

Cross-Fitting Under Nonregularity: Normality and Inference via Locality

Bruno Fava*

Stanford University

October 2, 2026

Abstract

Cross-fitting is routine in much of applied research. While conventional confidence intervals that ignore cross-fold dependence are asymptotically valid in several settings, they undercover in many applications that share a common form of nonregularity: from the classic cross-validation problem of testing whether a fitted model outperforms another, to testing for heterogeneous treatment effects with machine learning, to estimating the value of a potentially non-unique optimal treatment regime. Exploiting a new locality condition, I show that a large class of cross-fitting estimators still satisfies a central limit theorem despite the nonregularity, but with an asymptotic variance that must be adjusted for the cross-fold correlation. Then, I propose a method for estimating this correlation and construct new confidence intervals that attain asymptotically nominal coverage. Finally, I show that the proposed confidence intervals attain approximately nominal coverage in a simulation study with random forests and neural networks.

*Contact: bfava@stanford.edu. I am really grateful to Stefan Wager for numerous very helpful conversations. 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

Cross-fitting is widely popular in applied research. By using every observation both for training and for testing, it improves efficiency over sample-splitting, but doing so introduces statistical dependence across folds that can make inference challenging. In many settings, including semiparametric estimation and debiased machine learning, this dependence is asymptotically negligible, and confidence intervals that treat folds as independent have approximately nominal coverage (Schick, 1986; Hubbard et al., 2016; Chernozhukov et al., 2018; Newey and Robins, 2018). In several important applications, however, a form of nonregularity keeps the dependence from vanishing, and conventional confidence intervals undercover: from the classic cross-validation problem of comparing the predictive performance of fitted models (Stone, 1977; Efron and Tibshirani, 1997; Bayle et al., 2026), to testing for treatment effect heterogeneity (Chernozhukov et al., 2025; Wager, 2025), to learning the value of a potentially non-unique optimal treatment regime (Robins, 2004; Luedtke and van der Laan, 2016).

I revisit the asymptotics of cross-fitting and make two contributions. First, exploiting a new locality argument, I show that a large class of split-sample estimators still satisfies a central limit theorem despite the nonregularity, but with an asymptotic variance that depends on the cross-fold correlation. Second, I propose a method for estimating this correlation and constructing confidence intervals based on cross-fitting that attain asymptotically nominal coverage in many nonregular settings.

To introduce the problem I study, and previous results, consider a simple example: estimating the conditional predictive performance with 2-fold cross-validation (Efron, 1983; Picard and Cook, 1984; Efron and Tibshirani, 1997; Blum et al., 1999; Dudoit and van der Laan, 2005; Yang, 2006; Goldstein et al., 2016; Bayle et al., 2020; Lei, 2020; Austern and Zhou, 2025; Luo and Barber, 2025), a routine exercise in applications such as clinical prediction (Jiang et al., 2008; Pirracchio et al., 2015). Given an i.i.d. sample of an outcome Y and covariates X , $(Y_i, X_i) \sim P$ for $i = 1, \dots, n$, an analyst randomly splits $\{1, \dots, n\}$ into two folds $\mathbf{s}_1$ and $\mathbf{s}_2$ of equal size, and uses a learning algorithm, such as a random forest, to train a model $\hat{\mu}_k$ to predict Y from X on each fold $\mathbf{s}_k$, $k = 1, 2$. The analyst then evaluates each predictor on the other fold, and averages the two estimates:

$$\widehat{\text{Err}} = \frac{1}{2} (\widehat{\text{Err}}_1 + \widehat{\text{Err}}_2), \quad \widehat{\text{Err}}_k = \frac{1}{n/2} \sum_{i \notin \mathbf{s}_k} L(Y_i, \hat{\mu}_k(X_i)),$$

where L is a loss function, such as $L(y, \hat{y}) = (y - \hat{y})^2$. A natural and popular target is the out-of-sample loss of the fitted predictors (Efron, 1983; Dudoit and van der Laan, 2005),

$$\text{Err} = \frac{1}{2} (\text{Err}_1 + \text{Err}_2), \quad \text{Err}_k = \mathbb{E} \left[L(\tilde{Y}, \hat{\mu}_k(\tilde{X})) \mid \{Y_i, X_i\}_{i \in \mathbf{s}_k} \right],$$

where $(\tilde{Y}, \tilde{X}) \sim P$ independent of the sample.

Conventional inference on Err uses a concentration argument that renders cross-fold dependence negligible. If $L(Y_i, \hat{\mu}_k(X_i)) = \bar{L}_n(Y_i, X_i) + o_p(1)$, with $\bar{L}_n(y, x) = \mathbb{E}[L(y, \hat{\mu}_k(x))]$, then, under standard regularity conditions,

$$\sqrt{n/2} (\widehat{\text{Err}}_k - \text{Err}_k) = \frac{1}{\sqrt{n/2}} \sum_{i \notin \mathbf{s}_k} \left\{ \bar{L}_n(Y_i, X_i) - \mathbb{E}[\bar{L}_n(\tilde{Y}, \tilde{X})] \right\} + o_p(1), \quad (1)$$

cross-fold dependence is first-order negligible, $\widehat{\text{Err}}$ is approximately an average of i.i.d. terms, and $\sqrt{n}(\widehat{\text{Err}} - \text{Err})$ is asymptotically normal. This is the core mechanism in much of the cross-validation literature with random centering (Bayle et al., 2020; Austern and Zhou, 2025; Lei, 2025), which often verifies concentration around $\bar{L}_n$ through algorithmic stability, and is very close to the arguments in Hubbard et al. (2016) and Fava (2025).

In practice, however, Err is usually not of interest on its own, but in comparison with the error of another model or a benchmark. Let $\hat{\nu}_k$ denote another model trained on fold $\mathbf{s}_k$ using a different learner, such as a neural network or the sample average $\hat{\nu}_k(x) = (n/2)^{-1} \sum_{i \in \mathbf{s}_k} Y_i$. In this case, the cross-validation estimator is

$$\widehat{\Delta} = \frac{1}{2} (\widehat{\Delta}_1 + \widehat{\Delta}_2), \quad \widehat{\Delta}_k = \frac{1}{n/2} \sum_{i \notin \mathbf{s}_k} \widehat{\delta}_k(Y_i, X_i),$$

$$\widehat{\delta}_k(y, x) = L(y, \hat{\mu}_k(x)) - L(y, \hat{\nu}_k(x)),$$

and the estimand is

$$\Delta = \frac{1}{2} (\Delta_1 + \Delta_2), \quad \Delta_k = \mathbb{E}[\widehat{\delta}_k(\tilde{Y}, \tilde{X}) \mid \{Y_i, X_i\}_{i \in \mathbf{s}_k}].$$

In this setting, the concentration argument in (1) does not immediately apply, especially under the leading null hypothesis $\Delta = 0$. If both $\hat{\mu}_k$ and $\hat{\nu}_k$ converge to the same limit, $\widehat{\delta}_k(y, x) = o_p(1)$, and hence $\sqrt{n}(\widehat{\Delta} - \Delta) = o_p(1)$, which is uninformative for inference. A more useful result would require a deterministic function $\bar{\delta}_n(y, x) = \mathbb{E}[\widehat{\delta}_k(y, x)]$ such that

$$\frac{\widehat{\delta}_k(Y_i, X_i)}{\sigma_n} = \frac{\bar{\delta}_n(Y_i, X_i)}{\sigma_n} + o_p(1), \quad (2)$$

with $\sigma_n^2 = \text{Var}[\widehat{\delta}_k(\tilde{Y}, \tilde{X})]$, $\sigma_n^2 \rightarrow 0$. This requirement is counterintuitive: if for example $E[Y|X]$ is constant, the difference in performances $\widehat{\delta}_k(Y_i, X_i)$ is pure sampling noise, and it would not be unnatural that $\bar{\delta}_n(y, x) = 0$.

The failure of (2) is a known form of nonregularity (Robins, 2004; Laber and Murphy, 2011), previously addressed with procedures other than cross-fitting. Yadlowsky

et al. (2025) use a single train/holdout split. Sequential schemes evaluate each observation with a rule trained only on earlier data (Luedtke and van der Laan, 2016; Wager, 2025), and hence train and evaluate models on less data than cross-fitting does. Repeated sample-splitting combined with rules such as the median or twice the median of p -values, or majority votes over intervals, controls size (van de Wiel et al., 2009; DiCiccio et al., 2020; Gasparin and Ramdas, 2024; Gasparin et al., 2025; Chernozhukov et al., 2025), as do confidence intervals that inflate the standard error or add noise (Rinaldo et al., 2019; Dai et al., 2024; Verdinelli and Wasserman, 2024; Bayle et al., 2026) and the multivariate approach of Fava (2025), but all of these procedures are conservative and/or sacrifice power.

I focus instead on cross-fitting itself, and construct confidence intervals that attain asymptotically nominal coverage by exploiting a novel locality condition. My first contribution is a central limit theorem (CLT) that accommodates this nonregularity. My CLT applies to a broad class of split-sample estimators, covering K -fold and leave-one-out cross-fitting. Unlike in previous cross-validation CLTs, the asymptotic variance in my result depends on the correlation between fold-level estimators, which turns out to be positive in simulations, so that omitting it would lead to undercoverage.

My CLT builds on the intuition that the average of dependent random variables can still be asymptotically normal as long as the dependence structure is “local”. In the model comparison example,

$$\frac{\hat{\Delta}}{\sigma_n} = \frac{1}{n} \left(\sum_{i \in \mathbf{s}_1} \frac{\hat{\delta}_2(Y_i, X_i)}{\sigma_n} + \sum_{j \in \mathbf{s}_2} \frac{\hat{\delta}_1(Y_j, X_j)}{\sigma_n} \right),$$

and the threat to normality is the statistical dependence between $\hat{\delta}_2(Y_i, X_i)$ and $\hat{\delta}_1(Y_j, X_j)$ for $i \in \mathbf{s}_1$ and $j \in \mathbf{s}_2$. Dependence, however, need not invalidate normality. Indeed, there is a long tradition of CLTs for dependent data. In the time series literature, for example, it is typical to control dependence by assuming that two observations are “nearly independent” if they are “far apart” in time. I formalize the notion of locality in the cross-fitting setting: changing an observation (Y_i, X_i) with $i \in \mathbf{s}_1$ can affect all of $\{\hat{\delta}_1(Y_j, X_j)\}_{j \in \mathbf{s}_2}$, but this effect must be “small” if (Y_i, X_i) is “far” from (Y_j, X_j) . This is natural for nonparametric learners: predictions $\hat{\mu}_1(X_j)$ are more sensitive to a training point (Y_i, X_i) when X_j is close to X_i . But the locality condition is not universal: for parametric learners, changing one training point can affect all test points equally.

My CLT paves the way for inference, but raises a new challenge: estimating the cross-fold correlation. My second contribution is to address this problem. I propose a resampling-based estimator that applies to any black-box learner and many estimators, achieving nominal coverage asymptotically under mild regularity conditions.

Section 2 lays out the setting and notation, and Section 3 derives the central limit theorem. Section 4 is devoted to inference and feasible confidence intervals, and Section 5 contains a Monte Carlo study supporting the theoretical results. Section 6 concludes.

The main text focuses on two-fold cross-fitting to simplify the exposition. In Appendix A, I extend the result to K -fold cross-fitting, including leave-one-out. All proofs are deferred to the appendix.

2 Setting and Notation

I define the setting and notation, introduce the cross-fitting estimator and its estimand, and present three applications that fit into this framework, including the cross-validation model comparison problem discussed in the introduction.

Consider an i.i.d. sample $\mathcal{D} = \{W_i\}_{i=1}^n$ of even size n , with $W_i \sim P$ taking values in $\mathcal{W}$. For example, $W = (Y, X)$ for an outcome Y and covariates X . Further, consider a random split of the indices $[n] = \{1, \dots, n\}$ into two disjoint folds $\mathbf{s}_1, \mathbf{s}_2$ of equal size, independently of the data. Without loss of generality, I treat the split as fixed. For any subset $\mathbf{s} \subseteq [n]$, let $\mathcal{D}_{\mathbf{s}} = \{W_i : i \in \mathbf{s}\}$ denote the corresponding subsample and $\tilde{\mathbf{s}} = [n] \setminus \mathbf{s}$ the complement of $\mathbf{s}$.

Let $\mathcal{A}$ denote a training algorithm, a measurable map $\mathcal{A} : \bigcup_{m \geq 1} \mathcal{W}^m \rightarrow \mathcal{H}$, invariant under permutations of its arguments, that takes a training sample $\mathcal{D}$ of any size and outputs a model $\mathcal{A}(\mathcal{D}) \in \mathcal{H}$. Moreover, let $\{f_\eta : \eta \in \mathcal{H}\}$ be a known family of functions $f_\eta : \mathcal{W} \rightarrow \mathbb{R}$, and let $\hat{\eta}_k = \mathcal{A}(\mathcal{D}_{\mathbf{s}_k})$ be the model trained on the k -th fold.

I study the cross-fitting estimator

$$\hat{\theta}_{\hat{\eta}} = \frac{1}{2} \sum_{k=1}^2 \hat{\theta}_{\hat{\eta}_k}, \quad \hat{\theta}_{\hat{\eta}_k} = \frac{1}{n/2} \sum_{i \notin \mathbf{s}_k} f_{\hat{\eta}_k}(W_i),$$

where $\hat{\eta} = (\hat{\eta}_1, \hat{\eta}_2)$. The estimand is

$$\theta_{\hat{\eta}} = \frac{1}{2} \sum_{k=1}^2 \theta_{\hat{\eta}_k}, \quad \theta_{\hat{\eta}_k} = P f_{\hat{\eta}_k} = \int f_{\hat{\eta}_k}(w) dP(w).$$

For example, if $f_\eta(w) = (y - \eta(x))^2$, $\theta_{\hat{\eta}}$ is the out-of-sample mean squared error of a predictor that selects one of the two fitted models uniformly at random.

The three examples below fit into this setting and exhibit the nonregularity that motivates this paper.

Example 1 (Model comparison with cross-validation (Blum et al., 1999; Dudoit and van der Laan, 2005; Bayle et al., 2020; Austern and Zhou, 2025; Fava, 2025; Bayle et al., 2026)).

Let $W = (Y, X)$, with outcome Y and covariates X , and let $\eta = (\mu, \nu)$ be a pair of predictors of Y from X . The fitted model $\hat{\eta}_k = (\hat{\mu}_k, \hat{\nu}_k)$ then consists of two predictors trained on fold $\mathbf{s}_k$ by different learners, for example a random forest and a neural network. The score is the loss difference

$$f_\eta(y, x) = L(y, \mu(x)) - L(y, \nu(x)),$$

where L is a loss function, such as the squared error $L(y, \hat{y}) = (y - \hat{y})^2$ or the classification error $L(y, \hat{y}) = \mathbf{1}(y \neq \hat{y})$ for discrete outcomes.

Each

$$\theta_{\hat{\eta}_k} = \int [L(y, \hat{\mu}_k(x)) - L(y, \hat{\nu}_k(x))] dP(y, x)$$

is the difference in out-of-sample loss between the two fitted models $\hat{\mu}_k$ and $\hat{\nu}_k$, and the estimand

$$\theta_{\hat{\eta}} = \frac{1}{2} (\theta_{\hat{\eta}_1} + \theta_{\hat{\eta}_2})$$

is its average over the two folds, or, equivalently, the out-of-sample loss difference when $(\hat{\mu}_k, \hat{\nu}_k)$ is drawn uniformly at random from the two fitted pairs.

Example 2 (Value of a learned treatment rule (Luedtke and van der Laan, 2016; Shi et al., 2020; Whitehouse et al., 2026)).

Let $W = (Y, A, X)$, with outcome Y , binary treatment assignment A , and covariates X , in a randomized experiment with known treatment probability $\mathbb{P}(A = 1 \mid X) = \pi \in (0, 1)$. Moreover, let $Y(0), Y(1)$ denote potential outcomes with $Y = Y(A)$, and let $\eta : \mathcal{X} \rightarrow \{0, 1\}$ be a treatment rule, so that $\hat{\eta}_k$ is a rule learned on fold $\mathbf{s}_k$, for example by empirical welfare maximization (Kitagawa and Tetenov, 2018) or by thresholding an estimate of the conditional average treatment effect $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$.

The value of a rule is $V(\eta) = \mathbb{E}[Y(\eta(X))]$, and $V(\eta) = Pf_\eta$ for

$$f_\eta(y, a, x) = \frac{ay}{\pi} \eta(x) + \frac{(1-a)y}{1-\pi} \{1 - \eta(x)\}.$$

Each $\theta_{\hat{\eta}_k} = V(\hat{\eta}_k)$ is the value of the rule learned on fold $\mathbf{s}_k$, and the estimand $\theta_{\hat{\eta}}$ is the average value of the two learned rules.

The value of the learned treatment regimes is also informative for the value of the optimal treatment regime. The learned rules are dominated by the optimal rule $\eta^*(x) = \mathbf{1}(\tau(x) > 0)$, so $\theta_{\hat{\eta}} \leq V(\eta^*)$, and a lower confidence bound for $\theta_{\hat{\eta}}$ is also a lower confidence bound for the optimal value. Moreover, inference on $\theta_{\hat{\eta}}$ is asymptotically equivalent to inference on $V(\eta^*)$ whenever the regret $V(\eta^*) - \theta_{\hat{\eta}} = o_p(n^{-1/2})$, which Luedtke and van der Laan (2016) establish under margin and rate conditions, without requiring $\hat{\eta}_k$ to converge.

Example 3 (Best linear predictor (BLP) test of predictive power (Chernozhukov et al., 2025; Wager, 2025)).

Let $W = (Y, X)$, and let $\eta : \mathcal{X} \rightarrow \mathbb{R}$ be a predictor of Y from X . A natural way to assess the predictive power and calibration of $\hat{\eta}_k$ is to regress Y on $\hat{\eta}_k(X)$ and a constant in the test fold $\tilde{\mathbf{s}}_k$. A slope of zero means no predictive power, and a slope of one means good calibration.

Holding $\hat{\eta}_k$ fixed, the population slope is

$$\theta_{\hat{\eta}_k} = Pf_{\hat{\eta}_k}, \quad f_{\eta}(y, x) = \frac{(y - \mathbb{E}[Y]) \{\eta(x) - P\eta\}}{P(\eta - P\eta)^2}.$$

The ordinary least squares estimator differs from $\hat{\theta}_{\hat{\eta}_k}$ due to the oracle scale $P(\hat{\eta}_k - P\hat{\eta}_k)^2$ and the oracle centerings $\mathbb{E}[Y]$ and $P\eta$, but this distinction is typically first-order negligible since the sample variance of $\hat{\eta}_k(X)$ is ratio-consistent for $P(\hat{\eta}_k - P\hat{\eta}_k)^2$ under mild moment conditions. Under the null that $\mathbb{E}[Y | X]$ is constant, if $\hat{\eta}_k$ is consistent for this constant, $P(\hat{\eta}_k - P\hat{\eta}_k)^2 \xrightarrow{P} 0$, and the score diverges. The standardized score $(f_{\hat{\eta}_k} - Pf_{\hat{\eta}_k}) / \text{Var}[f_{\hat{\eta}_k}(W)]^{1/2}$ may then have no deterministic limit, which is the nonregularity my CLT addresses.

The same construction tests for treatment effect heterogeneity in a randomized controlled trial by replacing Y with its Horvitz–Thompson transform, which yields the best linear predictor of Chernozhukov et al. (2025, Theorem 3.2).

3 A CLT for Two-Fold Cross-Fitting

I establish a central limit theorem for two-fold cross-fitting that accommodates the examples of the previous section. First, I discuss the required assumptions of the CLT, including a novel locality condition. Then, I establish a qualitative CLT, followed by a quantitative result that establishes convergence to normality at the i.i.d. rate $n^{-1/2}$ in the Wasserstein metric under stronger conditions.

To simplify exposition, I focus on data-generating processes (DGPs) where

$$Pf_{\eta} = 0 \quad \text{and} \quad Pf_{\eta}^2 = 1$$

for every $\eta \in \mathcal{H}$. Alternatively, the general result follows from replacing each score $f_{\hat{\eta}_k}(w)$ by $(f_{\hat{\eta}_k}(w) - Pf_{\hat{\eta}_k}) / \sigma_{\hat{\eta}_k}$, with $\sigma_{\hat{\eta}_k}^2 = P(f_{\hat{\eta}_k} - Pf_{\hat{\eta}_k})^2$. While this normalization is infeasible, it is a natural target for a CLT, since the centering reflects the estimand of interest $\theta_{\hat{\eta}_k}$, and $\sigma_{\hat{\eta}_k}^2$ can often be estimated consistently from the test sample $\tilde{\mathbf{s}}_k$, which I address when deriving feasible standard errors. Importantly, this normalization allows the variance of the unstandardized score $\sigma_{\hat{\eta}_k}^2$ to converge to zero in probability, requiring only that it be almost surely positive for every n .

This section's main result is a new CLT establishing

$$T_n = \frac{\sqrt{n}\hat{\theta}_{\hat{\eta}}}{\sqrt{1 + \rho_n}} \rightsquigarrow \mathcal{N}(0, 1),$$

with

$$\rho_n = \text{Corr}(\hat{\theta}_{\hat{\eta}_1}, \hat{\theta}_{\hat{\eta}_2}) > -1.$$

I derive feasible standard errors in Section 4.

I collect in Assumption 1 the conditions I use for the CLT. I use the convention, without loss of generality, that $1, 2 \in \mathbf{s}_1$. The assumption uses three further definitions. First, define the mean-square stability coefficient

$$\gamma_n = \mathbb{E}\left[\left\{f_{\hat{\eta}_1}(W) - f_{\hat{\eta}'_1}(W)\right\}^2\right],$$

where $W \sim P$ independent of the data, $\mathcal{D}'_{\mathbf{s}_1} = \{W'_1\} \cup \{W_i : i \in \mathbf{s}_1 \setminus \{1\}\}$ is the training sample with W_1 replaced by an independent draw $W'_1 \sim P$, and $\hat{\eta}'_1 = \mathcal{A}(\mathcal{D}'_{\mathbf{s}_1})$ is the corresponding refitted model. γ_n is the mean-square stability coefficient studied by Kale et al. (2011) and Bayle et al. (2020), applied here to the centered and conditionally normalized score. Second, define the local covariance coefficients

$$r_i = \frac{n}{2} \text{Cov}\left[\sqrt{\frac{n}{2}}\hat{\theta}_{\hat{\eta}_1}, \sqrt{\frac{n}{2}}\hat{\theta}_{\hat{\eta}_2} \mid \mathcal{D}_{-i}\right], \quad i = 1, \dots, n,$$

where $\mathcal{D}_{-i} = \{W_j : j \neq i\}$ contains all observations except W_i . Each coefficient measures the covariance between the fold estimators when only observation i varies, isolating the interaction between using W_i to train $\hat{\eta}_1$ and to evaluate $\hat{\theta}_{\hat{\eta}_2}$, when $i \in \mathbf{s}_1$. Note that this normalization makes r_i a local analogue of the cross-fold correlation, since

$$\mathbb{E}[r_i] = \rho_n.$$

Moreover, for a random variable X , write

$$\|X\|_p = \begin{cases} \mathbb{E}[|X|^p]^{1/p}, & 1 \leq p < \infty, \\ \inf\{M \geq 0 : \mathbb{P}(|X| \leq M) = 1\}, & p = \infty. \end{cases}$$

Assumption 1. *For some $\bar{c} < \infty$, $\underline{c} > 0$, $p \in (2, \infty]$, with the convention $1/p = 0$ when $p = \infty$, and sufficiently large n , the following conditions hold:*

(i) (*Moments*).

$$\|f_{\hat{\eta}_1}(W)\|_p \leq \bar{c};$$

(ii) (*Non-degeneracy*).

$$\rho_n \geq -1 + \underline{c};$$

(iii) (*Mean-square stability*).

$$n^{\frac{1}{2(1-1/p)}} \gamma_n \rightarrow 0;$$

(iv) (*Locality*).

$$\text{Cov}[r_1, r_2] \rightarrow 0.$$

Lemma 1 (Sufficient Conditions). *When $\text{Var}[r_1] > 0$,*

$$|\text{Cov}[r_1, r_2]| \leq n\gamma_n |\text{Corr}(r_1, r_2)|.$$

In particular, Assumptions 1(iii) and 1(iv) hold under any of the following:

- (a) $n\gamma_n = O(1), \text{Corr}(r_1, r_2) = o(1);$
- (b) $p = \infty, \sqrt{n} \gamma_n = o(1), \sqrt{n} \text{Corr}(r_1, r_2) = O(1);$
- (c) $n\gamma_n = o(1).$

Assumption 1(i) is a uniform L^p moment bound for some $p > 2$, with $p = \infty$ for bounded scores. Assumption 1(ii) rules out near-perfect negative correlation between the fold-level estimators. It holds, for example, whenever $\rho_n \geq 0$, as in all simulations of Section 5. Assumption 1(iii) is a mean-square stability condition, restricting the sensitivity of the fitted score to replacing one training observation. Its exponent $1/[2(1 - 1/p)]$ decreases from 1 as $p \downarrow 2$ to $\frac{1}{2}$ at $p = \infty$, so that lighter tails permit a less stable learner. Related conditions have been used to establish asymptotic normality of cross-validation estimators of test error under the stronger requirement $n\gamma_n = o(1)$ (Bayle et al., 2020; Austern and Zhou, 2025; Lei, 2025), which implies $\rho_n \rightarrow 0$ (Remark 1). This condition can often be verified using algorithm-specific leave-one-out or replace-one stability bounds, available under suitable regularity and tuning conditions (for a discussion, see, e.g., Lei, 2025, Section 6).

Finally, Assumption 1(iv) is a locality condition. It requires the dependence created by using the same data point W_i to train model $\hat{\eta}_1$ and evaluate $f_{\hat{\eta}_2}(W_i)$ to be local across observations. Note that r_i measures the dependence between $\hat{\theta}_{\hat{\eta}_1}$ and $\hat{\theta}_{\hat{\eta}_2}$ as only observation W_i varies, isolating the interaction between using W_i to train $\hat{\eta}_1$ and to evaluate $\hat{\theta}_{\hat{\eta}_2}$, when $i \in \mathbf{s}_1$. Under the stability requirement $n\gamma_n = O(1)$, it suffices that $\text{Corr}(r_1, r_2) \rightarrow 0$ at any rate (Lemma 1(a)). This condition is plausible for nonparametric learners, including machine learning algorithms, because changing a training observation mostly affects predictions near its own covariate value. The coefficient r_i is then driven by the test observations whose covariates lie close to those of

W_i , so that r_1 and r_2 depend on nearly disjoint parts of the test fold and are approximately uncorrelated. In parametric models, however, all observations act through the same few parameters, so the contributions r_i can remain correlated, and asymptotic normality may fail. Note that $\mathbb{E}[r_i] = \rho_n$, and that Assumption 1(iv) allows ρ_n to be non-zero asymptotically.

Lemma 1 gives three sufficient conditions for Assumptions 1(iii) and 1(iv), trading off stability against locality. Since $\text{Var}[r_1] \leq n\gamma_n$, Assumption 1(iv) holds whenever $n\gamma_n \text{Corr}(r_1, r_2) \rightarrow 0$. Under condition (a), stability alone keeps $n\gamma_n$ bounded, so the correlation may vanish at any rate. Condition (b) relaxes stability by strengthening locality: once the correlation vanishes at rate $n^{-1/2}$, stability is needed only at the $\sqrt{n}$ rate for bounded scores, $\sqrt{n}\gamma_n \rightarrow 0$. Condition (c) is the stability requirement of previous cross-validation CLTs, and it implies Assumption 1(iv) with no restriction on the correlation. Hence, my CLT nests the regime $n\gamma_n \rightarrow 0$, in which $\rho_n \rightarrow 0$.

Theorem 1 (Cross-Fitting Central Limit Theorem). *Let Assumption 1 hold. Then,*

$$T_n \rightsquigarrow \mathcal{N}(0, 1).$$

Theorem 1 is, to my knowledge, the first to establish a central limit theorem for cross-fitting estimators that accommodates the nonregularity that the standardized influence functions may not converge to a fixed limit. Compared to previous CLTs in the cross-validation (Dudoit and van der Laan, 2005; Bayle et al., 2020; Austern and Zhou, 2025; Lei, 2025) and semiparametric estimation/double machine learning (Schick, 1986; Chernozhukov et al., 2018) literatures, the asymptotic variance of $\sqrt{n}\hat{\theta}_{\hat{\eta}}$ in my result depends on the cross-fold correlation, which may be nonzero. Note that the stronger stability condition $n\gamma_n = o(1)$, required by previous cross-validation CLTs, implies $\rho_n = o(1)$. Compared to those results, I weaken the stability requirement (Assumption 1(iii)), and introduce the locality Assumption 1(iv).

The next result strengthens Theorem 1. Under stronger conditions, the normal approximation holds at the rate $n^{-1/2}$, the Berry–Esseen rate for averages of independent random variables with finite third moments. I state the result in the Wasserstein distance, defined for random variables X and Y by

$$d_W(X, Y) = \sup_{h \in \text{Lip}(1)} \left| \mathbb{E}[h(X)] - \mathbb{E}[h(Y)] \right|,$$

where

$$\text{Lip}(1) = \{h: \mathbb{R} \rightarrow \mathbb{R} : |h(x) - h(y)| \leq |x - y| \text{ for all } x, y \in \mathbb{R}\}.$$

Theorem 2 (Quantitative Cross-Fitting Central Limit Theorem). *Let Assumption 1 hold with $p \geq 3$. Suppose, in addition, that*

$$n^{\frac{1}{1-1/p}}\gamma_n = O(1), \quad n \text{Cov}[r_1, r_2] = O(1).$$

Then,

$$d_W(T_n, \mathcal{N}(0, 1)) = O(n^{-1/2}).$$

Theorem 2 shows that the dependence between folds does not slow the normal approximation as long as the locality condition holds at a fast enough rate. Its conditions trade off tails against stability. The moment condition requires a uniformly bounded L^p norm of the score with $p \geq 3$, as in the Berry–Esseen theorem. The stability condition requires $\gamma_n = O(n^{-1/(1-1/p)})$, which ranges from $\gamma_n = O(n^{-3/2})$ at $p = 3$ to $\gamma_n = O(n^{-1})$ for bounded scores, so that heavier tails require a more stable learner. For bounded scores, the stability condition has the same order as the sufficient condition $n\gamma_n = O(1)$ in Lemma 1(a), and the $O(n^{-1/2})$ rate then costs only a strengthening of locality, from $\text{Cov}[r_1, r_2] = o(1)$ to $O(n^{-1})$. Since $\text{Cov}[r_1, r_2] = \text{Var}[r_1] \text{Corr}(r_1, r_2)$ and $\text{Var}[r_1] \leq n\gamma_n$ (Lemma B.1), the latter holds whenever $\text{Corr}(r_1, r_2) = O(n^{-1})$.

Remark 1 (Comparison with previous stability conditions). *Previous stability-based CLTs for cross-validation require first-order loss stability at the $o(n^{-1})$ scale relative to the variance of the leading score (Bayle et al., 2020; Austern and Zhou, 2025; Lei, 2025). Applied to the centered and conditionally normalized score used here, this condition is equivalent to*

$$n\gamma_n \rightarrow 0.$$

Indeed, if $\tilde{\sigma}_n^2 = \text{Var}[\mathbb{E}[f_{\hat{\eta}_1}(W) \mid W]]$, their condition is $n\gamma_n/\tilde{\sigma}_n^2 \rightarrow 0$, while

$$0 \leq 1 - \tilde{\sigma}_n^2 \leq \frac{n}{4}\gamma_n.$$

Hence, previous results cover the regime $n\gamma_n = o(1)$, which also forces $\rho_n \rightarrow 0$. By contrast, Theorem 1 allows weaker stability, for example $n\gamma_n = O(1)$ when the locality condition makes the covariance contributions asymptotically uncorrelated.

4 Estimation of ρ and Inference

I propose a resampling-based procedure to construct feasible confidence intervals based on Theorem 1, and show that they attain nominal coverage asymptotically under regularity conditions.

Let $n = 4m$ be the total sample size, and return to the unnormalized scores of Section 2, that is, $f_{\hat{\eta}_k}(w)$ may have non-zero mean and non-unit variance. Throughout this section, ρ_n denotes the cross-fold correlation of the centered and normalized fold estimators,

$$\rho_n = \text{Corr}\left(\frac{\hat{\theta}_{\hat{\eta}_1} - \theta_{\hat{\eta}_1}}{\sigma_{\hat{\eta}_1}}, \frac{\hat{\theta}_{\hat{\eta}_2} - \theta_{\hat{\eta}_2}}{\sigma_{\hat{\eta}_2}}\right),$$

which is the correlation of Section 3 computed with each score $f_{\hat{\eta}_k}$ replaced by $(f_{\hat{\eta}_k} - \theta_{\hat{\eta}_k})/\sigma_{\hat{\eta}_k}$. I consider directly observable scores, which fit Examples 1 and 2, and give an implementation for the BLP Example 3 in Appendix C. Keep the original folds fixed. For $b = 1, \dots, B$ and $k = 1, 2$, independently draw a uniform subset $\mathbf{s}_{k,b} \subset \mathbf{s}_k$ of size m and fit

$$\hat{\eta}_{k,b} = \mathcal{A}(\mathcal{D}_{\mathbf{s}_{k,b}}).$$

Define

$$\hat{\sigma}_{\hat{\eta}_{k,b}}^2 = \frac{1}{2m-1} \sum_{i \in \mathbf{s}_k} \left[f_{\hat{\eta}_{k,b}}(W_i) - \frac{1}{2m} \sum_{j \in \mathbf{s}_k} f_{\hat{\eta}_{k,b}}(W_j) \right]^2, \quad (3)$$

$$Z_{k,b} = \frac{\sum_{i \in \mathbf{s}_{3-k,b}} f_{\hat{\eta}_{k,b}}(W_i) - \sum_{i \in \mathbf{s}_k \setminus \mathbf{s}_{3-k,b}} f_{\hat{\eta}_{k,b}}(W_i)}{\sqrt{m} \hat{\sigma}_{\hat{\eta}_{k,b}}}, \quad (4)$$

$$\hat{\rho}_n = \max \left\{ -1, \min \left\{ 1, \frac{1}{B} \sum_{b=1}^B Z_{1,b} Z_{2,b} \right\} \right\}. \quad (5)$$

Set $Z_{k,b} = 0$ if its denominator is zero. Note that $Z_{k,b}$ evaluates the same fitted score on two disjoint samples, so its expected value is zero.

Let $\hat{\sigma}_{\hat{\eta}_k}^2$ be the test-fold sample variance

$$\hat{\sigma}_{\hat{\eta}_k}^2 = \frac{1}{2m-1} \sum_{i \in \mathbf{s}_k} [f_{\hat{\eta}_k}(W_i) - \hat{\theta}_{\hat{\eta}_k}]^2, \quad k = 1, 2. \quad (6)$$

The proposed standard error and confidence interval are

$$\widehat{\text{se}}(\hat{\theta}_{\hat{\eta}}) = \sqrt{\frac{\hat{\sigma}_{\hat{\eta}_1}^2 + \hat{\sigma}_{\hat{\eta}_2}^2 + 2\hat{\rho}_n \hat{\sigma}_{\hat{\eta}_1} \hat{\sigma}_{\hat{\eta}_2}}{2n}}, \quad (7)$$

$$\widehat{\text{CI}}_{\alpha} = [\hat{\theta}_{\hat{\eta}} \pm z_{1-\alpha/2} \widehat{\text{se}}(\hat{\theta}_{\hat{\eta}})], \quad (8)$$

where $z_{1-\alpha/2}$ is the standard normal quantile. The original estimator and its target are unchanged.

Theorem 3 (Asymptotic Validity of Confidence Interval). *Let Assumption 1 hold after centering and normalizing the scores. Suppose $0 < \sigma_{\eta}^2 < \infty$ for every $\eta \in \mathcal{H}$. If, for some $\rho > -1$,*

$$\text{Var}[\mathbb{E}[Z_{1,1} Z_{2,1} \mid \mathcal{D}]] \rightarrow 0, \quad \rho_n \rightarrow \rho, \quad B \rightarrow \infty, \quad (9)$$

then $\hat{\rho}_n - \rho_n \xrightarrow{L^2} 0$. If, additionally,

$$\frac{\sigma_{\hat{\eta}_1}}{\sigma_{\hat{\eta}_2}} \xrightarrow{p} 1, \quad (10)$$

then, for every fixed $\alpha \in (0, 1)$,

$$\mathbb{P}(\theta_{\hat{\eta}} \in \widehat{\text{CI}}_{\alpha}) \rightarrow 1 - \alpha.$$

The first condition in (9) requires the covariance averaged over resampling splits to stabilize across datasets. The second requires the cross-fold correlation to converge to some limit greater than -1 .

5 Simulation Study

I examine the finite-sample performance of the confidence intervals derived in Section 4 in a simulation study. In particular, I consider the cross-validation (CV) model comparison (Example 1) and the BLP test of predictive power (Example 3). I use sample sizes $n \in \{1000, 2000\}$ and 2,400 Monte Carlo replications for each data-generating process and sample size. The correlation estimator uses $B = 500$ repetitions.

Observations are independent and identically distributed, with $X_i \sim \mathcal{N}(0, I_5)$ and $U_i \sim \mathcal{N}(0, 1)$ independent of X_i . I consider two no-signal designs. In the homoskedastic Gaussian design,

$$Y_i = U_i.$$

In the heteroskedastic, conditionally skewed design,

$$Y_i = \omega(X_i) \frac{\exp(\lambda(X_i)U_i) - \exp(\lambda(X_i)^2/2)}{\sqrt{\exp(2\lambda(X_i)^2) - \exp(\lambda(X_i)^2)}},$$

where, writing $\Lambda(t) = \{1 + \exp(-t)\}^{-1}$,

$$\omega(x) = 0.5 + 1.5\Lambda(x_1 + 0.5x_2), \quad \lambda(x) = 0.25 + 0.75\Lambda(x_2 - x_3).$$

Both designs satisfy $\mathbb{E}[Y_i | X_i] = 0$. The conditional variance is one in the Gaussian design and $\omega(X_i)^2$ in the skewed design. The latter also permits conditional skewness to vary with the covariates.

I estimate the predictive model with two learners. The first is a single-hidden-layer neural network from the R package **nnet**, with 10 hidden units, a linear output, no skip connections, and weight decay equal to 0.1. The second is a random forest from the R package **ranger**, with 200 trees, three candidate covariates per split, and a minimum node size of five.

For CV, I compare the squared-error loss of the fitted learner with that of the training fold outcome mean. In the notation of Example 1, $L(y, \hat{y}) = (y - \hat{y})^2$, $\hat{\mu}_k$ is

Table 1: Coverage rates of nominal 95% BLP and CV confidence intervals.

		BLP		CV	
Design	n	Estimated $\hat{\rho}_n$	$\hat{\rho}_n = 0$	Estimated $\hat{\rho}_n$	$\hat{\rho}_n = 0$
Neural network					
Gaussian	1000	94.38	90.42	94.21	91.67
	2000	95.04	91.13	94.75	91.54
Skewed	1000	94.33	91.54	94.38	92.75
	2000	95.13	92.38	94.42	92.13
Random forest					
Gaussian	1000	94.13	84.50	94.54	85.50
	2000	95.67	86.88	95.54	87.54
Skewed	1000	94.75	87.96	95.29	89.00
	2000	95.33	89.33	95.79	90.38

Notes: Entries are coverage percentages of $\theta_{\hat{\eta}}$, based on 2,400 replications per design and sample size. Gaussian denotes the homoskedastic Gaussian design, Skewed denotes the heteroskedastic, conditionally skewed design. Estimated $\hat{\rho}_n$ uses (5) with $B = 500$; $\hat{\rho}_n = 0$ ignores the cross-fold correlation.

the learner trained on $\mathbf{s}_k$, and $\hat{\nu}_k$ is the constant predictor equal to the average of Y_i over $i \in \mathbf{s}_k$.

For both applications, I report the coverage rate $\mathbb{P}(\theta_{\hat{\eta}} \in \widehat{\text{CI}}_\alpha)$ at the nominal 95% level. Since $\mathbb{E}[Y_i | X_i] = 0$ in both designs, the BLP estimand is $\theta_{\hat{\eta}} = 0$. For CV, the estimand of Example 1 is

$$\theta_{\hat{\eta}} = \frac{1}{2} \sum_{k=1}^2 \int [\{y - \hat{\mu}_k(x)\}^2 - \{y - \hat{\nu}_k\}^2] dP(y, x),$$

the average out-of-sample loss difference, holding the fitted models fixed. This target need not be zero under a no-signal design. I compute it by integrating analytically over the outcome noise and numerically over 65,000 independent covariate draws per fitted model.

Table 1 reports the results. With the estimated correlation, coverage rates range from 94.13% to 95.67% for BLP and from 94.21% to 95.79% for CV, close to the nominal 95% level across the learners, designs, and sample sizes considered. Setting $\hat{\rho}_n = 0$, which ignores the cross-fold correlation, yields undercoverage in every case, with coverage rates from 84.50% to 92.38% for BLP and from 85.50% to 92.75% for CV. The undercoverage is most severe for the random forest, whose coverage falls below 90% in the Gaussian design for both applications and sample sizes.

6 Conclusion

I revisited the problem of inference with cross-fitting when the standardized fitted score need not converge to a deterministic limit, a nonregularity shared by cross-validation model comparison, tests of predictive power and treatment effect heterogeneity, and the value of learned treatment rules. Exploiting a new locality condition, I established a central limit theorem for cross-fitting estimators with an asymptotic variance that depends on the cross-fold correlation, and I proposed confidence intervals based on a resampling estimator of this correlation that attain asymptotically nominal coverage. In simulations with random forests and neural networks, these intervals delivered coverage close to the nominal level, whereas intervals that ignore the cross-fold correlation undercovered.

References

- AUSTERN, M. AND W. ZHOU (2025): “Asymptotics of Cross-Validation,” *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, 61, 2804–2865.
- BAYLE, A., L. JANSON, AND L. MACKEY (2026): “The Relative Instability of Model Comparison with Cross-Validation,” arXiv:2508.04409.
- BAYLE, P., A. BAYLE, L. JANSON, AND L. MACKEY (2020): “Cross-Validation Confidence Intervals for Test Error,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol. 33, 16339–16350.
- BLUM, A., A. KALAI, AND J. LANGFORD (1999): “Beating the Hold-out: Bounds for K-Fold and Progressive Cross-Validation,” in *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, Santa Cruz, CA: ACM, 203–208.
- CHATTERJEE, S. (2008): “A New Method of Normal Approximation,” *The Annals of Probability*, 36, 1584–1610.
- 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 (2025): “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.
- DAI, B., X. SHEN, AND W. PAN (2024): “Significance Tests of Feature Relevance for a Black-Box Learner,” *IEEE Transactions on Neural Networks and Learning Systems*, 35, 1898–1911.
- DiCICCIO, C. J., T. J. DiCICCIO, AND J. P. ROMANO (2020): “Exact Tests via Multiple Data Splitting,” *Statistics & Probability Letters*, 166, 108865.
- 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.
- EFRON, B. (1983): “Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation,” *Journal of the American Statistical Association*, 78, 316–331.

- EFRON, B. AND R. TIBSHIRANI (1997): “Improvements on Cross-Validation: The 632+ Bootstrap Method,” *Journal of the American Statistical Association*, 92, 548–560.
- FAVA, B. (2025): “Training and Testing with Multiple Splits: A Central Limit Theorem for Split-Sample Estimators,” arXiv:2511.04957.
- GASPARIN, M. AND A. RAMDAS (2024): “Merging Uncertainty Sets via Majority Vote,” arXiv:2401.09379.
- GASPARIN, M., R. WANG, AND A. RAMDAS (2025): “Combining Exchangeable p -Values,” *Proceedings of the National Academy of Sciences*, 122, e2410849122.
- GOLDSTEIN, B. A., E. C. POLLEY, F. B. S. BRIGGS, M. J. VAN DER LAAN, AND A. HUBBARD (2016): “Testing the Relative Performance of Data Adaptive Prediction Algorithms: A Generalized Test of Conditional Risk Differences,” *The International Journal of Biostatistics*, 12, 117–129.
- HUBBARD, A. E., S. KHERAD-PAJOUH, AND M. J. VAN DER LAAN (2016): “Statistical Inference for Data Adaptive Target Parameters,” *The International Journal of Biostatistics*, 12, 3–19.
- JIANG, W., S. VARMA, AND R. SIMON (2008): “Calculating Confidence Intervals for Prediction Error in Microarray Classification Using Resampling,” *Statistical Applications in Genetics and Molecular Biology*, 7, Article 8.
- KALE, S., R. KUMAR, AND S. VASSILVITSKII (2011): “Cross-Validation and Mean-Square Stability,” in *Innovations in Computer Science – ICS 2011*, Beijing, China: Tsinghua University Press, 487–495.
- KITAGAWA, T. AND A. TETENOV (2018): “Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice,” *Econometrica*, 86, 591–616.
- LABER, E. B. AND S. A. MURPHY (2011): “Adaptive Confidence Intervals for the Test Error in Classification,” *Journal of the American Statistical Association*, 106, 904–913.
- LEI, J. (2020): “Cross-Validation With Confidence,” *Journal of the American Statistical Association*, 115, 1978–1997.
- (2025): “A Modern Theory of Cross-Validation through the Lens of Stability,” arXiv:2505.23592.

- LUEDTKE, A. R. AND M. J. VAN DER LAAN (2016): “Statistical Inference for the Mean Outcome under a Possibly Non-Unique Optimal Treatment Strategy,” *The Annals of Statistics*, 44, 713–742.
- LUO, Y. AND R. F. BARBER (2025): “The Limits of Assumption-Free Tests for Algorithm Performance,” arXiv:2402.07388.
- NEWAY, W. K. AND J. R. ROBINS (2018): “Cross-Fitting and Fast Remainder Rates for Semiparametric Estimation,” arXiv:1801.09138.
- PICARD, R. R. AND R. D. COOK (1984): “Cross-Validation of Regression Models,” *Journal of the American Statistical Association*, 79, 575–583.
- PIRRACCHIO, R., M. L. PETERSEN, M. CARONE, M. RESCHE-RIGON, S. CHEVRET, AND M. J. VAN DER LAAN (2015): “Mortality Prediction in Intensive Care Units with the Super ICU Learner Algorithm (SICULA): A Population-Based Study,” *The Lancet Respiratory Medicine*, 3, 42–52.
- RINALDO, A., L. WASSERMAN, AND M. G’SSELL (2019): “Bootstrapping and Sample Splitting for High-Dimensional, Assumption-Lean Inference,” *The Annals of Statistics*, 47, 3438–3469.
- ROBINS, J. M. (2004): “Optimal Structural Nested Models for Optimal Sequential Decisions,” in *Proceedings of the Second Seattle Symposium in Biostatistics*, ed. by P. Bickel, P. Diggle, S. Fienberg, U. Gather, I. Olkin, S. Zeger, D. Y. Lin, and P. J. Heagerty, New York, NY: Springer, vol. 179 of *Lecture Notes in Statistics*, 189–326.
- SCHICK, A. (1986): “On Asymptotically Efficient Estimation in Semiparametric Models,” *The Annals of Statistics*, 14, 1139–1151.
- 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.
- STEIN, C. (1972): “A Bound for the Error in the Normal Approximation to the Distribution of a Sum of Dependent Random Variables,” in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, ed. by L. M. Le Cam, J. Neyman, and E. L. Scott, University of California Press, 583–602.
- (1986): *Approximate Computation of Expectations*, IMS.
- STONE, M. (1977): “Asymptotics for and against Cross-Validation,” *Biometrika*, 64, 29–35.

- VAN DE WIEL, M. A., J. BERKHOF, AND W. N. VAN WIERINGEN (2009): “Testing the Prediction Error Difference between 2 Predictors,” *Biostatistics*, 10, 550–560.
- VERDINELLI, I. AND L. WASSERMAN (2024): “Decorrelated Variable Importance,” *Journal of Machine Learning Research*, 25, 1–27.
- WAGER, S. (2025): “A Comment on: ‘Fisher–Schultz Lecture: Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, With an Application to Immunization in India’ by Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val,” *Econometrica*, 93, 1171–1176.
- WHITEHOUSE, J., Q. CHEN, M. AUSTERN, AND V. SYRGKANIS (2026): “Inference on Optimal Policy Values and Other Irregular Functionals via Softmax Smoothing,” arXiv:2507.11780.
- YADLOWSKY, S., S. FLEMING, N. SHAH, E. BRUNSKILL, AND S. WAGER (2025): “Evaluating Treatment Prioritization Rules via Rank-Weighted Average Treatment Effects,” *Journal of the American Statistical Association*, 120, 38–51.
- YANG, Y. (2006): “Comparing Learning Methods for Classification,” *Statistica Sinica*, 16, 635–657.

A Extension to K_n -Fold Cross-Fitting

I extend Theorem 1 to a possibly growing number of folds, including leave-one-out cross-fitting.

Let $K_n \in \{2, \dots, n\}$ divide n , let $m_n = n/K_n$, and partition $[n]$ into equal-sized folds

$$\mathbf{s}_1, \dots, \mathbf{s}_{K_n}.$$

For $k = 1, \dots, K_n$, let

$$\hat{\eta}_{-k} = \mathcal{A}(\mathcal{D}_{[n] \setminus \mathbf{s}_k})$$

denote the model trained on all observations outside fold $\mathbf{s}_k$. In this section, $k(i)$ denotes the fold containing observation i , so that W_i is evaluated by $\hat{\eta}_{-k(i)}$. The cross-fitting estimator and its estimand are

$$\hat{\theta}_{\hat{\eta}} = \frac{1}{K_n} \sum_{k=1}^{K_n} \hat{\theta}_{\hat{\eta}_{-k}}, \quad \hat{\theta}_{\hat{\eta}_{-k}} = \frac{1}{m_n} \sum_{i \in \mathbf{s}_k} f_{\hat{\eta}_{-k}}(W_i), \quad \theta_{\hat{\eta}} = \frac{1}{K_n} \sum_{k=1}^{K_n} P f_{\hat{\eta}_{-k}},$$

which reduce to those of Section 2 when $K_n = 2$.

Unlike in the two-fold result, I use a deterministic unconditional normalization. Let

$$\sigma_n^2 = \mathbb{E} \left[P \left\{ f_{\hat{\eta}_{-1}} - P f_{\hat{\eta}_{-1}} \right\}^2 \right] \in (0, \infty), \quad (\text{A.1})$$

and, for every $\eta \in \mathcal{H}$, define the standardized score

$$\tilde{f}_{\eta}(w) = \frac{f_{\eta}(w) - P f_{\eta}}{\sigma_n}.$$

Thus $P \tilde{f}_{\eta} = 0$ for every $\eta \in \mathcal{H}$, and, by fold symmetry, for $k = 1, \dots, K_n$,

$$P \tilde{f}_{\hat{\eta}_{-k}} = 0, \quad \mathbb{E} \left[P \tilde{f}_{\hat{\eta}_{-k}}^2 \right] = 1. \quad (\text{A.2})$$

Define the conditional variances of the standardized scores and their average across folds,

$$v_{k,n} = P \tilde{f}_{\hat{\eta}_{-k}}^2, \quad \bar{v}_n = \frac{1}{K_n} \sum_{k=1}^{K_n} v_{k,n}. \quad (\text{A.3})$$

Hence $\mathbb{E}[v_{k,n}] = \mathbb{E}[\bar{v}_n] = 1$, but $v_{k,n}$ need not equal one for a realized fit.

For each fold, let

$$S_{k,n} = \frac{1}{\sqrt{m_n}} \sum_{i \in \mathbf{s}_k} \tilde{f}_{\hat{\eta}_{-k}}(W_i). \quad (\text{A.4})$$

Conditional on $\hat{\eta}_{-k}$, the summands are i.i.d. with mean zero and variance $v_{k,n}$, so $\mathbb{E}[S_{k,n}] = 0$ and $\text{Var}[S_{k,n}] = \mathbb{E}[v_{k,n}] = 1$. By fold symmetry,

$$\rho_n = \text{Cov}[S_{k,n}, S_{\ell,n}], \quad k \neq \ell,$$

does not depend on k and ℓ . Since $\sqrt{n}(\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}})/\sigma_n = K_n^{-1/2} \sum_{k=1}^{K_n} S_{k,n}$,

$$\text{Var}\left[\frac{\sqrt{n}(\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}})}{\sigma_n}\right] = V_n, \quad V_n = 1 + (K_n - 1)\rho_n. \quad (\text{A.5})$$

Define

$$T_n = \frac{\sqrt{n}(\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}})}{\sigma_n \sqrt{V_n}}. \quad (\text{A.6})$$

Without loss of generality, label the observations so that

$$\mathbf{s}_k = \{(k-1)m_n + 1, \dots, km_n\}$$

for $k = 1, \dots, K_n$. In particular, $1 \in \mathbf{s}_1$, $2 \in \mathbf{s}_1$ when $m_n \geq 2$, and $m_n + 1 \in \mathbf{s}_2$. Let $W'_1 \sim P$ be independent of the data, and let $\hat{\eta}'_{-2}$ denote the model obtained after replacing W_1 by W'_1 in the training sample of $\hat{\eta}_{-2}$. Define the mean-square stability coefficient

$$\gamma_n = \mathbb{E}\left[\left\{\tilde{f}_{\hat{\eta}_{-2}}(W) - \tilde{f}_{\hat{\eta}'_{-2}}(W)\right\}^2\right], \quad (\text{A.7})$$

where $W \sim P$ is independent of all observed and replacement data.

For $i = 1, \dots, n$, define

$$r_i = \sqrt{m_n} \text{Cov}\left[\tilde{f}_{\hat{\eta}_{-k(i)}}(W_i), \frac{1}{K_n - 1} \sum_{\ell \neq k(i)} S_{\ell,n} \middle| \mathcal{D}_{-i}\right]. \quad (\text{A.8})$$

This is the K_n -fold analogue of the local covariance coefficient of Section 3, and

$$\mathbb{E}[r_i] = \rho_n. \quad (\text{A.9})$$

Indeed, since $\hat{\eta}_{-k(i)}$ is $\mathcal{D}_{-i}$ -measurable, $\mathbb{E}[\tilde{f}_{\hat{\eta}_{-k(i)}}(W_i) \mid \mathcal{D}_{-i}] = 0$, so

$$\mathbb{E}[r_i] = \sqrt{m_n}(K_n - 1)^{-1} \sum_{\ell \neq k(i)} \mathbb{E}[\tilde{f}_{\hat{\eta}_{-k(i)}}(W_i) S_{\ell,n}],$$

and exchangeability within $\mathbf{s}_{k(i)}$ gives

$$\sqrt{m_n} \mathbb{E}[\tilde{f}_{\hat{\eta}_{-k(i)}}(W_i) S_{\ell,n}] = \text{Cov}[S_{k(i),n}, S_{\ell,n}] = \rho_n.$$

The approximate Stein kernel is

$$\mathcal{K}_n = \frac{\bar{v}_n + (K_n - 1)\bar{r}_n}{V_n}, \quad \bar{r}_n = \frac{1}{n} \sum_{i=1}^n r_i, \quad (\text{A.10})$$

and $\mathbb{E}[\mathcal{K}_n] = 1$ by (A.3) and (A.9).

Theorem A.1 (K_n -Fold Cross-Fitting Central Limit Theorem). *Suppose that there exist $\bar{c} < \infty$, $\underline{c} > 0$, and $p \in (2, \infty]$, with $1/p = 0$ when $p = \infty$, such that the first two conditions below hold for all sufficiently large n and the last three hold as $n \rightarrow \infty$:*

$$\|\tilde{f}_{\hat{\eta}_{-1}}(W)\|_p \leq \bar{c}, \quad (\text{A.11})$$

$$V_n \geq \underline{c}, \quad (\text{A.12})$$

$$\frac{\text{Var}[\bar{v}_n]}{V_n^2} \rightarrow 0, \quad (\text{A.13})$$

$$n^{\frac{1}{2(1-1/p)}} (K_n - 1) \gamma_n \rightarrow 0, \quad (\text{A.14})$$

$$\frac{(K_n - 1)^2 \text{Var}[\bar{r}_n]}{V_n^2} \rightarrow 0. \quad (\text{A.15})$$

Then

$$T_n \rightsquigarrow \mathcal{N}(0, 1).$$

(A.13) is an additional condition required by the unconditional normalization. It requires the average conditional variance $\bar{v}_n$ to concentrate around its mean. By fold symmetry, $\text{Var}[\bar{v}_n] = \text{Var}[v_{1,n}]/K_n + (1 - 1/K_n) \text{Cov}[v_{1,n}, v_{2,n}]$, so, when $K_n \rightarrow \infty$, the condition can hold even if each $v_{k,n}$ remains random, provided $\text{Var}[v_{1,n}]/(K_n V_n^2) \rightarrow 0$ and $\text{Cov}[v_{1,n}, v_{2,n}]/V_n^2 \rightarrow 0$. Under the conditional normalization used in Section 3, $v_{k,n} = 1$ almost surely, so the condition holds immediately. (A.14) is the analogue of Assumption 1(iii), with the same exponent. The factor $K_n - 1$ reflects that one observation enters the training samples of $K_n - 1$ fitted models. For fixed K_n , the condition has the same order as Assumption 1(iii). (A.15) ensures that the local covariance term of (A.10) has vanishing variance. By fold symmetry and exchangeability within folds,

$$\text{Var}[\bar{r}_n] = \frac{\text{Var}[r_1]}{n} + \frac{m_n - 1}{n} \text{Cov}[r_1, r_2] + \left(1 - \frac{1}{K_n}\right) \text{Cov}[r_1, r_{m_n+1}],$$

where the second term vanishes when $m_n = 1$. The three terms control the diagonal contribution, the dependence between distinct observations in the same fold, and the dependence between observations in different folds, respectively.

When $K_n = 2$ and the scores satisfy the normalization of Section 3, $\sigma_n = 1$, $\tilde{f}_\eta = f_\eta$, and $v_{k,n} = 1$ almost surely, so (A.13) holds trivially. Moreover, ρ_n , γ_n , and

$r_i, i \in \mathbf{s}_1$, coincide with those of Section 3, with $\hat{\eta}_{-1}$ and $\hat{\eta}_{-2}$ in the roles of $\hat{\eta}_2$ and $\hat{\eta}_1$, so (A.11), (A.12) and (A.14) reduce to Assumptions 1(i) to 1(iii). For (A.15), $\bar{r}_n$ is the average of the fold means $R_k = m_n^{-1} \sum_{i \in \mathbf{s}_k} r_i, k = 1, 2$, which have equal variances by fold symmetry. Minkowski's inequality and Lemma B.1 then give

$$\text{Var}[\bar{r}_n] \leq \text{Var}[R_1] = \frac{2}{n} \text{Var}[r_1] + \left(1 - \frac{2}{n}\right) \text{Cov}[r_1, r_2] \leq \frac{\gamma_n}{2} + |\text{Cov}[r_1, r_2]|,$$

so Assumptions 1(ii) to 1(iv) imply (A.15). Thus Theorem A.1 contains Theorem 1 as the special case $K_n = 2$.

Proof. For $i = 1, \dots, n$, write $f_i = \tilde{f}_{\hat{\eta}_{-k(i)}}(W_i)$, so that $T_n = (nV_n)^{-1/2} \sum_{i=1}^n f_i$. Since $\hat{\eta}_{-k(i)}$ is $\mathcal{D}_{-i}$ -measurable, (A.2) and (A.3) give

$$\mathbb{E}[f_i \mid \mathcal{D}_{-i}] = 0, \quad \mathbb{E}[f_i^2 \mid \mathcal{D}_{-i}] = v_{k(i),n}. \quad (\text{A.16})$$

When $p < \infty$, the conditional Jensen inequality gives $v_{k,n}^{p/2} \leq \mathbb{E}[|\tilde{f}_{\hat{\eta}_{-k}}(W)|^p \mid \hat{\eta}_{-k}]$, and when $p = \infty$, $v_{k,n} \leq \bar{c}^2$ directly. Hence (A.11) and fold symmetry give, for all i and k ,

$$\|f_i\|_p \leq \bar{c}, \quad \|v_{k,n}^{1/2}\|_p \leq \bar{c}. \quad (\text{A.17})$$

For $i = 1, \dots, n$, let $\Delta_i = T_n - \mathbb{E}[T_n \mid \mathcal{D}_{-i}]$ and, for $\ell \neq k(i)$, let $A_{i,\ell} = S_{\ell,n} - \mathbb{E}[S_{\ell,n} \mid \mathcal{D}_{-i}]$. Let $q = 1/(1 - 1/p) \in [1, 2)$ denote the conjugate exponent of p .

Step 1. Increment bound. Conditional on $\mathcal{D}_{-i}$, the scores $f_j, j \in \mathbf{s}_{k(i)} \setminus \{i\}$, are fixed, while f_i and the models $\hat{\eta}_{-\ell}, \ell \neq k(i)$, vary. Hence, by (A.16),

$$\Delta_i = \frac{f_i}{\sqrt{nV_n}} + \frac{1}{\sqrt{K_n V_n}} \sum_{\ell \neq k(i)} A_{i,\ell}. \quad (\text{A.18})$$

Let $\hat{\eta}'_{-\ell}$ denote the model obtained after replacing W_i by an independent copy in the training sample of $\hat{\eta}_{-\ell}$, and let $S'_{\ell,n}$ denote $S_{\ell,n}$ computed with $\hat{\eta}'_{-\ell}$ in place of $\hat{\eta}_{-\ell}$. Conditional on $\mathcal{D}_{-i}$, $S_{\ell,n}$ and $S'_{\ell,n}$ are independent and identically distributed, so $\mathbb{E}[A_{i,\ell}^2] = \frac{1}{2} \mathbb{E}[(S_{\ell,n} - S'_{\ell,n})^2]$. The observations in $\mathbf{s}_\ell$ are independent of $(\hat{\eta}_{-\ell}, \hat{\eta}'_{-\ell})$, so, conditional on these models, the differences $\tilde{f}_{\hat{\eta}_{-\ell}}(W_j) - \tilde{f}_{\hat{\eta}'_{-\ell}}(W_j), j \in \mathbf{s}_\ell$, are i.i.d. with mean zero, since $P\tilde{f}_\eta = 0$ for every $\eta \in \mathcal{H}$. Hence the cross terms of the square vanish, and fold symmetry and exchangeability within folds give $\mathbb{E}[(S_{\ell,n} - S'_{\ell,n})^2] = \gamma_n$, so $\mathbb{E}[A_{i,\ell}^2] = \gamma_n/2$. By (A.18), the Cauchy-Schwarz inequality, $\mathbb{E}[f_i^2] = 1, (K_n - 1)^2/K_n \leq K_n - 1$, and (A.12),

$$\mathbb{E}[\Delta_i^2] \leq \frac{2}{nV_n} + \frac{(K_n - 1)^2}{K_n V_n} \gamma_n \leq \frac{2}{\underline{c}} \left\{ \frac{1}{n} + (K_n - 1) \gamma_n \right\}. \quad (\text{A.19})$$

Step 2. Kernel concentration. For $i = 1, \dots, n$, let

$$c_i = \text{Cov} \left[f_i, \sum_{j=1}^n f_j \middle| \mathcal{D}_{-i} \right] = \sqrt{nV_n} \mathbb{E}[f_i \Delta_i \mid \mathcal{D}_{-i}],$$

where the second equality uses $\mathbb{E}[f_i \mid \mathcal{D}_{-i}] = 0$. The term $j = i$ contributes $v_{k(i),n}$ by (A.16), the terms $j \in \mathbf{s}_{k(i)} \setminus \{i\}$ are $\mathcal{D}_{-i}$ -measurable and contribute zero, and $\sum_{j \in \mathbf{s}_\ell} f_j = \sqrt{m_n} S_{\ell,n}$ for $\ell \neq k(i)$, so (A.8) gives

$$c_i = v_{k(i),n} + (K_n - 1)r_i. \quad (\text{A.20})$$

Since $\sum_{i=1}^n v_{k(i),n} = m_n \sum_{k=1}^{K_n} v_{k,n}$, (A.10) yields $\mathcal{K}_n = (nV_n)^{-1} \sum_{i=1}^n c_i$. Since $\mathbb{E}[\mathcal{K}_n] = 1$ and $\text{Var}[X + Y] \leq 2 \text{Var}[X] + 2 \text{Var}[Y]$,

$$\mathbb{E}[|\mathcal{K}_n - 1|]^2 \leq \text{Var}[\mathcal{K}_n] \leq \frac{2 \text{Var}[\bar{v}_n]}{V_n^2} + \frac{2(K_n - 1)^2 \text{Var}[\bar{r}_n]}{V_n^2}, \quad (\text{A.21})$$

which converges to zero by (A.13) and (A.15).

Step 3. Stein remainder. Fix $h \in \text{Lip}(1)$, and let φ_h be the solution to (B.13), so that $\|\varphi'_h\|_\infty \leq 1$ and φ'_h is 2-Lipschitz. As in (B.15),

$$d_W(T_n, \mathcal{N}(0, 1)) \leq \mathbb{E}[|\mathcal{K}_n - 1|] + \sup_{h \in \text{Lip}(1)} |R_n(\varphi_h)|, \quad R_n(\varphi_h) = \mathbb{E}[T_n \varphi_h(T_n)] - \mathbb{E}[\mathcal{K}_n \varphi'_h(T_n)]. \quad (\text{A.22})$$

For each i , since $\mathbb{E}[f_i \mid \mathcal{D}_{-i}] = 0$ and $T_n - \Delta_i = \mathbb{E}[T_n \mid \mathcal{D}_{-i}]$ is $\mathcal{D}_{-i}$ -measurable,

$$\begin{aligned} \mathbb{E}[f_i \varphi_h(T_n)] &= \mathbb{E}[f_i \{ \varphi_h(T_n) - \varphi_h(T_n - \Delta_i) \}] \\ &= \mathbb{E}[f_i \Delta_i \varphi'_h(T_n - \Delta_i)] + \varepsilon_{i,1}(h) \\ &= \frac{1}{\sqrt{nV_n}} \mathbb{E}[c_i \varphi'_h(T_n - \Delta_i)] + \varepsilon_{i,1}(h) \\ &= \frac{1}{\sqrt{nV_n}} \mathbb{E}[c_i \varphi'_h(T_n)] + \varepsilon_{i,1}(h) + \varepsilon_{i,2}(h), \end{aligned}$$

where the second and fourth equalities define the remainders $\varepsilon_{i,1}(h)$ and $\varepsilon_{i,2}(h)$. Multiplying by $(nV_n)^{-1/2}$, summing over i , and using $\mathcal{K}_n = (nV_n)^{-1} \sum_{i=1}^n c_i$ gives

$$|R_n(\varphi_h)| \leq \frac{1}{\sqrt{nV_n}} \sum_{i=1}^n \{ |\varepsilon_{i,1}(h)| + |\varepsilon_{i,2}(h)| \}.$$

As in (B.25), the Taylor remainder satisfies

$$|\varepsilon_{i,1}(h)| \leq \mathbb{E}[|f_i| \min\{\Delta_i^2, 2|\Delta_i|\}] \leq 2 \mathbb{E}[|f_i| |\Delta_i| \min\{|\Delta_i|, 1\}] \leq 2\bar{c} \mathbb{E}[\Delta_i^2]^{1-1/p},$$

where the last inequality uses Hölder's inequality with conjugate exponents p and q , (A.17), and $(|d| \min\{|d|, 1\})^q \leq d^2$ for $d \in \mathbb{R}$.

For the shift remainder, $|\varphi'_h(x) - \varphi'_h(y)| \leq 2 \min\{|x - y|, 1\}$ and the $\mathcal{D}_{-i}$ -measurability of c_i give

$$|\varepsilon_{i,2}(h)| \leq \frac{2}{\sqrt{nV_n}} \mathbb{E}[|c_i| \mid \mathcal{D}_{-i}] \mathbb{E}[\min\{|\Delta_i|, 1\} \mid \mathcal{D}_{-i}].$$

Let $Y_i = \mathbb{E}[\Delta_i^2 \mid \mathcal{D}_{-i}]$. By the conditional Cauchy–Schwarz inequality and (A.16), $|c_i| \leq \sqrt{nV_n} v_{k(i),n}^{1/2} Y_i^{1/2}$ and $\mathbb{E}[\min\{|\Delta_i|, 1\} \mid \mathcal{D}_{-i}] \leq \min\{Y_i, 1\}^{1/2}$, so

$$|\varepsilon_{i,2}(h)| \leq 2 \mathbb{E}[v_{k(i),n}^{1/2} G_i], \quad G_i = (Y_i \min\{Y_i, 1\})^{1/2}.$$

Since $q \in [1, 2)$, $G_i^q = Y_i^q \leq Y_i$ when $Y_i \leq 1$, and $G_i^q = Y_i^{q/2} \leq Y_i$ when $Y_i > 1$. Hölder's inequality with conjugate exponents p and q and (A.17) then give

$$|\varepsilon_{i,2}(h)| \leq 2 \left\| v_{k(i),n}^{1/2} \right\|_p \mathbb{E}[Y_i]^{1-1/p} \leq 2\bar{c} \mathbb{E}[\Delta_i^2]^{1-1/p}.$$

Both bounds are uniform over $h \in \text{Lip}(1)$, so (A.12) and (A.19) give

$$\sup_{h \in \text{Lip}(1)} |R_n(\varphi_h)| \leq 4\bar{c} \sqrt{\frac{n}{V_n}} \max_i \mathbb{E}[\Delta_i^2]^{1-1/p} \leq C \sqrt{n} \left\{ \frac{1}{n} + (K_n - 1)\gamma_n \right\}^{1-1/p}, \quad (\text{A.23})$$

where $C = 4\bar{c} c^{-1/2} (2/c)^{1-1/p}$.

Step 4. Conclusion. By (A.21) to (A.23), it remains to show that the right-hand side of (A.23) vanishes. Since $(a + b)^s \leq a^s + b^s$ for $a, b \geq 0$ and $s \in (0, 1]$,

$$\sqrt{n} \left\{ \frac{1}{n} + (K_n - 1)\gamma_n \right\}^{1-1/p} \leq n^{-(\frac{1}{2} - \frac{1}{p})} + \left\{ n^{\frac{1}{2(1-1/p)}} (K_n - 1)\gamma_n \right\}^{1-1/p},$$

which converges to zero because $p > 2$ and by (A.14). Hence $d_W(T_n, \mathcal{N}(0, 1)) \rightarrow 0$, which implies $T_n \rightsquigarrow \mathcal{N}(0, 1)$. $\blacksquare$

B Proofs

Throughout this section, let $k(i)$ denote the fold not containing observation i , $i \in \tilde{\mathbf{s}}_{k(i)}$, so that W_i evaluates the model $\hat{\eta}_{k(i)}$ trained on the other fold, and define

$$f_i = f_{\hat{\eta}_{k(i)}}(W_i), \quad \Delta_i = T_n - \mathbb{E}[T_n \mid \mathcal{D}_{-i}], \quad i = 1, \dots, n,$$

so that

$$T_n = \frac{\sum_{i=1}^n f_i}{\sqrt{n(1+\rho_n)}}, \quad \mathbb{E}[T_n] = 0, \quad \text{Var}[T_n] = 1.$$

Define also the approximate Stein kernel

$$\mathcal{K}_n = \frac{1 + \frac{2}{n} \sum_{i \in \mathbf{s}_1} r_i}{1 + \rho_n}.$$

For a function $g: \mathbb{R} \rightarrow \mathbb{R}$, let $\|g\|_\infty = \sup_{x \in \mathbb{R}} |g(x)|$.

B.1 Auxiliary Results

Lemma B.1 (Stability Bounds). *It holds that*

$$\mathbb{E}[\text{Var}[\hat{\theta}_{\hat{\eta}_1} \mid \mathcal{D}_{-1}]] = \frac{\gamma_n}{n}, \quad (\text{B.1})$$

and

$$\text{Var}[r_1] \leq \mathbb{E}[r_1^2] \leq \frac{n\gamma_n}{4}. \quad (\text{B.2})$$

Moreover,

$$\mathbb{E}[r_1] = \rho_n, \quad \text{and hence} \quad |\rho_n| \leq \frac{1}{2}\sqrt{n\gamma_n}. \quad (\text{B.3})$$

Proof. Let

$$\hat{\theta}_{\hat{\eta}'_1} = \frac{2}{n} \sum_{j \in \mathbf{s}_2} f_{\hat{\eta}'_1}(W_j)$$

denote the estimator $\hat{\theta}_{\hat{\eta}'_1}$ computed with $\hat{\eta}'_1$ in place of $\hat{\eta}_1$. Then

$$\mathbb{E}[\text{Var}[\hat{\theta}_{\hat{\eta}_1} \mid \mathcal{D}_{-1}]] = \frac{1}{2} \mathbb{E}[\mathbb{E}[(\hat{\theta}_{\hat{\eta}_1} - \hat{\theta}_{\hat{\eta}'_1})^2 \mid \mathcal{D}_{-1}]] \quad (\text{B.4})$$

$$\begin{aligned} &= \frac{1}{2} \mathbb{E}[(\hat{\theta}_{\hat{\eta}_1} - \hat{\theta}_{\hat{\eta}'_1})^2] \\ &= \frac{2}{n^2} \mathbb{E}\left[\left(\sum_{j \in \mathbf{s}_2} \{f_{\hat{\eta}_1}(W_j) - f_{\hat{\eta}'_1}(W_j)\}\right)^2\right] \\ &= \frac{2}{n^2} \sum_{j \in \mathbf{s}_2} \mathbb{E}[(f_{\hat{\eta}_1}(W_j) - f_{\hat{\eta}'_1}(W_j))^2] \\ &= \frac{2}{n^2} \cdot \frac{n}{2} \gamma_n = \frac{\gamma_n}{n}. \end{aligned} \quad (\text{B.5})$$

(B.4) uses that, conditional on $\mathcal{D}_{-1}$, $\hat{\theta}_{\hat{\eta}_1}$ and $\hat{\theta}_{\hat{\eta}'_1}$ are independent and identically distributed, so that $\text{Var}[X \mid \mathcal{D}_{-1}] = \frac{1}{2} \mathbb{E}[(X - X')^2 \mid \mathcal{D}_{-1}]$. (B.5) expands the square; the cross terms vanish because, conditional on $(\hat{\eta}_1, \hat{\eta}'_1)$, the differences $f_{\hat{\eta}_1}(W_j) - f_{\hat{\eta}'_1}(W_j)$ are independent across $j \in \mathbf{s}_2$ with mean $Pf_{\hat{\eta}_1} - Pf_{\hat{\eta}'_1} = 0$. The last equality uses that W_j , $j \in \mathbf{s}_2$, is independent of $(\hat{\eta}_1, \hat{\eta}'_1)$ and $|\mathbf{s}_2| = n/2$.

For (B.2), the first inequality is $\text{Var}[X] \leq \mathbb{E}[X^2]$, and

$$\mathbb{E}[r_1^2] \leq \frac{n^2}{4} \mathbb{E}\left[\text{Var}\left[\sqrt{n/2} \hat{\theta}_{\hat{\eta}_1} \mid \mathcal{D}_{-1}\right] \text{Var}\left[\sqrt{n/2} \hat{\theta}_{\hat{\eta}_2} \mid \mathcal{D}_{-1}\right]\right] \quad (\text{B.6})$$

$$\begin{aligned} &= \frac{n^2}{4} \mathbb{E}\left[\text{Var}\left[\hat{\theta}_{\hat{\eta}_1} \mid \mathcal{D}_{-1}\right]\right] \\ &= \frac{n\gamma_n}{4}. \end{aligned} \quad (\text{B.7})$$

(B.6) is conditional Cauchy–Schwarz. (B.7) uses that $\hat{\eta}_2$ and W_i , $i \in \mathbf{s}_1 \setminus \{1\}$, are $\mathcal{D}_{-1}$ -measurable, so that only $f_{\hat{\eta}_2}(W_1)$ varies and, by the standardization,

$$\text{Var}\left[\sqrt{n/2} \hat{\theta}_{\hat{\eta}_2} \mid \mathcal{D}_{-1}\right] = \frac{n}{2} \left(\frac{2}{n}\right)^2 \text{Var}\left[f_{\hat{\eta}_2}(W_1) \mid \mathcal{D}_{-1}\right] = \frac{2}{n}.$$

The last equality is (B.1).

For (B.3),

$$\begin{aligned} \mathbb{E}[r_1] &= \frac{n^2}{4} \mathbb{E}\left[\text{Cov}\left[\hat{\theta}_{\hat{\eta}_1}, \hat{\theta}_{\hat{\eta}_2} \mid \mathcal{D}_{-1}\right]\right] \\ &= \frac{n}{2} \mathbb{E}\left[\text{Cov}\left[f_{\hat{\eta}_2}(W_1), \hat{\theta}_{\hat{\eta}_1} \mid \mathcal{D}_{-1}\right]\right] \\ &= \frac{n}{2} \mathbb{E}\left[f_{\hat{\eta}_2}(W_1) \hat{\theta}_{\hat{\eta}_1}\right] \\ &= \frac{n}{2} \text{Cov}\left[\hat{\theta}_{\hat{\eta}_1}, \hat{\theta}_{\hat{\eta}_2}\right] \\ &= \rho_n. \end{aligned} \quad (\text{B.8})$$

The second equality uses that $\hat{\eta}_2$ and W_i , $i \in \mathbf{s}_1 \setminus \{1\}$, are $\mathcal{D}_{-1}$ -measurable. (B.8) uses $\mathbb{E}[f_{\hat{\eta}_2}(W_1) \mid \mathcal{D}_{-1}] = Pf_{\hat{\eta}_2} = 0$. The fourth equality uses exchangeability within $\mathbf{s}_1$ and $\mathbb{E}[\hat{\theta}_{\hat{\eta}_2}] = 0$, and the last uses $\text{Var}[\hat{\theta}_{\hat{\eta}_k}] = 2/n$, which follows from the standardization. Finally, by Jensen’s inequality and (B.2), $|\rho_n| = |\mathbb{E}[r_1]| \leq \sqrt{\mathbb{E}[r_1^2]} \leq \frac{1}{2} \sqrt{n\gamma_n}$. $\blacksquare$

B.2 Proof of Lemma 1

Proof. By exchangeability within $\mathbf{s}_1$, $\text{Var}[r_1] = \text{Var}[r_2]$, so $\text{Cov}[r_1, r_2] = \text{Var}[r_1] \text{Corr}(r_1, r_2)$, where both sides are zero if $\text{Var}[r_1] = 0$. Hence, by (B.2) and $|\text{Corr}(r_1, r_2)| \leq 1$,

$$|\text{Cov}[r_1, r_2]| \leq \frac{n\gamma_n}{4} |\text{Corr}(r_1, r_2)| \leq \frac{n\gamma_n}{4}.$$

Under (a), $n^{\frac{1}{2(1-1/p)}}\gamma_n = n^{-\frac{1-2/p}{2(1-1/p)}} \cdot n\gamma_n \rightarrow 0$ since $p > 2$, and the middle term is $O(1) \cdot o(1)$. Under (b), the first condition is Assumption 1(iii) with $p = \infty$, and the middle term equals $\frac{1}{4}\sqrt{n}\gamma_n \cdot \sqrt{n}|\text{Corr}(r_1, r_2)| = o(1) \cdot O(1)$. Under (c), $n^{\frac{1}{2(1-1/p)}}\gamma_n \leq n\gamma_n \rightarrow 0$ since $\frac{1}{2(1-1/p)} \leq 1$, and the right-hand side is $o(1)$. $\blacksquare$

B.3 A Stein Bound

The following theorem is the common core of the proofs of Theorems 1 and 2. It uses only the setup of Section 3, not Assumption 1.

Theorem B.1 (Stein Bound). *For every n ,*

$$\mathbb{E}[\mathcal{K}_n] = 1, \quad (\text{B.9})$$

$$\text{Var}[\mathcal{K}_n] = \frac{2}{(1 + \rho_n)^2} \frac{\text{Var}[r_1]}{n} + \frac{1}{(1 + \rho_n)^2} \left(1 - \frac{2}{n}\right) \text{Cov}[r_1, r_2], \quad (\text{B.10})$$

$$\mathbb{E}[\Delta_1^2] = \frac{1 + 2\rho_n + \frac{n}{4}\gamma_n}{n(1 + \rho_n)}, \quad (\text{B.11})$$

and

$$d_W(T_n, \mathcal{N}(0, 1)) \leq \sqrt{\text{Var}[\mathcal{K}_n]} + \sqrt{\frac{n}{1 + \rho_n}} \left[\mathbb{E}[|f_1| \min\{\Delta_1^2, 2|\Delta_1|\}] + 2\mathbb{E}[\Delta_1^2] \right]. \quad (\text{B.12})$$

Proof. The proof uses Stein's method of normal approximation (Stein, 1972, 1986). Its starting point is that a random variable Z follows a standard normal distribution, $Z \sim \mathcal{N}(0, 1)$, if, and only if,

$$\mathbb{E}[Z\varphi(Z)] = \mathbb{E}[\varphi'(Z)]$$

for every φ in a sufficiently rich class. This characterization suggests that if there is a random variable $\mathcal{K}_n$, often called an approximate Stein kernel, satisfying two properties,

$$\mathbb{E}[T_n\varphi(T_n)] - \mathbb{E}[\mathcal{K}_n\varphi'(T_n)] \approx 0,$$

and

$$\mathcal{K}_n \xrightarrow{L^1} 1,$$

then T_n converges to a normal distribution. The main novelty of this proof is in constructing a specific $\mathcal{K}_n$ that satisfies both requirements under the locality and regularity conditions.

First, for any $h \in \text{Lip}(1)$, let φ_h be the solution to Stein's equation

$$h(t) - \mathbb{E}[h(Z)] = \varphi'_h(t) - t\varphi_h(t), \quad t \in \mathbb{R}, \quad (\text{B.13})$$

given by

$$\varphi_h(t) = e^{t^2/2} \int_{-\infty}^t [h(x) - \mathbb{E}[h(Z)]] e^{-x^2/2} dx.$$

In particular, $\|\varphi'_h\|_\infty \leq 1$ and $|\varphi'_h(x) - \varphi'_h(y)| \leq 2|x - y|$ for all $x, y \in \mathbb{R}$ (Chatterjee, 2008, Lemma 4.2). Second, define the Stein remainder

$$R_n(\varphi_h) = \mathbb{E}[T_n \varphi_h(T_n)] - \mathbb{E}[\mathcal{K}_n \varphi'_h(T_n)]. \quad (\text{B.14})$$

Combining (B.13), (B.14), and $\|\varphi'_h\|_\infty \leq 1$ yields the inequality

$$\begin{aligned} |\mathbb{E}[h(T_n)] - \mathbb{E}[h(Z)]| &= |\mathbb{E}[T_n \varphi_h(T_n)] - \mathbb{E}[\varphi'_h(T_n)]| \\ &= |\mathbb{E}[(1 - \mathcal{K}_n) \varphi'_h(T_n)] - R_n(\varphi_h)| \\ &\leq \mathbb{E}[|\mathcal{K}_n - 1|] + |R_n(\varphi_h)|. \end{aligned}$$

Hence,

$$d_W(T_n, \mathcal{N}(0, 1)) \leq \mathbb{E}[|\mathcal{K}_n - 1|] + \sup_{h \in \text{Lip}(1)} |R_n(\varphi_h)|. \quad (\text{B.15})$$

The proof proceeds in two steps. Step 1 establishes (B.9) and (B.10) and bounds the first term in (B.15). Step 2 bounds the second term and establishes (B.11).

Step 1. Kernel Concentration.

$$\mathbb{E}[|\mathcal{K}_n - 1|]^2 \leq \text{Var}[\mathcal{K}_n] \quad (\text{B.16})$$

$$\begin{aligned} &= \frac{4}{n^2(1 + \rho_n)^2} \text{Var} \left[\sum_{i \in \mathbf{s}_1} r_i \right] \\ &= \frac{4}{n^2(1 + \rho_n)^2} \left(\frac{n}{2} \text{Var}[r_1] + \frac{n}{2} \left(\frac{n}{2} - 1 \right) \text{Cov}[r_1, r_2] \right) \\ &= \frac{2}{(1 + \rho_n)^2} \frac{\text{Var}[r_1]}{n} + \frac{1}{(1 + \rho_n)^2} \left(1 - \frac{2}{n} \right) \text{Cov}[r_1, r_2]. \end{aligned} \quad (\text{B.17})$$

(B.16) follows since

$$\mathbb{E}[\mathcal{K}_n] = (1 + \rho_n)^{-1} (1 + \mathbb{E}[r_1]) = 1,$$

by exchangeability of observations within $\mathbf{s}_1$ and (B.3). (B.17) uses exchangeability within $\mathbf{s}_1$. This proves (B.9) and (B.10).

Step 2. Vanishing Stein Remainder. Fix $h \in \text{Lip}(1)$. The key identity is

$$\mathbb{E}[T_n \varphi_h(T_n)] = \sqrt{\frac{n}{1 + \rho_n}} \mathbb{E}[f_1 \varphi_h(T_n)] \quad (\text{B.18})$$

$$= \sqrt{\frac{n}{1 + \rho_n}} \mathbb{E}[f_1 \{\varphi_h(T_n) - \varphi_h(T_n - \Delta_1)\}] \quad (\text{B.19})$$

$$= \sqrt{\frac{n}{1 + \rho_n}} \mathbb{E}[f_1 \Delta_1 \varphi'_h(T_n - \Delta_1)] + \varepsilon_{n,1}(h) \quad (\text{B.20})$$

$$= \frac{1}{1 + \rho_n} \mathbb{E}[(1 + r_1) \varphi'_h(T_n - \Delta_1)] + \varepsilon_{n,1}(h) \quad (\text{B.21})$$

$$= \frac{1}{1 + \rho_n} \mathbb{E}[(1 + r_1) \varphi'_h(T_n)] + \varepsilon_{n,1}(h) + \varepsilon_{n,2}(h) \quad (\text{B.22})$$

$$= \mathbb{E}[\mathcal{K}_n \varphi'_h(T_n)] + \varepsilon_{n,1}(h) + \varepsilon_{n,2}(h), \quad (\text{B.23})$$

where the deterministic errors $\varepsilon_{n,1}(h)$ and $\varepsilon_{n,2}(h)$ are defined by (B.20) and (B.22), respectively. Thus,

$$R_n(\varphi_h) = \varepsilon_{n,1}(h) + \varepsilon_{n,2}(h).$$

(B.18) follows from fold symmetry and exchangeability within $\mathbf{s}_1$. (B.19) subtracts

$$\mathbb{E}[f_1 \varphi_h(T_n - \Delta_1)] = \mathbb{E}[\mathbb{E}[f_1 \mid \mathcal{D}_{-1}] \varphi_h(T_n - \Delta_1)] = 0.$$

(B.20) defines the Taylor expansion remainder $\varepsilon_{n,1}(h)$. (B.21) uses

$$\begin{aligned} \sqrt{n(1 + \rho_n)} \mathbb{E}[f_1 \Delta_1 \mid \mathcal{D}_{-1}] &= \mathbb{E} \left[f_1 \left\{ \sum_{j=1}^n f_j - \mathbb{E} \left[\sum_{j=1}^n f_j \mid \mathcal{D}_{-1} \right] \right\} \mid \mathcal{D}_{-1} \right] \\ &= \text{Cov} \left[f_1, \sum_{j=1}^n f_j \mid \mathcal{D}_{-1} \right] \\ &= 1 + \frac{n}{2} \text{Cov} [f_1, \hat{\theta}_{\hat{n}_1} \mid \mathcal{D}_{-1}] \\ &= 1 + r_1. \end{aligned} \quad (\text{B.24})$$

(B.22) defines the remainder $\varepsilon_{n,2}(h)$, and (B.23) follows from exchangeability within $\mathbf{s}_1$ and the definition of $\mathcal{K}_n$.

To bound $\varepsilon_{n,1}(h)$, note that, for all $t, d \in \mathbb{R}$,

$$\begin{aligned} |\varphi_h(t) - \varphi_h(t - d) - d \varphi'_h(t - d)| &= |d| \left| \int_0^1 [\varphi'_h(t - d + sd) - \varphi'_h(t - d)] \, ds \right| \\ &\leq |d| \int_0^1 \min\{2s|d|, 2\} \, ds \\ &\leq \min\{d^2, 2|d|\} \\ &\leq 2|d| \min\{|d|, 1\}, \end{aligned}$$

where the first inequality uses the 2-Lipschitz property of φ'_h and the bound $\|\varphi'_h\|_\infty \leq 1$. Consequently,

$$\begin{aligned} |\varepsilon_{n,1}(h)| &= \sqrt{\frac{n}{1+\rho_n}} |\mathbb{E}[f_1 \{\varphi_h(T_n) - \varphi_h(T_n - \Delta_1) - \Delta_1 \varphi'_h(T_n - \Delta_1)\}]| \\ &\leq \sqrt{\frac{n}{1+\rho_n}} \mathbb{E}[|f_1| \min\{\Delta_1^2, 2|\Delta_1|\}] \end{aligned} \quad (\text{B.25})$$

For the remainder $\varepsilon_{n,2}(h)$, since φ'_h is 2-Lipschitz,

$$\begin{aligned} |\varepsilon_{n,2}(h)| &= \frac{1}{1+\rho_n} |\mathbb{E}[(1+r_1) \{\varphi'_h(T_n - \Delta_1) - \varphi'_h(T_n)\}]| \\ &\leq \frac{2}{1+\rho_n} \mathbb{E}[|1+r_1| |\Delta_1|] \\ &\leq 2\sqrt{\frac{n}{1+\rho_n}} \mathbb{E}[\Delta_1^2]. \end{aligned} \quad (\text{B.26})$$

The last inequality follows from Cauchy–Schwarz, since (B.24) and conditional Cauchy–Schwarz give

$$\begin{aligned} \mathbb{E}[(1+r_1)^2] &= n(1+\rho_n) \mathbb{E}[\mathbb{E}[f_1 \Delta_1 \mid \mathcal{D}_{-1}]^2] \\ &\leq n(1+\rho_n) \mathbb{E}[\mathbb{E}[f_1^2 \mid \mathcal{D}_{-1}] \mathbb{E}[\Delta_1^2 \mid \mathcal{D}_{-1}]] \\ &= n(1+\rho_n) \mathbb{E}[\Delta_1^2], \end{aligned}$$

where $\mathbb{E}[f_1^2 \mid \mathcal{D}_{-1}] = 1$ by the standardization.

Since both error bounds are uniform over $h \in \text{Lip}(1)$, combining (B.25) and (B.26) gives

$$\begin{aligned} \sup_{h \in \text{Lip}(1)} |R_n(\varphi_h)| &= \sup_{h \in \text{Lip}(1)} |\varepsilon_{n,1}(h) + \varepsilon_{n,2}(h)| \\ &\leq \sup_{h \in \text{Lip}(1)} |\varepsilon_{n,1}(h)| + \sup_{h \in \text{Lip}(1)} |\varepsilon_{n,2}(h)| \\ &\leq \sqrt{\frac{n}{1+\rho_n}} [\mathbb{E}[|f_1| \min\{\Delta_1^2, 2|\Delta_1|\}] + 2\mathbb{E}[\Delta_1^2]]. \end{aligned} \quad (\text{B.27})$$

Together with (B.16) and (B.15), this proves (B.12).

It remains to establish (B.11). Since the scores of all other fold-1 observations are $\mathcal{D}_{-1}$ -measurable and $\mathbb{E}[f_1 \mid \mathcal{D}_{-1}] = 0$,

$$\Delta_1 = \frac{f_1 + \frac{n}{2} \left\{ \hat{\theta}_{\hat{\eta}_1} - \mathbb{E}[\hat{\theta}_{\hat{\eta}_1} \mid \mathcal{D}_{-1}] \right\}}{\sqrt{n(1+\rho_n)}}.$$

Expanding the square,

$$\begin{aligned}\mathbb{E}[\Delta_1^2] &= \frac{1}{n(1+\rho_n)} \left[\mathbb{E}[f_1^2] + n \mathbb{E}[f_1 \{\hat{\theta}_{\hat{\eta}_1} - \mathbb{E}[\hat{\theta}_{\hat{\eta}_1} | \mathcal{D}_{-1}]\}] + \frac{n^2}{4} \mathbb{E}[\text{Var}[\hat{\theta}_{\hat{\eta}_1} | \mathcal{D}_{-1}]] \right] \\ &= \frac{1 + 2\rho_n + \frac{n}{4}\gamma_n}{n(1+\rho_n)}.\end{aligned}$$

The second equality uses $\mathbb{E}[f_1^2] = 1$, the identity $\mathbb{E}[f_1 \hat{\theta}_{\hat{\eta}_1}] = 2\rho_n/n$ from (B.8), $\mathbb{E}[f_1 \mathbb{E}[\hat{\theta}_{\hat{\eta}_1} | \mathcal{D}_{-1}]] = \mathbb{E}[\mathbb{E}[f_1 | \mathcal{D}_{-1}] \mathbb{E}[\hat{\theta}_{\hat{\eta}_1} | \mathcal{D}_{-1}]] = 0$, and (B.1). $\blacksquare$

B.4 Proof of Theorem 1

Proof. The main steps of the proof are first established in Theorem B.1. After that, it suffices to show that $\text{Var}[\mathcal{K}_n] \rightarrow 0$ and that $\sqrt{n} [\mathbb{E}[|f_1| \min\{\Delta_1^2, 2|\Delta_1|\}] + 2\mathbb{E}[\Delta_1^2]] \rightarrow 0$.

For the first claim, (B.10) and (B.2) give

$$\text{Var}[\mathcal{K}_n] \leq \frac{\gamma_n}{2(1+\rho_n)^2} + \frac{|\text{Cov}[r_1, r_2]|}{(1+\rho_n)^2} \rightarrow 0,$$

since $1+\rho_n$ is bounded away from zero by Assumption 1(ii), $\gamma_n \rightarrow 0$ by Assumption 1(iii), and $\text{Cov}[r_1, r_2] \rightarrow 0$ by Assumption 1(iv).

For the second claim, first note that (B.11), $\rho_n \leq 1$, and Assumption 1(ii) give

$$\mathbb{E}[\Delta_1^2] \leq \frac{2}{n} + \frac{\gamma_n}{4\underline{c}} \leq C_0 \left(\frac{1}{n} + \gamma_n \right), \quad C_0 = \max \left\{ 2, \frac{1}{4\underline{c}} \right\}. \quad (\text{B.28})$$

Next, let $q = 1/(1 - 1/p) \in [1, 2)$ denote the conjugate exponent of p . By Hölder's inequality,

$$\begin{aligned}\mathbb{E}[|f_1| \min\{\Delta_1^2, 2|\Delta_1|\}] &\leq 2 \mathbb{E}[|f_1| |\Delta_1| \min\{|\Delta_1|, 1\}] \\ &\leq 2\bar{c} \mathbb{E}[\Delta_1^2]^{1-1/p},\end{aligned}$$

where the last inequality uses conjugate exponents p and q , $\|f_1\|_p = \|f_{\hat{\eta}_1}(W)\|_p \leq \bar{c}$ by fold symmetry and Assumption 1(i), and

$$(|d| \min\{|d|, 1\})^q \leq d^2,$$

which holds for all $d \in \mathbb{R}$ since $q \in [1, 2)$. Combining these bounds gives

$$\begin{aligned} \sqrt{n} \left[\mathbb{E}[|f_1| \min\{\Delta_1^2, 2|\Delta_1|\}] + 2\mathbb{E}[\Delta_1^2] \right] &\leq 2(\bar{c} + 1) \sqrt{n} \mathbb{E}[\Delta_1^2]^{1-1/p} \\ &\leq C\sqrt{n} \left(\frac{1}{n} + \gamma_n \right)^{1-1/p} \\ &= C \left[n^{-\frac{1-2/p}{2(1-1/p)}} + n^{\frac{1}{2(1-1/p)}} \gamma_n \right]^{1-1/p} \longrightarrow 0, \end{aligned}$$

for some constant $C < \infty$ depending only on $\bar{c}$, $\underline{c}$, and p . The first inequality uses

$$\mathbb{E}[\Delta_1^2] \leq \mathbb{E}[T_n^2] = 1 \implies \mathbb{E}[\Delta_1^2] \leq \mathbb{E}[\Delta_1^2]^{1-1/p},$$

where $\mathbb{E}[\Delta_1^2] \leq \mathbb{E}[T_n^2]$ follows because conditional expectation is an L^2 projection. The second inequality uses (B.28). The convergence follows from $p > 2$ and Assumption 1(iii). Since $1 + \rho_n \geq \underline{c}$, the factor $(1 + \rho_n)^{-1/2}$ in (B.12) is bounded, which completes the proof. $\blacksquare$

B.5 Proof of Theorem 2

Proof. Let $M < \infty$ be such that $n^{1/(1-1/p)}\gamma_n \leq M$ and $n|\text{Cov}[r_1, r_2]| \leq M$ for all n . Since $1/(1-1/p) \geq 1$, also $n\gamma_n \leq M$. Throughout, $C < \infty$ denotes a constant depending only on $\bar{c}$, $\underline{c}$, p , and M , whose value may change from line to line. By (B.12) and $1 + \rho_n \geq \underline{c}$, it suffices to show that each of $\text{Var}[\mathcal{K}_n]$, $\sqrt{n} \mathbb{E}[\Delta_1^2]$, and $\sqrt{n} \mathbb{E}[|f_1| \min\{\Delta_1^2, 2|\Delta_1|\}]$ is at most C/n , $C/\sqrt{n}$, and $C/\sqrt{n}$, respectively.

First, by (B.10), (B.2), and Assumption 1(ii),

$$\text{Var}[\mathcal{K}_n] \leq \frac{\gamma_n}{2\underline{c}^2} + \frac{|\text{Cov}[r_1, r_2]|}{\underline{c}^2} \leq \frac{M}{2\underline{c}^2 n} + \frac{M}{\underline{c}^2 n} \leq \frac{C}{n}.$$

Second, by (B.11), $\rho_n \leq 1$, and Assumption 1(ii),

$$\mathbb{E}[\Delta_1^2] \leq \frac{2}{n} + \frac{\gamma_n}{4\underline{c}} \leq \frac{2}{n} + \frac{M}{4\underline{c}n} \leq \frac{C}{n}. \quad (\text{B.29})$$

Third, let $\phi(x) = \min\{x^2, 2|x|\}$ and decompose $\Delta_1 = A + B$, where

$$A = \frac{f_1}{\sqrt{n(1 + \rho_n)}}, \quad B = \frac{\frac{n}{2} \left\{ \hat{\theta}_{\hat{\eta}_1} - \mathbb{E}[\hat{\theta}_{\hat{\eta}_1} \mid \mathcal{D}_{-1}] \right\}}{\sqrt{n(1 + \rho_n)}}.$$

Since ϕ is even, nondecreasing in $|x|$, and satisfies $\phi(2x) \leq 4\phi(x)$, for all $a, b \in \mathbb{R}$,

$$\phi(a + b) \leq \phi(2 \max\{|a|, |b|\}) \leq 4 \max\{\phi(a), \phi(b)\} \leq 4\phi(a) + 4\phi(b),$$

and hence

$$\mathbb{E}[|f_1| \phi(\Delta_1)] \leq 4 \mathbb{E}[|f_1| \phi(A)] + 4 \mathbb{E}[|f_1| \phi(B)]. \quad (\text{B.30})$$

For the first term in (B.30), $\phi(A) \leq A^2$ and Lyapunov's inequality with $p \geq 3$ give

$$\mathbb{E}[|f_1| \phi(A)] \leq \frac{\mathbb{E}[|f_1|^3]}{n(1 + \rho_n)} \leq \frac{\bar{c}^3}{\underline{c}n}.$$

For the second term, set $\alpha = 1 - 2/p \in [1/3, 1]$ and $q = 1/(1 - 1/p) \in [1, 3/2]$, so that $(1 + \alpha)q = 2$ and $1/q = 1 - 1/p$. Since $\min\{x^2, 2|x|\} \leq (x^2)^\alpha (2|x|)^{1-\alpha} = 2^{1-\alpha} |x|^{1+\alpha}$, Hölder's inequality with conjugate exponents p and q gives

$$\mathbb{E}[|f_1| \phi(B)] \leq 2^{1-\alpha} \|f_1\|_p \mathbb{E}[|B|^{(1+\alpha)q}]^{1/q} = 2^{2/p} \|f_1\|_p \mathbb{E}[B^2]^{1-1/p}.$$

By (B.1) and Assumption 1(ii),

$$\mathbb{E}[B^2] = \frac{n}{4(1 + \rho_n)} \mathbb{E}[\text{Var}[\hat{\theta}_{\hat{\eta}_1} \mid \mathcal{D}_{-1}]] = \frac{\gamma_n}{4(1 + \rho_n)} \leq \frac{\gamma_n}{4\underline{c}},$$

and therefore, since $\|f_1\|_p = \|f_{\hat{\eta}_1}(W)\|_p \leq \bar{c}$ by fold symmetry and $\gamma_n \leq Mn^{-1/(1-1/p)}$,

$$\mathbb{E}[|f_1| \phi(B)] \leq 2^{2/p} \bar{c} \left(\frac{\gamma_n}{4\underline{c}} \right)^{1-1/p} \leq C \gamma_n^{1-1/p} \leq \frac{C}{n}.$$

Substituting both bounds into (B.30) yields $\sqrt{n} \mathbb{E}[|f_1| \phi(\Delta_1)] \leq C/\sqrt{n}$.

Combining the three bounds with (B.12),

$$d_W(T_n, \mathcal{N}(0, 1)) \leq \sqrt{\frac{C}{n}} + \frac{1}{\sqrt{\underline{c}}} \left[\frac{C}{\sqrt{n}} + \frac{2C}{\sqrt{n}} \right] = O(n^{-1/2}).$$

■

B.6 Proof of Theorem 3

Proof. Write Z_k° for $Z_{k,1}$ with its denominator replaced by $\sqrt{m} \sigma_{\hat{\eta}_{k,1}}$. After subtracting the fitted population mean from each score, write $Z_k^\circ = U_k - V_k$, where U_k and V_k are the primary and reference normalized sums. Unconditionally, the four quarters are independent i.i.d. samples. Conditional on the two training quarters, U_1, U_2 are fixed, while V_1, V_2 are independent, centered, and have variance one. Therefore,

$$\mathbb{E}[Z_1^\circ Z_2^\circ] = \rho_{n/2}, \quad \mathbb{E}[(Z_k^\circ)^2] = 2. \quad (\text{B.31})$$

For fixed values $x_1, \dots, x_{2m}$, let $Z(x)$ denote the contrast in (4) under a uniform balanced split. If $v(x)$ is the centered vector divided by its Euclidean norm $\|\cdot\|_2$, with $v(x) = 0$ for a constant vector, then

$$Z(x) = \sqrt{\frac{2m-1}{m}} \epsilon^\top v(x),$$

where ϵ_i is 1 on the selected half and -1 on its complement. The balanced-sign identities

$$\mathbb{E}[\epsilon_i] = 0, \quad \mathbb{E}[\epsilon_i \epsilon_j] = -\frac{1}{2m-1}, \quad \mathbb{E}[\epsilon_i \epsilon_j \epsilon_\ell \epsilon_r] = \frac{3}{(2m-1)(2m-3)}$$

for distinct indices give

$$\mathbb{E}[Z(x)] = 0, \quad \mathbb{E}[Z(x)^2] = 2\mathbb{1}(v(x) \neq 0), \quad \mathbb{E}[Z(x)^4] \leq 12. \quad (\text{B.32})$$

The case $m = 1$ follows directly.

Conditional on the training quarter, the pooled test observations are i.i.d. The moment condition gives uniform integrability of the squared normalized scores, so truncation and the conditional law of large numbers imply

$$\mathbb{E} \left[\left| \frac{\hat{\sigma}_{\hat{\eta}_{k,1}}^2}{\sigma_{\hat{\eta}_{k,1}}^2} - 1 \right| \right] \rightarrow 0.$$

Conditioning on the pooled scores and randomizing their balanced split then gives

$$\begin{aligned} \mathbb{E}[(Z_{k,1} - Z_k^\circ)^2] &\leq 2 \mathbb{E} \left[\left(\frac{\hat{\sigma}_{\hat{\eta}_{k,1}}}{\sigma_{\hat{\eta}_{k,1}}} - 1 \right)^2 \right] \\ &\leq 2 \mathbb{E} \left[\left| \frac{\hat{\sigma}_{\hat{\eta}_{k,1}}^2}{\sigma_{\hat{\eta}_{k,1}}^2} - 1 \right| \right] \rightarrow 0. \end{aligned}$$

When the sample variance is zero, both contrasts are zero. By Cauchy–Schwarz and (B.31),

$$\mathbb{E}[Z_{1,1} Z_{2,1}] = \rho_{n/2} + o(1).$$

Put $q_n = \mathbb{E}[Z_{1,1} Z_{2,1} \mid \mathcal{D}]$. Applying (B.32) conditionally on the data and the training subset gives

$$\mathbb{E}[Z_{k,1}^2 \mid \mathcal{D}] \leq 2, \quad \mathbb{E}[Z_{k,1}^4 \mid \mathcal{D}] \leq 12,$$

and hence $|q_n| \leq 2$ and $\text{Var}[Z_{1,1}Z_{2,1} \mid \mathcal{D}] \leq 12$. Conditional independence of the repetitions and the fact that projection onto $[-1, 1]$ cannot increase the distance to ρ_n yield

$$\begin{aligned} \mathbb{E}[(\hat{\rho}_n - \rho_n)^2] &\leq \frac{12}{B} + \text{Var}[q_n] + \{\mathbb{E}[q_n] - \rho_n\}^2 \\ &= \frac{12}{B} + \text{Var}[q_n] + \{\rho_{n/2} - \rho_n + o(1)\}^2 \rightarrow 0, \end{aligned}$$

where the last term vanishes because $\rho_{n/2}$ and ρ_n both converge to ρ .

The same moment argument gives $\hat{\sigma}_{\hat{\eta}_k}/\sigma_{\hat{\eta}_k} \xrightarrow{p} 1$ for the original fits. Since $\rho > -1$, $1 + \rho_n$ is bounded away from zero for all large n . Consequently, (10) and correlation consistency imply

$$\begin{aligned} \frac{\hat{\theta}_{\hat{\eta}} - \theta_{\hat{\eta}}}{\widehat{\text{se}}(\hat{\theta}_{\hat{\eta}})} &= \frac{\sqrt{n/2} \sum_{k=1}^2 (\hat{\theta}_{\hat{\eta}_k} - \theta_{\hat{\eta}_k}) / \sigma_{\hat{\eta}_k}}{\sqrt{2(1 + \rho_n)}} + o_p(1) \\ &\rightsquigarrow \mathcal{N}(0, 1), \end{aligned}$$

where the last step uses Theorem 1. Continuity of the normal distribution proves the coverage statement. $\blacksquare$

C Inference for the BLP Example

For the OLS estimator in Example 3, use the slope influence score

$$\phi_\eta(y, x) = \frac{\{\eta(x) - P\eta\} [y - \mathbb{E}[Y] - \theta_\eta \{\eta(x) - P\eta\}]}{P(\eta - P\eta)^2}.$$

Under the zero-slope null this equals the score in Example 3. Away from the null, the residual term also accounts for estimation of the prediction variance.

For each refit, define

$$\begin{aligned} (\hat{\alpha}_{k,b}, \hat{\beta}_{k,b}) &= \underset{a,t}{\operatorname{argmin}} \sum_{i \in \tilde{\mathbf{s}}_k} \left\{ Y_i - a - t\hat{\eta}_{k,b}(X_i) \right\}^2, \\ \hat{z}_{k,b,i} &= \hat{\eta}_{k,b}(X_i) - \frac{1}{2m} \sum_{j \in \tilde{\mathbf{s}}_k} \hat{\eta}_{k,b}(X_j), \\ \hat{e}_{k,b,i} &= Y_i - \hat{\alpha}_{k,b} - \hat{\beta}_{k,b} \hat{\eta}_{k,b}(X_i), \\ \hat{\phi}_{k,b,i} &= \frac{\hat{z}_{k,b,i} \hat{e}_{k,b,i}}{(2m)^{-1} \sum_{j \in \tilde{\mathbf{s}}_k} \hat{z}_{k,b,j}^2}. \end{aligned}$$

Use $\hat{\phi}_{k,b,i}$ in place of $f_{\hat{\eta}_{k,b}}(W_i)$ in (3) and (4). Its common positive denominator cancels from $Z_{k,b}$, so the implementation may equivalently use $\hat{z}_{k,b,i}\hat{e}_{k,b,i}$. These pooled quantities do not depend on the test fold's primary/reference labels.

The correlation result extends provided

$$\frac{\sum_{i \in \tilde{s}_k} \left\{ \hat{\phi}_{k,1,i} - \phi_{\hat{\eta}_{k,1}}(W_i) \right\}^2}{2m P \phi_{\hat{\eta}_{k,1}}^2} \xrightarrow{p} 0,$$

with the theorem's normalized-score assumptions applied to ϕ_η . Indeed, with $Z(x)$ and $v(x)$ as in the proof of Theorem 3 in Appendix B.6, for two pooled score vectors x, y , the balanced contrast satisfies

$$\mathbb{E}[\{Z(x) - Z(y)\}^2] = 2 \|v(x) - v(y)\|_2^2.$$

Relative score consistency makes the normalized vectors converge in probability. Their squared distance is bounded by four, giving convergence in mean square of the contrasts and leaving the correlation proof unchanged.

For coverage using the original OLS estimators, also require

$$\hat{\beta}_k - \theta_{\hat{\eta}_k} = \frac{2}{n} \sum_{i \in \tilde{s}_k} \phi_{\hat{\eta}_k}(W_i) + o_p\left(\sqrt{\frac{P \phi_{\hat{\eta}_k}^2}{n}}\right),$$

and relative consistency of their marginal standard errors. In this application, the scale-balance condition concerns the slope influence scales $\sqrt{P \phi_{\hat{\eta}_k}^2}$.