Maximum Moment Restriction for Instrumental Variable Regression

Rui Zhang §
Australian National University
Canberra, Australia

Masaaki Imaizumi †
University of Tokyo
Tokyo, Japan

Bernhard Schölkopf ‡
MPI for Intelligent Systems
Tübingen, Germany

Krikamol Muandet ‡

Abstract

We propose a simple framework for nonlinear instrumental variable (IV) regression based on a kernelized conditional moment restriction (CMR) known as a maximum moment restriction (MMR). The MMR is formulated by maximizing the interaction between the residual and functions of IVs that belong to a unit ball of reproducing kernel Hilbert space (RKHS). This allows us to tackle the IV regression as an empirical risk minimization where the risk depends on the reproducing kernel on the instrument and can be estimated by a U-statistic or V-statistic. This simplification not only enables us to derive elegant theoretical analyses in both parametric and non-parametric settings, but also results in easy-to-use algorithms with a justified hyper-parameter selection procedure. We demonstrate the advantages of our framework over existing ones using experiments on both synthetic and real-world data.

1 INTRODUCTION

Instrumental variables (IV) have become standard tools for economists, epidemiologists, and social scientists to uncover causal relationships from observational data without an unconfoundedness assumption (Angrist and Pischke, 2008; Klungel et al., 2015). Randomization of treatments or policies has been perceived as the gold standard for such tasks, but 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, the use of IV enables 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 (Card, 1999). In genetic epidemiology, the idea to use genetic variants as IVs, known as Mendelian randomization, has also gained popularity in recent years (Burgess et al., 2017, 2020).

Classical methods for IV regression often rely on a linearity assumption in which a two-stage least squares (2SLS) is the most popular technique (Angrist et al., 1996). The generalized method of moments (GMM) of Hansen (1982), which imposes the orthogonality restrictions, can also be used for IV regression. However, linearity assumption is often violated in real-world applications, motivating several works to extend both 2SLS and GMM to the nonlinear setting. Unfortunately, existing approaches suffer from various drawbacks such as the need to solve an ill-posed integral equation or to estimate a conditional density (Newey and Powell, 2003; Hall and Horowitz, 2005; Blundell et al., 2007; Horowitz, 2011), complicated first-stage estimation (Hartford et al., 2017; Singh et al., 2019), and cumbersome minimax optimization (Lewis and Syrgkanis, 2018; Bennett et al., 2019; Muandet et al., 2020b), which limit the prospects of the methods.

To overcome these drawbacks, we propose a simple framework for nonlinear IV regression based on the maximum moment restriction (MMR) framework of Muandet et al. (2020a), which reformulates the original conditional moment restriction (CMR) by maximizing the interaction between the residual and functions of IVs that belong to a unit ball in a reproducing kernel Hilbert space (RKHS). We call this framework MMR-IV for brevity. Based on this formulation, we can view IV regression as an empirical risk minimization (ERM) problem (Vapnik, 1998) where the empirical risk depends on the reproducing kernel on instruments and can be expressed as a U-statistic or V-statistic. Not only does this formulation allow us to perform elegant theoretical analyses in both parametric and non-parametric settings, but it also results in easy-to-use algorithms with a justified hyper-parameter selection procedure. In particular, we can...
Figure 1: A causal graph depicting an instrumental variable \( Z \) that satisfies an exclusion restriction and unconfoundedness (there may be a confounder \( \varepsilon \) acting on \( X \) and \( Y \), but it is independent of \( Z \)).

Figure 1

even solve the problem in closed-form if the solution is assumed to belong to a RKHS. We demonstrate the effectiveness of the proposed approach through experiments on both synthetic and real data.

2 PRELIMINARIES

2.1 Instrumental Variable Regression

Following standard setting in the literature (Hartford et al., 2017; Lewis and Syrgkanis, 2018; Bennett et al., 2019; Singh et al., 2019; Muandet et al., 2020b), 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 \( f : \mathcal{X} \rightarrow \mathbb{R} \) from a structural equation model (SEM) 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), \( \varepsilon \) 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 when there exist unobserved confounders (i.e., common causes) between \( X \) and \( Y \).

To illustrate the problem, let us consider an example taken from Hartford et al. (2017) which aims at predicting sales of airline ticket price of the 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, etc. This creates a correlation between \( \varepsilon \) and \( X \) in (1) that prevents us from applying the standard regression toolboxes directly on observational data.

Instrumental variable(s). 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., \( \varepsilon \perp \perp Z \). For example, the instrument \( Z \) may be the cost of fuel, which influences sales only via price. Intuitively, the instrument \( Z \) acts as a “natural experiment” that induces variation in \( X \), as illustrated in Figure 1.

Generalized method of moment (GMM). The aforementioned conditions imply that \( \mathbb{E}[\varepsilon|Z] = 0 \) for \( P_Z \)-almost all \( z \). This is a conditional moment restriction (CMR) which we can use to estimate \( f \) (Newey, 1993). For any measurable function \( h \), the CMR implies a continuum of unconditional moment restrictions (Lewis and Syrgkanis, 2018; Bennett et al., 2019):

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

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 conditions should be used as a basis for estimating the function \( f \) (Donald and Newey, 2001; Hall, 2005). In this work, we show that, for the purpose of 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

Throughout this paper, we assume that \( h \) is a real-valued function on \( Z \) which belongs to a RKHS \( \mathcal{H}_k \) endowed with a reproducing kernel \( k : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R} \). \( \mathcal{H}_k \) satisfies two important properties: (i) for all \( z \in \mathcal{Z} \) and \( h \in \mathcal{H}_k \), we have \( k(z, \cdot) \in \mathcal{H}_k \) and (ii) (reproducing property of \( \mathcal{H}_k \)) \( h(z) = \langle h, k(z, \cdot) \rangle_{\mathcal{H}_k} \) 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 any \( z, z' \in \mathcal{Z} \), i.e., an inner product between the feature maps of \( z \) and \( z' \) can be evaluated implicitly through the kernel evaluation.

Every positive definite kernel \( k \) uniquely determines the RKHS for which \( k \) is a reproducing kernel (Aronszajn, 1950). For detailed exposition on kernel methods, see, e.g., Schölkopf and Smola (2002), Berlinet and Thomas-Agnan (2004), and Muandet et al. (2017).

Instead of (2), we define a risk in terms of a maximum moment restriction (MMR) (Muandet et al., 2020a):

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

That is, instead of considering all measurable functions \( h \) as instruments, we only restrict to functions that lie within a unit ball of the RKHS \( \mathcal{H}_k \). The risk is then defined as a maximum value of the moment restriction with respect to this function class. The benefits of our formulation are two-fold. 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

\[
E(|Y - f(X)|^2) = 0. \tag{2}
\]
1 below. Second, this restriction still preserves the consistency of parameter estimated using $R_k(f)$. In other words, nothing is lost in terms of statistical estimation by this simplification (cf. Theorem 1).

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

**Assumption 1.** The kernel $k$ is continuous, bounded (i.e., \( \sup_{z \in Z} \sqrt{k(z, z)} < \infty \)) and satisfies the following result, which is a special case of Muandet et al. (2020a, Thm 3.2), showing that the population risk $R$ has an analytic solution.

\[
\text{It is instructive to note that population risk } R_k(f) \text{ is integrally strictly positive definite (ISPD).}
\]

An important property of such kind of kernels is that $k(z, z') > 0$ for any $z, z' \in Z$. They are also related to the notion of characteristic and universal kernels; see, e.g., Simon-Gabriel and Schölkopf (2018). More details on ISPD kernels are given in Appendix A.

### 3 OUR METHOD

We propose to learn $f$ by minimizing $R_k(f)$ in (3). To this end, we define an optimal function $f^*$ as a minimizer of the above population risk with respect to a function class $\mathcal{F}$ of real-valued functions on $X$, i.e.,

\[
f^* \in \arg\min_{f \in \mathcal{F}} R_k(f).
\]

It is instructive to note that population risk $R_k$ depends on the choice of the kernel $k$. The important insight for our work is that the population risk $R_k(f)$ has an analytic solution.

**Lemma 1** (Muandet et al. (2020a), Thm 3.3). Assume that $E[(Y - f(X))^2k(Z, Z)] < \infty$. Then, we have

\[
R_k(f) = E[(Y - f(X))(Y' - f(X'))k(Z, Z')]
\]

where $(X', Y', Z')$ is an independent copy of $(X, Y, Z)$.

Based on Assumption 1 and Lemma 1, we obtain the following result, which is a special case of Muandet et al. (2020a, Theorem 3.2), showing that $R_k(f) = 0$ if and only if $f$ satisfies the original conditional moment restriction (see Appendix B.2 for the proof).

**Theorem 1.** Assume that the kernel $k$ is integrally strictly positive definite (ISPD). Then, for any real-valued measurable function $f$, $R_k(f) = 0$ if and only if $E[Y - f(X) | z] = 0$ for $P_z$-almost all $z$.

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

### 3.1 Empirical Risk Minimization

The previous results pave the way for an empirical risk minimization (ERM) framework (Vapnik, 1998) to be used in our work. That is, given a sample \( \{(x_i, y_i, z_i)\}_{i=1}^n \sim P^n(X, Y, Z) \) of size $n$, an empirical estimate of the risk $R_k(f)$ can be obtained as

\[
\hat{R}_U(f) := \sum_{1 \leq i \neq j \leq n} \frac{(y_i - f(x_i))(y_j - f(x_j))k(z_i, z_j)}{n(n-1)},
\]

which is in the form of U-statistic (Serfling, 1980, Section 5). Alternatively, an empirical risk based on V-statistic can also be used, i.e.,

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

Both forms of empirical risk can be used as a basis for a consistent estimation of $f$. The advantage of (5) is that it is a minimum-variance unbiased estimator with appealing asymptotic properties, whereas (6) is a biased estimator of the population risk (4), i.e., $E[\hat{R}_V] \neq R_k$. However, the estimator based on V-statistics employs a full sample and hence may yield better estimate of the risk than the U-statistic counterpart.

Let $x := [x_1, \ldots, x_n]^{\top}$, $y := [y_1, \ldots, y_n]^{\top}$ and $z := [z_1, \ldots, z_n]^{\top}$ be column vectors. Let $K_z$ be the kernel matrix $K(z, z) = [k(z_i, z_j)]_{ij}$ evaluated on the instruments $z$. Then, both (5) and (6) can be rewritten as

\[
\hat{R}_{V(U)}(f) = (y - f(x))^{\top} W_{V(U)} (y - f(x)),
\]

where $f(x) := [f(x_1), \ldots, f(x_n)]^{\top}$ and $W_{V(U)} \in \mathbb{R}^{n \times n}$ is a symmetric weight matrix that depends on the kernel matrix $K_z$. Specifically, the weight matrix $W$ corresponding to (5) is given by $W_U = (K_z - \text{diag}(k(z_1, z_1), \ldots, k(z_n, z_n))/(n(n-1))$ where $\text{diag}(a_1, \ldots, a_n)$ denotes an $n \times n$ diagonal matrix whose diagonal elements are $a_1, \ldots, a_n$. As shown in Appendix B.5, $W_U$ is indefinite and may cause problematic inferences. The weight matrix $W$ for (6) is given by $W_V := K_z/n^2$ which is a positive definite matrix for the ISPD kernel $k$. Finally, our objective (7) also resembles the well-known generalized least regression with correlated noise (Kariya and Kurata, 2004, Chapter 2) where the covariance matrix is the $Z$-dependent invertible matrix $W_{V(U)}^{-1}$. Based on (5) and (6), we estimate $f^*$ by minimizing the regularized empirical risk over a function class $\mathcal{F}$:

\[
\hat{f}_{V(U)} \in \arg\min_{f \in \mathcal{F}} \hat{R}_{V(U)}(f) + \lambda \Omega(f)
\]

where $\lambda > 0$ is a regularization constant satisfying $\lim_{n \to \infty} \lambda = 0$, and $\Omega(f)$ is the regularizer. Since $\hat{R}_U$ and $\hat{R}_V$ respectively minimizes objectives which are regularized U-statistic and V-statistic, they can
be viewed a specific form of M-estimators; see, e.g., Van der Vaart (2000a, Ch. 5). In this work, we focus on the V-statistic empirical risk and provide practical algorithms when \( F \) is parametrized by deep neural networks (NNs) and an RHKS of real-valued functions.

**Kernelized GMM.** We may view the objective (7) from the GMM perspective (Hall, 2005). The assumption that the instruments \( Z \) are exogenous implies that \( \mathbb{E}[\Phi_k(Z)\varepsilon] = 0 \) where \( \Phi_k \) denotes the canonical feature map associated with the kernel \( k \). This gives us an infinite number of moments, \( g(f) = \Phi_k(Z)(Y - f(X)) \). 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 \( J(f) := \|\hat{g}(f)\|^2_{H_k} = \|\hat{g}(f), \hat{g}(f)\|_{H_k} = \frac{1}{n^2} \sum_{i,j=1}^n (y_i - f(x_i))(\Phi_k(z_i), \Phi_k(z_j))_{H_k}(\Phi_k(y_j - f(x_j)) = (y - f(x))^\top W V (y - f(x)). \) That is, our objective (7) defined using V-statistic is a special case of the GMM objective when the weighting matrix is the identity operator. Discovering more efficient weighting operators is also an interesting research direction; see, e.g., Carrasco et al. (2007, Ch. 6) for an earlier endeavour.

### 3.2 Practical MMR-IV Algorithms

A workflow of our algorithm based on \( \hat{R}_V \) is summarized in Algorithm 1; we leave the \( \hat{R}_U \) based method to future work to solve the inference issues caused by indefinite \( W_U \). We provide examples of the class \( F \) in both parametric and non-parametric settings below.

**Deep neural networks.** In the parametric setting, the function class \( F \) can often be expressed as \( F_\Theta = \{ f_\theta : \theta \in \Theta \} \) 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 \( \Phi : x \mapsto \sigma_h(W_h \sigma_{h-1}(\cdots \sigma_1(W_1 x))) \) denotes a nonlinear feature map of a depth-\( h \) NN. Here, \( W_i \) for \( i = 1, \ldots, h \) are parameter matrices and each \( \sigma_i \) denotes the entry-wise activation function of the \( i \)-th 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

\[
\hat{\theta}_V \in \arg \min_{\theta \in \Theta} \hat{R}_V(f_\theta) + \lambda \|\theta\|^2_2
\]

where \( f_\theta \in F_\Theta \). 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 3 in Appendix D.1.

**Kernel machines.** In a non-parametric setting, the function class \( F \) becomes an infinite dimensional space. In this work, we consider \( F \) to be an RKHS \( \mathcal{H}_L \) of real-valued functions on \( X \) with a reproducing kernel \( l : X \times X \rightarrow \mathbb{R} \). Then, the regularized solution can be obtained by

\[
\hat{f}_V = \arg \min_{\alpha \in \mathbb{R}^n} (y - L \alpha)^\top W_V (y - L \alpha) + \lambda \alpha^\top L \alpha
\]

where \( L = [l(x_i, x_j)]_{ij} \) is the kernel matrix on \( x \). For U-statistic version, the quadratic program (8) substitutes indefinite \( W_U \) for \( W_V \), so it may not be positive definite. The value of \( \lambda \) needs to be sufficiently large to ensure that (8) is definite. On the other hand, the V-statistic based estimate (8) is definite for all non-zero \( \lambda \) since \( W_V \) is positive semi-definite. Thus, the optimal \( \hat{\alpha} \) can be obtained by solving the first-order stationary condition and if \( L \) is positive definite, the solution has a closed form expression, \( \hat{\alpha} = (L_W V + \lambda L)^{-1} L W_V y \).

In our experiments, we focus on the V-statistic version because of these advantages. 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 (Williams and Seeger, 2001) to accelerate the matrix inversion in \( \hat{\alpha} = (L W_V + \lambda L)^{-1} L W_V y \). First, we randomly select a subset of \( m (\ll n) \) samples from the original dataset and construct the corresponding sub-matrices of \( W_U \), namely, \( W_{nm} \) and \( W_{mn} \) based on this subset. Second, let \( V \) and \( U \) be the eigenvalue vector and the eigenvector matrix of \( W_{nm} \). Then, the Nyström approximation is obtained as

\[
\hat{\alpha} = \hat{\alpha} \approx (L W_V + \lambda L)^{-1} L W_V y
\]

where \( \hat{\alpha} \) is the matrix inversion using the Woodbury formula (Flannery et al., 1992, p. 75) to obtain \( (L W_V + \lambda L)^{-1} L W_V = L^{-1}(U V U^\top + \lambda L^{-1})^{-1} U V U^\top = \lambda^{-1}(I - U (\lambda^{-1} U^\top L U + V^{-1} - 1)^{-1} U V U^\top \lambda^{-1} L) U V U^\top \). We will refer to this algorithm as MMR-IV (Nyström); Algorithm 2 in Appendix D.1. The runtime complexity of this algorithm is \( O(nm^2 + n^2) \).

**Algorithm 1 MMR-IV**

**Input:** Dataset \( D = \{ x, y, z \} \), kernel function \( k \) with parameters \( \theta_k \), a function class \( F \), regularization functional \( \Omega(\cdot) \), and regularization parameter \( \lambda \)

**Output:** The estimate of \( f^* \) in \( F \).

1. Compute the kernel matrix \( K = k(z, z; \theta_k) \).
2. Define the residual \( \varepsilon(f) = y - f(x) \).
3. \( \hat{f}_\lambda \leftarrow \arg \min_{f \in F} \frac{1}{n^2} \varepsilon(f)^\top K \varepsilon(f) + \lambda \Omega(\|f\|_F) \)
4. return \( \hat{f}_\lambda \)
3.3 Hyper-parameter Selection

The V-statistic empirical risks (6) provides a convenient hyper-parameter selection through cross validation (CV) procedures. In particular, one can observe that our RKHS model in (8) is similar to the kernel ridge regression, except the V-statistic empirical risk $\hat{R}(f)$. Based on this observation, we provide an efficient way for MMR-IV (Nystöm) to select the regularization parameter $\lambda$ and parameters of $l$ using CV, and the derivation is inspired by Vehatari et al. (2016). While for NN, we follow the standard CV.

Gaussian process (GP) interpretation. We first introduce a GP over $f$, and we show that maximizing its posterior distribution minimizes the regularized V-statistic objective (8), and equivalence between the predictive mean and the optimal function of (8) in Appendix C. The GP model has the prior of $f(x) \sim \text{GP}(0, \delta(x, x))$ and the likelihood $p(D|f(x)) = (2\pi)^{n/2}K_z^{-1/2}N(f(x) | y, K_z)$, where $l$ is the kernel function, $\delta = (\lambda z^2)^{-1}$ plays the role of the regularization parameter justified in Appendix C, and $D := \{ (x, y, z) \}$. The posterior distribution of $f(x)$ can be straightforwardly derived from the Gaussian-like prior and likelihood,

$$p(f(x)|D) = N(f(x)|c, C) \propto p(f(x))p(D|f(x)),$$

where $C = (K_z + (\delta L)^{-1})^{-1} \approx \delta[L - L\tilde{U}(n^2\delta\tilde{U}^T\tilde{L}\tilde{U} + \tilde{V}^{-1})^{-1}\tilde{U}(n^2\delta)L]$ by Nyström approx. and $c = CK_zy$.

Analytical cross-validation error. Now, we derive the analytical error of the leave-M-out cross validation (LOMCV) from the perspective of the GP. We use $D_{-M}$ and $D_M \equiv \{ (x_M, y_M, z_M) \}$ to denote the training and validation datasets, and clearly $D = D_M \cup D_{-M}$. Given $D_{-M}$, the predictive distribution on $D_M$ can be obtained by Bayes’ rules $p(f(x_M)|D_{-M}) := N(f(x_M)|b, B) \propto p(f(x_M)|D)/p(D_{-M}|f(x_M))$. $p(f(x_M)|D_{-M})$ is Gaussian because of the Gaussian posterior and likelihood, i.e., $p(f(x_M)|D) = N(f(x_M)|c_M, C_M)$ and $p(D_{-M}|f(x_M)) = N(f(x_M)|y_M, K_{-M})$, where $c_M, C_M$ are the mean and covariance of $f(x_M)$ in the posterior $p(f(x)|D)$ and $K_{-M} = k(z_M, z_M)$. We obtain $b$ and $B$ by matching moments of normalized $p(D_{-M}|f(x_M))p(f(x_M)|D_{-M})$ and $p(f(x_M)|D)$:

$$B^{-1} = C_M^{-1} - K_M, b = B(C_M^{-1}c_M - K_My_M).$$

As the predictive function value is the predictive mean $\hat{f}(x_M) = b$, error of m LOMCV is expressed as

$$\text{LOMCV Error} := \sum_{i=1}^{M}(r(i))^T K_M^{-1}r(i),$$

where we denote by $(i)$ the LOMCV index and the residual $r(i) := \hat{y}^{(i)} - y^{(i)} = (1 - C_M^{-1}K_M^{-1})(c_M - y_M)$. The analytical error enables an efficient parameter selection with the CV procedure.

4 THEORETICAL RESULTS

We provide the consistency and asymptotic normality of $\hat{f}_V$. We also develop the same results for $\hat{f}_U$, but we defer them to the appendix due to space limitation. All proofs are presented in the appendix.

4.1 Consistency

We first show the consistency of $\hat{f}_V$, which depends on the convexity and uniform convergence of the risk functions. The result holds regardless of the shape of $\Omega(f)$, so we can utilize the regularization $\|\theta\|_2^2$ which is common for NN but non-convex in terms of $f$.

Theorem 2 (Consistency of $\hat{f}_V$). Assume that the distribution of $Y$ has bounded support, $F$ is compact, and Assumption 1 holds, $\Omega(f)$ is a bounded function, $\lambda \xrightarrow{p} 0$, and $F$ is convex. Then $\hat{f}_V \xrightarrow{p} f^*$.

If the $\Omega(f)$ is convex in $f$, the consistency can be obtained more easily by Newey and McFadden (1994, Theorem 2.7). In this case, we can avoid several conditions. We provide an additional result with this setting in Section B.6 in the appendix.

4.2 Asymptotic Normality

We analyze asymptotic normality of the estimator $\hat{f}_V$, which is important to advanced statistical analysis such as tests. Here, we investigate two different cases: the estimator has finite- and infinite-dimension.

Finite-dimension case. We consider the $\hat{f}_V$ is characterized by a finite-dimensional parameter from a parameter space $\Theta$. We rewrite the regularized V-statistic risk as a compact form $\hat{R}_{V,\lambda}(f) := \frac{1}{n} \sum_{i=1}^{n} h^u(u_i, u_j) + \lambda \Omega(\theta)$, $h^u(u_i, u_j) := (y_i - f(x_i))(y_j - f(x_j))k(z_i, z_j)$, and consider $R_k(f)$ is uniquely minimized at $\theta^* \in \Theta$.

Theorem 3 (Asymptotic normality of $\hat{f}_V$). Suppose that $H = \mathbb{E}[\nabla^2_\theta x_h(U, U')]$ is non-singular, $\Theta$ is compact, $p(Y)$ has bounded support, $\sqrt{n}\lambda \xrightarrow{P} 0$, $f_0$ and $\Omega(\theta)$ are twice continuously differentiable about $\theta$, $\sqrt{n}\lambda \xrightarrow{P} 0$, $\theta^*$ is an interior point of $\Theta$, and Assumption 1 holds. Then, $\sqrt{n}(\hat{f}_V - \theta^*) \xrightarrow{D} N(0, \Sigma_V)$ holds, where $\Sigma_V = 4H^{-1}\mathbb{E}[H_2, [h^u(U, U')]|H^{-1}$ and $\xrightarrow{D}$ denotes a convergence in law.

Infinite-dimension case. We show the asymptotic normality of an infinite-dimensional estimator $\hat{f}_V$. That is, we show that an error of $f_V$ weakly converges to a Gaussian process that takes values in a function space $\mathcal{H}_l$. We set $\Omega(f) = \|f\|^2_{\mathcal{H}_l}$ and consider a minimizer:

$$f_{\lambda_0}^{\star} \in \arg\min_{f \in \mathcal{F}} R_k(f) + \lambda_0\|f\|^2_{\mathcal{H}_l}$$

with arbitrary $\lambda_0 > 0$. We also define $N(\mathcal{H}, \|\cdot\|)$ as an
\( \varepsilon \)-covering number of a set \( \mathcal{H} \) in terms of \( \| \cdot \| \). Then, we obtain the following:

**Theorem 4.** 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 there exists \( s \in (0, 2) \) and a constant \( C_H > 0 \) such that \( \log \mathcal{N}(\varepsilon, \mathcal{H}_l, \| \cdot \|_{L^s}) \leq C_H \varepsilon^{-s} \) for any \( \varepsilon \in (0, 1) \). Then, there exists a Gaussian process \( \mathbb{G}_p^\ast \) such that \( \sqrt{n}(f_Y - f_\lambda^\ast) \sim \mathbb{G}_p^\ast \) in \( \mathcal{H}_l \).

This result allows statistical inference on functional estimators such as kernel machines. Although there are many conditions, all of them are valid for many well-known kernels; see Appendix B.11 for examples.

### 5 RELATED WORK

When the relationships between \( Z \) and \( Y \) as well as between \( X \) and \( Y \) are nonlinear, several extensions of 2SLS and GMM have been proposed. In the two-stage approach, the function \( f(x) \) has often been obtained by solving a Fredholm integral equation of the first kind \( \mathbb{E}[Y|Z] = \int f(x) dP(X|Z) \). In Newey and Powell (2003); Blundell et al. (2007); Horowitz (2011); Chen and Pouzo (2012), linear regression is replaced by a linear projection onto a set of known basis functions. A uniform convergence rate of this approach is provided in Chen and Christensen (2018). However, it remains an open question how to choose the appropriate set of basis functions. In Hall and Horowitz (2005) and Darolles et al. (2011), the first-stage regression is replaced by a conditional density estimation of Darolles et al. (2011), the first-stage regression is re-

6 EXPERIMENTAL RESULTS

We present the experimental results in a wide range of settings for IV estimation. Following Lewis and Syrgkanis (2018) and Bennett et al. (2019), we consider both low and high-dimensional scenarios. In our experiments, we compare our algorithms to the following baselines: (i) DirectNN (ii) 2SLS (iii) Poly2SLS (iv) DeepIV (Hartford et al., 2017) (v) KernelIV (Singh et al., 2019) (vi) GMM+NN (vii) ACM (Lewis and Syrgkanis, 2018) (viii) DeepGMM (Bennett et al., 2019). We refer readers to Appendix D for more experimental details due to space limitation, and our code is publicly available\(^1\).

6.1 Low-dimensional Scenarios

Following Bennett et al. (2019), we employ the following data generation process:

\[
Y = f^\ast(X) + \varepsilon + \delta, \quad X = Z_1 + \varepsilon + \gamma,
\]

where \( Z := (Z_1, Z_2) \sim \text{Uniform}([-3, 3]^2), \varepsilon \sim \mathcal{N}(0, 1), \) and \( \gamma, \delta \sim \mathcal{N}(0, 0.1^2). \) In words, \( Z \) is a two-dimensional IV, but only the first instrument \( Z_1 \) has an

\(^1\)https://github.com/RuiZhang2016/MMRIV
Table 1: The mean square error (MSE) ± one standard deviation in the large-sample regime (n = 2000).

| Algorithm         | True Function f∗ | Setting MNISTx | Setting MNISTx | Setting MNISTxx |
|-------------------|------------------|----------------|----------------|-----------------|
| DirectNN          | 0.116 ± 0.000    | 0.355 ± 0.000 | 0.190 ± 0.000 | 0.168 ± 0.011   |
| 2SLS              | 0.135 ± 0.000    | 0.306 ± 0.000 | 0.183 ± 0.000 | 0.156 ± 0.012   |
| Poly2SLS          | 0.192 ± 0.000    | 0.372 ± 0.000 | 0.215 ± 0.000 | 0.175 ± 0.012   |
| GMM-NN            | 0.214 ± 0.000    | 0.393 ± 0.000 | 0.238 ± 0.000 | 0.202 ± 0.012   |
| AGMM              | 0.260 ± 0.000    | 0.436 ± 0.000 | 0.281 ± 0.000 | 0.240 ± 0.012   |
| DeepIV            | 0.271 ± 0.000    | 0.446 ± 0.000 | 0.293 ± 0.000 | 0.250 ± 0.012   |
| Kernel2IV         | 0.297 ± 0.000    | 0.472 ± 0.000 | 0.318 ± 0.000 | 0.265 ± 0.012   |
| MMR-IV (NN)       | 0.311 ± 0.000    | 0.496 ± 0.000 | 0.340 ± 0.000 | 0.299 ± 0.012   |
| MMR-IV (Rys)      | 0.311 ± 0.000    | 0.496 ± 0.000 | 0.340 ± 0.000 | 0.299 ± 0.012   |

Table 2: The mean square error (MSE) ± one standard deviation on high-dimensional structured data. We run each method 10 times.

| Algorithm            | Setting MNISTx | Setting MNISTx | Setting MNISTxx |
|----------------------|----------------|----------------|-----------------|
| DirectNN             | 0.134 ± 0.000  | 0.229 ± 0.000  | 0.196 ± 0.011   |
| 2SLS                 | 0.563 ± 0.001  | >1000          | >1000           |
| Ridge2SLS            | 0.567 ± 0.001  | 0.431 ± 0.000  | 0.705 ± 0.000   |
| GMM+NN               | 0.121 ± 0.004  | 0.235 ± 0.002  | 0.240 ± 0.016   |
| AGMM                 | n/a            | n/a            | n/a             |
| DeepGMM              | 0.114 ± 0.005  | n/a            | n/a             |
| MMR-IV (NN)          | 0.024 ± 0.006  | 0.124 ± 0.021  | 0.130 ± 0.009   |
| MMR-IV (Rys)         | 0.015 ± 0.002  | 0.442 ± 0.000  | 0.425 ± 0.002   |

Figure 2: Runtime comparison in the large-sample regime (n = 2000). The computational time of parameter selection is excluded from the comparison.

Figure 2: Runtime comparison in the large-sample regime (n = 2000). The computational time of parameter selection is excluded from the comparison.

The 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) \(\text{step}\): \(f^*(x) = 1_{(x \geq 0)}\). (iii) \(\text{abs}\): \(f^*(x) = |x|\). (iv) \(\text{linear}\): \(f^*(x) = x\). We consider both small-sample (n = 200) and large-sample (n = 2000) regimes.

The results are reported in Table 1 for the large-sample regime (and Table 3 in Appendix D for the small-sample regime). Our findings are as follows: (i) MMR-IVs perform reasonably well in most cases. (ii) The linear methods (2SLS and Poly2SLS) perform best when the linearity assumption is satisfied. (iii) Some nonlinear but complicated methods (GMM+NN, DeepIV and DeepGMM) are unstable due to the sensitivity to hyper-parameters. It shows that MMR-IV (Nyström) has the advantage of adaptive hyper-parameter selection. Section D.2 provides more comments.

We also record the runtimes of all methods on the large-sample regime and report them in Figure 2. 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. 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). Lastly, the two RKHS-based methods, namely, KernelIV and MMR-IV (RKHS), have similar runtime complexity. We observe that both of them scale poorly to large datasets.

6.2 High-dimensional Structured Scenarios

In high-dimensional setting, we employ the same data generating process as in Section 6.1. We consider only the absolute function for \(f^*\), but map \(Z, X\), or both \(X\) and \(Z\) to MNIST images (784-dim) (LeCun et al., 1998). Let us denote the original outputs in Section 6.1 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 \(RI(d)\) be a function that selects a random MNIST image from the digit class \(d\). We set \(n = 10,000\). Then, the scenarios we consider are (i) \(\text{MINIST}_Z\): \(Z \leftarrow RI(\pi(Z^{low}))\), (ii) \(\text{MINIST}_X\): \(X \leftarrow RI(\pi(X^{low}))\), and (iii) \(\text{MINIST}_{XZ}\): \(X \leftarrow RI(\pi(X^{low})), Z \leftarrow RI(\pi(Z^{low}))\).

We report the results in Table 2. We summarize our findings: (i) MMR-IV (NN) performs well across scenarios. (ii) MMR-IV (Nyström) fails except MINIST\(_Z\), because the ARD kernel with PCA are not representative enough for MINIST\(_X\). (iii) Some methods (AGMM and DeepIV) do not work when \(X\) is high-dimensional, similar to Bennett et al. (2019). (iv) The two-step methods (Poly2SLS and Ridge2SLS) has large errors because the first-stage regression from \(Z\) to \(X\) is ill-posed. Section D.3 provides more comments.

6.3 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 Hartwig et al. (2017):

\[
Y = \beta X + c_1 e + \delta, \quad X = \sum_{i=1}^{m} \alpha_i Z_i + c_2 e + \gamma,
\]
Figure 3: 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.

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 $\beta, c_1$ control the strength of exposures and confounders to outcomes, while $c_2, \alpha_i$ 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 instrument becomes weaker while the overall strength of instruments ($\sum_i \alpha_i$) remains roughly constant.

There are several reasons genetic variants may not be valid IVs (Burgess et al., 2020). Most notably, a variant might affect exposures on different causal pathways (pleiotropy). Linkage disequilibrium and population stratification are another common reasons for invalid IVs. 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 experiments: (i) $d' = 8, 16, 32$; (ii) $c_1 = 0.5, 1, 2$; (iii) $c_2 = 0.5, 1, 2$; unmentioned parameters use default values: $\beta = 1$, $d' = 16$, $c_1 = 1$, $c_2 = 1$. 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-dimensional scenario.

Figure 3 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 (Angrist and Pischke, 2008). 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. Section D.5 contains additional findings.

**The Vitamin D data.** Lastly, we apply our algorithm to the Vitamin D data (Sjolander and Martinussen, 2019, Sec. 5.1). The data were collected from a 10-year study on 2571 individuals aged 40–71 and 4 variables are employed: age (at baseline), filaggrin (binary indicator of filaggrin mutations), VitD (Vitamin D level at baseline) and death (binary indicator of death during study). The goal is to evaluate the potential effect of VitD on death. We follow Sjolander and Martinussen (2019) by controlling age in the analyses, using filaggrin as instrument, and then applying the MMR-IV (Nyström) algorithm. Sjolander and Martinussen (2019) modeled the effect of VitD on death by a generalized linear model and found the effect is insignificant by 2SLS ($p$-value on the estimated coefficient is 0.13 with the threshold of 0.05). More details can be found in Appendix D.6. The estimated effect is illustrated in Figure 4 in the appendix. We observe that: (i) the result without instruments is counterintuitive: at a young age, deficiency of VitD results in low mortality rate; (ii) In contrast, the result with instruments is more reasonable: a low VitD level at a young age has a slight effect on death, but a more adverse effect at an old age (Meehan and Penckofer, 2014).

7 CONCLUSION

Learning causal relations when hidden confounders are present is a cornerstone of decision making. IV regression is a standard tool to tackle this task, but currently faces challenges in nonlinear settings. The present work presents a simple framework that overcomes some of these challenges. We employ RKHS theory in the reformulation of conditional moment restriction (CMR) as a maximum moment restriction (MMR) based on which we can approach the problem from the empirical risk minimization (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. The paper also shows a way of combining the elegant theoretical approach of kernel methods with practical merits of deep neural networks. Despite these advantages, the optimal choice of the kernel $k$ remains an open question which we hope to address in future work.
References

M. G. Akritas et al. Empirical processes associated with V-statistics and a class of estimators under random censoring. *The Annals of Statistics*, 14(2): 619–637, 1986.

J. D. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press, 2008.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434): 444–455, 1996.

M. A. Arcones and E. Gine. Limit theorems for U-processes. *The Annals of Probability*, pages 1494–1542, 1993.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3): 337–404, 1950.

A. Bennett, N. Kallus, and T. Schnabel. Deep generalized method of moments for instrumental variable analysis. In *Advances in Neural Information Processing Systems 32*, pages 3564–3574. Curran Associates, Inc., 2019.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

R. Blundell, X. Chen, and D. Kristensen. Semiparametric IV estimation of shape-invariant Engel curves. *Econometrica*, 75(6): 1613–1669, 2007.

S. Burgess, D. S. Small, and S. G. Thompson. A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research*, 26(5): 2333–2355, 2017.

S. Burgess, C. N. Foley, E. Allara, J. R. Staley, and J. M. M. Howson. A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nature Communications*, 11(1): 376, 2020.

D. Card. The causal effect of education on earnings. In O. Ashenfelter and D. Card, editors, *Handbook of Labor Economics*, volume 3 of *Handbook of Labor Economics*, chapter 30, pages 1801–1863. Elsevier, 1999.

M. Carrasco, J.-P. Florens, and E. Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. In J. Heckman and E. Leamer, editors, *Handbook of Econometrics*, volume 6B, chapter 77. Elsevier, 1 edition, 2007.

X. Chen and T. M. Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quantitative Economics*, 9(1): 39–84, 2018.

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

S. Daron, Y. Fan, J. P. Florens, and E. Renault. Nonparametric instrumental regression. *Econometrica*, 79(5): 1541–1565, 2011.

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

S. G. Donald and W. K. Newey. Choosing the number of instruments. *Econometrica*, 69(5): 1161–1191, 2001.

B. P. Flannery, W. H. Press, S. A. Teukolsky, and W. Vetterling. Numerical recipes in C. Press Sydicate of the University of Cambridge, New York, 24: 78, 1992.

K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, Dec. 2004.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

R. Hable. Asymptotic normality of support vector machine variants and other regularized kernel methods. *Journal of Multivariate Analysis*, 106:92–117, 2012.

A. Hall. *Generalized Method of Moments*. Advanced texts in econometrics. Oxford University Press, 2005.

P. Hall and J. L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33(6): 2904–2929, 12 2005.

L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4): 1029–1054, 1982.

J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy. Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1414–1423. PMLR, 2017.

F. P. Hartwig, G. Davey Smith, and J. Bowden. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology*, 46(6): 1985–1998, 2017.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301): 13–30, 1963.

J. L. Horowitz. Applied nonparametric instrumental variables estimation. *Econometrica*, 79(2): 347–394, 2011.
Maximum Moment Restriction for Instrumental Variable Regression

V. K. Ivanov, V. V. Vasin, and V. Tanana. Theory of Linear Ill-posed Problems and Its Applications. Inverse and ill-posed problems series. VSP, 2002.

T. Kariya and H. Kurata. Generalized least squares. John Wiley & Sons, 2004.

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

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

G. Lewis and V. Syrgkanis. Adversarial generalized method of moments. 03 2018.

D. G. Luenberger. Optimization by vector space methods. John Wiley & Sons, 1997.

H. B. Mann and A. Wald. On stochastic limit and order relationships. The Annals of Mathematical Statistics, 14(3):217–226, 09 1943.

M. Meehan and S. Penckofer. The role of vitamin D in the aging adult. Journal of aging and gerontology, 2(2):60, 2014.

K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyond. Foundations and Trends in Machine Learning, 10(1-2):1–141, 2017.

K. Muandet, W. Jitkrittum, and J. Kühler. Kernel conditional moment test via maximum moment restriction. In Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence, volume 124 of Proceedings of Machine Learning Research, pages 41–50. PMLR, 2020a.

K. Muandet, A. Mehrjou, S. K. Lee, and A. Raj. Dual instrumental variable regression. In Advances in Neural Information Processing Systems 33. Curran Associates, Inc., 2020b. Forthcoming.

W. Newey. Efficient estimation of models with conditional moment restrictions. In Handbook of Statistics, volume 11, chapter 16, pages 419–454. 1993.

W. K. Newey and D. McFadden. Chapter 36 large sample estimation and hypothesis testing. volume 4 of Handbook of Econometrics, pages 2111 – 2245. Elsevier, 1994.

W. K. Newey and J. L. Powell. Instrumental variable estimation of nonparametric models. Econometrica, 71(5):1565–1578, 2003.

C. E. Rasmussen and C. K. I. Williams. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press, 2005.

B. Schölkopf and A. Smola. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA, 2002.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In International conference on computational learning theory, pages 416–426. Springer, 2001.

R. Serfling. Approximation theorems of mathematical statistics. John Wiley & Sons, 1980.

C.-J. Simon-Gabriel and B. Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. Journal of Machine Learning Research, 19(44):1–29, 2018.

R. Singh, M. Sahani, and A. Gretton. Kernel instrumental variable regression. In Advances in Neural Information Processing Systems 32, pages 4593–4605, 2019.

A. Sjolander and T. Martinussen. Instrumental variable estimation with the R package ivtools. Epidemiologic Methods, 8(1), 2019.

L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In Proceedings of the 26th International Conference on Machine Learning (ICML), June 2009.

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

B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. Journal of Machine Learning Research, 12:2389–2410, July 2011.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. Journal of Machine Learning Research, 12:2389–2410, July 2011.

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

A. Tsybakov. Introduction to Nonparametric Estimation. Springer Publishing Company, Incorporated, 1st edition, 2008.

A. Van der Vaart. Asymptotic Statistics. Cambridge University Press, 2000a.

A. Van der Vaart. Asymptotic statistics, volume 3. Cambridge university press, 2000b.

A. Van der Vaart and J. Wellner. Weak convergence and empirical processes, 1996.

V. N. Vapnik. Statistical Learning Theory. Wiley-Interscience, 1998.
A. Vehtari, T. Mononen, V. Tolvanen, T. Sivula, and O. Winther. Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *The Journal of Machine Learning Research*, 17(1): 3581–3618, 2016.

C. J. Walder and A. N. Bishop. Fast Bayesian intensity estimation for the permanental process. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3579–3588. JMLR.org, 2017.

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

D.-X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.
Appendix

Maximum Moment Restriction for Instrumental Variable Regression

Table of Contents

A INTEGRALLY STRICTLY POSITIVE DEFINITE (ISPD) KERNELS 12
B DETAILED PROOFS 13
   B.1 Proof of Lemma 1 ................................................................. 13
   B.2 Proof of Theorem 1 ................................................................. 13
   B.3 Convexity Result ................................................................. 13
   B.4 Uniform Convergence of Risk Functionals .................................. 14
   B.5 Indefiniteness of Weight Matrix $W_U$ ...................................... 15
   B.6 Consistency of $\hat{f}_V$ with Convex $\Omega(f)$ ............................ 15
   B.7 Proof of Theorem 2 ................................................................. 15
   B.8 Consistency of $\hat{f}_U$ ............................................................. 16
   B.9 Asymptotic Normality of $\hat{\theta}_U$ ............................................. 16
   B.10 Proof of Theorem 3 ............................................................... 18
   B.11 Asymptotic Normality in the Infinite-dimension Case .................. 19
C GAUSSIAN PROCESS (GP) INTERPRETATION 24
   C.1 Likelihood Function ............................................................... 25
   C.2 Maximum A Posteriori (MAP) ................................................... 25
D ADDITIONAL DETAILS OF EXPERIMENTS 26
   D.1 Baseline Algorithms ............................................................... 26
   D.2 Experimental Settings on Low-dimensional Scenario .................... 27
   D.3 Experimental Settings on High-dimensional Scenario .................... 28
   D.4 More Details of MMR-IVs ......................................................... 28
   D.5 Additional Comments on Mendelian Randomization ..................... 29
   D.6 Experimental Settings on Vitamin D Data .................................... 29

A INTEGRALLY STRICTLY POSITIVE DEFINITE (ISPD) KERNELS

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

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

where $\sigma$ is a positive bandwidth parameter. Another important kernel is an inverse multiquadric (IMQ) kernel

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

where $c$ and $\gamma$ are positive parameters (Steinwart and Christmann, 2008, Ch. 4). This class of kernel functions is closely related to the notions of universal kernels (Steinwart, 2002) and characteristic kernels (Fukumizu et al., 2004). 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 Sriperumbudur et al. (2011) and Simon-Gabriel and Schölkopf (2018) for more details.
B DETAILED PROOFS

This section contains detailed proofs of the results that are missing in the main paper.

B.1 Proof of Lemma 1

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

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

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 (Steinwart and Christmann, 2008, Def. A.5.20). Hence, we can write (10) as

$$\|\mathbb{E}[(Y - f(X))k(Z, \cdot)]\|_{\mathcal{H}_k}^2 = \mathbb{E}[(Y - f(X))k(Z, \cdot), \mathbb{E}[(Y - f(X))k(Z, \cdot)]_{\mathcal{H}_k}$$

as required.

B.2 Proof of Theorem 1

Proof. First, the law of iterated expectation implies that

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

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

$$R_k(f) = \int \int_Z g(z)k(z, z')g(z')dzdz' = 0,$$

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

B.3 Convexity Result

Theorem 5. If $\mathcal{F}$ is a convex set, then the risk $R_k$ given in (4) is strictly convex on $\mathcal{F}$.

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

$$R_k(\alpha f + (1 - \alpha)g) - \alpha R_k(f) - (1 - \alpha)R_k(g) < 0.$$
\[
\begin{align*}
&= \alpha(\alpha - 1) \|E[(f(X) - g(X))k(Z, \cdot)]\|_{H_k}^2 < 0 \\
&\quad \text{for some } \alpha > 0, \quad \text{and } \sup_{\alpha > 0} \left\{ E[(f(X) - g(X))k(Z, \cdot)] \right\} < 0.
\end{align*}
\]

The equality (a) is obtained by considering \( Y = \alpha Y + (1 - \alpha)Y \) in \( \|E[(Y - \alpha f(X) - (1 - \alpha)g(X))k(Z, \cdot)]\|_{H_k}^2 \) on the left hand side of (a). We note that the right hand side of (a) is quadratic in \( \|E[(Y - f(X))k(Z, \cdot)]\|_{H_k}^2 \) and \( \|E[(Y - g(X))k(Z, \cdot)]\|_{H_k}^2 \), 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 and \( \alpha(\alpha - 1) < 0 \). \( \square \)

### B.4 Uniform Convergence of Risk Functionals

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

**Lemma 2** (Uniform consistency of \( \hat{f}_V \)). Assume that the distribution of \( Y \) has bounded support, \( \mathcal{F} \) is compact, and Assumption 1 holds. Then, the risk \( R_k(f) \) is continuous about \( f \in \mathcal{F} \) and \( \sup_{f \in \mathcal{F}} |\hat{R}_V(f) - R_k(f)| \xrightarrow{P} 0 \).

**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 that \( \hat{R}_V \) converges uniformly to \( R_k \), we need to show that (i) \( h_f(u, u') \) is continuous at each \( f \) with probability one; (ii) \( E_{u, u'} \left[ \sup_{f \in \mathcal{F}} |h_f(u, u')| \right] < \infty \); and (iii) \( E_{u, u'} \left[ \sup_{f \in \mathcal{F}} |h_f(u, u')| \right] \) exists for some constant \( M \in \mathbb{R}^+ \). Since the distribution of \( Y \) has a bounded support, \( \mathcal{F} \) is compact, and Assumption 1 holds, then \( E[|Y|] < \infty \). The kernel \( k(\cdot, \cdot) \) is continuous about \( (x, y, z) \) and \( \sup_{x,y,z} \sqrt{k(x, z)k(z', z')} < \infty \). Consequently, we have \( h_f \) is continuous at each \( f \) with probability one and

\[
E_{u, u'} \left[ \sup_{f \in \mathcal{F}} |h_f(u, u')| \right] < \infty \quad \text{and} \quad E_{u, u'} \left[ \sup_{f \in \mathcal{F}} |h_f(u, u')| \right] < \infty.
\]

Hence, our assertion follows from Newey and McFadden (1994, Lemma 8.5). \( \square \)

**Lemma 3** (Uniform consistency of \( \hat{R}_U \)). Assume that the distribution of \( Y \) has bounded support, \( \mathcal{F} \) is compact and Assumption 1 holds. Then, \( R_k(f) \) is continuous about \( f \) and \( \sup_{f \in \mathcal{F}} |\hat{R}_U(f) - R_k(f)| \xrightarrow{P} 0 \).

**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 \( \hat{R}_U \), we need to show that (i) \( h_f(u, u') \) is continuous at each \( f \) with probability one; (ii) \( \sup_{u, u'} |h_f(u, u')| \) exists for all \( f \in \mathcal{F} \) and \( E_{u, u'}[|h_f(u, u')|] < \infty \); (iii) \( (u, u') \rightarrow (x, y, z) \) is continuous and bounded as per Assumption 1, i.e.,
sup_{z \in \mathbb{Z}} \sqrt{k(z,z)} < \infty$, so $\sqrt{k(z,z)}k(z',z')$ is integrable. Consequently, (i) $h_f(u, u')$ is continuous at each $f$ with probability one and (ii) $\mathbb{E}_{u,u'}[d(u, u')] < \infty$.

Furthermore, we show that $(u_i, u_j)_{i \neq j}$ has strict stationarity and ergodicity. Strict stationarity means that the distribution of a set of data $(u_i, u_{j \neq i})_{i=1}^{i+m+j+m'}$ does not depend on the starting indices $i, j$ for any $m$ and $m'$, which is easy to check. Ergodicity means that $\tilde{R}_U(f) \xrightarrow{p} R_k(f)$ 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 $\tilde{R}_U(f) \xrightarrow{p} R_k(f)$ follows by Hoeffding (1963, P.25). Therefore, ergodicity holds, and we have shown all conditions required by extended results of Newey and McFadden (1994, Lemma 2.4). Then, it follows that $\sup_{f \in \mathcal{F}} |\tilde{R}_U(f) - R_k(f)| \xrightarrow{p} 0$ and $R_k(f)$ is continuous. □

### B.5 Indefiniteness of Weight Matrix $W_U$

**Theorem 6.** If Assumption 1 holds, $W_U$ is indefinite.

**Proof.** By definition, we have

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

where $\text{diag}(a_1, \ldots, a_n)$ 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 $\text{trace}(W_U) = 0$. Let us denote the eigenvalues of $W_U$ by $\{\lambda_i\}_{i=1}^n$. Since $\sum_{i=1}^n \lambda_i = \text{trace}(W_U)$, we conclude that there exist both positive and negative eigenvalues (all eigenvalues being zeros yields trivial $W_U = 0$). As a result, $W_U$ is indefinite. □

### B.6 Consistency of $\hat{f}_V$ with Convex $\Omega(f)$

**Theorem 7 (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 $f$, $\lambda \xrightarrow{p} 0$ and Assumption 1 holds. Then, $\hat{f}_V$ exists with probability approaching one and $\hat{f}_V \xrightarrow{p} f^*$.

**Proof.** Given $\Omega(f)$ is convex about $f$, we prove the consistency based on Newey and McFadden (1994, Theorem 2.7) which requires (i) $R_k(f)$ is uniquely maximized at $f^*$; (ii) $\tilde{R}_V(f) + \lambda \Omega(f)$ is convex; (iii) $\tilde{R}_V(f) + \lambda \Omega(f) \xrightarrow{p} R_k(f)$ for all $f \in \mathcal{F}$.

Recall that $\tilde{R}_V(f) = \left\| \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))k(z_i, \cdot) \right\|_{\tilde{R}_k}^2$, and by the law of large number, we have that $\frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))k(z_i, \cdot) \xrightarrow{D} \mathbb{E}[(Y - f(X))k(Z, \cdot)]$. Then $\tilde{R}_V(f) \xrightarrow{p} R_k(f)$ follows from the Continuous Mapping Theorem (Mann and Wald, 1943) based on the fact that the function $g(\cdot) = \| \cdot \|_{\tilde{R}_k}$ is continuous. As $\lambda \xrightarrow{p} 0$, we obtain (iii) $\tilde{R}_V(f) + \lambda \Omega(f) \xrightarrow{p} R_k(f)$ by Slutsky’s theorem (Van der Vaart, 2000b, Lemma 2.8). Besides, it is easy to see that $\tilde{R}_V(f)$ is convex because the weight matrix $W_V$ is positive definite, and (ii) $\tilde{R}_V(f) + \lambda \Omega(f)$ is convex due to convex $\Omega(f)$. Further, the condition (i) directly follows from Theorem 5, and given that $f^*$ is an interior point of the convex set $\mathcal{F}$, our assertion follows from Newey and McFadden (1994, Theorem 2.7). □

### B.7 Proof of Theorem 2

**Proof.** From the conditions of Lemma 2, we know that $\mathcal{F}$ is compact, $R_k(f)$ is continuous about $f$ and $\sup_{f \in \mathcal{F}} |\hat{R}_V(f) - R_k(f)| \xrightarrow{p} 0$. As $\mathcal{F}$ is convex, $R_k(f)$ is convex on $\mathcal{F}$ and is uniquely minimized at $f^*$. Based on the conditions that $\Omega(f)$ is bounded and $\lambda \xrightarrow{p} 0$, we obtain by Slutsky’s theorem that

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

Consequently, we assert the conclusion by Newey and McFadden (1994, Theorem 2.1). □
B.8 Consistency of $\hat{f}_U$

**Theorem 8** (Consistency of $\hat{f}_U$). Assume that conditions of Lemma 3 hold, $F$ is a convex function and $\lambda \xrightarrow{p} 0$. Then $\hat{f}_U \xrightarrow{p} f^*$. 

**Proof.** By the conditions of Lemma 3, we know that $F$ is compact, $R_k(f)$ is continuous about $f$ and $\sup_{f \in F} |\hat{R}_U(f) - R_k(f)| \xrightarrow{p} 0$. As $F$ is convex, $R_k(f)$ is convex on $F$ and is uniquely minimized at $f^*$. Based on the conditions that $\Omega(f)$ is bounded and $\lambda \xrightarrow{p} 0$, we obtain by Slutsky’s theorem that

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

Consequently, we assert the conclusion by Newey and McFadden (1994, Theorem 2.1).

B.9 Asymptotic Normality of $\hat{\theta}_U$

In this section, we consider the regularized U-statistic risk $\hat{R}_{U,\lambda}(f_\theta)$. 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

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

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

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

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

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

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

**Lemma 4.** Suppose that $f_\theta$ and $\Omega(\theta)$ are continuously differentiable about $\theta$, $E[||\nabla_\theta h_\theta^*(U, U')||^2] < \infty$, there exists $z \in Z$ such that $E_{X}[(\nabla_\theta f_\theta^*)(X) | z]p(z) \neq 0$ or $E_{XY}[(Y - f_\theta^*)(X) | z]p(z) \neq 0$, and $\sqrt{n}\lambda \xrightarrow{p} 0$. Then,

$$
\sqrt{n}\nabla_\theta \hat{R}_{U,\lambda}(f_\theta^*) \xrightarrow{D} N(0, 4E_{U'}[\nabla_\theta h_\theta^*(U, U')]).
$$

**Proof.** The proof follows from Serfling (1980, Section 5.5.1 and Section 5.5.2) and we need to show that (i) $\nabla_\theta \hat{R}_U(f_\theta^*) \xrightarrow{p} 0$ and (ii) whether $\Var_{U}[\nabla_\theta h_\theta^*(U, U')]>0$ or not. (i) can be obtained by the law of large numbers because $\nabla_\theta \hat{R}_U(f_\theta^*)$ is a sample average of $\nabla_\theta R_k(f_\theta^*) = 0$.

To prove (ii), we first note that $E_{U'}[\nabla_\theta h_\theta^*(U, U')] = 0$. Then,

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

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

$$
E_{X'Z'}[(\nabla_\theta f_\theta^*(X')k(Z', Z)] = E_{Z'}[E_{X'}[(\nabla_\theta f_\theta^*(X'))Z']k(Z', Z)] = 0,
$$

Maximum Moment Restriction for Instrumental Variable Regression
where we note that \( \mathbb{E}[\nabla_{\theta}f_0(X) \mid Z']p(Z') = 0 \) for any \( Z' \) implied by the second function above. Similarly, the coefficient of \( \nabla_{\theta}f_0(X) \) must be zero, which implies that \( \mathbb{E}_{X,Y'}[|Y' - f_0(X')| \mid Z']p(Z') = 0 \) for any \( Z' \). The two coefficients cannot be zero at the same time (otherwise against the given conditions, so \( \mathbb{V}_{U'}[\mathbb{E}_{U'}[\nabla_{\theta}h_{U'}(U,U')] > 0 \). Further due to the given condition \( \mathbb{E}[\|\nabla_{\theta}h(U,U')\|^2] < \infty \), we obtain \( \sqrt{n}\nabla_{\theta}\hat{R}_{U,L}(f_0) \xrightarrow{p} N(0,4\mathbb{E}_{U'}[\mathbb{E}_{U'}[\nabla_{\theta}h_{U'}(U,U')]]) \) as per Serfling (1980, 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,

\[
\sqrt{n}\nabla_{\theta}\hat{R}_{U,L}(f_0) = \sqrt{n}\nabla_{\theta}\hat{R}_{U,L}(f_0) + \sqrt{n}\nabla_{\theta}\Omega(\theta^*) \xrightarrow{p} N(0,4\mathbb{E}_{U'}[\mathbb{E}_{U'}[\nabla_{\theta}h_{U'}(U,U')]])
\]

This concludes the proof.

**Uniform consistency of \( \nabla_{\theta}^{2}\hat{R}_{U,L}(f_0) \)**. Next, we consider the second derivative \( \nabla_{\theta}^{2}\hat{R}_{U,L}(f_0) \) and show its uniform consistency. In what follows, we denote by \( \| \cdot \|_F \) the Frobenius norm. We can express \( \nabla_{\theta}^{2}\hat{R}_{U,L}(f_0) \) as

\[
\nabla_{\theta}^{2}\hat{R}_{U,L}(f_0) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \nabla_{\theta}^{2}h_{0}(u_i, u_j) + \lambda \nabla_{\theta}^{2}\Omega(\theta)
\]

\[
\nabla_{\theta}^{2}h_{0}(u_i, u_j) = [\nabla_{\theta}f_{0}(x_i)\nabla_{\theta}f_{0}^{T}(x_j) - (y_i - f_{0}(x_i))\nabla_{\theta}^{2}f_{0}(x_j) + \nabla_{\theta}f_{0}(x_j)\nabla_{\theta}f_{0}^{T}(x_i) - (y_j - f_{0}(x_j))\nabla_{\theta}^{2}f_{0}(x_i)]k(z_i, z_j).
\]

**Lemma 5.** Suppose that \( f_{0} \) and \( \Theta(\theta) \) are twice continuously differentiable about \( \theta \), \( \Theta \) is compact, \( p(Y) \) has bounded support, \( \lambda \xrightarrow{p} 0 \) and Assumption 1 holds. Then, \( \mathbb{E}[\nabla_{\theta}^{2}h_{0}(U,U')] \) is continuous about \( \theta \) and

\[
\sup_{\theta \in \Theta} \left\| \nabla_{\theta}^{2}\hat{R}_{U,L}(f_0) - \mathbb{E}[\nabla_{\theta}^{2}h_{0}(U,U')] \right\|_F \xrightarrow{p} 0.
\]

**Proof.** The proof is similar to that of Lemma 3 and both applies extended results of Newey and McFadden (1994, Lemma 2.4). As \( (u_i, u_j) \neq (u_j, u_i) \) being strictly stationary in the sense of Newey and McFadden (1994, Footnote 18 in P.2129) has been shown in Lemma 3, we only need to show that there exists \( d(u,u') \geq |\nabla_{\theta}^{2}h_{0}(u,u')| \) for all \( u, u' \in X \times Y \times Z \) and \( \mathbb{E}[d(U,U')] < \infty \). As \( f_{0} \) is twice continuously differentiable about \( \theta \) and \( \Theta \) is compact, \( f_{0} \) is bounded. We can easily find a constant function \( a(x) = A \in \mathbb{R} \) such that \( |f_{0}(x)| < |a(x)| < \infty \). By the same condition, we have each entry of \( \nabla_{\theta}f_{0} \) bounded as \( |(\nabla_{\theta}f_{0}(x))_{i}| < b(x) < \infty \) where \( b(x) = B \in \mathbb{R} \), and each entry of \( \nabla_{\theta}^{2}f_{0}(x) \) bounded as \( |(\nabla_{\theta}^{2}f_{0}(x))_{ij}| < c(x) < \infty \) where \( c(x) = C \in \mathbb{R} \). Consequently, we exploit the triangle inequality of the Frobenius norm and obtain

\[
\left\| \nabla_{\theta}^{2}h_{0}(u,u') \right\|_F \leq [2|\nabla_{\theta}f_{0}(x)\nabla_{\theta}f_{0}^{T}(x')|_F + |y' + f_{0}(x')||\nabla_{\theta}^{2}f_{0}(x')|_F + |y' + f_{0}(x')||\nabla_{\theta}^{2}f_{0}(x')|_F]k(z, z') \leq [2mB^2 + 2(A + M)mCk(z, z')] \equiv d(u, u'),
\]

where we assume that \( |y| < M < \infty \) for all \( y \in Y \) in the second inequality because \( p(Y) \) has bounded support and that \( \theta \) is m-dimensional. Since \( k(z, z') \) is bounded as implied by Assumption 1, \( \mathbb{E}[d(U,U')] < \infty \) and \( \mathbb{E}[\nabla_{\theta}^{2}h_{0}(U,U')] \) is continuous about \( \theta \). Therefore, we obtain \( \sup_{\theta \in \Theta} \left\| \nabla_{\theta}^{2}\hat{R}_{U,L}(f_0) - \mathbb{E}[\nabla_{\theta}^{2}h_{0}(U,U')] \right\|_F \xrightarrow{p} 0 \) following from the extended results in the remarks of Newey and McFadden (1994, Lemma 2.4). Furthermore, from the conditions that \( \Theta(\theta) \) is twice continuously differentiable and the parameter space \( \Theta \) is compact, we obtain that \( \left\| \nabla_{\theta}^{2}\Omega(\theta) \right\|_F < \infty \) for any \( \theta \in \Theta \). Finally, it follows from the Slutsky’s theorem that

\[
\sup_{\theta \in \Theta} \left\| \nabla_{\theta}^{2}\hat{R}_{U,L}(f_0) - \mathbb{E}[\nabla_{\theta}^{2}h_{0}(U,U')] \right\|_F \leq \sup_{\theta \in \Theta} \left\| \nabla_{\theta}^{2}\hat{R}_{U,L}(f_0) - \mathbb{E}[\nabla_{\theta}^{2}h_{0}(U,U')] \right\|_F + \lambda \sup_{\theta \in \Theta} \left\| \nabla_{\theta}^{2}\Omega(\theta) \right\|_F \xrightarrow{p} 0.
\]

This concludes the proof.

**Theorem 9 (Asymptotic normality of \( \hat{\theta}_{U} \)).** Suppose that \( H = \mathbb{E}[\nabla_{\theta}^{2}h_{0}(U,U')] \) is non-singular, \( \Theta \) is convex and compact, \( p(Y) \) has bounded support, \( f_{0} \) and \( \Omega(\theta) \) are twice continuously differentiable about \( \theta \), \( \sqrt{n}\lambda \xrightarrow{p} 0 \), \( R_{k}(f_{0}) \) is uniquely minimized at \( \theta^* \) and \( \theta^* \) is an interior point of \( \Theta \), and Assumption 1 holds. Then

\[
\sqrt{n}(\hat{\theta}_{U} - \theta^*) \xrightarrow{p} N(0,4H^{-1}\mathbb{E}_{U'}[\mathbb{E}_{U'}[h_{\theta}^{2}(U,U')]])H^{-1}.
\]
The proof is the same as that of Theorem 9 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 \( \sup_{\theta \in \Theta} |\hat{R}_{U,\lambda}(f_\theta) - R_k(f_\theta)| \overset{p}{\to} 0 \) and that \( R_k(f_\theta) \) is continuous about \( \theta \) similarly to Lemma 3, and then \( \hat{\theta}_U \overset{p}{\to} \theta^* \) similarly to Theorem 8.

From the conditions that \( \Theta \) is compact, \( f_\theta \) is twice continuously differentiable about \( \theta \), \( p(Y) \) has bounded support and \( k(z, z') \) is bounded as implied by Assumption 1, we can obtain (ii) \( \hat{R}_{U,\lambda}(\theta) \) is twice continuously differentiable and \( \mathbb{E}[\|\nabla^2 h_{\theta^*} (U, U')\|^2] < \infty \) by a similar proof to that of Lemma 5. Given \( H = \mathbb{E}[\nabla^2 h_{\theta^*} (U, U')] = \nabla^2 R_k(\theta) \) is non-singular and \( R_k(f_\theta) \) is uniquely minimized at \( \theta^* \), we can obtain that the Hessian matrix \( H \) is positive definite because \( R_k(\theta) \) is convex as per Theorem 5,

\[
H = 2\mathbb{E}_{(X,Y,Z)\sim (X,Y,Z')} \left[ (\nabla_{\theta} f_{\theta^*}(X) \nabla_{\theta} f_{\theta^*}(X') - (Y - f_{\theta^*}(X)) \nabla^2_{\theta} f_{\theta^*}(X')) k(Z, Z') \right] > 0.
\]

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

Finally by Lemma 5, we know that \( H(\theta) = \mathbb{E}[\nabla^2 h_{\theta^*} (U, U')] \) and \( H(\theta^*) = H \), so (iv) and (v) are satisfied. Now, conditions of Newey and McFadden (1994, Theorem 3.1) are all satisfied, so we assert the conclusion.

### B.10 Proof of Theorem 3

We restate the notations

\[
\hat{R}_{V,\lambda}(f_\theta) := \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h_\theta(u_i, u_j) + \lambda \Omega(\theta) \quad \hat{R}_V(f_\theta) := \sum_{i=1}^{n} h_\theta(u_i, u_j) k(z_i, z_j)(y_j - f_\theta(x_j)).
\]

**Lemma 6.** Suppose that conditions of Lemma 4 hold. Then \( \sqrt{n}\nabla \hat{R}_{V,\lambda}(f_\theta) \overset{p}{\to} N(0, 4\mathbb{E}_U[\|\nabla^2 h_{\theta^*} (U, U')\|^2]) \).

**Proof.** As \( \mathbb{E}[\|\nabla^2 h_{\theta^*} (U, U')\|^2] < \infty \), \( \sqrt{n}\nabla \hat{R}_V(f_\theta) \) has the same limit distribution as that of \( \sqrt{n}\nabla \hat{R}_U(f_\theta) \) by Serfling (1980, Section 5.7.3). Furthermore, by \( \sqrt{n}\lambda \overset{p}{\to} 0 \) and \( \sqrt{n}\Omega(\theta^*) < \infty \) from that \( \Omega(\theta) \) is first continuously differentiable, we assert the conclusion by Slutsky’s theorem

\[
\sqrt{n}\nabla \hat{R}_{V,\lambda}(f_\theta) = \sqrt{n}\nabla \hat{R}_V + \sqrt{n}\lambda \nabla \theta \Omega(\theta^*) \overset{p}{\to} N(0, 4\mathbb{E}_U[\|\nabla^2 h_{\theta^*} (U, U')\|^2]).
\]

**Lemma 7.** Suppose that conditions of Lemma 5 hold. Then, \( \mathbb{E}[\nabla^2 h_{\theta^*} (U, U')] \) is continuous about \( \theta \) and \( \sup_{\theta \in \Theta} \|\nabla^2 \hat{R}_U(f_\theta) - \mathbb{E}[\nabla^2 h_{\theta^*} (U, U')]\|_F \overset{p}{\to} 0 \).

**Proof.** From conditions of Lemma 5, we can obtain that \( \|\nabla^2 h_{\theta^*} (u, u')\|_F \) and \( \|\nabla^2 h_{\theta^*} (u, u)\|_F \) are bounded, and \( \mathbb{E}[\nabla^2 h_{\theta^*} (U, U')] \) is continuous about \( \theta \) as proved in Lemma 5. By \( \|\nabla^2 h_{\theta^*} (u, u')\|_F \) and \( \|\nabla^2 h_{\theta^*} (u, u)\|_F \) are bounded, it follows that \( \nabla^2 h_{\theta^*} (u, u') \) is continuous about \( \theta \), \( \mathbb{E}[\sup_{\theta \in \Theta} \|\nabla^2 h_{\theta^*} (U, U')\|_F] < \infty \) and \( \mathbb{E}[\sup_{\theta \in \Theta} \|\nabla^2 h_{\theta^*} (U, U')\|_F] < \infty \). Then, we assert the conclusion by Newey and McFadden (1994, Lemma 8.5).

**Proof of Theorem 3.** The proof is the same as that of Theorem 9 except that \( \hat{R}_U \) is replaced by \( \hat{R}_V \).
Further, we define a functional associated with measure. We consider functional spaces where
\[ P \] such that
\[ \log \mathcal{N}(\varepsilon, H_1, \| \cdot \|_{L^2}) \leq C_H \varepsilon^{-s} \] for any \( \varepsilon \in (0, 1) \). If \( \lambda - \lambda_0 = o(n^{-1/2}) \) holds, then there exists a Gaussian process \( G^*_p \) such that
\[ \sqrt{n}(\hat{f}_U - f^*_U) \rightsquigarrow G^*_p \text{ in } H_1. \]

An exact covariance of \( G^*_p \) is described in the proof. The proof is based on the uniform convergence of U-processes on the function space (Arcones and Gine, 1993) and the functional delta method using the asymptotic expansion of the loss function (Hable, 2012). This asymptotic normality allows us to perform statistical inference, such as tests, even in the non-parametric case.

We discuss the boundedness and covering number assumptions in Theorem 4 and Theorem 10. For the boundedness assumption, many common kernels, such as the Gaussian RBF kernel, the Laplacian kernel, and the Mercer kernel, satisfy it. For the covering number, the common kernels above also satisfy it with a certain parameter configuration. For example, the Gaussian kernel (see Section 4 in Steinwart and Christmann (2008)) and the Mercer kernel (explained in Zhou (2002)).

To prove the theorem, we provide some notation. Let \( P \) be a probability measure which generates \( u = (x, y, z) \) and \( W = X \times Y \times Z \). Also, we define a function \( h_f(u, u') = (y - f(x))(y' - f(x'))k(z, z') \). Let \( \mathbb{H} := \{ h_f : W \times W \to \mathbb{R} | f \in H_1 \} \). For preparation, we define \( P^1 h_f : W \to \mathbb{R} \) 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 \( W \), we define a measure \( Q^2 := Q \otimes Q \) on \( W \times W \). Then, we can rewrite the U-statistic risk as
\[ \hat{R}_U(f) = \frac{(n - 2)!}{n!} \sum_{i=1}^{n} \sum_{j \neq i} h_f(u_i, u_j) =: U^2_n h_f, \]
where \( U_n \) is an empirical measure for the U-statistics. Similarly, we can rewrite the V-statistic risk as
\[ \hat{R}_V(f) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h_f(u_i, u_j) =: P^2_n h_f, \]
where \( P_n \) is an empirical measure of \( u \).

Further, we define a functional associated with measure. We consider functional spaces \( G_1 := \{ g : W \times W \to [0, 1] | \text{a convex set } \omega \text{ s.t. } g = 1_{\{ \cdot \leq \omega \}} \} \) and \( G_2 := \{ g : W \times W \to [0, 1] | \exists f, f' \in H_1, g(u, u') = h_f(u, \cdot)(f'(x) + f'(x')) \} \). Note that \( G_1 \) contains a functional which corresponds to the \( U^2_n \). Then, we consider a set of functionals
\[ B_2 := \left\{ F : G_1 \cup G_2 \to \mathbb{R} | \exists \text{non-zero finite } Q^2 \text{ s.t. } F(g) = \int g \, dQ^2, \quad \forall g \in G_1 \cup G_2 \right\}, \]
and also let \( B_0 \) be the closed linear span of \( B_2 \). For a functional \( F \in B_2 \), let \( \iota(F) \) be a measure which satisfies
\[ F(g) = \int g \, d\iota(F). \]

Uniform Central Limit Theorem: We firstly 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 11 (Theorem 4.4 in Arcones and Gine (1993)). 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
\[ \lim_{n \to \infty} \mathbb{E}[n^{-1/2} \log \mathcal{N}(n^{-1/2} \varepsilon, \mathbb{H}, \| \cdot \|_{L^2(U_2^n)})] = 0. \]
holds, for all \( \varepsilon > 0 \). Then, we obtain
\[
\sqrt{n}(U_n^2 - P^2) \xrightarrow{a.s.} 2G_{P^1}, \text{ in } \ell^\infty(\mathbb{H}).
\]

Here, \( G_{P^1} \) denotes a Brownian bridge, which is a Gaussian process 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 11. We firstly provide the following bound:

**Lemma 8.** For any \( f, f' \in \mathcal{H}_i \) such that \( \|f\|_{L^\infty} + \|f'\|_{L^\infty} \leq B \) 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 8.** We simply obtain the following:
\[
| h_f(u, u') - h_f(u', u') |
\]
\[
= |(y - f(x))k(z, z')(y' - f(x')) - (y - f'(x))k(z, z')(y' - f'(x'))|
\]
\[
= |k(z, z')(y' - f(x)) - (y - f'(x))(y' - f'(x'))|
\]
\[
= |k(z, z')(y'f(x) - f(x)) + y(f'(x') - f(x')) + f(x)f'(x') - f'(x)f'(x')|
\]
\[
\leq |k(z, z')(u'|f'(x') - f(x)| + |y||f'(x') - f(x')| + |f(x)||f'(x') - f'(x')| - |f'(x)||f'(x') - f'(x')|
\]
\[
\leq 4B |k(z, z')| \| f - f' \|_{L^\infty},
\]
as required. \( \square \)

**Lemma 9.** Suppose the assumptions of Theorem 4 hold. Then, the followings hold:

1. \( \{ P^1 h_f \mid h_f \in \mathbb{H} \} \) is a Donsker class.
2. For any \( \varepsilon > 0 \), the following holds:
\[
\lim_{n \to \infty} E[n^{-1/2} \log \mathcal{N}(n^{-1/2}\varepsilon, \mathbb{H}, \cdot \| L^2(\mathcal{U}_n^2))] = 0.
\]

**Proof of Lemma 9.** For preparation, fix \( \varepsilon > 0 \) and set \( N = \mathcal{N}(\varepsilon, \mathcal{H}_i, \cdot \| L^\infty) \). Also, let \( Q \) be an arbitrary finite discrete measure. Then, by the definition of a bracketing number, there exist \( N \) functions \( \{ f_i \in \mathcal{H}_i \}_{i=1}^N \) such that for any \( f \in \mathcal{H}_i \) there exists \( i \in \{1, 2, ..., N\} \) such as \( \| f - f_i \|_{L^\infty} \leq \varepsilon \).

For the first condition, as shown in Equation (2.17) in Van der Vaart and Wellner (1996), it is sufficient to show
\[
\sup_Q \log \mathcal{N}(\varepsilon, \{ P^1 h_f \mid h_f \in \mathbb{H}, \cdot \| L^2(\mathcal{U}_n)\} \leq c \varepsilon^{-2\delta} (\varepsilon \to 0),
\]
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 \( \log \mathcal{N}(\varepsilon, \{ P^1 h_f \mid h_f \in \mathbb{H}, \cdot \| L^2(\mathcal{Q})\} \leq c' \log \mathcal{N}(\varepsilon, \mathcal{H}_i, \cdot \| L^\infty) \) with a constant \( c' > 0 \). Fix \( P^1 h_f \in \{ P^1 h_f \mid h_f \in \mathbb{H} \} \) arbitrary, and set \( f_i \) which satisfies \( \| f - f_i \|_{L^2(\mathcal{Q})} \leq \varepsilon \). Then, we have
\[
\| P^1 h_f - P^1 h_{f_i} \|_{L^2(\mathcal{Q})}^2 = \int \left\{ \int (h_f(u, u') + h_{f_i}(u', u))/2 - (h_{f_i}(u, u') + h_f(u', u))/2 \right\} dP(u')^2 dQ(u)
\]
\[
= \int \left( \int h_f(u, u') - h_{f_i}(u, u') dP(u') \right)^2 dQ(u)
\]
\[
\leq C \int \left( \int |k(z, z')| dP(u') \right)^2 dQ(u) \| f - f_i \|_{L^\infty}^2
\]
For the second condition, we have the similar strategy. For any first statement.

Second inequality follows the bounded condition of \( k \) with constants \( C, C' \).

\[ \| f - f_i \|_{L^\infty} \leq C' \varepsilon \]

with constants \( C, C' > 0 \). The first inequality follows Lemma 8 with the bounded property of \( f, f' \) and \( \mathcal{Y} \). The second inequality follows the bounded condition of \( k \) in Theorem 10. 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

\[ \| h_f - h_{f_i} \|_{L^1(U^2)} = \int | h_f(u, u') - h_{f_i}(u, u') | \, dU^2(u, u') \leq C'' \| f - f_i \|_{L^\infty} \leq C'' \varepsilon, \]

with a constant \( C'' > 0 \). Hence, we have

\[ \mathbb{E}[n^{-1/2} \log \mathcal{N}(n^{-1/2} \varepsilon, \mathbb{H}, \| \cdot \|_{L^1(U^2)})] \leq n^{-1/2} \log \mathcal{N}(n^{-1/2} \varepsilon, \mathcal{H}_l, \| \cdot \|_{L^\infty}) \]

\[ \leq Cn^{-1/2} \left( \frac{n^{1/2}}{\varepsilon} \right)^s = Cn^{(s-1)/2} \rightarrow 0, \quad (n \rightarrow \infty), \]

since \( s \in (0, 1) \).

From Theorem 11 and Lemma 9, we rewrite the central limit theorem utilizing terms of functionals. Note that \( \iota^{-1}(U^2), \iota^{-1}(P^2) \in B_S \) holds. Then, we can obtain

\[ \sqrt{n}(\iota^{-1}(U^2_n) - \iota^{-1}(P^2)) \Rightarrow 2G_{P_2} \text{ in } \ell^\infty(\mathbb{H}). \]

**Learning Map and Functional Delta Method:** We consider a learning map \( S : B_S \rightarrow \mathcal{H}_l \). For a functional \( F \in B_S \), we define

\[ S_\lambda(F) := \arg \min_{f \in \mathcal{H}_l} \iota(F) h_f + \lambda \| f \|_{\mathcal{H}_l}^2. \]

Obviously, we have

\[ \hat{f} = S_\lambda(\iota^{-1}(U^2_n)) \text{, and } f^*_l = S_{\lambda_0}(\iota^{-1}(P^2)). \]

We consider a derivative of \( S_\lambda \) in the sense of the Gateau differentiation by the following steps.

Firstly, we define a partial derivative of the map \( R_{Q^2}(f) \). To investigate the optimality of the minimizer of

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

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

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

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

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

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

**Lemma 10.** If the assumptions in Theorem 4 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 10.** We consider a sequence of functions \( h_n \in \mathcal{H}_l \) for \( n \in \mathbb{N} \), such that \( h_n(x) \neq 0, \forall x \in X \) and \( \| h_n \|_{L^\infty} \rightarrow 0 \) as \( n \rightarrow \infty \). Then, for \( f \in \mathcal{H}_l \), a simple calculation yields

\[ \left| \frac{R_{Q^2, \lambda}(f + h_n) - R_{Q^2, \lambda}(f) - \nabla R_{Q^2, \lambda}[f](h_n)}{\| h_n \|_{L^\infty}} \right| \]
Next, we apply the implicit function theorem to obtain an explicit formula of the derivative of $S$. This equivalence follows Theorem 7.4.1 and Lemma 8.7.1 in Luenberger (1997).

Here, we consider its RKHS-type formulation of $\nabla R_{Q^2,\lambda}$, which is convenient to describe a minimizer. Let $\Phi_1 : X \to H_1$ be the feature map associated with the RKHS $H_1$, such that $\langle \Phi_1[x], f \rangle_{H_1} = f(x)$ for any $x \in X$ and $f \in H_1$. Let $\nabla \tilde{R}_{Q^2,\lambda} : H_1 \to H_1$ be an operator such that

$$\nabla \tilde{R}_{Q^2,\lambda}(f) := 2\lambda f + \int \partial_{f,1} h f(u, u') \Phi_1[x](-) + \partial_{f,2} h f(u, u') \Phi_1[x'](-) dQ^2(u, u').$$

Obviously, $\nabla R_{Q^2,\lambda}(f) = \langle \nabla \tilde{R}_{Q^2,\lambda}(f), \cdot \rangle_{H_1}$. Now, we can describe the first-order condition of the minimizer of the risk. Namely, we can state that

$$\hat{f} = \arg\min_{f \in H_1} R_{Q^2,\lambda}(f) \Leftrightarrow \nabla \tilde{R}_{Q^2,\lambda}(\hat{f}) = 0.$$ 

This equivalence follows Theorem 7.4.1 and Lemma 8.7.1 in Luenberger (1997).

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 $\nabla^2 \tilde{R}_{Q^2,\lambda} : H_1 \to H_1$ as

$$\nabla^2 \tilde{R}_{Q^2,\lambda}(f) := 2\lambda f + \int k(z, z') f(x) \Phi_1[x](-) + f(x') \Phi_1[x'](-) dQ^2(u, u'),$$

which follows (b) in Lemma A.2 in Hable (2012). Its basic properties are provided in the following result:

**Lemma 11.** If Assumption 1 and the assumptions in Theorem 4 hold, then $\nabla^2 \tilde{R}_{Q^2,\lambda}$ is a continuous linear operator and it is invertible.

**Proof of Lemma 11.** By (b) in Lemma A.2 in Hable (2012), $\nabla^2 \tilde{R}_{Q^2,\lambda}$ is a continuous linear operator. In the following, we define $A : H_1 \to H_1$ as $A(f) = \int k(z, z') f(x) \Phi_1[x](-) + f(x') \Phi_1[x'](-) dQ^2(u, u')$. To show $\nabla^2 \tilde{R}_{Q^2,\lambda}$ 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 non-zero $f \in H_1$ and obtain

$$\|\nabla^2 \tilde{R}_{Q^2,\lambda}(f)\|^2_H = (2\lambda f + A(f), 2\lambda f + A(f))_{H_1}$$

$$= 4\lambda^2 \|f\|^2_{H_1} + 4\lambda \langle f, A(f) \rangle_{H_1} + \|A(f)\|^2_{H_1}$$

$$> 4\lambda \langle f, A(f) \rangle_{H_1}$$

$$= 4\lambda \left( f, \int k(z, z') f(x) \Phi_1[x] dQ^2(u, u') \right)_{H_1} + 4\lambda \left( f, \int k(z, z') f(x') \Phi_1[x'] dQ^2(u, u') \right)_{H_1}$$

$$= 4\lambda \int k(z, z') f(x)^2 dQ^2(u, u') + 4\lambda \int k(z, z') f(x')^2 dQ^2(u, u')$$

$$\geq 0.$$ 

The last equality follows the property of $\Phi_1$ and the last inequality follows the ISPD property in Assumption 1.

For the compactness, we follow Lemma A.5 in Hable (2012) and obtain that operators ($f \mapsto \int k(z, z') f(x) \Phi_1[x](-) dQ^2(u, u')$) and ($f \mapsto \int k(z, z') f(x') \Phi_1[x'](-) dQ^2(u, u')$) 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

$$\nabla S_{Q^2,\lambda}(F') := -\nabla^2 \hat{R}_{Q^2,\lambda}^{-1} \left( \int \partial_{f,1} h_f(u, u') \Phi_l[x](\cdot) + \partial_{f,2} h_f(u, u') \Phi_l[x'](\cdot) \, du(F')(u, u') \right),$$

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

**Proposition 1.** Suppose the assumptions in Theorem 4 hold. For $F \in B_S$, $F' \in \ell^\infty(\mathcal{G}_1 \cup \mathcal{G}_2)$, and $s \in \mathbb{R}$ such that $F + sF' \in B_S$, $\nabla S_{\ell(F),\lambda}(F')$ is a Gateau-derivative of $S_{\lambda}$, namely,

$$\lim_{s \to 0} \left\| \frac{S_{\lambda}(F + sF') - S_{\lambda}(F)}{s} - \nabla S_{\ell(F),\lambda}(F') \right\|_{\mathcal{H}_1} = 0.$$

**Proof of Proposition 1.** This proof has the following two steps, (i) define a proxy operator $\Gamma$, then (ii) prove the statement by the implicit function theorem.

**(i) Define $\Gamma$:** Note that $\ell(F')$ exists since $F + sF' \in B_S$ implies $F' \in B_0$. We define the following operator $\Gamma(s, f, \lambda) : \mathcal{H}_1 \to \mathcal{H}_1$ for $f \in \mathcal{H}_1$:

$$\Gamma(s, f, \lambda) := \nabla \hat{R}_{\ell(F) + st(F'),\lambda}$$

$$= 2\lambda f + \int \partial_{f,1} h_f(u, u') \Phi_l[x](\cdot) + \partial_{f,2} h_f(u, u') \Phi_l[x'](\cdot) \, du(F)(u, u')$$

$$+ s \int \partial_{f,1} h_f(u, u') \Phi_l[x](\cdot) + \partial_{f,2} h_f(u, u') \Phi_l[x'](\cdot) \, du(F')(u, u').$$

For simple derivatives, Lemma A.2 in Hable (2012) provides

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

and

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

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

$$\Gamma(s, f, \lambda) = 0 \Leftrightarrow f = S_{\lambda}(\ell(F) + st(F')).$$

Also, we obtain

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

Then, for each $\lambda > 0$, by the implicit function theorem, there exists a smooth map $\varphi_{\lambda} : \mathbb{R} \to \mathcal{H}_1$ such that

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

and it satisfies

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

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

$$\lim_{s \to 0} \left\| \frac{S_{\lambda}(F + sF') - S_{\lambda}(F')}{s} - \nabla S_{\ell(F),\lambda}(F') \right\|_{\mathcal{H}_1} = \lim_{s \to 0} \left\| \frac{\varphi_{\lambda}(s) - \varphi_{\lambda}(0)}{s} - \nabla_s \varphi_{\lambda}(0) \right\|_{\mathcal{H}_1} = 0.$$

Then, we obtain the statement.

Now, we are ready to prove Theorem 4 and Theorem 10.
Proof of Theorem 10. As a preparation, we mention that $S_{\lambda}$ is differentiable in the Hadamard sense, which is Gateau differentiable by Proposition 1. Lemma A.7 and A.8 in Hable (2012) show that $\nabla S_{(F),\lambda,\mu(G)}$ is Hadamard-differentiable for any $\lambda$, $G$ and $F$.

Then, we apply the functional delta method. As shown in Theorem 11 and Lemma 9, we have

$$\sqrt{n}(\ell^{-1}(U^2_n) - \ell^{-1}(P^2)) \rightsquigarrow 2G_{P1}.$$

Hence, we obtain

$$\sqrt{n}(\ell^{-1}(U^2_n) - \ell^{-1}(P^2)) = \frac{\lambda_0}{\lambda} \sqrt{n}(\ell^{-1}(U^2_n) - \ell^{-1}(P^2)) + \frac{\sqrt{n}(\lambda - \lambda_0)}{\lambda} \rightsquigarrow 2G_{P1},$$

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

$$\sqrt{n}(\tilde{f} - f^*_n) = \sqrt{n}(S_{\lambda_0}(\ell^{-1}(U^2_n)) - S_{\lambda_0}(\ell^{-1}(P^2))) + o_P(1) \rightsquigarrow \nabla S_{P^2,\lambda_0}(2G_{P1}),$$

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

Proof of Theorem 4. This proof is completed by substituting the Central Limit Theorem part (Theorem 11) of the proof of Theorem 10. From Section 3 in Akritas et al. (1986), the V- and U- processes have the same limit distribution asymptotically, so the same result holds.

C GAUSSIAN PROCESS (GP) INTERPRETATION

We present the close connection between the non-parametric model and the Gaussian process (GP) in this section. We will show that the RKHS solution of our objective function (7) is equivalent to the maximum function of the posterior distribution of the GP. The relationship is inspired by the similarity of our objective function to that of GLS, and it will be used to derive an efficient cross validation error.

By Mercer theorem, the kernel $l(x,x')$ can be expanded as $l(x,x') = \sum_{i=1}^{N} \lambda_i \phi_i(x)\phi_i(x')$, where $\phi_i(x)$ are orthonormal in $L^2(\mathbb{R}^d)$, a space of square-integrable real-valued functions, and $N < \infty$ for degenerate kernels and $N = \infty$ otherwise. To satisfy for arbitrary $f(x) = \sum_{i=1}^{N} \alpha_i \phi_i(x)$ the reproducing property

$$\langle \sum_{i=1}^{N} \alpha_i \phi_i, l(\cdot, \cdot) \rangle_{\mathcal{H}^l} = f(x),$$

we choose $\phi_i(x)$ such that $\phi_i$ is orthogonal in $\mathcal{H}^l$ as the Mercer expansion is not unique. Based on the reproducing property, we can obtain $\langle \phi_i, \phi_j \rangle_{\mathcal{H}^l} = \delta_{ij} \lambda_i^{-1}$ where $\delta_{ij} = 1$, if $i = j$, and zero otherwise. Similar settings can be found in Walder and Bishop (2017), for example.

Let us consider a GP over a space of functions

$$f(x) = \Phi(x)w,$$

where $\Phi(x) = [\phi_1(x), \phi_2(x), \ldots, \phi_N(x)]$ is the feature vector and $w$ is the parameter vector. We assume a prior distribution of $w \sim N(0, \delta I)^2$ where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ is a diagonal matrix with eigenvalues $\lambda_1, \ldots, \lambda_N$, and $\delta > 0$ is a hyper-parameter, which plays the same role as the regularization hyper-parameter (as we show later). This definition is equivalent to assuming that on a set of input $x$, $f(x) \sim GP(0, \delta l(x, x))$ because

$$E[f(x)] = \Phi(x)E[w] = 0,$$

$$\text{Var}[f(x)] = \delta \Phi(x)\Lambda\Phi(x)^T = \delta l(x, x).$$

We refer the interested readers to Rasmussen and Williams (2005) for further details on Gaussian process models.

2Throughout the paper, we denote by $N(\mu, \sigma^2)$ a Gaussian distribution with the mean $\mu$ and variance $\sigma^2$. 


### C.1 Likelihood Function

Given the prior distribution on \( w \), we aim to characterize the posterior distribution \( p(w | D) \propto p(D | w)p(w) \) where \( D = \{(x_i, y_i, z_i)\}_{i=1}^n \) is an i.i.d. sample of size \( n \), and \( f(x) = \Phi(x)w \) where the \( i \)'s row of \( \Phi(x) \) contains the feature vector of \( x_i \), namely, \( \Phi(x_i)^\top \). To define the likelihood \( p(D | w) \), 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

\[
p((x, y, z), (x', y', z') | w) \propto \exp \left[ -\frac{1}{2} (y - \Phi(x)w)k(z, z')(y' - \Phi(x')w) \right]
\]

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

Hence, the likelihood on both \( D \) and \( D' \) can be expressed as

\[
p((D, D') | w) \propto \prod_{i=1}^n \prod_{j=1}^n p((x_i, y_i, z_i), (x'_j, y'_j, z'_j) | w)
\]

\[= \exp \left[ -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (y_i - f(x_i))k(z_i, z_j)(y_j - f(x_j)) \right]. \tag{28}
\]

In practice, however, we only have access to a single copy of 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 used to form \( D' \). Unfortunately, this approach reduces the effective sample size that can be used for learning. Alternatively, we propose to estimate the original likelihood (28) by using \( M \)-estimators: given the full dataset \( D = \{(x_i, y_i, z_i)\}_{i=1}^n \), our approximated likelihood can be defined as

\[
p(D | w) \propto \prod_{i=1}^n \prod_{j=1}^n p((x_i, y_i, z_i), (x_j, y_j, z_j) | w)
\]

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

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

\[= (2\pi)^{-n/2} |K_z|^{-1/2} N(f(x) | y, K_z^{-1}).
\]

The matrix \( K_z \), as a kernel matrix defined above, plays a role of correlating the residuals \( y_i - f(x_i) \). Besides, the standard GP regression has a similar likelihood, which is Gaussian but the covariance matrix is computed on \( x \) rather than \( z \). Based on the likelihood \( p(D | w) \), the maximum likelihood (ML) estimator hence coincides with the minimizer of the unregularized version of our objective (7).

### C.2 Maximum A Posteriori (MAP)

Combining the above likelihood function with the prior on \( w \), the maximum a posteriori (MAP) estimate of \( w \) can be obtained by solving

\[
\arg\max_w \log p(w | D) = \arg\max_w \left[ \log p(D | w) + \log p(w) - \log p(D) \right]
\]

\[= \arg\max_w \left[ -\frac{1}{2} (y - \Phi(x)w)^\top K_z (y - \Phi(x)w)
\]

\[\quad - \frac{n}{2} \log 2\pi - \frac{1}{2} \log |\Delta| - \frac{1}{2} w^\top (\Delta)^{-1}w - \log p(D) \right]
\]

\[= \arg\max_w \left[ -\frac{1}{2} (y - \Phi(x)w)^\top K_z (y - \Phi(x)w) - \frac{1}{2} w^\top (\Delta)^{-1}w \right]. \tag{33}
\]
Specifically, setting the first derivative of (33) to zero yields the first-order condition

$$-(\delta \Lambda)^{-1} \hat{w} - \Phi(x)^\top K_z \Phi(x) \hat{w} + \Phi(x)^\top K_z y = 0$$

and the MAP estimate

$$\hat{w} = \delta \Lambda \Phi(x)^\top K_z (y - \Phi(x)^\top \hat{w}). \quad (34)$$

Note that we express $\hat{w}$ using the implicit expression. Given a test data point $x^*$, the corresponding prediction can be obtained by computing the posterior expectation, i.e.,

$$\hat{f}(x^*) \equiv E_{f|D}[f(x^*)] = \Phi(x^*) \hat{w} = \Phi(x^*) (\delta \Lambda) \Phi(x)^\top K_z (y - \Phi(x)^\top \hat{w}) \equiv l(x^*, x) \hat{\alpha},$$

where we define $\hat{\alpha} \equiv \delta K_z (y - \Phi(x)^\top \hat{w})$. The second equality holds because $\hat{w}$ is the posterior mean. The third equality is obtained by substituting (34) into the second equation. Finally, the last equation follows from the Mercer theorem: $\Phi(x') \Lambda \Phi(x)^\top = l(x', x)$.

We can see that without using the representer theorem here, we get the same expression of the optimal solution $\hat{f}$ as the one we obtain by invoking the representer theorem in Section 3.2. By substituting $\hat{\alpha} = \delta K_z (y - \Phi(x)^\top \hat{w})$ back into (34), we obtain

$$\hat{w} = \Lambda \Phi(x)^\top \hat{\alpha}. \quad (35)$$

Furthermore, putting (35) into (33) also yields the following optimization problem:

$$\hat{\alpha} = \arg \min_{\alpha} \left[ \frac{1}{\delta} - 1 \alpha^\top L \alpha + (y - L \alpha)^\top K_z (y - L \alpha) \right]$$

$$= \arg \min_{\alpha} \left[ (n^2 \delta)^{-1} \alpha^\top L \alpha + (y - L \alpha)^\top W (y - L \alpha) \right],$$

where $L = l(x, x)$ is the kernel matrix. This problem has the same form as that of the V-statistic objective (8) except that $(n^2 \delta)^{-1}$ replaces $\lambda$. Therefore, we conclude that the predictive mean is equivalent to the optimal function of the V-statistic objective (8) given that $(n^2 \delta)^{-1} = \lambda$, and $\delta$ plays the role of the regularization parameter. Such a GP interpretation is used for an elegant derivation of the efficient cross validation error in Section 3.3.

**D ADDITIONAL DETAILS OF EXPERIMENTS**

In this section, we provide additional details about the experiments as well as more discussions on the experimental results presented in Section 6.

**D.1 Baseline Algorithms**

The details of MMR-IV (Nyström) and MMR-IV (NN) algorithms are given in Algorithm 2 and Algorithm 3, respectively.

In the experiments, we compare our algorithms to the following baseline algorithms:

- **DirectNN**: A standard least square regression on $X$ and $Y$ using a neural network (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.
- **DeepIV** (Hartford et al., 2017): A nonlinear extension of 2SLS using deep NNs. We use the implementation available at https://github.com/microsoft/EconML.
We combine the training and validation sets to perform leave-2-out CV to select parameters of the kernel. In contrast to our NN-based method, the RKHS-based method in Algorithm 2 has the analytic form of CV error. So we use them without changes. We fix the random seed to 527 for all data generation and model initialization. Parameter selection of baselines has been tuned reasonably well in the package provided by Bennett et al. (2019), for fair comparisons with the baselines. These hyper-parameters are provided in Appendix D.4. Besides, the parameter of the above baselines are provided in the package except KernelIV.

Algorithm 2 MMR-IV (Nyström)

Input: Dataset \( D = \{ (x, y, z) \} \), kernel functions \( k \) and \( l \) with parameters \( \theta_k \) and \( \theta_l \), regularization parameter \( \lambda \), leave out data size \( M \), Nyström approximation sample size \( M' \), LMOCV times \( m \)

Output: predictive value at \( x \)∗
1: \( K = k(z, z; \theta_k) \)
2: \( \tilde{\delta}, \tilde{\theta}_l = \arg\min_{\delta, \theta_l} \sum_{i=1}^{m} (b^{(i)} - y_M^{(i)})^\top K_M^{(i)} (b^{(i)} - y_M^{(i)}) \)
3: Equation (9) Nyström Approx with \( M' \) samples
4: \( K/n^2 \approx \tilde{U} \tilde{V} \tilde{U}^\top \)
5: \( \tilde{\alpha} = \lambda^{-1} [I - \tilde{U} (\lambda^{-1} \tilde{U} ^\top \tilde{L} \tilde{U} + \tilde{V}^{-1})^{-1} \tilde{U} ^\top \lambda^{-1} \tilde{L}] \tilde{U} \tilde{V} \tilde{U}^\top y \)
6: return \( \hat{f}(x) = l(x, x; \tilde{\theta}_l) \tilde{\alpha} \)

Algorithm 3 MMR-IV (NN)

Input: Dataset \( D = \{ x, y, z \} \), kernel function \( k \) with parameters \( \theta_k \), NN \( f_{NN} \) with parameters \( \theta_{NN} \), regularization parameter \( \lambda \)

Output: predictive value at \( x \)∗
1: Compute \( K = k(z, z; \theta_k) \)
2: \( \hat{f}_{NN} = \arg\min_{f_{NN}} (y - f_{NN}(x))^\top K(y - f_{NN}(x))/n^2 + \lambda \| \theta_{NN} \|^2_2 \)
3: return \( \hat{f}_{NN}(x) \)

- KernelIV (Singh et al., 2019): 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/r4hul-5in9h/KIV.
- GMM+NN: An optimally-weighted GMM (Hansen, 1982) is combined with a NN \( f(X) \). The details of this algorithm can be found in Bennett et al. (2019, Section 5).
- AGMM (Lewis and Syrgkanis, 2018): 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 (Bennett et al., 2019): 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 above baselines are provided in the package except KernelIV.

D.2 Experimental Settings on Low-dimensional Scenario

For the experiments in Section 6.1, we consider both small-sample \( (n = 200) \) and large-sample \( (n = 2000) \) regimes, in which \( n \) points are sampled for training, validation and test sets, respectively. In both regimes, we standardize the values of \( Y \) to have zero mean and unit variance for numerical stability. Hyper-parameters of NNs in Algorithm 3, including the learning rate and the regularization parameter, are chosen by 2-fold CV for fair comparisons with the baselines. These hyper-parameters are provided in Appendix D.4. Besides, the parameter selection of baselines has been tuned reasonably well in the package provided by Bennett et al. (2019), so we use them 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 2 has the analytic form of CV error. We combine the training and validation sets to perform leave-2-out CV to select parameters of the kernel \( l \) and the regularizer parameter \( \lambda \). We choose the sum of Gaussian kernels for \( k \) and 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.

Detailed comments on the experimental results. 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.
For the kernel function on instruments, we employ the sum of Gaussian kernels.

D.4 More Details of MMR-IVs

DirectNN suspects that hard-to-optimize objective and complicated optimization procedure are the causes. Compared with enough. In addition, we observe that DeepGMM fails because the features from PCA are not representative information. Furthermore, MMR-IV (NN) performs competitively across scenarios. On MNIST, MMR-IVs perform better than other methods, which implies using the sum of Gaussian kernels for the kernel $k$ is proper. DeepGMM has competitive performance as well. On MNIST and MNISTXZ, MMR-IV (NN) outperforms all other methods. 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 hard-to-optimize objective and complicated optimization procedure are the causes. Compared with DirectNN, most of baselines can hardly deal with high-dimensional structured instrumental regressions.

D.3 Experimental Settings on High-dimensional Scenario

For experiments in Section 6.2, 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 in this experiment since its kernel parameter selection is not suitable for image data.

Detailed comments on the experimental results. Like Bennett et al. (2019), we observe that the AGMM code crashes, and the DeepIV code returns NaN when $X$ is high-dimensional. Besides, we run Ridge2SLS, which is Poly2SLS with fixed linear degree, in place of Poly2SLS to reduce computational workload. 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 is insufficient to extract useful information. Furthermore, MMR-IV (NN) performs competitively across scenarios. On MNIST, MMR-IVs perform better than other methods, which implies using the sum of Gaussian kernels for the kernel $k$ is proper. DeepGMM has competitive performance as well. On MNISTX and MNISTXZ, MMR-IV (NN) outperforms all other methods. 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 hard-to-optimize objective and complicated optimization procedure are the causes. Compared with DirectNN, most of baselines can hardly deal with high-dimensional structured instrumental regressions.

D.4 More Details of MMR-IVs

For the kernel function on instruments, we employ the sum of Gaussian kernels

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

and the Gaussian kernel for $l(x, x') = \exp(-\|x - x'\|_2^2/(2\sigma_l^2))$, where $\sigma_{k_1}$ is chosen as the median interpoint distance of $z = \{z_i\}_{i=1}^{n}$ and $\sigma_{k_2} = 0.1\sigma_{k_1}$, $\sigma_{k_3} = 10\sigma_{k_1}$. 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.

Table 3: The mean square error (MSE) ± one standard deviation in the small-sample regime ($n = 200$).

| Algorithm          | Function $f^*$ |
|--------------------|-----------------|
|                    | abs | linear | sin | step |
| DirectNN           | .143 ± .000    | .046 ± .000 | .404 ± .006 | .253 ± .000 |
| 2SLS               | .654 ± .000    | .003 ± .000 | .304 ± .000 | .076 ± .000 |
| Poly2SLS           | .125 ± .000    | .003 ± .000 | .164 ± .000 | .077 ± .000 |
| GMM+NN             | .792 ± .000    | .203 ± .000 | 1.560 ± .001 | .550 ± .000 |
| AGM-MM             | .031 ± .000    | .011 ± .000 | .330 ± .000 | .080 ± .000 |
| DeepIV             | .204 ± .008    | .047 ± .004 | .197 ± .004 | .039 ± .001 |
| DeepGMM            | .022 ± .003    | .032 ± .016 | .143 ± .030 | .039 ± .002 |
| KernelIV           | .063 ± .000    | .024 ± .000 | .086 ± .000 | .055 ± .000 |
| MMR-IV (NN)        | .019 ± .003    | .004 ± .001 | .292 ± .024 | .075 ± .008 |
| MMR-IV (RKHS)      | .030 ± .000    | .011 ± .000 | .075 ± .000 | .057 ± .000 |
As per the dimensions of $X$, we parametrize $f$ either as a fully connected neural network with leaky ReLU activations and 2 hidden layers, each of which has 100 cells, for non-image data, or a deep convolutional neural network (CNN) architecture for MNIST data. We denote the fully connected neural network 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 4.

### D.5 Additional Comments on Mendelian Randomization

For experiments in Section 6.3, 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 presents significant instability with the number of instruments in Figure 3(left) because of failure of the hyper-parameter selection. In Figure 3(middle) and Figure 3(right), we can observe that the performance of most approaches deteriorates as the effect of confounders becomes stronger. MMR-IV (Nyström) is more sensitive to $c_2$ than $c_1$.

### D.6 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 instrument and (ii) with instruments. By without instrument, we mean that the $W_V$ matrix of MMR-IV (Nyström) becomes an identity matrix. Following Sjolander and Martinussen (2019), we assess the effect of Vitamin D (exposure) on mortality rate (outcome), control the age in the analyses, and use filagrin as the instrument. We illustrate original Vitamin D, age and death in Figure 5. We randomly pick (random seed is 527) 300 Nyström samples and use leave-2-out cross validation to select hyper-parameters.

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

$$Y = f(X,C) + \epsilon, \quad X = t(Z,C) + g(\epsilon) + \nu$$

where $\mathbb{E}[\epsilon] = 0$ 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., $Y \perp \perp Z|X,\epsilon,C$;

(iii) **Unconfounded instrument(s):** $Z$ is conditionally independent of the error, i.e., $\epsilon \perp \perp Z|C$.

Unlike the conditions specified in the main text, (ii) and (iii) also include the controlled variable $C$. A similar model is employed in Hartford et al. (2017). From Assumption (iii), we can see that $\mathbb{E}[\epsilon|C,Z] = \mathbb{E}[\epsilon|C]$, and based on this, we further obtain

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

for any measurable function $h$. Note that $\mathbb{E}[\epsilon|C]$ is only conditioned on $C$, remains constant on arbitrary values of $C$, and is typically non-zero. 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 $f'(X,C) := f(X,C) + \mathbb{E}[\epsilon|C]$ instead of just $f(X,C)$.
Figure 4: Estimated effect of vitamin D level on mortality rate, controlled by age. Both plots depict normalized contours of $\hat{f}$ defined in Section D.6 where blue represents low mortality rate and yellow the opposite. We can divide each plot roughly into left (young age) and right (old age) parts. While the right parts reflect similar information (i.e., lower vitamin D level at an old age leads to higher mortality rate), the left parts are different. In (a), a high level of vitamin D at a young age can result in a high mortality rate, which is counter-intuitive. A plausible explanation is that it is caused by some unobserved confounders between vitamin D level and mortality rate. In (b), on the other hand, this spurious effect disappears when the filaggrin mutation is used as 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 (Meehan and Penckofer, 2014). This comparison demonstrates the benefit of an instrument variable.

Figure 5: Plot of Vitamin D data.