text{Var}_P[\psi_{\theta_{\eta_P^*},\eta_P^*}(W)] = 0$. Constructing $p_c(\tau)$ is context-specific, but a conservative, trivially valid option is $p_c(\tau) = 1$. Note that this option does not lead to an unbounded CI since $a_n(\tau) = 1$ with probability approaching one for values of $\tau$ far from $\theta_{\hat{\eta}}$. Assumption F.1(ii) requires $\gamma_n$ to converge to zero at the $n^{-1}$ rate. Assumption F.1(iii) substitutes and weakens Assumption B.1(iii) and Assumption B.1(v). It allows $\dot{\Psi}_{\eta_P^*}$ to be singular and $\|\Psi_{\eta_P^*}(\theta)\| = 0$ to have multiple solutions for $\theta$ when the variance of $\psi_{\theta_{\eta_P^*},\eta_P^*,j}(W)$ is zero for some $j$. Valid inference is achieved in these cases since $a_n = 0$ with probability approaching one. Note that Assumption B.1(iii) and Assumption B.1(v) imply Assumption F.1(iii) since $\mathcal{P} = \mathcal{P}_+$, and if

$$\inf_{P\in\mathcal{P}} \min_{j\in[d]} \text{Var}_P\left[\psi_{\theta_{\eta_P^*},\eta_P^*,j}(W)\right] > 0,$$

Assumption F.1(iii) implies both Assumption B.1(iii) and Assumption B.1(v). I establish the uniform asymptotic validity of $\hat{C}_{1-\alpha}$, and explore its power properties.

**Theorem F.1.** *(Uniform Asymptotic Validity of $\hat{C}_{1-\alpha}$)*
*Let Assumptions B.1(i)-B.1(iv) and F.1 hold. Then,*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} P(\theta_{\hat{\eta}} \in \hat{C}_{1-\alpha}) \geqslant 1-\alpha.$$

□

I show that the hypothesis test (F.2), where $H_{0,\hat{\eta}}$ is rejected if $p(\tau) > \alpha$, has power approaching 1 for fixed alternatives and non-trivial power for some sequences of local alternative hypotheses. I compare my test with an oracle test that correctly picks $p_e(\tau)$ or $p_c(\tau)$ depending on the asymptotic behavior of $\hat{\theta}_{\hat{\eta}}$. In order to study local power, I consider sequences $(P_n)_{n\geqslant 1} \subseteq \mathcal{P}$ under different regimes for the limit behavior of $\sqrt{n}\,\|\theta_{\hat{\eta}} - \tau\|$ and the variance of $\psi_{\theta_{\eta^*_{P_n}},\eta^*_{P_n}}(W)$. Let

$$v_n^2 = \min_{j\in[d]} \mathrm{Var}_{P_n}\left[\psi_{\theta_{\eta^*_{P_n}},\eta^*_{P_n},j}(W)\right].$$

The oracle test is defined by

$$p^*(\tau) = \begin{cases} p_c(\tau), \text{ if } v_n \to 0 \text{ and } \sqrt{n}\,\|\theta_{\hat{\eta}} - \tau\| = O_{P_n}(1), \\ p_e(\tau), \text{ otherwise.} \end{cases}$$

This test is infeasible since it depends on the sequence of DGPs $(P_n)_{n\geqslant 1}$. For the different regimes, I compare the limits

$$\pi_\alpha = \lim_{n\to\infty} P_n\left(p(\tau) \leqslant \alpha\right),$$
$$\pi^*_\alpha = \lim_{n\to\infty} P_n\left(p^*(\tau) \leqslant \alpha\right).$$

**Theorem F.2.** *Let Assumption B.1(i)-Assumption B.1(iv) and Assumption F.1 hold, $\tau \in \mathbb{R}$, $\alpha \in (0,1)$, and $(P_n)_{n\geqslant 1}$ be a sequence such that the limits $v$, $\pi_\alpha$ and $\pi^*_\alpha$ exist. Assume $p_c(\tau)$ is an independent Bernoulli random variable taking value 0 with probability $\alpha$ and 1 with probability $1-\alpha$ (that is, it rejects the null with probability $\alpha$). Then, the relationships in Table 5 hold, where each row defines a separate regime for $\sqrt{n}\,\|\theta_{\hat{\eta}} - \tau\|$.* □

## F.3 Proofs and Extra Definitions

*Proof of Theorem F.1.* Let $(P_n)_{n\geqslant 1} \subseteq \mathcal{P}$ be such that

$$v = \lim_{n\to\infty} \min_{j\in[d]} \mathrm{Var}_{P_n}\left[\psi_{\theta_{\eta^*_{P_n}},\eta^*_{P_n},j}(W)\right]$$

Table 5: Power Comparison by Regime

| | $nv_n^2 = o(1)$ | $nv_n^2 = O(1)^*$ | $nv_n^2 \to \infty^{**}$ | $v_n^2 \to v^2 > 0$ |
|---|---|---|---|---|
| $\sqrt{n}\,\|\theta_{\hat{\eta}} - \tau\| \xrightarrow{P_n} \infty$ | $\alpha = \pi_\alpha < \pi_\alpha^*$ | $\alpha < \pi_\alpha < \pi_\alpha^*$ | $\pi_\alpha = \pi_\alpha^* = 1$ | $\pi_\alpha = \pi_\alpha^* = 1$ |
| $\sqrt{n}\,\|\theta_{\hat{\eta}} - \tau\| = O_{P_n}(1)^{***}$ | $\alpha = \pi_\alpha < \pi_\alpha^*$ | $\alpha = \pi_\alpha < \pi_\alpha^*$ | $\alpha = \pi_\alpha < \pi_\alpha^*$ | $\alpha < \pi_\alpha < \pi_\alpha^*$ |
| $\sqrt{n}\,\|\theta_{\hat{\eta}} - \tau\| \xrightarrow{P_n} 0$ | $\alpha = \pi_\alpha = \pi_\alpha^*$ | $\alpha = \pi_\alpha = \pi_\alpha^*$ | $\alpha = \pi_\alpha = \pi_\alpha^*$ | $\alpha = \pi_\alpha = \pi_\alpha^*$ |

* Assumes $nv_n^2 \nrightarrow 0$; ** Assumes $v_n^2 \to 0$; *** Assumes $\sqrt{n}\,\|\theta_{\hat{\eta}} - \tau\| \neq o_{P_n}(1)$.

exists.

If $v > 0$, $p(\tau) \geqslant p_e(\tau)$, and

$$P_n(\theta_{\hat{\eta}} \in \hat{C}_{1-\alpha}) \geqslant 1 - \alpha + o(1)$$

by Theorem 3.1.

For $v = 0$, note that by Theorem E.1,

$$\begin{aligned}
&\sqrt{n}\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\hat{\Psi}_{\mathsf{s},\hat{\eta}_{\bar{\mathsf{s}}}}(\theta_{\hat{\eta}}) \\
&= \sqrt{n}\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\left(\hat{\Psi}_{\mathsf{s},\hat{\eta}_{\bar{\mathsf{s}}}}(\theta_{\hat{\eta}}) - \Psi_{\hat{\eta}_{\bar{\mathsf{s}}}}(\theta_{\hat{\eta}})\right) \\
&= \sqrt{n}\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\left(\hat{\Psi}_{\mathsf{s},\eta^*_{P_n}}(\theta_{\hat{\eta}}) - \Psi_{\eta^*_{P_n}}(\theta_{\hat{\eta}})\right) + o_{P_n}(1) \\
&= \sqrt{n}\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\left(\hat{\Psi}_{\mathsf{s},\eta^*_{P_n}}(\theta_{\eta^*_{P_n}}) - \Psi_{\eta^*_{P_n}}(\theta_{\eta^*_{P_n}})\right) + o_{P_n}(1) \\
&= \sqrt{n}\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\hat{\Psi}_{\mathsf{s},\eta^*_{P_n}}(\theta_{\eta^*_{P_n}}) + o_{P_n}(1),
\end{aligned}$$

and, for any $j \in [d]$,

$$\mathrm{Var}_{P_n}\left[\frac{\sqrt{n}}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\hat{\Psi}_{\mathsf{s},\hat{\eta}_{\bar{\mathsf{s}}},j}(\theta_{\hat{\eta}})\right] \leqslant \mathrm{Var}_{P_n}\left[\psi_{\theta_{\eta^*_{P_n}},\eta^*_{P_n},j}(W)\right] + o(1).$$

If $v = 0$,

$$\mathrm{Var}_{P_n}\left[\frac{\sqrt{n}}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\hat{\Psi}_{\mathsf{s},\hat{\eta}_{\bar{\mathsf{s}}},j_n}(\theta_{\hat{\eta}})\right] \to 0$$

for some $(j_n)_{n\geqslant 1}$, and hence

$$\mathrm{Var}_{P_n}\left[\min_{j\in[d]}\left|\frac{\sqrt{n}}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\hat{\Psi}_{\mathsf{s},\hat{\eta}_{\bar{\mathsf{s}}},j}(\theta_{\hat{\eta}})\right|\right]\to 0.$$

As a consequence,

$$P_n\Big(\underbrace{\sqrt{n}\hat{\Psi}_{\min}(\theta_{\hat{\eta}})}_{=o_{P_n}(1)}\underbrace{\sqrt{n}\hat{\Psi}(\tau)}_{=O_{P_n}(1)} > \underbrace{n\gamma_n}_{=\gamma+o(1)}\Big)\to 0,$$

and $a_n(\theta_{\hat{\eta}})\xrightarrow{P_n}0$, which concludes the proof. ■

*Proof of Theorem F.2.* Define

$$\Psi_n(\tau)=\left\|P_n\frac{1}{MK}\sum_{r\in\mathcal{R}}\sum_{\mathsf{s}\in r}\hat{\Psi}_{\mathsf{s},\hat{\eta}_{\bar{\mathsf{s}}}}(\tau)\right\|.$$

First, let $\sqrt{n}\left\|\theta_{\hat{\eta}}-\tau\right\|\xrightarrow{P_n}\infty^*$. If $nv_n\to\infty$,

$$P_n\Big(\underbrace{\sqrt{n}\hat{\Psi}_{\min}(\theta_{\hat{\eta}})}_{=O_{P_n}(1)}\Big[\underbrace{\sqrt{n}\hat{\Psi}(\tau)-\sqrt{n}\Psi_n(\tau)}_{=O_{P_n}(1)}+\underbrace{\sqrt{n}\Psi_n(\tau)}_{\xrightarrow{P_n}\infty}\Big]\underbrace{\sqrt{n}\hat{\Psi}(\tau)}_{=O_{P_n}(1)} > \underbrace{n\gamma_n}_{=\gamma+o(1)}\Big)\xrightarrow{P_n}1,$$

since

$$\sqrt{n}\left(\Psi_{\hat{\eta}_{\bar{\xi}}}(\tau)-\Psi_{\hat{\eta}_{\bar{\xi}}}(\theta_{\hat{\eta}})\right)=\left(\dot{\Psi}_{\eta^*_{P_n}}+o_{P_n}(1)\right)\sqrt{n}\left(\tau-\theta_{\hat{\eta}}\right).$$

If $nv_n=O(1)$,

$$P_n\Big(\underbrace{v_n^{-1}\sqrt{n}\hat{\Psi}_{\min}(\theta_{\hat{\eta}})}_{=O_{P_n}(1)}\Big[\underbrace{v_n\sqrt{n}(\hat{\Psi}(\tau)-\Psi_n(\tau))}_{=o_{P_n}(1)}+\underbrace{v_n\sqrt{n}\Psi_n(\tau)}_{=O_{P_n}(1)}\Big]\underbrace{\sqrt{n}\hat{\Psi}(\tau)}_{=O_{P_n}(1)} > \underbrace{n\gamma_n}_{=\gamma+o(1)}\Big)$$

is $O_{P_n}(1)$, and $a_n(\tau)=O_{P_n}(1)$. If $nv_n=o(1)$, $v_n\sqrt{n}\Psi_n(\tau)=o_{P_n}(1)$, and $a_n\xrightarrow{P_n}0$.

Second, let $\sqrt{n}\left\|\theta_{\hat{\eta}}-\tau\right\|=O_{P_n}(1)$, and hence $\sqrt{n}\hat{\Psi}(\tau)=O_{P_n}(1)$. It follows that

$$P_n\Big(\underbrace{v_n^{-1}\sqrt{n}\hat{\Psi}_{\min}(\theta_{\hat{\eta}})}_{=O_{P_n}(1)}\underbrace{\sqrt{n}\hat{\Psi}(\tau)}_{=O_{P_n}(1)} > v_n^{-1}\underbrace{n\gamma_n}_{=\gamma+o(1)}\Big)$$

is $O_{P_n}(1)$ if $v_n^2\to v^2>0$, and converges to zero if $v_n\to 0$.

Finally, the last row follows since $p_e(\theta_{\hat{\eta}})-p_e(\tau)=o_{P_n}(1)$. Note that $p_e(\theta_{\hat{\eta}})-p_e(\tau)>o_{P_n}(1)$ for the second row. ■

# G Note on Comparing Two Nonparametric Models

I discuss an extension of the setting of Section 4.2 for comparing $\theta_{\hat{\eta}}$ to the performance of another model $\hat{\eta}'$, computed with the same split-sample approach as $\hat{\eta}$.

Let

$$\mathcal{S} = (\mathsf{s}_{m,k})_{m\in[M],k\in[K]}$$

and denote the split-specific models $\hat{\eta} = (\hat{\eta}_{\mathsf{s}})_{\mathsf{s}\in\mathcal{S}}$ and $\hat{\eta}' = (\hat{\eta}'_{\mathsf{s}})_{\mathsf{s}\in\mathcal{S}}$, where $\hat{\eta}_{\mathsf{s}} = \mathcal{A}(D_{\tilde{\mathsf{s}}})$ and $\hat{\eta}'_{\mathsf{s}} = \mathcal{A}'(D_{\tilde{\mathsf{s}}})$, that is, the two models are trained using the same sample but different algorithms. For example, $\hat{\eta}$ could be estimated with random forests while $\hat{\eta}'$ could be estimated with neural networks. Denote

$$\hat{\delta}^{(1)}_{\hat{\eta},\hat{\eta}'} = \left(\hat{\theta}_{\hat{\eta}_{\mathsf{s}}} - \hat{\theta}_{\hat{\eta}'}\right)_{\mathsf{s}\in\mathcal{S}}$$

and

$$\hat{\delta}^{(2)}_{\hat{\eta},\hat{\eta}'} = \left(\hat{\theta}_{\hat{\eta}'_{\mathsf{s}}} - \hat{\theta}_{\hat{\eta}}\right)_{\mathsf{s}\in\mathcal{S}}.$$

$\hat{\delta}^{(1)}_{\hat{\eta},\hat{\eta}'}$ can be used for testing whether $\hat{\theta}_{\hat{\eta}_{\mathsf{s}}} \geqslant \hat{\theta}_{\hat{\eta}'}$ for all $\mathsf{s}\in\mathcal{S}$ versus the alternative that $\hat{\theta}_{\hat{\eta}_{\mathsf{s}}} < \hat{\theta}_{\hat{\eta}'}$ for at least one $\mathsf{s}\in\mathcal{S}$, similarly to Section 4.2.1 and Theorem 4.2. Note that the Donsker and rate conditions in Assumption 4.1(ii) are not required for Theorem 4.2. They are used only for the pointwise Theorem 4.3 to cover the case $\theta_{\eta^*_P} = \theta_{b_P}$. Similarly, $\hat{\delta}^{(2)}_{\hat{\eta},\hat{\eta}'}$ can be used to test whether $\hat{\theta}_{\hat{\eta}'_{\mathsf{s}}} \geqslant \hat{\theta}_{\hat{\eta}}$ for all $\mathsf{s}\in\mathcal{S}$ versus the alternative that $\hat{\theta}_{\hat{\eta}'_{\mathsf{s}}} < \hat{\theta}_{\hat{\eta}}$ for at least one $\mathsf{s}\in\mathcal{S}$.

# H Additional Results

**Proposition H.1.** *In the context of Assumption E.1, let $T = \{t\}$ and $\mathcal{P} = \{P\}$ be singletons, and let Assumption D.1 hold for $P$ and $f_{\eta,t}$ measurable with $f_{\eta,t}(w) < \infty$ for all $w$. Then, Assumption E.1 holds.* □

*Proof of Proposition H.1.* Assumption E.1(i) and Assumption E.1(ii) hold trivially. Assumption E.1(iii) holds by taking $f_{\eta,t}$ as its own envelope. The uniform integrability condition Assumption E.1(iv) is implied by the $2+\delta$ assumption Assumption D.1(i). Assumption E.1(v) holds trivially. Assumption E.1(vi) holds since both covering and bracketing numbers are equal to 1 with singleton $T$. Finally, Assumption E.2 follows since

$$\mathbb{E}_P\left[\mathrm{Var}_P\left[f_{\hat{\eta}_{\tilde{\xi}}}(W) - f_{\eta^*}(W) \;\middle|\; D_{\tilde{\xi}}\right]\right] \to 0,$$

as established under Assumption D.1(ii) in the proof of Theorem D.1 (D.13), since convergence in $L_1$ implies convergence in probability. ■

**Lemma H.1.** *In the context of Theorem E.1,*

$$Z_n = n^{-1} \sum_{i=1}^{n} \mathbb{I}\left(M_i = c\right) \xrightarrow{P_n} \binom{M}{c} \pi^c (1-\pi)^{M-c}.$$

*Proof.* I show that $\mathbb{E}_{P_n}\left[Z_n\right] = \binom{M}{c} {\pi_n}^c (1-\pi_n)^{M-c}$ and $\operatorname{Var}_{P_n}\left[Z_n\right] \to 0$ as $n \to \infty$. By definition, $M_i = \left|\left\{s \in \{s_{m,1}\}_{m \in [M]} : i \in s\right\}\right|$. $\mathbb{E}_{P_n}\left[Z_n\right] = \mathbb{E}_{P_n}\left[\mathbb{I}\left(M_1 = c\right)\right] = P_n\left(M_1 = c\right)$ since all $M_i$ are equally distributed for any $i$. The event $\{M_i = c\}$ is equivalent to the event that observation $i$ is chosen in exactly $c$ of the $M$ splits of the sample. Since the splits are independent, $M_i$ follows a binomial distribution with parameters $M$ and $\pi_n$. Hence, the probability of this event is $\binom{M}{c} {\pi_n}^c (1-\pi_n)^{M-c}$.

To show that $\operatorname{Var}_{P_n}\left[Z_n\right] \to 0$, I use the fact that

$$\begin{aligned}
\operatorname{Var}_{P_n}\left[Z_n\right] &= n^{-2} \sum_{i=1}^{n} \operatorname{Var}_{P_n}\left[\mathbb{I}\left(M_i = c\right)\right] + n^{-2} \sum_{i \neq j} \operatorname{Cov}_{P_n}\left[\mathbb{I}\left(M_i = c\right), \mathbb{I}\left(M_j = c\right)\right] \\
&= n^{-1} \operatorname{Var}_{P_n}\left[\mathbb{I}\left(M_1 = c\right)\right] + n^{-2} n(n-1) \operatorname{Cov}_{P_n}\left[\mathbb{I}\left(M_1 = c\right), \mathbb{I}\left(M_2 = c\right)\right].
\end{aligned}$$

Hence, it's enough to show that

$$\operatorname{Cov}_{P_n}\left[\mathbb{I}\left(M_1 = c\right), \mathbb{I}\left(M_2 = c\right)\right] = P_n\left(M_1 = c, M_2 = c\right) - P_n\left(M_1 = c\right)^2 \to 0.$$

I show that $P_n\left(M_1 = c \,\middle|\, M_2 = c\right) \to P_n\left(M_1 = c\right)$. Note $b = \pi_n n$ is the number of draws in each split. Using combinatorial arguments, the conditional probability is given by

$$\begin{aligned}
P_n\left(M_1 = c \,\middle|\, M_2 = c\right) = \sum_{t=0}^{c} \binom{c}{t} \binom{M-c}{c-t} &\left(\frac{\binom{n-2}{b-2}}{\binom{n-1}{b-1}}\right)^t \left(1 - \frac{\binom{n-2}{b-2}}{\binom{n-1}{b-1}}\right)^{c-t} \\
&\times \left(\frac{\binom{n-2}{b-1}}{\binom{n-1}{b}}\right)^{c-t} \left(1 - \frac{\binom{n-2}{b-1}}{\binom{n-1}{b}}\right)^{M-2c+t}.
\end{aligned}$$

$t$ represents the number of splits that contain both observations 1 and 2. Since observation 2 is chosen in $c$ splits, $0 \leqslant t \leqslant c$. There are $\binom{c}{t}$ ways of choosing

among the $c$ splits that contain observation 2, which $t$ will also contain observation 1. There are $\binom{M-c}{c-t}$ ways of choosing the remaining $c-t$ splits that contain observation 1 but not 2. $\left(\frac{\binom{n-2}{b-2}}{\binom{n-1}{b-1}}\right)^t$ is the probability of choosing observation 1 in the $t$ splits that contain both observations. $\left(1-\frac{\binom{n-2}{b-2}}{\binom{n-1}{b-1}}\right)^{c-t}$ is the probability of not choosing observation 1 in the remaining $c-t$ splits that contain observation 2. $\left(\frac{\binom{n-2}{b-1}}{\binom{n-1}{b}}\right)^{c-t}$ is the probability of choosing observation 1 in the $c-t$ splits that contain observation 1 but not 2. Finally, $\left(1-\frac{\binom{n-2}{b-1}}{\binom{n-1}{b}}\right)^{M-2c+t}$ is the probability of not choosing observation 1 in the remaining $M-2c+t$ splits that contain neither observation.

For large $n$, we can approximate the combinatorial terms:

$$\frac{\binom{n-2}{b-2}}{\binom{n-1}{b-1}} = \frac{(n-2)!}{(b-2)!(n-b)!}\left(\frac{(n-1)!}{(b-1)!(n-b)!}\right)^{-1} = \frac{b-1}{n-1} = \pi_n + o(1).$$

Similarly,

$$\frac{\binom{n-2}{b-1}}{\binom{n-1}{b}} = \frac{(n-2)!}{(b-1)!(n-b-1)!}\left(\frac{(n-1)!}{b!(n-b-1)!}\right)^{-1} = \frac{b}{n-1} = \pi_n + o(1).$$

It follows that

$$\begin{aligned} P_n\left(M_1 = c \,\middle|\, M_2 = c\right) &= \sum_{t=0}^{c} \binom{c}{t}\binom{M-c}{c-t} {\pi_n}^t (1-\pi_n)^{c-t} {\pi_n}^{c-t} (1-\pi_n)^{M-2c+t} + o(1) \\ &= {\pi_n}^c (1-\pi_n)^{M-c} \sum_{t=0}^{c} \binom{c}{t}\binom{M-c}{c-t} + o(1) \\ &= {\pi_n}^c (1-\pi_n)^{M-c} \binom{M}{c} + o(1) \\ &= P_n\left(M_1 = c\right) + o(1), \end{aligned}$$

where the third equality uses Vandermonde's Identity. ■

# I Covariate Adjustment in Randomized Trials

Let $W = (Y, A, X)$, where $Y \in \mathbb{R}$ is an observed outcome, $A$ is a binary (randomized) treatment assignment indicator, and $X \in \mathcal{X} \subseteq \mathbb{R}^d$ is a set of covariates, for some $d \geqslant 1$. Let $Y(1), Y(0)$ denote potential outcomes respectively under treatment and control, and $Y = AY(1) + (1-A)Y(0)$. In the simplest form of an RCT, $A \perp (X, Y(1), Y(0))$. In this setting, the ATE $\theta$ can be identified from the regression

$$Y = \alpha + \theta A + \varepsilon. \tag{I.1}$$

The covariates are not necessary for identification of $\theta$. However, adding regressors in (I.1) can lead to power improvement by reducing the variance of the error term $\varepsilon$ and thus the asymptotic variance of the least squares estimator of $\theta$. One approach to incorporating covariates is through a covariate-adjustment term $\eta(X)$:

$$Y = \alpha_\eta + \theta_\eta D + \beta_\eta \eta(X) + \varepsilon. \tag{I.2}$$

If $A \perp (X, Y(1), Y(0))$, $\theta_\eta = \theta$ does not depend on $\eta$. Still, its OLS estimator $\hat{\theta}_\eta$ does depend on $\eta$. In practice, one needs to estimate $\eta$ with a model $\hat{\eta}$. Inference becomes challenging if the same data is used to estimate both $\hat{\eta}$ and $\hat{\theta}_{\hat{\eta}}$ because the observations in (I.2) become no longer iid. The asymptotic distribution of $\sqrt{n}(\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}})$ can be characterized following Section 3, specifically Theorem 3.1.