Research Article

Rui Zhang, Masaaki Imaizumi, Bernhard Schölkopf, and Krikamol Muandet*

Instrumental variable regression via kernel maximum moment loss

Abstract: We investigate a simple objective for nonlinear instrumental variable (IV) regression based on a kernelized conditional moment restriction known as a maximum moment restriction (MMR). The MMR objective is formulated by maximizing the interaction between the residual and the instruments belonging to a unit ball in a reproducing kernel Hilbert space. First, it allows us to simplify the IV regression as an empirical risk minimization problem, where the risk function depends on the reproducing kernel on the instrument and can be estimated by a U-statistic or V-statistic. Second, on the basis this simplification, we are able to provide consistency and asymptotic normality results in both parametric and nonparametric settings. Finally, we provide easy-to-use IV regression algorithms with an efficient hyperparameter selection procedure. We demonstrate the effectiveness of our algorithms using experiments on both synthetic and real-world data.

Keywords: nonlinear instrumental variable regression, conditional moment restriction, kernel method, reproducing kernel Hilbert space, U/V-statistics

MSC 2020: 46E22, 62G08, 34A55

1 Introduction

Instrumental variables (IVs) have become standard tools for economists, epidemiologists, and social scientists to uncover causal relationships from observational data [1,2]. Randomization of treatments or policies has been perceived as the gold standard for such tasks, but it is generally prohibitive in many real-world scenarios due to time constraints or ethical concerns. When treatment assignment is not randomized, it is generally impossible to discern between the causal effect of treatments and spurious correlations that are induced by unobserved factors. Instead, IVs enable the investigators to incorporate *natural* variation through an IV that is associated with the treatments, but not with the outcome variable, other than through its effect on the treatments. In economics, for instance, the season of birth was used as an IV to study the return from schooling, which measures causal effect of education on labor market earning [3]. In genetic epidemiology, the idea to use genetic variants as IVs, known as Mendelian randomization, has also gained increasing popularity [4,5].

* **Corresponding author: Krikamol Muandet,** CISPA–Helmholtz Center for Information Security, Saarbrücken, Saarland, Germany, e-mail: muandet@cispa.de

Rui Zhang: Research School of Computer Science, Australian National University, Canberra ACT 2601, Australia, e-mail: rui.zhang@anu.edu.au

Masaaki Imaizumi: Komaba Institute for Science (KIS), The University of Tokyo, 7 Chome-3-1 Hongo, Bunkyo City, Tokyo 113-8654, Japan, e-mail: imaizumi@g.ecc.u-tokyo.ac.jp

Bernhard Schölkopf: Empirical Inference Department, Max Planck Institute for Intelligent Systems, Max-Planck-Ring 4, 72076 Tübingen, Germany, e-mail: bs@tuebingen.mpg.de

There are numerous variations of methods for dealing with IV. Classical methods for IV regression often rely on a linearity assumption in which a two-stage least squares (2SLS) is the most popular technique [6]. The generalized method of moments (GMM) of ref. [7], which imposes the orthogonality restrictions, can also be used for the linear IV regression. For nonlinear regression, numerous methodologies have emerged in the field of nonparametric IV [8–11] and recent machine learning [12–18]. These estimators can be categorized into two general approaches. The former generalizes a two-stage optimization procedure [12,16], whereas the latter requires solving a minimax optimization problem [13,14,17,19,20]. While the equivalence between 2SLS and GMM is known in the linear setting, the connection between two-stage procedure and minimax optimization in the nonlinear setting was established recently, see, e.g., refs [15], [17, Appendix F], and [21, Sec 3.3]. For statistical inference, numerous approaches have been developed using asymptotic distributions of each point of uniform bands of the target functions by the sieve method [22–26].

In this article, we propose a novel nonlinear IV regression framework named *Maximum Moment Restriction-IV* (MMR-IV), which utilizes the reproducing kernel Hilbert spaces (RKHSs) and an associated positive definite kernel from the machine learning domain.

The MMR-IV possesses mainly two important features: (i) It reformulates the original conditional moment restriction (CMR) into a single-step empirical risk minimization (ERM) problem, by using the MMR framework of ref. [27] with the kernel and RKHSes. (ii) Based on the ERM problem, it reveals a closed form of the solution by U/V-statistics techniques [28]. Based on this framework, we propose two practical algorithms based on kernel functions and neural networks (NNs), which we call MMR-IV (RKHS) and MMR-IV (NN), respectively.

Our MMR-IV has the following advantages. First, the derived closed-form expression is computationally simple and stable. Further, owing to the generic structure of ERM, we can substitute NNs for kernel functions, which is suitable for highly nonlinear modeling. Second, MMR-IV can identify the optimal solution, i.e., the minimizer of the risk function is guaranteed to be unique, by using the positive definite property of the kernel function. Third, we can identify a distribution that the solution follows in the large sample limit, i.e., the asymptotic distribution. The identified asymptotic distribution allows us to analyze uncertainties of algorithms and estimators, such as statistical tests and confidence intervals. This result is applicable to the case where the model is both parametric and nonparametric, and especially in the nonparametric case, we utilize the functional differentiation technique [29] to derive the results. Moreover, on the basis of the asymptotic distribution, we discuss how the choice of kernel functions affects the efficiency of MMR-IV. Fourth, we provide a hyperparameter selection scheme, by developing a novel interpretation of MMR-IV from a Gaussian process (GP) perspective. This interpretation is obtained by mapping our ERM form to a likelihood of GPs. Since hyperparameter selection for IV regression is a nontrivial problem in both minimax and two-stage formulation, our framework sheds light on efficient methods to solve this problem. In the experimental section, we demonstrate the superiority of the proposed framework by validating its accuracy and effectiveness and show its usefulness in real data analysis.

Our contributions can be summarized as follows:

(i) We study the simple ERM problem converted from the minimax optimization, and then propose the MMR-IV algorithms to minimize the risk based on kernel functions and NNs. This algorithm avoids the difficulties of minimax problems and is computationally simple and stable.

(ii) Our method can identify unique solutions by selecting positive definite kernel functions. Our analysis also reveals how the asymptotic efficiency varies with the choice of kernel functions.

(iii) We derive the asymptotic normal distribution of the MMR-IV estimators in both parametric and non-parametric settings. This result is derived from the ERM form and allows for statistical inference.

(iv) We develop the efficient framework to select hyperparameters of MMR-IV. This approach is based on the interpretation of MMR-IV with kernel functions from a GP perspective.

(v) We compare our methods to a wide range of baselines on different settings and show that MMR-IV has competitive performance.

The rest of this article is organized as follows. Section 2 introduces the IV regression problem and the MMR framework. Our methods are presented in Section 3, followed by the hyperparameter selection in Section 4. Next, we provide the consistency and asymptotic normality results in Section 5. We then provide a thorough discussion on related works in Section 6. Finally, the experimental results are presented in Section 7. All proofs of the main results can be found in the appendix.

2 Preliminaries

2.1 IV regression

Following standard setting in the literature [12–16], let X be a treatment (endogenous) variable taking value in $\mathcal{X} \subseteq \mathbb{R}^d$ and Y a real-valued outcome variable. Our goal is to estimate a function $f : \mathcal{X} \to \mathbb{R}$ from a structural equation model of the form

$$Y = f(X) + \varepsilon, \quad X = t(Z) + g(\varepsilon) + \nu, \tag{1}$$

where we assume that $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\nu] = 0$. Unfortunately, as we can see from (1), ε is correlated with the treatment X, i.e., $\mathbb{E}[\varepsilon|X] \neq 0$, and hence, standard regression methods cannot be used to estimate f. This setting arises, for example, when there exist unobserved confounders (i.e., common causes) between X and Y. To illustrate the problem, let us consider an example taken from ref. [12], which aims at predicting sales of airline ticket Y under an intervention in price of the ticket X. However, there exist unobserved variables that may affect both sales and ticket price, e.g., conferences, COVID-19 pandemic. This creates a correlation between ε and X in (1) that prevents us from applying the standard regression toolboxes directly on observational data.

Instrumental variable (**IV**). To address this problem, we assume access to an *instrumental* variable Z taking value in $\mathcal{Z} \subseteq \mathbb{R}^{d'}$. As we can see in (1), the instrument Z is associated with the treatments X, but not with the outcome Y, other than through its effect on the treatments. Formally, Z must satisfy (i) *Relevance*: Z has a causal influence on X, (ii) *Exclusion restriction: Z* affects Y only through X, i.e., $Y \perp\!\!\!\perp Z|X, \varepsilon$, and (iii) *Unconfounded instrument(s)*: Z is independent of the error, i.e., $\varepsilon \perp\!\!\!\perp Z$. See Figure 1 for a visualization of the three conditions. For example, the instrument Z may be the cost of fuel, which influences sales Y only via price X. For detailed exposition on IV, we refer the readers to refs [1,8].

Conditional moment restriction (**CMR**). Together with the structural assumption in (1), these conditions imply that $\mathbb{E}[\varepsilon \mid Z] = 0$ for P_Z-almost all z. This is known as a *CMR*, which we can use to estimate f [30]. For every measurable function h, the CMR implies a continuum of unconditional moment restrictions [13,14][1]

$$\mathbb{E}[(Y - f(X))h(Z)] = 0. \tag{2}$$

That is, there exists an infinite number of moment conditions, each of which is indexed by the function h. One of the key questions in econometrics is which moment condition should be used as a basis for estimating the function f [31,32]. In this work, we show that, for the purpose of consistently estimating f, it is sufficient to restrict h to be within a unit ball of a RKHS of real-valued functions on Z.

2.2 Maximum moment restriction (MMR)

Throughout this article, we assume that h is a real-valued function on $\mathcal{Z}$, which belongs to an RKHS $\mathcal{H}_k$ endowed with a reproducing kernel $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$. The RKHS $\mathcal{H}_k$ satisfies two important properties: for

---

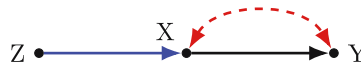**1** Also, IV assumptions imply ε is independent of Z, i.e., $\mathbb{E}[(Y - f(X))h(Z)] = 0$ for all measurable h.

Figure 1: A causal graph depicting an IV Z that satisfies an exclusion restriction and unconfoundedness (there may be a confounder ε acting on X and Y, but it is independent of Z).

all $z \in \mathcal{Z}$ and $h \in \mathcal{H}_k$, (i) $k(z, \cdot) \in \mathcal{H}_k$ and (ii) (reproducing property) $h(z) = \langle h, k(z, \cdot) \rangle_{\mathcal{H}_k}$ where $k(z, \cdot)$ is a function of the second argument. Furthermore, we define $\Phi_k(z)$ as a *canonical* feature map of z in $\mathcal{H}_k$. It follows from the reproducing property that $k(z, z') = \langle \Phi_k(z), \Phi_k(z') \rangle_{\mathcal{H}_k}$ for every $z, z' \in \mathcal{Z}$, i.e., an inner product between the feature maps of z and z' can be evaluated through the kernel evaluation. Every positive definite kernel k uniquely determines the RKHS for which k is a reproducing kernel [33]. For detailed exposition on kernel methods, see, e.g., [34–36].

Instead of considering all measurable functions h in (2) as instruments, we only restrict to functions that lie within a unit ball of the RKHS $\mathcal{H}_k$. The risk functional is then defined as a maximum value of the moment restriction with respect to this function class:

$$R_k(f) := \sup_{h \in \mathcal{H}_k, \|h\| \leq 1} (\mathbb{E}[(Y - f(X))h(Z)])^2. \tag{3}$$

The benefits of this formulation are twofold. First, it is computationally intractable to learn f from (2) using *all* measurable functions as instruments. By restricting the function class to a unit ball of the RKHS, the problem becomes computationally tractable, as will be shown in Lemma 1. Second, this restriction still preserves the consistency of parameters estimated using $R_k(f)$. In other words, the RKHS is a sufficient class of instruments for the nonlinear IV problem (cf. Theorem 1). A crucial insight for our approach is that the population risk $R_k(f)$ has an analytic solution.

Lemma 1. ([27], Theorem 3.3) *Assume that* $\mathbb{E}[(Y - f(X))^2 k(Z, Z)] < \infty$. *Then, we have the following closed-form expression, where* (X', Y', Z') *is an independent copy of* (X, Y, Z):

$$R_k(f) = \mathbb{E}[(Y - f(X))(Y' - f(X'))k(Z, Z')]. \tag{4}$$

We assume throughout that the reproducing kernel k is integrally strictly positive definite (ISPD).

Assumption 1. The kernel k is continuous, bounded (i.e., $\sup_{z \in \mathcal{Z}} \sqrt{k(z, z)} < \infty$), and satisfies the condition of ISPD kernels, i.e., for every function g that satisfies $0 < \|g\|_2^2 < \infty$, we have $\iint_{\mathcal{Z}} g(z)k(z, z')g(z')\mathrm{d}z\mathrm{d}z' > 0$.

Popular kernel functions that satisfy Assumption 1 are the Gaussian RBF kernel and Laplacian kernel

$$k(z, z') = \exp\left(-\frac{\|z - z'\|_2^2}{2\sigma^2}\right), \qquad k(z, z') = \exp\left(-\frac{\|z - z'\|_1}{\sigma}\right),$$

where σ is a positive bandwidth parameter. Another important kernel satisfying this assumption is an inverse multiquadric kernel:

$$k(z, z') = (c^2 + \|z - z'\|_2^2)^{-\gamma}$$

where c and γ are positive parameters [37, Ch. 4]. This class of kernel functions is closely related to the notions of universal kernels [38] and characteristic kernels [39]. The former ensures that kernel-based classification/regression algorithms can achieve the Bayes risk, whereas the latter ensures that the kernel mean embeddings can distinguish different probability measures. In principle, they guarantee that the corresponding RKHSs induced by these kernels are sufficiently rich for the tasks at hand. We refer the readers to refs [40,41] for more details.

Next, we further assume the identification for the minimizer of $R_k(f)$.

Assumption 2. Consider the function space $\mathcal{F}$ and $f^* \in \operatorname{argmin}_{f \in \mathcal{F}} R_k(f)$. Then for every $g \in \mathcal{F}$ with $|\mathbb{E}[g(X)]| < \infty$, $\mathbb{E}[g(X) - f^*(X) \mid Z] = 0$ implies $g = f^*$ almost everywhere with respect to P_X.

A sufficient condition for identification follows from the *completeness* property of X for Z, e.g., the conditional distribution of X given Z belongs to the exponential family [8]. See ref. [42] for more general sufficient conditions. Provided identification, it is straightforward to obtain consistency.

3 Our method

We propose to learn f by minimizing $R_k(f)$ in (4). To this end, we define an optimal function f^* as a minimizer of the aforementioned population risk with respect to a function class $\mathcal{F}$ of real-valued functions on X, i.e.,

$$f^* \in \operatorname*{argmin}_{f \in \mathcal{F}} R_k(f).$$

It is instructive to note that population risk R_k depends on the choice of the kernel k. On the basis of Assumption 1 and Lemma 1, we obtain the following result by applying [27, Theorem 3.2], showing that $R_k(f) = 0$ if and only if f satisfies the original CMR (see Appendix A.2 for the proof).

Proposition 1. *Assume that the kernel k is ISPD. Then for every real-valued measurable function f, $R_k(f) = 0$ if and only if $\mathbb{E}[Y - f(X) \mid Z = z] = 0$ for P_Z-almost every z.*

Proposition 1 holds as long as the kernel k belongs to a class of ISPD kernels. Hence, it allows for more flexibility in terms of kernel choice. Moreover, it is not difficult to show that $R_k(f)$ is strictly convex in f; see the proof in Appendix A.3.

Proposition 2. *If $\mathcal{F}$ is a convex set and Assumptions 1 and 2 hold, then the risk R_k given in (4) is strictly convex on $\mathcal{F}$.*

3.1 Empirical risk minimization

The previous results pave the way for an ERM framework [43] to be used in our work. That is, given an i.i.d. sample $\{(x_i, y_i, z_i)\}_{i=1}^n \sim P^n(X, Y, Z)$ of size n, empirical estimates of the risk $R_k(f)$ can be obtained as in the form of U-statistic or V-statistic [28, Section 5]:

$$\widehat{R}_U(f) := \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} (y_i - f(x_i))(y_j - f(x_j)) k(z_i, z_j)$$

and

$$\widehat{R}_V(f) := \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n (y_i - f(x_i))(y_j - f(x_j)) k(z_i, z_j).$$

Both forms of empirical risk can be used as a basis for a consistent estimation of f. The advantage of $\widehat{R}_U$ is that it is a minimum-variance unbiased estimator with appealing asymptotic properties, whereas $\widehat{R}_V$ is a biased estimator of the population risk (4), i.e., $\mathbb{E}[\widehat{R}_V] \neq R_k$. However, the estimator based on V-statistics employs full pairs of samples and hence may yield a better estimate of the risk than the U-statistic counterpart. Let $\mathbf{x} := [x_1, \ldots, x_n]^\top$, $\mathbf{y} := [y_1, \ldots, y_n]^\top$ and $\mathbf{z} := [z_1, \ldots, z_n]^\top$ be column vectors. Let $K_{\mathbf{z}}$ be the kernel matrix $K(\mathbf{z}, \mathbf{z}) = [k(z_i, z_j)]_{ij}$ evaluated on the instruments $\mathbf{z}$ and $f(\mathbf{x}) := [f(x_1), \ldots, f(x_n)]^\top$. Then, both $\widehat{R}_U$ and $\widehat{R}_V$ can be rewritten as follows:

$$\widehat{R}_{V(U)}(f) = (\boldsymbol{y} - f(\boldsymbol{x}))^{\top} W_{V(U)}(\boldsymbol{y} - f(\boldsymbol{x})), \tag{5}$$

where $W_{V(U)}$ is a $n \times n$ symmetric weight matrix that depends on the kernel matrix $K_{\boldsymbol{z}}$. Specifically, $W_U = (K_{\boldsymbol{z}} - \mathrm{diag}(k(z_1, z_1), \ldots, k(z_n, z_n)))/(n(n-1))$ corresponds to $\widehat{R}_U$, where $\mathrm{diag}(a_1, \ldots, a_n)$ denotes an $n \times n$ diagonal matrix whose diagonal elements are $a_1, \ldots, a_n$. As shown in Appendix A.5, W_U is indefinite and may cause problematic inferences. For $\widehat{R}_V$, $W_V \coloneqq K_{\boldsymbol{z}}/n^2$ is positive definite because the kernel k is ISPD. Finally, our objective (5) also resembles the well-known generalized least regression with correlated noise [44, Chapter 2], where the covariance matrix is the Z-dependent invertible matrix $W_{V(U)}^{-1}$.

On the basis of $\widehat{R}_U$ and $\widehat{R}_V$, we estimate f^* by minimizing the *regularized* empirical risk over $\mathcal{F}$:

$$\hat{f}_{V(U)} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \ \widehat{R}_{V(U)}(f) + \lambda \Omega(f),$$

where $\lambda > 0$ is a regularization constant satisfying $\lim_{n \to \infty} \lambda = 0$, and $\Omega(f)$ is the regularizer. Since $\hat{f}_U$ and $\hat{f}_V$ minimize objectives that are regularized U-statistic and V-statistic, they can be viewed as a specific form of M-estimators; see, e.g., [45, Ch. 5]. In this work, we focus on the V-statistic empirical risk and provide practical algorithms when $\mathcal{F}$ is parametrized by deep NNs and an RHKS of real-valued functions.

Kernelized GMM. We may view the objective (5) from the GMM perspective [32]. The assumption that the instruments Z are exogenous implies that $\mathbb{E}[\Phi_k(Z)\varepsilon] = 0$, where Φ_k denotes the canonical feature map associated with the kernel k. This gives us an infinite number of moments, $g(f) = \mathbb{E}[\Phi_k(Z)(Y - f(X))]$. Hence, we can write the sample moments as $\hat{g}(f) = (1/n)\sum_{i=1}^n \Phi_k(z_i)(y_i - f(x_i))$. The intuition behind GMM is to choose a function f that sets these moment conditions as close to zero as possible, motivating the objective function

$$J(f) \coloneqq \|\hat{g}(f)\|_{\mathcal{H}_k}^2 = \langle \hat{g}(f), \hat{g}(f) \rangle_{\mathcal{H}_k} = \frac{1}{n^2} \sum_{i,j} (y_i - f(x_i)) \langle \Phi_k(z_i), \Phi_k(z_j) \rangle_{\mathcal{H}_k} (y_j - f(x_j)) = \widehat{R}_V(f).$$

Hence, our objective $\widehat{R}_V(f)$ (5) is a special case of the GMM objective when the weighting matrix is the identity operator. Carrasco et al. [46, Ch. 6] showed the optimal weighting operator in terms of the inversed covariance operator.

3.1.1 Theoretical perspective: Identification and efficiency

Before introducing practical algorithms, we discuss the theoretical aspects of our framework, specifically, the *identifiability* and *efficiency* issues by changing the norm constraint in the formulation. The key difference between our risk $R_k(f)$ in (3) and the ordinary framework of the unconstrained moment restriction is that, in the ordinary framework, the CMR (2) is rewritten as the following unconditional moment restriction:

$$\sup_{h:\mathbb{E}[(f(X))^2]<\infty} (\mathbb{E}[(Y - f(X))h(Z)])^2. \tag{6}$$

Note that the test function h is maximized over a class of square-integrable functions, i.e., $\mathbb{E}[(h(X))^2] < \infty$. In contrast, our risk $R_k(f)$ utilizes a maximized test function h over a unit ball of the RKHS, that is, $h \in \mathcal{H}_k, \|h\| \leq 1$ with the RKHS norm $\|\cdot\|$. This arbitrary design with the RKHS plays a critical role in our method.

Identification: An advantage of our design (3) with RKHSes is that it uniquely defines the function that minimizes the risk. Since RKHSes equips a positive definite kernel, and we further assume the kernel is ISPD in Assumption 1, the risk $R_k(f)$ has a positive definite Hessian matrix (operator) at the minimum, and hence, it has the unique minimizer. In short, the loss design with RKHSes has the advantage that the function class of h can be chosen so that the solution is uniquely identified.

Efficiency of IV: Another important issue is the efficiency of the estimator, i.e., the variance of the asymptotic distribution of estimators. For the linear regression problem with finite-dimensional parameters, we can define the *optimal IV*, i.e., the estimator with the IV that minimizes the asymptotic variance [8]. In contrast, our regression model in this setup is nonlinear, infinite-dimensional, and also based on U-statistics, hence it is not trivial to discuss optimality.

To address the latter issue, we provide the variance of the asymptotic distribution of our estimator as a way to measure efficiency. In Proposition 5 in Section 5, we derive asymptotic Gaussian approximations of the estimator, both pointwise and uniformly, that reveal their variance. The result in Proposition 5 is summarized as follows. We first define coefficients $\Lambda_{\max}$ and $\Lambda_{\min}$ as follows:

$$\Lambda_{\max} := \sup_{g:\|g\|_{L^2}=1} \iint_{\mathcal{Z}} g(z)k(z,z')g(z')\mathrm{d}z\mathrm{d}z', \quad \text{and} \quad \Lambda_{\min} := \inf_{g:\|g\|_{L^2}=1} \iint_{\mathcal{Z}} g(z)k(z,z')g(z')\mathrm{d}z\mathrm{d}z',$$

where $\|g\|_{L^2} = \int_{\mathcal{Z}} g(z)^2 \mathrm{d}z$ is the L2 norm. The value $\Lambda_{\max}$ and $\Lambda_{\min}$ is the largest/smallest eigenvalue of an embedding operator to the RKHS. Here, we define $\delta_\Lambda := \max\{|\Lambda_{\max} - 1|, |1 - \Lambda_{\min}|\}$. Then, we show that an upper bound of the asymptotic variance with the proposed estimator is increased by

$$O(\Lambda_{\max}\delta_\Lambda + \{\Lambda_{\max}\delta_\Lambda\}^2),$$

compared to the case without kernel functions. The case with $\delta_\Lambda = 1$ corresponds to the unconditional moment restriction with the L2 norm, as displayed in (6). This indicates that kernels that are concentrated in the direction of a particular eigenfunction may have worse efficiency. More details will be provided in Section 5.2.1.

3.2 Practical MMR-IV algorithms

A workflow of our algorithm based on $\widehat{R}_V$ is summarized in Algorithm 1; we leave the $\widehat{R}_U$, based method to future work to solve the inference issues caused by indefinite W_U. We provide examples of the class $\mathcal{F}$ in both parametric and nonparametric settings.

Algorithm 1 MMR-IV

Input: Dataset $D = \{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}\}$, kernel k with parameters θ_k, a function class $\mathcal{F}$, regularizer $\Omega(\cdot)$ and λ
Output: The estimate of f^* in $\mathcal{F}$.
1: Compute the kernel matrix $K = k(\boldsymbol{z}, \boldsymbol{z}; \theta_k)$.
2: Define the residual $\boldsymbol{\varepsilon}(f) = \boldsymbol{y} - f(\boldsymbol{x})$.
3: $\hat{f}_\lambda \leftarrow \mathrm{argmin}_{f \in \mathcal{F}} \; n^{-2}\boldsymbol{\varepsilon}(f)^\top K\boldsymbol{\varepsilon}(f) + \lambda\Omega(\|f\|_{\mathcal{F}})$.
4: **return** $\hat{f}_\lambda$.

Deep NNs. In the parametric setting, the function class $\mathcal{F}$ can often be expressed as $\mathcal{F}_\Theta = \{f_\theta : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^m$ denotes a parameter space. We consider a very common nonlinear model in machine learning $f(x) = W_0\Phi(x) + b_0$ where, $\Phi : x \longmapsto \sigma_h(W_h\sigma_{h-1}(\cdots \sigma_1(W_1 x)))$ denotes a nonlinear feature map of a depth-h NN. Here, W_i for $i = 1, \ldots, h$ are parameter matrices, and each σ_i denotes the entry-wise activation function of the ith layer. In this case, $\theta = (b_0, W_0, W_1, \ldots, W_h)$. As a result, we can rewrite $\hat{f}_V$ in terms of their parameters as $\hat{\theta}_V \in \arg\min_{\theta \in \Theta} \widehat{R}_V(f_\theta) + \lambda\|\theta\|_2^2$, where $f_\theta \in \mathcal{F}_\Theta$. We denote $\theta^* \in \arg\min_{\theta \in \Theta} R_k(f_\theta)$. In what follows, we refer to this algorithm as **MMR-IV (NN)**; see Algorithm 2.

Algorithm 2 MMR-IV (NN)

Input: Dataset $D = \{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}\}$, kernel function k with parameters θ_k, NN f_{NN} with parameters θ_{NN}, regularization parameter λ.
Output: predictive value at $\boldsymbol{x}_*$.
1: Compute $K = k(\boldsymbol{z}, \boldsymbol{z}; \theta_k)$.
2: $\hat{f}_{NN} = \mathrm{argmin}_{f_{NN}}(\boldsymbol{y} - f_{NN}(\boldsymbol{x}))^\top K(\boldsymbol{y} - f_{NN}(\boldsymbol{x}))/n^2 + \lambda\|\theta_{NN}\|_2^2$.
3: **return** $\hat{f}_{NN}(\boldsymbol{x}_*)$

Kernel machines. In a nonparametric setting, the function class $\mathcal{F}$ becomes an infinite dimensional space. In this work, we consider $\mathcal{F}$ to be an RKHS $\mathcal{H}_l$ of real-valued functions on $\mathcal{X}$ with a reproducing kernel $l : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then, the regularized solution can be obtained by $\arg\min_{f \in \mathcal{H}_l} \hat{R}_V(f) + \lambda \|f\|_{\mathcal{H}_l}^2$. As per the representer theorem, every optimal $\hat{f}$ admits a form $\hat{f}(x) = \sum_{i=1}^n \alpha_i l(x, x_i)$ for some $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ [47], and on the basis of this representation, we rewrite the objective as follows:

$$\hat{f}_V = \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} (\boldsymbol{y} - L\boldsymbol{\alpha})^\top W_V(\boldsymbol{y} - L\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^\top L\boldsymbol{\alpha}, \tag{7}$$

where $L = [l(x_i, x_j)]_{ij}$ is the kernel matrix on $\boldsymbol{x}$. For the U-statistic version, the quadratic program (7) substitutes indefinite W_U for W_V, so it may not be positive definite. The value of λ needs to be sufficiently large to ensure that (7) is definite. On the other hand, the V-statistic-based estimate (7) is definite for all nonzero λ since W_V is positive semidefinite. Thus, the optimal $\hat{\boldsymbol{\alpha}}$ can be obtained by solving the first-order stationary condition, and if L is positive definite, the solution has a closed-form expression, $\hat{\boldsymbol{\alpha}} = (LW_V L + \lambda L)^{-1} LW_V \boldsymbol{y}$. Thus, we will focus on the V-statistic version in our experiments. In the following, we refer to this algorithm as **MMR-IV (RKHS)**.

Nyström approximation. The MMR-IV (RKHS) algorithm is computationally costly for large datasets as it requires a matrix inversion. To improve the scalability, we resort to Nyström approximation [48] to accelerate the matrix inversion in

$$\hat{\boldsymbol{\alpha}} = (LW_V L + \lambda L)^{-1} LW_V \boldsymbol{y}.$$

First, we randomly select a subset of $m(\ll n)$ samples from the original dataset and construct the corresponding submatrices of W_V, namely, W_{mm} and W_{nm} based on this subset. Second, let V and U be the eigenvalue vector and the eigenvector matrix of W_{mm}, respectively. Then, the Nyström approximation is obtained as $W_V \approx \widetilde{U}\widetilde{V}\widetilde{U}^\top$, where $\widetilde{U} := \sqrt{\frac{m}{n}} W_{nm} U V^{-1}$ and $\widetilde{V} := \frac{n}{m} V$. We finally apply the Woodbury formula [49, p. 75] to obtain

$$(LW_V L + \lambda L)^{-1} LW_V = L^{-1}(W_V + \lambda L^{-1})^{-1} W_V \approx \lambda^{-1}[I - \widetilde{U}(\lambda^{-1}\widetilde{U}^\top L\widetilde{U} + \widetilde{V}^{-1})^{-1}\widetilde{U}^\top \lambda^{-1} L]\widetilde{U}\widetilde{V}\widetilde{U}^\top.$$

We will refer to this algorithm as **MMR-IV (Nyström)**; Algorithm 3. The runtime complexity of this algorithm is $O(nm^2 + n^2)$.

Algorithm 3 MMR-IV (Nyström)

Input: Dataset $D = \{(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})\}_{i=1}^n$, kernel functions k and l with parameters θ_k and θ_l, regularization parameter λ, leave out data size M, Nyström approximation sample size M', leave-out-cross-validation (LMOCV) times m
Output: predictive value at $\boldsymbol{x}_*$
1: $K = k(\boldsymbol{z}, \boldsymbol{z}; \theta_k)$
2: $\hat{\delta}, \hat{\theta}_l = \text{argmin}_{\delta, \theta_l} \sum_{i=1}^m (\boldsymbol{b}^{(i)} - \boldsymbol{y}_M^{(i)})^\top K_M^{(i)}(\boldsymbol{b}^{(i)} - \boldsymbol{y}_M^{(i)})$ Equation (18)
3: $K/n^2 \approx \widetilde{U}\widetilde{V}\widetilde{U}^\top$ Nyström Approx with M' samples
4: $\hat{\alpha} = \lambda^{-1}[I - \widetilde{U}(\lambda^{-1}\widetilde{U}^\top L\widetilde{U} + \widetilde{V}^{-1})^{-1}\widetilde{U}^\top \lambda^{-1} L]\widetilde{U}\widetilde{V}\widetilde{U}^\top \boldsymbol{y}$
5: **return** $\hat{f}(\boldsymbol{x}_*) = l(\boldsymbol{x}_*, \boldsymbol{x}; \hat{\theta}_l)\hat{\alpha}$

---

4 Hyperparameter selection

We discuss the selection of hyperparameters such as the regularization coefficient λ. The hyperparameters in the IV regression problem significantly affect the finite sample performance, but the selection process is nontrivial. We first give an overview of our cross-validation (CV) approach in Section 4.1, and then develop

some techniques that make the approach efficient in Sections 4.2 and 4.3. Figure 2 gives an overview of the contents of this section.

4.1 Leave-M-Out Cross Validation (LMOCV)

We utilize the LMOCV approach for the hyperparameter selection, which is described as follows. Let M be an integer such that $M \leq n$, and Λ be a set of hyperparameter candidates.

1:     $K \leftarrow n/M$
2:     Split the data into K parts: $[D_1, D_2, \ldots, D_K]$ where $|D_i| \approx M$
3:     **for** each $(\lambda, \mathcal{F}) \in \Lambda$ **do**
4:        **for** each $D_i \in [D_1, D_2, \ldots, D_K]$ **do**
5:          Solve $\hat{f}_\lambda \leftarrow \arg\min_{f \in \mathcal{F}} \widehat{R}_V(f) + \lambda \Omega(f)$ using $D_{-i} = [D_1, \ldots, D_{i-1}, D_{i+1}, \ldots, D_K]$
6:          Evaluate $S(\lambda, i) \leftarrow \widehat{R}_V(\hat{f}_\lambda)$ using D_i
7:        **end for**
8:       $S(\lambda) \leftarrow (1/K)\sum_{k=1}^{K} S(\lambda, k)$
9:     **end for**
10:    Output $\lambda^*, \mathcal{F}^* \leftarrow \arg\min_{\lambda, \mathcal{F}} S(\lambda)$

One disadvantage of this approach is that it is computationally expensive because the empirical risks are optimized many times. To alleviate this problem, we study an analytical form of the score function $S(\lambda)$ for MMR-IV (RKHS) and MMR-IV (Nyström). The analytical error enables any fold of CV and in comparison, related solutions rely on a few fold of CV [15,50] or median heuristic [16] for this task, which, however, have rather limited selection capability. We propose this method in the next subsections.

4.2 GP interpretation

Our interest is to obtain an analytical empirical risk for MMR-IV (Nyström) with the V-statistic objective (7). Toward the goal, we relate our ERM form to a stochastic model with a GP, inspired by ref. [51]. Different from the counterpart in the ordinary kernel regression, we need to apply additional analysis to deal with the additional weight matrix W_V in the V-statistic risk (7).

We build the connection between the V-statistic risk and the posterior distribution of a GP model in this subsection, after which we will show that the MMR-IV (Nyström) estimator is equivalent to the Maximum-A-Posteriori estimator based on the GP. The relationship is inspired by the similarity of our objective function to that of GLS.
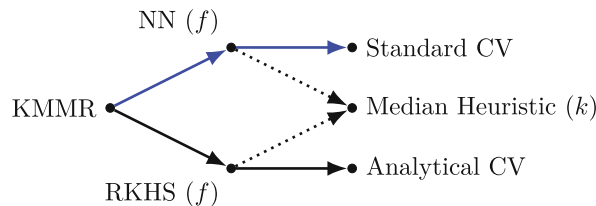

Figure 2: A summary of proposed algorithms. Following the KMMR framework, two methods – MMR-IV (NN) and MMR-IV (RKHS) – are proposed, where f is parameterized by a NN and an RKHS function, respectively. An analytical CV error is proposed for MMR-IV (RKHS), and both methods select the kernel k by the median heuristic.

Prior and likelihood. Let us consider a GP prior over a space of functions $f(x) \sim \text{GP}(\mathbf{0}, \delta l(x, x))$, where $\delta > 0$ is a real constant, and a set of i.i.d. data $D = \{(x_i, y_i, z_i)\}_{i=1}^{n}$. To define the likelihood $p(D \mid f)$, we recall from Lemma 1 that the risk $R_k(f)$ can be expressed in terms of two independent copies of random variables (X, Y, Z) and (X', Y', Z'). To this end, let $D' := \{(x_i', y_i', z_i')\}_{i=1}^{n}$ be an independent sample of size n with an identical distribution to D. Given a pair of samples (x, y, z) and (x', y', z') from D and D', respectively, we then define the likelihood as follows:

$$p(\{(x, y, z), (x', y', z')\} \mid f) = \frac{\exp\left[-\frac{1}{2}(y - f(x))k(z, z')(y' - f(x'))\right]}{\int \exp\left[-\frac{1}{2}(\bar{y} - f(\bar{x}))k(\bar{z}, \bar{z}')(\bar{y}' - f(\bar{x}'))\right]\mathrm{d}P \otimes P((\bar{x}, \bar{y}, \bar{z}), (\bar{x}', \bar{y}', \bar{z}'))}.$$

Here, P is a joint distribution of (X, Y, Z), and $P \otimes P$ is a product of the distribution. Let $t = \bar{y} - f(\bar{x})$ and $t' = \bar{y}' - f(\bar{x}')$ and assume that the kernel k is bounded. Since $\int \exp(-\frac{1}{2}tk(\bar{z}, \bar{z}')t')\mathrm{d}P \otimes P \leq \max\{\int \exp(-\frac{1}{2}t^2 C)\mathrm{d}P, \exp(-\frac{1}{2}(t')^2 C)\mathrm{d}P\}$ with some $C > 0$, the likelihood function is integrable, and hence well defined. Since the integration in the denominator does not depend on (x, y, z), (x', y', z'), we regard the denominator as a normalizing constant and simply denote

$$p(\{(x, y, z), (x', y', z')\} \mid f) \propto \exp\left[-\frac{1}{2}(y - f(x))k(z, z')(y' - f(x'))\right].$$

Hence, the likelihood on both D and D' can be expressed as follows:

$$p(\{D, D'\} \mid f) = \prod_{i=1}^{n}\prod_{j=1}^{n} p(\{(x_i, y_i, z_i), (x_j', y_j', z_j')\} \mid f)$$

$$\propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - f(x_i))k(z_i, z_j')(y_j' - f(x_j'))\right]. \tag{8}$$

This likelihood function is regarded as a generalized version of a loss function of generalized least squares with heteroskedastic noise [52]. That is, if we set $k(z_i, z_j) = \sigma_i^2$ if $i = j$ with $\sigma_i^2 > 0$, and $k(z_i, z_j) = 0$ otherwise, the likelihood corresponds to that of a linear regression model with Gaussian noise having heteroskedastic variance. In practice, however, we only have access to a single copy of the sample, i.e., D, but not D'. One way of constructing D' is through data splitting: the original dataset is split into two halves of equal size, where the former is used to construct D and the latter is for D'. Unfortunately, this approach reduces the effective sample size that can be used for learning. Alternatively, we propose to estimate the original likelihood (8) by using M-estimators: given the full dataset $D = \{(x_i, y_i, z_i)\}_{i=1}^{n}$, our approximated likelihood is defined as follows:

$$p(D \mid f) = \prod_{i=1}^{n}\prod_{j=1}^{n} p(\{(x_i, y_i, z_i), (x_j, y_j, z_j)\} \mid f(x_i), f(x_j))$$

$$\propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - f(x_i))k(z_i, z_j)(y_j - f(x_j))\right] \tag{9}$$

$$= \exp\left[-\frac{1}{2}(\mathbf{y} - f(\mathbf{x}))^{\top}K_z(\mathbf{y} - f(\mathbf{x}))\right].$$

This likelihood assumes correlated errors with the normal distribution. It is closely related to the standard GP regression that has a similar Gaussian likelihood, but the covariance matrix is computed on $\mathbf{x}$ rather than $\mathbf{z}$, and the generalized ridge regression as well as the weighted least regression. The matrix K_z, as a kernel matrix defined earlier, plays a role in correlating the residuals $y_i - f(x_i)$. Based on the likelihood $p(D \mid f)$, the maximum likelihood estimator hence coincides with the minimizer of the unregularized version of our objective (5).

Maximum *a posteriori* (MAP). Combining the aforementioned likelihood function with the prior on $f(\boldsymbol{x})$, the posterior probability of $f(\boldsymbol{x})$ is expressed as follows:

$$p(f(\boldsymbol{x})|D) = N(f(\boldsymbol{x})|\boldsymbol{c}, C) \propto p(f(\boldsymbol{x}))p(D|f),$$
$$C = (K_{\boldsymbol{z}} + (\delta L)^{-1})^{-1} = L(LW_V L + (\delta n^2)^{-1}L)^{-1}Ln^{-2},$$
$$\boldsymbol{c} = CK_{\boldsymbol{z}}\boldsymbol{y} = L(LW_V L + (\delta n^2)^{-1}L)^{-1}LW_V\boldsymbol{y}.$$

The MAP estimate of $f(\boldsymbol{x})$ is simply $\boldsymbol{c}$,

$$\underset{f(\boldsymbol{x})}{\mathrm{argmax}}\log p(f(\boldsymbol{x}) \mid D) = L\underbrace{(LW_V L + (\delta n^2)^{-1}L)^{-1}LW_V\boldsymbol{y}}_{=\hat{\alpha}}.$$

We can see that the MAP estimator returns the same result as that of the MMR-IV (RKHS) estimator given that $(n^2\delta)^{-1} = \lambda$ (see the context around (7)), and δ plays the role of the regularization parameter. For MMR-IV (Nyström), we consider the GP model with W_V approximated by $\widehat{U}\widehat{V}\widehat{U}^{\top}$ in C and $\boldsymbol{c}$, and the conclusion still holds,

$$p(f(\boldsymbol{x})|D) = N(f(\boldsymbol{x})|\boldsymbol{c}', C')$$
$$C' = \delta L[I - \widetilde{U}(\lambda^{-1}\widetilde{U}^{\top}L\widetilde{U} + \widetilde{V}^{-1})^{-1}\widetilde{U}^{\top}\lambda^{-1}L],$$
$$\boldsymbol{c}' = \delta n^2 L[I - \widetilde{U}(\lambda^{-1}\widetilde{U}^{\top}L\widetilde{U} + \widetilde{V}^{-1})^{-1}\widetilde{U}^{\top}\lambda^{-1}L]\widehat{U}\widehat{V}\widehat{U}^{\top}\boldsymbol{y}.$$

We summarize the result in Proposition 3. Such a GP interpretation is used for an elegant derivation of the analytical CV error in the following subsection.

Proposition 3. *Given* $\delta = (\lambda n^2)^{-1}$ *and* $\hat{f} = \mathrm{argmin}_{f \in \mathcal{H}_l}$ *(equation (7), including Nyström approximated),*

$$\underset{f(\boldsymbol{x}_*)}{\mathrm{argmax}}\, p(f(\boldsymbol{x}_*)|D) = \hat{f}(\boldsymbol{x}_*).$$

4.3 Analytical form of error

Now, we derive the analytical error of LMOCV via Bayes' rules based on the posterior GP. We split the whole dataset D into training and development datasets, denoted as D_{tr} and $D_{de} \coloneqq \{\boldsymbol{x}_{de}, \boldsymbol{y}_{de}, \boldsymbol{z}_{de}\}$, respectively, where D_{de} has M triplets of data points. Given D_{tr}, the predictive probability of $f(D_{de})$ can be obtained by Bayes' rules (see also equation (22) in ref. [51]):

$$p(f(\mathbf{x}_{de})|D) = p(f(\mathbf{x}_{de})|D_{tr}, D_{de})$$
$$= p(D_{de}|f(\boldsymbol{x}_{de}), D_{tr})p(f(\boldsymbol{x}_{de})|D_{tr})/p(D_{de}|D_{tr})$$
$$= p(D_{de}|f(\boldsymbol{x}_{de}))p(f(\boldsymbol{x}_{de})|D_{tr})/p(D_{de}|D_{tr}),$$

which implies that

$$p(f(\boldsymbol{x}_{de})|D_{tr}) = p(f(\boldsymbol{x}_{de})|D)p(D_{de}|D_{tr})/p(D_{de}|f(\boldsymbol{x}_{de})), \qquad (10)$$

where $p(f(\boldsymbol{x}_{de})|D) = N(\boldsymbol{c}_{de}, C_{de})$ with $\boldsymbol{c}_{de}$, C_{de} the mean and covariance of $f(\boldsymbol{x}_{de})$ in $p(f(\boldsymbol{x})|D)$ and $p(D_{de}|f(\boldsymbol{x}_{de})) \propto N(\boldsymbol{y}_{de}, K_{de}^{-1})$, and $K_{de} = k(\boldsymbol{z}_{de}, \boldsymbol{z}_{de})$. From this result, we can see that $p(D_{de}|f(\boldsymbol{x}_{de}))$ and $p(f(\boldsymbol{x}_{de})|D_{tr})$ are in Gaussian forms and $p(D_{de}|D_{tr})$ is a constant term, so $p(f(\boldsymbol{x}_{de})|D_{tr})$ must be Gaussian:

$$p(f(\boldsymbol{x}_{de})|D_{tr}) = N(\boldsymbol{b}, B),$$

where $B^{-1} = C_{de}^{-1} - K_{de}$ and $\boldsymbol{b} = B(C_{de}^{-1}\boldsymbol{c}_{de} - K_{de}\boldsymbol{y}_{de})$ can be obtained by solving equation (10). By Proposition 3, we know that the prediction on the validation data $f(x_{de})$ given the training data D_{tr} is $\boldsymbol{b}$. Therefore, the error of m repeated LMOCV is computed as shown in (11), where we use (i) to denote the ith split of the whole dataset into D_{de} and D_{tr} and $C_{de}^{(i)}$ and as mentioned earlier, $\boldsymbol{c}_{de}^{(i)}$ are straightforwardly obtained from $p(f(\boldsymbol{x})|D)$. As a result, this enables an efficient parameter selection with the CV principle: the standard LMOCV procedure requires multiple experiments (e.g., m here) on different D_{tr} and D_{de}, while our result

needs only a single training on D and the validation error of m experiments is simply gained by plugging the training result into the analytical error (11):

$$\text{LMOCV Error} \coloneqq \sum_{i=1}^{m} (r^{(i)})^{\top} K_{de}^{(i)} r^{(i)}, \quad r^{(i)} \coloneqq \boldsymbol{b}^{(i)} - \boldsymbol{y}_{M}^{(i)} = (I - C_{de}^{(i)} K_{de}^{(i)})^{-1} (\boldsymbol{c}_{de}^{(i)} - \boldsymbol{y}_{de}^{(i)}). \quad (11)$$

5 Theoretical result: Consistency and asymptotic normality

We provide the consistency and asymptotic normality of $\hat{f}_V$. Similar results for $\hat{f}_U$, together with all the proofs, are provided in the appendix. Readers interested in the finite sample rate should consult [50, E.4] for further details.

5.1 Consistency

We first show the consistency of $\hat{f}_V$, which depends on the uniform convergence of the risk functions. The result holds for both parametric and nonparametric cases regardless of the shape of $\Omega(f)$, so we can utilize the regularization $\|\theta\|_2^2$, which is common for NN but nonconvex in terms of f. In the following result on the consistency, we consider that $\mathcal{F}$ has topology induced by distance for parameters $\theta \in \Theta$ that characterize functions, that is, for functions $f_\theta, f_{\theta'} \in \mathcal{F}$, their distance is measured by that of their parameters θ and θ': $d(f_\theta, f_{\theta'}) = \|\theta - \theta'\|_2$.

Proposition 4. (Consistency of $\hat{f}_V$) *Assume that* $\mathbb{E}[|Y|^2] < \infty$, $\mathbb{E}[\sup_{f \in \mathcal{F}} |f(X)|^2] < \infty$, $\mathcal{F}$ *is compact, Assumptions* 1 *and* 2 *hold,* $\Omega(f)$ *is a bounded function, and* $\lambda \xrightarrow{\text{p}} 0$. *Then* $\hat{f}_V \xrightarrow{\text{p}} f^*$.

If $\Omega(f)$ is convex in f, the consistency can be obtained more easily by [53, Thm. 2.7], which relies on the convexity of the risk functions. In this case, we can avoid several conditions. We provide an additional result with this setting in Appendix. A.6.

5.2 Asymptotic normality with finite-dimension case

We analyze the asymptotic normality of the estimator $\hat{f}_V$, which is important to advance statistical analysis such as tests. The results are obtained by applying the delta method with finite parameters and functionals, respectively.

We first consider the $\hat{f}_V$ that is characterized by a finite-dimensional parameter from a parameter space Θ. We rewrite the regularized V-statistic risk as a compact form $\widehat{R}_{V,\lambda}(f_\theta) \coloneqq \frac{1}{n^2} \sum_{i,j} h_\theta(u_i, u_j) + \lambda \Omega(\theta)$, $h_\theta(u_i, u_j) \coloneqq (y_i - f_\theta(x_i))(y_j - f_\theta(x_j)) k(z_i, z_j)$, and consider $R_k(f_\theta)$ is uniquely minimized at $\theta^* \in \Theta$ and $\rightsquigarrow$ denotes a convergence in law.

Theorem 1. (Asymptotic normality of $\hat{\theta}_V$) *Suppose that* f_θ *and* $\Omega(\theta)$ *are twice continuously differentiable about* θ, Θ *is compact,* $H = \mathbb{E}[\nabla_\theta^2 h_{\theta^*}(U, U')]$ *is nonsingular,* $\mathbb{E}[|Y|^2] < \infty$, $\mathbb{E}[\sup_{\theta \in \Theta} |f_\theta(X)|^2] < \infty$, $\mathbb{E}[\sup_{\theta \in \Theta} \|\nabla_\theta f_\theta(X)\|_2^2] < \infty$, $\mathbb{E}[\sup_{\theta \in \Theta} \|\nabla_\theta^2 f_\theta(X)\|_F^2] < \infty$, $\sqrt{n}\lambda \xrightarrow{\text{p}} 0$, $R_k(f_\theta)$ *is uniquely minimized at* θ^* *which is an interior point of* Θ, *and Assumption* 1 *holds. Then,*

$$\sqrt{n}(\hat{\theta}_V - \theta^*) \rightsquigarrow N(\boldsymbol{0}, \Sigma_V)$$

holds, where $\Sigma_V = 4H^{-1} \text{diag}(\mathbb{E}_U[\mathbb{E}_{U'}^2[h_{\theta^*}(U, U')]]) H^{-1}$.

5.2.1 Connection to efficiency

The variance of the asymptotic distribution varies with the choice of kernel. To see this, we consider the constants $\Lambda_{\min} \le \Lambda_{\max}$ such that

$$\Lambda_{\min} = \inf_{g:\|g\|_{L^2}=1} \iint_{\mathcal{Z}} g(z)k(z, z')g(z')\mathrm{d}z\mathrm{d}z' \le \sup_{g:\|g\|_{L^2}=1} \iint_{\mathcal{Z}} g(z)k(z, z')g(z')\mathrm{d}z\mathrm{d}z' = \Lambda_{\max}.$$

Their positiveness and existence are guaranteed by Assumption 1. By Theorem 4.51 in ref. [37], $\Lambda_{\min}$ and $\Lambda_{\max}$ correspond to the smallest and largest eigenvalue of the embedding operator $T : L^2 \to \mathcal{H}_k$, where $\mathcal{H}_k$ is the RKHS whose reproducing kernel is k. Equivalently, by the Mercer's theorem (Theorem 4.49 in ref. [37]), we have the form of k as follows:

$$k(z, z') = \sum_j \lambda_j \phi_j(x)\phi_j(z'),$$

where $\{\lambda_j\}_j$ are coefficients and $\phi_j(\cdot)$ is an orthonormal basis function, and $\Lambda_{\max} = \max_j \lambda_j$ and $\Lambda_{\min} = \min_j \lambda_j$. We also define an asymptotic variance *without* the kernel k. That is, we define a loss without the kernel as $\check{h}_\theta(u_i, u_j) := (y_i - f_\theta(x_i))(y_j - f_\theta(x_j))$, and its associated matrices $\check{H} = \mathbb{E}[\nabla_\theta^2 \check{h}_{\theta^*}(U, U')]$ and $\check{\Sigma}_V = 4\check{H}^{-1}\mathrm{diag}(\mathbb{E}_U[\mathbb{E}_{U'}^2[\check{h}_{\theta^*}(U, U')]])\check{H}^{-1}$. Let $\|\Sigma_V\|_F = \sqrt{\mathrm{tr}(\Sigma_V^2)}$ be the Frobenius norm for square matrices.

Proposition 5. *Suppose that Assumption* 1 *holds. Define* $\delta_\Lambda := \max\{|\Lambda_{\max} - 1|, |1 - \Lambda_{\min}|\}$. *Then, we have*

$$\|\Sigma_V\|_F \le \|\check{\Sigma}_V\|_F + O(\Lambda_{\max}\delta_\Lambda + (\Lambda_{\max}\delta_\Lambda)^2).$$

This result evaluates the increase in the asymptotic variance due to the choice of the kernel by the upper bound. Strictly speaking, as the eigenvalues $\Lambda_{\min}$ and $\Lambda_{\max}$ of the kernel k approach 1, i.e., as the kernel function approaches a constant function, the asymptotic variance decreases at the speed of the first and second power of δ_Λ. Therefore, this result demonstrates that utilizing the simplest possible kernel, one that is not focused on a specific eigenspace, is the method for preventing an increase in the asymptotic variance.

Intuitively, this result represents the difficulty of choosing the kernel function as the optimal IV in nonlinear IV regression. Unlike the case of linear regression [30], the nonlinear IV regression does not provide information such as which eigenspaces of data have a large variance. Thus, the anisotropy of the eigenvalues of kernels may worsen the efficiency in the worst case. However, efficiency can be improved if the anisotropy of the eigenvalues fits the data. If such properties could be known, one can develop an optimal kernel selection procedure based on them.

5.3 Asymptotic normality of infinite-dimension case

We show the asymptotic normality of an infinite-dimensional estimator $\hat{f}_V$. That is, we show that an error of $\hat{f}_V$ weakly converges to a GP that takes values in a function space $\mathcal{H}_l$. We set $\Omega(f) = \|f\|_{\mathcal{H}_l}^2$ and consider a minimizer: $f_{\lambda_0}^* \in \mathrm{argmin}_{f \in \mathcal{F}}\ R_k(f) + \lambda_0\|f\|_{\mathcal{H}_l}^2$ with arbitrary $\lambda_0 > 0$. We also define $\mathcal{N}(\varepsilon, \mathcal{H}, \|\cdot\|)$ as an ε-covering number of $\mathcal{H}$ in terms of $\|\cdot\|$. Then, we obtain the following result, which allows statistical inference on functional estimators such as kernel machines.

Assumption 3. (Low-entropy condition) There exists $s \in (0, 2)$ and a constant $C_H > 0$ such that

$$\log\mathcal{N}\left(\varepsilon, \mathcal{H}_l, \|\cdot\|_{L^\infty}\right) \le C_H\varepsilon^{-s},$$

for every $\varepsilon \in (0, 1)$.

Although the covering number condition $\log \mathcal{N}(\varepsilon, \mathcal{H}_l, \|\cdot\|_{L^\infty}) \le C_H \varepsilon^{-s}$ in Theorem 2 is strong, we can present some examples that satisfy the condition. A representative example is the set of Lipschitz functions on a bounded interval. The condition is then satisfied under $s = 1$ (Example 5.10 in ref. [54]). Although the support is one dimensional, it can cover nondifferentiable functions. Another example is the set of b-times differentiable functions on a compact set in $\mathbb{R}^d$ with $b > 2d$ (Theorem 2.7.1 in ref. [55]). Although requires higher-order smoothness, it can handle functions with multidimensional inputs. The last example is the RKHS with kernel functions characterized by the spectrum: the condition is satisfied when s is a decay rate of the eigenvalues of the embedding operator into the RKHS. Several common kernels, such as the Gaussian kernel (see Section 4 in ref. [37]) and the Mercer kernel (explained in ref. [56]), satisfy it with a certain parameter configuration. This characterization is used in several machine learning studies, e.g., [57,58].

Theorem 2. (Asymptotic normality of $\hat{f}_V$) *Suppose Assumption* 1 *holds, l is a bounded kernel, k is a uniformly bounded function, and $\lambda - \lambda_0 = o(n^{-1/2})$ holds. Also, suppose that $\mathcal{X}$, $\mathcal{Z}$, and $\mathcal{Y}$ are compact spaces, and Assumption* 3 *holds. We define a Gaussian process $\mathbb{G}_P^*$ on $\mathcal{H}_l$ with zero mean and its covariance*

$$\text{Cov}(G_P^*(f), G_P^*(f')) = \mathbb{E}_U[P^1 h_f(U) P^1 h_{f'}(U)] - \mathbb{E}_U[P^1 h_f(U)] \mathbb{E}_U[P^1 h_{f'}(U)]$$

with $h_f(u, u') = (y - f(x))(y' - f(x'))k(z, z')$ and $P^1 h_f(\cdot) = (\int h_f(u, \cdot) + h_f(\cdot, u) \, dP(u))/2$. Then, there exists a linear operator $\nabla S_{P^2, \lambda_0} : \mathcal{H}_l \to \mathcal{H}_l$ such that

$$\sqrt{n}(\hat{f}_V - f_{\lambda_0}^*) \rightsquigarrow \nabla S_{P^2, \lambda_0}(\mathbb{G}_P^*) \text{ in } \mathcal{H}_l.$$

Note that $\nabla S_{P^2, \lambda_0}(\mathbb{G}_P^*)$ is also a GP on $\mathcal{H}_l$ owing to the linearly of $\nabla S_{P^2, \lambda_0}$ (Section 3.9.2 in ref. [55]).

We provide the detailed derivation in the appendix. A specific form of the operator $\nabla S_{P^2, \lambda_0}$ is also given in the appendix. All conditions for the theorem are valid for many well-known kernels. For the boundedness assumption, many common kernels, such as the Gaussian RBF kernel, the Laplacian kernel, and the Mercer kernel, satisfy it.

This nonparametric asymptotic normality is a generic result that leads to finite-dimensional asymptotic normality. That is, for every finite set $\{x_j\}_{j=1}^N \subset \mathcal{X}$, Theorem 2 immediately implies the following convergence:

$$\sqrt{n} \begin{pmatrix} \hat{f}_V(x_1) - f_{\lambda_0}^*(x_1) \\ \vdots \\ \hat{f}_V(x_N) - f_{\lambda_0}^*(x_N) \end{pmatrix} \rightsquigarrow \mathcal{N}_N(0, \Sigma),$$

where $\mathcal{N}_N(0, \Sigma)$ is the N-variate normal distribution with zero mean and a corresponding covariance matrix Σ. This result is a convenient generalization of asymptotic normality in a parametric setting.

Remark 1. (On asymptotic normality) Theorem 2 and $\lambda_0 > 0$ results in a biased center $f_{\lambda_0}^*$ while it is essential to obtain the limit distribution in infinite dimensional space. This is one of the reasons why $\sqrt{n}$-convergence can be achieved in Theorem 2. Although it is not very satisfactory, we note that deriving the limit distribution has a different nature than obtaining error bounds, and it is more restrictive than developing a learner with bounded generalization errors. One can find that a similar difficulty appears in deriving a limit with the non-Donsker class [55]. To remove this constraint, we need to introduce some strong assumptions, such as the low noise condition in the classification problem, as discussed in ref. [29]. However, we do not solve this problem in this study because such a response would go far beyond the scope of this work.

6 Related work

Several extensions of 2SLS and GMM exist for the nonlinear IV problem. In the two-stage approach, the function $f(x)$ has often been obtained by solving a Fredholm integral equation of the first kind

$\mathbb{E}[Y|Z] = \int f(x)\mathrm{d}P(x|Z)$. In refs [8,10,11,59], linear regression is replaced by a linear projection onto a set of known basis functions. Like the kernel choice problem in our work, this approach requires choosing the appropriate set of basis functions. A uniform convergence rate of this approach is provided in ref. [60]. In refs [9,61], the first-stage regression is replaced by a conditional density estimation of $\mathbb{P}(X|Z)$ using a kernel density estimator. On the contrary, our approach does not rely on the conditional density estimation problem.

The IV regression has also recently received attention in the machine learning community. Hartford et al. [12] proposed to solve the integral equation by first estimating $P(X|Z)$ with a mixture of deep generative models on which the function $f(x)$ can be learned with other deep NNs. Instead of NNs, [16] proposed to model the first-stage regression using the conditional mean embedding of $P(X|Z)$ [36,62,63], which is then used in the second-step kernel ridge regression. In other words, the first-stage estimation in ref. [16] becomes a vector-valued regression problem. In an attempt to alleviate the two-stage-estimation spirit, [15] and [17] reformulate the two-stage procedure as a convex–concave saddle-point problem, which is in a minimax form, via studying the Fenchel and Lagrange dual reformulations of the population risk, respectively. By using RKHSes for the inner maximization, DualIV [15] obtains a quadratic objective function similar to ours, but its RKHS is applied over (Z, Y), which cannot be interpreted as a valid instrument. In contrast, the approach of [17, Appendix F] provides an exact dual reformulation of our method and obtains our objective if RKHSes are used for the inner maximization. Besides, the starting objectives of the aforementioned works differ from ours: in refs [12,15,16] and [17, Appendix F], they started from minimizing the L_2 norm of the CMR, whereas we start from minimizing the squared UMR. The fact that they arrive at the same objective suggests a deeper connection that requires further investigation.

Our work follows in spirit many GMM-based approaches for IV regression, namely, [13,14,27]. We adopt the MMR framework of ref. [27], which only considers a conditional moment testing problem, to the parameter estimation problem of IV regression, and hence, our approaches and both theoretical and experimental analyses are different from theirs. In fact, this framework was initially inspired by refs [13,14], which instead parametrize the instruments by deep NNs. By combining the GMM framework with RKHS functions, the objective function can be evaluated in a closed form as shown in our work. As a result, our IV estimate can be obtained by minimizing the empirical risk, as opposed to an adversarial optimization [13,14]. It is important to note that recently these adversarial approaches have been improved similarly with RKHSs by refs [18,50], while there is a major difference. First, ref. [50] extends the work of ref. [13] to a single-stage algorithm similar to our MMR-IV (RKHS) [50, Section 4]. Although both works employ RKHSs in the minimax frameworks, ref. [50] incorporate a Tikhonov regularization on h in (3) and resort to the representer theorem [64] to develop the analytical objective function, whereas we impose a unit-ball constraint which is a form of Ivanov regularization [65]. The basis of our objective is the adversarial GMM objective of ref. [13]. Furthermore, prior works do not study the analytical CV error.

We also discuss connections to existing works in the areas outside of the IV setting in Section B in Appendix.

7 Experimental results

We present the experimental results in a wide range of settings for IV estimation. Following refs [13,14], we consider both low- and high-dimensional scenarios. In the experiments, we compare our algorithms to the following baseline algorithms:

- `DirectNN`: A standard least square regression on X and Y using a NN.
- `2SLS`: A vanilla 2SLS on raw X and Z.
- `Poly2SLS`: The 2SLS that is performed on polynomial features of X and Z via ridge regressions.
- `SieveIV` [26]: This algorithm solves the CMR via the sieve method, which uses orthonormal basis functions. In the implementation, we utilize the trigonometric basis, and the number of basis functions is selected by Lepski's method. The details are found in ref. [26] and its related studies.

– `DeepIV` [12]: A nonlinear extension of 2SLS using deep NNs. We use the implementation available at https://github.com/microsoft/EconML.
– `KernelIV` [16]: A generalization of 2SLS by modeling relations among X, Y, and Z as nonlinear functions in RKHSs. We use the publicly available implementation at https://github.com/r4hu1-5in9h/KIV.
– `GMM+NN`: An optimally weighted GMM [7] is combined with a NN $f(X)$. The details of this algorithm can be found in [14, Section 5].
– `AGMM` [13]: This algorithm models $h(Z)$ by a deep NN and employs a minimax optimization to solve for $f(X)$. The implementation we use is available at https://github.com/vsyrgkanis/adversarial_gmm.
– `DeepGMM` [14]: This algorithm is a variant of `AGMM` with optimal inverse-covariance weighting matrix. The publicly available implementation at https://github.com/CausalML/DeepGMM is used, and all of the aforementioned baselines are provided in the package except KernelIV.
– `AGMM-K` [50]: This algorithm extends AGMM by modeling $h(Z)$ and $f(X)$ as RKHSs. Nyström approximation is applied for fast computation. The publicly available implementation at https://github.com/microsoft/AdversarialGMM is used.
– `DualIV` [15]: This algorithm solves the CMR via a saddle-point reformulation and obtains a minimax problem. By using an RKHS for the inner maximization, an objective similar to ours is obtained, while the central kernel matrix is evaluated on Z and Y, which thus cannot be viewed as instruments and different from ours. The publicly available implementation at https://github.com/krikamol/DualIV-NeurIPS2020 is used.

7.1 Low-dimensional scenarios

Following [14], we employ the following data generation process:

$$Y = f^*(X) + e + \delta, \quad X = Z_1 + e + \gamma, \quad Z := (Z_1, Z_2),$$

where $e \sim \mathcal{N}(0, 1)$, $\gamma, \delta \sim \mathcal{N}(0, 0.1^2)$, and $Z \sim \text{Uniform}([-3, 3]^2)$ is a two-dimensional IV, but Z_1 has an effect on X. The variable e is the confounding variable that creates the correlation between X and the residual $Y - f^*(X)$. We vary the true function f^* between the following cases to enrich the datasets: (i) `sin`: $f^*(x) = \sin(x)$. (ii) `step`: $f^*(x) = \mathbb{1}_{\{x \geq 0\}}$. (iii) `abs`: $f^*(x) = |x|$. (iv) `linear`: $f^*(x) = x$. We consider both small-sample ($n = 200$) and large-sample ($n = 2,000$) regimes.

7.1.1 Experimental settings

For the experiments on low-dimensional data, we consider both small-sample ($n = 200$) and large-sample ($n = 2,000$) regimes, in which n points are sampled for training, validation, and test sets. In both regimes, we standardize the values of Y to have zero mean and unit variance for numerical stability.

Hyperparameters of NNs in Algorithm 2, including the learning rate and the regularization parameter, are chosen by a two-fold CV for fair comparisons with the baselines. As per the dimensions of X, we parametrize f either as a fully connected NN with leaky ReLU activations and 2 hidden layers, each of which has 100 cells, for nonimage data, or a deep convolutional neural network architecture for MNIST data. We denote the fully connected NN as FCNN(100,100) and refer readers to our code release for exact details on our CNN construction. Learning rates and regularization parameters are summarized in Table 1. Besides, we use the well-tuned hyperparameter selections of baselines provided in their packages without changes. We fix the random seed to 527 for all data generation and model initialization.

In contrast to our NN-based method, the RKHS-based method in Algorithm 3 has the analytic form of CV error. For the kernel function on instruments, we employ the sum of Gaussian kernels

$$k(z, z') = \frac{1}{3} \sum_{i=1}^{3} \exp\left(-\frac{\|z - z'\|_2^2}{2\sigma_{ki}^2}\right)$$

Table 1: Hyperparameters of NNs used in the experiments

Scenario	Model f	Learning rates	λ
Low dimensional	FCNN(100,100)	$(10^{-12}, 10^{-11}, 10^{-10}, 10^{-9}, 10^{-8}, 10^{-7}, 10^{-6})$	$(5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4})$
MNIST$_z$	FCNN(100,100)	$(10^{-12}, 10^{-11}, 10^{-10}, 10^{-9}, 10^{-8}, 10^{-7}, 10^{-6})$	$(5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4})$
MNIST$_x$	CNN	$(10^{-12}, 10^{-11}, 10^{-10}, 10^{-9}, 10^{-8}, 10^{-7}, 10^{-6})$	$(5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4})$
MNIST$_{xz}$	CNN	$(10^{-12}, 10^{-11}, 10^{-10}, 10^{-9}, 10^{-8}, 10^{-7}, 10^{-6})$	$(5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4})$
Mendelian	FCNN(100,100)	$(10^{-12}, 10^{-11}, 10^{-10}, 10^{-9}, 10^{-8}, 10^{-7}, 10^{-6})$	$(5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4})$

and the Gaussian kernel for $l(x, x') = \exp(-\|x - x'\|_2^2/(2\sigma_l^2))$, where σ_{k1} is chosen as the median interpoint distance of $\boldsymbol{z} = \{z_i\}_{i=1}^n$ and $\sigma_{k2} = 0.1\sigma_{k1}$, $\sigma_{k3} = 10\sigma_{k1}$. The motivation of such a kernel k is to optimize f on multiple kernels, and we leave parameter selection of k to the future work. We combine the training and validation sets to perform leave-2-out CV to select parameters of the kernel l and the regularization parameter λ. We choose the Gaussian kernel for l. For the Nyström approximation, we subsample 300 points from the combined set. As a small subset of the Gram matrix is used as Nyström samples, which misses much information of data, we avoid outliers in test results by averaging 10 test errors with different Nyström approximations in each experiment. All methods are repeated 10 times on each dataset with random initialization.

7.1.2 Experiment results

Tables 2 and 3 report the results for the large-sample and small-sample regimes respectively. First, under the influence of confounders, DirectNN performs worst as it does not use instruments. Second, MMR-IVs perform reasonably well in both small-sample and large-sample regimes. For the linear function, 2SLS and Poly2SLS tend to outperform other algorithms as the linearity assumption is satisfied in this case. For nonlinear functions, some NN-based methods show competitive performance in certain cases. Notably, GMM+NN has unstable performance as the function h in (2) is designed manually. KernelIV performs well because of simple simulation models. Its kernel parameters are selected by a median heuristic which is relatively naive to the CV, and in later more complicated experiments, the performance of KernelIV deteriorates. AGMM-K is similar in principle to our method while its errors are high. We suspect that this is due to

Table 2: The mean square error (MSE) $\pm$ one standard deviation in the large-sample regime ($n = 2{,}000$). Italics denote the second best

Algorithm	True function f^*			
	abs	**linear**	**sin**	**step**
DirectNN	0.116 ± 0.000	0.035 ± 0.000	0.189 ± 0.000	0.199 ± 0.000
2SLS	0.522 ± 0.000	$\mathbf{0.000 \pm 0.000}$	0.254 ± 0.000	0.050 ± 0.000
Poly2SLS	0.083 ± 0.000	$\mathbf{0.000 \pm 0.000}$	0.133 ± 0.000	0.039 ± 0.000
GMM+NN	0.318 ± 0.000	0.044 ± 0.000	0.694 ± 0.000	0.500 ± 0.000
AGMM	0.600 ± 0.001	0.025 ± 0.000	0.274 ± 0.000	0.047 ± 0.000
DeepIV	0.247 ± 0.004	0.056 ± 0.003	0.165 ± 0.003	0.038 ± 0.001
DeepGMM	0.027 ± 0.009	0.005 ± 0.001	0.160 ± 0.025	0.025 ± 0.006
KernelIV	0.019 ± 0.000	0.009 ± 0.000	0.046 ± 0.000	0.026 ± 0.000
AGMM-K	181 ± 0.000	2.34 ± 0.000	19.4 ± 0.000	4.13 ± 0.000
DualIV	0.344 ± 0.000	0.034 ± 0.000	0.379 ± 0.000	0.345 ± 0.000
SieveIV	0.170 ± 0.009	0.279 ± 0.001	0.021 ± 0.001	6.89 ± 0.044
MMR-IV$_{NN}$	$\mathbf{0.011 \pm 0.002}$	0.005 ± 0.000	0.153 ± 0.019	0.040 ± 0.004
MMR-IV$_{Nys}$	$\mathbf{0.011 \pm 0.001}$	0.001 ± 0.000	$\mathbf{0.006 \pm 0.002}$	$\mathbf{0.020 \pm 0.002}$

Table 3: The MSE ± one standard deviation in the small-sample regime ($n = 200$). Italics denote the second best

Algorithm	True function f^*			
	abs	**linear**	**sin**	**step**
DirectNN	0.143 ± 0.000	0.046 ± 0.000	0.404 ± 0.006	0.253 ± 0.000
2SLS	0.564 ± 0.000	**0.003** ± **0.000**	0.304 ± 0.000	0.076 ± 0.000
Poly2SLS	0.125 ± 0.000	**0.003** ± **0.000**	0.164 ± 0.000	0.077 ± 0.000
GMM+NN	0.792 ± 0.000	0.203 ± 0.000	1.56 ± 0.001	0.550 ± 0.000
AGMM	0.031 ± 0.000	0.011 ± 0.000	0.330 ± 0.000	0.080 ± 0.000
DeepIV	0.204 ± 0.008	0.047 ± 0.004	0.197 ± 0.004	**0.039** ± **0.001**
DeepGMM	*0.022* ± *0.003*	0.032 ± 0.016	0.143 ± 0.030	*0.039* ± *0.002*
KernelIV	0.063 ± 0.000	0.024 ± 0.000	*0.086* ± *0.000*	0.055 ± 0.000
AGMM-K	12.3 ± 0.000	1.32 ± 0.000	1.57 ± 0.000	1.71 ± 0.000
DualIV	0.202 ± 0.000	0.103 ± 0.000	0.251 ± 0.000	0.362 ± 0.000
SieveIV	0.466 ± 0.002	0.452 ± 0.001	0.363 ± 0.002	9.07 ± 0.038
MMR-IV (NN)	**0.019** ± **0.003**	*0.004* ± *0.001*	0.292 ± 0.024	0.075 ± 0.008
MMR-IV (RKHS)	0.030 ± 0.000	0.011 ± 0.000	**0.075** ± **0.000**	0.057 ± 0.000

its hyperparameter selection, and a few folds of CV over a small set of hyperparameters are not effective enough, because we observed that the AGMM-K's performance becomes better as the numbers of CV folds and hyperparameter candidates increase. A similar observation was also obtained on DualIV. Thus, it shows that it is desirable to have the analytical CV error, which is an advantage of our method against other RKHS baselines. Besides, the selection of the weight matrix remains an open question, and although DualIV, AGMM-K, and MMRIV (RKHS) rely on the median heuristic to select the bandwidth, we believe that the selection has different effects on different methods and it is difficult to obtain fair selection. We will leave this problem to the future work. Moreover, the performances of the complicated methods like DeepIV and DeepGMM deteriorate more in the large-sample regimes than the small-sample regimes. We suspect that this is because these methods rely on two NNs and thus require proper hyperparameters (e.g., model structures) on random samples. For SieveIV, it performs well when the true function is the sin function, because we utilize the trigonometric basis for the sieve method. It suggests that the choice of basis functions is important for the performance of this method. In contrast, MMR-IV (Nyström) has the advantage of adaptive hyperparameter selection.

We also record the runtimes of all methods on the large-sample regime and report them in Figure 3. Compared to the NN-based methods, i.e., AGMM, DeepIV, DeepGMM, our MMR-IV (NN) is the most computationally efficient method, which is clearly a result of a simpler objective. By using a minimax optimization between two NNs, AGMM is the least efficient method. DeepGMM and DeepIV are more efficient than AGMM, but are less efficient than MMR-IV (NN). Finally, all four RKHS-based methods, namely, KernelIV, AGMM-K, DualIV, and MMR-IV (RKHS), have similar computational time. All four methods are observed to scale poorly on large datasets. In contrast, SieveIV is the most computationally efficient method in this experiment.
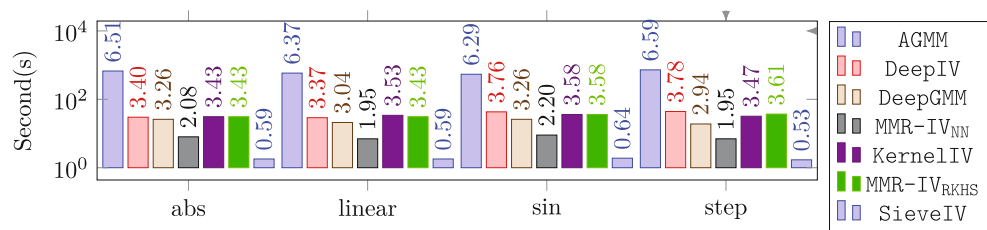
Figure 3: Runtime in the large-sample regime ($n = 2,000$). Time for parameter selection is excluded. AGMM-K (No Nyström), DualIV, and MMR-IV (RKHS) overlap due to the same runtime. Python and library versions are included in the "README" and "requirement" files in the software. We use 2.9 GHz Intel Core i5 for the experiment.

7.2 High-dimensional structured scenarios

In a high-dimensional setting, we employ the same data-generating process as in Section 7.1. We consider only the absolute function for $f^*(x) = |x|$, but map Z, X, or both X and Z to MNIST images (784-dim) [66]. Let us denote the original outputs in Section 7.1 by X^{low}, Z^{low}, and let $\pi(u) := \text{round}(\min(\max(1.5u + 5, 0), 9))$ be a transformation function mapping inputs to an integer between 0 and 9, and let $\text{RI}(d)$ be a function that selects a random MNIST image from the digit class d. Then, the scenarios we consider are (i) **MNIST$_Z$**: $Z \leftarrow \text{RI}(\pi(Z_1^{low}))$, (ii) **MNIST$_X$**: $X \leftarrow \text{RI}(\pi(X^{low}))$, and (iii) **MNIST$_{XZ}$**: $X \leftarrow \text{RI}(\pi(X^{low}))$, $Z \leftarrow \text{RI}(\pi(Z_1^{low}))$.

7.2.1 Experimental settings

We sample $n = 10,000$ points for the training, validation, and test sets, and run each method 10 times. For MMR-IV (Nyström), we run it only on the training set to reduce computational workload and use principal component analysis (PCA) to reduce dimensions of X from 728 to 8. We still use the sum of Gaussian kernels for $k(z, z')$, but use the automatic relevance determination (ARD) kernel for $l(x, x')$. We omit KernelIV and DualIV in this experiment since their kernels are not suitably designed for structured data and as we will see, so is MMR-IV (Nyström). We also omit SieveIV in this setting because the method requires a huge number of basis functions in the high-dimensional setting.

7.2.2 Experiment results

We report the results in Table 4. Like [14], we observe that the DeepIV code returns NaN when X is high dimensional, so the implementation has to be improved before being used for similar scenarios. Besides, we run Ridge2SLS, which is Poly2SLS with a fixed linear degree, in place of Poly2SLS to reduce computational workload, and as a result, its performance is unsatisfactory. 2SLS has large errors for high-dimensional X because the first-stage regression from Z to X is ill posed. The average performance of GMM+NN suggests that manually designed functions for instruments are insufficient to extract useful information. Furthermore, MMR-IV (NN) performs competitively across scenarios. On MNIST$_Z$, MMR-IVs perform better than other methods, which implies that using the sum of Gaussian kernels for the kernel k is proper. DeepGMM has competitive performance as well. On MNIST$_X$ and MNIST$_{XZ}$, MMR-IV (NN) outperforms all other methods. AGMM-K+NN, i.e., the KLayerTrained method by ref. [50], is a variant of AGMM-K, which uses NNs to model $f(X)$ and to extract features of Z as inputs of the kernel k. To reinforce the flexibility, AGMM-K+NN also trains the kernel hyperparameter, i.e., the scale of the RBF kernel k. Compared with MMR-IV (NN), AGMM-K+NN has a

Table 4: The MSE ± one standard deviation on high-dimensional structured data. We run each method 10 times

Algorithm	Setting		
	MNIST$_z$	MNIST$_x$	MNIST$_{xz}$
DirectNN	0.134 ± 0.000	*0.229 ± 0.000*	*0.196 ± 0.011*
2SLS	0.563 ± 0.001	>1,000	>1,000
Ridge2SLS	0.567 ± 0.000	0.431 ± 0.000	0.705 ± 0.000
GMM+NN	0.121 ± 0.004	0.235 ± 0.002	0.240 ± 0.016
AGMM	*0.017 ± 0.007*	0.732 ± 0.107	0.529 ± 0.163
DeepIV	0.114 ± 0.005	n/a	n/a
DeepGMM	0.038 ± 0.004	0.315 ± 0.130	0.333 ± 0.168
AGMM-K+NN	0.021 ± 0.007	1.05 ± 0.366	0.327 ± 0.192
MMR-IV (NN)	0.024 ± 0.006	**0.124 ± 0.021**	**0.130 ± 0.009**
MMR-IV (Nys)	**0.015 ± 0.002**	0.442 ± 0.000	0.425 ± 0.002

more flexible kernel k but lacks good kernel hyperparameter selection. So, it is vulnerable to a local minimum and shows unstable performance. The ARD kernel with PCA of `MMR-IV (Nyström)` fails because the features from PCA are not representative enough. In addition, we observe that `DeepGMM` often produces results that are unreliable across all settings. We suspect that the hard-to-optimize objective and the complicated optimization procedure are the causes. Compared with `DirectNN`, most of the baselines can hardly deal with high-dimensional structured instrumental regressions.

7.3 Mendelian randomization

We demonstrate our method in the setting of Mendelian randomization, which relies on genetic variants that satisfy the IV assumptions. The "exposure" X and outcome Y are univariate and generated from the simulation process by ref. [67]:

$$Y = \beta X + c_1 e + \delta, \quad X = \sum_{i=1}^{m} \alpha_i Z_i + c_2 e + \gamma,$$

where $Z \in \mathbb{R}^{d'}$ with each entry $Z_i \sim B(2, p_i)$, $p_i \sim \text{unif}(0.1, 0.9)$, $e \sim \mathcal{N}(0, 1)$, $\alpha_i \sim \text{unif}([0.8/d', 1.2/d'])$, and $\gamma, \delta \sim \mathcal{N}(0, 0.1^2)$. $Z_i \sim B(2, p_i)$ mimics the frequency of an individual getting one or more genetic variants. The parameters β, c_1 control the strength of exposures and confounders to outcomes, while c_2, α_i control the strength of instruments and confounders to exposures. We set $\alpha_i \sim \text{unif}([0.8/d', 1.2/d'])$ so that as the number of IVs increases, each IV becomes weaker while the overall strength of instruments ($\sum_i^{d'} \alpha_i$) remains constant.

In Mendelian randomization, it is known that genetic variants may act as *weak* IVs [5,68,69], so this experiment aims to evaluate the sensitivity of different methods to the number of instruments (d') and confounder strengths (c_1, c_2). We consider three scenarios: (i) $d' = 8, 16, 32$; (ii) $c_1 = 0.5, 1, 2$; (iii) $c_2 = 0.5, 1, 2$; unmentioned parameters use default values: $\beta = 1$, $d' = 16$, $c_1 = 1$, and $c_2 = 1$. We draw 10,000 samples for the training, validation, and test sets, respectively, and train `MMR-IV (Nyström)` only on the training set. Other settings are the same as those of the low-dim scenario.

Figure 4 depicts the experimental results. Overall, `2SLS` performs well on all settings due to the linearity assumption, except particular sensitivity to the number of (weak) instruments, which is a well-known property of `2SLS` [1]. Although imposing no such assumption, `MMR-IVs` perform competitively and even more stably, since the information of instruments is effectively captured by the kernel k and we only need to deal with a simple objective, and also the analytical CV error plays an essential role. `KernelIV` also achieves competitive and stable performance across all settings, but not as good as ours. `DirectNN` is always among the worst approaches on all settings as no instrument is used. `Poly2SLS` performs accurately on the last two experiments, while it presents significant instability with the number of instruments in Figure 4 (left) because of failure of the hyperparameter selection. In Figure 4 (middle) and Figure 4 (right), we can observe that the performance of most approaches deteriorates as the effect of confounders becomes stronger. `MMR-IV`
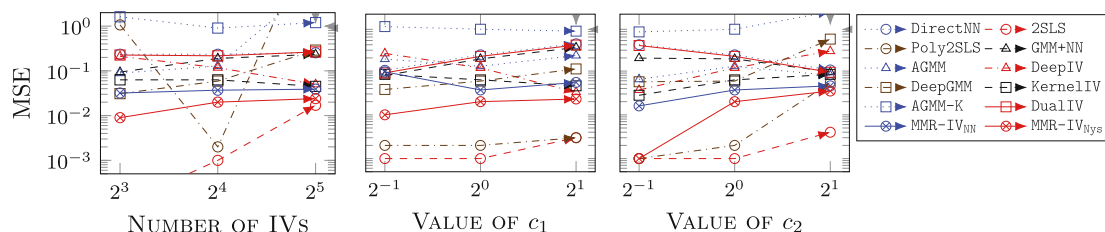

Figure 4: The MSE of different methods on Mendelian randomization experiments as we vary the numbers of instruments (left), the strength of confounders to exposures c_1 (middle), and the strength of confounders to instruments c_2 (right). The MSE is obtained from 10 repetitions of the experiment.

(Nyström) has promising performance and shows a bit more sensitivity to c_2 than c_1, and the good performance takes the advantage of the hyperparameter selection compared with AGMM-K and DualIV.

7.4 Vitamin D data

Finally, we apply our algorithm to the Vitamin D data [70, Section 5.1]. The data were collected from a 10-year study on 2,571 individuals aged 40 to 71 and 4 variables are employed: *age* (at baseline), *filaggrin* (a binary indicator of filaggrin mutations), *VitD* (Vitamin D level at baseline), and *death* (a binary indicator
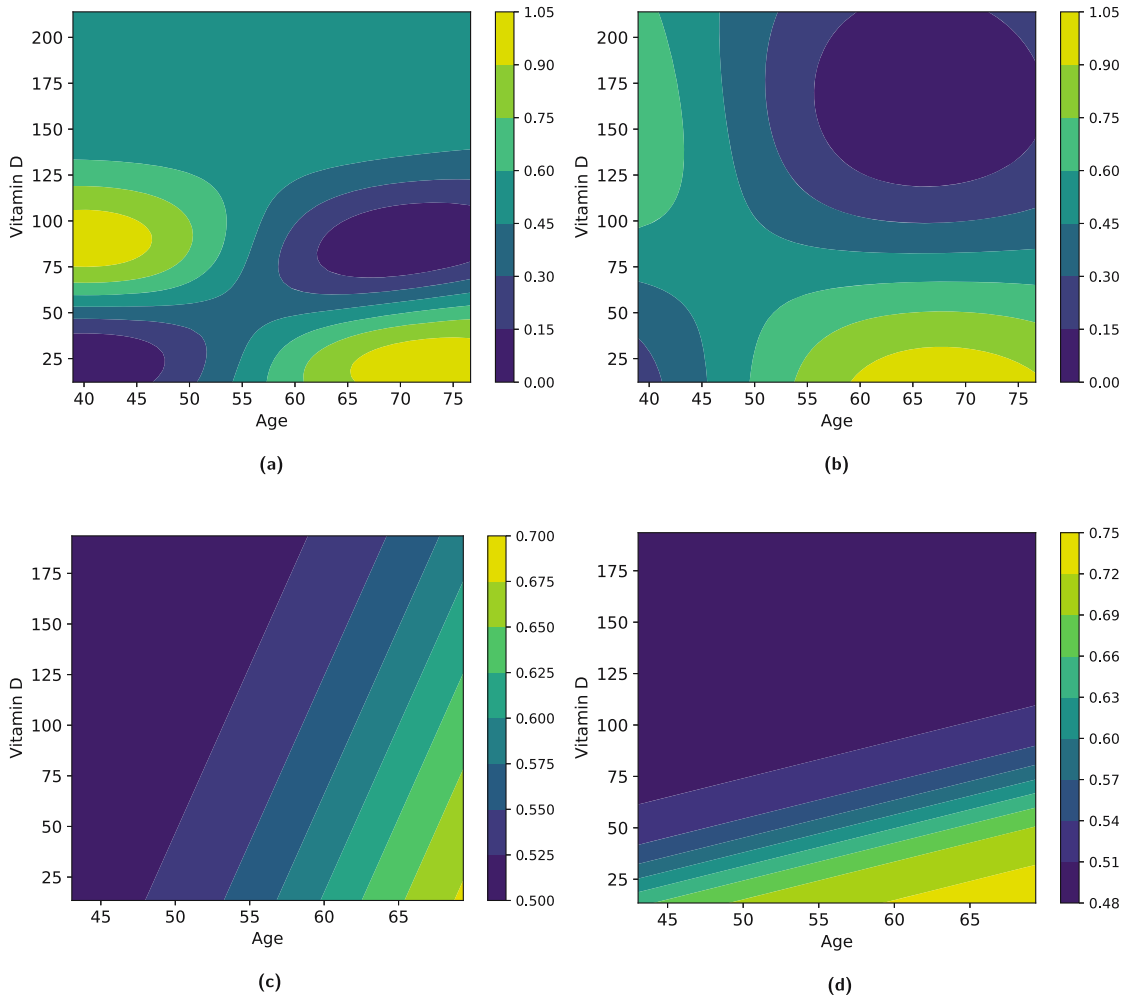
Figure 5: Estimated effect of vitamin D level on mortality rate, controlled by age. All plots depict normalized contours of $f'(X, C)$ defined in Section 7.4.1, where blue represents low mortality rate and yellow the opposite. We can divide each plot roughly into the left (young age) and right (old age) parts. While the right parts reflect similar information (i.e., lower vitamin D levels and old age lead to a higher mortality rate), the left parts are different. (a) A high level of vitamin D at a young age can result in a high mortality rate, which is counterintuitive. A plausible explanation is that it is caused by some unobserved confounders between vitamin D levels and mortality rates. (b) On the other hand, this spurious effect disappears when the filaggrin mutation is used as an instrument, i.e., a low vitamin D level at a young age has only a slight effect on death, but a more adverse effect at an old age [71]. This comparison demonstrates the benefit of an IV. (c) and (d) correspond to the results obtained by using [70]' GLM, from which we can draw similar conclusions. It is noteworthy that MMR-IV allows more flexible nonlinearity for a causal effect. (a) Kernel ridge regression (IV: none). (b) MMRIV (IV: Filaggrin mutation). (c) GLM (IV: none). (d) 2SLS + GLM (IV: Filaggrin mutation).

of death during the study). The goal is to evaluate the potential effect of VitD on death. We follow [70] by controlling age in the analyses, using filaggrin as an instrument, and then by applying the MMR-IV (Nyström) algorithm. Sjolander and Martinussen [70] modeled the effect of VitD on death by a generalized linear model (GLM) and found that the effect is insignificant by 2SLS (p-value on the estimated coefficient is 0.13 with the threshold of 0.05). More details are shown later. The estimated effect is illustrated in Figure 5 in the appendix. We observe that: (i) by using instruments, both our method and [70] output more reasonable results compared with those without instruments: a low VitD level at a young age has a slight effect on death, but a more adverse effect at an old age [71]; (ii) Unlike [70], our method allows more flexible nonlinearity for causal effect.

7.4.1 Experimental settings on Vitamin D data

We normalize each variable to have a zero mean and unit variance to reduce the influence of different scales. We consider two cases: (i) without instruments and (ii) with instruments. By without instrument, we mean that the W_V matrix of MMR-IV (Nyström) becomes an identity matrix. Following [70], we assess the effect of vitamin D (exposure) on mortality rate (outcome), control the age in the analyses, and use filaggrin as the instrument. We illustrate original vitamin D, age, and death in Figure 6. We randomly pick (random seed is 527) 300 Nyström samples and use leave-2-out CV to select hyperparameters. The GLMs in ref. [70] are a linear function in the first step and a logistic regression model in the second step.
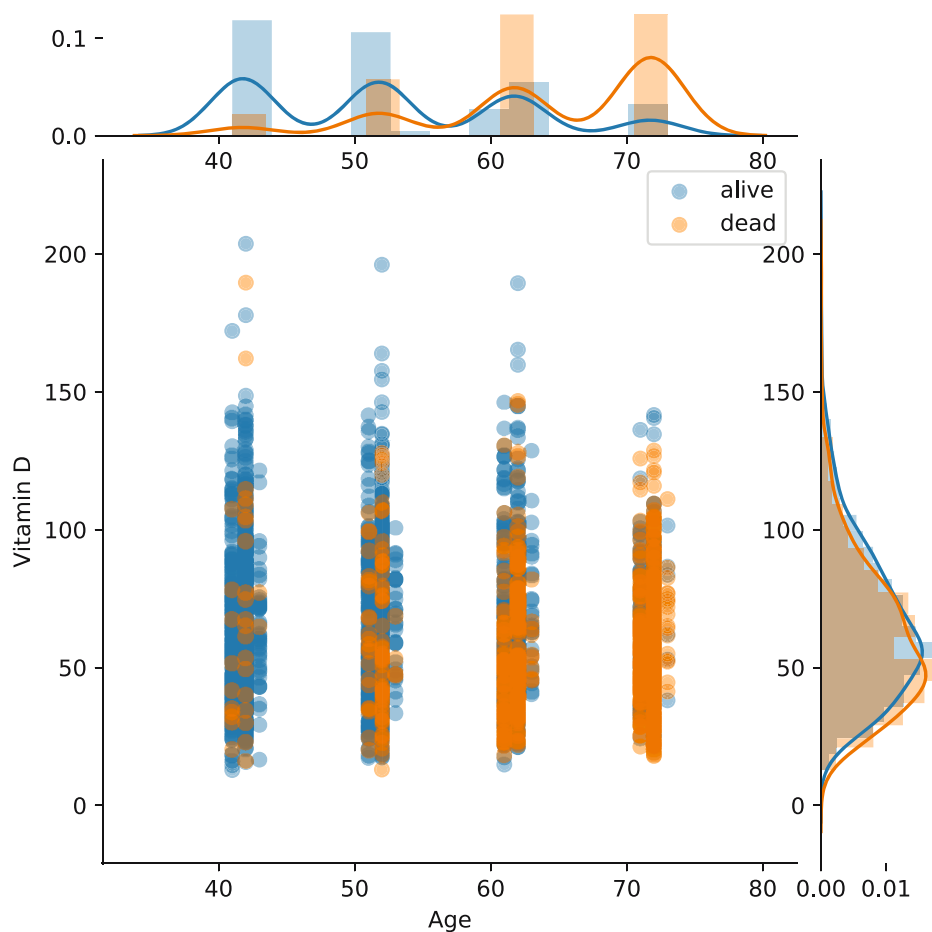

Figure 6: Distribution of Vitamin D data. Data points are plotted in the middle, the solid curve and histogram on the right describe the kernel density estimation and histogram of Vitamin D, and those on the top are for age.

In this experiment, we use *age* as a control variable by considering a structural equation model, which is similar to the model (1) except for the presence of the controlled (exogenous) variable C,

$$Y = f(X, C) + \varepsilon, \quad X = t(Z, C) + g(\varepsilon) + \nu, \tag{12}$$

where $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\nu] = 0$. We further assume that the instrument Z satisfies the following three conditions:

 (i) *relevance: Z has a causal influence on X*;
 (ii) *exclusion restriction: Z affects Y only through X, i.e., $Y \perp\!\!\!\perp Z|X, \varepsilon, C$;*
(iii) *unconfounded instrument(s): Z is conditionally independent of the error, i.e., $\varepsilon \perp\!\!\!\perp Z \mid C$.*

Unlike the conditions specified in the main text, (ii) and (iii) also include the controlled variable C. A similar model is employed in ref. [12]. From Assumption (iii), we can see that $\mathbb{E}[\varepsilon \mid C, Z] = \mathbb{E}[\varepsilon \mid C]$, and on the basis of this, we further obtain

$$\mathbb{E}[(Y - f(X, C) - \mathbb{E}[\varepsilon|C])h(Z, C)] = 0 \tag{13}$$

for every measurable function h. Note that $\mathbb{E}[\varepsilon \mid C]$ is only conditioned on C, remains constant on arbitrary values of C, and is typically nonzero. To adapt our method to this model, we only need to use the kernel k with (Z, C) as inputs and be aware that the output of the method is an estimate of $f'(X, C) \coloneqq f(X, C) + \mathbb{E}[\varepsilon|C]$ instead of just $f(X, C)$. For simplicity, we directly fit f' to binary Y without using such as the logistic transform $g(x) = e^x/(1 + e^x)$. This is because the transform requires nontrivial modification to the analytical CV error. We want to test MMR-IV (Nyström) with the exact analytical error proposed in the article.

8 Conclusion and discussion

Learning causal relations when hidden confounders are present is a cornerstone of reliable decision-making. IV regression is a standard tool to tackle this task, but currently faces challenges in nonlinear settings. This work presents a simple kernel-based framework that overcomes some of these challenges. We employ RKHS theory in the reformulation of CMR as a MMR based on which we can approach the problem from the ERM perspective. As we demonstrate, this framework not only facilitates theoretical analysis but also results in easy-to-use algorithms that perform well in practice compared to existing methods. This article also shows a way of combining the elegant theoretical approach of kernel methods with the practical merits of deep NNs. Despite these advantages, the optimal choice of the kernel k in the MMR objective remains an open question that we aim to address in future work.

The efficiency issue, i.e., the volume of the error by our estimator, is one of the major unsolved problems. Importantly, since the efficiency heavily depends on the choice of kernel function k, we have to discuss an appropriate method of kernel selection. This dependence has been clarified by the analysis of errors in the minimax problem by ref. [50], which shows that the error is upper bounded by a metric entropy of RKHS associated with the kernel k. Hence, the selection of optimal kernel functions is one of the important future directions.

In the classical IV problem with parametric models, the optimal choice of IV is based on the variance of the asymptotic distribution of estimators [30]. However, our setup corresponds to a nonparametric regression problem, which requires a more complicated approach. A recent work [72] tackles this problem, but some arbitrariness remains. Another challenge is recent works on a variational method of moments [73]. This work provides the optimal weighted objective and the closed-form expression of the estimator, which can achieve the semiparametric efficiency bound for solving the CMR. This approach would be highly relevant to the discussion of efficiency in our setting.

Author contributions: All authors have accepted responsibility for the entire content of this article and approved its submission. The study and manuscript have received significant contributions from all authors whose contributions are reflected in the list of authors.

Conflict of interest: The authors have a conflict of interest with the Australian National University, Data61 of Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia, Max Planck Institute (MPI), CISPA–Helmholtz Center for Information Security, and the University of Tokyo.

Ethical approval: The conducted research is not related to either human or animal use.

Data availability statement: The datasets generated during and analyzed during the current study are available at https://github.com/RuiZhang2016/MMRIV, where the implementation of the approaches is also provided.

References

[1] Angrist JD, Pischke JS. Mostly harmless econometrics: an empiricistas companion. Princeton and Oxford: Princeton University Press; 2008.

[2] Klungel O, Jamal Uddin M, de Boer A, Belitser S, Groenwold R, Roes K. Instrumental variable analysis in epidemiologic studies: an overview of the estimation methods. Pharm Anal Acta. 2015;6(353):2.

[3] Card D. The causal effect of education on earnings. In: Ashenfelter O, Card D, editors. Handbook of labor economics. vol. 3 of Handbook of labor economics. Elsevier; 1999. p. 1801–63.

[4] Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for Mendelian randomization. Statist Meth Med Res. 2017;26(5):2333–55.

[5] Burgess S, Foley CN, Allara E, Staley JR, Howson JMM. A robust and efficient method for Mendelian randomization with hundreds of genetic variants. Nature Commun. 2020;11(1):376.

[6] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Amer Stat Assoc. 1996;91(434):444–55.

[7] Hansen LP. Large sample properties of generalized method of moments estimators. Econometrica. 1982;50(4):1029–54.

[8] Newey WK, Powell JL. Instrumental variable estimation of nonparametric models. Econometrica. 2003;71(5):1565–78.

[9] Hall P, Horowitz JL. Nonparametric methods for inference in the presence of instrumental variables. Ann Stat. 2005 12;33(6):2904–2929.

[10] Blundell R, Chen X, Kristensen D. Semi-nonparametric IV estimation of shape-invariant Engel curves. Econometrica. 2007;75(6):1613–69.

[11] Horowitz JL. Applied nonparametric instrumental variables estimation. Econometrica. 2011;79(2):347–94.

[12] Hartford J, Lewis G, Leyton-Brown K, Taddy M. Deep IV: a flexible approach for counterfactual prediction. In: Proceedings of the 34th ICML. vol. 70. PMLR; 2017. p. 1414–23.

[13] Lewis G, Syrgkanis V. Adversarial generalized method of moments; ArXiv Preprint: 1803.07164; 2018.

[14] Bennett A, Kallus N, Schnabel T. Deep generalized method of moments for instrumental variable analysis. In: NeurIPS 32. Curran Associates, Inc., 2019. p. 3564–74.

[15] Muandet K, Mehrjou A, Lee SK, Raj A. Dual instrumental variable regression. In: NeurIPS 33. Curran Associates, Inc.; 2020.

[16] Singh R, Sahani M, Gretton A. Kernel instrumental variable regression. In: NeurIPS 32. Curran Associates, Inc.; 2019. p. 4593–605.

[17] Liao L, Chen Y, Yang Z, Dai B, Wang Z, Kolar M. Provably efficient neural estimation of structural equation model: an adversarial approach. In: NeurIPS 33. Curran Associates, Inc.; 2020.

[18] Bennett A, Kallus N. The variational method of moments; ArXiv Preprint: 2012.09422; 2020.

[19] Daskalakis C, Panageas I. The limit points of (optimistic) gradient descent in min-max optimization. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems; Curran Associates, Inc. 2018. p. 9256–66.

[20] Bailey JP, Gidel G, Piliouras G. Finite regret and cycles with fixed step-size via alternating gradient descent-ascent. In: Conference on Learning Theory. PMLR; 2020. p. 391–407.

[21] Mastouri A, Zhu Y, Gultchin L, Korba A, Silva R, Kusner MJ, et al. Proximal causal learning with kernels: two-stage estimation and moment restriction. In: Proceedings of the 38th International Conference on Machine Learning. vol. 139 of Proceedings of Machine Learning Research. PMLR; 2021. p. 7512–23.

[22] Horowitz JL, Lee S. Uniform confidence bands for functions estimated nonparametrically with instrumental variables. J Econom. 2012;168(2):175–88.

[23] Chen X, Christensen TM. Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. J Econom. 2015;188(2):447–65.

[24] Chen X, Christensen TM. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. Quant Econom. 2018;9(1):39–84.

[25] Babii A. Honest confidence sets in nonparametric IV regression and other ill-posed models. Econometric Theory. 2020;36(4):658–706.

[26] Chen X, Christensen T, Kankanala S. Adaptive estimation and uniform confidence bands for nonparametric iv. 2021. arXiv: http://arXiv.org/abs/arXiv:210711869.

[27] Muandet K, Jitkrittum W, Kübler J. Kernel conditional moment test via maximum moment restriction. In: Proceedings of the 36th Conference on UAI. vol. 124 of Proceedings of Machine Learning Research. PMLR; 2020. p. 41–50.

[28] Serfling R. Approximation theorems of mathematical statistics. New York, NY: John Wiley & Sons; 1980.

[29] Hable R. Asymptotic normality of support vector machine variants and other regularized kernel methods. J Multivariate Anal. 2012;106:92–117.

[30] Newey W. Efficient estimation of models with conditional moment restrictions. In: Handbook of statistics. vol. 11. Elsevier; 1993. p. 419–54.

[31] Donald SG, Newey WK. Choosing the number of instruments. Econometrica. 2001;69(5):1161–91.

[32] Hall AR. Generalized method of moments. Advanced texts in econometrics. Oxford: Oxford University Press; 2005.

[33] Aronszajn N. Theory of reproducing kernels. Trans Am Math Soc. 1950;68(3):337–404.

[34] Schölkopf B, Smola A. Learning with Kernels: support vector machines, regularization, optimization, and beyond. Cambridge, MA, USA: MIT Press; 2002.

[35] Berlinet A, Thomas-Agnan C. Reproducing Kernel Hilbert spaces in probability and statistics. Amsterdam: Kluwer Academic Publishers; 2004.

[36] Muandet K, Fukumizu K, Sriperumbudur B, Schölkopf B. Kernel mean embedding of distributions: a review and beyond. Foundations Trends Machine Learn. 2017;10(1–2):1–141.

[37] Steinwart I, Christmann A. Support vector machines. New York, NY: Springer Science & Business Media; 2008.

[38] Steinwart I. On the influence of the kernel on the consistency of support vector machines. JMLR 2002 Mar;2:67–93.

[39] Fukumizu K, Bach FR, Jordan MI. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. JMLR. 2004 Dec;5:73–99.

[40] Sriperumbudur BK, Fukumizu K, Lanckriet GRG. Universality, characteristic kernels and RKHS embedding of measures. JMLR. 2011 Jul;12:2389–410.

[41] Simon-Gabriel CJ, Schölkopf B. Kernel distribution embeddings: universal kernels, characteristic kernels and kernel metrics on distributions. JMLR. 2018;19(44):1–29.

[42] D'Haultfoeuille X. On the completeness condition in nonparametric instrumental problems. Econom Theory. 2011;27(3):460–71.

[43] Vapnik VN. Statistical learning theory. Hoboken, NJ, USA: Wiley-Interscience; 1998.

[44] Kariya T, Kurata H. Generalized least squares. Hoboken, NJ, USA: John Wiley & Sons; 2004.

[45] Van der Vaart A. Asymptotic statistics. Cambridge, UK: Cambridge University Press; 2000.

[46] Carrasco M, Florens JP, Renault E. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. In: Heckman JJ, Leamer EE, editors. Handbook of econometrics. vol. 6B. 1st ed. Elsevier; 2007.

[47] Schölkopf B, Herbrich R, Smola AJ. A generalized representer theorem. In: International Conference on Computational Learning Theory. Springer; 2001. p. 416–26.

[48] Williams CKI, Seeger M. Using the Nyström method to speed up kernel machines. In: NeurIPS 13. Cambridge, MA: MIT Press; 2001. p. 682–8.

[49] Flannery BP, Press WH, Teukolsky SA, Vetterling W. Numerical recipes in C. vol. 24. New York: Press Syndicate of the University of Cambridge; 1992. p. 78.

[50] Dikkala N, Lewis G, Mackey L, Syrgkanis V. Minimax estimation of conditional moment models. CoRR. 2020;abs/2006.07201.

[51] Vehtari A, Mononen T, Tolvanen V, Sivula T, Winther O. Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. JMLR. 2016;17(1):3581–618.

[52] Greene WH. Econometric analysis. Delhi: Pearson Education India; 2003.

[53] Newey WK, McFadden D. Chapter 36 Large sample estimation and hypothesis testing. In: Handbook of econometrics. vol. 4. Elsevier; 1994. p. 2111–245.

[54] Wainwright MJ. High-dimensional statistics: a non-asymptotic viewpoint. vol. 48. Cambridge, UK: Cambridge University Press; 2019.

[55] Van der Vaart A, Wellner J. Weak convergence and empirical processes. New York: Springer; 1996.

[56] Zhou DX. The covering number in learning theory. J Complexity. 2002;18(3):739–67.

[57] Steinwart I, Hush DR, Scovel C. Optimal rates for regularized least squares regression. COLT; 2009. p. 79–93.

[58] Caponnetto A, De Vito E. Optimal rates for the regularized least-squares algorithm. Foundation Comput Math. 2007;7(3):331–68.

[59] Chen X, Pouzo D. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. Econometrica. 2012;80(1):277–321.

[60] Chen X, Christensen TM. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. Quant Econ. 2018;9(1):39–84.

[61] Darolles S, Fan Y, Florens JP, Renault E. Nonparametric instrumental regression. Econometrica. 2011;79(5):1541–65.

[62] Song L, Huang J, Smola A, Fukumizu K. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In: Proceedings of the 26th ICML (ICML); 2009. p. 961–8.

[63] Song L, Fukumizu K, Gretton A. Kernel embeddings of conditional distributions: a unified kernel framework for nonparametric inference in graphical models. IEEE Signal Processing Magazine. 2013;30(4):98–111.

[64] Schölkopf B, Herbrich R, Smola AJ. A generalized representer theorem. In: Helmbold D, Williamson B, editors. Computational learning theory. Berlin, Heidelberg: Springer Berlin Heidelberg; 2001. p. 416–26.

[65] Ivanov VK, Vasin VV, Tanana VP. Theory of linear ill-posed problems and its applications. Inverse and ill-posed problems series. Boston: De Gruyter; 2002.

[66] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–324.

[67] Hartwig FP, Davey Smith G, Bowden J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. Int J Eepidemiol. 2017;46(6):1985–98.

[68] Kuang Z, Sala F, Sohoni N, Wu S, Córdova-Palomera A, Dunnmon J, et al. Ivy: instrumental variable synthesis for causal inference. In: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. vol. 108 of Proceedings of Machine Learning Research. PMLR; 2020. p. 398–410.

[69] Hartford JS, Veitch V, Sridhar D, Leyton-Brown K. Valid causal inference with (some) invalid instruments. CoRR. 2020;abs/2006.11386.

[70] Sjolander A, Martinussen T. Instrumental variable estimation with the R package ivtools. Epidemiol Meth. 2019;8(1):20180024.

[71] Meehan M, Penckofer S. The role of vitamin D in the aging adult. J Aging Gerontol. 2014;2(2):60.

[72] Zhang R, Muandet K, Schölkopf B, Imaizumi M. Instrument space selection for kernel maximum moment restriction. 2021. arXiv: http://arXiv.org/abs/arXiv:210603340.

[73] Bennett A, Kallus N. The variational method of moments. 2020. arXiv: http://arXiv.org/abs/arXiv:201209422.

[74] Hoeffding W. Probability inequalities for sums of bounded random variables. J Am Stat Assoc. 1963;58(301):13–30.

[75] Mann HB, Wald A. On stochastic limit and order relationships. Ann Math Stat. 1943 Sep;14 (3):217–26.

[76] Arcones MA, Gine E. Limit theorems for U-processes. Ann Probab. 1993;21(3):1494–542.

[77] Luenberger DG. Optimization by vector space methods. Hoboken, NJ: John Wiley & Sons; 1997.

[78] Akritas MG. Empirical processes associated with V-statistics and a class of estimators under random censoring. The Annals of Statistics. 1986;14(2):619–37.

[79] Feng Y, Li L, Liu Q. A kernel loss for solving the bellman equation. NeurIPS. Curran Associates, Inc. 2019;32.

[80] Liu Q, Li L, Tang Z, Zhou D. Breaking the curse of horizon: infinite-horizon off-policy estimation. In: NeurIPS. vol. 31. Curran Associates, Inc.; 2018.

[81] Uehara M, Huang J, Jiang N. Minimax weight and q-function learning for off-policy evaluation. In: ICML. PMLR; 2020. p. 9659–68.

[82] Kallus N. Balanced policy evaluation and learning. NeurIPS. Curran Associates, Inc. 2018;31.

[83] Kallus N. Generalized optimal matching methods for causal inference. JMLR. 2020;21(62):1–54.

[84] Wong RK, Chan KCG. Kernel-based covariate functional balancing for observational studies. Biometrika. 2018;105(1):199–213.

Appendix

A Detailed proofs

This section contains detailed proofs of the results that are missing in the main article. Most of the proofs on consistency and asymptotic normality take advantages of the useful resource by ref. [53]. Readers are referred to it for more detailed discussions on assumptions. Note that our proofs are based on the normed $\mathcal{F}$ space.

A.1 Proof of Lemma 1

Proof. Since $\mathcal{H}_k$ is the RKHS, we can rewrite (3) as follows:

$$
\begin{aligned}
R_k(f) &= \sup_{h \in \mathcal{H}_k, \|h\| \leq 1} (\mathbb{E}[(Y - f(X))\langle h, k(Z, \cdot)\rangle_{\mathcal{H}}])^2 \\
&= \sup_{h \in \mathcal{H}_k, \|h\| \leq 1} (\langle h, \mathbb{E}[(Y - f(X))k(Z, \cdot)]\rangle_{\mathcal{H}})^2 \qquad\qquad (A1) \\
&= \|\mathbb{E}[(Y - f(X))k(Z, \cdot)]\|^2_{\mathcal{H}_k},
\end{aligned}
$$

where we used the reproducing property of $\mathcal{H}_k$ in the first equality and the fact that $\mathcal{H}_k$ is a vector space in the last equality. By assumption, $\mathbb{E}[(Y - f(X))k(Z, \cdot)]$ is Bochner integrable [37, Def. A.5.20]. Hence, we can write (A1) as follows:

$$
\begin{aligned}
\|\mathbb{E}[(Y - f(X))k(Z, \cdot)]\|^2_{\mathcal{H}_k} &= \langle \mathbb{E}[(Y - f(X))k(Z, \cdot)], \mathbb{E}[(Y - f(X))k(Z, \cdot)]\rangle_{\mathcal{H}_k} \\
&= \mathbb{E}[\langle (Y - f(X))k(Z, \cdot), \mathbb{E}[(Y - f(X))k(Z, \cdot)]\rangle_{\mathcal{H}_k}] \\
&= \mathbb{E}[\langle (Y - f(X))k(Z, \cdot), (Y' - f(X'))k(Z', \cdot)\rangle_{\mathcal{H}_k}] \\
&= \mathbb{E}[(Y - f(X))(Y' - f(X'))k(Z, Z')],
\end{aligned}
$$

as required. □

A.2 Proof of Proposition 1

Proof. First, the law of iterated expectation implies that

$$
\mathbb{E}[(Y - f(X))k(Z, \cdot)] = \mathbb{E}_Z[\mathbb{E}_{XY}[(Y - f(X))k(Z, \cdot)|Z]] = \mathbb{E}_Z[\mathbb{E}_{XY}[Y - f(X)|Z]k(Z, \cdot)].
$$

By Lemma 1, we know that $R_k(f) = \|\mathbb{E}[(Y - f(X))k(Z, \cdot)]\|^2_{\mathcal{H}_k}$. As a result, $R_k(f) = 0$ if $\mathbb{E}[Y - f(X) \mid z] = 0$ for P_Z-almost all z. To show the converse, we assume that $R_k(f) = 0$ and rewrite it as follows:

$$
R_k(f) = \iint_{\mathcal{Z}} g(z)k(z, z')g(z')\mathrm{d}z\mathrm{d}z' = 0,
$$

where we define $g(z) := \mathbb{E}_{XY}[Y - f(X) \mid z]p(z)$. Since k is ISPD by assumption, this implies that g is a zero function with respect to P_Z, i.e., $\mathbb{E}[Y - f(X) \mid z] = 0$ for P_Z-almost all z. □

A.3 Proof of Proposition 2

Proof. Given $\alpha \in (0, 1)$ and any functions $f, g : \mathcal{X} \to \mathbb{R}$, we will show that

$$
R_k(\alpha f + (1 - \alpha)g) - \alpha R_k(f) - (1 - \alpha)R_k(g) < 0.
$$

By Lemma 1, we know that $R_k(f) = \|\mathbb{E}[(Y - f(X))k(Z, \cdot)]\|^2_{\mathcal{H}_k}$. Hence, we can rewrite the aforementioned function as follows:

$$R_k(\alpha f + (1-\alpha)g) - \alpha R_k(f) - (1-\alpha)R_k(g)$$
$$= \|\mathbb{E}[(Y - \alpha f(X) - (1-\alpha)g(X))k(Z,\cdot)]\|_{\mathcal{H}_k}^2 - \alpha\|\mathbb{E}[(Y - f(X))k(Z,\cdot)]\|_{\mathcal{H}_k}^2 - (1-\alpha)\|\mathbb{E}[(Y - g(X))k(Z,\cdot)]\|_{\mathcal{H}_k}^2$$
$$\overset{(a)}{=} \alpha(\alpha-1)\|\mathbb{E}[(Y - f(X))k(Z,\cdot)]\|_{\mathcal{H}_k}^2 + \alpha(\alpha-1)\|\mathbb{E}[(Y - g(X))k(Z,\cdot)]\|_{\mathcal{H}_k}^2$$
$$\quad - 2\alpha(\alpha-1)\langle\mathbb{E}[(Y - f(X))k(Z,\cdot)], \mathbb{E}[(Y - g(X))k(Z,\cdot)]\rangle_{\mathcal{H}_k}$$
$$\overset{(b)}{=} \alpha(\alpha-1)\underbrace{\|\mathbb{E}[(f(X) - g(X))k(Z,\cdot)]\|_{\mathcal{H}_k}^2}_{>0} < 0.$$

The equality (a) is obtained by considering $Y = \alpha Y + (1-\alpha)Y$ in $\|\mathbb{E}[(Y - \alpha f(X) - (1-\alpha)g(X))k(Z,\cdot)]\|_{\mathcal{H}_k}^2$ on the left-hand side of (a). We note that the right-hand side of (a) is quadratic in $\|\mathbb{E}[(Y - f(X))k(Z,\cdot)]\|_{\mathcal{H}_k}$ and $\|\mathbb{E}[(Y - g(X))k(Z,\cdot)]\|_{\mathcal{H}_k}$, and can be further expressed as a square binomial as the right-hand side of (b). Therefore, the convexity follows from the fact that k is the ISPD kernel, $\|\mathbb{E}[(f(X) - g(X))|Z]\|_2 \neq 0$ and $\alpha(\alpha-1) < 0$. □

A.4 Uniform convergence of risk functionals

The results presented in this section are used to prove the consistency of $\hat{f}_V$ and $\hat{f}_U$.

Lemma A1. (Uniform consistency of $\widehat{R}_V(f)$) *Assume that* $\mathbb{E}[|Y|^2] < \infty$, $\mathcal{F}$ *is compact,* $\mathbb{E}[\sup_{f\in\mathcal{F}}|f(X)|^2] < \infty$, *and Assumption* 1 *holds. Then, the risk* $R_k(f)$ *is continuous about* $f \in \mathcal{F}$ *and* $\sup_{f\in\mathcal{F}}|\widehat{R}_V(f) - R_k(f)| \overset{p}{\to} 0$.

Proof. First, let $u := (x, y, z)$, $u' := (x', y', z')$, and $h_f(u, u') := (y - f(x))(y' - f(x'))k(z, z')$ for some $(x, y, z), (x', y', z') \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. To prove that $\widehat{R}_V$ converges uniformly to R_k, we need to show that (i) $h_f(u, u')$ is continuous at each f with probability one; (ii) $\mathbb{E}_{U,U'}[\sup_{f\in\mathcal{F}}|h_f(U, U')|] < \infty$, and $\mathbb{E}_{U,U}[\sup_{f\in\mathcal{F}}|h_f(U, U)|] < \infty$ [53, Lemma 8.5]. To this end, it is easy to see that

$$|h_f(u, u')| = |(y - f(x))(y' - f(x'))k(z, z')|$$
$$\leq |y - f(x)||y' - f(x')||k(z, z')|$$
$$\leq |y - f(x)||y' - f(x')|\sqrt{k(z, z)k(z', z')}$$
$$\leq (|y| + |f(x)|)(|y'| + |f(x')|)\sqrt{k(z, z)k(z', z')}.$$

The third inequality follows from the Cauchy-Schwarz inequality. Since $\mathcal{F}$ is compact, every $f \in \mathcal{F}$ has $f(x)$ bounded for $\|x\| < \infty$. In term of $k(\cdot, \cdot)$ is bounded as per Assumption 1, we have $h_f(u, u') < \infty$, and thus, $h_f(u, u')$ continuous at each f with probability one. Furthermore, we obtain the following inequalities:

$$\mathbb{E}_{U,U'}\left[\sup_{f\in\mathcal{F}}|h_f(U, U')|\right] \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}(|Y| + |f(X)|)(|Y'| + |f(X')|)\sqrt{k(Z, Z)k(Z', Z')}\right]$$
$$\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}(|Y| + |f(X)|)\sup_{f\in\mathcal{F}}(|Y'| + |f(X')|)\right]\sup_z k(z, z)$$
$$= \mathbb{E}\left[\sup_{f\in\mathcal{F}}(|Y| + |f(X)|)\right]^2 \sup_z k(z, z)$$
$$= \mathbb{E}\left[|Y| + \sup_{f\in\mathcal{F}}|f(X)|\right]^2 \sup_z k(z, z) < \infty$$
$$\mathbb{E}_{U,U}\left[\sup_{f\in\mathcal{F}}|h_f(U, U)|\right] \leq \mathbb{E}_u\left[\sup_{f\in\mathcal{F}}(|Y| + |f(X)|)^2\right]\sup_z k(z, z)$$
$$= \mathbb{E}\left[(|Y| + \sup_{f\in\mathcal{F}}|f(X)|)^2\right]\sup_z k(z, z)$$
$$\leq 2(\mathbb{E}[|Y|^2] + \mathbb{E}\left[\sup_{f\in\mathcal{F}}|f(X)|^2\right])\sup_z k(z, z) < \infty.$$

Hence, our assertion follows from [53, Lemma 8.5].   □

Lemma A2. (Uniform consistency of $\widehat{R}_U(f)$) *Assume that* $\mathbb{E}[\|Y\|] < \infty$, $\mathcal{F}$ *is compact,* $\mathbb{E}[\|f(X)\|] < \infty$ *and Assumption* 1 *holds. Then,* $R_k(f)$ *is continuous about* f *and* $\sup_{f \in \mathcal{F}} |\widehat{R}_U(f) - R_k(f)| \xrightarrow{p} 0$.

Proof. First, let $u := (x, y, z)$, $u' := (x', y', z')$, and $h_f(u, u') := (y - f(x))(y' - f(x'))k(z, z')$ for some $(x, y, z), (x', y', z') \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. To prove the uniform consistency of $\widehat{R}_U$, we need to show that (i) $h_f(u, u')$ is continuous at each f with probability one; (ii) there is $d(u, u')$ with $|h_f(u, u')| \leq d(u, u')$ for all $f \in \mathcal{F}$ and $\mathbb{E}_{U, U'}[d(U, U')] < \infty$ [53, Lemma 2.4]; (iii) $(u_i, u_j)_{i \neq j}^{n, n}$ has strict stationarity and ergodicity in the sense of [53, Footnote 18 in p. 2129]. To this end, it is easy to see that

$$
\begin{aligned}
|h_f(u, u')| &= |(y - f(x))(y' - f(x'))k(z, z')| \\
&\leq |y - f(x)||y' - f(x')||k(z, z')| \\
&\leq |y - f(x)||y' - f(x')|\sqrt{k(z, z)k(z', z')} \\
&\leq (|y| + |f(x)|)(|y'| + |f(x')|)\sqrt{k(z, z)k(z', z')} \equiv d(u, u').
\end{aligned}
$$

The third inequality follows from the Cauchy-Schwarz inequality. Since $\mathcal{F}$ is compact, every $f \in \mathcal{F}$ has $f(x)$ bounded for $\|x\| < \infty$. In terms of $k(\cdot, \cdot)$ is bounded as per Assumption 1, we have $h_f(u, u') < \infty$, and thus, it proves that (i) $h_f(u, u')$ is continuous at each f with probability one. To prove (ii) $\mathbb{E}_{U, U'}[d(U, U')] < \infty$, we show that

$$
\mathbb{E}_{U, U'}[d(U, U')] \leq \mathbb{E}[|Y| + |f(X)|]^2 \sup_z k(z, z) < \infty.
$$

Furthermore, we show that $(u_i, u_j)_{i \neq j}^{n, n}$ has strict stationarity and ergodicity. Strict stationarity means that the distribution of a set of data $(u_i, u_{j \neq i})_{i=1, j=1}^{i+m, j+m'}$ does not depend on the starting indices i, j for any m and m', which is easy to check. Ergodicity means that $\widehat{R}_U(f) \xrightarrow{p} R_k(f)$ for all $f \in \mathcal{F}$ and $\mathbb{E}[|h_f(U, U')|] < \infty$. We have already shown that $h_f(u, u')$ is bounded and $\mathbb{E}[|h_f(U, U')|] < \infty$, so $\widehat{R}_U(f) \xrightarrow{p} R_k(f)$ follows by [74, P.25]. Therefore, ergodicity holds, and we have shown all conditions required by extended results of [53, Lemma 2.4]. Then, it follows that $\sup_{f \in \mathcal{F}} |\widehat{R}_U(f) - R_k(f)| \xrightarrow{p} 0$ and $R_k(f)$ is continuous.   □

A.5 Indefiniteness of weight matrix W_U

Theorem A1. *If Assumption* 1 *holds,* W_U *is indefinite.*

Proof. By definition, we have

$$
W_U = \frac{1}{n(n-1)}[K(\boldsymbol{z}, \boldsymbol{z}) - \mathrm{diag}(k(z_1, z_1), \ldots, k(z_n, z_n))] = \frac{1}{n(n-1)}K_U,
$$

where $\mathrm{diag}(a_1, \ldots, a_n)$ denotes an $n \times n$ diagonal matrix whose diagonal elements are $a_1, \ldots, a_n$. We can see that the diagonal elements of K_U are zeros, and therefore, $\mathrm{trace}(W_U) = 0$. Let us denote the eigenvalues of W_U by $\{\lambda_i\}_{i=1}^n$. Since $\sum_{i=1}^n \lambda_i = \mathrm{trace}(W_U)$, we conclude that there exist both positive and negative eigenvalues (all eigenvalues being zeros yields trivial $W_U = \mathbf{0}$). As a result, W_U is indefinite.   □

A.6 Consistency of $\hat{f}_V$ with convex $\Omega(f)$

Theorem A2. (Consistency of $\hat{f}_V$ with convex $\Omega(f)$) *Assume that $\mathcal{F}$ is a convex set, f^* is an interior point of $\mathcal{F}$, $\Omega(f)$ is convex about f, $\lambda \xrightarrow{\mathrm{p}} 0$, and Assumptions 1 and 2 hold. Then, $\hat{f}_V$ exists with probability approaching one and $\hat{f}_V \xrightarrow{\mathrm{p}} f^*$.*

Proof. Given $\Omega(f)$ is convex about f, we prove the consistency based on [53, Theorem 2.7], which requires (i) $R_k(f)$ is uniquely maximized at f^*; (ii) $\widehat{R}_V(f) + \lambda\Omega(f)$ is convex; (iii) $\widehat{R}_V(f) + \lambda\Omega(f) \xrightarrow{\mathrm{p}} R_k(f)$ for all $f \in \mathcal{F}$.

Recall that $\widehat{R}_V(f) = \|\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))k(z_i, \cdot)\|_{\mathcal{H}_k}^2$, and by the law of large number, we have that $\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))k(z_i, \cdot) \xrightarrow{\mathrm{p}} \mathbb{E}[(Y - f(X))k(Z, \cdot)]$. Then $\widehat{R}_V(f) \xrightarrow{\mathrm{p}} R_k(f)$ follows from the continuous mapping theorem [75] based on the fact that the function $g(\cdot) = \|\cdot\|_{\mathcal{H}_k}^2$ is continuous. As $\lambda \xrightarrow{\mathrm{p}} 0$, we obtain (iii) $\widehat{R}_V(f) + \lambda\Omega(f) \xrightarrow{\mathrm{p}} R_k(f)$ by Slutsky's theorem [45, Lemma 2.8]. Besides, it is easy to see that $\widehat{R}_V(f)$ is convex because the weight matrix W_V is positive definite, and (ii) $\widehat{R}_V(f) + \lambda\Omega(f)$ is convex due to convex $\Omega(f)$. Further, the condition (i) directly follows from Proposition 2, and given that f^* is an interior point of the convex set $\mathcal{F}$, our assertion follows from [53, Theorem 2.7]. □

A.7 Proof of Proposition 4

Proof. From the conditions of Lemma A1, we know that $\mathcal{F}$ is compact, $R_k(f)$ is continuous about f and $\sup_{f \in \mathcal{F}}|\widehat{R}_V(f) - R_k(f)| \xrightarrow{p} 0$. As Assumptions 1 and 2 hold, $R_k(f)$ is uniquely minimized at f^*. On the basis of the conditions that $\Omega(f)$ is bounded and $\lambda \xrightarrow{\mathrm{p}} 0$, we obtain by Slutsky's theorem that

$$\sup_{f \in \mathcal{F}}|\widehat{R}_V(f) + \lambda\Omega(f) - R_k(f)| \leq \sup_{f \in \mathcal{F}}|\widehat{R}_V(f) - R_k(f)| + \lambda\sup_{f \in \mathcal{F}}\Omega(f) \xrightarrow{\mathrm{p}} 0.$$

Consequently, we assert the conclusion by [53, Theorem 2.1]. □

A.8 Consistency of $\hat{f}_U$

Theorem A3. (Consistency of $\hat{f}_U$) *Assume that conditions of Lemma A2 and Assumption 2 hold, $\Omega(f)$ is a bounded function and $\lambda \xrightarrow{\mathrm{p}} 0$. Then $\hat{f}_U \xrightarrow{\mathrm{p}} f^*$.*

Proof. By the conditions of Lemma A2, we know that $\mathcal{F}$ is compact, $R_k(f)$ is continuous about f, and $\sup_{f \in \mathcal{F}}|\widehat{R}_U(f) - R_k(f)| \xrightarrow{p} 0$. As Assumptions 1 and 2 hold, $R_k(f)$ is uniquely minimized at f^*. On the basis of the conditions that $\Omega(f)$ is bounded and $\lambda \xrightarrow{\mathrm{p}} 0$, we obtain by Slutsky's theorem that

$$\sup_{f \in \mathcal{F}}|\widehat{R}_U(f) + \lambda\Omega(f) - R_k(f)| \leq \sup_{f \in \mathcal{F}}|\widehat{R}_U(f) - R_k(f)| + \lambda\sup_{f \in \mathcal{F}}\Omega(f) \xrightarrow{\mathrm{p}} 0.$$

Consequently, we assert the conclusion by [53, Theorem 2.1]. □

A.9 Asymptotic normality of $\hat{\theta}_U$

In this section, we consider the regularized U-statistic risk $\widehat{R}_{U,\lambda}(f_\theta)$. For $u_i := (x_i, y_i, z_i)$ and $u_j := (x_j, y_j, z_j)$, we express it in a compact form:

$$\widehat{R}_{U,\lambda}(f_\theta) := \underbrace{\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} h_\theta(u_i, u_j)}_{\widehat{R}_U(f_\theta)} + \lambda\Omega(\theta)$$

$$h_\theta(u_i, u_j) := (y_i - f_\theta(x_i))k(z_i, z_j)(y_j - f_\theta(x_j)).$$

We will assume that f_θ and $\Omega(\theta)$ are twice continuously differentiable about θ. The first-order derivative $\nabla_\theta\widehat{R}_U(f_\theta)$ can also be written as follows:

$$\nabla_\theta\widehat{R}_{U,\lambda}(f_\theta) = \underbrace{\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \nabla_\theta h_\theta(u_i, u_j)}_{\nabla_\theta\widehat{R}_U(f_\theta)} + \lambda\nabla_\theta\Omega(\theta)$$

$$\nabla_\theta h_\theta(u_i, u_j) = -[(y_i - f_\theta(x_i))\nabla_\theta f_\theta(x_j) + (y_j - f_\theta(x_j))\nabla_\theta f_\theta(x_i)]k(z_i, z_j).$$

A.10 Asymptotic normality of $\nabla_\theta\widehat{R}_{U,\lambda}(f_{\theta^*})$

We first show the asymptotic normality of $\nabla_\theta\widehat{R}_{U,\lambda}(f_{\theta^*})$. We assume that there exists $z \in \mathcal{Z}$ such that $\mathbb{E}_X[\nabla_\theta f_{\theta^*}(X) \mid z]p(z) \neq 0$ or $\mathbb{E}_{XY}[Y - f_{\theta^*}(X) \mid z]p(z) \neq 0$. Both terms being equal to zeros for all $z \in \mathcal{Z}$ leads to a singular $\nabla_\theta^2\widehat{R}_U(f_{\theta^*})$ and the asymptotic distribution therefore becomes much more complicated to analyze.

Lemma A3. *Suppose that f_θ and $\Omega(\theta)$ are first continuously differentiable about θ, $\mathbb{E}[\|\nabla_\theta h_{\theta^*}(U, U')\|_2^2] < \infty$, there exists $z \in \mathcal{Z}$ such that $\mathbb{E}_X[\nabla_\theta f_{\theta^*}(X) \mid z]p(z) \neq 0$ or $\mathbb{E}_{XY}[Y - f_{\theta^*}(X) \mid z]p(z) \neq 0$, and $\sqrt{n}\lambda \xrightarrow{p} 0$. Then,*

$$\sqrt{n}\nabla_\theta\widehat{R}_{U,\lambda}(f_{\theta^*}) \xrightarrow{p} N\big(\mathbf{0}, 4\mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[\nabla_\theta h_{\theta^*}(U, U')\big]\big]\big)\big).$$

Proof. The proof follows from [28, Sections 5.5.1 and 5.5.2] and we need to show that (i) $\nabla_\theta\widehat{R}_U(f_{\theta^*}) \xrightarrow{p} \mathbf{0}$ and (ii) whether $\mathrm{V\,ar}_U[\mathbb{E}_{U'}[\nabla_\theta h_{\theta^*}(U, U')]] > 0$. (i) can be obtained by the law of large numbers because $\nabla_\theta\widehat{R}_U(f_{\theta^*})$ is a sample average of $\nabla_\theta R_k(f_{\theta^*}) = \mathbf{0}$.

To prove (ii), we first note that $\mathrm{Var}_U[\mathbb{E}_{U'}[\nabla_\theta h_{\theta^*}(U, U')]] = \mathbb{E}_U[\mathbb{E}_{U'}^2[\nabla_\theta h_{\theta^*}(U, U')]] - \underbrace{\mathbb{E}_{UU'}^2[\nabla_\theta h_{\theta^*}(U, U')]}_{=\mathbf{0}} \geq \mathbf{0}$,

where equality holds if for any U, there is $\mathbb{E}_{U'}[\nabla_\theta h_{\theta^*}(U, U')] = 0$, i.e.,

$$\mathbb{E}_{U'}\big[\nabla_\theta h_{\theta^*}(U, U')\big] = -\mathbb{E}_{X'Z'}\big[\nabla_\theta f_{\theta^*}(X')k(Z', Z)\big]\big(Y - f_{\theta^*}(X)\big) - \mathbb{E}_{X'Y'Z'}\big[\big(Y' - f_{\theta^*}(X')\big)k(Z', Z)\big]\nabla_\theta f_{\theta^*}(X) = \mathbf{0}.$$

As the aforementioned equation holds for any Y, the coefficient of Y must be 0:

$$\mathbb{E}_{X'Z'}\big[\nabla_\theta f_{\theta^*}(X')k(Z', Z)\big] = \mathbb{E}_{Z'}\big[\mathbb{E}_{X'}\big[\nabla_\theta f_{\theta^*}(X')|Z'\big]k(Z', Z)\big] = 0,$$

where we note that $\mathbb{E}[\nabla_\theta f_{\theta^*}(X') \mid Z']p(Z') = 0$ for any Z' implied by the second function earlier. Similarly, the coefficient of $\nabla_\theta f_{\theta^*}(X)$ must be zero, which implies that $\mathbb{E}_{X'Y'}[(Y' - f_{\theta^*}(X')) \mid Z']p(Z') = 0$ for any Z'. The two coefficients cannot be zero at the same time (otherwise against the given conditions), so $\mathrm{Var}_U[\mathbb{E}_{U'}[\nabla_\theta h_{\theta^*}(U, U')]] > 0$. Further, due to the given condition $\mathbb{E}[\|\nabla_{\theta^*} h(U, U')\|_2^2] < \infty$, we obtain $\sqrt{n}\nabla_\theta\widehat{R}_U(f_{\theta^*}) \xrightarrow{p} N(\mathbf{0}, 4\mathbb{E}_U[\mathbb{E}_{U'}^2[\nabla_\theta h_{\theta^*}(U, U')]])$ as per [28, Section 5.5.1]. Finally, as $\sqrt{n}\lambda \xrightarrow{p} 0$ and $\nabla_\theta\Omega(\theta^*) < \infty$ by the condition that $\Omega(\theta)$ is first continuously differentiable, we assert the conclusion by Slutsky's theorem,

$$\sqrt{n}\nabla_\theta\widehat{R}_{U,\lambda}(f_{\theta^*}) = \sqrt{n}\nabla_\theta\widehat{R}_U(f_{\theta^*}) + \sqrt{n}\lambda\nabla_\theta\Omega(\theta^*) \xrightarrow{p} N\big(\mathbf{0}, 4\mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[\nabla_\theta h_{\theta^*}(U, U')\big]\big]\big)\big).$$

This concludes the proof.　　　　□

A.11 Uniform consistency of $\nabla^2_\theta \widehat{R}_{U,\lambda}(f_\theta)$

Next, we consider the second derivative $\nabla^2_\theta \widehat{R}_{U,\lambda}(f_\theta)$ and show its uniform consistency. In what follows, we denote by $\|\cdot\|_F$ the Frobenius norm. We can express $\nabla^2_\theta \widehat{R}_{U,\lambda}(f_\theta)$ as follows:

$$\nabla^2_\theta \widehat{R}_{U,\lambda}(f_\theta) = \underbrace{\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n \nabla^2_\theta h_\theta(u_i, u_j)}_{\nabla^2_\theta \widehat{R}_U(f_\theta)} + \lambda \nabla^2_\theta \Omega(\theta)$$

$$\nabla^2_\theta h_\theta(u_i, u_j) = [\nabla_\theta f_\theta(x_i)\nabla_\theta f_\theta^\top(x_j) - (y_i - f_\theta(x_i))\nabla^2_\theta f_\theta(x_j) + \nabla_\theta f_\theta(x_j)\nabla_\theta f_\theta^\top(x_i) - (y_j - f_\theta(x_j))\nabla^2_\theta f_\theta(x_i)]k(z_i, z_j).$$

Lemma A4. *Suppose that f_θ and $\Omega(\theta)$ are twice continuously differentiable about θ, Θ is compact, $\mathbb{E}[|f_\theta(X)|] < \infty$, $\mathbb{E}[\|\nabla_\theta f_\theta(X)\|_2] < \infty$, $\mathbb{E}[\|\nabla^2_\theta f_\theta(X)\|_F] < \infty$, $\mathbb{E}[|Y|] < \infty$, $\lambda \overset{p}{\to} 0$ and Assumption 1 holds. Then, $\mathbb{E}[\nabla^2_\theta h_\theta(U, U')]$ is continuous about θ and*

$$\sup_{\theta \in \Theta} \|\nabla^2_\theta \widehat{R}_{U,\lambda}(f_\theta) - \mathbb{E}[\nabla^2_\theta h_\theta(U, U')]\|_F \overset{p}{\to} 0.$$

Proof. The proof is similar to that of Lemma A2 and both applies extended results of [53, Lemma 2.4]. As $(u_i, u_j)_{i \neq j}$ being strictly stationary in the sense of [53, Footnote 18 in p. 2129] has been shown in Lemma A2, we only need to show that (i) $\nabla^2_\theta h_\theta(u, u')$ is continuous at each $\theta \in \Theta$ with probability one and (ii) there exists $d(u, u') \geq \|\nabla^2_\theta h_\theta(u, u')\|_F$ for all $\theta \in \Theta$ and $\mathbb{E}[d(U, U')] < \infty$. We exploit the triangle inequality of the Frobenius norm and obtain

$$\|\nabla^2_\theta h_\theta(u, u')\|_F \leq [2\|\nabla_\theta f_\theta(x)\nabla_\theta f_\theta^\top(x')\|_F + (|y| + |f_\theta(x)|)\|\nabla^2_\theta f_\theta(x')\|_F + (|y'| + |f_\theta(x')|)\|\nabla^2_\theta f_\theta(x)\|_F]k(z, z') \equiv d(u, u'),$$

We first show $d(u, u')$ is bounded for bounded u, u'. As f_θ is twice continuously differentiable about θ and Θ is compact, we have $f_\theta(x)$ bounded as well as each entry of $\nabla_\theta f_\theta(x)$ and $\nabla^2_\theta f_\theta(x)$ for $\|x\| < \infty$. Further taking into account that $k(\cdot, \cdot)$ is bounded as per Assumption 1, we know that $d(u, u') < \infty$ if u, u' are bounded, and it follows that (i) $\nabla^2_\theta h_\theta(u, u')$ is continuous at each $\theta \in \Theta$ with probability one as f_θ is twice continuously differentiable.

We then show that (ii) $\mathbb{E}_{U,U'}[d(U, U')] < \infty$ by the following inequalities:

$$\mathbb{E}_{U,U'}[d(U, U')] \leq 2\mathbb{E}[\|\nabla_\theta f_\theta(X)\nabla_\theta f_\theta^\top(X')\|_F + (|Y| + |f_\theta(X)|)\|\nabla^2_\theta f_\theta(X')\|_F] \sup_z k(z, z)$$

$$= 2(\underbrace{\mathbb{E}[\|\nabla_\theta f_\theta(X)\|_F]\mathbb{E}[\|\nabla_\theta f_\theta^\top(X')\|_F]}_{<\infty} + \underbrace{\mathbb{E}[|Y| + |f_\theta(X)|]}_{<\infty}\underbrace{\mathbb{E}[\|\nabla^2_\theta f_\theta(X')\|_F]}_{<\infty}) \sup_z k(z, z)$$

$$< \infty.$$

Therefore, we obtain $\sup_{\theta \in \Theta} \|\nabla^2_\theta \widehat{R}_U(f_\theta) - \mathbb{E}[\nabla^2_\theta h_\theta(U, U')]\|_F \overset{p}{\to} 0$ following from the extended results in the remarks of [53, Lemma 2.4]. Furthermore, from the conditions that $\Omega(\theta)$ is twice continuously differentiable and the parameter space Θ is compact, we obtain that $\|\nabla^2_\theta \Omega(\theta)\|_F < \infty$ for any $\theta \in \Theta$. Finally, it follows from the Slutsky's theorem that

$$\sup_{\theta \in \Theta} \|\nabla^2_\theta \widehat{R}_{U,\lambda}(f_\theta) - \mathbb{E}[\nabla^2_\theta h_\theta(U, U')]\|_F \leq \sup_{\theta \in \Theta} \|\nabla^2_\theta \widehat{R}_U(f_\theta) - \mathbb{E}[\nabla^2_\theta h_\theta(U, U')]\|_F + \lambda \sup_{\theta \in \Theta} \|\nabla^2_\theta \Omega(\theta)\|_F \overset{p}{\to} 0.$$

This concludes the proof.　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　□

Theorem A4. *(Asymptotic normality of $\hat{\theta}_U$) Suppose that $H = \mathbb{E}[\nabla^2_\theta h_{\theta^*}(U, U')]$ is nonsingular, Θ compact, $\mathbb{E}[|f_\theta(X)|] < \infty$, $\mathbb{E}[|Y|] < \infty$, f_θ and $\Omega(\theta)$ are twice continuously differentiable about θ, $\mathbb{E}[\|\nabla_\theta f_\theta(X)\|_2] < \infty$, $\mathbb{E}[\|\nabla^2_\theta f_\theta(X)\|_F] < \infty$, $\sqrt{n}\lambda \overset{p}{\to} 0$, $R_k(f_\theta)$ is uniquely minimized at θ^* which is an interior point of Θ, $\mathbb{E}[\|\nabla_\theta h_{\theta^*}(U, U')\|_2^2] < \infty$ and Assumption 1 hold. Then*

$$\sqrt{n}(\hat{\theta}_U - \theta^*) \overset{p}{\to} N\left(\mathbf{0}, 4H^{-1}\text{diag}\left(\mathbb{E}_U\left[\mathbb{E}^2_{U'}\left[h_{\theta^*}(U, U')\right]\right]\right)H^{-1}\right).$$

Proof. The proof follows by [53, Theorem 3.1], and we need to show that (i) $\widehat{\theta}_U \overset{p}{\to} \theta^\star$; (ii) $\widehat{R}_{U,\lambda}(\theta)$ is twice continuously differentiable; (iii) $\sqrt{n}\,\nabla_\theta\widehat{R}_{U,\lambda}(f_{\theta^*}) \overset{p}{\to} N(\mathbf{0}, 4\mathbb{E}_U[\mathbb{E}^2_{U'}[h_{\theta^*}(U, U')]])$; (iv) there is $H(\theta)$ that is continuous at θ^* and $\sup_{\theta\in\Theta}\|\nabla^2_\theta\widehat{R}_{U,\lambda}(f_\theta) - H(\theta)\|_F \overset{p}{\to} 0$; (v) $H(\theta^*)$ is nonsingular.

The proof of (i) is very similar to Theorem A3 except that we consider finite dimensional parameter space instead of functional space. For a neat proof, we would like to omit the detailed proof here. We can first show the uniform consistency $\sup_{\theta\in\Theta}|\widehat{R}_{U,\lambda}(f_\theta) - R_k(f_\theta)| \overset{p}{\to} 0$ and $R_k(f_\theta)$ is continuous about θ similarly to Lemma A2. Here, the proof is based on the conditions $\mathbb{E}[|Y|] < \infty$, Θ is compact, $\mathbb{E}[|f_\theta(X)|] < \infty$, and f_θ is twice continuously differentiable about θ, and Assumption 1 holds. Then, $\widehat{\theta}_U \overset{p}{\to} \theta^*$ similarly to Theorem A3, because of the extra condition $R_k(f_\theta)$ is uniquely minimized at θ^*.

Furthermore, from the conditions that Θ is compact, f_θ is twice continuously differentiable about θ, $\mathbb{E}[|f_\theta(X)|] < \infty$, $\mathbb{E}[\|\nabla_\theta f_\theta(X)\|_2] < \infty$, $\mathbb{E}[\|\nabla^2_\theta f_\theta(X)\|_F] < \infty$, $\mathbb{E}[|Y|] < \infty$, and $k(z, z')$ is bounded as implied by Assumption 1, we can obtain (ii) $\widehat{R}_{V,\lambda}(\theta)$ is twice continuously differentiable about θ. Given $H = \mathbb{E}[\nabla^2_\theta h_{\theta^*}(U, U')] = \nabla^2_\theta R_k(\theta)$ is nonsingular and $R_k(f_\theta)$ is uniquely minimized at θ^*, we can obtain that the Hessian matrix H is positive definite,

$$H = 2\mathbb{E}_{(XYZ),(X'Y'Z')}\left[\left(\nabla_\theta f_{\theta^*}(X)\nabla_\theta f^\top_{\theta^*}(X') - \left(Y - f_{\theta^*}(X)\right)\nabla^2_\theta f_{\theta^*}(X')\right)k(Z, Z')\right] \succ \mathbf{0}.$$

If for all $z \in \mathcal{Z}$, there is $\mathbb{E}_X[\nabla_\theta f_{\theta^*}(X) \mid z]p(z) = \mathbf{0}$ and $\mathbb{E}_{XY}[Y - f_{\theta^*}(X) \mid z]p(z) = 0$, then we can see that the above function $H = \mathbf{0}$, which contradicts $H \succ \mathbf{0}$. Therefore, there must exist z s.t. $\mathbb{E}_X[\nabla_\theta f_{\theta^*}(X) \mid z]p(z) \neq \mathbf{0}$ or $\mathbb{E}_{XY}[Y - f_{\theta^*}(X) \mid z]p(z) \neq 0$. Then, it follows by Lemma A3 that (iii) $\sqrt{n}\,\nabla_\theta\widehat{R}_{U,\lambda}(f_{\theta^*}) \overset{p}{\to} N(\mathbf{0}, 4\mathbb{E}_U[\mathbb{E}^2_{U'}[h_{\theta^*}(U, U')]])$.

Finally by Lemma A4, we know that $H(\theta) = \mathbb{E}[\nabla^2_\theta h_\theta(U, U')]$ and $H(\theta^*) = H$, so (iv) and (v) are satisfied. Now, conditions of [53, Theorem 3.1] are all satisfied, so we assert the conclusion. $\qquad\square$

A.12 Proof of Theorem 1

We restate the notations

$$\widehat{R}_{V,\lambda}(f_\theta) \coloneqq \underbrace{\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n h_\theta(u_i, u_j)}_{\widehat{R}_V(f_\theta)} + \lambda\Omega(\theta)$$

$$h_\theta(u_i, u_j) \coloneqq (y_i - f_\theta(x_i))k(z_i, z_j)(y_j - f_\theta(x_j)).$$

Lemma A5. *Suppose that conditions of Lemma* A3 *hold. Then* $\sqrt{n}\,\nabla_\theta\widehat{R}_{V,\lambda}(f_{\theta^*}) \overset{p}{\to} N(\mathbf{0}, 4\mathbb{E}_U[\mathbb{E}^2_{U'}[\nabla_\theta h_{\theta^*}(U, U')]])$.

Proof. As $\mathbb{E}[\|\nabla_\theta h_{\theta^*}(U, U')\|^2_2] < \infty$, $\sqrt{n}\,\nabla_\theta\widehat{R}_V(f_{\theta^*})$ has the same limit distribution as that of $\sqrt{n}\,\nabla_\theta\widehat{R}_U(f_{\theta^*})$ by [28, Section 5.7.3]. Furthermore, by $\sqrt{n}\lambda \overset{p}{\to} 0$ and $\nabla_\theta\Omega(\theta^*) < \infty$ from that $\Omega(\theta)$ is first continuously differentiable, we assert the conclusion by Slutsky's theorem

$$\sqrt{n}\,\nabla_\theta\widehat{R}_{V,\lambda}(f_{\theta^*}) = \sqrt{n}\,\nabla_\theta\widehat{R}_V + \sqrt{n}\lambda\nabla_\theta\Omega(\theta^*) \overset{p}{\to} N\left(\mathbf{0}, 4\mathbb{E}_U\left[\mathbb{E}^2_{U'}\left[\nabla_\theta h_{\theta^*}(U, U')\right]\right]\right). \qquad\square$$

Lemma A6. *Suppose that* f_θ *and* $\Omega(\theta)$ *are twice continuously differentiable about* θ, Θ *is compact,* $\mathbb{E}[\sup_{\theta\in\Theta}|f_\theta(X)|^2] < \infty$, $\mathbb{E}[\sup_{\theta\in\Theta}\|\nabla_\theta f_\theta(X)\|^2_2] < \infty$, $\mathbb{E}[\sup_{\theta\in\Theta}\|\nabla^2_\theta f_\theta(X)\|^2_F] < \infty$, $\mathbb{E}[|Y|^2] < \infty$, $\lambda \overset{p}{\to} 0$, *and Assumption* 1 *holds. Then,* $\mathbb{E}[\nabla^2_\theta h_\theta(U, U')]$ *is continuous about* θ *and* $\sup_{\theta\in\Theta}\|\nabla^2_\theta\widehat{R}_V(f_\theta) - \mathbb{E}[\nabla^2_\theta h_\theta(U, U')]\|_F \overset{p}{\to} 0$.

Proof. We apply [53, Lemma 8.5] for this proof and need to show (i) $\nabla_\theta^2 h_\theta(u, u')$ is continuous about each $\theta \in \Theta$ with probability one, and (ii) $\mathbb{E}[\sup_{\theta\in\Theta}\|\nabla_\theta^2 h_\theta(U, U')\|_F] < \infty$ and $\mathbb{E}[\sup_{\theta\in\Theta}\|\nabla_\theta^2 h_\theta(U, U)\|_F] < \infty$.

We first see that $\|\nabla_\theta^2 h_\theta(u, u')\|_F$ and $\|\nabla_\theta^2 h_\theta(u, u)\|_F$ are bounded for finite u, u' because f_θ is twice continuously differentiable about θ. It follows that (i) $\nabla_\theta^2 h_\theta(u, u')$ is continuous about θ with probability one. We then derive upper bounds for $\mathbb{E}[\sup_{\theta\in\Theta}\|\nabla_\theta^2 h_\theta(U, U')\|_F]$ and $\mathbb{E}[\sup_{\theta\in\Theta}\|\nabla_\theta^2 h_\theta(U, U)\|_F]$ so as to show their boundedness,

$$
\mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta^2 h_\theta(U, U')\|_F\right]
$$

$$
\leq 2\mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta f_\theta(X)\nabla_\theta f_\theta^\top(X')\|_F + (|Y| + |f_\theta(X)|)\|\nabla_\theta^2 f_\theta(X')\|_F\right]\sup_z k(z, z)
$$

$$
\leq 2\mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta f_\theta(X)\|_2\right]^2 + 2\mathbb{E}\left[|Y| + \sup_{\theta\in\Theta}|f_\theta(X)|\right]\mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta^2 f_\theta(X')\|_F\right]\sup_z k(z, z) < \infty,
$$

and

$$
\mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta^2 h_\theta(U, U)\|_F\right]
$$

$$
\leq 2\mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta f_\theta(X)\nabla_\theta f_\theta^\top(X)\|_F + (|Y| + |f_\theta(X)|)\|\nabla_\theta^2 f_\theta(X)\|_F\right]\sup_z k(z, z)
$$

$$
\leq 2\mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta f_\theta(X)\|_2^2\right] + 2\mathbb{E}\left[(|Y| + \sup_{\theta\in\Theta}|f_\theta(X)|)^2\right]\mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta^2 f_\theta(X')\|_F^2\right]\sup_z k(z, z)
$$

$$
\leq 2\mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta f_\theta(X)\|_2^2\right] + \mathbb{E}\left[(|Y| + \sup_{\theta\in\Theta}|f_\theta(X)|)^2\right] + \mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta^2 f_\theta(X')\|_F^2\right]\sup_z k(z, z)
$$

$$
\leq 2\mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta f_\theta(X)\|_2^2\right] + 2\mathbb{E}\left[|Y|^2 + \sup_{\theta\in\Theta}|f_\theta(X)|^2\right] + \mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta^2 f_\theta(X')\|_F^2\right]\sup_z k(z, z) < \infty.
$$

Thus, we assert the conclusion by [53, Lemma 8.5]. $\square$

Proof of Theorem 1. The proof is the same as that of Theorem A4 except that $\widehat{R}_U$ is replaced by $\widehat{R}_V$. $\square$

A.13 Asymptotic normality in the infinite-dimension case

We first state the asymptotic normality theorem for $\hat{f}_U$ and its proof. Afterward, we provide the proof of Theorem 2 whose proof is a slightly modified version of that of $\hat{f}_U$.

Theorem A5. *Suppose Assumption* 1 *holds, l is a bounded kernel, k is a uniformly bounded function, and $\lambda \geq \lambda_0$ holds. Also, suppose that X, Z, and Y are compact spaces, and there exists $s \in (0, 2)$ and a constant $C_H > 0$ such that $\log \mathcal{N}(\varepsilon, \mathcal{H}_l, \|\cdot\|_{L^\infty}) \leq C_H \varepsilon^{-s}$ for any $\varepsilon \in (0, 1)$. If $\lambda - \lambda_0 = o(n^{-1/2})$ holds, then there exists a Gaussian process $\mathbb{G}_P^*$ such that*

$$
\sqrt{n}(\hat{f}_U - f_{\lambda_0}^*) \rightsquigarrow \mathbb{G}_P^* \quad \text{in } \mathcal{H}_l.
$$

An exact covariance of $\mathbb{G}_P^*$ is described in the proof. The proof is based on the uniform convergence of U-processes on the function space [76] and the functional delta method using the asymptotic expansion of the loss function [29]. This asymptotic normality allows us to perform statistical inference, such as tests, even in the nonparametric case.

To prove the theorem, we provide some notation. Let P be a probability measure that generates $u = (x, y, z)$ and $\mathcal{W} = \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Also, we define a function $h_f(u, u') = (y - f(x))(y' - f(x'))k(z, z')$. Let $\mathbb{H} := \{h_f : \mathcal{W} \times \mathcal{W} \to \mathbb{R} | f \in \mathcal{H}_l\}$. For preparation, we define $P^1 h_f : \mathcal{W} \to \mathbb{R}$ as $P^1 h_f(\cdot) = (\int h_f(u, \cdot) + h_f(\cdot, u) dP(u))/2$ for $h_f \in \mathbb{H}$. For a signed measure Q on $\mathcal{W}$, we define a measure $Q^2 := Q \otimes Q$ on $\mathcal{W} \times \mathcal{W}$. Then, we can rewrite the U-statistic risk as follows:

$$\widehat{R}_U(f) = \frac{(n-2)!}{n!} \sum_{i=1}^{n} \sum_{j \neq i}^{n} h_f(u_i, u_j) =: U_n^2 h_f,$$

where U_n is an empirical measure for the U-statistics. Similarly, we can rewrite the V-statistic risk as follows:

$$\widehat{R}_V(f) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h_f(u_i, u_j) =: P_n^2 h_f,$$

where P_n is an empirical measure of u.

Further, we define a functional associated with measure. We consider functional spaces $\mathcal{G}_1 := \{g : \mathcal{W} \times \mathcal{W} \to [0, 1] | \text{a convex set } \omega \text{ s.t. } g = \mathbf{1}\{\cdot \leq \omega\}\}$ and $\mathcal{G}_2 := \{g : \mathcal{W} \times \mathcal{W} \to [0, 1] | \exists f, f' \in \mathcal{H}_l, g(u, u') = h_f(u, u')(f'(x) + f(x'))\}$. Note that $\mathcal{G}_1$ contains a functional which corresponds to the U_n^2. Then, we consider a set of functionals

$$B_S := \left\{ F : \mathcal{G}_1 \cup \mathcal{G}_2 \to \mathbb{R} | \exists \text{nonzero finite } Q^2 \text{ s.t. } F(g) = \int g \mathrm{d}Q^2, \ \forall g \in \mathcal{G}_1 \cup \mathcal{G}_2 \right\},$$

and also let B_0 be the closed linear span of B_S. For a functional $F \in B_S$, let $\iota(F)$ be a measure that satisfies

$$F(g) = \int g \mathrm{d}\iota(F).$$

Uniform central limit theorem: We first achieve the uniform convergence. Note that a measure P^2 satisfies $P^2 h_f := \mathbb{E}_{(U, U')}[h_f(U, U')]$. The convergence theorem for U-processes is as follows.

Theorem A6. (Theorem 4.4 in ref. [76]) *Suppose $\mathbb{H}$ is a set of uniformly bounded class and symmetric functions, such that $\{P^1 h_f | h_f \in \mathbb{H}\}$ is a Donsker class and*

$$\lim_{n \to \infty} \mathbb{E}\left[n^{-1/2} \log \mathcal{N}\left(n^{-1/2}\varepsilon, \mathbb{H}, \|\cdot\|_{L^1(U_n^2)} \right) \right] = 0$$

holds, for all $\varepsilon > 0$. Then, we obtain

$$\sqrt{n}(U_n^2 - P^2) \rightsquigarrow 2\mathbb{G}_{P^1}, \quad in \ \ell^\infty(\mathbb{H}).$$

Here, $\mathbb{G}_{P^1}$ denotes a Brownian bridge, which is a GP on $\mathbb{H}$ with zero mean and a covariance

$$\mathbb{E}_U[P^1 h_f(U) P^1 h_{f'}(U)] - \mathbb{E}_U[P^1 h_f(U)]\mathbb{E}_U[P^1 h_{f'}(U)],$$

with $h_f, h_{f'} \in \mathbb{H}$.

To apply the theorem, we have to show that $\mathbb{H}$ satisfies the condition in Theorem A6. We first provide the following bound.

Lemma A7. *For any $f, f' \in \mathcal{H}_l$ such that $\|f\|_{L^\infty} \vee \|f'\|_{L^\infty} \leq B$ holds, $y, y' \in [-B, B]$ and $u, u' \in \mathcal{W}$, we have*

$$|h_f(u, u') - h_{f'}(u, u')| \leq 4B|k(z, z')|\|f - f'\|_{L^\infty}.$$

Proof of Lemma A7. We simply obtain the following:

$$
\begin{aligned}
|h_f(u, u') - h_{f'}(u, u')| &= |(y - f(x))k(z, z')(y' - f(x')) - (y - f'(x))k(z, z')(y' - f'(x'))| \\
&= |k(z, z')||(y - f(x))(y' - f(x')) - (y - f'(x))(y' - f'(x'))| \\
&= |k(z, z')||y'(f'(x) - f(x)) + y(f'(x') - f(x')) + f(x)f(x') - f'(x)f'(x')| \\
&= |k(z, z')||y'(f'(x) - f(x)) + y(f'(x') - f(x')) + f(x')(f(x) - f'(x)) \\
&\quad - f'(x)(f(x') - f'(x'))| \\
&\leq |k(z, z')|\{|y'||f'(x) - f(x)| + |y||f'(x') - f(x')| + |f(x')||f(x) - f'(x)| - |f'(x)||f(x') \\
&\quad - f'(x')|\} \\
&\leq 4B|k(z, z')|\|f - f'\|_{L^\infty},
\end{aligned}
$$

as required. □

Lemma A8. *Suppose the assumptions of Theorem 2 hold. Then, the followings hold*:
(1) $\{P^1 h_f | h_f \in \mathbb{H}\}$ *is a Donsker class.*
(2) *For any $\varepsilon > 0$, the following holds*:

$$
\lim_{n \to \infty} E\left[n^{-1/2} \log \mathcal{N}\left(n^{-1/2}\varepsilon, \mathbb{H}, \|\cdot\|_{L^1(U_n^2)} \right) \right] = 0.
$$

Proof of Lemma A8. For preparation, fix $\varepsilon > 0$ and set $N = \mathcal{N}(\varepsilon, \mathcal{H}_l, \|\cdot\|_{L^\infty})$. Also, let Q be an arbitrary finite discrete measure. Then, by the definition of a bracketing number, there exist N functions $\{f_i \in \mathcal{H}_l\}_{i=1}^N$ such that for any $f \in \mathcal{H}_l$ there exists $i \in \{1, 2, \ldots, N\}$ such as $\|f - f_i\|_{L^\infty} \leq \varepsilon$.

For the first condition, as shown in equation (2.1.7) in ref. [55], it is sufficient to show that

$$
\sup_Q \log \mathcal{N}\left(\varepsilon, \{P^1 h_f | h_f \in \mathbb{H}\}, \|\cdot\|_{L^2(Q)} \right) \leq c\varepsilon^{-2\delta} (\varepsilon \to 0),
$$

for arbitrary $\delta \in (0, 1)$. Here, $c > 0$ is some constant, and Q is taken from all possible finite discrete measure. To this end, it is sufficient to show that $\log \mathcal{N}(\varepsilon, \{P^1 h_f | h_f \in \mathbb{H}\}, \|\cdot\|_{L^2(Q)}) \leq c' \log \mathcal{N}(\varepsilon, \mathcal{H}_l, \|\cdot\|_{L^\infty})$ with a constant $c' > 0$. Fix $P^1 h_f \in \{P^1 h_f | h_f \in \mathbb{H}\}$ arbitrary, and set f_i which satisfies $\|f - f_i\|_{L^2(Q)} \leq \varepsilon$. Then, we have

$$
\begin{aligned}
\|P^1 h_f - P^1 h_{f_i}\|_{L^2(Q)}^2 &= \int \left\{ \int (h_f(u, u') + h_f(u', u))/2 - \left(h_{f_i}(u, u') + h_{f_i}(u', u) \right)/2 \mathrm{d}P(u') \right\}^2 dQ(u) \\
&= \int \left(\int h_f(u, u') - h_{f_i}(u, u') \mathrm{d}P(u') \right)^2 dQ(u) \\
&\leq C \int \left(\int |k(z, z')| \mathrm{d}P(u') \right)^2 dQ(u) \|f - f_i\|_{L^\infty}^2 \\
&\leq C' \|f - f_i\|_{L^\infty}^2 \\
&\leq C'\varepsilon^2,
\end{aligned}
$$

with constants $C, C' > 0$. The first inequality follows Lemma A7 with the bounded property of f, f', and $\mathcal{Y}$. The second inequality follows the bounded condition of k in Theorem A5. Hence, the entropy condition shows the first statement.

For the second condition, we have the similar strategy. For any $h_f \in \mathbb{H}$, we consider $i \in \{1, 2, \ldots, N\}$ such that $\|f - f_i\|_{L^\infty} \leq \varepsilon$. Then, we measure the following value:

$$
\|h_f - h_{f_i}\|_{L^1(U_n^2)} = \int |h_f(u, u') - h_{f_i}(u, u')| \mathrm{d}U_n^2(u, u') \leq C'' \|f - f_i\|_{L^\infty} \leq C''\varepsilon,
$$

with a constant $C'' > 0$. Hence, we have

$$
\begin{aligned}
\mathbb{E}\left[n^{-1/2} \log \mathcal{N}\left(n^{-1/2}\varepsilon, \mathbb{H}, \|\cdot\|_{L^1(U_n^2)} \right) \right] &\leq n^{-1/2} \log \mathcal{N}\left(n^{-1/2}\varepsilon, \mathcal{H}_l, \|\cdot\|_{L^\infty} \right) \\
&\leq C n^{-1/2} \left(\frac{n^{1/2}}{\varepsilon} \right)^s = C n^{(s-1)/2} \to 0, (n \to \infty),
\end{aligned}
$$

since $s \in (0, 1)$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

From Theorem A6 and Lemma A8, we rewrite the central limit theorem utilizing terms of functionals. Note that $\iota^{-1}(U_n^2), \iota^{-1}(P^2) \in B_S$ holds. Then, we can obtain

$$\sqrt{n}(\iota^{-1}(U_n^2) - \iota^{-1}(P^2)) \rightsquigarrow 2\mathbb{G}_{P^1} \quad \text{in } \ell^\infty(\mathbb{H}).$$

Learning map and functional delta method: We consider a learning map $S : B_S \to \mathcal{H}_l$. For a functional $F \in B_S$, we define

$$S_\lambda(F) := \underset{f \in \mathcal{H}_l}{\operatorname{argmin}} \, \iota(F)h_f + \lambda\|f\|_{\mathcal{H}_l}^2.$$

Obviously, we have

$$\hat{f} = S_\lambda(\iota^{-1}(U_n^2)), \text{ and} f_{\lambda_0}^* = S_{\lambda_0}(\iota^{-1}(P^2)).$$

We consider a derivative of S_λ in the sense of the Gateau differentiation by the following steps.

First, we define a partial derivative of the map $R_{Q^2}(f)$. We investigate the optimality of the minimizer of

$$R_{Q^2, \lambda}(f) := \int h_f(u, u') \mathrm{d}Q^2(u, u') + \lambda\|f\|_{\mathcal{H}_l}^2.$$

To this end, we consider the following derivative $\nabla R_{Q^2, \lambda}[f] : \mathcal{H}_l \to \mathcal{H}_l$ with a direction f as follows:

$$\nabla R_{Q^2, \lambda}[f](f') := 2\lambda f + \int \partial_{f,1} h_f(u, u')f'(x) + \partial_{f,2} h_f(u, u')f'(x') \mathrm{d}Q^2(u, u').$$

Here, $\partial_{f,1} h_f$ is a partial derivative of h_f in terms of the input $f(x)$ as follows:

$$\partial_{f,1} h_f(u, u') = \partial_{t|t=f(x)}(y - t)k(z, z')(y' - f(x')) = -(y' - f(x'))k(z, z'),$$

and $\partial_{f,2} h_f$ follows it respectively. The following lemma validates the derivative.

Lemma A9. *If the assumptions in Theorem 2 hold, then $\nabla R_{Q^2, \lambda}[f]$ is a Gateau-derivative of $R_{Q^2, \lambda}$ with the direction $f \in \mathcal{H}_l$.*

Proof of Lemma A9. We consider a sequence of functions $h_n \in \mathcal{H}_l$ for $n \in \mathbb{N}$, such that $h_n(x) \neq 0, \forall x \in \mathcal{X}$ and $\|h_n\|_{L^\infty} \to 0$ as $n \to \infty$. Then, for $f \in \mathcal{H}_l$, a simple calculation yields

$$\left| \frac{R_{Q^2, \lambda}(f + h_n) - R_{Q^2, \lambda}(f) - \nabla R_{Q^2, \lambda}[f](h_n)}{\|h_n\|_{L^\infty}} \right|$$

$$\leq \int \|h_n\|_{L^\infty}^{-1} |k(z, z')((y - f(x))h_n(x) + (y' - f(x'))h_n(x') + h_n(x)h_n(x')) - \nabla R_{Q^2, \lambda}[f](h_n)| \mathrm{d}Q^2(u, u')$$

$$\leq \int \|h_n\|_{L^\infty}^{-1} |k(z, z')h_n(x)h_n(x')| \mathrm{d}Q^2(u, u')$$

$$\leq \int \|h_n\|_{L^\infty}^{-1} |k(z, z')\|h_n\|_{L^\infty}^2| \mathrm{d}Q^2(u, u')$$

$$\leq \|h_n\|_{L^\infty} \int |k(z, z')| \mathrm{d}Q^2(u, u') \to 0, (n \to \infty).$$

The convergence follows the definition of h_n and the absolute integrability of k, which follows the bounded property of k and compactness of $\mathcal{Z}$. Then, we obtain the statement. $\qquad\qquad$ $\square$

Here, we consider its RKHS-type formulation of $\nabla R_{Q^2, \lambda}$, which is convenient to describe a minimizer. Let $\Phi_l : \mathcal{X} \to \mathcal{H}_l$ be the feature map associated with the RKHS $\mathcal{H}_l$, such that $\langle \Phi_l[x], f \rangle_{\mathcal{H}_l} = f(x)$ for any $x \in \mathcal{X}$ and $f \in \mathcal{H}_l$. Let $\nabla \tilde{R}_{Q^2, \lambda} : \mathcal{H}_l \to \mathcal{H}_l$ be an operator such that

$$\nabla \tilde{R}_{Q^2,\lambda}(f) := 2\lambda f + \int \partial_{f,1} h_f(u, u') \Phi_l[x](\cdot) + \partial_{f,2} h_f(u, u') \Phi_l[x'](\cdot) \mathrm{d}Q^2(u, u').$$

Obviously, $\nabla R_{Q^2,\lambda}[f](\cdot) = \langle \nabla \tilde{R}_{Q^2,\lambda}(f), \cdot \rangle_{\mathcal{H}_l}$. Now, we can describe the first-order condition of the minimizer of the risk. Namely, we can state that

$$\hat{f} = \underset{f \in \mathcal{H}_l}{\arg\min} R_{Q^2,\lambda}(f) \Leftrightarrow \nabla \tilde{R}_{Q^2,\lambda}(\hat{f}) = 0.$$

This equivalence follows Theorem 7.4.1 and Lemma 8.7.1 in ref. [77].

Next, we apply the implicit function theorem to obtain an explicit formula of the derivative of S. To this end, we consider a second-order derivative $\nabla^2 \tilde{R}_{Q^2,\lambda} : \mathcal{H}_l \to \mathcal{H}_l$ as follows:

$$\nabla^2 \tilde{R}_{Q^2,\lambda}(f) := 2\lambda f + \int k(z, z')(f(x)\Phi_l[x](\cdot) + f(x')\Phi_l[x'](\cdot)) \mathrm{d}Q^2(u, u'),$$

which follows (b) in Lemma A.2 in ref. [29]. Its basic properties are provided in the following result.

Lemma A10. *If Assumption* 1 *and the assumptions in Theorem* 2 *hold, then* $\nabla^2 \tilde{R}_{Q^2,\lambda}$ *is a continuous linear operator and it is invertible.*

Proof of Lemma A10. By (b) in Lemma A.2 in ref. [29], $\nabla^2 \tilde{R}_{Q^2,\lambda}$ is a continuous linear operator. In the following, we define $A : \mathcal{H}_l \to \mathcal{H}_l$ as $A(f) = \int k(z, z') f(x) \Phi_l[x](\cdot) + f(x')\Phi_l[x'](\cdot) \mathrm{d}Q^2(u, u')$. To show that $\nabla^2 \tilde{R}_{Q^2,\lambda}$ is invertible, it is sufficient to show that (i) $\nabla^2 \tilde{R}_{Q^2,\lambda}$ is injective, and (ii) A is a compact operator.

For the injectivity, we fix nonzero $f \in \mathcal{H}_l$ and obtain

$$
\begin{aligned}
\|\nabla^2 \tilde{R}_{Q^2,\lambda}(f)\|_{\mathcal{H}_l}^2 &= \langle 2\lambda f + A(f), 2\lambda f + A(f) \rangle_{\mathcal{H}_l} \\
&= 4\lambda^2 \|f\|_{\mathcal{H}_l}^2 + 4\lambda \langle f, A(f) \rangle_{\mathcal{H}_l} + \|A(f)\|_{\mathcal{H}_l}^2 \\
&> 4\lambda \langle f, A(f) \rangle_{\mathcal{H}_l} \\
&= 4\lambda \left\langle f, \int k(z, z') f(x) \Phi[x] \mathrm{d}Q^2(u, u') \right\rangle_{\mathcal{H}_l} + 4\lambda \left\langle f, \int k(z, z') f(x') \Phi[x'] \mathrm{d}Q^2(u, u') \right\rangle_{\mathcal{H}_l} \\
&= 4\lambda \int k(z, z') f(x)^2 \mathrm{d}Q^2(u, u') + 4\lambda \int k(z, z') f(x')^2 \mathrm{d}Q^2(u, u') \\
&\geq 0.
\end{aligned}
$$

The last equality follows the property of Φ_l, and the last inequality follows the ISPD property in Assumption 1.

For the compactness, we follow Lemma A.5 in ref. [29] and obtain that operators $(f \mapsto \int k(z, z') f(x) \Phi_l[x](\cdot) \mathrm{d}Q^2(u, u'))$ and $(f \mapsto \int k(z, z') f(x') \Phi_l[x'](\cdot) \mathrm{d}Q^2(u, u'))$ are compact. □

We define the Gateau derivative of S. For a functional $F' \in \ell^\infty(\mathcal{G}_1 \cup \mathcal{G}_2)$, we define the following function:

$$\nabla S_{Q^2,\lambda}(F') := -\nabla^2 \tilde{R}_{Q^2,\lambda}^{-1} \left(\int \partial_{f,1} h_{f_{Q^2}}(u, u') \Phi_l[x](\cdot) + \partial_{f,2} h_{f_{Q^2}}(u, u') \Phi_l[x'](\cdot) \mathrm{d}\iota(F')(u, u') \right),$$

where $f_{Q^2} = S_\lambda(\iota^{-1}(Q^2))$ and Q^2 is a signed measure on $\mathcal{W} \times \mathcal{W}$. Then, we provide the following derivative theorem.

Proposition A1. *Suppose the assumptions in Theorem* 2 *hold. For* $F \in B_S$, $F' \in \ell^\infty(\mathcal{G}_1 \cup \mathcal{G}_2)$, *and* $s \in \mathbb{R}$ *such that* $F + sF' \in B_S$, $\nabla S_{\iota(F),\lambda}(F')$ *is a Gateau-derivative of* S_λ, *namely,*

$$\lim_{s \to 0} \left\| \frac{S_\lambda(F + sF') - S_\lambda(F)}{s} - \nabla S_{\iota(F),\lambda}(F') \right\|_{\mathcal{H}_l} = 0.$$

Proof of Proposition A1. This proof has the following two steps, (i) define a proxy operator Γ and then (ii) prove the statement by the implicit function theorem.

    (i) Define Γ: Note that $\iota(F')$ exists since $F + sF' \in B_S$ implies $F' \in B_0$. We define the following operator $\Gamma(s, f, \lambda) : \mathcal{H}_l \to \mathcal{H}_l$ for $f \in \mathcal{H}_l$:

$$
\begin{aligned}
\Gamma(s, f, \lambda) &:= \nabla \tilde{R}_{\iota(F) + s\iota(F'), \lambda} \\
&= 2\lambda f + \int \partial_{f,1} h_f(u, u') \Phi_l[x](\cdot) + \partial_{f,2} h_f(u, u') \Phi_l[x'](\cdot) \mathrm{d}\iota(F)(u, u') + s \int \partial_{f,1} h_f(u, u') \Phi_l[x](\cdot) \\
&\quad + \partial_{f,2} h_f(u, u') \Phi_l[x'](\cdot) \mathrm{d}\iota(F')(u, u').
\end{aligned}
$$

For simple derivatives, Lemma A.2 in ref. [29] provides

$$
\nabla_s \Gamma(s, f, \lambda) = \int \partial_{f,1} h_f(u, u') \Phi_l[x](\cdot) + \partial_{f,2} h_f(u, u') \Phi_l[x'](\cdot) \mathrm{d}\iota(F')(u, u')
$$

and

$$
\nabla_f \Gamma(s, f, \lambda) = \nabla^2 \tilde{R}_{\iota(F) + s\iota(F'), \lambda}.
$$

    (ii) Apply the implicit function theorem: By its definition and the optimal conditions, we have

$$
\Gamma(s, f, \lambda) = 0 \Leftrightarrow f = S_\lambda(\iota(F) + s\iota(F')).
$$

Also, we obtain

$$
\nabla_f \Gamma(0, S_\lambda(F), \lambda) = \nabla^2 \tilde{R}_{\iota(F), \lambda}.
$$

Then, for each $\lambda > 0$, by the implicit function theorem, there exists a smooth map $\varphi_\lambda : \mathbb{R} \to \mathcal{H}_l$ such that

$$
\Gamma(s, \varphi_\lambda(s), \lambda) = 0, \quad \forall s,
$$

and it satisfies

$$
\nabla_s \varphi_\lambda(0) = -(\nabla_f \Gamma(0, \varphi_\lambda(0), \lambda))^{-1}(\nabla_s \Gamma(0, \varphi_\lambda(0), \lambda)) = \nabla S_{\iota(F), \lambda}(F').
$$

Also, we have $\varphi_\lambda(s) = S(Q^2 + s\mu^2)$. Then, we have

$$
\lim_{s \to 0} \left\| \frac{S_\lambda(F + sF') - S_\lambda(F')}{s} - \nabla S_{\iota(F), \lambda}(F') \right\|_{\mathcal{H}_l} = \lim_{s \to 0} \left\| \frac{\varphi_\lambda(s) - \varphi_\lambda(0)}{s} - \nabla_s \varphi_\lambda(0) \right\|_{\mathcal{H}_l} = 0.
$$

Then, we obtain the statement.   □

    Now, we are ready to prove Theorems 2 and A5.

Proof of Theorem A5. As a preparation, we mention that S_λ is differentiable in the Hadamard sense, which is Gateau differentiable by Proposition A1. Lemmas A.7 and A.8 in ref. [29] show that $\nabla S_{\iota(F), \lambda, \iota(G)}$ is Hadamard-differentiable for any λ, G and F.

    Then, we apply the functional delta method. As shown in Theorem A6 and Lemma A8, we have

$$
\sqrt{n}(\iota^{-1}(U_n^2) - \iota^{-1}(P^2)) \rightsquigarrow 2\mathbb{G}_{P^1}.
$$

Hence, we obtain

$$
\sqrt{n}((\lambda_0 / \lambda)\iota^{-1}(U_n^2) - \iota^{-1}(P^2)) = \frac{\lambda_0}{\lambda}\sqrt{n}(\iota^{-1}(U_n^2) - \iota^{-1}(P^2)) + \frac{\sqrt{n}(\lambda - \lambda_0)}{\lambda} \rightsquigarrow 2\mathbb{G}_{P^1},
$$

since $\lambda - \lambda_0 = o(n^{-1/2})$. By utilizing the result, we can obtain

$$
\sqrt{n}(\hat{f} - f_{\lambda_0}^*) = \sqrt{n}\left(S_{\lambda_0}((\lambda_0 / \lambda)\iota^{-1}(U_n^2)) - S_{\lambda_0}(\iota^{-1}(P^2))\right) + o_P(1) \rightsquigarrow \nabla S_{P^2, \lambda_0}\left(2\mathbb{G}_{P^1}\right),
$$

in $\ell^\infty(\mathbb{H})$. The convergence follows the functional delta method.   □

Proof of Theorem 2. This proof is completed by substituting the Central Limit Theorem part (Theorem A6) of the proof of Theorem A5. From Section 3 in ref. [78], the V- and U-processes have the same limit distribution asymptotically, so the same result holds. □

A.14 Proof of Proposition 5

Proof of Proposition 5. For short, we write $\Lambda = \Lambda_{\min}$ and $\Lambda' = \Lambda_{\max}$.

We start by investigating H and $\check{H}$. By using the independent property between $U = (X, Y, Z)$ and $U' = (X', Y', Z')$, we have

$$
\begin{aligned}
H &= \mathbb{E}[\nabla_\theta^2(Y - f_\theta(X))(Y' - f_\theta(X'))k(Z, Z')] \\
&= \mathbb{E}\left[\nabla_\theta^2 f_{\theta^*}(X)\varepsilon' k(Z, Z')\right] + \mathbb{E}\left[\nabla_\theta^2 f_{\theta^*}(X')\varepsilon k(Z, Z')\right] + 2\mathbb{E}\left[\nabla_\theta f_{\theta^*}(X)k(Z, Z')\nabla_\theta f_{\theta^*}(X')\right] \\
&= 2\mathbb{E}_{Z,Z'}\left[\mathbb{E}_X\left[\nabla_\theta f_{\theta^*}(X)|Z\right]k(Z, Z')\nabla_\theta\mathbb{E}_{X'}[f_{\theta^*}(X')|Z']\right] \\
&= \iint q(z)q(z')k(z, z')\mathrm{d}P(z, z') \\
&= \iint q(z)p(z)q(z')p(z')k(z, z')\mathrm{d}z\mathrm{d}z',
\end{aligned}
$$

where $q : \mathcal{Z} \to \mathbb{R}^d, z \mapsto \mathbb{E}_X[\nabla_\theta f_{\theta^*}(X)|z]$ with the parameter dimension d. The third equality follows that $\varepsilon = Y - f_{\theta^*}(X)$ is an independent noise variable, and the last equality follows the independent property between Z and Z'. By the same way, $\check{H}$ is rewritten as follows:

$$
\check{H} = \iint q(z)p(z)q(z')p(z')\mathrm{d}z\mathrm{d}z'.
$$

On the basis of these facts, we obtain the following form

$$
H = \check{H} \circ W,
$$

where W is a $d \times d$ matrix whose all elements belong to $[\Lambda, \Lambda']$, and $\circ$ denotes the Haramard product of matrices. In words, the elements of H are obtained by multiplying some coefficients to their corresponding elements of $\check{H}$, and the coefficients belongs to $[\Lambda, \Lambda']$. By the same calculation, we obtain $\mathbb{E}_U[\mathbb{E}_{U'}^2[h_{\theta^*}(U, U')]] = \mathbb{E}_U[\mathbb{E}_{U'}^2[\check{h}_{\theta^*}(U, U')]]w^2$ with some $w \in [\Lambda, \Lambda']$.

By using the result, we study H^{-1} with the decomposition

$$
H^{-1} = \check{H}^{-1} + (H^{-1} - \check{H}^{-1}).
$$

Let $\mathbb{I}$ be a $d \times d$ matrices whose all elements are 1. We have the following difference

$$
\begin{aligned}
H^{-1} - \check{H}^{-1} &= H^{-1}(\check{H} - H)\check{H}^{-1} \\
&= H^{-1}(\check{H} - W \circ H)\check{H}^{-1} \\
&= H^{-1}((\mathbb{I} - W) \circ \check{H})\check{H}^{-1} \\
&= H^{-1}\check{H}\check{H}^{-1}(\mathbb{I} \circ (\mathbb{I} - W)) \\
&= H^{-1} \circ (\mathbb{I} - W).
\end{aligned}
$$

Finally, we study Σ_V. It is decomposed as follows:

$$\Sigma_V = 4H^{-1}\mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[h_{\theta^*}(U, U')\big]\big]\big)H^{-1}$$

$$= 4(\check{H}^{-1} + (H^{-1} - \check{H}^{-1}))\big(\mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[\check{h}_{\theta^*}(U, U')\big]\big]\big) + \big(\mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[h_{\theta^*}(U, U')\big]\big]\big)$$

$$- \mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[\check{h}_{\theta^*}(U, U')\big]\big]\big)\big)\big)(\check{H}^{-1} + (H^{-1} - \check{H}^{-1}))$$

$$= 4\check{H}^{-1}\mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[\check{h}_{\theta^*}(U, U')\big]\big]\big)\check{H}^{-1} + 4(H^{-1} - \check{H}^{-1})\mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[\check{h}_{\theta^*}(U, U')\big]\big]\big)\check{H}^{-1}$$

$$+ 8\check{H}^{-1}\big(\mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[h_{\theta^*}(U, U')\big]\big]\big) - \mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[\check{h}_{\theta^*}(U, U')\big]\big]\big)\big)\check{H}^{-1}$$

$$+ 8\check{H}^{-1}\big(\mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[h_{\theta^*}(U, U')\big]\big]\big) - \mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[\check{h}_{\theta^*}(U, U')\big]\big]\big)\big)(H^{-1} - \check{H}^{-1})$$

$$+ 4(H^{-1} - \check{H}^{-1})^2\mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[\check{h}_{\theta^*}(U, U')\big]\big]\big) + 4(H^{-1} - \check{H}^{-1})^2\big(\mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[h_{\theta^*}(U, U')\big]\big]\big)$$

$$- \mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[\check{h}_{\theta^*}(U, U')\big]\big]\big)\big)$$

$$= \check{\Sigma}_V + 8(w^2 - 1)\mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[\check{h}_{\theta^*}(U, U')\big]\big]\big)\check{H}^{-2} + 4(2 - w^2)\mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[\check{h}_{\theta^*}(U, U')\big]\big]\big)\check{H}^{-1}(H^{-1} - \check{H}^{-1})$$

$$+ 4(H^{-1} - \check{H}^{-1})^2\mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[h_{\theta^*}(U, U')\big]\big]\big)$$

$$= \check{\Sigma}_V + 8(w^2 - 1)\mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[\check{h}_{\theta^*}(U, U')\big]\big]\big)\check{H}^{-2}$$

$$+ 4(2 - w^2)\mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[\check{h}_{\theta^*}(U, U')\big]\big]\big)\check{H}^{-1}(H^{-1} \circ (\mathbb{I} - W))$$

$$+ 4w^2\mathrm{diag}\big(\mathbb{E}_U\big[\mathbb{E}_{U'}^2\big[\check{h}_{\theta^*}(U, U')\big]\big]\big)(H^{-1} \circ (\mathbb{I} - W))^2.$$

By using the statement, we study the Frobenisu norm $\|\Sigma_V\|_F$. As preparation, we set $c := \mathrm{diag}(\mathbb{E}_U[\mathbb{E}_{U'}^2[\check{h}_{\theta^*}(U, U')]])$ and recall the inequality $\|A \circ B\|_F \le \|A\|_F\|B\|_F$. Also, recall the definition $\delta_\Lambda := \max\{|\Lambda' - 1|, |1 - \Lambda|\}$. We obtain

$$\|\Sigma_V\|_F \le \|\check{\Sigma}_V\|_F + 8(w^2 - 1)c\|\check{H}^{-2}\|_F + 4(2 - w^2)c\|\check{H}^{-1}\|_F^2\|\mathbb{I} - W\|_F + 4w^2c\|\check{H}^{-1}\|_F^2\|\mathbb{I} - W\|_F^2$$

$$\le \|\check{\Sigma}_V\|_F + 8\Lambda'\delta_\Lambda c\|\check{H}^{-2}\|_F + 8c\|\check{H}^{-1}\|_F^2 d\delta_\Lambda + 4(\Lambda')^2c\|\check{H}^{-1}\|_F^2 d^2\delta_\Lambda^2$$

$$= \|\check{\Sigma}_V\|_F + O(\Lambda'\delta_\Lambda + (\Lambda'\delta_\Lambda)^2).$$

Landau's Big O notation in the last equality picks up the terms depends only Λ and Λ'.   □

B Related studies beyond IV regression setting

Beyond the IV regression, there are numerous prior studies that are related to ours, especially in policy evaluation, reinforcement learning, and causal inference. First, the idea of "kernel loss" was proposed in ref. [79] to estimate the value function in reinforcement learning. Similar ideas have been used to estimate the importance ratio of two state or state-action distributions in refs [80,81] and to estimate the average policy effect (APE) and policy learning in refs [82,83]. Second, in the area of causal inference, ref. [84] employed a similar technique to estimate an average treatment effect. Despite the methodological similarity to our work, these works did not consider the IV regression setting and the analytical cross validation error. We will first introduce the objective functions of the aforementioned works for a better understanding of the connection and then highlight the differences, challenges, and novelties of our work.

Although estimating different subjects f, refs [80] and [79] employ similar population objective functions in the following form:

$$\min_f \mathbb{E}_X\mathbb{E}_{X'}[((\mathcal{A}f)(X) - f(X))k(X, X')((\mathcal{A}f)(X') - f(X'))], \tag{A2}$$

where $\mathcal{A}$ is a task-specific operator acting on the subject f and X' is an independent copy of X. More specifically, ref. [80] aims to estimate the importance ratio of state distributions of two policies as $f(X)$. In the task of value function estimation by ref. [79], $f(X)$ is the value function. In both works, X represents the state variable in the context of reinforcement learning. Further, ref. [82] studies the estimation of the APE

similar to that of the average treatment effect, which estimate an average effect of a provided policy for a treatment conditioned on a covariate.

Compared with our population and empirical risks (4) and (5), the kernel loss (A2) has a similar quadratic form. The difference is that the kernel loss has variables involved in both the kernel function and the residual, whereas the kernel function in our objective depends on the instrument Z which does not appear in the residual. Besides, our analytical cross validation error was not studied by these works, and the similar forms of the objectives allow to adapt our result to their approaches.

The difference in losses also simplifies the estimation in the related work and requires us to perform new analyses. Specifically, the consistency of the estimators in the aforementioned related work holds under mild conditions. That is, in the estimation of the value function, minimizing the population objective function to zero guarantees a solution for the Bellman equation, which is consistent to the value function according to the unique solution property of the Bellman equation [79]; for the importance ratio and the APE estimation, the consistency holds under mild conditions of data distributions [80,82]. In contrast, the estimator obtained by minimizing the MMR risk (4) to zero is hardly consistent to the true $f(X)$ under mild conditions because there can be $\hat{f} \neq f$ satisfying $\mathbb{E}[Y - \hat{f}(X) \mid Z] = 0$ almost surely. Therefore, we first introduce the completeness condition in Assumption 2 for the estimation, which is classical in the IV regression literature and is needed to identify $f(X)$ from the CMR. Second, a new theoretical analysis is necessary to clarify the nature of the estimation on $f(X)$, such as consistency and asymptotic normality. We develop a novel theory for this point in Section 5, which has not been studied by these related works.