Semiparametrically Efficient Inference for Kernel Measures of Noise Heterogeneity

Jakub Wornbard [*1] **Zikai Shen** [*2] **Dimitri Meunier** [1] **Arthur Gretton** [1]

Abstract

We develop semiparametrically efficient inference for kernel measures of noise heterogeneity in additive noise models. In many applications, the regression function is estimated using flexible machine learning methods. Downstream procedures based on the resulting residuals can then inherit first-stage bias: regression error may induce spurious dependence between covariates and residuals, invalidating the assumptions needed for standard analysis. We construct a novel Hilbert-valued one-step estimator of the kernel covariance operator between covariates and residuals. Our estimator yields bootstrap-calibrated tests for residual independence and goodness of fit in additive noise models, while also providing asymptotically efficient confidence intervals for the kernel dependence measure under noise heterogeneity. The framework extends to settings with additional covariates, enabling inference on distributional heterogeneity of residual noise across treatment groups. Simulations show improved calibration and power relative to naive plug-in residual methods.

1. Introduction

Many statistical and machine learning workflows use residuals to assess whether a fitted model has captured all systematic structure in the data. In additive noise models, this idea takes a particularly simple form: after regressing an outcome on covariates, the remaining noise should be independent of the variables whose effect has been explained. More broadly, one may ask whether the distribution of residual noise is homogeneous across covariates, treatment groups, or other features of interest. Kernel dependence measures provide a flexible way to quantify such residual noise heterogeneity, since they can detect general distributional differences beyond changes in mean or variance.

Classical nonparametric independence testing assumes that the variables to be tested are directly observed (Gretton et al., 2007; Heller et al., 2013; Berrett & Samworth, 2019; Shekhar et al., 2023). In residual-based problems, however, the noise is not observed directly. It is generated by a first-stage regression estimator, often fit using flexible machine learning methods. This creates a multi-stage inference problem: downstream kernel procedures may inherit bias from the first-stage nuisance estimator. In this work, we bridge two strands of the statistical inference literature. On the one hand, recent work combines machine learning with semiparametric efficiency theory to estimate finite-dimensional target parameters (Chernozhukov et al., 2018; Hines et al., 2022; Kennedy, 2024). On the other hand, reproducing kernel Hilbert space (RKHS)-based dependence measures have become standard tools for nonparametric dependence testing (Gretton et al., 2005, 2007; Chwialkowski et al., 2014; Zhang et al., 2018; Albert et al., 2022; Shekhar et al., 2023) and related tasks in machine learning (Song et al., 2007; Shen et al., 2009; Peters et al., 2013; Li et al., 2021). This raises the question of whether semiparametric debiasing can be used to construct valid kernel inference procedures when the variables entering the kernel statistic are themselves estimated.

To make the exposition concrete, we begin with the additive noise model studied by Peters et al. (2013) in the context of causal discovery. The dataset consists of n i.i.d. pairs $(x_i, y_i)_{i=1}^n$ drawn from the same distribution as $(X, Y) \sim P$. The null hypothesis is that P is correctly specified by an additive noise model: there exists a function m, whose functional form is unrestricted, and a centered noise variable ϵ independent of X such that $Y = m(X) + \epsilon$. The null is not a restriction on the functional form of the regression function; rather, it is the requirement that the regression residual be independent of the covariates.

If the true residuals $\epsilon_i = y_i - m(x_i)$ were observed, goodness of fit could be assessed by testing whether X and ϵ are independent. In practice, m is unknown, so

[*]Equal contribution [1]Gatsby Computational Neuroscience Unit, University College London [2]Department of Statistical Science, University College London. Correspondence to: Jakub Wornbard <jakub.wornbard.24@ucl.ac.uk>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

an investigator estimates it by $\widehat{m}$ and constructs residuals $\widehat{\epsilon}_i = y_i - \widehat{m}(x_i)$. A natural plug-in approach is then to apply an off-the-shelf nonparametric independence test to the sample $\{(x_i, \widehat{\epsilon}_i)\}_{i=1}^n$.

We focus on kernel dependence measures based on the RKHS covariance operator between covariates and residuals. The squared Hilbert–Schmidt norm of this operator is the Hilbert–Schmidt Independence Criterion (HSIC), a standard kernel measure of dependence (Gretton et al., 2005). With characteristic kernels (Sriperumbudur et al., 2010, 2011), this squared norm vanishes if and only if the covariate and residual noise are independent. Thus, in the oracle setting where ϵ is observed, HSIC provides a natural test statistic for additive-noise goodness of fit and, more generally, for residual noise heterogeneity.

The setting with plug-in estimator $\widehat{m}$ poses significant challenges. Estimated residuals satisfy $\widehat{\epsilon}_i = \epsilon_i + m(x_i) - \widehat{m}(x_i)$, so regression error can induce dependence between X_i and $\widehat{\epsilon}_i$ even when the additive noise model is correctly specified. As a simple illustration of the challenges for hypothesis testing, suppose the data is split into two disjoint folds: one fold is used to estimate m, and the other is used to construct the independence test. If the fitting fold is small, $\widehat{m}$ may have substantial prediction error, and the kernel test may interpret this first-stage error as evidence of dependence, leading to inflated Type I error. If the fitting fold is large, the residuals better approximate the oracle residuals, but the holdout sample used for testing is small, reducing power. These observations highlight the practical difficulty of controlling Type I error while retaining power when nuisance functions are estimated. Related empirical failures have been observed in prior work on conditional independence testing (Pogodin et al., 2024; Wang et al., 2025; He et al., 2025). Sample splitting removes the direct dependence between the fitted regression algorithm and the test sample, but conditional on the fitted function the residual still contains the systematic error $m(X) - \widehat{m}(X)$, which is a function of X.

The discussion above motivates a debiased estimator for kernel residual dependence. A standard route to debiasing is to use the efficient influence function of a target functional (Bickel et al., 1993; Kennedy, 2024). However, direct scalar debiasing of HSIC presents a complication: the empirical unbiased HSIC is a degenerate U-statistic under the null (Lee, 1990). At the population level, this first-order degeneracy appears as a vanishing efficient influence function: the first-order efficient influence function of

$$\mathrm{HSIC}(X, Y - \mathbb{E}[Y \mid X])$$

is identically zero under the null. This is not a pathology of HSIC, but a consequence of its squared-norm structure: at independence, the underlying kernel covariance operator is zero, and the derivative of $x \mapsto \|x\|^2$ vanishes at zero. Therefore, first-order scalar debiasing of HSIC cannot remove the first-stage bias that matters for testing near the null. The insufficiency of first-order linearization for Maximum Mean Discrepancy based test statistics has been observed in Luedtke et al. (2019); Williamson et al. (2023). It has also been observed in a related setting by the concurrent work Agarwal & Luedtke (2026).

Our key idea is to debias before taking the squared norm. We construct a Hilbert-valued one-step estimator of the kernel covariance operator, building on recent work on Hilbert-valued semiparametric efficiency theory (Luedtke & Chung, 2024; Morzywolek et al., 2025). We then obtain scalar kernel dependence measures by mapping the debiased RKHS-valued estimator to $\mathbb{R}$ through its squared Hilbert–Schmidt norm. This operator-level construction yields a calibrated test under the null, where scalar HSIC is degenerate, while also giving efficient confidence intervals for the value of the kernel dependence measure under alternatives.

Contributions. We propose a Hilbert-valued one-step estimator for the kernel covariance operator between covariates and residuals, whose squared norm gives the corresponding kernel dependence measure. We prove asymptotic linearity and weak convergence of the estimator under nonparametric nuisance estimation. These results yield a bootstrap-calibrated independence test with asymptotic validity under the null and consistency against fixed alternatives. Under alternatives, our approach also provides asymptotically efficient confidence intervals for the value of the kernel dependence measure. We further extend the framework to settings with additional covariates, enabling covariate-adjusted inference on distributional heterogeneity of residual noise across treatment groups. Finally, we conduct simulations showing that the proposed method improves calibration and power relative to naive plug-in residual methods.

Outline. In Section 2, we introduce the problem setup and the RKHS covariance operator target. In Section 3, we derive the Hilbert-valued efficient influence function, define the one-step estimator, and establish asymptotic linearity, weak convergence, and bootstrap calibration. Section 4 presents synthetic experiments illustrating calibration, power, and confidence-interval coverage. Section 5 discusses limitations and future directions.

2. Problem Setup

We study independence testing between random variables X and Y, after adjusting for a covariate, denoted W. We apply our method to two practical settings, namely goodness-of-fit testing for additive noise models and testing for noise homo-

geneity across different treatment groups, see Section 4. We refer to these as ***Settings 1*** and ***2***, respectively. In the former case, $X = W$, and our data consist of (X, Y)-pairs. In the latter case, X is a treatment indicator, and $W = (X, T)$, where T denotes additional covariates. We do not test for homogeneity with respect to the covariates T. For the remainder of the paper, we make the following assumption, although our theory readily generalizes to the case where X is not measurable with respect to the covariate W we adjust for.

Assumption 1. *There exists a measurable map $\pi : \mathcal{W} \to \mathcal{X}$ such that $X = \pi(W)$.*

Let $\mathcal{O} = \mathcal{X} \times \mathcal{W} \times \mathbb{R}^{d_y}$, and let $\mathcal{M}$ denote a nonparametric model of probability laws on $\mathcal{O}$. We write $P_0 \in \mathcal{M}$ for the true data-generating distribution. For $P \in \mathcal{M}$, define $m_P(w) \doteq \mathbb{E}_P[Y \mid W = w]$ and $\xi_P(w, y) \doteq y - m_P(w)$. We test $H_0 : P_0 \in \{P \in \mathcal{M} : \xi_P(W, Y) \perp X\}$. In ***Setting 2***, testing for H_0 addresses the question of whether the marginal distribution of the residual is independent of X, after adjusting for X and additional covariates T.

We next introduce necessary RKHS concepts. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric positive definite kernel and let $\mathcal{H}_k$ be its reproducing kernel Hilbert space (RKHS) (Berlinet & Thomas-Agnan, 2011). We denote the canonical feature map of $\mathcal{H}_k$ as $\varphi_k(x) = k(\cdot, x)$. Let $l : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \to \mathbb{R}$ be a symmetric positive definite kernel with RKHS $\mathcal{H}_l$ and feature map $\varphi_l(u) = l(\cdot, u)$. We refer the reader to Appendix B for additional background on RKHSs. We assume all RKHSs mentioned are separable. This is satisfied, e.g. if the reproducing kernel is continuous and the domain is separable.

Throughout, we assume the kernels are measurable and satisfy the moment conditions needed for the following Bochner expectations to exist; for example, the following assumption is sufficient.

Assumption 2 (Kernel boundedness)**.** $\sup_{x \in \mathcal{X}} k(x, x) \leq \kappa_k^2$ *and* $\sup_{u \in \mathbb{R}^{d_y}} l(u, u) \leq \kappa_l^2$.

The dependence between X and the residual $\xi_P(W, Y)$ is encoded by first featurizing them with infinite-dimensional kernel features and then taking the expectation of their outer product. In particular, we use the language of kernel mean embeddings (Smola et al., 2007). We denote the kernel mean embeddings of X and $\xi_P(W, Y)$ as $\mu_{X,P} \doteq \mathbb{E}_P[\varphi_k(X)] \in \mathcal{H}_k$ and $\mu_{\xi,P} \doteq \mathbb{E}_P[\varphi_l(\xi_P(W, Y))] \in \mathcal{H}_l$. For characteristic kernels, testing for dependence between X and $\xi_P(W, Y)$ nonparametrically is equivalent to testing whether the kernel mean embedding of the tensor product feature map $\mathbb{E}_P[\varphi_k(X) \otimes \varphi_l(\xi_P(W, Y))]$ is equal to $\mu_{X,P} \otimes \mu_{\xi,P}$ (Gretton, 2015). This is algebraically equivalent to testing for $\Psi(P) = 0$, where

$$\Psi(P) = \mathbb{E}_P\left[\widetilde{\varphi}_{k,P}(X) \otimes \widetilde{\varphi}_{l,P}(\xi_P(W, Y))\right] \in \mathcal{H}_{kl},$$

where we define the *centered* feature maps $\widetilde{\varphi}_{k,P}(x) \doteq \varphi_k(x) - \mu_{X,P}$ and $\widetilde{\varphi}_{l,P}(\xi_P(w, y)) \doteq \varphi_l(\xi_P(w, y)) - \mu_{\xi,P}$. $\mathcal{H}_{kl} \doteq \mathcal{H}_k \otimes \mathcal{H}_l$ denotes the tensor product RKHS induced by the product kernel $(x, u), (x', u') \mapsto k(x, x')l(u, u')$. $\|\Psi(P)\|^2_{\mathcal{H}_{kl}}$ is the Hilbert–Schmidt Independence Criterion (HSIC) between X and $Y - m_P(W)$ (Gretton et al., 2005). We define the hypothesis test

$$H_0 : \Psi(P_0) = 0 \qquad \text{against} \qquad H_1 : \Psi(P_0) \neq 0.$$

Throughout, a subscript 0 denotes evaluation at the true distribution P_0; for example, $m_0 := m_{P_0}$ and $\Psi_0 := \Psi(P_0)$.

3. One-step Estimator

The key challenge is that regression error can be indistinguishable from dependence between the estimated residual and the covariate X in the downstream hypothesis test. Semiparametric efficiency theory offers an attractive possibility of debiased two-stage learning, by computing an efficient influence function (EIF) and then removing the first-order bias of the first-stage estimator. However, the scalar-valued HSIC is not regular at the null H_0 as its EIF at the null is identically zero. The correct object to debias is thus the Hilbert-valued cross-covariance operator $\Psi(P)$, and we should compute the squared norm ***after*** debiasing. We now derive the EIF of the cross-covariance. Let $L_0^2(P_0)$ denote the space of square-integrable functions with zero mean under P_0.

Assumption 3 (Model, moments, and differentiability)**.** *Let $\mathcal{M}$ be a nonparametric model with tangent space $\dot{\mathcal{M}}_{P_0} = L_0^2(P_0)$. Assume that $\mathbb{E}_0\|Y\|^2 < \infty$, and that $u \mapsto \varphi_l(u)$ is continuously Fréchet differentiable as a map $\mathbb{R}^{d_y} \to \mathcal{H}_l$, with $D\varphi_l(u)[a] = \sum_{j=1}^{d_y} a_j\, \partial_j \varphi_l(u), a \in \mathbb{R}^{d_y}$. Moreover, for each j, we assume that $\sup_{u \in \mathbb{R}^{d_y}} \|\partial_j \varphi_l(u)\|_{\mathcal{H}_l} < \infty$.*

Intuition about the form of the EIF. The correction term in the EIF can be understood from a first-order expansion of the residual feature map. Let $\widehat{\delta}(w) \doteq \widehat{m}(w) - m_0(w)$. Then

$$\begin{aligned}\varphi_l\{Y - \widehat{m}(W)\} &= \varphi_l\{\xi_0(W, Y) - \widehat{\delta}(W)\} \\ &\approx \varphi_l\{\xi_0(W, Y)\} \\ &\quad - \sum_{j=1}^{d_y} \widehat{\delta}_j(W) \partial_j \varphi_l\{\xi_0(W, Y)\}.\end{aligned}$$

Therefore, conditional on W, the first-order shift in the

residual feature is

$$\text{first-order shift} = -\sum_{j=1}^{d_y} \widehat{\delta}_j(W) v_{0,j}(W),$$
$$v_{0,j}(w) \doteq \mathbb{E}_0[\partial_j \varphi_l\{\xi_0(W,Y)\} \mid W = w].$$

This is the component of the plug-in error that can create spurious dependence with X. The EIF subtracts $\sum_{j=1}^{d_y} \xi_{0,j}(W,Y) v_{0,j}(W)$. For the plug-in residual $\widehat{\xi}_j = \xi_{0,j} - \widehat{\delta}_j(W)$, this correction changes by $+\sum_{j=1}^{d_y} \widehat{\delta}_j(W) v_{0,j}(W)$, which cancels the conditional first-order Taylor bias above. Equivalently, after the correction, the leading perturbation is proportional to $\sum_{j=1}^{d_y} \widehat{\delta}_j(W) \left[\partial_j \varphi_l\{\xi_0(W,Y)\} - v_{0,j}(W)\right]$, whose conditional mean given W is zero. Thus the first-order effect of regression error is removed, leaving only higher-order remainder terms.

Theorem 1 (Efficient influence function)**.** *Suppose Assumptions 1 to 3 hold. Then $\Psi : \mathcal{M} \to \mathcal{H}_{kl}$ is pathwise differentiable at P_0. For any regular parametric submodel $t \mapsto P_t$ through P_0 with score $s \in L_0^2(P_0)$,*

$$\dot{\Psi}_0(s) \doteq \lim_{t \to 0} \frac{\Psi(P_t) - \Psi(P_0)}{t} = \mathbb{E}_0[\phi_0(O) s(O)] \in \mathcal{H}_{kl}.$$

For $o = (x, w, y)$,

$$\phi_0(o) \doteq \widetilde{\varphi}_{k,0}(x) \otimes \Bigg[\widetilde{\varphi}_{l,0}(\xi_0(w,y)) - \sum_{j=1}^{d_y} \xi_{0,j}(w,y) v_{0,j}(w)\Bigg] - \Psi_0.$$

ϕ_0 belongs to $L_0^2(P_0; \mathcal{H}_{kl})$, therefore it is the $\mathcal{H}_{kl}$-valued EIF of Ψ at P_0 in the sense of Luedtke & Chung (2024).

A slightly more general form of Theorem 1 is proved in Appendix C.

3.1. One-step estimator and asymptotic normality

We now use the EIF to perform one-step debiasing for Ψ_0. We refer to the debiased estimator as a one-step estimator (Schick, 1986; Luedtke & Chung, 2024). We will establish Hilbert-valued asymptotic normality of this estimator under a set of reasonable assumptions. We may use the asymptotic normality theorem to construct a confidence interval for any scalar-valued transformation of Ψ_0. We choose the RKHS norm of Ψ_0, which is sufficient for testing for whether $\Psi_0 = 0$. A hypothesis test with Type I error control directly leads to a confidence interval for the true value Ψ_0. We refer to such a test as calibrated. By including all values of Ψ_0 not rejected by a test with Type I error control, one obtains a confidence set with the desired coverage level. We leverage this duality to simultaneously obtain a confidence interval for $\|\Psi_0\|_{\mathcal{H}_{kl}}$ and a hypothesis test for H_0.

We assume the dataset is $\{o_i = (x_i, w_i, y_i)\}_{i=1}^n$, where n is even. Let $I_1 = \{1, \ldots, n/2\}$ and $I_2 = \{n/2+1, \ldots, n\}$. We fit the nuisance functions on I_2 and evaluate the one-step estimator on I_1. Write $\widehat{\xi}_i \doteq y_i - \widehat{m}(w_i)$, where $\widehat{m}$ is *any* machine learning estimator for m_0. Define the empirical kernel mean embeddings $\widehat{\mu}_X \doteq \frac{2}{n} \sum_{i \in I_2} \varphi_k(x_i)$ and $\widehat{\mu}_\xi \doteq \frac{2}{n} \sum_{i \in I_2} \varphi_l(\widehat{\xi}_i)$.

For $j = 1, \ldots, d_y$, we estimate the $\mathcal{H}_l$-valued nuisance $v_{0,j}$ by vector-valued kernel ridge regression (Caponnetto & De Vito, 2007; Park & Muandet, 2020; Li et al., 2024). Let $\mathcal{H}_{ql}$ denote the vector-valued RKHS induced by the kernel $q(\cdot,\cdot)\mathrm{Id}_{\mathcal{H}_l}$. For details on vector-valued kernels and their associated RKHS, we refer the reader to Li et al. (2024) and Appendix I. We define, for each $j = 1, \ldots, d_y$,

$$\widehat{v}_j = \underset{v \in \mathcal{H}_{ql}}{\arg\min} \left\{ \frac{2}{n} \sum_{i \in I_2} \left\| (\partial_j \varphi_l)(\widehat{\xi}_i) - v(w_i) \right\|_{\mathcal{H}_l}^2 + \lambda_{j,n} \|v\|_{\mathcal{H}_{ql}}^2 \right\}.$$

For $i \in I_1$, define $\widehat{M}_i \doteq \varphi_l(\widehat{\xi}_i) - \widehat{\mu}_\xi - \sum_{j=1}^{d_y} \widehat{\xi}_{i,j} \widehat{v}_j(w_i)$. The one-step estimator evaluated on I_1 is

$$\widehat{\Psi}_n \doteq \frac{2}{n} \sum_{i \in I_1} (\varphi_k(x_i) - \widehat{\mu}_X) \otimes \widehat{M}_i. \tag{1}$$

Remark 1 (Cross-fitting)**.** *To simplify our exposition and notation, in the main document we work exclusively with a sample-split estimator $\widehat{\Psi}_n$, introduced above in Eq. (1). Since $\widehat{\Psi}_n$ is constructed by evaluating half of the samples, it leads to a loss of statistical power. To remedy this, we construct a K-fold cross-fitted estimator, where we use all K folds except one for fitting, one fold for evaluation, and average over all K choices of the evaluation fold. We emphasize that all simulation experiments and theoretical guarantees in our work apply to the K-fold cross-fitted estimator, as detailed in the Appendix. The cross-fitted estimator allows us to remove restrictive Donsker-type conditions on the function class (Schick, 1986; Chernozhukov et al., 2018). We provide generalized results for the cross-fitted estimator in the Appendix, denoted with* ***reverse hat*** *notation as $\check{\Psi}_n$. We develop the core asymptotic results for this estimator in Section C assuming $K = 2$. Obtaining analogues for higher values is not harder but requires more careful notation. This simplification is standard in the literature (Chernozhukov et al., 2018)*

Assumption 4 (Main kernel differentiability and Lipschitz continuity assumption)**.** *We assume that the kernel $l : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \to \mathbb{R}$ is such that, for all $j \in [d_y]$, the mixed*

partial derivative $\partial_j \partial_{j+d_y} l$ exists and is continuous, where the derivative is taken with respect to the jth coordinate of the first argument and the jth coordinate of the second argument. We further assume that there exists a global constant $L > 0$ such that $\|\varphi_l(u) - \varphi_l(v)\|_{\mathcal{H}_l} \leq L\|u - v\|_{\mathbb{R}^{d_y}}$ for all $u, v \in \mathbb{R}^{d_y}$. We assume that $\varphi_l : \mathbb{R}^{d_y} \to \mathcal{H}_l$ is twice Fréchet differentiable and that its second derivative is uniformly bounded: there exists $L_2 < \infty$ such that $\|D^2\varphi_l(u)\|_{\mathrm{op}} \leq L_2$ for all $u \in \mathbb{R}^{d_y}$.

We provide an assumption stated solely in terms of partial derivatives of l that implies Assumption 4 in Assumption 5 in the Appendix. This allows us to show that Assumption 4 is satisfied by popular kernels, including Gaussian kernels and sufficiently smooth Matérn kernels. By Steinwart & Christmann (2008, Lemma 4.34, Corollary 4.36), Assumption 4 in particular implies that, for each $j = 1, \ldots, d_y$, the partial derivative $\partial_j \varphi_l : \mathbb{R}^{d_y} \to \mathcal{H}_l$ exists and is continuous, and for all $y, y' \in \mathbb{R}^{d_y}$ we have $\langle \partial_j \varphi_l(y), \partial_j \varphi_l(y') \rangle_{\mathcal{H}_l} = \partial_j \partial_{j+d_y} l(y, y')$.

Requirement 1 (Nuisance-level requirements)**.**

(A) There exists a constant $C > 0$ such that

$$\|m_0\|_{L^\infty(P_{0,W};\mathbb{R}^{d_y})} \leq C,$$
$$\operatorname*{ess\,sup}_{w \in \mathrm{supp}(P_{0,W})} \mathbb{E}_0[\|Y - m_0(W)\|^2_{\mathbb{R}^{d_y}} \mid W = w] \leq C.$$

(B) We have

$$\|\widehat{m} - m_0\|_{L^\infty(P_{0,W};\mathbb{R}^{d_y})} = o_p(1),$$
$$\|\widehat{m} - m_0\|_{L^2(P_{0,W};\mathbb{R}^{d_y})} = o_p(1),$$
$$\|\widehat{v} - v_0\|_{L^2(P_{0,W};\mathcal{H}_l^{d_y})} = o_p(1),$$
$$\|\widehat{\mu}_\xi - \mu_{\xi,0}\|_{\mathcal{H}_l} = o_p(1).$$

(C) We have

$$\begin{aligned}&\|\widehat{m} - m_0\|^2_{L^2(P_{0,W};\mathbb{R}^{d_y})}\\ &+\|\widehat{m} - m_0\|_{L^2(P_{0,W};\mathbb{R}^{d_y})} \cdot \|\widehat{v} - v_0\|_{L^2(P_{0,W};\mathcal{H}_l^{d_y})}\\ &=o_p(n^{-1/2}).\end{aligned}$$

Theorem 2 (Asymptotic linearity and weak convergence of the sample-split one-step estimator)**.** *Suppose the conclusion of Theorem 1 holds and let ϕ_0 be given as in Theorem 1. Under Assumption 4 and Requirement 1, the sample-split one-step estimator $\widehat{\Psi}_n$ is regular and asymptotically linear:*

$$\sqrt{n/2}\left(\widehat{\Psi}_n - \Psi_0\right) = \sqrt{\frac{2}{n}} \sum_{i \in I_1} \phi_0(o_i) + o_p(1) \rightsquigarrow \mathbb{H}, \tag{2}$$

where $\mathbb{H}$ is a tight $\mathcal{H}_{kl}$-valued Gaussian random variable such that, for each $h \in \mathcal{H}_{kl}$, $\langle \mathbb{H}, h \rangle_{\mathcal{H}_{kl}} \sim \mathcal{N}(0, \mathbb{E}_0[\langle \phi_0(X, W, Y), h \rangle^2_{\mathcal{H}_{kl}}])$.

When is Requirement 1 satisfied? Requirements 1B and 1C are a set of sufficient conditions that hold under more fundamental assumptions, such as membership in a Hölder smoothness class. $\widehat{m}$ is a standard nonparametric regression estimator for m_0. Its minimax rate of convergence is well understood (Tsybakov, 2009). Both $\widehat{v}$ and $\widehat{\mu}_\xi$ depend on the nonparametric first-stage $\widehat{m}$. In the second stage, $\widehat{v}$ is obtained by performing a vector-valued kernel ridge regression, while $\widehat{\mu}_\xi$ is obtained by computing a sample average. In the former case, the convergence rates depend on the smoothness of the vector-valued target v_0 and the capacity of the vector-valued RKHS $\mathcal{H}_{ql}$. In the latter case, the second-stage error is a Monte Carlo error. We emphasize that thanks to the second-order nature of the nuisance rates in Requirement 1C, nowhere do we require a nuisance function to converge at the parametric rate $n^{-\frac{1}{2}}$. In Appendix I, we provide the necessary background on vector-valued kernel learning, state and prove technical theorems on the convergence rates of $\widehat{v}$ and $\widehat{\mu}_\xi$, and provide a concrete example where Requirement 1 is satisfied.

3.2. Debiased confidence set and hypothesis test

We now construct a confidence interval estimator for $\|\Psi_0\|_{\mathcal{H}_{kl}}$, and then convert it into a hypothesis test for the null H_0 that $\Psi_0 = 0$ with desired Type I error control. We introduce two methods, both of which rely on the asymptotic normality result of Theorem 2. By the continuous mapping theorem applied to Eq. (2), we deduce that $(n/2)\|\widehat{\Psi}_n - \Psi_0\|^2_{\mathcal{H}_{kl}} \rightsquigarrow \|\mathbb{H}\|^2_{\mathcal{H}_{kl}}$. We can approximate $\|\mathbb{H}\|^2_{\mathcal{H}_{kl}}$ by the bootstrapped distribution of $\|\widehat{\Psi}_n - \widehat{\Psi}_n^{(b)}\|^2_{\mathcal{H}_{kl}}$ (Efron, 1992). We briefly sketch our procedure, and refer the reader to Appendix F for a formal exposition. For $b \in [B]$, we let $o_1^{(b)}, \ldots, o_{\frac{n}{2}}^{(b)} \overset{\text{i.i.d.}}{\sim} \frac{2}{n}\sum_{i \in I_1} \delta_{o_i}$, conditionally on the original data. Let ζ_n denote the $(1-\alpha)$-quantile of $\left\|\sqrt{n/2}(\widehat{\Psi}_n - \widehat{\Psi}_n^{(b)})\right\|^2_{\mathcal{H}_{kl}}$, where $\widehat{\Psi}_n^{(b)}$ is given as in Eq. (1) except evaluated on the bth bootstrap sample. Then ζ_n converges to the $(1-\alpha)$-quantile $\zeta_{1-\alpha}$ of $\|\mathbb{H}\|^2_{\mathcal{H}_{kl}}$, provided the estimated EIF is a consistent estimator for the true EIF. We refer the reader to Luedtke & Chung (2024, Theorem 4) for the proof. In simulations, we replace ζ_n by its finite-B approximation $\widehat{\zeta}_n$.

Armed with a consistent estimate ζ_n for the $(1-\alpha)$-quantile of $\|\mathbb{H}\|^2_{\mathcal{H}_{kl}}$, we propose the following confidence interval for $\|\Psi_0\|_{\mathcal{H}_{kl}}$:

$$\begin{aligned}\mathcal{I}_{\zeta_n} &\doteq [L_{\zeta_n}, U_{\zeta_n}],\\ L_{\zeta_n} &\doteq \left\{\|\widehat{\Psi}_n\|_{\mathcal{H}_{kl}} - \sqrt{2\zeta_n/n}\right\}_+,\\ U_{\zeta_n} &\doteq \|\widehat{\Psi}_n\|_{\mathcal{H}_{kl}} + \sqrt{2\zeta_n/n}.\end{aligned} \tag{3}$$

It is a $(1-\alpha)$-confidence interval for $\|\Psi_0\|_{\mathcal{H}_{kl}}$, due to the reverse triangle inequality $\left|\|\widehat{\Psi}_n\|_{\mathcal{H}_{kl}} - \|\Psi_0\|_{\mathcal{H}_{kl}}\right| \leq$

$\|\widehat{\Psi}_n - \Psi_0\|_{\mathcal{H}_{kl}}$. It is asymptotically exact at the null in the sense that, when $\Psi_0 = 0$, the reverse triangle inequality becomes an equality. We obtain a calibrated hypothesis test for H_0 by rejecting if and only if $0 \notin \mathcal{I}_{\zeta_n}$. In general, we both expect and observe in simulations that Eq. (3) behaves conservatively when the null does not hold. We emphasize the key role of Theorem 2 in the validity of this construction.

However, when $\Psi_0 \neq 0$, we may construct a confidence interval that performs even better in terms of asymptotic variance. We observe the following decomposition

$$\begin{aligned}
&\sqrt{n/2}\Big(\|\widehat{\Psi}_n\|^2_{\mathcal{H}_{kl}} - \|\Psi_0\|^2_{\mathcal{H}_{kl}}\Big) \\
&= \sqrt{n/2}\Big\langle \widehat{\Psi}_n - \Psi_0, \widehat{\Psi}_n + \Psi_0 \Big\rangle_{\mathcal{H}_{kl}} \\
&= \sqrt{n/2}\|\widehat{\Psi}_n - \Psi_0\|^2_{\mathcal{H}_{kl}} + \sqrt{2n}\langle \widehat{\Psi}_n - \Psi_0, \Psi_0\rangle_{\mathcal{H}_{kl}}.
\end{aligned} \tag{4}$$

For fixed $\Psi_0 \neq 0$, the second term is asymptotically normal with mean zero and asymptotic variance $\mathbb{E}_0[4\langle \phi_0(X,W,Y), \Psi_0\rangle^2_{\mathcal{H}_{kl}}]$, whereas the first term is $O_p(1/\sqrt{n})$. As a result,

$$\begin{aligned}
\Delta_n &\doteq \sqrt{n/2}\left(\|\widehat{\Psi}_n\|^2_{\mathcal{H}_{kl}} - \|\Psi_0\|^2_{\mathcal{H}_{kl}}\right), \\
\Delta_n &\rightsquigarrow \mathcal{N}(0, \sigma^2_\delta), \\
\sigma^2_\delta &\doteq \mathbb{E}_0[4\langle \phi_0(X,W,Y), \Psi_0\rangle^2_{\mathcal{H}_{kl}}].
\end{aligned}$$

Since $\widehat{\Psi}_n$ is the one-step estimator for Ψ_0, our method is an example of the functional delta method. The fact that the derivative vanishes at $\Psi_0 = 0$ is precisely what prevents us from debiasing the scalar-valued HSIC directly. We define the following confidence interval for $\|\Psi_0\|^2_{\mathcal{H}_{kl}}$.

$$\begin{aligned}
\eta_n &\doteq 2 z_{1-\alpha/2}\sqrt{\frac{2\widehat{\sigma}^2_n}{n}}, \\
\mathcal{J}^{1-\alpha}_\delta &\doteq [L_\delta, U_\delta], \\
L_\delta &\doteq \|\widehat{\Psi}_n\|^2_{\mathcal{H}_{kl}} - \eta_n, \\
U_\delta &\doteq \|\widehat{\Psi}_n\|^2_{\mathcal{H}_{kl}} + \eta_n.
\end{aligned}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a standard normal distribution, and $\widehat{\sigma}_n$ is an estimator for $\mathbb{E}_0[\langle \phi_0(X,W,Y), \Psi_0\rangle^2_{\mathcal{H}_{kl}}]$. From van der Vaart (2000, Theorem 25.47), we know that $\mathbb{E}_0[8\langle \phi_0(X,W,Y), \Psi_0\rangle^2_{\mathcal{H}_{kl}}]$ is the semiparametric efficiency bound for estimating $\|\Psi_0\|^2_{\mathcal{H}_{kl}}$ with $\frac{n}{2}$ samples.

A further application of the delta method for the square-root map, yields the confidence set for $\|\Psi_0\|_{\mathcal{H}_{kl}}$,

$$\begin{aligned}
\mathcal{I}^{1-\alpha}_\delta &\doteq [\ell_\delta, u_\delta], \\
\ell_\delta &\doteq \|\widehat{\Psi}_n\|_{\mathcal{H}_{kl}} - \frac{\eta_n}{2\|\widehat{\Psi}_n\|_{\mathcal{H}_{kl}}}, \\
u_\delta &\doteq \|\widehat{\Psi}_n\|_{\mathcal{H}_{kl}} + \frac{\eta_n}{2\|\widehat{\Psi}_n\|_{\mathcal{H}_{kl}}},
\end{aligned}$$

defined whenever $\|\widehat{\Psi}_n\|^2 > 0$, which is eventually true whenever Ψ_0 is non-zero. This construction retains asymptotic efficiency. This suggests that if $\Psi \neq 0$, then $\mathcal{I}^{1-\alpha}_\delta$ may be much narrower than $\mathcal{I}_{\zeta_n}$ while being asymptotically calibrated. We confirm this empirically in Section 4. In the Appendix, we construct such an estimator in Eq. (11), and establish its consistency in Proposition 3.

As we have access to an asymptotically always valid confidence interval, we can adopt the following approach. We define a data-adaptive union confidence interval that is $\mathcal{I}^{1-\alpha}_\delta$ augmented with the singleton set $\{0\}$ if $0 \in \mathcal{I}_{\zeta_n}$, for $\mathcal{I}_{\zeta_n}$ defined in Eq. (3) (see Morzywolek et al. (2025, Theorem 4) for a similar approach). We would nevertheless like to understand the validity of $\mathcal{I}^{1-\alpha}_\delta$ in finite samples, where the sample size n is small and the true Ψ_0 is close to zero. This is motivated by the asymptotics where we consider a sequence of experiments where $\Psi_n \asymp \frac{1}{\sqrt{n}}$. Along this sequence, the two terms in Eq. (4) are of the same order of magnitude. Therefore a confidence interval based on the asymptotic variance of the second term may not be trustworthy. We propose a lightweight data-driven diagnostic that determines when it is plausible that

$$2\sqrt{n/2}\langle \Psi_0, \widehat{\Psi}_n - \Psi_0\rangle_{\mathcal{H}_{kl}} \gg_p \sqrt{n/2}\|\widehat{\Psi}_n - \Psi_0\|^2_{\mathcal{H}_{kl}}.$$

The details of our construction can be found in Appendix H.

4. Experiments

We verify the validity of the proposed procedures on a range of synthetic experiments, where we are able to explicitly control the extent to which noise homogeneity is violated, and to adjust the statistical difficulty of all the intermediate procedures.

Baselines. We compare against plug-in residual procedures that estimate m and then apply standard HSIC inference without the EIF correction. The split-sample baseline follows the ANM diagnostic of Peters et al. (2013): fit $\hat{m}$ on a training split, compute residuals $Y_i - \hat{m}(W_i)$ on held-out observations, and run the permutation HSIC test (Gretton et al., 2007). We report several training fractions because the split controls a calibration–power tradeoff.

We also consider a full-sample, cross-fitted plug-in permutation test. Let $I_1, \ldots, I_K$ be the folds, and write $\hat{m}^{(-k)}$ for the regression estimator trained on all observations outside fold $\mathcal{I}_k$. For $i \in \mathcal{I}_k$, define the out-of-fold residual $\hat{\xi}_i = Y_i - \hat{m}^{(-k)}(W_i)$, form Gram matrices $K_{ij} = k(X_i, X_j)$ and $L_{ij} = l(\hat{\xi}_i, \hat{\xi}_j)$, and perform a full-sample permutation test using these. This uses all observations and avoids in-sample residuals, but remains heuristic because the permutation null does not account for first-stage estimation error.

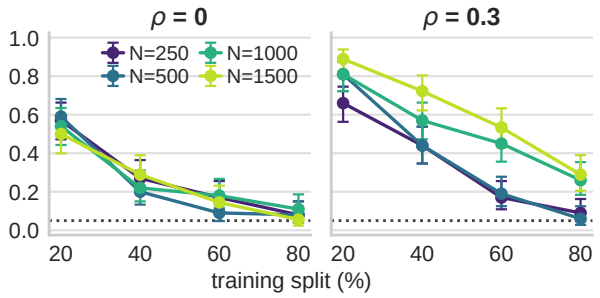


Figure 1. Split fit-test baseline at $s_m = 1, s_\epsilon = 0.25$.

For confidence intervals, we also compare against plug-in methods, which use the same bootstrap or delta-method construction as our estimator, but with the EIF correction term omitted. This helps us evaluate the effect of bias correction on the performance.

X-dependent residuals. We first evaluate goodness-of-fit testing for the scalar additive-noise model $Y = m(X) + \epsilon$, $\epsilon \perp X$. Data are generated from

$$X \sim \mathrm{Unif}[-3, 3],$$
$$Y = m(X) + \sigma(X)Z, \qquad Z \sim \mathcal{N}(0, 1).$$

Both m and $\log \sigma$ are sampled as truncated random Fourier series. The decay parameters s_m and s_ϵ control smoothness, and hence first-stage estimation difficulty: s_m is the exponent in the power-law decay rate of the Fourier coefficients of m, and s_ϵ plays an analogous role in $\log \sigma(X)$. ρ controls the strength of residual heterogeneity. Smaller values of these parameters lead to rougher regression functions and less regular variance structure in the noise. ρ multiplies the non-constant X-dependent term in $\log \sigma(X)$: $\rho = 0$ gives homogeneous noise and larger ρ yields stronger X-dependence in the residual distribution. In each setting, we estimate $\sqrt{\mathrm{HSIC}(X, \epsilon)}$ using Gaussian kernels with bandwidths chosen by the median heuristic. We keep $d_x = d_y = 1$ in the reported experiments to isolate first-stage residual-estimation bias from the additional effects of high-dimensional kernel testing and nuisance estimation. The exact data-generating mechanism is given in Appendix G.1.

Figure 2 reports rejection curves over ρ for a representative setting. At $\rho = 0$, the debiased bootstrap test is close to nominal, while the cross-fitted permutation baseline is conservative. As ρ increases, the debiased test gains power without requiring a split-ratio choice. So does the baseline, though it attains lower power. Figure 1 shows the split-sample plug-in baseline for the same synthetic residual experiment. Under the null, rejection should be near the nominal 5% line, but small training splits leave enough residual-estimation bias to inflate rejection. Under the alternative, larger training splits fit the regression better but

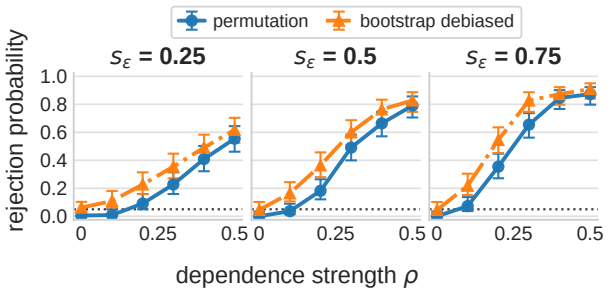


Figure 2. Rejection probabilities at $s_m = 1$ and $n = 500$ across noise-dependence levels and smoothness parameters s_ϵ of $\sigma(x)$, for the debiased CI of Eq. (3) and the cross-fitted permutation baseline.

leave fewer observations for the HSIC test, so power can fall sharply. This calibration–power tradeoff is the tuning problem that Figure 2 avoids: the debiased test uses the full sample and gains power without requiring a split-ratio choice.

Residual heterogeneity with additional covariates. We next consider a setting where the magnitude of Ψ_0, rather than only the null $\Psi_0 = 0$, is scientifically meaningful. Let $W = (X, T)$, where $X \in \{0, 1\}$ is a binary group indicator and T is an additional covariate. Writing $R = Y - \mathbb{E}[Y \mid X, T]$, the target measures whether the covariate-adjusted residual law differs across the two groups.

This connection is especially transparent with the discrete kernel $k(x, x') = \mathbf{1}\{x = x'\}$ on X. When the two groups are balanced,

$$\mathrm{HSIC}(X, R) = \frac{1}{8}\mathrm{MMD}_l^2(\mathcal{L}(R \mid X = 0), \mathcal{L}(R \mid X = 1)),$$

is the Maximum Mean Discrepancy (Gretton et al., 2007) associated with the residual kernel l, between the laws of the residuals under $X = 0$ and under $X = 1$. We evaluate this target in a synthetic model

$$Y = m(X, T) + \sigma_X Z, \qquad Z \sim \mathcal{N}(0, 1),$$

where m is a finite trigonometric series in T, and l is Gaussian. The null has equal residual scales across groups, while the alternative has $\sigma_0 \neq \sigma_1$. Full data-generating details are given in Appendix G.4.

Figure 3 confirms that confidence intervals based on plug-in Ψ are highly suboptimal in this setting. They are very conservative and too wide, while the delta-method CIs track their nominal coverage near-perfectly and are substantially narrower.

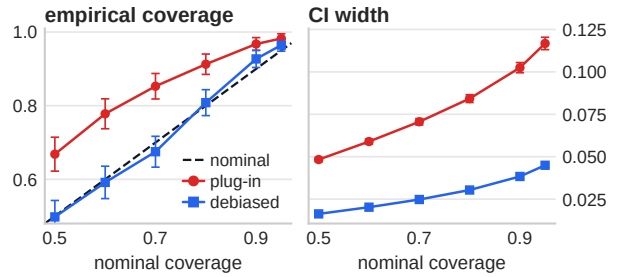


Figure 3. Delta-method CI calibration for Ψ_0 when $W = (X, T)$.

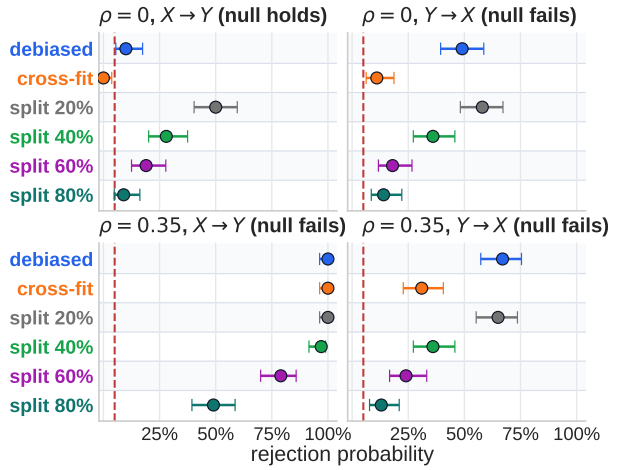


Figure 4. Rejection probabilities for nominal level-5% causal-discovery tests. "Debiased" is our bootstrap test, "cross-fit" is the full-sample cross-fitted permutation baseline, and "split x%" fits nuisances on x% of the sample and tests on the rest.

Causal discovery by two-way goodness-of-fit testing. Additive-noise causal discovery compares the two regression directions by fitting

$$Y = m_{Y|X}(X) + \epsilon_{Y|X},$$
$$X = m_{X|Y}(Y) + \epsilon_{X|Y},$$

and testing whether the fitted residuals are independent of the corresponding covariate (Peters et al., 2013). In the ideal additive-noise setting, the causal direction should pass this goodness-of-fit test, while the reverse direction in general should fail, except for special non-identifiable cases such as linear m and Gaussian noise. This requires calibration under a fitted-residual null and power against a non-additive reverse direction.

We study a synthetic bivariate example with known structural direction $X \to Y$. The structural function is nonlinear and resembles the low-dimensional cause-effect pairs used in the Tübingen cause-effect pair benchmark (Mooij et al., 2016). We consider two regimes. In the first, the forward model has homogeneous additive noise, so the ANM null holds for the regression of Y on X and fails in the reverse direction. In the second, the structural direction remains $X \to Y$, but the forward residual law depends on X; in this case both ANM goodness-of-fit nulls should be rejected. Full data-generating details and example datasets are given in Appendix G.3. Figure 4 shows that the debiased bootstrap gives the most favourable two-way goodness-of-fit profile, and highlights the failure mode of split-sample tests. The latter only become calibrated under the null when 80% of the sample is used for fitting, but the remaining small testing sample leads to very low power.

5. Discussion

We developed an operator-level debiasing approach for kernel measures of residual noise heterogeneity. Rather than applying a semiparametric correction directly to scalar HSIC, whose first-order derivative degenerates at independence, we target the RKHS cross-covariance operator between covariates and residuals. The resulting Hilbert-valued one-step estimator yields a bootstrap-calibrated test of residual independence and a confidence interval for the root-HSIC signal $\|\Psi_0\|_{H_{kl}}$. When dependence holds, the same estimator gives delta-method confidence intervals for $\|\Psi_0\|^2_{H_{kl}}$, attaining the semiparametric efficiency bound for this scalar functional. In synthetic experiments, the debiased procedure improves calibration and power relative to plug-in residual baselines.

A number of areas warrant further investigation. First, it would be valuable to develop a non-asymptotic theory for our estimators and confidence sets. A second direction is suggested by our findings in Appendix F, where centering the delta-method confidence sets at U-statistics leads to better finite sample performance despite being asymptotically equivalent. The same idea does not trivially extend to our more conservative confidence sets based on the triangle inequality, but it would be interesting to understand how their small-sample performance can be optimized. Beyond this, there are many alternative methods of constructing everywhere-valid confidence intervals from an efficient estimator of Ψ. It would be interesting to investigate the optimal way of constructing a confidence set which has good coverage when $\Psi = 0$ and reduces to the derived asymptotically efficient one when $\Psi \neq 0$.

Impact Statement

This paper presents work whose goal is to advance statistical methodology and its applications in causal discovery. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Adams, R. A. and Fournier, J. J. F. ***Sobolev spaces***, volume 140. Elsevier Science & Technology, Chantilly, 2nd edition, 2003. Web.

Agarwal, M. and Luedtke, A. Sinkhorn treatment effects: A causal optimal transport measure. ***arXiv preprint arXiv:2605.08485***, 2026.

Albert, M., Laurent, B., Marrel, A., and Meynaoui, A. Adaptive test of independence based on HSIC measures. ***The Annals of Statistics***, 50(2):858–879, apr 2022.

Aubin, J.-P. ***Applied functional analysis***. John Wiley & Sons, 2000.

Berlinet, A. and Thomas-Agnan, C. ***Reproducing Kernel Hilbert Spaces in Probability and Statistics***. Springer, 2011.

Berrett, T. B. and Samworth, R. J. Nonparametric independence testing via mutual information. ***Biometrika***, 106 (3):547–566, 2019.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. ***Efficient and adaptive estimation for semiparametric models***, volume 4. Johns Hopkins University Press Baltimore, 1993.

Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. ***Foundations of Computational Mathematics***, 7(3):331–368, 2007.

Carmeli, C., De Vito, E., and Toigo, A. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. ***Analysis and Applications***, 4(04): 377–408, 2006.

Carmeli, C., De Vito, E., Toigo, A., and Umanitá, V. Vector valued reproducing kernel Hilbert spaces and universality. ***Analysis and Applications***, 8(01):19–61, 2010.

Chen, Z., Naslidnyk, M., and Briol, F.-X. Nested expectations with kernel quadrature. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J. (eds.), ***Proceedings of the 42nd International Conference on Machine Learning***, volume 267, pp. 8760–8793. PMLR, 13–19 Jul 2025.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. ***The Econometrics Journal***, 21(1), 2018.

Chwialkowski, K., Sejdinovic, D., and Gretton, A. A wild bootstrap for degenerate kernel tests. In ***NeurIPS***, 2014.

De Vito, E., Rosasco, L., and Caponnetto, A. Discretization error analysis for tikhonov regularization. ***Analysis and Applications***, 4(01):81–99, 2006.

Efron, B. Bootstrap methods: another look at the jackknife. In ***Breakthroughs in statistics: Methodology and distribution***, pp. 569–593. Springer, 1992.

Fischer, S. and Steinwart, I. Sobolev norm learning rates for regularized least-squares algorithms. ***Journal of Machine Learning Research***, 21(205):1–38, 2020.

Gretton, A. A simpler condition for consistency of a kernel independence test. ***arXiv preprint arXiv:1501.06103***, 2015.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In ***International conference on algorithmic learning theory***, pp. 63–77. Springer, 2005.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. A kernel statistical test of independence. ***Advances in neural information processing systems***, 20, 2007.

He, Z., Pogodin, R., Li, Y., Deka, N., Gretton, A., and Sutherland, D. J. On the hardness of conditional independence testing in practice. ***arXiv preprint arXiv:2512.14000***, 2025.

Heller, R., Heller, Y., and Gorfine, M. A consistent multivariate test of association based on ranks of distances. ***Biometrika***, 100(2):503–510, 2013.

Hines, O., Dukes, O., Diaz-Ordaz, K., and Vansteelandt, S. Demystifying statistical learning based on efficient influence functions. ***The American Statistician***, 76(3): 292–304, 2022.

Kennedy, E. H. Semiparametric doubly robust targeted double machine learning: a review. ***Handbook of statistical methods for precision medicine***, pp. 207–236, 2024.

Lee, A. J. ***U-Statistics: Theory and Practice***. Routledge, 1 edition, 1990. doi: 10.1201/9780203734520. URL https://doi.org/10.1201/9780203734520.

Li, Y., Pogodin, R., Sutherland, D. J., and Gretton, A. Self-supervised learning with kernel dependence maximization. ***Advances in Neural Information Processing Systems***, 34:15543–15556, 2021.

Li, Z., Meunier, D., Mollenhauer, M., and Gretton, A. Optimal rates for regularized conditional mean embedding learning. In ***Advances in Neural Information Processing Systems***, volume 35, pp. 4433–4445, 2022.

Li, Z., Meunier, D., Mollenhauer, M., and Gretton, A. Towards optimal sobolev norm rates for the vector-valued regularized least-squares algorithm. ***Journal of Machine Learning Research***, 25(181):1–51, 2024.

Lin, J. and Cevher, V. Optimal distributed learning with multi-pass stochastic gradient methods. In ***International Conference on Machine Learning***, pp. 3092–3101. PMLR, 2018.

Lin, J., Rudi, A., Rosasco, L., and Cevher, V. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. ***Applied and Computational Harmonic Analysis***, 48(3):868–890, 2020.

Luedtke, A. Simplifying debiased inference via automatic differentiation and probabilistic programming. ***Journal of the Royal Statistical Society Series B: Statistical Methodology***, 2025. doi: 10.1093/jrsssb/qkaf052. URL https://par.nsf.gov/biblio/10649160.

Luedtke, A. and Chung, I. One-step estimation of differentiable Hilbert-valued parameters. ***The Annals of Statistics***, 52(4):1534 – 1563, 2024. doi: 10.1214/24-AOS2403. URL https://doi.org/10.1214/24-AOS2403.

Luedtke, A., Carone, M., and van der Laan, M. J. An omnibus non-parametric test of equality in distribution for unknown functions. ***Journal of the Royal Statistical Society Series B: Statistical Methodology***, 81(1):75–99, 2019.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. Distinguishing cause from effect using observational data: Methods and benchmarks. ***Journal of Machine Learning Research***, 17(32):1–102, 2016.

Morzywolek, P., Gilbert, P. B., and Luedtke, A. Inference on local variable importance measures for heterogeneous treatment effects. ***arXiv preprint arXiv:2510.18843***, 2025.

Park, J. and Muandet, K. A measure-theoretic approach to kernel conditional mean embeddings. ***Advances in neural information processing systems***, 33:21247–21259, 2020.

Peters, J., Mooij, J., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. ***Journal of Machine Learning Research***, 15, 09 2013. doi: 10.15496/publikation-1672.

Pogodin, R., Schrab, A., Li, Y., Sutherland, D. J., and Gretton, A. Practical kernel tests of conditional independence. ***arXiv preprint arXiv:2402.13196***, 2024.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In ***Advances in Neural Information Processing Systems***, volume 20, pp. 1177–1184, 2007.

Schick, A. On asymptotically efficient estimation in semiparametric models. ***The Annals of Statistics***, pp. 1139–1151, 1986.

Schrab, A., Kim, I., Guedj, B., and Gretton, A. Efficient aggregated kernel tests using incomplete u-statistics. In ***Advances in Neural Information Processing Systems***, 2022.

Serfling, R. J. ***Approximation theorems of mathematical statistics***. John Wiley & Sons, 2009.

Shekhar, S., Kim, I., and Ramdas, A. A permutation-free kernel independence test. ***Journal of Machine Learning Research***, 24(369):1–68, 2023.

Shen, H., Jegelka, S., and Gretton, A. Fast kernel-based independent component analysis. ***IEEE Transactions on Signal Processing***, 57:3498–3511, 2009.

Smale, S. and Zhou, D.-X. Shannon sampling and function reconstruction from point values. ***Bulletin of the American Mathematical Society***, 41(3):279–305, 2004.

Smale, S. and Zhou, D.-X. Shannon sampling II: Connections to learning theory. ***Applied and Computational Harmonic Analysis***, 19(3):285–302, 2005.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A hilbert space embedding for distributions. In ***International conference on algorithmic learning theory***, pp. 13–31. Springer, 2007.

Song, L., Smola, A., Gretton, A., Borgwardt, K. M., and Bedo, J. Supervised feature selection via dependence estimation. In ***Proceedings of the 24th international conference on Machine learning***, pp. 823–830, 2007.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. Hilbert space embeddings and metrics on probability measures. ***Journal of Machine Learning Research***, 11:1517–1561, 2010.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. Universality, characteristic kernels and RKHS embedding of measures. ***Journal of Machine Learning Research***, 12(Jul):2389–2410, 2011.

Steinwart, I. and Christmann, A. ***Support vector machines***. Springer Science & Business Media, 2008.

Steinwart, I. and Scovel, C. Mercer's theorem on general domains: On the interaction between measures, kernels, and rkhss. ***Constructive Approximation***, 35(3):363–417, 2012.

Tsybakov, A. B. ***Introduction to nonparametric estimation***. Springer, 2009.

van der Vaart, A. W. ***Asymptotic statistics***, volume 3 of ***Cambridge Series in Statistical and Probabilistic Mathematics***. Cambridge university press, 2000.

Wang, W., Gong, M., Huang, B., Bailey, J., Han, B., Zhang, K., and Liu, F. Practical kernel selection for kernel-based conditional independence test. In ***The Thirty-ninth Annual Conference on Neural Information Processing Systems***, 2025.

Williams, C. K. I. and Seeger, M. Using the nyström method to speed up kernel machines. In ***Advances in Neural Information Processing Systems***, volume 13, pp. 682–688. MIT Press, 2001.

Williamson, B. D., Gilbert, P. B., Simon, N. R., and Carone, M. A general framework for inference on algorithm-agnostic variable importance. ***Journal of the American Statistical Association***, 118(543):1645–1658, 2023.

Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. Large-scale kernel methods for independence testing. ***Statistics and Computing***, 28(1):113–130, 2018.

A. Guide to appendix

In this section, we provide a guide to the contents of the appendix. At the top of each section, we provide a table listing the notation used in that section, with the exception of Section B, where we provide a table for notation used across the paper.

Sample splitting allows us to avoid restrictive Donsker-type complexity restrictions on the function class, and significantly simplifies proofs. We refer the reader to Kennedy (2024, Section 4.2) for an exposition on sample splitting and cross-fitting. The main drawback of the sample-splitting estimator is that it only uses half of the total sample for evaluation, so its asymptotic variance is two times larger than the semiparametric efficiency bound for the full sample. The virtue of the sample-splitting estimator is that it is significantly less notationally burdensome to present, as there is no need to perform bookkeeping for the fold indices. In order to strike a balance between pedagogical clarity and generality, we opted to present the sample-splitting estimator, as it already reveals the essential features of our debiased procedure and the nuisance requirements (Requirements 1B and 1C).

As explained in Remark 1, by swapping the role of evaluation and fitting samples, we can recover full statistical efficiency by using a K-fold cross-fitted estimator. The prototypical example is the 2-fold cross-fitted estimator. It is standard in theoretical works on semiparametric efficiency theory to present guarantees for the 2-fold cross-fitted estimator, as the extension of the proofs to the K-fold cross-fitted one presents no essential difficulty while requiring more burdensome notation, see e.g. Luedtke & Chung (2024, Section 2.5). Furthermore, the asymptotic variance of the 2-fold cross-fitted estimator already attains the semiparametric efficiency bound.

The contents of the appendix can therefore be summarized as follows. In Section B, we introduce additional background material. In Section C we prove Theorem 1 in the main text, in the more general setting where Assumption 1 may not hold. Our results therein reduce immediately to Theorem 1 when Assumption 1 does in fact hold.

In Section D, we generalize the guarantees in Luedtke & Chung (2024) for generic Hilbert-valued parameters to a two-fold cross-fitted variant of the estimator presented in Eq. (1). This allows us to exhibit the required rates on the nuisance functions, and show that they are exactly given by Requirements 1B and 1C, with a straightforward notational extension from the sample-splitting to the cross-fitting case.

The K-fold cross-fitted estimator often enjoys superior finite-sample performance to the 2-fold cross-fitted estimator, and as a result, in all of our empirical simulations we use a K-fold cross-fitted estimator for $K \geq 2$. To support this and to ensure reproducibility of our research, we derive the closed-form solution for the K-fold cross-fitted estimator in Section E, where we are fully explicit about how to perform tensor slice operations that let us perform cross-fitting in full generality.

In Section F, we provide supplementary theoretical results to the two methods of constructing confidence intervals proposed in Section 3.2 in the main text, for the general K-fold cross-fitting setting. In particular, we provide a closed-form estimator $\widehat{\sigma}^2$ for the asymptotic variance in the delta method in the K-fold cross-fitting case, fully reusing expressions computed in Section E. We then establish the consistency of $\widehat{\sigma}^2$. In Section H, we construct the data-adaptive diagnostic for the trustworthiness of the delta method introduced in Section 3.2. In Section I, we provide a detailed introduction to statistical learning theory for vector-valued kernel ridge regression, which allows us to obtain convergence rates for the nuisance $\widehat{v}_j$, and thereby show when Requirements 1B and 1C is satisfied by more fundamental assumptions on Sobolev smoothness.

In Section G, we provide all experimental results not included in the main text and the experimental configuration.

B. Notation and background

Notation	Meaning
$O = (X, W, Y), o_i = (x_i, w_i, y_i)$	Observation and realized observation.
$\mathcal{M}$, P_0, P	Nonparametric model, true law, and generic law. Except for H_0, subscript 0 always denotes an object indexed by the true data generating process.
m_P, ξ_P	Regression function $m_P(w) = \mathbb{E}_P[Y \mid W = w]$ and $\xi_P(w, y) = y - m_P(w)$.
$k, l; \mathcal{H}_k, \mathcal{H}_l, \mathcal{H}_{kl}$	Reproducing kernels on $\mathcal{X}$ and $\mathbb{R}^{d_y}$; corresponding reproducing kernel Hilbert spaces; tensor-product RKHS $\mathcal{H}_k \otimes \mathcal{H}_l$.
φ_k, φ_l, $\varphi_{\xi,P}$	Canonical feature maps, with $\varphi_{\xi,P}(w, y) = \varphi_l\{\xi_P(w, y)\}$.
$\mu_{X,P}$, $\mu_{\xi,P}$	Kernel mean embeddings $\mathbb{E}_P[\varphi_k(X)]$ and $\mathbb{E}_P[\varphi_{\xi,P}(W, Y)]$.
$\widetilde{\varphi}_{k,P}$, $\widetilde{\varphi}_{\xi,P}$	Centered feature maps obtained by subtracting the corresponding kernel mean embeddings.
$\Psi(P)$, Ψ_0	RKHS cross-covariance operator between X and the residual $\xi_P(W, Y)$; $\Psi_0 = \Psi(P_0)$.
HSIC	Signal magnitude $\theta_0 = \|\Psi_0\|_{\mathcal{H}_{kl}}$; its square is HSIC.
$\partial_j \varphi_l$	Derivative of the residual feature map with respect to residual coordinate j.
$\widehat{\cdot}$, $\widetilde{\cdot}$, $\cdot^{(-k)}$	Estimated object, cross-fitted object, and object fitted without fold k.

This table records only notation used across the paper. Section-specific notation is introduced again, locally, in the appendix section where it is used.

B.1. Background on semiparametric efficiency theory

Pathwise differentiability for Hilbert-valued parameters. Let $\mathcal{P}$ be a collection of distributions defined on a common complete metric space $(\mathcal{Z}, \mathcal{B})$. Let $\mathcal{H}$ denote a reproducing kernel Hilbert space over a set $\mathcal{T}$ and let $\nu : \mathcal{P} \to \mathcal{H}$ be a parameter whose value is to be estimated. We let the feature map of $\mathcal{H}$ be denoted as ϕ. The parameter ν is said to be pathwise differentiable at $P \in \mathcal{P}$ if and only if there exists a continuous linear operator $\dot{\nu}_P : \dot{\mathcal{P}}_P \to \mathcal{H}$ such that for all parametric submodels $\{P_\epsilon : \epsilon \in [0, \delta)\}$ with score function s and centred at P with $P_0 = P$, we have

$$\|\nu(P_\epsilon) - \nu(P) - \epsilon \dot{\nu}_P(s)\|_{\mathcal{H}} = o(\epsilon).$$

The operator $\dot{\nu}_P$ is called the local parameter of ν at P and its Hermitian adjoint $\dot{\nu}_P^*$ is called the efficient influence operator. We refer the reader to Luedtke & Chung (2024, Section 2.2) and Section B.1 in the Appendix for a full definition of a parametric submodel and score function.

Hilbert-valued one-step estimator. In the notation of the above paragraph, we define $\phi_P : \mathcal{Z} \to \mathcal{H}$ as follows for each $t \in \mathcal{T}$:

$$\phi_P(z)(t) = \dot{\nu}_P^*(\phi(t))(z), \qquad P\text{-a.s. } z.$$

If ϕ_P is square integrable in the sense of $\|\phi_P\|_{L^2(P;\mathcal{H})} < \infty$, then we say that ϕ_P is the EIF of ν. In this case, we can construct a one-step estimator of the form $\nu(\widehat{P}) + P_n \phi_{\widehat{P}}$, where P_n denotes the sample average and $\widehat{P}$ is an initial data-driven estimate of the data-generating distribution P_0.

B.2. Background on reproducing kernel Hilbert spaces

Assumption 5 (Regularity assumptions on l)**.** *The kernel l satisfies the following conditions.*

(K1) l is symmetric, positive definite, and bounded on the diagonal:

$$\kappa_0^2 := \sup_{u \in \mathbb{R}^d} l(u, u) < \infty.$$

(K2) For every pair of multi-indices $\alpha, \beta \in \mathbb{N}_0^d$ with $|\alpha| \le 2$ and $|\beta| \le 2$, the derivative $\partial_1^\alpha \partial_2^\beta l$ exists and is continuous on $\mathbb{R}^d \times \mathbb{R}^d$.

(K3) The first and second diagonal derivative seminorms are finite:

$$\kappa_1^2 := \sup_{u \in \mathbb{R}^d} \sup_{\|a\|=1} \partial_{1,a}\partial_{2,a} l(u,u) < \infty,$$
$$\kappa_2^2 := \sup_{u \in \mathbb{R}^d} \sup_{\|a\|=\|b\|=1} \partial_{1,a}\partial_{1,b}\partial_{2,a}\partial_{2,b} l(u,u) < \infty.$$

Derivatives with respect to the first and second arguments of l are denoted by ∂_1 and ∂_2. For multi-indices $\alpha, \beta \in \mathbb{N}_0^d$, write

$$\partial_1^\alpha \partial_2^\beta l(u,v) := \frac{\partial^{|\alpha|+|\beta|}}{\partial u_1^{\alpha_1} \cdots \partial u_d^{\alpha_d} \partial v_1^{\beta_1} \cdots \partial v_d^{\beta_d}} l(u,v),$$

whenever this derivative exists. For directions $a, b, c, e \in \mathbb{R}^d$, write, for example,

$$\partial_{1,a}\partial_{2,b} l(u,v) := \sum_{j,r=1}^{d} a_j b_r \, \partial_{1,j}\partial_{2,r} l(u,v).$$

Lemma 1 (Feature-map consequences)**.** *Let $\phi(u) := l(\cdot, u)$ be the canonical feature map into $\mathcal{H}_l$. Under Assumption 5, ϕ is twice continuously Fréchet differentiable and, for all $u, v \in \mathbb{R}^d$ and $a, b \in \mathbb{R}^d$,*

$$\langle D\phi(u)[a], D\phi(v)[b] \rangle_{\mathcal{H}_l} = \partial_{1,a}\partial_{2,b} l(u,v),$$

and

$$\left\langle D^2\phi(u)[a,b], D^2\phi(v)[a,b] \right\rangle_{\mathcal{H}_l} = \partial_{1,a}\partial_{1,b}\partial_{2,a}\partial_{2,b} l(u,v).$$

Consequently,

$$\sup_u \|\phi(u)\|_{\mathcal{H}_l} \le \kappa_0, \qquad \sup_u \|D\phi(u)\|_{\mathrm{op}} \le \kappa_1, \qquad \sup_u \|D^2\phi(u)\|_{\mathrm{op}} \le \kappa_2.$$

Hence, for all $u, v, h \in \mathbb{R}^d$,

$$\|\phi(u) - \phi(v)\|_{\mathcal{H}_l} \le \kappa_1 \|u - v\|,$$
$$\|D\phi(u) - D\phi(v)\|_{\mathrm{op}} \le \kappa_2 \|u - v\|,$$

and

$$\|\phi(u+h) - \phi(u) - D\phi(u)[h]\|_{\mathcal{H}_l} \le \frac{\kappa_2}{2}\|h\|^2.$$

In coordinates,

$$\langle \partial_j \phi(u), \partial_r \phi(v) \rangle_{\mathcal{H}_l} = \partial_{1,j}\partial_{2,r} l(u,v).$$

Proof. The differentiability and inner-product identities follow from the RKHS differentiability theorem for kernels with continuous mixed derivatives, applied up to order two. The bounds are obtained by evaluating the identities at $u = v$:

$$\|D\phi(u)[a]\|_{\mathcal{H}_l}^2 = \partial_{1,a}\partial_{2,a} l(u,u) \le \kappa_1^2 \|a\|^2,$$

and

$$\|D^2\phi(u)[a,b]\|_{\mathcal{H}_l}^2 = \partial_{1,a}\partial_{1,b}\partial_{2,a}\partial_{2,b} l(u,u) \le \kappa_2^2 \|a\|^2 \|b\|^2.$$

The Lipschitz and Taylor bounds follow from the Banach-valued fundamental theorem of calculus. □

Proposition 1 (Gaussian and Sobolev kernels satisfy Assumption 5)**.** *The following standard kernels satisfy Assumption 5.*

(i) For the Gaussian kernel

$$l_\sigma(u,v) = \exp\{-\|u-v\|^2/(2\sigma^2)\},$$

one may take

$$\kappa_0 = 1, \qquad \kappa_1 = \sigma^{-1}, \qquad \kappa_2 = \sqrt{3}\,\sigma^{-2}.$$

(ii) For the Sobolev/Bessel-potential kernel

$$l_{s,\rho}(u,v) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{i\omega^\top(u-v)}(1+\rho^2\|\omega\|^2)^{-s}\,\mathrm{d}\omega,$$

Assumption 5 holds whenever

$$s > d/2 + 2.$$

Equivalently, Matérn kernels with smoothness $\nu > 2$ satisfy Assumption 5 when their spectral density is proportional to

$$(\lambda^2 + \|\omega\|^2)^{-(\nu+d/2)}.$$

Proof. For the Gaussian kernel, smoothness and boundedness are immediate. Direct differentiation gives, for unit a, b,

$$\partial_{1,a}\partial_{2,a} l_\sigma(u,u) = \sigma^{-2},$$

and

$$\partial_{1,a}\partial_{1,b}\partial_{2,a}\partial_{2,b} l_\sigma(u,u) = \sigma^{-4}\{1 + 2(a^\top b)^2\} \le 3\sigma^{-4}.$$

For the Sobolev kernel, positive definiteness follows from its nonnegative spectral density. Derivatives up to order two in each argument are obtained by differentiating under the integral. This is justified whenever

$$\int_{\mathbb{R}^d} \|\omega\|^4(1+\rho^2\|\omega\|^2)^{-s}\,\mathrm{d}\omega < \infty,$$

which is equivalent to $s > d/2 + 2$. The same fourth-moment bound gives $\kappa_1 < \infty$ and $\kappa_2 < \infty$. For the Matérn family, $s = \nu + d/2$, so the condition is $\nu > 2$. □

C. Efficient influence function of the cross-covariance operator

We prove Theorem 1 in the main text. In this work, this section is the only section in which we do not impose Assumption 1. This is because it is essentially for free to compute the EIF in the more general setting where X need not be W-measurable. At the end of this section, we derive the simplification when X is W-measurable, which recovers the statement of Theorem 1.

Proof of Theorem 1 in the general setting where we drop Assumption 1. Following a standard convention in semiparametric statistical theory, we use the subscript 0 to denote objects indexed by the true distribution P_0. We have

$$\begin{aligned}
&\mathbb{E}_P\left[\widetilde{\phi}_{X,P}(X) \otimes \widetilde{\varphi}_{\xi,P}(W,Y)\right] \\
&= \mathbb{E}_P[(\varphi_k(X) - \mathbb{E}_P[\varphi_k(X')]) \otimes (\varphi_{\xi,P}(W,Y) - \mathbb{E}_P[\varphi_{\xi,P}(W',Y')])] \\
&= \mathbb{E}_P[\varphi_k(X) \otimes \varphi_{\xi,P}(W,Y)] - \mathbb{E}_P[\varphi_k(X)] \otimes \mathbb{E}_P[\varphi_{\xi,P}(W,Y)].
\end{aligned}$$

We define the following RKHS-valued parameters on $\mathcal{M}$,

$$\begin{aligned}
\Psi_1(P) &:= \mathbb{E}_P[\varphi_k(X) \otimes \varphi_l(Y - \mathbb{E}_P[Y \mid W])] = \mathbb{E}_P[\varphi_k(X) \otimes \varphi_{\xi,P}(W,Y)] \\
\Psi_2(P) &:= \mathbb{E}_P[\varphi_k(X)] \\
\Psi_3(P) &:= \mathbb{E}_P[\varphi_l(Y - \mathbb{E}_P[Y \mid W])] = \mathbb{E}_P[\varphi_{\xi,P}(W,Y)] \\
\Psi(P) &= \Psi_1(P) - \Psi_2(P) \otimes \Psi_3(P).
\end{aligned}$$

We write $\Psi_1(P)$ using the computation graph of primitives

$$\begin{aligned}
\theta_1 : \mathcal{M} \times \emptyset &\to L^2(P_0; \mathbb{R}^{d_y}) \\
(P,) &\mapsto ((x, w, y) \mapsto y) \\
\theta_2 : \mathcal{M} \times L^2(P_0; \mathbb{R}^{d_y}) &\to L^2(P_{0,W}; \mathbb{R}^{d_y}) \\
(P, h_1) &\mapsto \mathbb{E}_P[h_1(X, W, Y) \mid W = w] \\
\theta_3 : \mathcal{M} \times L^2(P_{0,W}; \mathbb{R}^{d_y}) &\to L^2(P_0; \mathbb{R}^{d_y}) \\
(P, h_2) &\mapsto ((x, w, y) \mapsto h_2(w)) \\
\theta_4 : \mathcal{M} \times L^2(P_0; \mathbb{R}^{d_y}) &\to L^2(P_0; \mathbb{R}^{d_y}) \\
(P, h_3) &\mapsto ((x, w, y) \mapsto y - h_3(x, w, y)) \\
\theta_5 : \mathcal{M} \times L^2(P_0; \mathbb{R}^{d_y}) &\to L^2(P_0; \mathcal{H}_{kl}) \\
(P, h_4) &\mapsto ((x, w, y) \mapsto \varphi_k(x) \otimes \varphi_l(h_4(x, w, y))) \\
\theta_6 : \mathcal{M} \times L^2(P_0; \mathcal{H}_{kl}) &\to \mathcal{H}_{kl} \\
(P, h_5) &\mapsto \mathbb{E}_P[h_5(X, W, Y)].
\end{aligned}$$

We let $h_1 = \theta_1(P,)$, $h_2 = \theta_2(P, h_1)$, $h_3 = \theta_3(P, h_2)$, $h_4 = \theta_4(P, h_3)$, $h_5 = \theta_5(P, h_4)$, $h_6 = \theta_6(P, h_5)$. Concretely, we have

$$\begin{aligned}
h_1(x, w, y) &= y \\
h_2(w) &= \mathbb{E}_P[Y \mid W = w] \\
h_3(x, w, y) &= h_2(w) = \mathbb{E}_P[Y \mid W = w] \\
h_4(x, w, y) &= y - \mathbb{E}_P[Y \mid W = w] \\
h_5(x, w, y) &= \varphi_k(x) \otimes \varphi_l(y - \mathbb{E}_P[Y \mid W = w]) \\
h_6 &= \mathbb{E}_P\left[\varphi_k(X) \otimes \varphi_l(Y - \mathbb{E}_P[Y \mid W])\right].
\end{aligned}$$

We have $\Psi_1(P) = h_6$. Note that all of the above primitives are contained in Luedtke (2025, Table 1). They are, respectively, ***constant map***, ***conditional mean***, ***lift to new domain***, ***pointwise operation***, ***pointwise operation*** and ***marginal mean***. For $\Psi_1(P)$, we initialize the adjoint variables $f_i = 0$ for $0 \leq i \leq 5$ and let $f_6 \in \mathcal{H}_{kl}$ be user-specified. We follow the notation convention in Luedtke & Chung (2024) by using subscript 0 to denote objects associated with the true data-generating process. We let $\dot{\mathcal{M}}_0$ denote the tangent space of the statistical model $\mathcal{M}$ at P_0, which we assume to be the space $L_0^2(P_0)$ of mean-zero square integrable functions. This corresponds to a nonparametric statistical model. We now describe the backpropagation procedure.

1. $j = 6$. h_6 is a ***marginal mean***. We have that the pathwise derivative of θ_6 at (P_0, h_5) is given by

$$D[\theta_6](P_0, h_5) : \dot{\mathcal{M}}_0 \oplus L^2(P_0; \mathcal{H}_{kl}) \to \mathcal{H}_{kl}.$$

   We backpropagate f_6 through θ_6 to f_0, f_5 as follows:

$$\begin{pmatrix} f_0 \\ f_5 \end{pmatrix} += D[\theta_6](P_0, h_5)^*(f_6) = \begin{pmatrix} (x, w, y) \mapsto \langle f_6, h_5(x, w, y) - \mathbb{E}_0[h_5(X, W, Y)] \rangle_{\mathcal{H}_{kl}} \\ (x, w, y) \mapsto f_6 \end{pmatrix}.$$

2. $j = 5$. We have

$$D[\theta_5](P_0, h_4) : \dot{\mathcal{M}}_0 \oplus L^2(P_0; \mathbb{R}^{d_y}) \to L^2(P_0; \mathcal{H}_{kl}).$$

   We backpropagate f_5 through θ_5 to f_4 as follows:

$$\begin{pmatrix} f_0 \\ f_4 \end{pmatrix} += D[\theta_5](P_0, h_4)^*(f_5) = \begin{pmatrix} 0 \\ (x, w, y) \mapsto (\langle f_5(x, w, y), \varphi_k(x) \otimes (\partial_j \varphi_l)(h_4(x, w, y)) \rangle_{\mathcal{H}_{kl}})_{j=1}^{d_y} \end{pmatrix},$$

   where $\partial_j \varphi_l$ refers to the jth-entry of the gradient of $\varphi_l : \mathbb{R}^{d_y} \to \mathcal{H}_l$.

3. $j = 4$. We have

$$D[\theta_4](P_0, h_3) : \dot{\mathcal{M}}_0 \oplus L^2(P_0; \mathbb{R}^{d_y}) \to L^2(P_0; \mathbb{R}^{d_y}).$$

We backpropagate f_4 through θ_4 to f_3 as follows:

$$\begin{pmatrix} f_0 \\ f_3 \end{pmatrix} += D[\theta_4](P_0, h_3)^*(f_4) = \begin{pmatrix} 0 \\ (x, w, y) \mapsto -f_4(x, w, y) \end{pmatrix}.$$

4. $j = 3$. We have

$$D[\theta_3](P_0, h_2) : \dot{\mathcal{M}}_0 \oplus L^2(P_{0,W}; \mathbb{R}^{d_y}) \to L^2(P_0; \mathbb{R}^{d_y}).$$

We backpropagate f_3 through θ_3 to f_2 as follows:

$$\begin{pmatrix} f_0 \\ f_2 \end{pmatrix} += D[\theta_3](P_0, h_2)^*(f_3) = \begin{pmatrix} 0 \\ w \mapsto \mathbb{E}_0[f_3(X, W, Y) \mid W = w] \end{pmatrix}.$$

5. $j = 2$. We have

$$D[\theta_2](P_0, h_1) : \dot{\mathcal{M}}_0 \oplus L^2(P_0; \mathbb{R}^{d_y}) \to L^2(P_{0,W}; \mathbb{R}^{d_y}).$$

We backpropagate f_2 through θ_2 to f_1 as follows:

$$\begin{pmatrix} f_0 \\ f_1 \end{pmatrix} += D[\theta_2](P_0, h_1)^*(f_2) = \begin{pmatrix} (x, w, y) \mapsto f_2(w)^\top (h_1(x, w, y) - \mathbb{E}_0[h_1(X, W, Y) \mid W = w]) \\ (x, w, y) \mapsto f_2(w) \end{pmatrix}.$$

6. $j = 1$. We have

$$D[\theta_1](P_0) : \dot{\mathcal{M}}_0 \to L^2(P_0; \mathbb{R}^{d_y}).$$

We backpropagate f_1 through θ_1 to f_0 as follows:

$$f_0 += D[\theta_1](P_0)^*(f_1) = 0.$$

We have

$$\begin{aligned}
f_5(x, w, y) &= f_6 \\
f_4(x, w, y) &= (\langle f_6, \varphi_k(x) \otimes (\partial_j \varphi_l)(h_4(x, w, y))\rangle_{\mathcal{H}_{kl}})_{j=1}^{d_y} \\
f_3(x, w, y) &= -(\langle f_6, \varphi_k(x) \otimes (\partial_j \varphi_l)(h_4(x, w, y))\rangle_{\mathcal{H}_{kl}})_{j=1}^{d_y} \\
f_2(w) &= (-\mathbb{E}_0 [\langle f_6, \varphi_k(X) \otimes (\partial_j \varphi_l)(h_4(X, W, Y))\rangle_{\mathcal{H}_{kl}} \mid W = w])_{j=1}^{d_y} \\
f_1(x, w, y) &= (-\mathbb{E}_0 [\langle f_6, \varphi_k(X) \otimes (\partial_j \varphi_l)(h_4(X, W, Y))\rangle_{\mathcal{H}_{kl}} \mid W = w])_{j=1}^{d_y}.
\end{aligned}$$

Finally, summing the adjoint contributions for f_0 yields

$$\begin{aligned}
f_0(x, w, y) &= \langle f_6, h_5(x, w, y) - \mathbb{E}_0[h_5(X, W, Y)]\rangle_{\mathcal{H}_{kl}} \\
&\quad + f_2(w)^\top (h_1(x, w, y) - \mathbb{E}_0[h_1(X, W, Y) \mid W = w]) \\
&= \langle f_6, \varphi_k(x) \otimes \varphi_l(y - \mathbb{E}_0[Y \mid W = w]) - \mathbb{E}_0 [\varphi_k(X) \otimes \varphi_l(Y - \mathbb{E}_0[Y \mid W])]\rangle_{\mathcal{H}_{kl}} \\
&\quad - \sum_{j=1}^{d_y} \mathbb{E}_0 [\langle f_6, \varphi_k(X) \otimes (\partial_j \varphi_l)(Y - \mathbb{E}_0[Y \mid W])\rangle_{\mathcal{H}_{kl}} \mid W = w] (y_j - \mathbb{E}_0[Y_j \mid W = w]) \\
&= \langle f_6, \varphi_k(x) \otimes \varphi_l(y - \mathbb{E}_0[Y \mid W = w]) - \mathbb{E}_0 [\varphi_k(X) \otimes \varphi_l(Y - \mathbb{E}_0[Y \mid W])]\rangle_{\mathcal{H}_{kl}} \\
&\quad - \mathbb{E}_0 [\langle f_6, \varphi_k(X) \otimes (\nabla \varphi_l)(Y - \mathbb{E}_0[Y \mid W])\rangle_{\mathcal{H}_{kl}} \mid W = w]^\top (y - \mathbb{E}_0[Y \mid W = w]),
\end{aligned}$$

where $(\nabla\varphi_l)(Y - \mathbb{E}_0[Y \mid W])$ denotes a tuple in $\mathcal{H}_l^{d_y}$, and the inner product with f_6 is understood componentwise.

Recall we defined $m_0(w) = \mathbb{E}_0[Y \mid W = w]$, $\xi_0(w, y) = y - m_0(w)$, and $\varphi_{\xi,0}(w, y) = \varphi_l\{\xi_0(w, y)\}$. We write $\mu_{X,0} := \mathbb{E}_0[\varphi_k(X)]$ and $\mu_{\xi,0} := \mathbb{E}_0[\varphi_{\xi,0}(W, Y)]$. For each $j = 1, \ldots, d_y$, define

$$\begin{aligned} v_j(w) &:= \mathbb{E}_0[(\partial_j\varphi_l)(\xi_0(W, Y)) \mid W = w] \in \mathcal{H}_l, \\ a_j(w) &:= \mathbb{E}_0[\varphi_k(X) \otimes (\partial_j\varphi_l)(\xi_0(W, Y)) \mid W = w] \in \mathcal{H}_{kl}. \end{aligned}$$

By Luedtke (2025, Theorem 1), $\Psi_1(P)$ is pathwise differentiable at P_0, and its efficient influence operator $\dot{\Psi}^*_{1,0} : \mathcal{H}_{kl} \to \dot{\mathcal{M}}_0$ is given by

$$\begin{aligned} \dot{\Psi}^*_{1,0}(f)(x, w, y) &= \langle f, \varphi_k(x) \otimes \varphi_{\xi,0}(w, y) - \mathbb{E}_0[\varphi_k(X) \otimes \varphi_{\xi,0}(W, Y)]\rangle_{\mathcal{H}_{kl}} \\ &\quad - \sum_{j=1}^{d_y} \xi_{0,j}(w, y)\langle f, a_j(w)\rangle_{\mathcal{H}_{kl}} \\ &= \langle f, \Gamma_1(x, w, y)\rangle_{\mathcal{H}_{kl}}, \end{aligned}$$

where we define

$$\Gamma_1(x, w, y) := \varphi_k(x) \otimes \varphi_{\xi,0}(w, y) - \mathbb{E}_0[\varphi_k(X) \otimes \varphi_{\xi,0}(W, Y)] - \sum_{j=1}^{d_y} \xi_{0,j}(w, y)a_j(w).$$

Note that $\Psi_2(P)$ is a ***kernel mean embedding*** (see Luedtke (2025, C.2.5)), since

$$\Psi_2(P) = \int_{\mathcal{X}} k_X(x, \cdot)P_X(dx).$$

By Luedtke (2025, Lemma S11), $\Psi_2(P)$ is pathwise differentiable at P_0, and its efficient influence operator $\dot{\Psi}^*_{2,0} : \mathcal{H}_k \to \dot{\mathcal{M}}_0$ is given by

$$\dot{\Psi}^*_{2,0}(f)(x, w, y) = f(x) - \mathbb{E}_0[f(X)] = \langle f, \Gamma_2(x)\rangle_{\mathcal{H}_k},$$

where we define

$$\Gamma_2(x) := \varphi_k(x) - \mu_{X,0}.$$

Note that $\Psi_3(P)$ can be obtained by modifying the primitive θ_5 of $\Psi_1(P)$ to

$$\begin{aligned} \widetilde{\theta}_5 : \mathcal{M} \times L^2(P_0; \mathbb{R}^{d_y}) &\to L^2(P_0; \mathcal{H}_l) \\ (P, h_4) &\mapsto ((x, w, y) \mapsto \varphi_l(h_4(x, w, y))). \end{aligned}$$

We conclude that $\Psi_3(P)$ is pathwise differentiable at P_0, with efficient influence operator $\dot{\Psi}^*_{3,0} : \mathcal{H}_l \to \dot{\mathcal{M}}_0$ given by

$$\begin{aligned} \dot{\Psi}^*_{3,0}(f)(x, w, y) &= \langle f, \varphi_{\xi,0}(w, y) - \mu_{\xi,0}\rangle_{\mathcal{H}_l} - \sum_{j=1}^{d_y} \xi_{0,j}(w, y)\langle f, v_j(w)\rangle_{\mathcal{H}_l} \\ &= \langle f, \Gamma_3(x, w, y)\rangle_{\mathcal{H}_l}, \end{aligned}$$

where we define

$$\Gamma_3(x, w, y) := \varphi_{\xi,0}(w, y) - \mu_{\xi,0} - \sum_{j=1}^{d_y} \xi_{0,j}(w, y)v_j(w).$$

By product rule, for $s \in L_0^2(P_0)$, we have

$$\dot{\Psi}_0[s] = \dot{\Psi}_{1,0}[s] - \dot{\Psi}_{2,0}[s] \otimes \mu_{\xi,0} - \mu_{X,0} \otimes \dot{\Psi}_{3,0}[s] \in \mathcal{H}_{kl}.$$

Let $f \in \mathcal{H}_{kl}$ be arbitrary. We have

$$\begin{aligned}\langle f, \dot{\Psi}_0[s]\rangle_{\mathcal{H}_{kl}} &= \langle \dot{\Psi}_0^* f, s\rangle_{L^2(P_0)} \\ &= \langle f, \dot{\Psi}_{1,0}[s] - \dot{\Psi}_{2,0}[s] \otimes \mu_{\xi,0} - \mu_{X,0} \otimes \dot{\Psi}_{3,0}[s]\rangle_{\mathcal{H}_{kl}}.\end{aligned}$$

We have, for all elementary tensors $u \otimes v \in \mathcal{H}_{kl}$,

$$\begin{aligned}\langle u \otimes v, \dot{\Psi}_{2,0}[s] \otimes \mu_{\xi,0}\rangle_{\mathcal{H}_{kl}} &= \langle u, \dot{\Psi}_{2,0}[s]\rangle_{\mathcal{H}_k} \langle v, \mu_{\xi,0}\rangle_{\mathcal{H}_l} \\ &= \langle \dot{\Psi}_{2,0}^* u, s\rangle_{L^2(P_0)} \langle v, \mu_{\xi,0}\rangle_{\mathcal{H}_l} \\ &= \langle \dot{\Psi}_{2,0}^* (\mathrm{Id}_{\mathcal{H}_k} \otimes_{\mathrm{op}} \mu_{\xi,0}^*)(u \otimes v), s\rangle_{L^2(P_0)}.\end{aligned}$$

Hence, extending by linearity and continuity, we have for all $f \in \mathcal{H}_{kl}$,

$$\begin{aligned}\langle f, \dot{\Psi}_{2,0}[s] \otimes \mu_{\xi,0}\rangle_{\mathcal{H}_{kl}} &= \langle \dot{\Psi}_{2,0}^* (\mathrm{Id}_{\mathcal{H}_k} \otimes_{\mathrm{op}} \mu_{\xi,0}^*) f, s\rangle_{L^2(P_0)} \\ \langle f, \mu_{X,0} \otimes \dot{\Psi}_{3,0}[s]\rangle_{\mathcal{H}_{kl}} &= \langle \dot{\Psi}_{3,0}^* (\mu_{X,0}^* \otimes_{\mathrm{op}} \mathrm{Id}_{\mathcal{H}_l}) f, s\rangle_{L^2(P_0)}.\end{aligned}$$

Therefore,

$$\begin{aligned}\dot{\Psi}_0^*(f)(x,w,y) &= \dot{\Psi}_{1,0}^*(f)(x,w,y) - \dot{\Psi}_{2,0}^*(\mathrm{Id}_{\mathcal{H}_k} \otimes_{\mathrm{op}} \mu_{\xi,0}^*) f(x,w,y) - \dot{\Psi}_{3,0}^*(\mu_{X,0}^* \otimes_{\mathrm{op}} \mathrm{Id}_{\mathcal{H}_l}) f(x,w,y) \\ &= \langle f, \Gamma_1(x,w,y)\rangle_{\mathcal{H}_{kl}} - \langle (\mathrm{Id}_{\mathcal{H}_k} \otimes_{\mathrm{op}} \mu_{\xi,0}^*) f, \Gamma_2(x)\rangle_{\mathcal{H}_k} - \langle (\mu_{X,0}^* \otimes_{\mathrm{op}} \mathrm{Id}_{\mathcal{H}_l}) f, \Gamma_3(x,w,y)\rangle_{\mathcal{H}_l} \\ &= \langle f, \Gamma_1(x,w,y) - (\mathrm{Id}_{\mathcal{H}_k} \otimes_{\mathrm{op}} \mu_{\xi,0}^*)^* \Gamma_2(x) - (\mu_{X,0}^* \otimes_{\mathrm{op}} \mathrm{Id}_{\mathcal{H}_l})^* \Gamma_3(x,w,y)\rangle_{\mathcal{H}_{kl}}.\end{aligned}$$

Assuming that $\Gamma_1(x,w,y) - (\mathrm{Id}_{\mathcal{H}_k} \otimes_{\mathrm{op}} \mu_{\xi,0}^*)^* \Gamma_2(x) - (\mu_{X,0}^* \otimes_{\mathrm{op}} \mathrm{Id}_{\mathcal{H}_l})^* \Gamma_3(x,w,y)$ belongs to $L_0^2(P_0; \mathcal{H}_{kl})$, the parameter Ψ has an EIF at P_0 in the sense of Luedtke & Chung (2024), given by

$$\phi_0(x,w,y) = \Gamma_1(x,w,y) - (\mathrm{Id}_{\mathcal{H}_k} \otimes_{\mathrm{op}} \mu_{\xi,0}^*)^* \Gamma_2(x) - (\mu_{X,0}^* \otimes_{\mathrm{op}} \mathrm{Id}_{\mathcal{H}_l})^* \Gamma_3(x,w,y).$$

We compute

$$\begin{aligned}(\mathrm{Id}_{\mathcal{H}_k} \otimes_{\mathrm{op}} \mu_{\xi,0}^*)^* \Gamma_2(x) &= (\mathrm{Id}_{\mathcal{H}_k} \otimes_{\mathrm{op}} \mu_{\xi,0}^*)^* (\varphi_k(x) - \mu_{X,0}) \\ &= (\varphi_k(x) - \mu_{X,0}) \otimes \mu_{\xi,0} \\ &= \widetilde{\phi}_{X,0}(x) \otimes \mu_{\xi,0},\end{aligned}$$

where the second line can be verified by a simple adjoint identity. Similarly, we have

$$\begin{aligned}(\mu_{X,0}^* \otimes_{\mathrm{op}} \mathrm{Id}_{\mathcal{H}_l})^* \Gamma_3(x,w,y) &= (\mu_{X,0}^* \otimes_{\mathrm{op}} \mathrm{Id}_{\mathcal{H}_l})^* \left(\varphi_{\xi,0}(w,y) - \mu_{\xi,0} - \sum_{j=1}^{d_y} \xi_{0,j}(w,y) v_j(w) \right) \\ &= \mu_{X,0} \otimes \left(\varphi_{\xi,0}(w,y) - \mu_{\xi,0} - \sum_{j=1}^{d_y} \xi_{0,j}(w,y) v_j(w) \right).\end{aligned}$$

Hence we have

$$\begin{aligned}\mathcal{H}_{kl} \ni \phi_0(x,w,y) = \varphi_k(x) \otimes \varphi_{\xi,0}(w,y) - \mathbb{E}_0[\varphi_k(X) \otimes \varphi_{\xi,0}(W,Y)] - \sum_{j=1}^{d_y} \xi_{0,j}(w,y) a_j(w) \\ - \widetilde{\phi}_{X,0}(x) \otimes \mu_{\xi,0} - \mu_{X,0} \otimes \left(\varphi_{\xi,0}(w,y) - \mu_{\xi,0} - \sum_{j=1}^{d_y} \xi_{0,j}(w,y) v_j(w) \right).\end{aligned}$$

Define

$$\begin{aligned} F_j(w) &:= a_j(w) - \mu_{X,0} \otimes v_j(w) \\ &= \mathbb{E}_0[\{\varphi_k(X) - \mu_{X,0}\} \otimes (\partial_j \varphi_l)(\xi_0(W,Y)) \mid W = w] \in \mathcal{H}_{kl}. \end{aligned}$$

Since

$$\Psi_0 = \mathbb{E}_0[\varphi_k(X) \otimes \varphi_{\xi,0}(W,Y)] - \mu_{X,0} \otimes \mu_{\xi,0},$$

we obtain

$$\phi_0(x,w,y) = \widetilde{\phi}_{X,0}(x) \otimes (\varphi_{\xi,0}(w,y) - \mu_{\xi,0}) - \sum_{j=1}^{d_y} \xi_{0,j}(w,y) F_j(w) - \Psi_0.$$

□

D. Asymptotic linearity of the two-fold cross-fitted estimator

Fix $P \in \mathcal{M}$. Recall that

$$m_P(w) = \mathbb{E}_P[Y \mid W = w], \qquad \xi_P(w,y) = y - m_P(w),$$

and

$$\mu_{X,P} = \mathbb{E}_P[\varphi_k(X)], \qquad \mu_{\xi,P} = \mathbb{E}_P[\varphi_l\{\xi_P(W,Y)\}].$$

For $j = 1, \ldots, d_y$, define

$$v_{P,j}(w) \doteq \mathbb{E}_P[(\partial_j \varphi_l)\{\xi_P(W,Y)\} \mid W = w] \in \mathcal{H}_l, \qquad v_P = (v_{P,1}, \ldots, v_{P,d_y}).$$

Set

$$\eta_P = (m_P, v_P, \mu_{X,P}, \mu_{\xi,P}).$$

For $\eta = (m, v, \mu_X, \mu_\xi)$, define

$$D_\eta(x,w,y) \doteq \{\varphi_k(x) - \mu_X\} \otimes \left[\varphi_l\{y - m(w)\} - \mu_\xi - \sum_{j=1}^{d_y} \{y_j - m_j(w)\} v_j(w)\right].$$

Under Assumption 1, the EIF in Theorem 1 satisfies

$$\phi_0(o) = D_{\eta_0}(o) - \Psi_0.$$

Let I_1, I_2 be an equal split of $[n]$, and define

$$\mathbb{P}_{n,r} \doteq \frac{2}{n} \sum_{i \in I_r} \delta_{o_i}, \qquad r = 1, 2.$$

Let

$$\widehat{\eta}_r = (\widehat{m}_r, \widehat{v}_r, \widehat{\mu}_{X,r}, \widehat{\mu}_{\xi,r}),$$

be fitted on I_r, where

$$\widehat{\mu}_{X,r} = \mathbb{P}_{n,r} \varphi_k(X), \qquad \widehat{\mu}_{\xi,r} = \mathbb{P}_{n,r} \varphi_l\{Y - \widehat{m}_r(W)\}.$$

The two-fold cross-fitted estimator is

$$\check{\Psi}_n \doteq \frac{1}{2}\{\mathbb{P}_{n,1} D_{\widehat{\eta}_2} + \mathbb{P}_{n,2} D_{\widehat{\eta}_1}\}.$$

Lemma 2 (Foldwise L^2-consistency of the estimated non-centered EIF)**.** *Suppose Assumptions 1 to 4 and requirement 1A hold. Let* $\widehat{\eta}_r = (\widehat{m}_r, \widehat{v}_r, \widehat{\mu}_{X,r}, \widehat{\mu}_{\xi,r})$, $r = 1, 2$, *and define*

$$D_\eta(x, w, y) \doteq \{\varphi_k(x) - \mu_X\} \otimes \left[\varphi_l\{y - m(w)\} - \mu_\xi - \sum_{j=1}^{d_y}\{y_j - m_j(w)\}v_j(w)\right].$$

If, for $r = 1, 2$,

$$\|\widehat{m}_r - m_0\|_{L^\infty(P_{0,W})} + \|\widehat{m}_r - m_0\|_{L^2(P_{0,W})} + \|\widehat{v}_r - v_0\|_{L^2(P_{0,W})} = o_p(1),$$

then

$$\|\widehat{\mu}_{X,r} - \mu_{X,0}\|_{\mathcal{H}_k} = O_p(n^{-1/2}), \qquad \|\widehat{\mu}_{\xi,r} - \mu_{\xi,0}\|_{\mathcal{H}_l} = o_p(1),$$

and

$$\|D_{\widehat{\eta}_r} - D_{\eta_0}\|_{L^2(P_0;\mathcal{H}_{kl})} = o_p(1).$$

Proof. The bound for $\widehat{\mu}_{X,r}$ follows from Assumption 2. For $\widehat{\mu}_{\xi,r}$,

$$\begin{aligned}\|\widehat{\mu}_{\xi,r} - \mu_{\xi,0}\|_{\mathcal{H}_l} \leq\ & \|\mathbb{P}_{n,r}\{\varphi_l(Y - \widehat{m}_r(W)) - \varphi_l(Y - m_0(W))\}\|_{\mathcal{H}_l} \\ & + \|\mathbb{P}_{n,r}\varphi_l(Y - m_0(W)) - P_0\varphi_l(Y - m_0(W))\|_{\mathcal{H}_l}.\end{aligned}$$

The first term is bounded by $L\|\widehat{m}_r - m_0\|_{L^\infty(P_{0,W})}$, and the second is $o_p(1)$ by Assumption 2. Hence $\|\widehat{\mu}_{\xi,r} - \mu_{\xi,0}\|_{\mathcal{H}_l} = o_p(1)$.

Write

$$A_\eta(X) \doteq \varphi_k(X) - \mu_X, \qquad M_\eta(O) \doteq \varphi_l\{Y - m(W)\} - \mu_\xi - \sum_{j=1}^{d_y}\{Y_j - m_j(W)\}v_j(W).$$

Then

$$D_{\widehat{\eta}_r} - D_{\eta_0} = (A_{\widehat{\eta}_r} - A_{\eta_0}) \otimes M_{\eta_0} + A_{\widehat{\eta}_r} \otimes (M_{\widehat{\eta}_r} - M_{\eta_0}).$$

By Assumptions 2 and 3 and requirement 1A, $M_{\eta_0} \in L^2(P_0; \mathcal{H}_l)$. Therefore

$$\|(A_{\widehat{\eta}_r} - A_{\eta_0}) \otimes M_{\eta_0}\|_{L^2(P_0)} \lesssim \|\widehat{\mu}_{X,r} - \mu_{X,0}\|_{\mathcal{H}_k} = o_p(1).$$

Moreover, $\|A_{\widehat{\eta}_r}\|_{\mathcal{H}_k} \leq 2\kappa_k$, and

$$\begin{aligned}\|M_{\widehat{\eta}_r} - M_{\eta_0}\|_{L^2(P_0)} \lesssim\ & \|\widehat{m}_r - m_0\|_{L^2(P_{0,W})} + \|\widehat{\mu}_{\xi,r} - \mu_{\xi,0}\|_{\mathcal{H}_l} \\ & + \|\xi_0^\top(\widehat{v}_r - v_0)\|_{L^2(P_0)} + \|(\widehat{m}_r - m_0)^\top\widehat{v}_r\|_{L^2(P_0)}.\end{aligned}$$

The third term is $o_p(1)$ by Requirement 1A and $\|\widehat{v}_r - v_0\|_{L^2(P_{0,W})} = o_p(1)$. The last term is $o_p(1)$, since $\widehat{v}_r = v_0 + o_p(1)$ in $L^2(P_{0,W})$ and $v_0 \in L^\infty(P_{0,W}; \mathcal{H}_l^{d_y})$ by Assumption 3. Thus $\|M_{\widehat{\eta}_r} - M_{\eta_0}\|_{L^2(P_0)} = o_p(1)$, proving the claim. □

Lemma 3 (Upper bound for the remainder term in each fold)**.** *Suppose Assumptions 1 to 4 hold. For* $\eta = (m, v, \mu_X, \mu_\xi)$, *define*

$$R_0(\eta) \doteq P_0 D_\eta - \Psi_0.$$

Then, for $r = 1, 2$,

$$\begin{aligned}\|R_0(\widehat{\eta}_r)\|_{\mathcal{H}_{kl}} \lesssim\ & \|\widehat{\mu}_{X,r} - \mu_{X,0}\|_{\mathcal{H}_k}\|\widehat{\mu}_{\xi,r} - \mu_{\xi,0}\|_{\mathcal{H}_l} \\ & + \|\widehat{m}_r - m_0\|^2_{L^2(P_{0,W})} \\ & + \|\widehat{m}_r - m_0\|_{L^2(P_{0,W})}\|\widehat{v}_r - v_0\|_{L^2(P_{0,W})}.\end{aligned}$$

Consequently, if

$$\|\widehat{m}_r - m_0\|^2_{L^2(P_{0,W})} + \|\widehat{m}_r - m_0\|_{L^2(P_{0,W})}\|\widehat{v}_r - v_0\|_{L^2(P_{0,W})} = o_p(n^{-1/2}),$$

then $R_0(\widehat{\eta}_r) = o_p(n^{-1/2})$.

Proof. Let $\delta = m - m_0$, $\xi_0 = Y - m_0(W)$, and $A_{\mu_X}(W) = \varphi_k\{\pi(W)\} - \mu_X$. By Assumption 1, $A_{\mu_X}(X) = A_{\mu_X}(W)$ almost surely. Using $\mathbb{E}_0[\xi_0 \mid W] = 0$,

$$\begin{aligned} R_0(\eta) = &(\mu_{X,0} - \mu_X) \otimes (\mu_{\xi,0} - \mu_\xi) \\ &+ P_0\left[A_{\mu_X}(W) \otimes \{\varphi_l(\xi_0 - \delta(W)) - \varphi_l(\xi_0)\}\right] \\ &+ P_0\left[A_{\mu_X}(W) \otimes \sum_{j=1}^{d_y} \delta_j(W) v_j(W)\right]. \end{aligned}$$

Add and subtract $\sum_{j=1}^{d_y} \delta_j(W)(\partial_j \varphi_l)(\xi_0)$. Since

$$v_{0,j}(W) = \mathbb{E}_0[(\partial_j \varphi_l)(\xi_0) \mid W],$$

the linear terms combine to

$$P_0\left[A_{\mu_X}(W) \otimes \sum_{j=1}^{d_y} \delta_j(W)\{v_j(W) - v_{0,j}(W)\}\right].$$

The Taylor remainder is bounded by Assumption 4:

$$\left\| \varphi_l(\xi_0 - \delta) - \varphi_l(\xi_0) + \sum_{j=1}^{d_y} \delta_j(\partial_j \varphi_l)(\xi_0) \right\|_{\mathcal{H}_l} \le L_2 \|\delta\|^2_{\mathbb{R}^{d_y}}.$$

Since $\|A_{\mu_X}(W)\|_{\mathcal{H}_k} \le 2\kappa_k$, Cauchy-Schwarz gives

$$\|R_0(\eta)\|_{\mathcal{H}_{kl}} \lesssim \|\mu_{X,0} - \mu_X\|_{\mathcal{H}_k} \|\mu_{\xi,0} - \mu_\xi\|_{\mathcal{H}_l} + \|m - m_0\|^2_{L^2(P_{0,W})} + \|m - m_0\|_{L^2(P_{0,W})} \|v - v_0\|_{L^2(P_{0,W})}.$$

Taking $\eta = \widehat{\eta}_r$ gives the result. The final statement follows from $\|\widehat{\mu}_{X,r} - \mu_{X,0}\|_{\mathcal{H}_k} = O_p(n^{-1/2})$ and $\|\widehat{\mu}_{\xi,r} - \mu_{\xi,0}\|_{\mathcal{H}_l} = o_p(1)$, as shown in Lemma 2. □

Analysis of the cross-fitted one-step estimator.

Theorem 3 (Two-fold cross-fitted asymptotic linearity)**.** *Let $o_1, \ldots, o_n \sim P_0^n$, and let $\check{\Psi}_n$ be defined above. Suppose Assumptions 1 to 4 and requirement 1A hold. In place of Requirements 1B and 1C, assume that, for $r = 1, 2$,*

$$\|\widehat{m}_r - m_0\|_{L^\infty(P_{0,W})} + \|\widehat{m}_r - m_0\|_{L^2(P_{0,W})} + \|\widehat{v}_r - v_0\|_{L^2(P_{0,W})} = o_p(1),$$

and

$$\|\widehat{m}_r - m_0\|^2_{L^2(P_{0,W})} + \|\widehat{m}_r - m_0\|_{L^2(P_{0,W})} \|\widehat{v}_r - v_0\|_{L^2(P_{0,W})} = o_p(n^{-1/2}).$$

Then

$$\check{\Psi}_n - \Psi_0 = \mathbb{P}_n \phi_0 + o_p(n^{-1/2}),$$

where $\phi_0 = D_{\eta_0} - \Psi_0$ is the EIF in Theorem 1. Hence

$$\sqrt{n}(\check{\Psi}_n - \Psi_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi_0(o_i) + o_p(1) \rightsquigarrow \mathbb{H} \quad \text{in } \mathcal{H}_{kl},$$

where $\mathbb{H}$ is a mean-zero Gaussian element in $\mathcal{H}_{kl}$ with

$$\mathrm{Var}\{\langle \mathbb{H}, h\rangle_{\mathcal{H}_{kl}}\} = \mathbb{E}_0[\langle \phi_0(O), h\rangle^2_{\mathcal{H}_{kl}}], \qquad h \in \mathcal{H}_{kl}.$$

Proof of Theorem 3. Set $R_0(\eta) \doteq P_0 D_\eta - \Psi_0$. Since $D_{\eta_0} - \Psi_0 = \phi_0$,

$$\begin{aligned} \check{\Psi}_n - \Psi_0 = &\mathbb{P}_n \phi_0 \\ &+ \frac{1}{2}(\mathbb{P}_{n,1} - P_0)(D_{\widehat{\eta}_2} - D_{\eta_0}) + \frac{1}{2}(\mathbb{P}_{n,2} - P_0)(D_{\widehat{\eta}_1} - D_{\eta_0}) \\ &+ \frac{1}{2}\{R_0(\widehat{\eta}_1) + R_0(\widehat{\eta}_2)\}. \end{aligned}$$

By Lemma 2 and sample splitting, we can control the stochastic equicontinuity term as follows

$$\frac{1}{2}(\mathbb{P}_{n,1} - P_0)(D_{\widehat{\eta}_2} - D_{\eta_0}) + \frac{1}{2}(\mathbb{P}_{n,2} - P_0)(D_{\widehat{\eta}_1} - D_{\eta_0}) = o_p(n^{-1/2}).$$

By Lemma 3,

$$R_0(\widehat{\eta}_1) + R_0(\widehat{\eta}_2) = o_p(n^{-1/2}).$$

Therefore

$$\check{\Psi}_n - \Psi_0 = \mathbb{P}_n \phi_0 + o_p(n^{-1/2}).$$

The weak convergence to $\mathbb{H}$ follows by applying a Hilbert-valued CLT, using the fact that $\mathcal{H}_{kl}$ is separable and $\phi_0 \in L_0^2(P_0; \mathcal{H}_{kl})$. □

E. Computation

Notation	Meaning
$\widehat{A}_i, \widehat{C}_i, \widehat{M}_i$	Centered X feature, centered residual feature, and corrected residual feature in the sample-split formulas.
$\widehat{D}_i, G_{ii'}$	Tensor-product one-step summand and its Gram entry.
$\mathcal{D}^k, \mathcal{D}^{-k}, n_k$	Fold k, its complement, and fold size in the cross-fitted computation.
$\mathbb{P}_k, \mathbb{P}_k^{(b)}$	Empirical law on fold k and its bootstrap resample.
$z = (w, y)$	Abbreviation used for the residual-side arguments.
K, $L^{(-k_1,-k_2)}$	Stored Gram matrices for the X kernel and fold-specific residual kernel.
$r_X, r_\xi, r_{\xi,v}, r_v$	Stored centered X term, residual-feature term, residual–v cross term, and v-diagonal term.
∂L, ΔL	Stored first- and mixed-second-derivative tensors for the residual kernel.
Ξ^{-k}, W^{-k}	Stored residual-coordinate tensor and vvKRR weight tensor.
$\mathfrak{w}_j^{-k}$	Scalar vvKRR weight for label coordinate j trained outside fold k.

We first explain the computations on the simpler setting instantiated in Section 3. Then we extend this example to the cross-fitted estimators, where the computations are described in more detail. Since our experiments focus on the setting where X is W-measurable (Assumption 1), we describe this procedure to facilitate the reproduction of our results, which follow this exact computational scheme. The extension to general W primarily differs in the need to estimate the $\mathcal{H}_{kl}$-valued F_j instead of the $\mathcal{H}_l$-valued v_j, which further leads to slightly different expansions of the inner products analysed below.

E.1. Sample-split computations

For $i \in I_1 \cup I_2$, define $\widehat{\xi}_i \doteq y_i - \widehat{m}(w_i)$. The residuals $\widehat{\xi}_i$ are evaluated for all $i \in I_1 \cup I_2$. The mean embeddings are computed on I_2:

$$\widehat{\mu}_X \doteq \frac{2}{n} \sum_{t \in I_2} \varphi_k(x_t), \qquad \widehat{\mu}_\xi \doteq \frac{2}{n} \sum_{t \in I_2} \varphi_l(\widehat{\xi}_t).$$

Thus $\widehat{\xi}_i$ is used both at evaluation points $i \in I_1$ and at centering points $t \in I_2$. Let $\widehat{v}_j$, $j = 1, \ldots, d_y$, be the remaining nuisance estimates fitted on I_2. For $i \in I_1$, define

$$\widehat{A}_i \doteq \varphi_k(x_i) - \widehat{\mu}_X, \qquad \widehat{C}_i \doteq \varphi_l(\widehat{\xi}_i) - \widehat{\mu}_\xi, \qquad \widehat{M}_i \doteq \widehat{C}_i - \sum_{j=1}^{d_y} \widehat{\xi}_{i,j} \widehat{v}_j(w_i).$$

Then

$$\widehat{D}_i \doteq \widehat{D}(o_i) = \widehat{A}_i \otimes \widehat{M}_i.$$

Hence, the tensor-product inner product factorizes:

$$G_{ii'} \doteq \langle \widehat{D}_i, \widehat{D}_{i'} \rangle_{\mathcal{H}_{kl}} = \langle \widehat{A}_i, \widehat{A}_{i'} \rangle_{\mathcal{H}_k} \langle \widehat{M}_i, \widehat{M}_{i'} \rangle_{\mathcal{H}_l}.$$

Thus $\widehat{Q}_V = \|\widehat{\Psi}_n\|^2_{\mathcal{H}_{kl}} = \frac{4}{n^2} \sum_{i,i' \in I_1} G_{ii'}$ can be computed using only inner products in $\mathcal{H}_k$ and $\mathcal{H}_l$. The residual-side inner product is

$$\left\langle \widehat{M}_i, \widehat{M}_{i'} \right\rangle_{\mathcal{H}_l} = \left\langle \widehat{C}_i, \widehat{C}_{i'} \right\rangle_{\mathcal{H}_l} - \sum_{j=1}^{d_y} \widehat{\xi}_{i',j} \left\langle \widehat{C}_i, \widehat{v}_j(w_{i'}) \right\rangle_{\mathcal{H}_l}$$
$$- \sum_{j=1}^{d_y} \widehat{\xi}_{i,j} \left\langle \widehat{v}_j(w_i), \widehat{C}_{i'} \right\rangle_{\mathcal{H}_l} + \sum_{j,r=1}^{d_y} \widehat{\xi}_{i,j} \widehat{\xi}_{i',r} \left\langle \widehat{v}_j(w_i), \widehat{v}_r(w_{i'}) \right\rangle_{\mathcal{H}_l}.$$

The first factor is the centered k-Gram entry

$$\left\langle \widehat{A}_i, \widehat{A}_{i'} \right\rangle_{\mathcal{H}_k} = k(x_i, x_{i'}) - \frac{2}{n} \sum_{t \in I_2} k(x_i, x_t) - \frac{2}{n} \sum_{t \in I_2} k(x_{i'}, x_t) + \frac{4}{n^2} \sum_{s,t \in I_2} k(x_s, x_t).$$

Similarly,

$$\left\langle \widehat{C}_i, \widehat{C}_{i'} \right\rangle_{\mathcal{H}_l} = l(\widehat{\xi}_i, \widehat{\xi}_{i'}) - \frac{2}{n} \sum_{t \in I_2} l(\widehat{\xi}_i, \widehat{\xi}_t) - \frac{2}{n} \sum_{t \in I_2} l(\widehat{\xi}_{i'}, \widehat{\xi}_t) + \frac{4}{n^2} \sum_{s,t \in I_2} l(\widehat{\xi}_s, \widehat{\xi}_t).$$

It remains to compute the terms involving $\widehat{v}_j$. Let $R_{j,t} \doteq (\partial_j \varphi_l)(\widehat{\xi}_t), t \in I_2$. The vector-valued KRR estimator has the representer form (Carmeli et al., 2006)

$$\widehat{v}_j(w) = \sum_{t \in I_2} \omega_{j,t}(w) R_{j,t},$$

where, writing $Q = (q(w_s, w_t))_{s,t \in I_2}$ and $q_w = (q(w, w_t))_{t \in I_2}$, $\omega_j(w)^\top = q_w^\top \left(Q + \frac{n}{2}\lambda_{j,n} I\right)^{-1}$. Therefore,

$$\left\langle \widehat{C}_i, \widehat{v}_j(w) \right\rangle_{\mathcal{H}_l} = \sum_{t \in I_2} \omega_{j,t}(w) \left[\partial_{2,j} l(\widehat{\xi}_i, \widehat{\xi}_t) - \frac{2}{n} \sum_{s \in I_2} \partial_{2,j} l(\widehat{\xi}_s, \widehat{\xi}_t) \right],$$

where $\partial_{2,j} l$ denotes the derivative of l with respect to the jth coordinate of its second argument. Finally, $\langle \widehat{v}_j(w), \widehat{v}_r(w') \rangle_{\mathcal{H}_l} = \sum_{s,t \in I_2} \omega_{j,s}(w) \omega_{r,t}(w') \partial_{1,j} \partial_{2,r} l(\widehat{\xi}_s, \widehat{\xi}_t)$. Thus the statistic can be computed from evaluations of k, l, first derivatives $\partial_{2,j} l$, mixed second derivatives $\partial_{1,j} \partial_{2,r} l$, and the scalar KRR weight matrices associated with the kernel q.

E.2. Cross-fitted estimators and confidence sets.

We move on to explain how to compute the point estimates and confidence sets in full generality for K-fold cross-fitted estimators.

Usage of superscripts and subscripts. For an object that depends on estimated nuisances, we always use ***superscripts*** to denote the fold used to fit the nuisances. In addition, we always use ***subscripts*** to denote data points at which an object is evaluated, or indices.

Usage of colour. Uncoloured symbols denote input/data, red denotes intermediate mathematical objects, and green denotes quantities that are stored and actually computed. As a rule, we never express objects in red in terms of objects in green. For objects in red, we denote evaluation at data points as function evaluation, namely we use $(\cdot)$ in normal size. In contrast, we view objects in green as tensors, therefore we consider points at which they are evaluated as indices, which are written inside parentheses in subscript. When uncoloured data symbols and green stored tensors clash, green takes precedence.

The data are $\mathcal{D} = (X_i, Z_i := (W_i, Y_i))_{i \in [n]}$, split into K folds $\mathcal{D}^k$ of sizes n_k and with complements $\mathcal{D}^{-k}$.

Assumption 6 (Equal fold sizes)**.** *Assume K divides n and all folds are of size $n_k = n/K$.*

Remark 2. *This is assumed purely for simplicity of exposition. It suffices that fold sizes are such that there exist constants $c < C$ for which $c < \frac{n_k}{n} < C$ for all sufficiently large n.*

The cross-fitted estimator of Ψ_0 is

$$\check{\Psi}_n \doteq K^{-1} \sum_{k=1}^{K} \mathbb{P}_k \left(\widehat{D}^{-k} \right).$$

where $\mathbb{P}_k$ is the empirical law of $\mathcal{D}^k$ and $\widehat{D}^{-k}$ is fitted entirely using $\mathcal{D}^{-k}$. Further, we will denote $\mathbb{P}_k^{(b)}$ as empirical measures obtained by bootstrapping $\mathbb{P}_k$ (Efron, 1992). We shall denote $\mathbb{P}_k^{(0)} \doteq \mathbb{P}_k$. A minimal objective is to compute quantities of the form $\left\langle \mathbb{P}_{k_1}^{(b_1)} \widehat{D}^{-k_1}, \mathbb{P}_{k_2}^{(b_2)} \widehat{D}^{-k_2} \right\rangle_{\mathcal{H}_{kl}}$, since all quantities that we need to compute are expressed in terms of inner products of (possibly bootstrapped) estimators of Ψ_0.

WLOG we will work with $b_1 = b_2 = 0$ for the rest of the derivations. Write $z = (w, y)$. Recall that for a set of nuisances η,

$$A_\eta(x) \doteq \varphi_k(x) - \mu_X, \qquad M_\eta(z) \doteq \varphi_l(\xi(z)) - \mu_\xi - \sum_{j=1}^{d_y} \xi_j(z) v_j(w), \qquad D_\eta(x, z) = A_\eta(x) \otimes M_\eta(z).$$

Our estimator should be a plug-in version of these. That is,

$$\widehat{A}^{-k}(x) \doteq \varphi_k(x) - \widehat{\mu}_X^{-k}, \qquad \widehat{M}^{-k}(z) \doteq \varphi_l\left(\widehat{\xi}^{-k}(z)\right) - \widehat{\mu}_\xi^{-k} - \sum_{j=1}^{d_y} \widehat{\xi}_j^{-k}(z) \widehat{v}_j^{-k}(w).$$

$$\widehat{D}^{-k}(x, z) = \widehat{A}^{-k}(x) \otimes \widehat{M}^{-k}(z).$$

Quantities, such as,

$$\|\check{\Psi}_n\|_{\mathcal{H}_{kl}}^2 = \frac{1}{K^2} \sum_{k_1, k_2} \left\langle \mathbb{P}_{k_1} \widehat{D}^{-k_1}, \mathbb{P}_{k_2} \widehat{D}^{-k_2} \right\rangle_{\mathcal{H}_{kl}},$$

expand as averages of the following V-statistics,

$$\left\langle \mathbb{P}_{k_1} \widehat{D}^{-k_1}, \mathbb{P}_{k_2} \widehat{D}^{-k_2} \right\rangle_{\mathcal{H}_{kl}} = \frac{1}{n_{k_1} n_{k_2}} \sum_{(x_1, z_1) \in \mathcal{D}^{k_1}} \sum_{(x_2, z_2) \in \mathcal{D}^{k_2}} \left\langle \widehat{D}^{-k_1}(x_1, z_1), \widehat{D}^{-k_2}(x_2, z_2) \right\rangle_{\mathcal{H}_{kl}}. \tag{5}$$

We will explain how to compute those summands. Since $\langle f_1 \otimes g_1, f_2 \otimes g_2 \rangle_{\mathcal{H}_{kl}} = \langle f_1, f_2 \rangle_{\mathcal{H}_k} \langle g_1, g_2 \rangle_{\mathcal{H}_l}$ (Aubin, 2000), we isolate two kinds of terms out of these inner products.

E.3. X-terms

We define the empirical kernel mean embedding

$$\widehat{\mu}_X^{-k} \doteq \frac{1}{n - n_k} \sum_{x \in \mathcal{X}^{-k}} \varphi_k(x).$$

Thus, for any x_1, x_2, we have

$$\begin{aligned}
r_{X,(x_1,x_2)}^{(-k_1,-k_2)} &\doteq \left\langle \varphi_k(x_1) - \widehat{\mu}_X^{-k_1}, \varphi_k(x_2) - \widehat{\mu}_X^{-k_2} \right\rangle_{\mathcal{H}_k} \\
&= k(x_1,x_2) - \frac{1}{n-n_{k_1}} \sum_{x \in \mathcal{X}^{-k_1}} k(x,x_2) - \frac{1}{n-n_{k_2}} \sum_{x \in \mathcal{X}^{-k_2}} k(x,x_1) \\
&+ \frac{1}{(n-n_{k_1})(n-n_{k_2})} \sum_{\substack{x \in \mathcal{X}^{-k_1} \\ x' \in \mathcal{X}^{-k_2}}} k(x,x') \\
&= K_{(x_1,x_2)} - \frac{1}{n-n_{k_1}} 1_{n-n_{k_1}}^\top K_{(-k_1,x_2)} - \frac{1}{n-n_{k_2}} K_{(x_1,-k_2)} 1_{n-n_{k_2}} \\
&+ \frac{1}{(n-n_{k_1})(n-n_{k_2})} 1_{n-n_{k_1}}^\top K_{(-k_1,-k_2)} 1_{n-n_{k_2}},
\end{aligned}$$

where $K \in \mathbb{R}^{n \times n}$ denotes the full-data kernel matrix on X, namely $K_{(x_1,x_2)} = k(x_1,x_2)$, and 1_d denotes the $d \times 1$ vector of all ones, for any $d \geq 1$. $K_{(-k_1,x_2)} \in \mathbb{R}^{(n-n_{k_1}) \times 1}$ denotes the matrix of $k(x,x_2)$, where x are the X observations in fold $\mathcal{X}^{-k_1}$. The remaining notation is defined analogously.

E.4. ξ-terms

We continue to write $z = (w,y)$, e.g. $z_1 = (w_1, y_1)$.

Analogously to the X-setting, we define

$$\widehat{\mu}_\xi^{-k} \doteq \frac{1}{n-n_k} \sum_{z \in \mathcal{Z}^{-k}} \varphi_l \left\{ \widehat{\xi}^{-k}(z) \right\}.$$

For any z, we partition $\widehat{M}^{-k}(z) \in \mathcal{H}_l$ into two summands,

$$\varphi_l(\widehat{\xi}^{-k}(z)) - \widehat{\mu}_\xi^{-k}, \qquad -\sum_{j=1}^{d_y} \underbrace{\widehat{\xi}_j^{-k}(z)}_{\mathbb{R}} \underbrace{\widehat{v}_j^{-k}(w)}_{\mathcal{H}_l}.$$

That yields two kinds of "diagonal terms" and one type of "cross-term" in the inner product decomposition

$$\left\langle \widehat{M}^{-k_1}(z_1), \widehat{M}^{-k_2}(z_2) \right\rangle_{\mathcal{H}_l}.$$

φ_l-diagonal terms. Let $L^{(-k_1,-k_2)} \in \mathbb{R}^{n \times n}$, have entries

$$L_{(z_1,z_2)}^{(-k_1,-k_2)} \doteq \left\langle \varphi_l\left(\widehat{\xi}^{-k_1}(z_1)\right), \varphi_l\left(\widehat{\xi}^{-k_2}(z_2)\right) \right\rangle_{\mathcal{H}_l} = l\left(\widehat{\xi}^{-k_1}(z_1), \widehat{\xi}^{-k_2}(z_2)\right).$$

Then exactly like for the X-terms, we have

$$\begin{aligned}
r_{\xi,(z_1,z_2)}^{(-k_1,-k_2)} &\doteq \left\langle \varphi_l\left(\widehat{\xi}^{-k_1}(z_1)\right) - \widehat{\mu}_\xi^{-k_1}, \varphi_l\left(\widehat{\xi}^{-k_2}(z_2)\right) - \widehat{\mu}_\xi^{-k_2} \right\rangle_{\mathcal{H}_l} \\
&= L_{(z_1,z_2)}^{(-k_1,-k_2)} - \frac{1}{n-n_{k_1}} \mathbf{1}_{n-n_{k_1}}^\top L_{(-k_1,z_2)}^{(-k_1,-k_2)} - \frac{1}{n-n_{k_2}} L_{(z_1,-k_2)}^{(-k_1,-k_2)} \mathbf{1}_{n-n_{k_2}} \\
&+ \frac{1}{(n-n_{k_1})(n-n_{k_2})} \mathbf{1}_{n-n_{k_1}}^\top L_{(-k_1,-k_2)}^{(-k_1,-k_2)} \mathbf{1}_{n-n_{k_2}},
\end{aligned}$$

where $r_\xi^{(-k_1,-k_2)} \in \mathbb{R}^{n_{k_1} \times n_{k_2}}$.

Cross-terms. We define

$$r^{(-k_1,-k_2)}_{\xi,v,(z_1,z_2)} \doteq \left\langle \varphi_l\left(\widehat{\xi}^{-k_1}(z_1)\right) - \widehat{\mu}_\xi^{-k_1}, \sum_{j=1}^{d_y} \widehat{\xi}_j^{-k_2}(z_2)\widehat{v}_j^{-k_2}(w_2)\right\rangle_{\mathcal{H}_l}, \tag{6}$$

and an analogous term with $1 \leftrightarrow 2$. Suppose that $\widehat{v}_j^{-k}$ is the solution to a vector-valued kernel ridge regression with separable kernel (Li et al., 2024)

$$Q(w_1, w_2) = q(w_1, w_2)\mathrm{Id}_{\mathcal{H}_l}.$$

Then it can be written as

$$\widehat{v}_j^{-k}(w) = \sum_{z' \in \mathcal{Z}^{-k}} \mathfrak{w}_j^{-k}(z', w)(\partial_j \varphi_l)\left(\widehat{\xi}^{-k}(z')\right), \tag{7}$$

where $z' = (w', y')$ and

$$\mathfrak{w}_j^{-k}(z', w) = \left[\left(Q^{-k} + (n - n_k)\lambda_{j,n} I\right)^{-1} q^{-k}(w)\right]_{z'},$$

with

$$(Q^{-k})_{z',z''} = q(w', w''), \qquad (q^{-k}(w))_{z''} = q(w'', w).$$

Remark 3. *The vvKRR fit above is only one way to estimate v_j^{-k}. Any regression method can be used provided that the inner products appearing below can be computed. A particularly useful variant is to center the regression targets. Let*

$$S_j^{-k}(z') := (\partial_j \varphi_l)\left(\widehat{\xi}^{-k}(z')\right), \qquad \bar{S}_j^{-k} := \frac{1}{n - n_k} \sum_{z' \in \mathcal{Z}^{-k}} S_j^{-k}(z').$$

If vvKRR is fit to the centered labels $S_j^{-k}(z') - \bar{S}_j^{-k}$, we add the mean back:

$$\widehat{v}_j^{-k}(w) = \bar{S}_j^{-k} + \sum_{z' \in \mathcal{Z}^{-k}} \widetilde{\mathfrak{w}}_j^{-k}(z', w)\{S_j^{-k}(z') - \bar{S}_j^{-k}\}.$$

Equivalently,

$$\widehat{v}_j^{-k}(w) = \sum_{z' \in \mathcal{Z}^{-k}} \mathfrak{w}_{j,\mathrm{cen}}^{-k}(z', w) S_j^{-k}(z'),$$

where

$$\mathfrak{w}_{j,\mathrm{cen}}^{-k}(z', w) = \widetilde{\mathfrak{w}}_j^{-k}(z', w) + \frac{1 - \sum_{u \in \mathcal{Z}^{-k}} \widetilde{\mathfrak{w}}_j^{-k}(u, w)}{n - n_k}.$$

Thus the formulas below are unchanged after replacing $\mathfrak{w}_j^{-k}$ by $\mathfrak{w}_{j,\mathrm{cen}}^{-k}$. Centering is particularly helpful under the $\Psi_0 = 0$ null in a setting where $W = X$. In that case, $v_{0,j}$ is constant. The constant component is then estimated by the empirical mean and is not penalized by the ridge regression.

We define

$$\partial L^{(-k_1,-k_2)}_{j,(z_1,z_2)} = \left(\partial_j^1 l\right)\left(\widehat{\xi}^{-k_1}(z_1), \widehat{\xi}^{-k_2}(z_2)\right),$$

where $\partial L^{(-k_1,-k_2)} \in \mathbb{R}^{d_y \times n \times n}$ is a tensor of rank 3, and ∂_j^1 denotes the partial derivative with respect to the j-th component of the first $\mathbb{R}^{d_y}$-valued argument. It is not in general symmetric in its second and third coordinates. We also emphasize that ∂L should be understood as a single symbol. Substituting Eq. (7) into Eq. (6) yields

$$\begin{aligned}
&r^{(-k_1,-k_2)}_{\xi,v,(z_1,z_2)} \\
&= \left\langle \varphi_l\left(\widehat{\xi}^{-k_1}(z_1)\right) - \widehat{\mu}_\xi^{-k_1}, \sum_{j=1}^{d_y} \sum_{z' \in \mathcal{Z}^{-k_2}} \mathfrak{w}_j^{-k_2}(z', w_2)(\partial_j \varphi_l)\left(\widehat{\xi}^{-k_2}(z')\right)\widehat{\xi}_j^{-k_2}(z_2)\right\rangle_{\mathcal{H}_l} \\
&= \sum_{j=1}^{d_y} \sum_{z' \in \mathcal{Z}^{-k_2}} \mathfrak{w}_j^{-k_2}(z', w_2)\widehat{\xi}_j^{-k_2}(z_2)\left\langle (\partial_j \varphi_l)\left(\widehat{\xi}^{-k_2}(z')\right), \varphi_l\left(\widehat{\xi}^{-k_1}(z_1)\right) - \widehat{\mu}_\xi^{-k_1}\right\rangle_{\mathcal{H}_l} \\
&= \sum_{j=1}^{d_y} \sum_{z' \in \mathcal{Z}^{-k_2}} \mathfrak{w}_j^{-k_2}(z', w_2)\widehat{\xi}_j^{-k_2}(z_2)\left\{\partial L^{(-k_2,-k_1)}_{j,(z',z_1)} - \frac{1}{n - n_{k_1}} \partial L^{(-k_2,-k_1)\top}_{j,(z',-k_1)} 1_{n-n_{k_1}}\right\}.
\end{aligned}$$

Next, we define tensors storing the quantities needed to compute $\mathfrak{w}_j^{-k_2}(z', w_2)\widehat{\xi}_j^{-k_2}(z_2)$. Define a rank 3-tensor $W^{-k} \in \mathbb{R}^{d_y \times (n-n_k) \times n}$ and rank 2-tensor $\Xi^{-k} \in \mathbb{R}^{d_y \times n}$ by

$$\Xi_{j,z}^{-k} \doteq \widehat{\xi}_j^{-k}(z)$$
$$W_{j,(z',w)}^{-k} \doteq \mathfrak{w}_j^{-k}(z', w).$$

We can therefore write

$$r_{\xi,v,(z_1,z_2)}^{(-k_1,-k_2)} = \sum_{j=1}^{d_y} \sum_{z' \in \mathcal{Z}^{-k_2}} W_{j,(z',w_2)}^{-k_2} \Xi_{j,z_2}^{-k_2} \left\{ \partial L_{j,(z',z_1)}^{(-k_2,-k_1)} - \frac{1}{n-n_{k_1}} \partial L_{j,(z',-k_1)}^{(-k_2,-k_1)\top} 1_{n-n_{k_1}} \right\}.$$

The above tensor contraction can be performed using one line of Einstein summation in NumPy or JAX.

v-diagonal terms. We define the rank-4 tensor $\Delta L^{(-k_1,-k_2)} \in \mathbb{R}^{d_y \times d_y \times (n-n_{k_1}) \times (n-n_{k_2})}$ as follows

$$\Delta L_{j_1,j_2,(z',z'')}^{(-k_1,-k_2)} = \partial_{j_1}^1 \partial_{j_2}^2 l \left(\widehat{\xi}^{-k_1}(z'), \widehat{\xi}^{-k_2}(z'') \right).$$

Therefore we have

$$\begin{aligned} r_{v,(z_1,z_2)}^{(-k_1,-k_2)} &= \left\langle \sum_{j_1=1}^{d_y} \widehat{\xi}_{j_1}^{-k_1}(z_1)\widehat{v}_{j_1}^{-k_1}(w_1), \sum_{j_2=1}^{d_y} \widehat{\xi}_{j_2}^{-k_2}(z_2)\widehat{v}_{j_2}^{-k_2}(w_2) \right\rangle_{\mathcal{H}_l} \\ &= \sum_{j_1=1}^{d_y} \sum_{j_2=1}^{d_y} \Xi_{j_1,z_1}^{-k_1} \Xi_{j_2,z_2}^{-k_2} \sum_{\substack{z' \in \mathcal{Z}^{-k_1} \\ z'' \in \mathcal{Z}^{-k_2}}} W_{j_1,(z',w_1)}^{-k_1} W_{j_2,(z'',w_2)}^{-k_2} \Delta L_{j_1,j_2,(z',z'')}^{(-k_1,-k_2)}, \end{aligned}$$

which can again be computed using Einstein summation.

For consistency with Section 3, define the fold-pair Gram entry

$$G_{((x_1,z_1),(x_2,z_2))}^{(-k_1,-k_2)} \doteq r_{X,(x_1,x_2)}^{(-k_1,-k_2)} \left\{ r_{\xi,(z_1,z_2)}^{(-k_1,-k_2)} - r_{\xi,v,(z_1,z_2)}^{(-k_1,-k_2)} - r_{\xi,v,(z_2,z_1)}^{(-k_2,-k_1)} + r_{v,(z_1,z_2)}^{(-k_1,-k_2)} \right\}.$$

Putting everything together, we have

$$\left\langle \mathbb{P}_{k_1} \widehat{D}^{-k_1}, \mathbb{P}_{k_2} \widehat{D}^{-k_2} \right\rangle_{\mathcal{H}_{kl}} \tag{8}$$
$$= \frac{1}{n_{k_1} n_{k_2}} \sum_{\substack{(x_1,z_1) \in \mathcal{D}^{k_1} \\ (x_2,z_2) \in \mathcal{D}^{k_2}}} G_{((x_1,z_1),(x_2,z_2))}^{(-k_1,-k_2)}. \tag{9}$$

F. Inference

Notation	Meaning
$\mathcal{O}$, Ω_n, Ω_0	Class of positive self-adjoint operators; empirical and population choices for the operator used to construct a confidence set.
$w(h;\Omega)$, $\mathcal{C}_n(\zeta)$	Quadratic form and resulting Hilbert-space confidence set.
$\mathbb{H}$, $\mathbb{H}_n^{(b)}$	Gaussian limit and bootstrap empirical-process draw.
$\zeta_{1-\alpha}$, $\widehat{\zeta}_n$	Population and bootstrap-estimated quantiles for the chosen confidence-set geometry.
σ^2, $\widehat{\sigma}^2$	Delta-method variance component and its plug-in estimator.
$\check{\Psi}_n$	Cross-fitted one-step estimator of Ψ_0.
$\widehat{D}^{-k}$	Uncentered EIF estimate fitted without fold k.
D_{η_0}, ϕ_0	Population uncentered and centered EIFs, with $\phi_0 = D_{\eta_0} - \Psi_0$.
G_{ij}	Cross-fitted Gram entry $\langle \widehat{D}^{-k(i)}(O_i), \widehat{D}^{-k(j)}(O_j)\rangle_{\mathcal{H}_{kl}}$, where $k(i)$ is the validation fold containing O_i.
$\widehat{Q}_V$, $\widehat{Q}_U$	V-statistic and U-statistic center $\|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}}$ and its diagonal-term-free analogue.

We build on Luedtke & Chung (2024, Section 4.2) to construct $(1-\alpha)$-confidence sets centered at the one-step estimator. Let $\mathcal{O}$ denote the set of continuous, self-adjoint, positive-definite linear operators from $\mathcal{H}_{kl}$ to $\mathcal{H}_{kl}$. For $\Omega \in \mathcal{O}$, define

$$w(h;\Omega) := \langle h, \Omega h\rangle_{\mathcal{H}_{kl}}.$$

Let $\Omega_n \in \mathcal{O}$ be an empirical estimator of some possibly P_0-dependent operator $\Omega_0 \in \mathcal{O}$. For a threshold $\zeta \geq 0$, define

$$\mathcal{C}_n(\zeta) := \left\{ h \in \mathcal{H}_{kl} : w(\check{\Psi}_n - h; \Omega_n) \leq \frac{\zeta}{n} \right\}.$$

A consistent estimator of the $(1-\alpha)$-quantile $\zeta_{1-\alpha}$ of $w(\mathbb{H};\Omega_0)$ is obtained by Efron's bootstrap (Efron, 1992). We instantiate the bootstrap in our notation. Here $\mathbb{P}_k := n_k^{-1}\sum_{i:k(i)=k}\delta_{O_i}$ is the empirical law of fold k, and $n_k = n/K$. For $b \in [B]$ and $k \in [K]$, draw

$$O_{1,k}^{(b)}, \ldots, O_{n_k,k}^{(b)} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_k.$$

The bootstrap observations are sampled independently of everything else. Let $\mathbb{P}_k^{(b)}$ be the empirical law of $O_{1,k}^{(b)}, \ldots, O_{n_k,k}^{(b)}$. For $b \in [B]$, define

$$\mathbb{H}_n^{(b)} := \sqrt{n}\,\frac{1}{K}\sum_{k=1}^{K}(\mathbb{P}_k^{(b)} - \mathbb{P}_k)\widehat{D}^{-k} \in \mathcal{H}_{kl}.$$

We define $\widehat{\zeta}_n$ as the empirical $(1-\alpha)$-quantile of $w(\mathbb{H}_n^{(b)};\Omega_n)$ over $b = 1, \ldots, B$, conditionally on the original sample. The uncentered EIF $\widehat{D}^{-k}$ is not refitted inside each bootstrap iteration. The following proposition is a direct application of Luedtke & Chung (2024, Theorem 4).

Proposition 2. *Suppose the conditions of Theorem 3 hold. Further suppose that $\Omega_n, \Omega_0 \in \mathcal{O}$, $\|\Omega_n - \Omega_0\|_{\mathrm{op}} = o_p(1)$, and*

$$\max_{k\in[K]} \left\| \widehat{D}^{-k} - D_{\eta_0} \right\|_{L^2(P_0;\mathcal{H}_{kl})} = o_p(1).$$

Then $\widehat{\zeta}_n \xrightarrow{p} \zeta_{1-\alpha}$.

In this paper, we focus on the case $\Omega_n = \Omega_0 = \mathrm{Id}_{\mathcal{H}_{kl}}$, for which

$$w(\mathbb{H}_n^{(b)};\Omega_n) = \|\mathbb{H}_n^{(b)}\|^2_{\mathcal{H}_{kl}}.$$

This quantity is computed using the V-statistic in Section E, with $\mathbb{P}_k$ replaced by $\mathbb{P}_k^{(b)}$ in the validation folds. Equivalently, the same foldwise summands $\langle \widehat{D}^{-k_1}(o_1), \widehat{D}^{-k_2}(o_2)\rangle_{\mathcal{H}_{kl}}$ are used, with bootstrap resampling applied only to the outer validation sums. The nuisance estimators $\widehat{m}^{-k}$, $\widehat{\mu}_X^{-k}$, $\widehat{\mu}_\xi^{-k}$, and $\widehat{v}_j^{-k}$ are not refitted inside bootstrap draws.

Our primary objective is to benchmark the proposed method against naive methods which are not robust to nuisance estimation error, and which may therefore be biased when flexible machine learning estimators are used. It remains an open question whether the procedure can be improved by data-driven choices of Ω_n. A particularly natural choice is the regularized Wald-type confidence set described in Luedtke & Chung (2024). We defer these investigations to future work.

Confidence intervals based on triangle inequality. Let $k(i)$ denote the validation fold containing O_i, and define

$$G_{ij} := \left\langle \widehat{D}^{-k(i)}(O_i), \widehat{D}^{-k(j)}(O_j) \right\rangle_{\mathcal{H}_{kl}}.$$

Then the plug-in squared norm of the cross-fitted estimator is the V-statistic

$$\widehat{Q}_V := \|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}} = \frac{1}{n^2}\sum_{i,j=1}^{n} G_{ij}. \tag{10}$$

The quantities G_{ij} are precisely the foldwise Gram entries computed in Section E.

If $N_i^{(b)}$ denotes the number of times O_i appears in the bootstrap draw of its validation fold, then, for $\Omega_n = \mathrm{Id}_{\mathcal{H}_{kl}}$,

$$\|\mathbb{H}_n^{(b)}\|^2_{\mathcal{H}_{kl}} = \frac{1}{n}\sum_{i,j=1}^{n}\{N_i^{(b)}-1\}\{N_j^{(b)}-1\}G_{ij}.$$

For $\Omega_n = \mathrm{Id}_{\mathcal{H}_{kl}}$, the confidence set is

$$\mathcal{C}_n(\zeta) = \left\{ h \in \mathcal{H}_{kl} : \|\check{\Psi}_n - h\|^2_{\mathcal{H}_{kl}} \le \frac{\zeta}{n} \right\}.$$

Hence, by the reverse triangle inequality, it induces the conservative interval

$$\mathcal{I}_\zeta := \left[\left\{ \sqrt{\widehat{Q}_V} - \sqrt{\zeta/n} \right\}_+, \sqrt{\widehat{Q}_V} + \sqrt{\zeta/n} \right].$$

In implementation we take $\zeta = \widehat{\zeta}_n$. The induced test of $H_0 : \Psi_0 = 0$ rejects when $0 \notin \mathcal{I}_{\widehat{\zeta}_n}$, equivalently when $n\widehat{Q}_V > \widehat{\zeta}_n$.

F.1. CI based on the delta method

Beyond testing $\Psi_0 = 0$, we may be interested in the magnitude of the signal, namely $\|\Psi_0\|_{\mathcal{H}_{kl}}$ or $\|\Psi_0\|^2_{\mathcal{H}_{kl}}$. When $\Psi_0 \neq 0$, the functional delta method (van der Vaart, 2000, Chapter 3) gives an asymptotic normal approximation. Under the assumptions of Theorem 3,

$$\sqrt{n}(\check{\Psi}_n - \Psi_0) \rightsquigarrow \mathbb{H}.$$

Since $h \mapsto \|h\|^2_{\mathcal{H}_{kl}}$ is Fréchet differentiable with derivative $u \mapsto 2\langle h, u\rangle_{\mathcal{H}_{kl}}$, we have

$$\sqrt{n}\left(\|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}} - \|\Psi_0\|^2_{\mathcal{H}_{kl}}\right) \rightsquigarrow 2\langle \Psi_0, \mathbb{H}\rangle_{\mathcal{H}_{kl}}.$$

The limiting variance is $4\sigma^2$, where

$$\sigma^2 := \mathbb{E}_0\left[\langle \Psi_0, \phi_0(O)\rangle^2_{\mathcal{H}_{kl}}\right] = \mathbb{E}_0\left[\langle \Psi_0, D_{\eta_0}(O)\rangle^2_{\mathcal{H}_{kl}}\right] - \|\Psi_0\|^4_{\mathcal{H}_{kl}}.$$

This motivates the estimator

$$\begin{aligned}\widehat{\sigma}^2 &:= \frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\left\langle \check{\Psi}_n, \widehat{D}^{-l}(o')\right\rangle_{\mathcal{H}_{kl}}^2 - \|\check{\Psi}_n\|_{\mathcal{H}_{kl}}^4 \\ &= \frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\left(\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{o\in\mathcal{D}^k}\left\langle \widehat{D}^{-k}(o), \widehat{D}^{-l}(o')\right\rangle_{\mathcal{H}_{kl}}\right)^2 \\ &\quad - \left(\frac{1}{K^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\frac{1}{n_k n_l}\sum_{\substack{o\in\mathcal{D}^k\\ o'\in\mathcal{D}^l}}\left\langle \widehat{D}^{-k}(o), \widehat{D}^{-l}(o')\right\rangle_{\mathcal{H}_{kl}}\right)^2.\end{aligned} \tag{11}$$

The inner products $\langle \widehat{D}^{-k}(o), \widehat{D}^{-l}(o')\rangle_{\mathcal{H}_{kl}}$ are precisely the quantities computed in Section E.

Proposition 3 (Consistency of the estimator of asymptotic variance)**.** *Suppose the assumptions of Theorem 3 hold, and assume K is fixed and $n_k = n/K$ for all $k \in [K]$. Suppose further that*

$$\max_{k\in[K]} \|\widehat{D}^{-k} - D_{\eta_0}\|_{L^2(P_0;\mathcal{H}_{kl})} = o_p(1),$$

and that

$$\max_{k\in[K]} \int \|\widehat{D}^{-k}(o)\|_{\mathcal{H}_{kl}}^4 \, dP_0(o) = O_p(1), \qquad \int \|D_{\eta_0}(o)\|_{\mathcal{H}_{kl}}^4 \, dP_0(o) < \infty.$$

Define

$$\widehat{\sigma}^2 := \frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\langle \check{\Psi}_n, \widehat{D}^{-l}(o')\rangle_{\mathcal{H}_{kl}}^2 - \|\check{\Psi}_n\|_{\mathcal{H}_{kl}}^4.$$

Then

$$\widehat{\sigma}^2 \xrightarrow{p} \sigma^2,$$

where

$$\sigma^2 := \mathbb{E}_0[\langle \Psi_0, \phi_0(O)\rangle_{\mathcal{H}_{kl}}^2] = \mathbb{E}_0[\langle \Psi_0, D_{\eta_0}(O)\rangle_{\mathcal{H}_{kl}}^2] - \|\Psi_0\|_{\mathcal{H}_{kl}}^4.$$

Proof. By Theorem 3, $\check{\Psi}_n \xrightarrow{p} \Psi_0$. Hence, by the continuous mapping theorem,

$$\|\check{\Psi}_n\|_{\mathcal{H}_{kl}}^4 \xrightarrow{p} \|\Psi_0\|_{\mathcal{H}_{kl}}^4.$$

It remains to show that

$$\frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\langle \check{\Psi}_n, \widehat{D}^{-l}(o')\rangle_{\mathcal{H}_{kl}}^2 \xrightarrow{p} \mathbb{E}_0[\langle \Psi_0, D_{\eta_0}(O)\rangle_{\mathcal{H}_{kl}}^2].$$

We consider the decomposition

$$\begin{aligned}&\frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\langle \check{\Psi}_n, \widehat{D}^{-l}(o')\rangle_{\mathcal{H}_{kl}}^2 - \mathbb{E}_0[\langle \Psi_0, D_{\eta_0}(O)\rangle_{\mathcal{H}_{kl}}^2] \\ &= \frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\left\{\langle \check{\Psi}_n, \widehat{D}^{-l}(o')\rangle_{\mathcal{H}_{kl}}^2 - \langle \Psi_0, \widehat{D}^{-l}(o')\rangle_{\mathcal{H}_{kl}}^2\right\} && \text{(A)} \\ &\quad + \frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\left\{\langle \Psi_0, \widehat{D}^{-l}(o')\rangle_{\mathcal{H}_{kl}}^2 - \langle \Psi_0, D_{\eta_0}(o')\rangle_{\mathcal{H}_{kl}}^2\right\} && \text{(B)} \\ &\quad + \frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\langle \Psi_0, D_{\eta_0}(o')\rangle_{\mathcal{H}_{kl}}^2 - \mathbb{E}_0[\langle \Psi_0, D_{\eta_0}(O)\rangle_{\mathcal{H}_{kl}}^2]. && \text{(C)}\end{aligned}$$

For term (A), we use $a^2 - b^2 = (a-b)(a+b)$. We have

$$|(\text{A})| \le \frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\left|\langle\check{\Psi}_n - \Psi_0, \widehat{D}^{-l}(o')\rangle_{\mathcal{H}_{kl}}\right|\left|\langle\check{\Psi}_n + \Psi_0, \widehat{D}^{-l}(o')\rangle_{\mathcal{H}_{kl}}\right|$$
$$\le \|\check{\Psi}_n - \Psi_0\|_{\mathcal{H}_{kl}}\|\check{\Psi}_n + \Psi_0\|_{\mathcal{H}_{kl}}\frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\|\widehat{D}^{-l}(o')\|^2_{\mathcal{H}_{kl}}.$$

We now show that the final factor is $O_p(1)$. Conditionally on the training sample $\mathcal{D}^{-l}$, the observations in $\mathcal{D}^l$ are i.i.d. under P_0 and independent of $\widehat{D}^{-l}$. Therefore, by Chebyshev's inequality,

$$\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\|\widehat{D}^{-l}(o')\|^2_{\mathcal{H}_{kl}} - \int\|\widehat{D}^{-l}(o)\|^2_{\mathcal{H}_{kl}}\,dP_0(o) = o_p(1),$$

because

$$\int\|\widehat{D}^{-l}(o)\|^4_{\mathcal{H}_{kl}}\,dP_0(o) = O_p(1),$$

and $n_l \to \infty$. Since K is fixed, averaging over l gives

$$\frac{1}{K}\sum_{l=1}^{K}\left\{\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\|\widehat{D}^{-l}(o')\|^2_{\mathcal{H}_{kl}} - \int\|\widehat{D}^{-l}(o)\|^2_{\mathcal{H}_{kl}}\,dP_0(o)\right\} = o_p(1).$$

On the other hand,

$$\frac{1}{K}\sum_{l=1}^{K}\int\|\widehat{D}^{-l}(o)\|^2_{\mathcal{H}_{kl}}\,dP_0(o) \le \frac{2}{K}\sum_{l=1}^{K}\|\widehat{D}^{-l} - D_{\eta_0}\|^2_{L^2(P_0;\mathcal{H}_{kl})} + 2\|D_{\eta_0}\|^2_{L^2(P_0;\mathcal{H}_{kl})}$$
$$= O_p(1).$$

Therefore

$$\frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\|\widehat{D}^{-l}(o')\|^2_{\mathcal{H}_{kl}} = O_p(1).$$

Since $\check{\Psi}_n \xrightarrow{p} \Psi_0$, we have

$$|(\text{A})| = o_p(1).$$

For term (B), we again use $a^2 - b^2 = (a-b)(a+b)$. We have

$$|(\text{B})| \le \frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\|\Psi_0\|^2_{\mathcal{H}_{kl}}\|\widehat{D}^{-l}(o') - D_{\eta_0}(o')\|_{\mathcal{H}_{kl}}\|\widehat{D}^{-l}(o') + D_{\eta_0}(o')\|_{\mathcal{H}_{kl}}$$
$$\le \|\Psi_0\|^2_{\mathcal{H}_{kl}}\left\{\frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\|\widehat{D}^{-l}(o') - D_{\eta_0}(o')\|^2_{\mathcal{H}_{kl}}\right\}^{1/2}$$
$$\times\left\{\frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\|\widehat{D}^{-l}(o') + D_{\eta_0}(o')\|^2_{\mathcal{H}_{kl}}\right\}^{1/2}.$$

Conditionally on $\mathcal{D}^{-l}$, another application of Chebyshev's inequality gives

$$\frac{1}{K}\sum_{l=1}^{K}\left\{\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\|\widehat{D}^{-l}(o') - D_{\eta_0}(o')\|^2_{\mathcal{H}_{kl}} - \int\|\widehat{D}^{-l}(o) - D_{\eta_0}(o)\|^2_{\mathcal{H}_{kl}}\,dP_0(o)\right\} = o_p(1),$$

and similarly with $+$ in place of $-$. Hence

$$\frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\|\widehat{D}^{-l}(o')-D_{\eta_0}(o')\|_{\mathcal{H}_{kl}}^2=\frac{1}{K}\sum_{l=1}^{K}\|\widehat{D}^{-l}-D_{\eta_0}\|_{L^2(P_0;\mathcal{H}_{kl})}^2+o_p(1)=o_p(1),$$

while

$$\frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\|\widehat{D}^{-l}(o')+D_{\eta_0}(o')\|_{\mathcal{H}_{kl}}^2=O_p(1).$$

Therefore

$$|(\text{B})|=o_p(1).$$

Finally, for term (C), since $D_{\eta_0}\in L^2(P_0;\mathcal{H}_{kl})$, we have

$$\mathbb{E}_0[\langle\Psi_0,D_{\eta_0}(O)\rangle_{\mathcal{H}_{kl}}^2]\le\|\Psi_0\|_{\mathcal{H}_{kl}}^2\mathbb{E}_0[\|D_{\eta_0}(O)\|_{\mathcal{H}_{kl}}^2]<\infty.$$

Therefore, by the weak law of large numbers,

$$(\text{C})=\mathbb{P}_n\{\langle\Psi_0,D_{\eta_0}(O)\rangle_{\mathcal{H}_{kl}}^2\}-\mathbb{E}_0[\langle\Psi_0,D_{\eta_0}(O)\rangle_{\mathcal{H}_{kl}}^2]=o_p(1).$$

Combining the bounds for (A), (B), and (C), we obtain

$$\frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\langle\check{\Psi}_n,\widehat{D}^{-l}(o')\rangle_{\mathcal{H}_{kl}}^2\xrightarrow{p}\mathbb{E}_0[\langle\Psi_0,D_{\eta_0}(O)\rangle_{\mathcal{H}_{kl}}^2].$$

Together with $\|\check{\Psi}_n\|_{\mathcal{H}_{kl}}^4\xrightarrow{p}\|\Psi_0\|_{\mathcal{H}_{kl}}^4$, this proves

$$\widehat{\sigma}^2\xrightarrow{p}\mathbb{E}_0[\langle\Psi_0,D_{\eta_0}(O)\rangle_{\mathcal{H}_{kl}}^2]-\|\Psi_0\|_{\mathcal{H}_{kl}}^4=\sigma^2.$$

□

Theorem 4 (Validity and efficiency of the delta method)**.** *Suppose the assumptions of Theorem 3 hold. Then, in $\mathcal{H}_{kl}$,*

$$\sqrt{n}(\check{\Psi}_n-\Psi_0)\rightsquigarrow\mathbb{H},$$

where $\mathbb{H}$ is the centered Gaussian element in Theorem 3. Consequently,

$$\sqrt{n}\left(\|\check{\Psi}_n\|_{\mathcal{H}_{kl}}^2-\|\Psi_0\|_{\mathcal{H}_{kl}}^2\right)\rightsquigarrow 2\langle\Psi_0,\mathbb{H}\rangle_{\mathcal{H}_{kl}}.$$

Equivalently,

$$\sqrt{n}\left(\|\check{\Psi}_n\|_{\mathcal{H}_{kl}}^2-\|\Psi_0\|_{\mathcal{H}_{kl}}^2\right)\rightsquigarrow N(0,4\sigma^2),$$

where

$$\sigma^2:=\mathbb{E}_0[\langle\Psi_0,\phi_0(O)\rangle_{\mathcal{H}_{kl}}^2].$$

The scalar efficient influence function of $P\mapsto\|\Psi(P)\|_{\mathcal{H}_{kl}}^2$ at P_0 is

$$2\langle\Psi_0,\phi_0(O)\rangle_{\mathcal{H}_{kl}}.$$

Therefore $\|\check{\Psi}_n\|_{\mathcal{H}_{kl}}^2$ is regular and efficient for $\|\Psi_0\|_{\mathcal{H}_{kl}}^2$. If $\Psi_0\neq 0$, then

$$\sqrt{n}\left(\|\check{\Psi}_n\|_{\mathcal{H}_{kl}}-\|\Psi_0\|_{\mathcal{H}_{kl}}\right)\rightsquigarrow N\left(0,\frac{\sigma^2}{\|\Psi_0\|_{\mathcal{H}_{kl}}^2}\right),$$

and the scalar efficient influence function of $P\mapsto\|\Psi(P)\|_{\mathcal{H}_{kl}}$ at P_0 is

$$\frac{\langle\Psi_0,\phi_0(O)\rangle_{\mathcal{H}_{kl}}}{\|\Psi_0\|_{\mathcal{H}_{kl}}}.$$

Proof. The first display is exactly Theorem 3. The map $T : \mathcal{H}_{kl} \to \mathbb{R}$ defined by $T(h) = \|h\|^2_{\mathcal{H}_{kl}}$ is Fréchet differentiable at every $h \in \mathcal{H}_{kl}$, with derivative

$$\dot{T}_h[u] = 2\langle h, u\rangle_{\mathcal{H}_{kl}}.$$

The functional delta method therefore gives

$$\sqrt{n}\left(\|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}} - \|\Psi_0\|^2_{\mathcal{H}_{kl}}\right) \rightsquigarrow 2\langle \Psi_0, \mathbb{H}\rangle_{\mathcal{H}_{kl}}.$$

Since $\mathbb{H}$ is centered Gaussian and $\langle \Psi_0, \cdot\rangle_{\mathcal{H}_{kl}}$ is a continuous linear functional,

$$2\langle \Psi_0, \mathbb{H}\rangle_{\mathcal{H}_{kl}} \sim N(0, 4\sigma^2),$$

where

$$\sigma^2 = \mathbb{E}_0[\langle \Psi_0, \phi_0(O)\rangle^2_{\mathcal{H}_{kl}}].$$

The efficient influence function follows by the chain rule for pathwise differentiable Hilbert-valued parameters:

$$\phi_0^{\Theta}(O) = \dot{T}_{\Psi_0}[\phi_0(O)] = 2\langle \Psi_0, \phi_0(O)\rangle_{\mathcal{H}_{kl}}.$$

Since ϕ_0 is the efficient influence function for Ψ, this scalar influence function is efficient for $P \mapsto \|\Psi(P)\|^2_{\mathcal{H}_{kl}}$.

If $\Psi_0 \neq 0$, the map $S : \mathcal{H}_{kl} \to \mathbb{R}$ defined by $S(h) = \|h\|_{\mathcal{H}_{kl}}$ is Fréchet differentiable at Ψ_0, with derivative

$$\dot{S}_{\Psi_0}[u] = \frac{\langle \Psi_0, u\rangle_{\mathcal{H}_{kl}}}{\|\Psi_0\|_{\mathcal{H}_{kl}}}.$$

Another application of the functional delta method gives

$$\sqrt{n}\left(\|\check{\Psi}_n\|_{\mathcal{H}_{kl}} - \|\Psi_0\|_{\mathcal{H}_{kl}}\right) \rightsquigarrow \frac{\langle \Psi_0, \mathbb{H}\rangle_{\mathcal{H}_{kl}}}{\|\Psi_0\|_{\mathcal{H}_{kl}}},$$

which is centered Gaussian with variance $\sigma^2/\|\Psi_0\|^2_{\mathcal{H}_{kl}}$. The efficient influence function for $P \mapsto \|\Psi(P)\|_{\mathcal{H}_{kl}}$ follows by the same chain rule. □

Delta-method implementation with reduced finite-sample bias. We contrast the plug-in V-statistic defined in Eq. (10) with the corresponding U-statistic $\widehat{Q}_U$ by removing the self-pairs (Serfling, 2009).

$$\widehat{Q}_U \doteq \frac{1}{n(n-1)} \sum_{i \neq j} G_{ij}.$$

Under the assumptions of Theorem 3, we have $\widehat{Q}_U - \widehat{Q}_V = O_p(n^{-1})$. Therefore, when $\Psi_0 \neq 0$, both centers have the same delta-method limit since $\sqrt{n}(\widehat{Q}_U - \widehat{Q}_V) \rightsquigarrow 0$. In terms of asymptotic validity of the confidence sets, we can use whichever one we prefer. Since the positive bias of $\widehat{Q}_V$ is not negligible in small-sample settings, we suggest centering the delta-method confidence sets on $\widehat{Q}_U$, which leads to empirically better performance. However, the same argument does not allow for a substitution of $\widehat{Q}_U$ into the triangle-inequality-based interval. This is because it relies on the approximation

$$n\left(\sqrt{\widehat{Q}_V} - \|\Psi_0\|_{\mathcal{H}_{kl}}\right)^2 \rightsquigarrow \|\mathbb{H}\|^2_{\mathcal{H}_{kl}}.$$

The $O_p(n^{-1})$ error is not negligible at this scale. We therefore propose to rely on $\widehat{Q}_V$ in the construction of the triangle-inequality-based interval. We believe that constructing analogous confidence sets with better small-sample behaviour and analysing them is an interesting direction for future work.

G. Experiment details

Notation	Meaning
n	Sample size used in experiment figures and tables.
s_m, s_ϵ	Fourier-decay parameters controlling smoothness of the mean and noise-scale functions.
ρ	Strength of residual heterogeneity; $\rho = 0$ is homogeneous noise.
$a_k, b_k, \phi_{m,k}, \phi_{\sigma,k}$	Random Fourier coefficients and phases in the synthetic data-generating mechanisms.
$c_m, c_\epsilon, \sigma_0$	Scaling constants for the mean/noise series and the baseline noise scale.
$g(x), h(x), c$	Auxiliary functions and scale constant used in heteroscedastic noise examples.
$m_{Y\|X}, m_{X\|Y}$	Direction-specific regressions in the causal-arrow experiment.
$T, \sigma(0), \sigma(1)$	Extra covariate and group-specific residual scales in the covariate-adjusted experiment.

Throughout our experiments, the core nuisance functions $\hat{m}$ and $\hat{v}$ were obtained via 5-fold cross-fitting, using kernel ridge regression (vector-valued where appropriate). We found that Matérn kernels perform slightly better than smoother alternatives at learning v. Permutation tests are calibrated using 1000 permutations. Bootstrap-based procedures use 1000 bootstrap samples. All experiments can be performed on a single machine with a single NVIDIA RTX 4500 GPU and 64GB of RAM. To accelerate the evaluation of our method across a range of settings and random seeds, we parallelised the computations across multiple machines. Each one had 64GB of RAM, and multiple different GPUs were used, but the GPU capacity is not a core bottleneck. We did not evaluate how quickly our procedures can be run on a CPU. On the aforementioned RTX 4500 GPU, fitting a 5-fold cross-fitted confidence set in $d_x = d_y = 1$ and with 1000 data points took on the order of 1 to 2 minutes. The core complexity of the procedures scales quadratically with sample size, as is typical for kernel methods. We did not investigate ways of mitigating this empirically; however, since the implementation is expressed through Gram matrices and their derivative analogues, standard kernel-acceleration ideas such as random Fourier features, Nyström approximations, and incomplete or block U-statistics are natural candidates for future scalability work (Rahimi & Recht, 2007; Williams & Seeger, 2001; Zhang et al., 2018; Schrab et al., 2022).

G.1. Datasets used in the synthetic calibration-evaluation experiments

This subsection gives the scalar data-generating mechanism used for the reported calibration-evaluation experiments. The same random Fourier construction can be extended componentwise to multivariate X or Y, but the reported main experiments keep $d_x = d_y = 1$ to focus on residual-estimation difficulty.

The data-generating process is

$$Y = m(X) + \sigma(X)Z,$$

where $Z \sim \mathcal{N}(0,1)$ is sampled independently of X. The structural function used to construct the synthetic datasets is

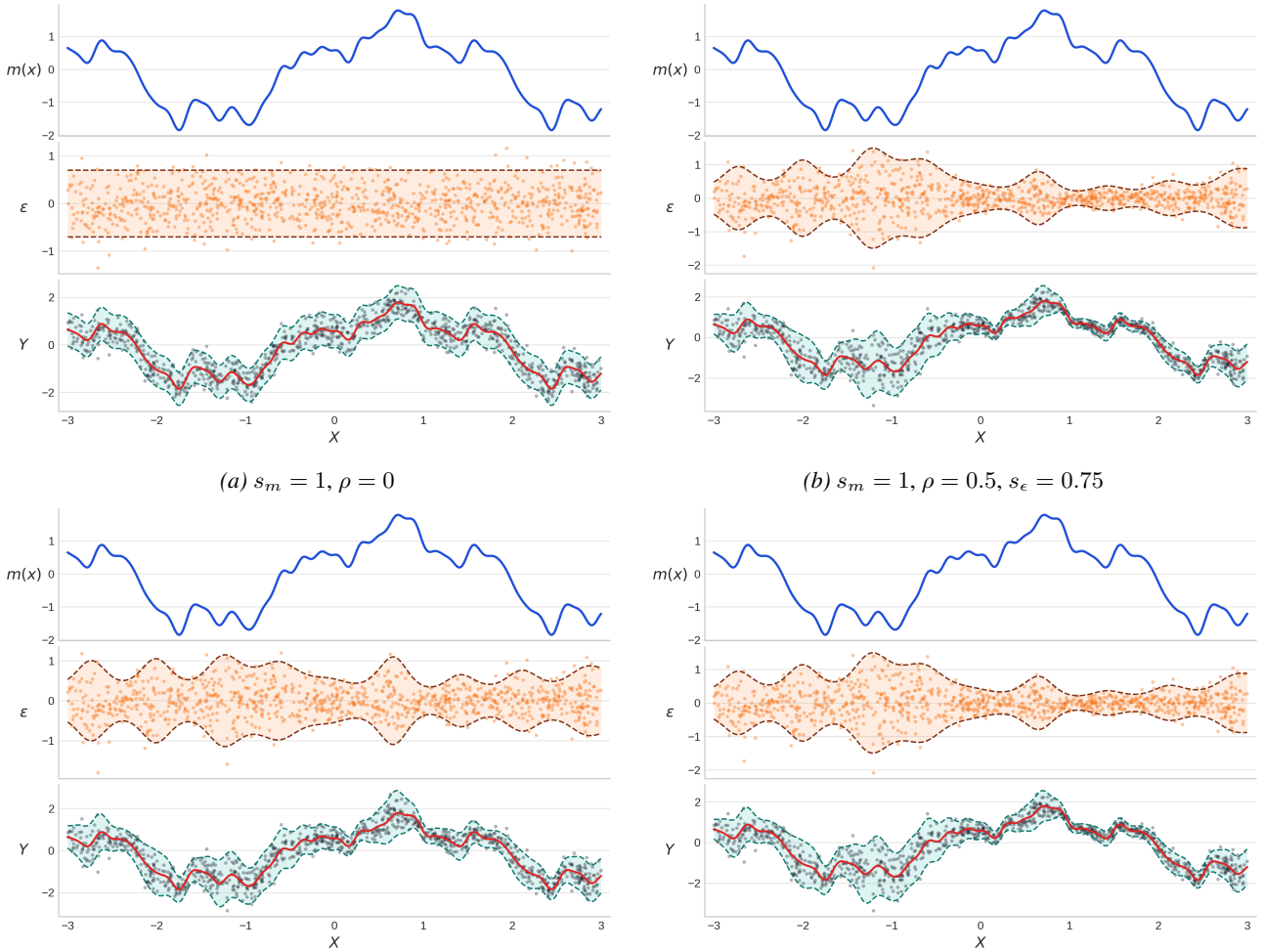
$$m(x) = c_m \sum_{k=1}^{20} a_k k^{-s_m} \sin(kx + \phi_{m,k}),$$

with $a_k \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $\phi_{m,k} \overset{i.i.d.}{\sim} \mathrm{Uniform}[-\pi, \pi]$, and c_m chosen post-hoc so that $\mathrm{Var}(m(X)) = 1$. Similarly,

$$g(x) = \sum_{k=1}^{20} b_k k^{-s_\epsilon} \sin(kx + \phi_{\sigma,k}),$$
$$\sigma(x) = \sigma_0 \exp(\rho c_\epsilon g(x)),$$

where $\rho \geq 0$ (a value of 0 corresponds to homogeneous noise with standard deviation σ_0). Weights b_k and phases $\phi_{\sigma,k}$ are sampled as above. c_ϵ is chosen to ensure $\mathrm{Var}(g(X)c_\epsilon) = 1$, and in all our experiments $\sigma_0 = 0.35$.

The HSIC kernels are Gaussian with bandwidths tuned by the median heuristic, applied to full X for the $\mathcal{H}_k$ kernel, and $\hat{\xi}$, the estimated residuals, on the $\mathcal{H}_l$ kernel. Example datasets are shown in Figure 5. We use the classical permutation



(a) $s_m = 1$, $\rho = 0$

(b) $s_m = 1$, $\rho = 0.5$, $s_\epsilon = 0.75$

(c) $s_m = 1$, $\rho = 0.3$, $s_\epsilon = 0.25$

(d) $s_m = 1$, $\rho = 0.5$, $s_\epsilon = 0.75$

Figure 5. Example synthetic datasets. Panel (a) is a homogeneous-noise null setting; panels (b)–(d) are alternatives with X-dependent noise scale.

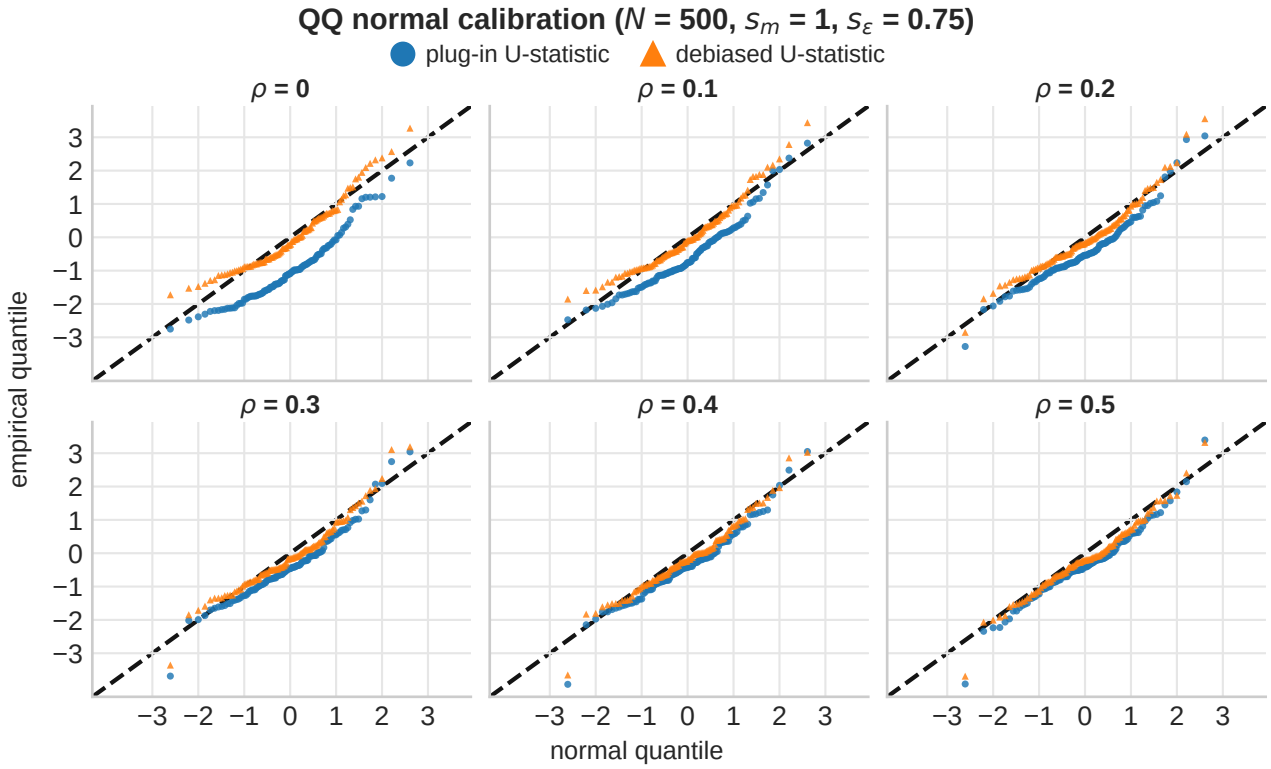
HSIC test (Gretton et al., 2007) as the plug-in baseline because it is the canonical residual-independence diagnostic in ANM experiments. Improvements to the underlying raw independence test, such as Schrab et al. (2022), are unrelated to the debiasing problem studied here, though it would be an interesting direction for future work to extend our estimation methodology to functionals that aggregate over cross-covariance operators in multiple RKHSs.

G.2. Normal convergence and coverage of the triangle inequality confidence sets.

We report empirical coverage for the nominal 95% confidence intervals in the same synthetic setting used for the main calibration plot, namely $s_m = 1$ and $s_\epsilon = 0.75$. The reverse-triangle-inequality intervals are valid at the null and remain conservative under alternatives. This is visible in Table 1: coverage is close to nominal when $\rho = 0$, but becomes essentially one as the signal increases. The delta-method intervals are not guaranteed to be valid at $\rho = 0$, where the delta method does not apply, but under alternatives they are much less conservative. For the debiased estimator, Table 2 shows coverage generally close to nominal, with the weakest performance at the smallest sample size and weakest alternative.

The QQ plots in Figure 6 tell the same story. Away from $\rho = 0$, the debiased statistic tracks the Gaussian reference line reasonably well, especially in the central quantiles. The plug-in statistic exhibits visible bias and slope distortions, while the debiased statistic is substantially better aligned with the target. At $\rho = 0$, the normal approximation is not expected to hold because of the norm functional's vanishing derivative; the departures in that panel are therefore a diagnostic of the

nonregular null rather than a failure of the alternative-regime delta method.



Note. At $\rho = 0$ there is no asymptotic normality under the null, so the QQ plot is not expected to fit the Gaussian reference line in that panel.

Figure 6. QQ normal calibration of the plug-in and debiased estimators for the X-dependent residual experiment at $n = 500$, across noise-dependence levels $\rho \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. The estimators are centered at the population-level $\|\Psi\|$ and scaled to have unit second moment.

Table 1. Coverage of the reverse-triangle-inequality confidence intervals for $s_m = 1$ and $s_\epsilon = 0.75$. Parentheses give 95% error bars.

ρ	$n = 250$	$n = 500$	$n = 1000$
0	0.930 (-0.050, +0.050)	0.955 (-0.039, +0.039)	0.916 (-0.053, +0.053)
0.1	0.970 (-0.033, +0.030)	0.982 (-0.025, +0.018)	0.991 (-0.018, +0.009)
0.2	0.990 (-0.020, +0.010)	0.991 (-0.018, +0.009)	1.000 (-0.000, +0.000)
0.3	1.000 (-0.000, +0.000)	1.000 (-0.000, +0.000)	1.000 (-0.000, +0.000)
0.4	1.000 (-0.000, +0.000)	1.000 (-0.000, +0.000)	1.000 (-0.000, +0.000)
0.5	1.000 (-0.000, +0.000)	1.000 (-0.000, +0.000)	1.000 (-0.000, +0.000)

Table 2. Coverage of the debiased delta-method confidence intervals for $s_m = 1$ and $s_\epsilon = 0.75$. The delta-method interval is only reported for $\rho > 0$. Parentheses give 95% error bars.

ρ	$n = 250$	$n = 500$	$n = 1000$
0.1	0.911 (-0.083, +0.083)	0.973 (-0.037, +0.027)	0.929 (-0.055, +0.055)
0.2	0.935 (-0.061, +0.061)	0.970 (-0.034, +0.030)	0.952 (-0.041, +0.041)
0.3	0.938 (-0.053, +0.053)	0.981 (-0.026, +0.019)	0.943 (-0.044, +0.044)
0.4	0.953 (-0.045, +0.045)	0.981 (-0.026, +0.019)	0.943 (-0.044, +0.044)
0.5	0.933 (-0.052, +0.052)	0.963 (-0.036, +0.036)	0.945 (-0.042, +0.042)

G.3. Causal arrow experiment details

The evaluated dataset is a simple example similar to those evaluated in the synthetic experiments of X***-dependent residuals*** in Section 4. The structural function is

$$m(x) = 0.55\,x + 0.95\,\sin(1.65\,x) + 0.38\,\sin(3.6\,x) + 0.1\,x^3.$$

The conditional standard deviation of the normally distributed noise is

$$\begin{aligned} \sigma(x) &= 0.45\exp(c\,\rho\{h(x) - \mathbb{E}[h(X)]\}), \\ h(x) &= \sin(1.6\,x) + 0.35\cos(2.4\,x), \\ c &= \sqrt{1/\mathrm{Var}(h(X))}. \end{aligned}$$

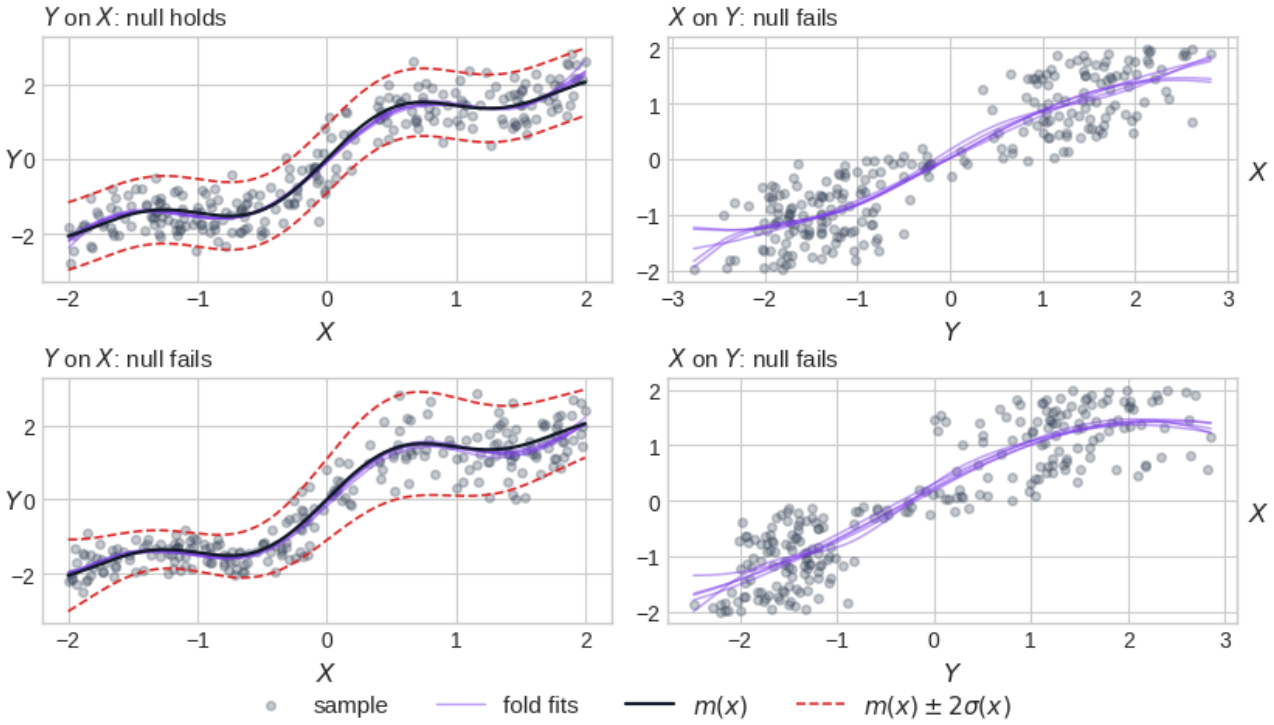


Figure 7. Example data from the causal ANM experiment, where the additive-noise direction is $X \to Y$.

We observe that the only procedures that are properly calibrated under the null are the debiased bootstrap procedure and the full-data cross-fitted permutation test, which was in fact very conservative, with its p-value distribution far from the uniform law that would correspond to perfect calibration. Among the tests that correctly fail to reject the null, only ours and the

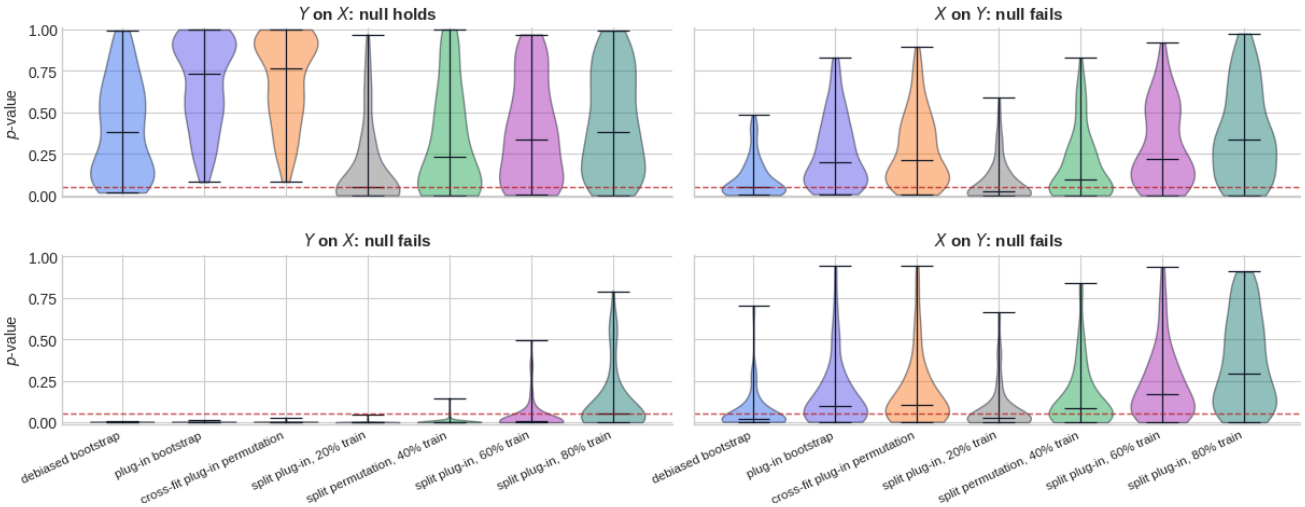


Figure 8. Distribution of causal-arrow scores across repetitions of the two-way goodness-of-fit experiment.

full-sample cross-fitted permutation baseline attain non-trivial power, with the former performing significantly better. The split-data tests that were calibrated attain very poor power under the alternative because the testing sample is too small.

G.4. Covariate-adjusted residual diagnostics

Each simulated dataset has $n = 120$ observations, with $X \sim \mathrm{Bernoulli}(1/2)$, $T \sim \mathrm{Unif}[-\pi, \pi]$, and

$$Y = m(X, T) + \sigma(X)\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, 1).$$

The mean is

$$m(x, t) = 1.5 \left\{ 0.2x + \sum_{k=1}^{5} \left[\frac{0.30}{k} \sin(kt) + \frac{0.25}{k} \cos(kt) + \frac{0.22}{k} x \sin(kt) + \frac{0.18}{k} x \cos(kt) \right] \right\}.$$

The null uses $\sigma(0) = \sigma(1) = 0.45$, while the alternative uses $\sigma(0) = 0.35, \sigma(1) = 0.55$.

The experiment uses 4-fold cross-fitting, with nuisances obtained via kernel ridge regression. Kernel bandwidths and ridge terms are obtained by cross-validation within the training folds.

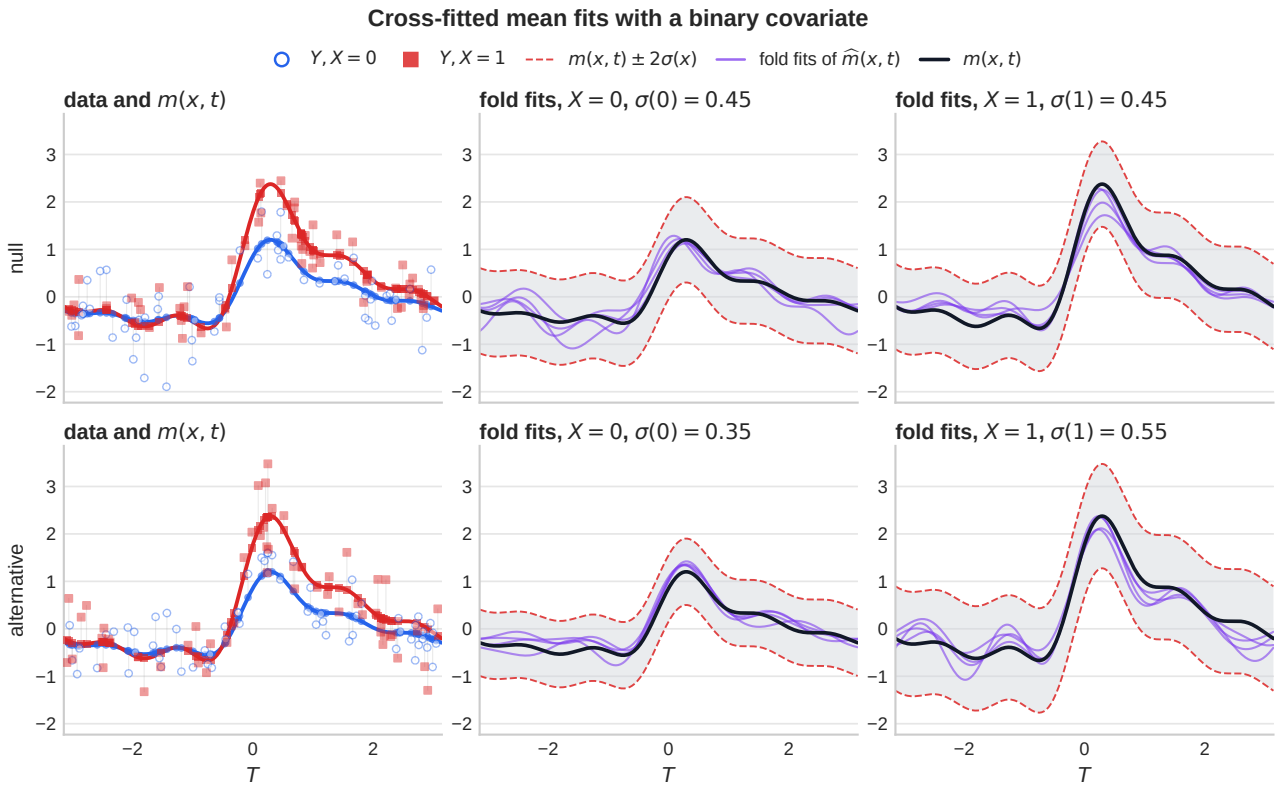


Figure 9. Fold-specific structural function fits in the covariate-adjusted residual experiment.

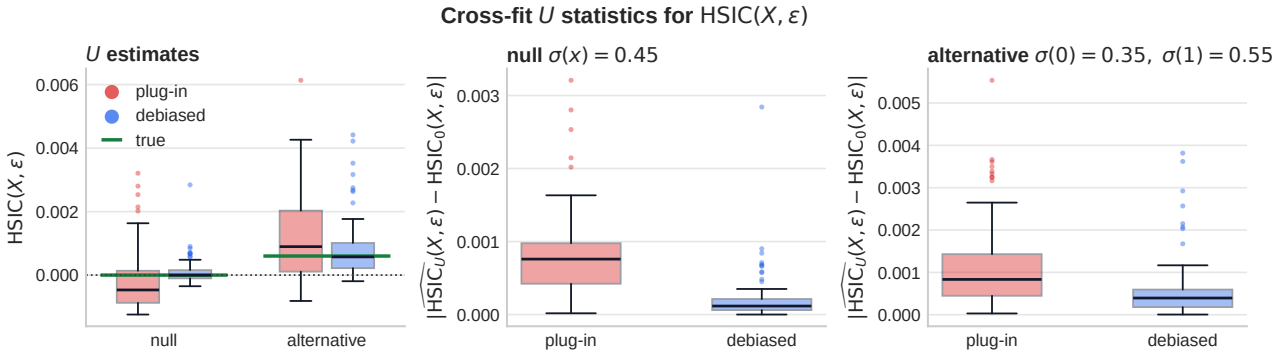


Figure 10. Comparison of the feasible U-statistic estimator with the oracle target in the covariate-adjusted residual experiment.

G.5. Ranges of empirical coverage point estimates

We report ranges of observed empirical coverage means across the indicated noise-dependence levels.

Table 3. Ranges of empirical coverage point estimates over ρ for the remaining synthetic calibration settings at $n = 250$, excluding the main setting $s_m = 1$ and $s_\epsilon = 0.75$. The reverse-triangle range is over $\rho \in \{0, 0.1, \ldots, 0.5\}$; the delta-method range is over $\rho \in \{0.1, \ldots, 0.5\}$.

s_m	s_ϵ	Reverse-triangle	Debiased delta method
1.0	0.25	0.940–0.990	0.838–0.941
1.0	0.5	0.930–1.000	0.881–0.951
1.5	0.25	0.920–1.000	0.886–0.987
1.5	0.5	0.890–1.000	0.900–0.989
1.5	0.75	0.890–1.000	0.940–0.989
2.0	0.25	0.920–1.000	0.887–0.972
2.0	0.5	0.910–0.990	0.893–0.975
2.0	0.75	0.910–1.000	0.931–0.973

Table 4. Ranges of empirical coverage point estimates over ρ for the remaining synthetic calibration settings at $n = 500$. The entries are ranges of observed means, not confidence intervals.

s_m	s_ϵ	Reverse-triangle	Debiased delta method
1.0	0.25	0.938–1.000	0.966–0.989
1.0	0.5	0.955–1.000	0.969–0.991
1.5	0.25	0.933–1.000	0.959–0.982
1.5	0.5	0.964–1.000	0.943–0.985
1.5	0.75	0.964–1.000	0.953–0.974
2.0	0.25	0.929–1.000	0.960–0.978
2.0	0.5	0.982–1.000	0.944–0.989
2.0	0.75	0.982–1.000	0.943–0.986

Table 5. Ranges of empirical coverage point estimates over ρ for the remaining synthetic calibration settings at $n = 1000$. The entries are ranges of observed means, not confidence intervals.

s_m	s_ϵ	Reverse-triangle	Debiased delta method
1.0	0.25	0.918–0.991	0.938–0.970
1.0	0.5	0.918–1.000	0.920–0.963
1.5	0.25	0.914–0.991	0.900–0.955
1.5	0.5	0.896–1.000	0.919–0.943
1.5	0.75	0.907–1.000	0.915–0.951
2.0	0.25	0.895–1.000	0.917–0.960
2.0	0.5	0.879–1.000	0.933–0.945
2.0	0.75	0.882–1.000	0.925–0.943

G.6. Additional QQ calibration plots at $n = 500$

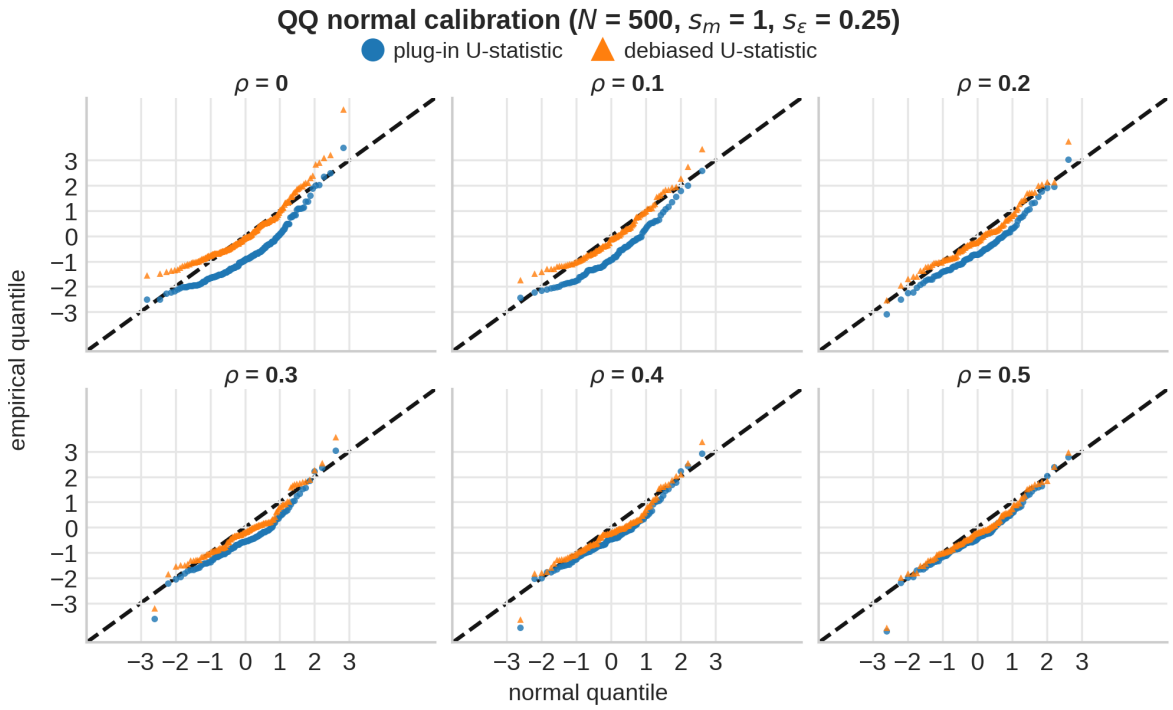


Figure 11. QQ calibration at $n = 500$, $s_m = 1$, and $s_\epsilon = 0.25$.

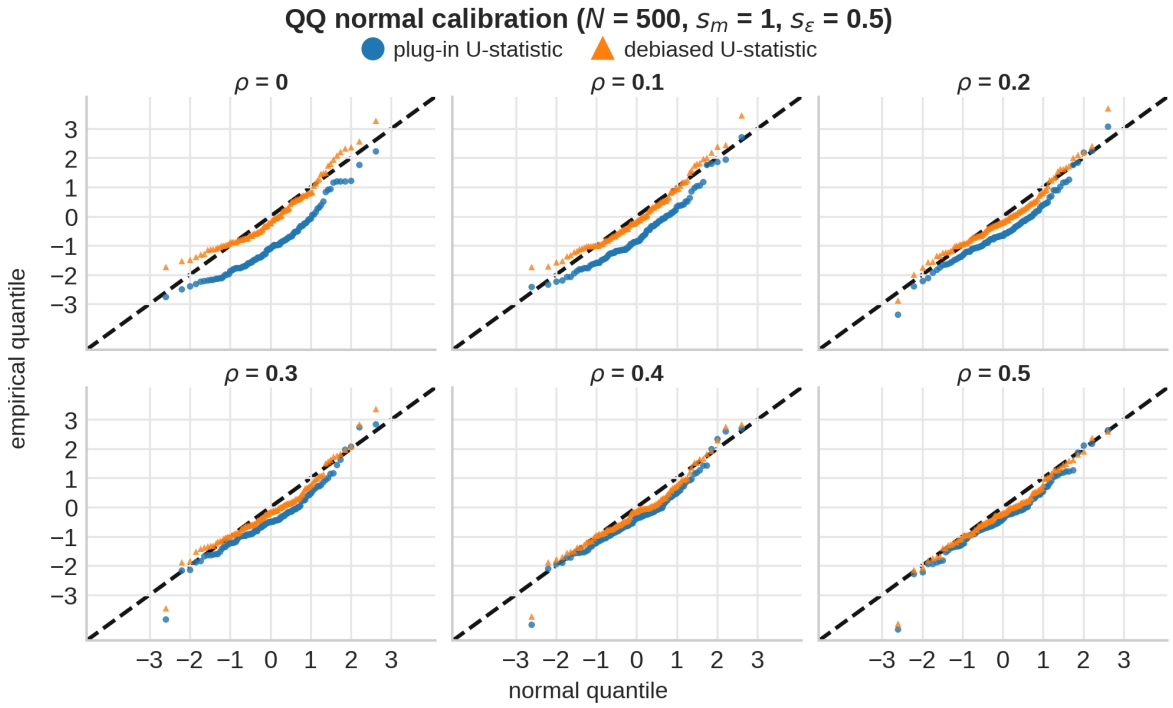


Figure 12. QQ calibration at $n = 500$, $s_m = 1$, and $s_\epsilon = 0.5$.

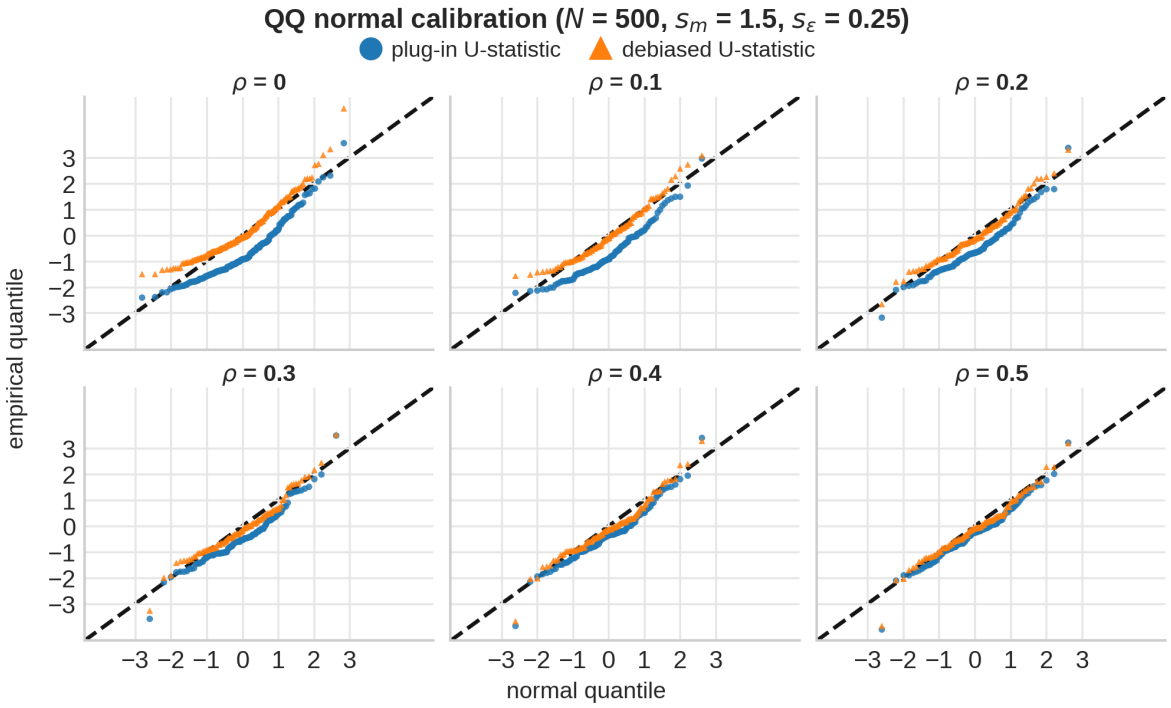


Figure 13. QQ calibration at $n = 500$, $s_m = 1.5$, and $s_\epsilon = 0.25$.

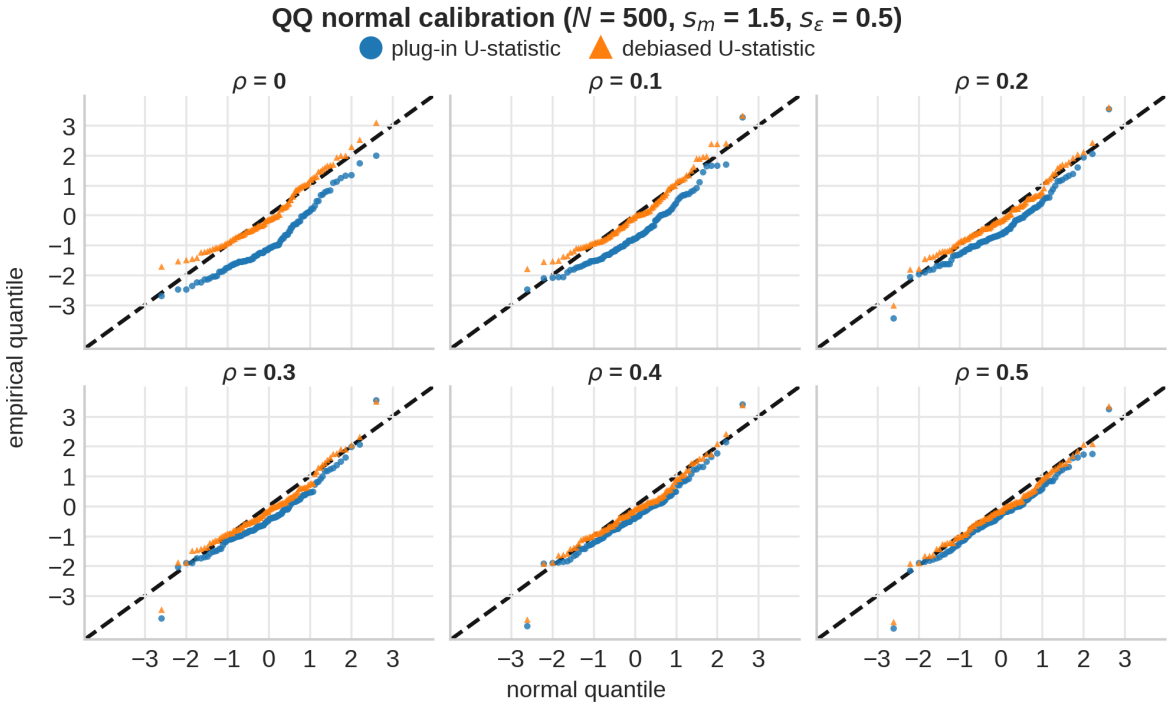


Figure 14. QQ calibration at $n = 500$, $s_m = 1.5$, and $s_\epsilon = 0.5$.

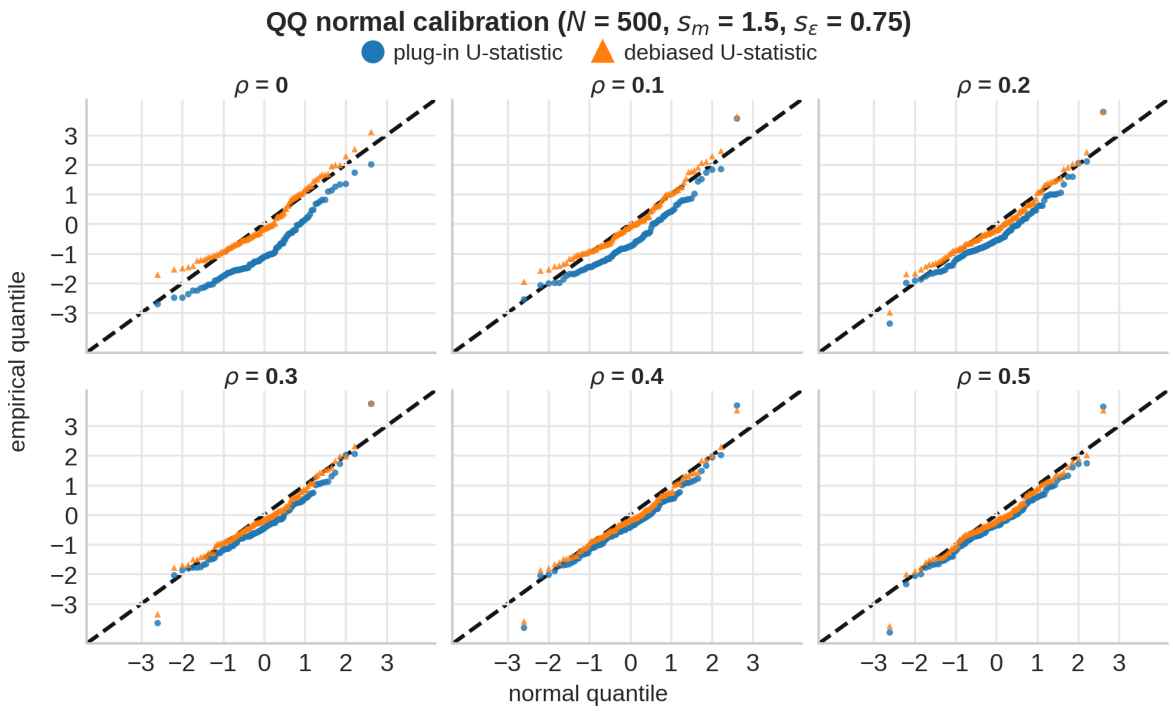


Figure 15. QQ calibration at $n = 500$, $s_m = 1.5$, and $s_\epsilon = 0.75$.

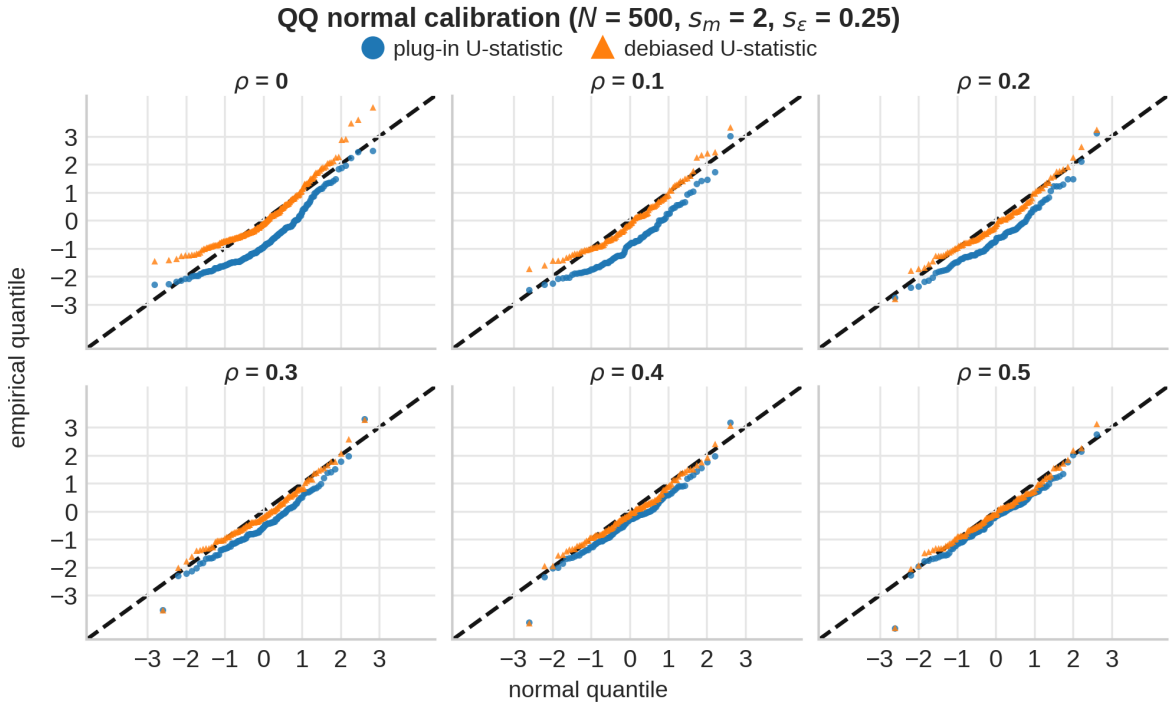


Figure 16. QQ calibration at $n = 500$, $s_m = 2$, and $s_\epsilon = 0.25$.

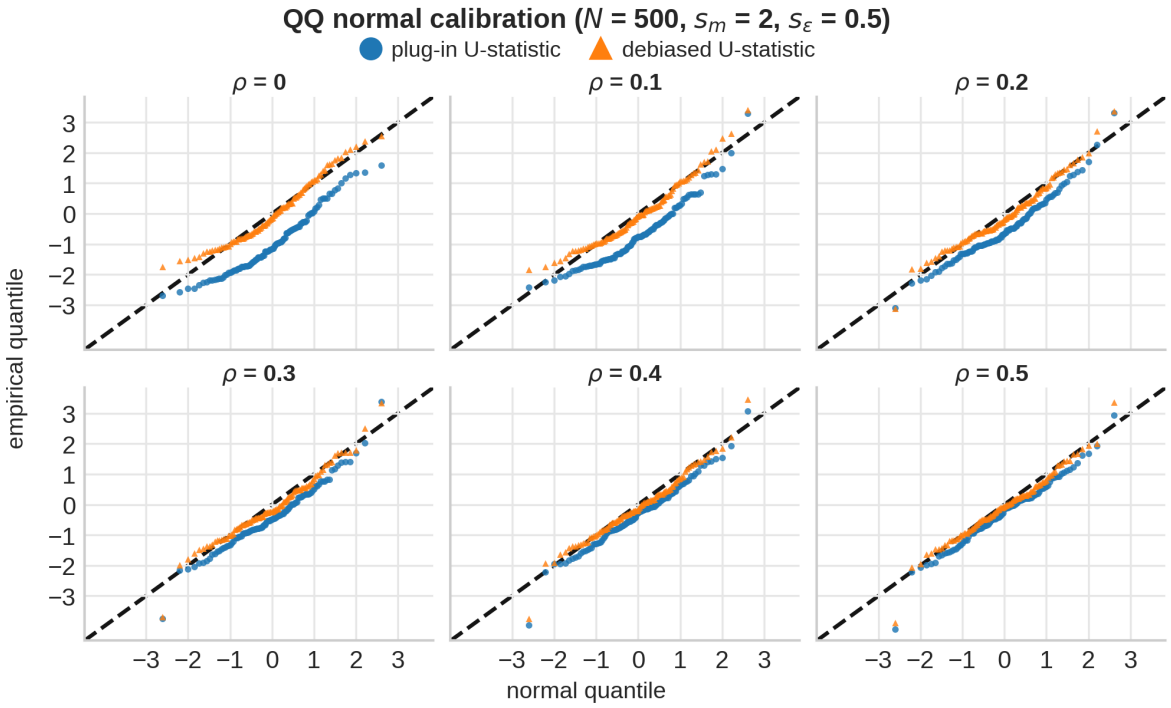


Figure 17. QQ calibration at $n = 500$, $s_m = 2$, and $s_\epsilon = 0.5$.

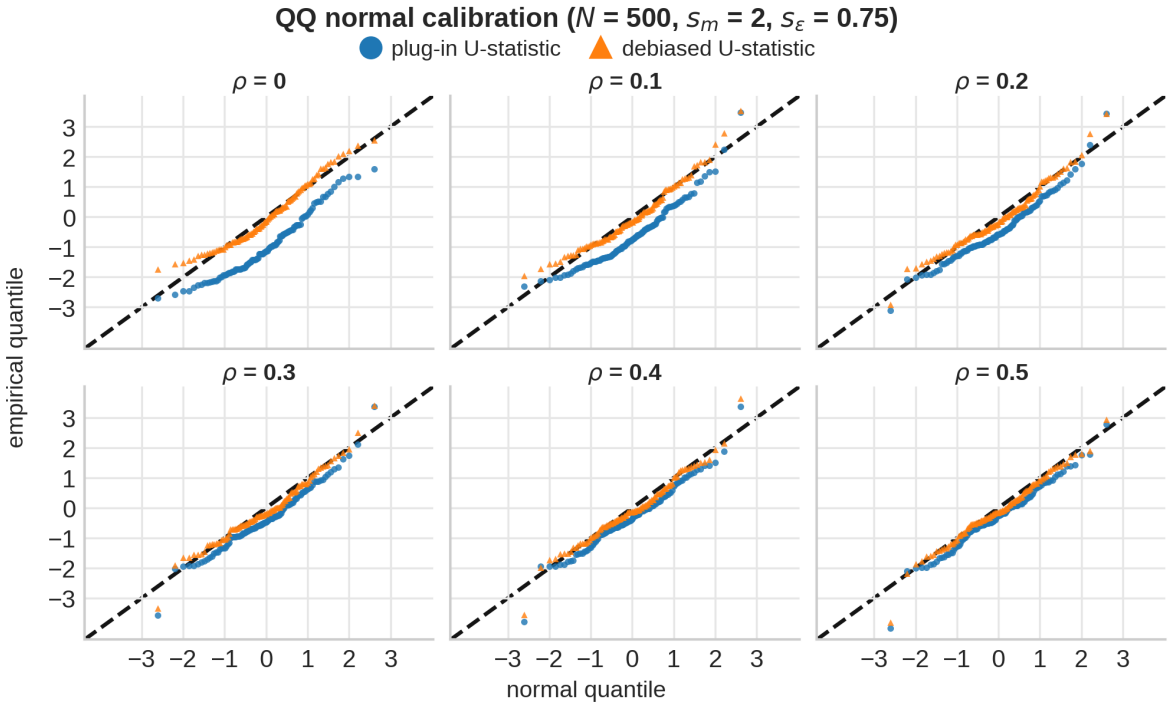


Figure 18. QQ calibration at $n = 500$, $s_m = 2$, and $s_\epsilon = 0.75$.

H. Validity of asymptotically normal confidence sets

Notation	Meaning
$\check{\Psi}_n$	Cross-fitted one-step estimator of Ψ_0 used in the delta-method diagnostic.
$\widehat{\zeta}_n$	Bootstrap quantile for $\|\mathbb{H}\|^2_{\mathcal{H}_{kl}}$.
$\hat{\sigma}^2$	Plug-in estimate of the delta-method variance component.
$z_{1-\beta/2}$	Standard normal quantile used for the diagnostic comparison.
$\widehat{\mathfrak{D}}_n$	Ratio of estimated quadratic remainder scale to estimated linear-term scale.

Delta method. We have the algebraic identity

$$\begin{aligned}&\sqrt{n}\left(\|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}} - \|\Psi_0\|^2_{\mathcal{H}_{kl}}\right)\\&= \sqrt{n}\left\langle \check{\Psi}_n - \Psi_0, \check{\Psi}_n + \Psi_0\right\rangle_{\mathcal{H}_{kl}}\\&= \sqrt{n}\|\check{\Psi}_n - \Psi_0\|^2_{\mathcal{H}_{kl}} + 2\sqrt{n}\langle \check{\Psi}_n - \Psi_0, \Psi_0\rangle_{\mathcal{H}_{kl}}.\end{aligned}$$

We refer to $2\sqrt{n}\langle \check{\Psi}_n - \Psi_0, \Psi_0\rangle_{\mathcal{H}_{kl}}$ as the *linear term* and $\sqrt{n}\|\check{\Psi}_n - \Psi_0\|^2_{\mathcal{H}_{kl}}$ as the *quadratic remainder*. For a fixed $\Psi_0 \neq 0$, we derive a distributional limit for $\sqrt{n}\left(\|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}} - \|\Psi_0\|^2_{\mathcal{H}_{kl}}\right)$. Suppose the conclusion of Theorem 3 holds, namely

$$\sqrt{n}(\check{\Psi}_n - \Psi_0) \rightsquigarrow \mathbb{H}.$$

By the Continuous Mapping Theorem, $n\|\check{\Psi}_n - \Psi_0\|^2_{\mathcal{H}_{kl}} \rightsquigarrow \|\mathbb{H}\|^2_{\mathcal{H}_{kl}}$. Thus $\sqrt{n}\|\check{\Psi}_n - \Psi_0\|^2_{\mathcal{H}_{kl}} \rightsquigarrow 0$. By another application of the Continuous Mapping Theorem, $\sqrt{n}\left\langle \check{\Psi}_n - \Psi_0, \Psi_0\right\rangle \rightsquigarrow \langle \mathbb{H}, \Psi_0\rangle_{\mathcal{H}_{kl}}$. Thus

$$\sqrt{n}\left(\|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}} - \|\Psi_0\|^2_{\mathcal{H}_{kl}}\right) \rightsquigarrow 2\langle \mathbb{H}, \Psi_0\rangle_{\mathcal{H}_{kl}}.$$

Since $2\langle \mathbb{H}, \Psi_0\rangle_{\mathcal{H}_{kl}} \sim \mathcal{N}(0, 4\mathbb{E}[\langle \phi(W,X,Y), \Psi_0\rangle^2_{\mathcal{H}_{kl}}])$, this yields a closed-form asymptotic variance that can be used to construct a calibrated confidence interval for $\sqrt{n}\left(\|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}} - \|\Psi_0\|^2_{\mathcal{H}_{kl}}\right)$.

However, in finite samples, the quality of the approximation heavily depends on the size of $\sqrt{n}\|\check{\Psi}_n - \Psi_0\|^2_{\mathcal{H}_{kl}}$ relative to $2\sqrt{n}\langle \check{\Psi}_n - \Psi_0, \Psi_0\rangle_{\mathcal{H}_{kl}}$. We therefore propose a data-driven diagnostic via the bootstrapped empirical measures $\mathbb{P}_n^{(b)}$. Recall that in Section 3, we defined $\widehat{\zeta}_n$ as the $(1-\alpha)$-quantile of $\|\mathbb{H}_n^{(b)}\|^2_{\mathcal{H}_{kl}}$, conditionally on the original sample $(O_1, \ldots, O_n)$. The confidence interval $\mathcal{C}_n(\widehat{\zeta}_n)$ contains 0 if and only if $n\|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}} \leq \widehat{\zeta}_n$.

If the linear term has reached its asymptotic regime, then

$$2\sqrt{n}\left\langle \check{\Psi}_n - \Psi_0, \Psi_0\right\rangle_{\mathcal{H}_{kl}} \overset{d}{\approx} N(0, 4\sigma^2), \qquad \sigma^2 = \mathrm{Var}\big(\langle \Psi_0, \mathbb{H}\rangle_{\mathcal{H}_{kl}}\big).$$

Thus the $(1-\beta)$ quantile of the absolute value of the linear term is approximated by

$$2\hat{\sigma} z_{1-\beta/2},$$

where $\hat{\sigma}^2$ is the plug-in variance estimator from Proposition 3. On the other hand, since $\hat{\zeta}_n$ estimates the $(1-\alpha)$ quantile of $\|\mathbb{H}\|^2_{\mathcal{H}_{kl}}$, the quantity $n^{-1/2}\hat{\zeta}_n$ estimates the $(1-\alpha)$ quantile of the quadratic remainder

$$\sqrt{n}\|\check{\Psi}_n - \Psi_0\|^2_{\mathcal{H}_{kl}}.$$

This gives the diagnostic

$$\widehat{\mathfrak{D}}_n = \frac{n^{-1/2}\hat{\zeta}_n}{2\hat{\sigma} z_{1-\beta/2}}.$$

A simple warning sign is that the triangle-inequality-based confidence set (Eq. (3)) contains zero:

$$0 \in \mathcal{C}_n(\hat{\zeta}_n) \quad \Longleftrightarrow \quad n\|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}} \leq \hat{\zeta}_n.$$

In this case the estimated signal is no larger than the bootstrap uncertainty for $\check{\Psi}_n$ itself. This is precisely the near-null regime where the derivative of $h \mapsto \|h\|^2_{\mathcal{H}_{kl}}$ at Ψ_0 is too small for the linear approximation to dominate.

I. Statistical learning theory

Notation	Meaning
$\mathcal{S}_j$, $\mathcal{S}$	$(\partial_j \varphi_l)\{Y - m_0(W)\} \in \mathcal{H}_l$; $\mathcal{S}_j$ with coordinate index suppressed in the derivations.
$v_{0,j}$, v	$\mathbb{E}_0[\mathcal{S}_j \mid W = w]$ vector-valued true regression function, where subscript 0 denotes dependence on true population distribution; $v_{0,j}$ with coordinate index suppressed in the derivations.
q, $\mathcal{H}_q$, κ_q	reproducing kernel on $\mathcal{W}$, its RKHS, and its uniform kernel bound.
I_w, S_w, L_W, C_W	L^2-embedding of $\mathcal{H}_q$, adjoint of I_w, integral operator $I_w \circ S_w$, covariance operator $S_w \circ I_w$.
I, μ_i, e_i	Index set, eigenvalues, and eigenfunctions of spectral decomposition of L_W used to define interpolation spaces. (see Eq. (12))
$[\mathcal{H}_q]^\alpha$, $[\mathcal{G}]^\alpha$	Scalar and vector-valued interpolation spaces with index α
$\mathcal{G}$, $S_2(H, H')$	Vector-valued RKHS (see Definition 1) and Hilbert–Schmidt operators from H to H'.
β, C_0	Exponent for v_0 in Eq. (SRC) (measures smoothness), and operator representer for v_0.
p	Exponent for μ_i in Eq. (EVD) (measures capacity of $\mathcal{H}_q$)
$p_{\mathcal{S}\mid W}$, σ, R	Conditional law of the label $\mathcal{S}$ and constants in Eq. (MOM).
$\widehat{v}_\lambda$, $\widetilde{v}_\lambda$	empirical vvKRR solution with estimated labels (i.e. plug-in regression function $\widehat{m}$) in Eq. (14), empirical vvKRR with oracle labels $\mathcal{S}_j$ in Eq. (15).
$\widehat{C}_W$, $C_{W,\lambda}$, $\widehat{C}_{W,\lambda}$	Empirical covariance operator $\frac{2}{n}\sum_{i=1}^{n/2} \varphi_q(w_i) \otimes \varphi_q(w_i)$, population regularized covariance operator $\mathbb{E}_0[\varphi_q(W) \otimes \varphi_q(W)] + \lambda I$, empirical regularized covariance operators $\widehat{C}_W + \lambda I$.

For the fixed coordinate j, define

$$\mathcal{S}_j := (\partial_j \varphi_l)\{Y - m_0(W)\} \in \mathcal{H}_l,$$

and let

$$v_{0,j}(w) := \mathbb{E}_0[\mathcal{S}_j \mid W = w].$$

We use $v_{0,j}$ both for a fixed measurable version of this conditional expectation and for its $L^2(P_{W,0}; \mathcal{H}_l)$-equivalence class. The analysis is identical for each j, so we fix j and suppress it from the notation. In other words, we write

$$\mathcal{S} := \mathcal{S}_j, \qquad v_0 := v_{0,j},$$

where

$$v_0(w) := \mathbb{E}_0[\mathcal{S} \mid W = w].$$

We use v_0 both for a fixed measurable version of this conditional expectation and for its $L^2(P_{W,0}; \mathcal{H}_l)$-equivalence class.

We make the following assumptions on the kernel q and its RKHS $\mathcal{H}_q \subseteq \mathbb{R}^{\mathcal{W}}$.

1. $\mathcal{H}_q$ is separable; this is satisfied if q is continuous, given that $\mathcal{W}$ is separable;[1]
2. $q(\cdot, w)$ is measurable for $P_{W,0}$-almost all $w \in \mathcal{W}$;

[1]This follows from Steinwart & Christmann (2008, Lemma 4.33). Note that the Lemma requires separability of $\mathcal{W}$, which is satisfied since we assume that $\mathcal{W}$ is a second countable locally compact Hausdorff space.

3. $q(w,w) \leq \kappa_q^2$ for $P_{W,0}$-almost all $w \in \mathcal{W}$.

The above assumptions are not restrictive in practice, as well-known kernels such as the Gaussian, Laplace and Matérn kernels satisfy them on $\mathcal{W} \subseteq \mathbb{R}^{d_w}$ (Sriperumbudur et al., 2011).

We now introduce some facts about the interplay between $\mathcal{H}_q$ and $L^2(P_{W,0})$, which has been extensively studied by (Smale & Zhou, 2004, 2005), (De Vito et al., 2006) and (Steinwart & Scovel, 2012). We first define the embedding $I_w : \mathcal{H}_q \to L^2(P_{W,0})$, mapping a function $f \in \mathcal{H}_q$ to its $P_{W,0}$-equivalence class $[f]$. Since the kernel q is bounded, we know that this embedding is a well-defined Hilbert-Schmidt operator. Indeed, by (Steinwart & Scovel, 2012, Lemma 2.2 & 2.3), its Hilbert-Schmidt norm can be computed as

$$\|I_w\|^2_{S_2(\mathcal{H}_q, L^2(P_{W,0}))} = \int_{\mathcal{W}} q(w,w) \mathrm{d}P_{W,0}(w) \leq \kappa_q^2.$$

The adjoint operator $S_w := I_w^* : L^2(P_{W,0}) \to \mathcal{H}_q$ is an integral operator with respect to the kernel q, i.e. for $f \in L^2(P_{W,0})$ and $w \in \mathcal{W}$ we have (Steinwart & Christmann, 2008, Theorem 4.27)

$$(S_w f)(w) = \int_{\mathcal{W}} q(w,w') f(w') \mathrm{d}P_{W,0}(w').$$

Next, we define the self-adjoint, positive semi-definite and trace class integral operators

$$L_W := I_w S_w : L^2(P_{W,0}) \to L^2(P_{W,0}) \quad \text{and} \quad C_W := S_w I_w : \mathcal{H}_q \to \mathcal{H}_q.$$

We now introduce the background required in order to characterize the smoothness of the target function v_0. We review the results of (Steinwart & Scovel, 2012) and (Fischer & Steinwart, 2020) in constructing scalar-valued interpolation spaces, and (Li et al., 2022) in defining vector-valued interpolation spaces.

Real-valued Interpolation Space: By the spectral theorem for self-adjoint compact operators, there exists an at most countable index set I, a non-increasing sequence $(\mu_i)_{i\in I} > 0$, and a family $(e_i)_{i\in I} \in \mathcal{H}_q$, such that $([e_i])_{i\in I}$ [2] is an orthonormal basis (ONB) of $\overline{\operatorname{ran} I_w} \subseteq L^2(P_{W,0})$ and $(\mu_i^{1/2} e_i)_{i\in I}$ is an ONB of $(\ker I_w)^\perp \subseteq \mathcal{H}_q$, and we have

$$L_W = \sum_{i\in I} \mu_i \langle \cdot, [e_i] \rangle_{L^2(P_{W,0})} [e_i], \qquad C_W = \sum_{i\in I} \mu_i \langle \cdot, \mu_i^{\frac{1}{2}} e_i \rangle_{\mathcal{H}_q} \left(\mu_i^{\frac{1}{2}} e_i \right). \tag{12}$$

For $\alpha \geq 0$, the α-interpolation space (Steinwart & Scovel, 2012) is defined by

$$[\mathcal{H}_q]^\alpha := \left\{ \sum_{i\in I} a_i \mu_i^{\alpha/2} [e_i] : (a_i)_{i\in I} \in \ell_2(I) \right\} \subseteq L^2(P_{W,0}),$$

equipped with the inner product

$$\left\langle \sum_{i\in I} a_i (\mu_i^{\alpha/2} [e_i]), \sum_{i\in I} b_i (\mu_i^{\alpha/2} [e_i]) \right\rangle_{[\mathcal{H}_q]^\alpha} = \sum_{i\in I} a_i b_i,$$

for $(a_i)_{i\in I}, (b_i)_{i\in I} \in \ell_2(I)$. The α-interpolation space defines a Hilbert space. Moreover, $\left(\mu_i^{\alpha/2} [e_i]\right)_{i\in I}$ forms an ONB of $[\mathcal{H}_q]^\alpha$ and consequently $[\mathcal{H}_q]^\alpha$ is a separable Hilbert space. In the following, we use the abbreviation $\|\cdot\|_\alpha := \|\cdot\|_{[\mathcal{H}_q]^\alpha}$.

Vector-valued Reproducing Kernel Hilbert Space. Let $K : \mathcal{W} \times \mathcal{W} \to \mathcal{L}(\mathcal{H}_l)$ be an operator valued positive-semidefinite (psd) kernel. Fix K, $w \in \mathcal{W}$, and $h \in \mathcal{H}_l$, then $(K_w h)(\cdot) := K(\cdot, w)h$ defines a function from $\mathcal{W}$ to $\mathcal{H}_l$. The completion of

$$\mathcal{G}_{\text{pre}} := \operatorname{span}\{K_w h \mid w \in \mathcal{W}, h \in \mathcal{H}_l\},$$

[2] We recall that the bracket $[\cdot]$ denotes the embedding that maps f to its equivalence class $I_w(f) \in L^2(P_{W,0})$.

with inner product on $\mathcal{G}_{\text{pre}}$ defined on the elementary elements as $\langle K_w h, K_{w'} h' \rangle_{\mathcal{G}} := \langle h, K(w, w') h' \rangle_{\mathcal{H}_l}$, defines a vRKHS denoted as $\mathcal{G}$. For a more complete overview of the vector-valued reproducing kernel Hilbert space, we refer the reader to (Carmeli et al., 2006), (Carmeli et al., 2010) and (Li et al., 2024, Section 2). In the following, we will denote $\mathcal{G}$ as the vRKHS induced by the kernel $K : \mathcal{W} \times \mathcal{W} \to \mathcal{L}(\mathcal{H}_l)$ with

$$K(w, w') := q(w, w') \operatorname{Id}_{\mathcal{H}_l}, \quad w, w' \in \mathcal{W}. \tag{13}$$

Theorem 5 (vRKHS isomorphism)**.** *For every function $F \in \mathcal{G}$ there exists a unique operator $C \in S_2(\mathcal{H}_q, \mathcal{H}_l)$ such that $F(\cdot) = C\varphi_q(\cdot) \in \mathcal{H}_l$ with $\|C\|_{S_2(\mathcal{H}_q,\mathcal{H}_l)} = \|F\|_{\mathcal{G}}$ and vice versa. Hence $\mathcal{G} \simeq S_2(\mathcal{H}_q, \mathcal{H}_l)$ and we denote the isometric isomorphism between $S_2(\mathcal{H}_q, \mathcal{H}_l)$ and $\mathcal{G}$ as $\mathfrak{I}$. It follows that $\mathcal{G}$ can be written as $\mathcal{G} = \{F : \mathcal{W} \to \mathcal{H}_l \mid F = C\varphi_q(\cdot), C \in S_2(\mathcal{H}_q, \mathcal{H}_l)\}$.*

Vector-valued Interpolation Space: Introduced in (Li et al., 2022), vector-valued interpolation spaces generalize the notion of scalar-valued interpolation spaces to vRKHS with a separable kernel of the form (13). We refer the reader to Steinwart & Scovel (2012); Fischer & Steinwart (2020) for an introduction to scalar-valued interpolation spaces. For two Hilbert spaces H, H', we use $S_2(H, H')$ to denote the space of Hilbert-Schmidt operators from H to H'.

Definition 1 (Vector-valued interpolation space)**.** *Let q be a real-valued kernel with associated RKHS $\mathcal{H}_q$ and let $[\mathcal{H}_q]^\alpha$ be the real-valued interpolation space associated to $\mathcal{H}_q$ with some $\alpha \geq 0$. The vector-valued interpolation space $[\mathcal{G}]^\alpha$ is defined as*

$$[\mathcal{G}]^\alpha := \mathfrak{I}\left(S_2([\mathcal{H}_q]^\alpha, \mathcal{H}_l)\right) = \{F \mid F = \mathfrak{I}(C),\ C \in S_2([\mathcal{H}_q]^\alpha, \mathcal{H}_l)\}.$$

The space $[\mathcal{G}]^\alpha$ is a Hilbert space equipped with the inner product

$$\langle F, G \rangle_\alpha := \langle C, L \rangle_{S_2([\mathcal{H}_q]^\alpha, \mathcal{H}_l)} \qquad (F, G \in [\mathcal{G}]^\alpha),$$

where $C = \mathfrak{I}^{-1}(F)$, $L = \mathfrak{I}^{-1}(G)$.

Remark 4 (Well-specified versus misspecified setting)**.** *We say that we are in the well-specified setting if $v_0 \in [\mathcal{G}]^1$. In this case, there exists $\bar{F} \in \mathcal{G}$ such that $v_0 = \bar{F}$ $P_{W,0}-$almost surely and $\|v_0\|_1 = \|\bar{F}\|_{\mathcal{G}}$, i.e. v_0 admits a representer in $\mathcal{G}$.*

Definition 1 and Remark 4 motivate the use of the following assumption on the smoothness of the target function: there exists $\beta > 0$ and a constant $B \geq 0$ such that $v_0 \in [\mathcal{G}]^\beta$

$$\|v_0\|_\beta \leq B. \tag{SRC}$$

We let $C_0 := \mathfrak{I}^{-1}(v_0) \in S_2([\mathcal{H}_q]^\beta, \mathcal{H}_l)$. (SRC) directly generalizes the notion of a so-called Hölder-type source condition in the learning literature (Caponnetto & De Vito, 2007; Fischer & Steinwart, 2020; Lin & Cevher, 2018; Lin et al., 2020) and allows us to characterize the misspecified learning scenario.

In addition to (SRC), we require standard assumptions to obtain the precise learning rate for kernel learning algorithms. We list them below. Recall μ_i is defined in Eq. (12). For constants $D_2 > 0$ and $p \in (0, 1]$ and for all $i \in I$,

$$\mu_i \leq D_2 i^{-1/p}. \tag{EVD}$$

(EVD) is a standard assumption on the *eigenvalue decay* of the integral operator: it describes the interplay between the marginal distribution $P_{W,0}$ and the RKHS $\mathcal{H}_q$ (see more details in Caponnetto & De Vito, 2007; Fischer & Steinwart, 2020).

Let $p_{\mathcal{S}|W}(w, dz)$ denote a regular conditional distribution of $\mathcal{S}$ given $W = w$. We assume that there exist constants $\sigma, R > 0$ such that, for $P_{W,0}$-almost every $w \in \mathcal{W}$ and every integer $r \geq 2$,

$$\int_{\mathcal{H}_l} \|z - v_0(w)\|^r_{\mathcal{H}_l} p_{\mathcal{S}|W}(w, dz) \leq \frac{1}{2} r! \sigma^2 R^{r-2}. \tag{MOM}$$

The (MOM) condition on the Markov kernel $p_{\mathcal{S}|W}(w, dz)$ is a *Bernstein moment condition* used to control the Hilbert-valued regression noise $\mathcal{S} - v_0(W)$. If $\|\mathcal{S}\|_{\mathcal{H}_l} \leq M$ almost surely, then (MOM) is satisfied with constants depending only on M. In particular, this holds if the derivative feature map $\partial_j \varphi_l$ is uniformly bounded.

Two-stage vvKRR rate. We now give the rate used to verify the nuisance conditions in the main asymptotic linearity theorem. The statement is written for a generic training fold I of size N. In the two-fold split of the main text, $N = n/2$; in K-fold cross-fitting with fixed K, $N \asymp n$. Throughout this paragraph, the coordinate $j \in [d_y]$ is fixed and the constants are uniform in j.

We overload the $\widehat{\cdot}$ notation to denote both quantities dependent on estimated regression function $\widehat{m}$ and the empirical measure $P^I_{W,N} := \frac{1}{N}\sum_{i\in I}\delta_{W_i}$. For $i \in I$, write

$$\widehat{\mathcal{S}}_i := (\partial_j\varphi_l)\{Y_i - \widehat{m}(W_i)\}, \qquad \mathcal{S}_i := (\partial_j\varphi_l)\{Y_i - m_0(W_i)\}.$$

Define the empirical and population covariance operators

$$\widehat{C}_W := \frac{1}{N}\sum_{i\in I}\varphi_q(W_i)\otimes\varphi_q(W_i), \qquad C_W := \mathbb{E}_0\{\varphi_q(W)\otimes\varphi_q(W)\},$$

and their regularized counterparts

$$\widehat{C}_{W,\lambda} := \widehat{C}_W + \lambda I_{\mathcal{H}_q}, \qquad C_{W,\lambda} := C_W + \lambda I_{\mathcal{H}_q}.$$

We recall the definition of the two-stage vvKRR estimator $\widehat{v}_\lambda$, and present the analogous oracle learner $\widetilde{v}_\lambda$:

$$\widehat{v}_\lambda = \arg\min_{v\in\mathcal{G}}\left\{\frac{1}{N}\sum_{i\in I}\|\widehat{\mathcal{S}}_i - v(W_i)\|^2_{\mathcal{H}_l} + \lambda\|v\|^2_{\mathcal{G}}\right\}, \tag{14}$$

$$\widetilde{v}_\lambda := \arg\min_{v\in\mathcal{G}}\left\{\frac{1}{N}\sum_{i\in I}\|\mathcal{S}_i - v(W_i)\|^2_{\mathcal{H}_l} + \lambda\|v\|^2_{\mathcal{G}}\right\}. \tag{15}$$

Here $\mathcal{G}$ is the vector-valued RKHS induced by $K(w,w') = q(w,w')I_{\mathcal{H}_l}$.

We define the effective dimension

$$\mathcal{N}(\lambda) := \mathrm{Tr}\{C_W(C_W+\lambda I)^{-1}\}.$$

We now establish an upper bound for $\mathcal{N}(\lambda)$ and use it to derive a useful inequality. We define the event with respect to the randomness in $P^I_{W,N}$

$$\mathfrak{E} = \left\{\left\|C^{-1/2}_{W,\lambda}(C_W - \widehat{C}_W)C^{-1/2}_{W,\lambda}\right\|_{\mathrm{op}} \le \frac{2}{3}\right\}.$$

Lemma 4. *Under (EVD), there is a constant $D < \infty$ such that the effective dimension $\mathcal{N}(\lambda) \le D\lambda^{-p}$. On the event $\mathfrak{E}$, we have*

$$\frac{1}{N}\sum_{i\in I}\left\|C^{1/2}_W\widehat{C}^{-1}_{W,\lambda}\varphi_q(W_i)\right\|^2_{\mathcal{H}_q} \le 3D\lambda^{-p}. \tag{16}$$

Furthermore, for every $\tau > 0$, if

$$N \ge 8\kappa_q^2\tau\lambda^{-1}\log\left(2e\lambda^{-p}\frac{\|C_W\|_{\mathrm{op}}+\lambda}{\|C_W\|_{\mathrm{op}}}\right), \tag{17}$$

then we have $\mathbb{P}(\mathfrak{E}) \ge 1 - 2e^{-\tau}$.

Proof. The existence of D such that inequality $\mathcal{N}(\lambda) \le D\lambda^{-p}$ holds follows from Fischer & Steinwart (2020, Lemma 11)

and (EVD). We have

$$\begin{aligned}
&\frac{1}{N}\sum_{i\in I}\left\|C_W^{1/2}\widehat{C}_{W,\lambda}^{-1}\varphi_q(W_i)\right\|_{\mathcal{H}_q}^2 \\
&= \operatorname{Tr}\left\{C_W^{1/2}\widehat{C}_{W,\lambda}^{-1}\widehat{C}_W\widehat{C}_{W,\lambda}^{-1}C_W^{1/2}\right\} \\
&= \operatorname{Tr}\left\{C_W^{1/2}\widehat{C}_{W,\lambda}^{-\frac{1}{2}}\widehat{C}_{W,\lambda}^{-\frac{1}{2}}\widehat{C}_W\widehat{C}_{W,\lambda}^{-\frac{1}{2}}\widehat{C}_{W,\lambda}^{-\frac{1}{2}}C_W^{1/2}\right\} \\
&\le \operatorname{Tr}\left\{C_W^{1/2}\widehat{C}_{W,\lambda}^{-\frac{1}{2}}\widehat{C}_{W,\lambda}^{-\frac{1}{2}}C_W^{1/2}\right\} \\
&= \operatorname{Tr}\left\{\widehat{C}_{W,\lambda}^{-1}C_W\right\}
\end{aligned}$$

where we use

$$0 \le \widehat{C}_{W,\lambda}^{-1/2}\widehat{C}_W\widehat{C}_{W,\lambda}^{-1/2} \le I_{\mathcal{H}_q},$$

by the spectral calculus, and the cyclicity of the trace in the last step. Now set

$$T_\lambda := C_{W,\lambda}^{-1/2}C_W C_{W,\lambda}^{-1/2}, \qquad R_\lambda := C_{W,\lambda}^{1/2}\widehat{C}_{W,\lambda}^{-1}C_{W,\lambda}^{1/2}.$$

The operator T_λ is positive trace class and R_λ is positive bounded, whence

$$\begin{aligned}
\operatorname{Tr}(\widehat{C}_{W,\lambda}^{-1}C_W) &= \operatorname{Tr}(R_\lambda T_\lambda) \le \|R_\lambda\|_{\mathrm{op}}\operatorname{Tr}(T_\lambda) \\
&= \left\|C_{W,\lambda}^{1/2}\widehat{C}_{W,\lambda}^{-1}C_{W,\lambda}^{1/2}\right\|_{\mathrm{op}}\operatorname{Tr}\{C_W(C_W+\lambda I)^{-1}\} \\
&\le D\lambda^{-p}\left\|C_{W,\lambda}^{1/2}\widehat{C}_{W,\lambda}^{-1}C_{W,\lambda}^{1/2}\right\|_{\mathrm{op}}.
\end{aligned}$$

We have the algebraic identity

$$C_{W,\lambda}^{1/2}\widehat{C}_{W,\lambda}^{-1}C_{W,\lambda}^{1/2} = \left[I_{\mathcal{H}_q} - C_{W,\lambda}^{-1/2}(C_W-\widehat{C}_W)C_{W,\lambda}^{-1/2}\right]^{-1}.$$

Writing the RHS as a Von Neumann series, we see on the event $\mathfrak{E}$, we have

$$\left\|C_{W,\lambda}^{1/2}\widehat{C}_{W,\lambda}^{-1}C_{W,\lambda}^{1/2}\right\|_{\mathrm{op}} \le 3.$$

Therefore Eq. (16) is proved. The last statement of the lemma on the lower bound of $\mathbb{P}(\mathfrak{E})$ under Eq. (17) follows from Fischer & Steinwart (2020, Lemma 17). □

Theorem 6 (Two-stage vvKRR rate)**.** *Assume Items 1 to 3 and Eqs.* (SRC) *to* (MOM)*. Suppose also that $r \mapsto (\partial_j\varphi_l)(r)$ is L-Lipschitz as a map $\mathbb{R}^{d_y} \to \mathcal{H}_l$. Let*

$$\lambda_N = N^{-1/(\beta+p)}.$$

Then, for every $\tau > 0$ and all sufficiently large N satisfying Eq. (17) *with $\lambda = \lambda_N$, there are constants $D, J < \infty$, independent of N and τ, such that, with probability at least $1 - 7e^{-\tau}$ under the joint law of the training fold,*

$$\left\|[\widehat{v}_{\lambda_N}] - v_0\right\|_{L^2(P_{W,0};\mathcal{H}_l)} \le L\sqrt{3D}\,\lambda_N^{-p/2}\|\widehat{m}-m_0\|_{L^2(P_{W,N}^I)} + \tau\sqrt{J}\,N^{-\beta/\{2(\beta+p)\}}. \tag{18}$$

Plugging in $\lambda_N = N^{-1/(\beta+p)}$ yields

$$\left\|[\widehat{v}_{\lambda_N}] - v_0\right\|_{L^2(P_{W,0};\mathcal{H}_l)} \le C_\tau\left\{N^{p/\{2(\beta+p)\}}\|\widehat{m}-m_0\|_{L^2(P_{W,N}^I)} + N^{-\beta/\{2(\beta+p)\}}\right\}, \tag{19}$$

where C_τ is independent of N.

Proof. By the representer theorem and the isomorphism of the vector-valued RKHS $\mathcal{G} \simeq S_2(\mathcal{H}_q,\mathcal{H}_l)$, there exist $\widehat{G}_\lambda, \widetilde{G}_\lambda \in S_2(\mathcal{H}_q,\mathcal{H}_l)$ such that

$$\widehat{v}_\lambda(w) = \widehat{G}_\lambda\varphi_q(w), \qquad \widetilde{v}_\lambda(w) = \widetilde{G}_\lambda\varphi_q(w),$$

with

$$\widehat{G}_\lambda = \left\{ \frac{1}{N} \sum_{i \in I} \widehat{\mathcal{S}}_i \otimes \varphi_q(W_i) \right\} \widehat{C}_{W,\lambda}^{-1}, \quad \widetilde{G}_\lambda = \left\{ \frac{1}{N} \sum_{i \in I} \mathcal{S}_i \otimes \varphi_q(W_i) \right\} \widehat{C}_{W,\lambda}^{-1}.$$

Consequently,

$$\widehat{G}_\lambda - \widetilde{G}_\lambda = \left\{ \frac{1}{N} \sum_{i \in I} (\widehat{\mathcal{S}}_i - \mathcal{S}_i) \otimes \varphi_q(W_i) \right\} \widehat{C}_{W,\lambda}^{-1}.$$

Using the $L^2(P_{W,0})$ representation of vector-valued RKHS functions,

$$\|[\widehat{v}_\lambda] - [\widetilde{v}_\lambda]\|_{L^2(P_{W,0};\mathcal{H}_l)} = \|(\widehat{G}_\lambda - \widetilde{G}_\lambda) C_W^{1/2}\|_{S_2(\mathcal{H}_q,\mathcal{H}_l)}.$$

Let $b_i = C_W^{1/2} \widehat{C}_{W,\lambda}^{-1} \varphi_q(W_i)$. Applying the Cauchy-Schwarz inequality in the vector space $S_2(\mathcal{H}_q, \mathcal{H}_l)$ gives

$$\|[\widehat{v}_\lambda] - [\widetilde{v}_\lambda]\|_{L^2(P_{W,0};\mathcal{H}_l)} \le \left(\frac{1}{N} \sum_{i \in I} \|\widehat{\mathcal{S}}_i - \mathcal{S}_i\|_{\mathcal{H}_l}^2 \right)^{1/2} \left(\frac{1}{N} \sum_{i \in I} \|b_i\|_{\mathcal{H}_q}^2 \right)^{1/2}.$$

Since $\partial_j \varphi_l$ is L-Lipschitz, we have

$$\left(\frac{1}{N} \sum_{i \in I} \|\widehat{\mathcal{S}}_i - \mathcal{S}_i\|_{\mathcal{H}_l}^2 \right)^{1/2} \le L \|\widehat{m} - m_0\|_{L^2(P_{W,N}^I)}.$$

Therefore, on the event $\mathfrak{E}$,

$$\begin{aligned} \|[\widehat{v}_\lambda] - [\widetilde{v}_\lambda]\|_{L^2(P_{W,0};\mathcal{H}_l)} &\le L \|\widehat{m} - m_0\|_{L^2(P_{W,N}^I)} \left(\frac{1}{N} \sum_{i \in I} \|C_W^{1/2} \widehat{C}_{W,\lambda}^{-1} \varphi_q(W_i)\|_{\mathcal{H}_q}^2 \right)^{1/2} \\ &\le L\sqrt{3D} \lambda^{-p/2} \|\widehat{m} - m_0\|_{L^2(P_{W,N}^I)}. \end{aligned} \tag{20}$$

Under (SRC), (EVD), and (MOM), Li et al. (2024, Theorem 3) says that, for $\lambda_N \asymp N^{-1/(\beta+p)}$, there exists $J < \infty$ such that with probability at least $1 - 5e^{-\tau}$,

$$\|[\widetilde{v}_{\lambda_N}] - v_0\|_{L^2(P_{W,0};\mathcal{H}_l)} \le \tau \sqrt{J}\, N^{-\beta/\{2(\beta+p)\}}.$$

Taking a union bound with event $\mathfrak{E}$, which satisfies $\mathbb{P}(\mathfrak{E}) \ge 1 - 2e^{-\tau}$ under Eq. (17), we have that Eq. (18) holds with probability at least $1 - 7e^{-\tau}$. □

Consistency of the residual mean embedding.

Lemma 5 (Consistency and rate for $\widehat{\mu}_\xi$)**.** *Assume* $\sup_u l(u,u) \le \kappa_l^2$ *and* $\|\varphi_l(u) - \varphi_l(v)\|_{\mathcal{H}_l} \le L_l \|u - v\|_{\mathbb{R}^{d_y}}$. *For a training fold I of size N, define*

$$\widehat{\mu}_\xi^I := \frac{1}{N} \sum_{i \in I} \varphi_l\{Y_i - \widehat{m}(W_i)\}, \qquad \mu_{\xi,0} := P_0 \varphi_l\{Y - m_0(W)\}.$$

Then

$$\|\widehat{\mu}_\xi^I - \mu_{\xi,0}\|_{\mathcal{H}_l} \le L_l \|\widehat{m} - m_0\|_{L^2(P_{W,N}^I)} + \left\| \frac{1}{N} \sum_{i \in I} [\varphi_l\{Y_i - m_0(W_i)\} - \mu_{\xi,0}] \right\|_{\mathcal{H}_l}. \tag{21}$$

The second term in Eq. (21) is $O_p(N^{-1/2})$. Hence, if $\|\widehat{m} - m_0\|_{L^2(P_{W,N}^I)} = o_p(1)$, then $\|\widehat{\mu}_\xi^I - \mu_{\xi,0}\|_{\mathcal{H}_l} = o_p(1)$; if $\|\widehat{m} - m_0\|_{L^2(P_{W,N}^I)} = O_p(a_N)$, then $\|\widehat{\mu}_\xi^I - \mu_{\xi,0}\|_{\mathcal{H}_l} = O_p(a_N + N^{-1/2})$.

Proof. By the triangle inequality,

$$\|\widehat{\mu}_\xi^I - \mu_{\xi,0}\|_{\mathcal{H}_l} \le \left\| \frac{1}{N}\sum_{i\in I}[\varphi_l\{Y_i - \widehat{m}(W_i)\} - \varphi_l\{Y_i - m_0(W_i)\}] \right\|_{\mathcal{H}_l} + \left\| \frac{1}{N}\sum_{i\in I}[\varphi_l\{Y_i - m_0(W_i)\} - \mu_{\xi,0}] \right\|_{\mathcal{H}_l}.$$

The first term is bounded by $L_l N^{-1}\sum_{i\in I}\|\widehat{m}(W_i) - m_0(W_i)\| \le L_l\|\widehat{m} - m_0\|_{L^2(P^I_{W,N})}$. For the second term, the summands are i.i.d., mean zero, and have norm at most $2\kappa_l$. Therefore its second moment is at most $4\kappa_l^2/N$, which implies the asserted $O_p(N^{-1/2})$ rate. □

I.1. Sufficient conditions for asymptotic normality

We now instantiate Requirements 1B and 1C.

Corollary 1 (Rate-based sufficient conditions)**.** *Consider a fixed number K of folds and let all training-fold sizes satisfy $N_k \asymp n$. Suppose Assumptions 1 to 4 and requirement 1A hold. For each fold k, let $\widehat{m}^{(-k)}$ be the first-stage estimator and let $\widehat{v}_j^{(-k)}$ be the feasible vvKRR estimator in (14), fitted with $\lambda_{N_k} \asymp N_k^{-1/(\beta+p)}$. Assume that, for some deterministic $a_n \downarrow 0$,*

$$\max_{k\le K}\|\widehat{m}^{(-k)} - m_0\|_{L^2(P_{W,0})} = O_p(a_n), \tag{22}$$

$$\max_{k\le K}\|\widehat{m}^{(-k)} - m_0\|_{L^2(P^{(-k)}_{W,N_k})} = O_p(a_n), \tag{23}$$

$$\max_{k\le K}\|\widehat{m}^{(-k)} - m_0\|_{L^\infty(P_{W,0})} = o_p(1). \tag{24}$$

Assume also that (SRC)–(MOM) hold for every $j \in [d_y]$, with constants uniform in j, and that $r \mapsto \partial_j\varphi_l(r)$ is Lipschitz uniformly in j. Assume additionally that $\beta + p > 1$, so that (17) holds for all sufficiently large N when $\lambda_N = N^{-1/(\beta+p)}$. We have

$$\max_{k\le K}\max_{j\le d_y}\|\widehat{v}_j^{(-k)} - v_{0,j}\|_{L^2(P_{W,0};\mathcal{H}_l)} = O_p\left(n^{\frac{p}{2(\beta+p)}}a_n + n^{-\frac{\beta}{2(\beta+p)}}\right). \tag{25}$$

Moreover, Requirements 1B and 1C hold provided

$$a_n^2 + a_n\left(n^{\frac{p}{2(\beta+p)}}a_n + n^{-\frac{\beta}{2(\beta+p)}}\right) = o(n^{-1/2}). \tag{26}$$

In particular, if $a_n = n^{-\gamma_m}$, then Eq. (26) is implied by

$$2\gamma_m + \frac{\beta}{2(\beta+p)} > 1. \tag{27}$$

Proof. Since K is fixed and $N_k \asymp n$, Theorem 6 and Eq. (23) give Eq. (25), uniformly over folds and coordinates. The L^2 and L^∞ consistency of $\widehat{m}^{(-k)}$ follow from Eq. (22) and Eq. (24). The consistency of $\widehat{v}^{(-k)}$ follows from Eq. (25) whenever the right-hand side is $o_p(1)$; this is implied by Eq. (26). By Lemma 5,

$$\max_{k\le K}\|\widehat{\mu}_\xi^{(-k)} - \mu_{\xi,0}\|_{\mathcal{H}_l} = O_p(a_n + n^{-1/2}) = o_p(1),$$

and the analogous bound $\max_k\|\widehat{\mu}_X^{(-k)} - \mu_{X,0}\|_{\mathcal{H}_k} = O_p(n^{-1/2})$ follows from boundedness of k. This verifies Requirement 1B.

For Requirement 1C, combine Eq. (22) and Eq. (25):

$$\|\widehat{m} - m_0\|^2_{L^2(P_{W,0})} + \|\widehat{m} - m_0\|_{L^2(P_{W,0})}\|\widehat{v} - v_0\|_{L^2(P_{W,0})} = O_p\Big(a_n^2 + a_n(n^{\frac{p}{2(\beta+p)}}a_n + n^{-\frac{\beta}{2(\beta+p)}})\Big).$$

Thus Eq. (26) gives Requirement 1C.

If $a_n = n^{-\gamma_m}$, the three terms in Eq. (26) have exponents $2\gamma_m$, $2\gamma_m + \frac{\beta}{2(\beta+p)} - 1/2$, and $\gamma_m + \frac{\beta}{2(\beta+p)}$. The condition $2\gamma_m + \frac{\beta}{2(\beta+p)} > 1$ implies each term is $o(n^{-1/2})$. □

Instantiation via Sobolev spaces. We now exhibit sufficient conditions for Corollary 1 in terms of smoothness conditions. Let d_w be the ambient dimension of W.

We use the following terminology from Chen et al. (2025). Let $\mathcal{W} \subseteq \mathbb{R}^{d_w}$. A positive definite kernel q on $\mathcal{W}$ is called a Sobolev reproducing kernel of smoothness $a > 0$ if its RKHS $\mathcal{H}_q$ is norm-equivalent to the real-valued Sobolev space $H^a(\mathcal{W}) = W_2^a(\mathcal{W})$. That is, $\mathcal{H}_q$ and $H^a(\mathcal{W})$ contain the same functions and their norms are equivalent. In particular, when $a > d_w/2$, the Sobolev embedding theorem (Adams & Fournier, 2003) implies that functions in $\mathcal{H}_q$ have continuous representatives and that point evaluation is bounded. Hence Sobolev reproducing kernels of smoothness $a > d_w/2$ are bounded on compact domains. A standard example is the Matérn-ν kernel, whose RKHS is norm-equivalent to a Sobolev space of smoothness $a = \nu + d_w/2$.

Assumption 7 (Sobolev smoothness)**.** *The following conditions hold.*

(S1) *$\mathcal{W} \subset \mathbb{R}^{d_w}$ is a compact Lipschitz domain, and $P_{W,0}$ has a density with respect to Lebesgue measure bounded above and below by positive constants on $\mathcal{W}$.*

(S2) *The regression function satisfies $m_0 \in H^{s_m}(\mathcal{W}; \mathbb{R}^{d_y})$ for some $s_m > d_w/2$. The first-stage estimator is a kernel ridge regression estimator with a Sobolev reproducing kernel, with regularization chosen so that, up to logarithmic factors,*

$$\|\widehat{m} - m_0\|_{L^2(P_{W,0})} = O_p(n^{-\gamma_m}), \qquad \gamma_m := \frac{s_m}{2s_m + d_w},$$

together with the empirical and sup-norm consistency requirements in Eq. (23) and Eq. (24).

(S3) *For every $j \in [d_y]$, the function*

$$v_{0,j}(w) = \mathbb{E}_0[(\partial_j \varphi_l)\{Y - m_0(W)\} \mid W = w]$$

belongs to $H^{s_v}(\mathcal{W}; \mathcal{H}_l)$ for some $s_v > 0$, uniformly in j.

(S4) *The second-stage kernel q is a Sobolev reproducing kernel of smoothness $a_v > d_w/2$ on $\mathcal{W}$. Assume $0 < s_v \le 2a_v$ and $a_v < s_v + d_w/2$. In the vvKRR estimator of $v_{0,j}$, use*

$$\lambda_{v,n} \asymp n^{-2a_v/(2s_v+d_w)}.$$

Corollary 2 (Sobolev sufficient conditions for asymptotic normality)**.** *Suppose Assumptions 1 to 4 and 7 and requirement 1A hold. If*

$$\frac{2s_m}{2s_m + d_w} + \frac{s_v}{2s_v + d_w} > 1, \tag{28}$$

then Requirements 1B and 1C hold.

Proof. Fix $j \in [d_y]$. Since q is a Sobolev reproducing kernel of smoothness $a_v > d_w/2$ on the compact Lipschitz domain $\mathcal{W}$, the preliminary kernel assumptions on q in Items 1 to 3 are satisfied.

By the Sobolev equivalence for $\mathcal{H}_q$, the interpolation space $[\mathcal{H}_q]^\beta$ is norm-equivalent to $H^{a_v\beta}(\mathcal{W})$. Taking $\beta = \frac{s_v}{a_v}$ therefore gives $[\mathcal{H}_q]^\beta \simeq H^{s_v}(\mathcal{W})$. Hence Item (S3) implies the source condition (SRC). The standard eigenvalue decay for Sobolev kernels on compact domains, together with the density assumption in Item (S1), gives (EVD) with $p = \frac{d_w}{2a_v}$.

It remains to check (MOM) and Lipschitz assumptions used in Theorem 6. By Assumption 3,

$$M_j := \sup_{r \in \mathbb{R}^{d_y}} \|\partial_j \varphi_l(r)\|_{\mathcal{H}_l} < \infty.$$

Thus

$$\|\mathcal{S}_j\|_{\mathcal{H}_l} = \|(\partial_j \varphi_l)\{Y - m_0(W)\}\|_{\mathcal{H}_l} \leq M_j$$

almost surely, and also $\|v_{0,j}(W)\|_{\mathcal{H}_l} \leq M_j$ almost surely. Hence

$$\|\mathcal{S}_j - v_{0,j}(W)\|_{\mathcal{H}_l} \leq 2M_j$$

almost surely, which implies the Bernstein moment condition (MOM). Moreover, the uniform bound on $D^2\varphi_l$ in Assumption 4 implies that $D\varphi_l$ is Lipschitz, and therefore each coordinate derivative $r \mapsto \partial_j \varphi_l(r)$ is Lipschitz as a map $\mathbb{R}^{d_y} \to \mathcal{H}_l$.

By Item (S2), the first-stage estimator satisfies Eq. (22)–Eq. (24) with

$$a_n = n^{-\gamma_m}, \qquad \gamma_m = \frac{s_m}{2s_m + d_w},$$

up to logarithmic factors. Thus Corollary 1 applies with

$$\gamma_m = \frac{s_m}{2s_m + d_w}, \qquad \frac{\beta}{2(\beta + p)} = \frac{s_v}{2s_v + d_w}.$$

The condition Eq. (28) is exactly

$$2\gamma_m + \frac{\beta}{2(\beta + p)} > 1,$$

which is sufficient for Requirement 1C by Corollary 1. □