

The variational method of moments

Andrew Bennett and Nathan Kallus

Cornell University, New York, NY, USA

Address for correspondence: Nathan Kallus, Cornell University, Bloomberg Center, Room 4552, W Loop Road, New York, NY 10044, USA. Email: kallus@cornell.edu

Abstract

The conditional moment problem is a powerful formulation for describing structural causal parameters in terms of observables, a prominent example being instrumental variable regression. We introduce a very general class of estimators called the *variational method of moments* (VMM), motivated by a variational minimax reformulation of optimally weighted generalized method of moments for finite sets of moments. VMM controls infinitely for many moments characterized by flexible function classes such as neural nets and kernel methods, while provably maintaining statistical efficiency unlike existing related minimax estimators. We also develop inference algorithms and demonstrate the empirical strengths of VMM estimation and inference in experiments.

Keywords: adversarial machine learning, conditional moment problem, inference, instrumental variable regression, kernel methods, structural estimation

1 Introduction

For many problems in fields such as economics, sociology, or epidemiology, we seek to use observational data to estimate *structural parameters*, which often describe some causal relationship. A common framework that unifies many such problems is the *conditional moment problem*, which assumes that the parameter of interest θ_0 is the unique element of some parameter space Θ such that

$$\mathbb{E}[\rho(X; \theta_0) | Z] = 0, \quad (1)$$

where $X \in \mathcal{X}$ denotes the observed data, $Z \in \mathcal{Z}$ is a random variable that is measurable with respect to X , and $\rho: \mathcal{X} \rightarrow \mathbb{R}^m$ is a vector-valued function indexed by Θ . Note that Equation (1) is an identity of *random variables*, not of numbers; that is, it holds almost surely with respect to the random Z . That Z is measurable with respect to X is without loss of generality, since given any \tilde{X} , Z we may define $X = (\tilde{X}, Z)$ as the observed data; thus ρ may potentially depend on all data. Note also that at this point we let Θ be general; for example, it may be finite dimensional or it may be a class of functions.

Example 1 Perhaps, the most common example of a conditional moment problem is the instrumental variable regression problem (see e.g., [Angrist & Pischke, 2008](#), and citations therein), where we seek to estimate the causal effect of some treatment T on an outcome Y , where the observed relationship between T and Y may be confounded by some unobserved variables, but we have an instrumental variable Z that affects T but only affects Y via its effect on T . Given some regression function g parameterized by $\theta \in \Theta$, the value θ_0 corresponding the true regression function is assumed to be the unique solution to

$$\mathbb{E}[Y - g(T; \theta_0) | Z] = 0.$$

This is an example of Equation (1) with $X = (T, Y, Z)$ and $\rho(X; \theta) = Y - g(T; \theta)$.

More intricate variants of this, for example, include [Berry et al. \(1995\)](#), which incorporate discrete choice and is widely used to formulate structural demand parameters in industrial organizations.

Example 2 A second example is instrumental quantile regression ([Chernozhukov et al., 2007](#); [Horowitz & Lee, 2007](#)). Here, again assume a treatment T , outcome Y , and instrumental variable Z , but now we seek to estimate the causal effect of T on the p^{th} quantile of Y , for some $0 < p < 1$. In this case, given a quantile regression function g parameterized by $\theta \in \Theta$, the value θ_0 corresponding to the true quantile regression function is assumed to be the unique solution to

$$\mathbb{E}[\mathbb{1}\{Y \leq g(T; \theta_0)\} - p \mid Z] = 0.$$

This is an example of Equation (1) with $X = (T, Y, Z)$ and $\rho(X; \theta) = \mathbb{1}\{Y \leq g(T; \theta)\} - p$.

Example 3 A third example of a problem is estimating the stationary state density ratio between two policies in offline reinforcement learning ([Bennett et al., 2021](#); [Kallus & Uehara, 2022](#); [Liu et al., 2018](#)). Consider a Markov decision process given by an unknown transition kernel $p(S' \mid S, A)$ describing the distribution of next state S' when action A is taken in previous state S . Suppose $\pi_e(A \mid S), \pi_b(A \mid S)$ are two known policies assumed to induce unknown stationary distributions on S , $p_e(S), p_b(S)$, which we assume exist. We are interested in their ratio (generally, Radon–Nikodym derivative), $d(S)$. Given observations (S, A, S') from $p_b(S)\pi_b(A \mid S)p(S' \mid S, A)$, we have $d(S; \theta_0) \propto d(S)$ if and only if

$$\mathbb{E}[d(S; \theta_0)\pi_e(A \mid S)\pi_b^{-1}(A \mid S) - d(S'; \theta_0) \mid S'] = 0.$$

Then, for example, if Θ satisfies $\int d(s; \theta)d\mu(s) = 1 \forall \theta \in \Theta$ for some fixed measure μ , and there exists some $\theta_0 \in \Theta$ such that $d(S; \theta_0) \propto d(S)$, then this conditional moment restriction will identify θ_0 . Note that although $d(S; \theta_0) \neq d(S)$ in general, estimates of θ_0 are still of interest, as they could be used to estimate $d(S)$ in downstream tasks, for example by dividing by a plug-in estimate of $\mathbb{E}[d(S; \theta_0)]$ using estimates of θ_0 and \mathbb{E} [since $d(S)$ is known to satisfy the normalization constraint $\mathbb{E}[d(S)] = 1$.] This is an example of Equation (1) with $X = (S, A, S')$, $Z = S'$, and $\rho(X; \theta) = d(S; \theta_0)\pi_e(A \mid S)\pi_b^{-1}(A \mid S) - d(S'; \theta_0)$.

The classic approach to the conditional moment problem is to reduce it to a system of k marginal moments, $\mathbb{E}[F(Z)\rho(X; \theta_0)] = 0$, where $F: \mathcal{Z} \mapsto \mathbb{R}^{k \times m}$ is a chosen matrix-valued function. Then, we can apply the optimally weighted generalized method of moments (OWGMM; [Hansen, 1982](#)), which we present in detail in Section 2.1. Since this marginal moment formulation is implied by Equation (1) but not necessarily vice versa, this requires us to we find a sufficiently rich $F(Z)$ such that the marginal moment problem still identifies θ_0 , that is, it is still the unique solution in Θ . Moreover, even if this identifies θ_0 and even though OWGMM is efficient in the model implied by $\mathbb{E}[F(Z)\rho(X; \theta_0)] = 0$, the result may not be efficient in the model implied by Equation (1).

There are a few general approaches to deal with this. There are classic nonparametric approaches that are sieve-based and simply grow k , the output dimension of $F(Z)$, with n by including additional functions from a basis for L_2 such as power series ([Chamberlain, 1987](#)). There are also classic nonparametric approaches that directly estimate some special identifying $F^*(Z)$ that also induces an efficient OWGMM ([Newey, 1990, 1993](#)). For example, in Example 1 with $g(T; \theta) = \theta^T T$, we have $F^*(Z) = \mathbb{E}[T \mid Z]$, which can be nonparametrically estimated and plugged into OWGMM. Furthermore, there are approaches that used sieve-based methods to simultaneously estimate $\mathbb{E}[\rho(X; \theta) \mid Z]$ for every $\theta \in \Theta$, and pick θ to minimize some weighted empirical

norm of these estimated conditional expectations (Ai & Chen, 2003; Chen & Pouzo, 2009, 2012; Newey & Powell, 2003).

A recent line of work instead focuses on tackling this with machine-learning-based approaches (Bennett et al., 2019; Dikkala et al., 2020; Hartford et al., 2017; Kallus et al., 2021; Lewis & Syrskanis, 2018; Muandet et al., 2019; Singh et al., 2019; Uehara et al., 2021). These approaches are varied, with some solving the general problem in Equation (1) and others solving the more specific instrumental variable regression problem or other specific problems, with approaches based on deep learning, kernel methods, or both. Most of these are based on an adversarial/minimax/saddle-point approach (Bennett et al., 2019; Dikkala et al., 2020; Kallus et al., 2021; Lewis & Syrskanis, 2018; Muandet et al., 2019; Uehara et al., 2021).

Currently, there is a disconnect between these two lines of approaches. On the one hand, the more classical approaches are well motivated by efficiency theory when we impose certain smoothness assumptions. This is in contrast with the recent machine-learning based approaches; while some provide consistency guarantees (Bennett et al., 2019) and even rates (Dikkala et al., 2020; Kallus et al., 2021; Singh et al., 2019; Uehara et al., 2021), none of these approaches are shown to be semiparametrically efficient for Equation (1) or can facilitate inference on θ_0 . On the other hand, however, the more recent line of work leverages modern machine learning approaches, which are commonly believed to have superior practical properties. For example, they have been empirically observed to be more stable, have easier parameter tuning, or be better able to adapt to the low-dimensional latent structure of complex data. Although our experiments do indeed seem to support this thesis, especially in more challenging settings, we emphasise that the point of this paper is *not* to demonstrate that modern machine learning-based approaches are superior to classical ones. Rather, we observe that for various reasons there is significant, growing interest in machine-learning based approaches to these problems within the community, and therefore extending this line of work to be semiparametrically efficient and to perform inference is of great importance.

In this paper, we study a general class of minimax approaches, which we call the *variational method of moments* (VMM). This generalizes the method of Bennett et al. (2019), who presented an estimator for instrumental variable regression using adversarial training of neural networks. Their proposal was motivated by a variational reformulation of OWGMM, aiming to combine the efficiency of more classical approaches with the flexibility of machine learning methods. This style of estimator has since been applied to a variety of other conditional moment problems including policy learning from observational data (Bennett & Kallus, 2020) and estimating stationary state density ratios (Bennett et al., 2021). However, this past work did not provide a general formulation of VMM and a detailed theoretical analysis. And, although they are motivated by efficiency considerations, it is not immediately clear that this actually leads to efficient estimators.

We present a unified theory for a general class of VMM estimators. In particular, for some specific versions of these estimators based on either deep learning or kernel methods, we provide appropriate assumptions under which these methods are consistent, asymptotically normal, and semiparametrically efficient. In addition, we provide inference algorithms for these estimators, which can be used to construct confidence intervals for the estimated parameters. These inference algorithms are based on the same kind of variational reformulation as the estimation algorithms themselves, again with varieties based on both kernel methods and deep learning. Finally, we provide a detailed series of experiments that demonstrate that these VMM algorithms obtain very good finite-sample estimation performance and that the corresponding inference algorithms produce high quality confidence intervals.

The rest of this paper is structured as follows: in Section 2, we define the VMM estimator and provide motivation for it by interpreting OWGMM as a specific case thereof; in Section 3, we provide our theory for *kernel VMM* estimators, which are a specific instance of VMM estimators based on kernel methods; in Section 4, we provide our theory for *neural VMM* estimators, which are an alternative instance of VMM based on deep learning methods; in Section 5, we present our inference theory, with proposed kernel- and neural net-based algorithms; in Section 7, we provide a detailed empirical evaluation of our proposed estimation and inference methods; and in Section 8 we provide a detailed discussion of past work on solving conditional moment problems and how these approaches relate to our VMM estimators.

Notation

We use uppercase letters such as X to denote random variables and lowercase ones to denote non-random quantities. The set of positive integers is \mathbb{N} , and for any $n \in \mathbb{N}$ we use $[n]$ to refer to the set $\{1, \dots, n\}$. We denote by $\|\cdot\|_{L_p}$ the usual L_p functional norm, defined as $\|f\|_{L_p} = \mathbb{E}[|f(X)|^p]^{1/p}$, where the probability measure is implicit from context.

2 Variational method of moments

We now define the class of VMM estimators. We consider data consisting of n independent and identically distributed observations of X , namely, $X_1, \dots, X_n \sim \mathcal{P}$, where \mathcal{P} denotes the data distribution. Let some sequence of function classes \mathcal{F}_n be given, such that each $f \in \mathcal{F}_n$ has signature $f: \mathcal{Z} \rightarrow \mathbb{R}^m$. Let a ‘prior estimate’ $\tilde{\theta}_n \in \Theta$ be given. In general, this may be *any* data-driven choice from Θ and need not necessarily be consistent for θ_0 ; in the theory that follows we will elaborate on what conditions $\tilde{\theta}_n$ needs to satisfy for our respective results. Furthermore, let $R_n: \mathcal{F}_n \rightarrow [0, \infty]$ be some optional regularizer, which measures the complexity of $f \in \mathcal{F}_n$. Then, we define the VMM estimate $\hat{\theta}_n^{\text{VMM}} = \hat{\theta}_n^{\text{VMM}}(\mathcal{F}_n, R_n, \tilde{\theta}_n)$ corresponding to these choices as follows:

$$\hat{\theta}_n^{\text{VMM}} = \arg \min_{\theta \in \Theta} \sup_{f \in \mathcal{F}_n} \mathbb{E}_n[f(Z)^\top \rho(X; \theta)] - \frac{1}{4} \mathbb{E}_n[(f(Z)^\top \rho(X; \tilde{\theta}_n))^2] - R_n(f), \tag{2}$$

where \mathbb{E}_n is an empirical average over the n data points.

In Section 3, we study the instantiation of this with \mathcal{F}_n being a reproducing kernel Hilbert space (RKHS). In Section 4, we study the instantiation with \mathcal{F}_n being a class of neural networks.

Before proceeding to study these new machine-learning-based instantiations of the VMM estimator with flexible choices for \mathcal{F}_n , we discuss a very simple instantiation that recovers OWGMM, which provides motivation and interpretation for each of the terms in Equation (2).

2.1 The optimally weighted generalized method of moments

First, we present the classic OWGMM method. Given $F(Z) = (f_1(Z), \dots, f_k(Z))$, we obtain the marginal moment conditions $\mathbb{E}[f_i(Z)^\top \rho(X; \theta_0)] = 0 \forall i \in [k]$. Let a ‘prior estimate’ $\tilde{\theta}_n$ be given and define the matrix Γ as

$$\Gamma_{i,j} = \mathbb{E}_n[f_i(Z)^\top \rho(X; \tilde{\theta}_n) \rho(X; \tilde{\theta}_n)^\top f_j(Z)].$$

Then, the OWGMM estimate $\hat{\theta}_n^{\text{OWGMM}} = \hat{\theta}_n^{\text{OWGMM}}(f_1, \dots, f_k, \tilde{\theta}_n)$ is defined as

$$\hat{\theta}_n^{\text{OWGMM}} = \arg \min_{\theta \in \Theta} \sum_{i=1}^k \sum_{j=1}^k (\Gamma^{-1})_{ij} \mathbb{E}_n[f_i(Z)^\top \rho(X; \theta)] \mathbb{E}_n[f_j(Z)^\top \rho(X; \theta)]. \tag{3}$$

Given certain regularity conditions and assuming the choice of functions f_1, \dots, f_k are sufficient such that the corresponding k moment conditions uniquely identify θ_0 , standard GMM theory says that $\hat{\theta}_n$ is consistent for θ_0 . Furthermore, if the prior estimate $\tilde{\theta}_n$ is consistent for θ_0 , then this estimator is efficient with respect to the model defined by these k moment conditions (Hansen, 1982).

OWGMM generalizes the method of moments, which solves $\mathbb{E}_n[f_i(Z)^\top \rho(X; \theta)] = 0$ for all $i \in [k]$. When there are many moments, we cannot make all of them zero due to finite-sample noise and instead we seek to make them *near* zero. But, it is not clear which moments are more important; for example, there may be duplicate or near-duplicate moments. The key to OWGMM’s efficiency is to *optimally* combine the k objectives of making each moment near zero into a single objective function. To get a consistent prior estimate, we can for example let $\tilde{\theta}_n$ itself be a OWGMM with any fixed prior estimate, leading to the two-step GMM estimator. This can be repeated, leading to the multi-step GMM estimator.

Unfortunately, estimators of this kind have many limitations. For one, in practice, it is difficult or impossible to verify that any such set of functions f_1, \dots, f_k are sufficient for identification. In addition, while such an estimator is efficient with respect to the model imposed by these k moment conditions, ideally we would like to be efficient with respect to the model given by Equation (1); that is, we would wish to be efficient with respect to the model given by *all* moment conditions of the form $\mathbb{E}[f(Z)^\top \rho(X; \theta_0)] = 0$ for square integrable f . Finally, in the case that k were very large and growing with n , as would be required to (at least approximately) alleviate the prior two concerns, the corresponding sieve-based estimator would require impractical tuning to select which basis of L_2 to use and to choose k as a function of n . As will be seen below, such an approach may be seen as equivalent to estimating the optimal instruments over a linear sieve, but unlike our variational approach that we propose below it is unclear how to appropriately regularize this sieve estimation and take advantage of modern machine learning advances on non-parametric function approximation.

2.2 Variational reformulation of OWGMM

One motivation for our VMM class of estimators, Equation (2), is that it recovers OWGMM with its efficient weighting.

The following result simply appeals to the optimization structures of Equations (2) and (3) and generalizes Bennett et al. (2019, Lemma 1). We include its proof as it is short and instructive.

Lemma 1 $\hat{\theta}_n^{\text{OWGMM}}(f_1, \dots, f_k, \tilde{\theta}_n) = \hat{\theta}_n^{\text{VMM}}(\text{span}(\{f_1, \dots, f_k\}), 0, \tilde{\theta}_n)$.

Proof of Lemma 1. Let $F(Z) = (f_1(Z), \dots, f_k(Z))$ be a map $\mathcal{Z} \rightarrow \mathbb{R}^{k \times m}$. Then,

$$\begin{aligned} & \hat{\theta}_n^{\text{OWGMM}}(f_1, \dots, f_k, \tilde{\theta}_n) \\ &= \arg \min_{\theta \in \Theta} \|\Gamma^{-1/2} \mathbb{E}_n[F(Z)\rho(X; \theta)]\|^2 \\ &= \arg \min_{\theta \in \Theta} \sup_{v \in \mathbb{R}^k} v^\top \mathbb{E}_n[F(Z)\rho(X; \theta)] - \frac{1}{4} v^\top \Gamma v \\ &= \arg \min_{\theta \in \Theta} \sup_{v \in \mathbb{R}^k} \mathbb{E}_n[(F(Z)^\top v)^\top \rho(X; \theta)] - \frac{1}{4} \mathbb{E}_n[(F(Z)^\top v)^\top \rho(X; \tilde{\theta}_n)]^2, \end{aligned}$$

where the second equality is a reformulation of the rotated Euclidean norm (see [Online Supplementary Material, Lemma 15](#) for the general Hilbert-space version). The conclusion follows by noting $\{F(Z)^\top v : v \in \mathbb{R}^k\} = \text{span}(\{f_1, \dots, f_k\})$. \square

Through the lens of Lemma 1, we can understand each term of Equation (2) as follows. The first term pushes θ to make $\mathbb{E}_n[f(Z)^\top \rho(X; \theta)]$ near zero for each $f \in \mathcal{F}_n$. The second term, $-\frac{1}{4} \mathbb{E}_n[(f(Z)^\top \rho(X; \tilde{\theta}_n)]^2)$, appropriately weights the relative importance of making each of these near zero. Finally, varying \mathcal{F}_n and/or $R_n(f)$ with n allows us to control the richness of moments that we consider, in analogy to sieve-based methods that grow the dimension of the space $\text{span}(\{f_1, \dots, f_k\})$ but admitting more flexible machine-learning approaches. This motivation is similar to Bennett and Kallus (2020); Bennett et al. (2019, 2021), but these did not study the problem in generality or establish properties such as asymptotic normality or efficiency.

3 Kernel VMM

First, we consider a class of VMM estimators where for every n we have $\mathcal{F}_n = \mathcal{F}$, where $\mathcal{F} = \bigoplus_{i=1}^m \mathcal{F}_i$ and each \mathcal{F}_i is a RKHS of functions $\mathcal{Z} \rightarrow \mathbb{R}$ given by a symmetric positive definite kernel $K_i : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$, and regularization is performed using the RKHS norm of \mathcal{F} , which we denote by

$\|(f_1, \dots, f_m)\|^2 = \sum_{i=1}^m \|f_i\|_{\mathcal{F}_i}^2$. We will call these estimators *kernel VMM* estimators, which we concretely define according to

$$\hat{\theta}_n^{\text{K-VMM}} = \arg \min_{\theta \in \Theta} J_n(\theta), \tag{4}$$

$$J_n(\theta) = \sup_{f \in \mathcal{F}} \mathbb{E}_n[f(Z)^\top \rho(X; \theta)] - \frac{1}{4} \mathbb{E}_n[(f(Z)^\top \rho(X; \tilde{\theta}_n))^2] - \frac{\alpha_n}{4} \|f\|^2,$$

and α_n is some non-negative sequence of regularization coefficients. Explicitly, this fits into our general VMM definition with $\mathcal{F}_n = \mathcal{F}$ for every n , and $R_n(f) = \alpha_n \|f\|^2$.

Before we provide our main theory for kernel VMM estimators, we provide a convenient reformulation of Equation (4). Let \mathcal{H} be the dual space of \mathcal{F} (that is, the space of all bounded linear functionals of the form $\mathcal{F} \mapsto \mathbb{R}$) and for each $\theta \in \Theta$ define the element $\bar{h}_n(\theta) \in \mathcal{H}$ according to

$$\bar{h}_n(\theta)(f) = \mathbb{E}_n[f(Z)^\top \rho(X; \theta)].$$

Furthermore, define the linear operator $C_n : \mathcal{H} \rightarrow \mathcal{H}$ according to

$$(C_n b)(f) = \mathbb{E}_n[\varphi(b)(Z)^\top \rho(X; \tilde{\theta}_n) \rho(X; \tilde{\theta}_n)^\top f(Z)],$$

where $\varphi : \mathcal{H} \rightarrow \mathcal{F}$ maps any element in \mathcal{H} to its Riesz representer in \mathcal{F} such that $b(f) = \langle \varphi(b), f \rangle$.

Lemma 2 The kernel VMM estimator defined in Equation (4) is equivalent to

$$\hat{\theta}_n^{\text{K-VMM}} = \arg \min_{\theta \in \Theta} \|(C_n + \alpha_n I)^{-1/2} \bar{h}_n(\theta)\|_{\mathcal{H}}^2,$$

where I is the identity operator $Ib = b$.

We note that comparing this result to Equation (3), this is a clear infinite-dimensional generalization of the OWGMM objective, where the matrix Γ defined there is replaced with a linear operator, and the inversion is performed using Tikhonov regularization. Note that this re-framing of our kernel VMM estimator also shows a connection to the continuum GMM estimators considered by Carrasco and Florens (2000). However, our estimator does not strictly fit within their framework. We discuss this in more detail in Section 8.

3.1 Consistency

We first provide some sufficient assumptions in order to ensure that our kernel VMM estimator is consistent; that is, $\hat{\theta}_n^{\text{K-VMM}} \rightarrow \theta_0$ in probability. Before we present these assumptions, we define the conditional covariance function of the moment problem

$$V(Z; \theta) = \mathbb{E}[\rho(X; \theta) \rho(X; \theta)^\top | Z]. \tag{5}$$

For our first assumption, we require each \mathcal{F}_i to be universally approximating with a smooth kernel. Recall for this definition that, a function is C^∞ -smooth if it is n -times continuously differentiable for every positive integer n . In addition, we recall that a kernel is *universal* if the corresponding RKHS is dense in the space of continuous real-valued functions on \mathcal{Z} under the supremum norm (Sriperumbudur et al., 2011). Note that all of the properties of the following assumption hold, for example, for the commonly used Gaussian kernel.

Assumption 1 (Universal RKHS). For each $i \in [m]$, K_i is C^∞ -smooth in both arguments and \mathcal{F}_i is universal.

Next, we require a basic regularity condition on the observed data distribution. This together with Assumption 1 ensure that \mathcal{F} is well-behaved with respect to ρ and satisfies some nice properties in terms of boundedness and metric entropy, as formalized by [Online Supplementary Material, Lemma 16](#) in the Appendix.

Assumption 2 (Regularity). \mathcal{Z} is a bounded subset of \mathbb{R}^{d_z} for some positive integer d_z .

Next, we require that the set of possible functions $\{\rho(\cdot; \theta) : \theta \in \Theta\}$ satisfies some basic boundedness, smoothness, and complexity properties. A simple example satisfying the below is for Θ to be a compact set in some finite-dimensional Euclidean space, and for $\rho(x; \theta)$ to be equi-Lipschitz continuous in θ for every x . Other examples that easily satisfy the second part of the below include $\{\theta(\cdot; \theta) : \theta \in \Theta\}$ having finite Vapnik–Chervonenkis dimension (see e.g., [Kosorok, 2007](#), Theorem 8.19 and Corollary 9.5), or be a bounded-norm subset of an RKHS (see [Online Supplementary Material, Lemma 17](#) in the Appendix for details). This assumption ensures that consistent estimation of θ_0 is possible, even though inversion of the conditional moment operator could be ill-posed.

Assumption 3 (Moment Class Complexity). $\sup_{x \in \mathcal{X}, \theta \in \Theta} |\rho(x; \theta)| < \infty$, and $\rho(X; \theta)$ is Lipschitz continuous in θ under the L_1 norm. Also, for each $i \in [m]$, the function set $\{\rho_i(\cdot; \theta) : \theta \in \Theta\}$ is \mathcal{P} -Donsker.

We also assume that the prior estimate $\tilde{\theta}_n$ is well-behaved, meaning that it converges sufficiently fast to some limit in probability. This limit need not be θ_0 for our consistency results. This will be used to ensure the convergence of the linear operator C_n defined above to some limiting operator C .

Assumption 4 (Convergent Prior Estimate). The prior estimate $\tilde{\theta}_n$ has a limit $\tilde{\theta}$ in probability, and satisfies $\|\rho_i(X; \tilde{\theta}_n) - \rho_i(X; \tilde{\theta})\|_2 = O_p(n^{-p})$ for every $i \in [m]$ and some $0 < p \leq 1/2$.

Finally, we assume a nonsingular covariance with bounded inverse moments.

Assumption 5 (Non-Degenerate Moments). For each $\theta \in \{\tilde{\theta}, \theta_0\}$, we have that $V(Z; \theta)$ is invertible almost surely, and also that $\|\sigma_{\min}(Z; \theta)^{-1}\|_\infty < \infty$, where $\sigma_{\min}(Z; \theta)$ denotes the minimum eigenvalue of $V(Z; \theta)$.

Assumption 5 is slightly subtle and is used to ensure that the objective J_n defined above converges to a well-behaved limiting objective J that is uniquely minimized by θ_0 , which is central to our consistency proof. In the absence of this assumption, it is possible that the limiting objective may diverge. We note that in the case of $m = 1$, the second part of the assumption is equivalent to requiring that $\|V(Z; \tilde{\theta})^{-1}\|_\infty, \|V(Z; \theta_0)^{-1}\|_\infty < \infty$, and in the case that the prior estimate $\tilde{\theta}_n$ is consistent we only need this condition to hold at $\theta_0 = \tilde{\theta}$. In general, it can be viewed in terms of certain moments defined in terms of the data distribution and ρ being bounded.

With these assumptions, we are prepared to state our consistency result.

Theorem 1 (Consistency). Let Assumptions 1–5 be given and suppose the regularization coefficient satisfies $\alpha_n = o(1)$ and $\alpha_n = \omega(n^{-p})$, where p is the constant referenced in Assumption 4. Then, for any $\hat{\theta}_n$ that satisfies $J_n(\hat{\theta}_n) = \inf_{\theta \in \Theta} J_n(\theta) + o_p(1)$, we have $\hat{\theta}_n \rightarrow \theta_0$ in probability.

Comparing this result to the corresponding consistency result given by [Bennett et al. \(2019, Theorem 2\)](#), we note that this result *does not* rely on any specific identification assumptions beyond Equation (1). Conversely, [Bennett et al. \(2019\)](#) assume that the class \mathcal{F} of neural nets that they take a supremum over is sufficient to uniquely identify θ_0 , which is a questionable assumption since this class is assumed to be fixed and not growing with n . Therefore, we argue that our VMM consistency here is given under much more reasonable assumptions.

Next, we make some observations about how this result compares with consistency results in the literature that tackles nonlinearities using sieves. First, note that Assumption 3 is weaker than the corresponding assumptions in [Ai and Chen \(2003\)](#) and [Newey and Powell \(2003\)](#),

who assume that $\rho(x; \theta)$ is point-wise Hölder-continuous and Θ is compact. Instead, we require the more general assumption of continuity in the L_1 -norm, along with a Donsker condition. Conversely, [Chen and Pouzo \(2009\)](#) and [Chen and Pouzo \(2012\)](#) similarly allow for non-smooth ρ , but they consider the setting where Θ can be non-compact, introducing ill-posedness issues that they tackle in their work. Rather, we specifically consider metrics on Θ under which ill-posedness is *not* an issue, given our Donsker assumption on $\{\rho(\cdot; \theta) : \theta \in \Theta\}$ and L_1 -continuity. Furthermore, we note that assumptions similar to Assumptions 2 and 5 are standard in these past works, and Assumptions 1 and 4 are straightforward technical conditions related to implementation choices for our method.

3.2 Asymptotic normality

We now present our theory for the asymptotic normality of kernel VMM estimates. Here, we consider the special case where Θ is a compact subset of \mathbb{R}^b for some positive integer b . We note that in this case, as discussed above, Assumption 3 follows under very simple additional conditions; e.g., $\rho(x; \theta)$ being equi-Lipschitz continuous in θ for every $x \in \mathcal{X}$. Under this setting, we will characterize the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$.

First, we require that $\rho(X; \theta)$ satisfies the following differentiability condition.

Assumption 6 (ρ Differentiable in Absolute Mean). For each $i \in [m]$, there exists some vector-valued function $D_i(X; \theta) \in \mathbb{R}^b$ indexed by θ , and some neighbourhood Θ_0 of θ_0 , such that, for every $\theta \in \Theta_0$, we have

$$\lim_{\theta' \rightarrow \theta} \frac{\left\| \rho_i(X; \theta') - \rho_i(X; \theta) - (\theta' - \theta)^\top D_i(X; \theta) \right\|_{L_1}}{\|\theta' - \theta\|} = 0.$$

In other words, D_i is a gradient-like function such that the first-order Taylor error decays to zero at a $o(\|\theta' - \theta\|)$ rate under the L_1 norm. For example, in the case that $\rho_i(x; \theta)$ is continuously differentiable in θ within some neighbourhood of θ_0 for all $x \in \mathcal{X}$, then Assumption 6 trivially follows from Taylor’s theorem. Furthermore, it is easy to see that for any x, θ where $\rho_i(x; \theta)$ is differentiable w.r.t. θ , we must have $D_i(x; \theta) = \nabla \rho_i(x; \theta)$. However, the above is more general and allows for situation where $\rho_i(x; \theta)$ is non-differentiable at some values of x and θ . In particular, the following lemma allows us to establish this assumption under more general conditions.

Lemma 3 Suppose there exist $\phi_i : \mathcal{X} \rightarrow \mathbb{R}$ indexed by Θ , and ‘gradient-like’ and ‘Hessian-like’ functions ρ'_i and ρ''_i such that: (1) $\rho_i(X; \theta)$ is twice differentiable in θ with gradient $\rho'_i(X; \theta)$ and Hessian $\rho''_i(X; \theta)$ whenever $\phi(X; \theta) \neq 0$; (2) $\sup_{x \in \mathcal{X}, \theta \in \Theta} \|\rho'_i(x; \theta)\|_2 \leq c'$ and $\sup_{x \in \mathcal{X}, \theta \in \Theta} \|\rho''_i(x; \theta)\|_{\text{op}} \leq c''$ for some $c', c'' < \infty$; (3) $\rho_i(x; \theta)$ is L_ρ -Lipschitz in θ for every $x \in \mathcal{X}$, for some $L_\rho < \infty$; and (4) $\phi(x; \theta)$ is $L_\phi(x)$ -Lipschitz in θ , for some $L_\phi(x)$ such that the probability density of the random variable $L_\phi(X)^{-1}\phi(X; \theta)$ is bounded within some neighbourhood of zero. Then, we have that Assumption 6 holds with $D_i(X; \theta) = \rho'_i(X; \theta)$.

This lemma allows us to establish Assumption 6 for a range of problems where $\rho(X; \theta)$ has some points of non-smoothness. Intuitively, the boundedness condition on ρ'' allows us to bound the first-order Taylor error whenever ρ is smooth, and the Lipschitz and bounded density assumptions on ϕ near $\phi = 0$ prevents non-smoothness from impacting the first-order Taylor expansion, up to an additional $o(\|\theta' - \theta\|)$ factor.

Next, we let $D(X; \theta) \in \mathbb{R}^{m \times b}$ denote the Jacobian-like function given by concatenating $D_i(X; \theta)$ for all $i \in [m]$. Similarly, we define $h'_n, h' \in \mathcal{H}^b$ according to

$$h'_n(f) = \mathbb{E}_n[D(X; \theta_0)^\top f(Z)] \quad h'(f) = \mathbb{E}[D(X; \theta_0)^\top f(Z)],$$

where \mathcal{H} is the dual space of \mathcal{F} , as above. We also define the analogue of the gradient of the objective $J'_n(\theta) \in \mathbb{R}^b$ for each $\theta \in \Theta_0$, according to

$$J'_n(\theta) = 2\langle (C_n + \alpha_n I)^{-1/2} b_n(\theta), b'_n(\theta) \rangle,$$

and note that in the case that $\rho(X_i; \theta)$ is differentiable at θ for $i \in [n]$ that $J'_n(\theta) = \nabla J_n(\theta)$. In addition, we define linear operators $C: \mathcal{H} \rightarrow \mathcal{H}$ and $C_0: \mathcal{H} \rightarrow \mathcal{H}$ according to

$$\begin{aligned} (Cb)(f) &= \mathbb{E}[\varphi(b)(Z)^\top \rho(X; \tilde{\theta}) \rho(X; \tilde{\theta})^\top f(Z)], \\ (C_0 b)(f) &= \mathbb{E}[\varphi(b)(Z)^\top \rho(X; \theta_0) \rho(X; \theta_0)^\top f(Z)], \end{aligned}$$

where $\tilde{\theta}$ is the probability limit of $\tilde{\theta}_n$ as specified by Assumption 4, and φ is defined as in the definition of C_n above. Given these definitions, we can now specify our additional assumptions and the asymptotic normality result.

This next additional assumption is a regularity condition on $D(X; \theta)$, which extends the properties of $\rho(X; \theta)$ specified in Assumption 3 to $D(X; \theta)_j$ for each $j \in [b]$.

Assumption 7 (Gradient Complexity). Let Θ_0 be the neighbourhood of θ_0 from Assumption 6. For each $i \in [m]$ and $j \in [b]$, we have $\sup_{x \in \mathcal{X}, \theta \in \Theta_0} |D_i(x; \theta)_j| < \infty$, and that $D_i(X; \theta)_j$ is Lipschitz continuous in θ under the L_1 norm. In addition, for each $i \in [m]$ and $j \in [b]$ the class $\{D_i(\cdot; \theta)_j; \theta \in \Theta\}$ is \mathcal{P} -Donsker.

Next, we assume a certain non-degeneracy in the parametrization of the problem, locally near θ_0 .

Assumption 8 (Non-degenerate Θ). For $\beta \in \mathbb{R}^b$, we have $\mathbb{E}[\sum_{j=1}^b \beta_j D(X; \theta_0)_j | Z] = 0$ almost surely if and only if $\beta = 0$.

Assumption 8 is needed to ensure that the limiting asymptotic variance is finite and that the matrix Ω defined in the theorem statement below is invertible. It can be interpreted as the assumption that the parametrization of Θ is non-degenerate, since it requires that the functions $\mathbb{E}[\rho'_i(X; \theta_0) | Z]$ are linearly independent. Note that this assumption is somewhat lax, since if it were violated, it is likely possible we could re-parameterize the problem with a lower-dimensional Θ in order to avoid this issue.

Finally, we will need to introduce a couple of important definitions. We say that an estimator $\hat{\theta}_n$ for θ_0 is *asymptotically linear* if $\hat{\theta}_n = \mathbb{E}_n[\psi(X)] + o_p(n^{-1/2})$, for some ψ satisfying $\mathbb{E}[\psi] = \theta_0$. In addition, we say that such an estimator is *asymptotically normal* if $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to a mean-zero Gaussian random variable, with some fixed covariance matrix.

With these additional assumptions and definitions, we are prepared to present our asymptotic normality result.

Theorem 2 (Asymptotic Normality). Let Assumptions 1–8 be given, and suppose the regularization coefficient satisfies $\alpha_n = o(1)$ and $\alpha_n = \omega(n^{-p})$, where p is the constant defined in Assumption 4. Then, for any $\hat{\theta}_n$ that satisfies $\|J'_n(\hat{\theta}_n)\| = o_p(n^{-1/2})$, we have that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically linear and asymptotically normal, with covariance matrix $\Omega^{-1} \Delta \Omega^{-1}$, where Δ and Ω are defined according to

$$\begin{aligned} \Delta_{i,j} &= \langle (C^{-1/2} C_0 C^{-1/2}) C^{-1/2} b'_i, C^{-1/2} b'_j \rangle, \\ \Omega &= \mathbb{E}[\mathbb{E}[D(X; \theta_0) | Z]^\top V(Z; \tilde{\theta})^{-1} \mathbb{E}[D(X; \theta_0) | Z]]. \end{aligned}$$

Note that this theorem requires an approximate first-order optimality condition, namely, that the gradient-like element J' satisfies $\|J'_n(\hat{\theta}_n)\| = o_p(n^{-1/2})$, which is stronger than the approximate

optimality condition in Theorem 1. Although this condition may be difficult to interpret or verify in general, the following lemma provides some sufficient conditions.

Lemma 4 (Sufficient Conditions for Approximate First-Order Optimality). Suppose that either (1) $\hat{\theta}_n \in \arg \min_{\theta} J_n(\theta)$; or (2) $\rho(x; \theta)$ is twice continuously differentiable in θ for every $x \in \mathcal{X}$, and $J_n(\hat{\theta}_n) = J_n(\theta_n^*) + o_p(1/n)$. Then, given the other conditions of Theorem 2, we have $\|J'_n(\hat{\theta}_n)\| = o_p(n^{-1/2})$.

Comparing this result to comparable results in the literature leveraging more classical non-parametric approaches, we note that our differentiability condition in Assumption 6 is weaker than the point-wise differentiability of $\rho(X; \theta)$ assumed by Ai and Chen (2003), but stronger than Chen and Pouzo (2009) who only require differentiability of $\mathbb{E}[\rho(X; \theta) | Z]$. We also note that, these two works further allow for non-parametric nuisance functions in addition to the asymptotically normal parametric component. Furthermore, we note that Assumption 8 is a standard condition in all of these works.

3.3 Efficiency

Next, we address the question of efficiency of these kernel VMM estimators. In order to present this theory, we first need to introduce the notions of *regularity* and *semiparametric efficiency*; we refer the reader to Van der Vaart (2000) for precise definitions. Roughly speaking, we say that an estimator $\hat{\theta}_n$ is *regular* with respect to some model of distributions if it is sufficiently well behaved such that its asymptotic behaviour is invariant to small perturbations [of size $O_p(n^{-1/2})$] to the data-generating distribution that remain inside the model. In addition, we say that $\hat{\theta}_n$ is *semiparametrically efficient* with respect to a model of distributions if it is regular and achieves the minimum asymptotic variance among *all* regular estimators (with respect to that model).

Given the complex form of the limiting covariance in Theorem 2 in terms of linear operators and inner products on \mathcal{H} , it is not immediately clear how large this covariance is and whether it is efficient under any conditions. Fortunately, the following theorem, which holds under no additional assumptions, justifies efficiency in the case that our prior estimate for θ_0 is consistent.

Theorem 3 (Efficiency). Let the assumptions of Theorem 2 be given with $\tilde{\theta} = \theta_0$, and let $\hat{\theta}_n$ be any estimator that satisfies the conditions of Theorem 2. Then, $\hat{\theta}_n$ is semiparametrically efficient with respect to the model given by Equation (1) and $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with asymptotic covariance matrix Ω_0^{-1} , where Ω_0 is defined according to

$$\Omega_0 = \mathbb{E}[\mathbb{E}[D(X; \theta_0) | Z]^T V(Z; \theta_0)^{-1} \mathbb{E}[D(X; \theta_0) | Z]].$$

This theorem immediately implies that such a kernel VMM estimator is not only efficient with respect to the class of all kernel VMM estimators, but that it achieves the semiparametric efficiency bound for solving Equation (1). This is a very strong result, which ensures that these kernel VMM estimators inherit the efficiency properties that OWGMM estimators possess for standard moment problems, as was hoped.

Comparing against the efficiency results of the continuum GMM estimators of Carrasco and Florens (2000), which is the most similar approach to kernel VMM, ours is stronger. Specifically, they only justified that their estimator is efficient compared with other estimators in their class of continuum GMM estimators, while we have proven efficiency relative to *all* possible regular estimators. That is, we achieve the same semiparametric efficiency as, e.g., Ai and Chen (2003) and Chen and Pouzo (2009), who use fundamentally different sieve-based approaches.

Finally, we note that although this limiting covariance matrix has a somewhat complicated form, this form has a variational interpretation similar to the Kernel VMM estimator itself. We discuss this interpretation and how to use it to estimate the efficient asymptotic variance in Section 5.

3.4 Implementing kernel VMM estimators

Finally, we address some implementation considerations for kernel VMM estimators.

Firstly, we note that the above theory does not provide any guidance on how to actually construct a prior estimate $\tilde{\theta}_n$ that has the required properties described in Assumption 4. In order to address this issue, we now present a concrete method for constructing such a $\tilde{\theta}_n$, which allows us to avoid explicitly assuming Assumption 4. Let us use the terminology that $\hat{\theta}_n$ is a 0-step kernel VMM estimate if $\hat{\theta}_n$ is chosen as some arbitrary fixed value, which does not depend on the observed data. Then, for any integer $k > 0$, we say that $\hat{\theta}_n$ is a k -step kernel VMM estimate if $\hat{\theta}_n$ is computed by approximately solving Equation (4) according to $J_n(\hat{\theta}_n) = \inf_{\theta \in \Theta} J_n(\theta) + o_p(1/n)$, with $\tilde{\theta}_n$ chosen as a $(k-1)$ -step kernel VMM estimate. In other words, $\hat{\theta}_n$ is a k -step kernel VMM estimate if it is computed by iteratively approximately solving Equation (4) k times, with $\tilde{\theta}_n$ chosen as the previous iterate solution, starting from some arbitrary constant value. This scheme is analogous to that of the k -step GMM estimator (Hansen et al., 1996). Given this definition, we have the following lemma:

Lemma 5 Suppose that $\tilde{\theta}_n$ is a k -step kernel VMM estimate for some $k > 0$. Then, given all assumptions of Theorem 2 except for Assumption 4, it follows that $\hat{\theta}_n$ satisfies the conditions of Assumption 4 with $p = 1/2$, and $\hat{\theta} = \theta_0$.

Therefore, as long as we construct $\hat{\theta}_n$ as a k -step kernel VMM estimator as described above for some $k > 1$, we are assured that Assumption 4 will be met with $p = 1/2$ and $\hat{\theta} = \theta_0$. Given this and Theorem 3, we immediately have the following corollary for k -step kernel VMM estimators.

Corollary 1 Suppose that $\hat{\theta}_n$ is calculated as a k -step kernel VMM estimate for some $k > 1$. Then given Assumptions 1–3 and 5–8, and assuming that the regularization coefficient satisfies $\alpha_n = o(1)$ and $\alpha_n = \omega(n^{-1/2})$, it follows that $\hat{\theta}_n$ is semiparametrically efficient for θ_0 .

This corollary ensures that, given our regularity assumptions about \mathcal{F} and the conditional moment problem itself, we can construct a specific k -step kernel VMM estimator that is semiparametrically efficient. The above also provides a valid specific choice of the regularization coefficient α_n that does not depend on unknown parameters.

Secondly, we address the fact that the cost function described in Equation (4) is given by a supremum over the infinite \mathcal{F} and provide a closed-form for the objective. By appealing to the representer theorem, and the factorization of \mathcal{F} into the direct sum of m RKHSs, we can establish the following lemma.

Lemma 6 Define the vector $\rho(\theta) \in \mathbb{R}^{n \cdot m}$ and the matrices $L \in \mathbb{R}^{(n \cdot m) \times (n \cdot m)}$ and $Q(\theta) \in \mathbb{R}^{(n \cdot m) \times (n \cdot m)}$ according to

$$\begin{aligned} \rho(\theta)_{i,k} &= \rho_k(X_i; \theta), & L_{(i,k),(i',k')} &= \mathbb{1}\{k = k'\} K_k(Z_i, Z_{i'}), \\ Q(\theta)_{(i,k),(i',k')} &= \frac{1}{n} \sum_{j=1}^n K_k(Z_i, Z_j) \rho_k(X_j; \theta) K_{k'}(Z_{i'}, Z_j) \rho_{k'}(X_j; \theta). \end{aligned}$$

Then, the cost function $J_n(\theta)$ being minimized by Equation (4) is equivalent to

$$J_n(\theta) = \frac{1}{n^2} \rho(\theta)^\top L (Q(\tilde{\theta}_n) + \alpha_n L)^{-1} L \rho(\theta).$$

In other words, the kernel VMM estimator can be computed by minimizing a simple closed-form cost function, which is given by a particular convex quadratic form on the terms of the form $\rho_k(X_i; \theta)$ for $i \in [n]$ and $k \in [m]$.

In the special case of instrumental variable regression, where we are fitting the regression function within an RKHS ball, we can not only find a closed-form solution for the cost function $J_n(\theta)$ to be minimized, but for the kernel VMM estimator itself. Specifically, we provide the following lemma, which follows by applying the representer theorem again.

Lemma 7 Consider the instrumental variable regression problem, where $m = 1$, $\rho(X; \theta) = Y - \theta(T)$, \mathcal{F} is the RKHS with kernel K_f , and Θ is a ball of the RKHS with kernel K_g with radius r and centred at zero. In addition, let Y denote the vector of outcomes (Y_1, \dots, Y_n) , let L_f and L_g denote the kernel Gram matrices of K_f and K_g on the data Z_1, \dots, Z_n and T_1, \dots, T_n , respectively, and define the $n \times n$ matrices $Q(\theta)$ and M according to

$$Q(\theta)_{i,i'} = \frac{1}{n} \sum_{j=1}^n K_f(Z_i, Z_j) K_f(Z_{i'}, Z_j) (W_j - \theta(T_j))^2,$$

$$M = \frac{1}{n^2} L_f (Q(\tilde{\theta}_n) + \alpha_n L_f)^{-1} L_f.$$

Then, we have $\hat{\theta}_n^{\text{K-VMM}} = \sum_{i=1}^n \beta_i^* K_g(\cdot, T_i)$, where

$$\beta^* = (L_g M L_g + \lambda_n L_g)^{-1} L_g M Y,$$

for some $\lambda_n \geq 0$ which depends implicitly on $r, K_f, K_g, \tilde{\theta}_n$, and the observed data.

The term λ_n enters into the above equation via Lagrangian duality, since minimizing $J_n(\theta)$ over the RKHS ball with radius r is mathematically equivalent to minimizing $J_n(\theta) + \lambda_n \|\theta\|^2$ over the entire RKHS, for some implicitly defined $\lambda_n \geq 0$. In practice, however, when performing IV regression according to Lemma 7 we could freely select λ_n as a hyperparameter instead of r . Superficially, the form of this estimator is similar to that of other recently proposed kernel-based estimators for IV regression (Muandet et al., 2019; Singh et al., 2019). However, unlike those estimators, ours incorporates optimal weighting using the prior estimate $\tilde{\theta}_n$.

4 Neural VMM estimators

We now consider a different class of VMM estimators, where the sequence of function classes \mathcal{F}_n is given by a class of neural networks with growing depth and width. We will refer to estimators in this class as *neural VMM* (N-VMM). Most generally, we will define the class of N-VMM estimators according to

$$\hat{\theta}_n^{\text{K-VMM}} = \arg \min_{\theta \in \Theta} \sup_{f \in \mathcal{F}_n} \mathbb{E}_n[f(Z)^\top \rho(X; \theta)] - \frac{1}{4} \mathbb{E}_n[(f(Z)^\top \rho(X; \tilde{\theta}_n))^2] - R_n(f), \tag{6}$$

where $R_n(f)$ is some regularizer. In this section, we analyze K-VMM for different choices of R_n . For simplicity, we will restrict our theoretical analysis to the case where \mathcal{F}_n is a fully connected neural network with ReLU activations and a common width in all layers, which allows us to use the universal approximation result of Yarotsky (2017, Theorem 1). Specifically, we fix a network architecture with D_n hidden layers, each with W_n neurons, with the final fully connected layer connecting to the m outputs. Then, the class \mathcal{F}_n is given by varying the weights on this network. We note that this choice is made for simplicity of exposition, but similar bounds could be given for different kinds of architectures, using other universal approximation results as in e.g., Yarotsky (2017, 2018).

4.1 Neural VMM with kernel regularizer

First, we consider the case where we regularize using some RKHS norm. Specifically, let \mathcal{F}_K be a product of m RKHSs satisfying Assumption 1, and let $\|f\|_{n,K} = \inf_{f' \in \mathcal{F}_K : f'(Z_i) = f(Z_i) \forall i} \|f'\|$ denote the minimum norm of any $f' \in \mathcal{F}_K$ that agrees with f at the points Z_1, \dots, Z_n . Note that

$$\|f\|_{n,K}^2 = \sum_{k=1}^m f_k^\top K_k^{-1} f_k,$$

where $f_k = (f(Z_1)_k, \dots, f(Z_n)_k)^\top$, and K_k is the kernel Gram matrix on the data Z_1, \dots, Z_n using the kernel for the k^{th} dimension of \mathcal{F}_K . Then, we will consider estimators of the form

$$\hat{\theta}_n^{\text{NK-VMM}} = \arg \min_{\theta \in \Theta} \sup_{f \in \mathcal{F}_n} \mathbb{E}_n[f(Z)^\top \rho(X; \theta)] - \frac{1}{4} \mathbb{E}_n[(f(Z)^\top \rho(X; \tilde{\theta}_n))^2] - \frac{\alpha_n}{4} \|f\|_{n,K}^2. \quad (7)$$

We note that if we were to replace \mathcal{F}_n with \mathcal{F}_K in equation (7), then this equation would be equivalent to equation (4), since by the representer theorem regularizing by $\|f\|_{\mathcal{F}_K}$ gives the same supremum over f as regularizing by $\|f\|_{n,K}$. Given this and the known universal approximation properties of neural networks, it may be hoped that if we grow the class \mathcal{F}_n sufficiently fast, then the objective we are minimizing over θ in equation (7) is approximately equal to that of equation (4) in a uniform sense over $\theta \in \Theta$. This, then, would hopefully imply that this neural VMM estimator is able to achieve the same desirable properties, in terms of consistency, asymptotic normality, and efficiency, as our kernel VMM estimators.

In order to formalize the above intuition, we first require the following assumption, which allows us to account for the rate of growth of the kernel Gram matrix inverses K_k^{-1} in the results we give below.

Assumption 9 (Inverse Kernel Growth). There exists some deterministic positive sequence $k_n = \Omega(1)$, such that $\|K_i^{-1}\|_2 = O_p(k_n)$ for each $i \in [m]$.

In addition, we require the following assumption on the rate of growth on the width W_n and depth D_n of \mathcal{F}_n , in order to ensure that we can approximate equation (4) sufficiently well.

Assumption 10 (Neural Network Size). There exist constants $q \geq 0$, $0 < a < 1/2$ and a sequence $r_n = o(n^{-1-q}k_n^{-1})$ such that $W_n = \omega(r_n^{-a} \log(r_n^{-1}))$ and $D_n = \omega(\log(r_n^{-1}))$.

Finally, in our results and discussion below, we will define $J_n(\theta)$ to be the loss in θ minimized by $\hat{\theta}_n^{\text{NK-VMM}}$, and $J_n^*(\theta)$ to be the corresponding oracle loss if we were to replace \mathcal{F}_n with \mathcal{F}_K .

Lemma 8 Let Assumptions 1–3, 9, and 10 be given. Then, we have

$$\sup_{\theta \in \Theta} |J_n(\theta) - J_n^*(\theta)| = o_p(n^{-q}).$$

This lemma follows by applying recent results on the size of a neural network required to uniformly approximate all functions of a given Sobolev norm (Yarotsky, 2017), and also older results that show that, under the conditions of Assumption 1, any RKHS ball has bounded Sobolev norm for any Sobolev space using more than $d_z/2$ derivatives (Cucker & Smale, 2002).

Given this, we can immediately state the following theorem, which ensures that the theoretical results of our kernel VMM estimators carry over to our neural VMM estimators with kernel regularization.

Theorem 4 Let the assumptions of Theorem 1 and Assumptions 9 and 10 be given. In addition, let $\hat{\theta}_n$ be any sequence that satisfies $J_n(\hat{\theta}_n) = \inf_{\theta \in \Theta} J_n(\theta) + o_p(n^{-q})$, where q is the constant referenced in Assumption 10. Then, in the case that these assumptions hold with $q = 0$, we have $\hat{\theta}_n \rightarrow \theta_0$ in probability.

Furthermore, suppose in addition that the assumptions of Theorem 2 hold, and the above assumptions are strengthened to hold with $q = 1$. Then, we have that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to a mean-zero Gaussian random variable, with covariance as given by Theorem 2.

Finally, assume that in addition $\hat{\theta} = \theta_0$. Then the asymptotic variance of $\hat{\theta}_n$ is given by Theorem 3, and the estimator is semiparametrically efficient.

The proof of this theorem follows immediately from Lemma 8, since this Lemma and the Theorem's conditions ensure that $J_n^*(\hat{\theta}_n) = \inf_{\theta \in \Theta} J_n^*(\theta) + o_p(n^{-q})$. Therefore, we can directly apply Theorems 1–3 to obtain these three results.

An immediate observation given this theorem is that, if we define k -step estimators as in Section 3.4, then by applying an identical argument we can construct efficient neural VMM estimators without having to explicitly make Assumption 4.

4.2 Neural VMM with other regularizers

Motivated by our theory above using kernel-based regularizers, we now provide some discussion of general neural VMM estimators of the form given by Equation (6) for other choices of $R_n(f)$, and in particular we discuss how these estimators may be justified.

First, consider the case where the kernel Gram matrices K_i for $i \in [m]$ are approximately equal to $\sigma_i I$, where σ_i is some scalar and I is the identity matrix. For example, this is the case if we use a Gaussian kernel with very small length scale parameter. In this case, we may reasonably approximate

$$\|f\|_{n,K} \approx \sum_{k=1}^m \frac{1}{\sigma_k} \sum_{i=1}^n f_k^2(Z_i). \tag{8}$$

That is, we could justify instead regularizing using some (possibly weighted) Frobenius norm of the matrix given by the values of the vector-valued f at the n data points. This form of regularization is much more attractive than that given by $\|f\|_{n,K}$, since it does not involve the computation of inverse kernel Gram matrices, and it more naturally fits into estimators for Equation (6) given by some form of alternating stochastic gradient descent. We also note that this form of regularization, based on the Frobenius norm of f , is similar to that used by [Dikkala et al. \(2020\)](#), although with some important differences; their proposed estimators do not include the $-(1/4)\mathbb{E}_n[(f(Z)^\top \rho(X; \tilde{\theta}_n))^2]$ term motivated by efficiency theory, and they only present theory on bounding the risk of their learned function given by $\tilde{\theta}_n$, not on the consistency or semiparametric efficiency of the estimated $\tilde{\theta}_n$. We discuss this comparison in more detail in Section 8.

Alternatively, we may heuristically justify leaving out the $R_n(f)$ term altogether, under the argument that neural network function classes naturally impose some smoothness constraints, and therefore optimizing over \mathcal{F}_n is morally similar to optimizing over \mathcal{F}_K with some norm constraint. This intuition can be made more concrete by noting that there is a rich literature showing equivalence between optimizing loss functions over neural network function classes, and optimizing the same loss over some norm-bounded RKHS class whose kernel is implicitly defined by the neural network architecture [see e.g., [Shankar et al. \(2020\)](#) and citations therein]. However, we leave more specific non-heuristic claims on the performance of our neural VMM algorithms with $R_n(f) = 0$ to future work.

4.3 Implementing neural VMM estimators

Regardless of the choice of the regularization term R_n , the question remains of how to actually solve Equation (6). Past work ([Bennett & Kallus, 2020](#); [Bennett et al., 2019](#)) has solved this problem using the Optimistic Adam (OAdam) algorithm, which is a form of alternating stochastic gradient descent (that is, alternating between first-order gradient steps minimizing the game objective with respect to θ and maximizing the game objective with respect to f) that has been designed to have good properties for solving minimax problems ([Daskalakis et al., 2017](#)). These past works have proposed to do this by continuously updating $\tilde{\theta}_n$; that is, at each iteration of alternating stochastic gradient descent they set $\tilde{\theta}_n$ as the previous iterate solution.

Alternatively, there is a rich recent literature on other, potentially more efficient, methods for solving smooth game optimization problems such as Equation (4). For example, see [Fiez et al. \(2020\)](#); [Gidel et al. \(2019\)](#); [Lin et al. \(2020a, 2020b\)](#); [Loizou et al. \(2020\)](#); [Thekumparampil et al. \(2019\)](#), and references therein. Some or all of the approaches suggested in these recent works may lead to successful neural VMM implementations. However, we leave this more empirical investigation to future work, and in our experiments we focus on approaches based on OAdam with continuously updated $\tilde{\theta}_n$, as discussed above.

5 Inference

So far, we have developed both theory and algorithms for kernel and neural VMM estimators, providing conditions under which such estimators are consistent, asymptotically normal, and/or efficient. We now extend our efficient estimation theory to efficient inferential theory, focusing on the case of $\Theta \subseteq \mathbb{R}^b$.

Now, suppose we want to construct confidence intervals for $\psi(\hat{\theta}_n)$, for some $\psi: \mathbb{R}^b \mapsto \mathbb{R}$. This is a very general kind of quantity to consider, since, for example, if we were interested in $(\hat{\theta}_n)_i$ for some $i \in [b]$, we could define $\psi(\theta) = \theta_i$. By the delta method, if $\hat{\theta}_n$ were an efficient estimate then the asymptotic variance of $\psi(\hat{\theta}_n)$ would be $\nabla\psi(\theta_0)^\top \Omega_0^{-1} \nabla\psi(\theta_0)$, where Ω_0^{-1} is the efficient covariance matrix defined in Theorem 3. Therefore, this suggests that we could construct asymptotically calibrated Wald confidence intervals by estimating $\hat{\beta}_n^\top \Omega_0^{-1} \hat{\beta}_n$, for some data-driven $\hat{\beta}_n$. In particular, if $\nabla\psi(\theta_0)$ were known, which would be the case if ψ were linear, then we could do this with $\hat{\beta}_n = \nabla\psi(\theta_0)$. Otherwise, we could do this using $\hat{\beta}_n = \nabla\psi(\hat{\theta}_n)$, where $\hat{\theta}_n$ is some consistent estimate of θ_0 (such as a VMM estimate), which would be consistent for $\nabla\psi(\theta_0)$ given Assumption 7.

In this section, we provide consistent algorithms for estimating $\beta^\top \Omega_0^{-1} \beta$ for arbitrary $\beta \in \mathbb{R}^b$, with analogous kernel and neural varieties of our algorithms. These consistent variance estimators can then immediately be used with the delta method, as discussed above, to construct asymptotically calibrated Wald confidence intervals for our efficient VMM estimators.

Our algorithms, which are presented in the next subsections, are motivated by the following key lemma.

Lemma 9 Let Ω_0 be defined as in Theorem 3, let the conditions of Theorem 3 hold, and let $\nabla\rho(X; \theta) \in \mathbb{R}^{m \times b}$ denote the Jacobian of $\rho(X; \theta)$ with respect to θ . Then, for any vector $\beta \in \mathbb{R}^b$, we have

$$\begin{aligned} \beta^\top \Omega_0^{-1} \beta &= \sup_{\gamma \in \mathbb{R}^b} \gamma^\top \beta - \frac{1}{4} \gamma^\top \Omega_0 \gamma \\ &= -\frac{1}{4} \inf_{\gamma \in \mathbb{R}^b} \sup_{f \in \mathcal{F}} \mathbb{E}[f(Z)^\top \nabla\rho(X; \theta_0) \gamma] - \frac{1}{4} \mathbb{E}[(f(Z)^\top \rho(X; \theta_0))^2] - 4\gamma^\top \beta. \end{aligned}$$

The first part of this lemma follows by applying a similar variational reformulation argument as in the proof of Lemma 1, and the second part follows by applying a similar argument again on the $\gamma^\top \Omega_0 \gamma$ term, given the definition of Ω_0 from Theorem 3. More details are given in the Appendix.

We note that the right hand side of Lemma 9 has a very similar structure to the game objective of our VMM algorithms. Given this, the previous argument suggests that the asymptotic variance of any such $\psi(\hat{\theta}_n)$ could be estimated using approaches similar to our kernel and neural VMM estimation algorithms presented previously. In the remainder of this section, we build on this intuition and present kernel- and neural-based algorithms for inference.

5.1 Kernel inference algorithm

First, we present an inference algorithm along the lines of our kernel VMM estimator. This algorithm is summarized by the following theorem.

Theorem 5 Let the conditions of Theorem 3 be given, and let $\hat{\theta}_n$ be any corresponding efficient estimate of θ_0 . In addition, let L and $Q(\theta)$ be defined as in Lemma 6, and define $D(\theta) \in \mathbb{R}^{(n \cdot m) \times b}$ and $\Omega_n \in \mathbb{R}^{b \times b}$ according to

$$D_{(i,k),j}(\theta) = \frac{\partial}{\partial \theta_j} \rho_k(X_i; \theta), \quad \Omega_n = \frac{1}{n^2} D^\top L(Q(\hat{\theta}_n) + \alpha_n L)^{-1} L D,$$

where α_n is any sequence satisfying the assumptions of Theorem 3. Then $\Omega_n \rightarrow \Omega_0$ in probability.

We note that an immediate corollary of this theorem is that, for any continuously differentiable ψ and an efficient $\hat{\theta}_n$ such as our VMM estimators, $\nabla\psi(\hat{\theta}_n)^\top \Omega_n^{-1} \nabla\psi(\hat{\theta}_n)$ is consistent for the

asymptotic variance of $\psi(\hat{\theta}_n)$, which is an efficient estimate of $\psi(\theta_0)$, where Ω_n^- denotes the pseudo-inverse of Ω_n . This follows trivially by the continuous mapping theorem and Slutsky's theorem, since by assumption Ω_0 is invertible. An advantage of this algorithm is that it allows easy estimation of the entire covariance matrix Ω_0^{-1} , from which the asymptotic variance of any single-dimensional function of $\hat{\theta}_n$ can instantly be estimated without applying any additional variational algorithms.

5.2 Neural inference algorithm

Our neural inference algorithm is similar in nature to our neural VMM estimator and is given by the following smooth game:

$$v_n(\beta) = -\frac{1}{4} \inf_{\gamma \in \mathbb{R}^b} \sup_{f \in \mathcal{F}_n} \mathbb{E}_n[f(Z)^\top \nabla \rho(X; \hat{\theta}_n) \gamma] - \frac{1}{4} \mathbb{E}_n[(f(Z)^\top \rho(X; \hat{\theta}_n))^2] - R_n(f) - 4\gamma^\top \beta, \quad (9)$$

where R_n is a regularizer for f and \mathcal{F}_n is a sequence of neural net classes. Then, given Lemma 9, we expect $v_n(\beta)$ to be a reasonable estimator for $\beta^\top \Omega_0^{-1} \beta$. Furthermore, following the argument presented at the beginning of Section 5, we expect $v_n(\hat{\beta}_n)$ to be a reasonable estimator for the (efficient) asymptotic variance of $\psi(\hat{\theta}_n)$ if $\hat{\beta}_n$ is consistent for $\nabla \psi(\theta_0)$. We note that, unlike for our neural VMM estimator, we do not provide any theoretical guarantees for this algorithm, due to some additional technical complications; unlike the game objective being solved by neural VMM, the space being minimized over for γ is unbounded, which complicates the technical argument by universal approximation we used for neural VMM. We leave this theoretical question to future work. However, we note that in our inference experiments in Section 7 this method seems to work well.

Unlike our kernel inference algorithm, this approach has the disadvantage that it requires solving a separate optimization problem for every given scalar parameter ψ . In practice, though, this may be alleviated by the practical strengths of neural methods, as discussed previously. In addition, as with our neural VMM algorithm, we may regularize for example by using a kernel-based norm or the Frobenius norm of $\{f(Z_1), \dots, f(Z_n)\}$, or we may omit this regularization term entirely.

6 Examples

Next, let us provide some concrete examples of our theory, in order to demonstrate how the assumptions for our consistency and asymptotic normality theory may be satisfied. For each example, we do not discuss Assumptions 1, 4, 9, and 10 explicitly, as these govern design choices for the algorithm that can be generically satisfied given the other assumptions.

6.1 Nonparametric instrumental-variable regression

First, let us consider a specific nonparametric instrumental regression example, which instantiates Example 1 from Section 1. Specifically, we will consider conditions under which our consistency result Theorem 1 applies. Let us consider the data generating process $Y = g(T; \theta_0) + \epsilon$, where $\mathbb{E}[\epsilon | Z] = 0$. We assume that $Z \in \mathbb{R}^{d_z}$, $\|Z\|_\infty < \infty$, $\|Y\|_\infty < \infty$, and $\mathbb{V}[\epsilon | Z] \geq \lambda$ almost surely, for some fixed $\lambda > 0$. Let us also suppose that $g(T; \theta)$ is $L(T)$ -Lipschitz continuous in θ , where $\mathbb{E}[L(T)^2] < \infty$, that $\sup_{\theta \in \Theta} \|g(T; \theta)\|_\infty < \infty$, and that $\mathcal{G} = \{g(\cdot; \theta) : \theta \in \Theta\}$ is a Donsker class. As one example, these conditions would be satisfied if \mathcal{G} were given by the class of all monotonic functions on T such that $\|g(X; \theta)\|_\infty \leq b$ for some fixed $b < \infty$, with the norm on Θ given by $\|\theta' - \theta\| = \|g(T; \theta') - g(T; \theta)\|_\infty$. As a second example, \mathcal{G} could be a norm-bounded RKHS satisfying the conditions of Assumption 1, with the norm on Θ given by the corresponding RKHS norm.

First, given the conditions on Z in this example, Assumption 2 is trivial. Second, given the conditions on the regression class \mathcal{G} , along with the assumption that $\|Y\|_\infty < \infty$, Assumption 3 trivially follows by applying Lemma 9.14 in Kosorok (2007). Finally, assuming that the prior estimate $\hat{\theta}_n$ comes from some arbitrary consistent methodology, then Assumption 5 only needs to hold for $\theta = \theta_0$. In this case, this is ensured under the above condition on the conditional variance of ϵ , since $V(Z; \theta_0) = \mathbb{V}[\epsilon | Z]$. Given this, we have consistency via Theorem 1.

6.2 Nonparametric instrumental-variable quantile regression

Next, let us consider a specific nonparametric instrumental regression example, which instantiates Example 2 from Section 1. Again, we will consider conditions under which consistency holds. Let us consider the data generating process $Y = g(T; \theta_0) + \epsilon$, where $\text{Prob}(\epsilon \leq 0 \mid Z) = p$ for almost everywhere Z , and $\rho(X; \theta) = \mathbb{1}\{Y \leq g(X; \theta)\} - p$. Again, we assume that $Z \in \mathbb{R}^{d_z}$, and $\|Z\|_\infty < \infty$. In this case, we will assume that $\mathcal{G} = \{g(\cdot; \theta) : \theta \in \Theta\}$ is some regression class that is Donsker under the supremum norm $\|\theta' - \theta\| = \|g(X; \theta') - g(X; \theta)\|_\infty$, and that Y has bounded density.

Again, given the conditions on Z in this example, Assumption 2 is trivial, and the Donsker part of Assumption 3 follows from the fact that \mathcal{G} is Donsker by Lemma 9.14 of Kosorok (2007). Also, we have $\mathbb{E}[|\rho(X; \theta') - \rho(X; \theta)|] = \text{Prob}(\min(g(X; \theta'), g(X; \theta)) \leq Y \leq \max(g(X; \theta'), g(X; \theta)))$. Now, since by assumption Y has bounded density, it easily follows that there exists some constant L such that $\text{Prob}(\min(g(X; \theta'), g(X; \theta)) \leq Y \leq \max(g(X; \theta'), g(X; \theta)) \leq L\|g(X; \theta') - g(X; \theta)\|_\infty$, which gives us the required Lipschitz continuity under L_1 norm. Also, the required boundedness is trivial since $|\rho(X; \theta)| \in \{-p, 1-p\}$, so we have Assumption 3. Finally, we have $V(Z; \theta_0) = \mathbb{E}[(\mathbb{1}\{\epsilon \leq 0\} - p)^2 \mid Z] = p - p^2$ almost surely, and therefore $\|V(Z; \theta_0)^{-1}\|_\infty \leq (p - p^2)^{-1} < \infty$, which gives us Assumption 5, again as long as the prior estimate $\tilde{\theta}_n$ is consistent. Therefore, again we have consistency via Theorem 1.

6.3 Parametric instrumental-variable mean and expectile regression

Next, we will consider a parametric expectile (including mean) regression example (Newey & Powell, 1987; Sobotka et al., 2013), where we can establish both consistency, asymptotic normality, and efficiency. For this example, we will assume that the data generating process is again given by $Y = g(T; \theta_0) + \epsilon$, where ϵ instead satisfies $p\mathbb{E}[\mathbb{1}\{\epsilon \geq 0\} \mid Z] = (1-p)\mathbb{E}[-\mathbb{1}\{\epsilon < 0\} \mid Z]$ for some $p \in (0, 1)$. For $p = 0.5$, we get the usual mean regression. Here, we let $\rho(X; \theta) = w(X; \theta)(Y - g(T; \theta))$, where $w(X; \theta) = p\mathbb{1}\{Y \geq g(T; \theta)\} + (1-p)\mathbb{1}\{Y < g(T; \theta)\}$, and the goal is to find the unique $\theta_0 \in \Theta$ such that $\mathbb{E}[\rho(X; \theta_0) \mid Z] = 0$. Note that this problem that can be seen as a mid-point between standard instrumental variable regression and instrumented quantile regression. Similar to the above example, let us suppose that $Z \in \mathbb{R}^{d_z}$, $\|Z\|_\infty < \infty$, $\|Y\|_\infty < \infty$, and $\forall \epsilon \mid Z \geq \lambda$ almost surely, for some fixed $\lambda > 0$. For this example, we will further assume that $T \in \mathbb{R}^{d_t}$, $\|T\|_\infty < \infty$, the regression class is given by $g(t; \theta) = \theta^\top t$, where $\Theta = \{\theta \in \mathbb{R}^{d_t} : \|\theta\|_2 \leq b\}$ for some $b < \infty$, and that the matrix $\mathbb{E}[\mathbb{E}[T \mid Z]\mathbb{E}[T \mid Z]^\top]$ is full-rank.

Again, given the conditions on Z in this example, Assumption 2 is trivial. Similarly, given the boundedness of Θ and T , and the fact that $\|w(X; \theta)\|_\infty \leq 1$, along with Lemma 9.14 of Kosorok (2007), we easily have that Assumption 3 holds. In addition, we have $V(Z; \theta_0) \geq \min(p, 1-p)^2 \forall \epsilon \mid Z$, and so Assumption 5 follows from our minimum conditional variance assumption as in the previous example. Therefore, we can establish consistency via Theorem 1.

Next, under the additional assumption that Y and T both have bounded probability density, then so does $Y - g(T; \theta)$ for every $\theta \in \Theta$. Therefore, we can apply Lemma 3 with $\phi(X; \theta) = Y - g(T; \theta)$ in order to establish Assumption 6 with $D(X; \theta) = T$, which we note trivially satisfies the conditions of Assumption 7. Finally, since $\mathbb{E}[\mathbb{E}[T \mid Z]\mathbb{E}[T \mid Z]^\top]$ is assumed to be full rank, we have $\beta^\top \mathbb{E}[\mathbb{E}[T \mid Z]\mathbb{E}[T \mid Z]^\top] \beta = \mathbb{E}[\mathbb{E}[\beta^\top D(X; \theta_0) \mid Z]^2] > 0$ for every non-zero β , which establishes Assumption 8. Therefore, we also have asymptotic normality via Theorem 2, and under the condition that the prior estimate $\tilde{\theta}_n$ was consistent we have semiparametric efficiency via Theorem 3.

7 Experiments

We now present a series of experiments to demonstrate our proposed methodologies. We present two kinds of experiments. First, we test the finite-sample performance of our kernel and neural VMM algorithms on a range of synthetic conditional moment problems. In this experiment, we compare their performance with the classical sieve minimum distance (SMD) approach of Ai and Chen (2003), which is a sieve-based method that has previously been proposed as a semiparametrically efficient approach to solving generic conditional moment problems. In addition, we

compare their performance with the recently proposed maximum moment restriction (MMR) algorithm of Zhang et al. (2020), which as discussed in Section 8 is equivalent to the limit of our kernel VMM algorithm in the limit as $\alpha_n \rightarrow \infty$. Second, we test our proposed inference algorithms on a subset of these scenarios, evaluating the quality of the resulting confidence intervals for different variations of our estimation and inference algorithms. Code for reproducing all experiments is available at <https://github.com/CausalML/VMM>.

7.1 Estimation experiments

7.1.1 Estimation scenarios

SimpleIV. This is a simple parametric instrumental variable regression scenario, based on a simple data generating process where

$$Z = \sin(\pi U/10), \quad T = -0.75U + 3.5H + 0.14\eta - 0.6, \quad Y = g(T; \theta_0) + -10H + \epsilon.$$

In this setup, η and H are exogenous iid $\mathcal{N}(0, 1)$ variables, and U is an exogenous iid Uniform $(-5, 5)$ random variable, and each of T, Z , and Y , are scalars. We note that the random variable H introduces endogeneity. Furthermore, we have $g(t; \theta) = \theta_1 + \theta_2 t + \theta_3 t^2$ where $\theta \in \mathbb{R}^3$, with the true parameter value given by $\theta_0 = [0.5, 3.0, -0.5]$. In this scenario, the conditional moment equation to be solved is $\mathbb{E}[Y - g(T; \theta_0) | Z] = 0$; that is, we have $X = (T, Y, Z)$, and $\rho(X; \theta) = Y - g(T; \theta)$. Note that in this scenario the relationship between treatment and instruments is nonlinear.

HeteroskedasticIV. This is a more challenging instrumental variable regression scenario, which introduces a more complex nonlinear regression function class and heteroskedastic noise. It follows a similar data generating process to the prior SimpleIV scenario, except here we have

$$Z = (U_1, U_2), \quad T = 0.75(Z_1 + |Z_2|) + 1.25H + 0.05\eta, \\ Y = g(T; \theta_0) + 5H + 0.1\text{softplus}(Z_1 + |Z_2|)\eta,$$

where again η and H are iid $\mathcal{N}(0, 1)$ distributed, and each of U_1 and U_2 are iid Uniform $(-5, 5)$ distributed. We also note the ‘softplus’ activation function is defined according to $\text{softplus}(x) = \log(1 + \exp(x))$. In this case, we have $\theta \in \mathbb{R}^4$, and our regression class is defined according to

$$g(t; \theta) = \theta_2 + \theta_3(t - \theta_1) + \frac{\theta_4 - \theta_3}{2} \text{softplus}(2(t - \theta_1)).$$

That is, our regression class is a smoothed version of a hinge function with slopes θ_3 and θ_4 and hinge point at (θ_1, θ_2) . The true parameter value is given by $\theta_0 = [2.0, 3.0, -0.5, 3.0]$. As with our SimpleIV scenario, the conditional moment restriction is given by $\mathbb{E}[Y - g(T; \theta_0) | Z] = 0$.

We note that although the regression residual is *not* independent of the instruments Z in this setting, it is mean-independent, since $\mathbb{E}[S\epsilon] = \mathbb{E}[\mathbb{E}[S\epsilon | Z]] = 0$. That is, we have heteroskedastic noise with respect to our instruments, which makes achieving efficiency more challenging.

PolicyLearning. Finally, this scenario is based on learning optimal binary treatment policies from surrogate loss reductions, following Bennett and Kallus (2020). Let $T \in \{-1, 1\}$ denote the binary treatment variable, Z denote individual covariates, $Y(t)$ denote the potential outcome for the individual that would occur if (possibly counter to fact) treatment t were assigned, and $Y = Y(T)$ denote the actual outcome. Then, given logged data where treatments were decided using some randomized policy, and some well-specified parametric class of deterministic treatment policies $\Pi = \{\pi_\theta : \theta \in \Theta\}$, the task is to estimate the parameters of the *optimal* policy within Π That is, we wish to estimate $\theta_0 = \arg \max_{\theta \in \Theta} \mathbb{E}[Y(\pi(Z; \theta))]$, where $\pi(z; \theta)$ denotes the treatment assigned by policy π_θ given $Z = z$. For this problem, we assume the following data generating process:

$$Z \sim \mathcal{N}(0, 1) \times \mathcal{N}(0, 1), \quad T \sim 2\text{Bernoulli}(e(Z)) - 1, \quad Y(t) = \mu_t(Z) + \sigma_t(Z)\epsilon_t \quad \forall t \in \{-1, 1\},$$

where ϵ_1 and ϵ_{-1} are iid $\mathcal{N}(0, 1)$ variables. The functions e , μ_{-1} , and μ_1 are all given by quadratic forms on Z , and the functions σ_{-1} and σ_1 are given by quadratic forms on Z with softplus activation; exact coefficients for these functions are provided in the [Online Supplementary Material](#).

Now, assume that the policy class Π is defined according to some parametric utility function $g(\cdot; \theta)$, where $\pi_\theta \in \Pi$ assigns an individual with covariates z to treatment 1 if and only if $g(z; \theta) \geq 0$, else it assigns the individual to treatment -1 . Then, under some assumptions outlined in [Bennett and Kallus \(2020\)](#) regarding correct specification with respect to the logistic regression surrogate loss, the problem of estimating θ_0 is described by the conditional moment problem $\mathbb{E}[W | (\text{expit}(g(Z; \theta_0)) - \mathbb{1}\{W > 0\}) | Z] = 0$, where the weighting variable W is defined according to

$$W = \mu_1(Z) - \mu_{-1}(Z) + \frac{T(Y - \mu_T(Z))}{Te(Z) + (1 - T)/2}.$$

In our experiments, we have $\theta \in \mathbb{R}^6$, and $g(z; \theta) = \theta^T \phi(z)$, where $\phi(z) = (1, z_1, z_2, z_1^2, z_2^2, z_1 z_2)$ gives a quadratic feature expansion of z . Given this and the fact that μ_1 and μ_{-1} are quadratic forms, it easily follows that the optimal parameters are given by $\theta_0 = \theta_1 - \theta_{-1}$, where θ_1 and θ_{-1} are the parameter vectors describing the quadratic forms μ_1 and μ_{-1} , respectively.

Note that, following [Bennett and Kallus \(2020\)](#), in practice when estimating θ_0 we estimate the weights W using plugin estimators for e , μ_{-1} , and μ_1 , which we fit using flexible neural nets. For fair comparison, in our experiments all methods use the same estimated weights W .

7.1.2 Estimation methods

KernelVMM. We implemented a two-step kernel VMM estimator, as described in Section 3.4. For the first step, $\hat{\theta}_n$ is chosen randomly. For a kernel function, we used the same kernel function as used by [Zhang et al. \(2020\)](#), which is given by the average of three Gaussian kernels with automatically selected data-driven bandwidths; we provide details in the [Online Supplementary Material](#). In all cases, we optimize $\hat{\theta}_n$ by minimizing the cost function described in Lemma 6 using L-BFGS, and we experimented with the range of values of α_n .

NeuralVMM. We implemented a neural VMM estimator by optimizing the minimax objective described by Equation (6) using alternating stochastic gradient descent, using the OAdam optimizer ([Daskalakis et al., 2017](#)), as discussed in Section 4.3. In all cases, we used a simple fixed three-layer fully connected architecture for our f network. As discussed in Section 4, we use a Frobenius-norm style regularization term of the form $R_n(f) = (\lambda_n/nm) \sum_{i=1}^n \sum_{k=1}^m f_k(Z_i)^2$, and we experimented with a wide range of λ_n . We also experimented with explicit kernel-based regularization as in Section 4.1, however we found that it performed extremely poorly in practice due to extreme gradient values, so we did not include it in our main experiments. We describe additional details, such as hyperparameters, network architectures, and early stopping, in the [Online Supplementary Material](#).

MMR. The MMR algorithm was originally developed by [Zhang et al. \(2020\)](#) for the instrumental variable regression problem but easily extends to other conditional moment problems. One particular form of their algorithm (we discuss the more general form in Section 8) is given by minimizing the objective function $J_n^{\text{MMR}}(\theta) = \frac{1}{n^2} \rho(\theta)^T L \rho(\theta)$, where L and $\rho(\theta)$ are defined as in Lemma 6. As discussed in Section 8, this is equivalent to KernelVMM in the limit as $\alpha_n \rightarrow \infty$. Given this deep connection, we also present results for this version of the MMR algorithm for all scenarios, using the same kernel function as for KernelVMM. As with our KernelVMM method, we minimize this objective using L-BFGS.

SMD. The SMD method of [Ai and Chen \(2003\)](#) applied to our problem given by Equation (1) (in fact it applies to a more general moment problem with an additional infinite-dimensional nuisance components; see Section 8 for details) is based on minimizing the objective function $J_n^{\text{SMD}}(\theta) = \mathbb{E}_n[q_n(Z; \theta)^T \Gamma_n(Z)^- q_n(Z; \theta)]$, where $q_n(z; \theta)$ is a nonparametric sieve regression estimate of $\mathbb{E}[\rho(X; \theta) | Z = z]$, $\Gamma_n(z)$ is any consistent estimate of $\mathbb{E}[\rho(X; \theta_0) \rho(X; \theta_0)^T | Z = z]$, and $\Gamma_n(Z)^-$ denotes the pseudo-inverse of $\Gamma_n(Z)$. The past work proposes various sieve-based approaches

for $q_n(Z; \theta)$, but is not prescriptive about the methodology for computing Γ_n . Given this, we experimented with various SMD estimators, using B-splines for $q_n(Z; \theta)$, and multiple approaches for Γ_n : (1) *Identity*, in which we simply set $\Gamma_n(z) = I \forall z$; (2) *Homoskedastic*, in which we set $\Gamma_n = \mathbb{E}_n[\rho(X; \hat{\theta}_n)\rho(X; \hat{\theta}_n)^\top] \forall z$; and (3) *Heteroskedastic*, in which we fit a diagonal $\Gamma_n(Z)$ by regressing $\rho(X; \hat{\theta}_n)_i^2$ on Z for each $i \in [m]$ using neural networks. We provide additional details in the [Online Supplementary Material](#).

OWGMM. The OWGMM estimator follows the method described in Section 2.1, for a flexible set of basis functions f_1, \dots, f_k . As with the SMD method we chose these sets of basis functions using B-splines, as this allowed for a very rich and flexible class of moment conditions. Again, we provide additional details in the [Online Supplementary Material](#).

NCB. Finally, we implemented a simple non-causal baseline (NCB) that estimates θ_0 by ignoring Z and instead trying to solve $\mathbb{E}[\rho(X; \theta_0) | X] = 0$. For example, for our instrumental variable regression scenarios, this corresponds to assumption that there is no endogeneity in the treatments T . For this baseline, we simply minimize the objective $J_n^{\text{NCB}}(\theta) = \mathbb{E}_n[\rho(X; \theta)^2]$, which we implement using L-BFGS.

7.1.3 Estimation results

For each scenario and each $n \in \{200, 500, 1,000, 2,000, 5,000, 10,000\}$, we repeated the following process 50 times: we drew a training set of n random iid data points using the respective scenario’s data generating process as well as an additional dataset of n random dev data points for early stopping, hyperparameter tuning, etc., and then we estimated $\hat{\theta}_n$ using all of our methods and baselines using the sampled dataset. Then, for each combination of scenario and n we computed the mean squared error (MSE) of the estimated $\hat{\theta}_n$ across these 50 replications. We summarize the results of this process in [Table 1](#). In addition, we computed additional results based on the risk (SimpleIV and HeteroskedasticIV) or regret (PolicyLearning) of the estimated $g(\cdot; \hat{\theta}_n)$. However, these broadly followed the same trend as the main results here, so we leave them to the [Online Supplementary Material](#). In addition, we provide additional tables of results that break down the MSE in terms of bias and standard deviation in the [Online Supplementary Material](#).

Overall, we can see that in all scenarios the best performing methods are our VMM methods, with the neural VMM method performing best in the SimpleIV and PolicyLearning scenarios, and the kernel VMM method performing best in the HeteroskedasticIV scenario. And, in all cases both the kernel and neural VMM methods significantly outperform the baselines. In particular, apart from the easy SimpleIV scenario, in the more complex HeteroskedasticIV and PolicyLearning scenarios, our VMM methods yield errors that are orders of magnitude smaller.

In terms of the values of the regularization hyperparameters, we note that kernel VMM can be sensitive to the choice of α_n when it takes extreme values. When α_n is too small, the algorithm appears to suffer from high variance and the occasional catastrophically bad results, whereas when α_n is too large the estimation becomes very biased, with performance converging to that of MMR. However, for α_n in the range of 10^{-2} – 10^{-6} , performance is good across all scenarios and n . It remains a question for future work how to automatically select this hyperparameter using observed data. However, we suspect that approaches based on the eigenvalues of $(Q(\hat{\theta}_n) + \alpha_n L)^{-1}$ appearing in [Lemma 6](#) might be productive.

Conversely, we note that our neural VMM algorithm is generally very insensitive to the choice of λ_n , with very little change in performance even for relatively large values of λ_n , and very strong and stable performance even when $\lambda_n = 0$. The one minor exception to this is in the challenging PolicyLearning scenario, where using the largest value of λ_n results in somewhat better performance than other choices for low values of n , but worse performance for large n . This reinforces the notion that the neural network function class and optimization algorithms are naturally regularizing, and that explicit regularization is not necessarily important.

In general, for both VMM algorithms, we note that there is a wide range of regularization hyperparameter values where performance is generally very good. Furthermore, we note that for both cases the choices of \mathcal{F} used were very generic and the same across all scenarios; either an RKHS with a completely generic data-driven kernel, or a very generic shallow MLP. Together, this

Table 1. For each combination of scenario, method, and n , the MSE of $\hat{\theta}_n$ is estimated over 50 replications, along with standard errors

Method		n					
		200	500	1,000	2,000	5,000	10,000
(a) SimpleIV							
K-VMM	$\alpha_n = 0$	>100	8.8 ± 42.7	>100	.67 ± 1.2	.23 ± .29	.14 ± .16
	$\alpha_n = 10^{-8}$	5.1 ± 7.0	2.8 ± 3.0	2.6 ± 5.3	3.2 ± 16.5	.25 ± .32	.17 ± .23
	$\alpha_n = 10^{-6}$	5.5 ± 7.0	2.5 ± 2.7	1.7 ± 3.0	.78 ± 1.3	.24 ± .33	.14 ± .16
	$\alpha_n = 10^{-4}$	5.5 ± 7.6	2.5 ± 3.2	1.8 ± 2.9	.72 ± 1.3	.25 ± .32	.14 ± .16
	$\alpha_n = 10^{-2}$	6.0 ± 8.3	2.7 ± 3.1	1.7 ± 2.4	.72 ± 1.2	.26 ± .34	.14 ± .17
	$\alpha_n = 1$	11.1 ± 21.2	4.1 ± 6.6	2.1 ± 2.8	.75 ± 1.1	.34 ± .41	.16 ± .21
N-VMM	$\lambda_n = 0$	2.5 ± 2.0	1.6 ± 1.9	.93 ± 1.2	.42 ± .65	.16 ± .21	.10 ± .14
	$\lambda_n = 10^{-2}$	2.2 ± 1.9	2.1 ± 2.6	.74 ± .99	.42 ± .66	.17 ± .23	.10 ± .12
	$\lambda_n = 1$	2.1 ± 2.0	2.1 ± 2.1	.94 ± 1.2	.39 ± .65	.18 ± .26	.11 ± .12
SMD	Identity	4.2 ± 6.5	2.5 ± 3.6	1.8 ± 3.0	.68 ± 1.0	.24 ± .31	.15 ± .19
	Homo	4.2 ± 6.5	2.5 ± 3.6	1.8 ± 3.0	.68 ± 1.0	.24 ± .32	.15 ± .19
	Hetero	4.3 ± 5.7	2.4 ± 3.3	1.7 ± 2.6	.66 ± 1.0	.24 ± .31	.15 ± .18
MMR		17.7 ± 28.0	5.6 ± 9.2	2.8 ± 3.7	.83 ± 1.1	.37 ± .45	.17 ± .23
OWGMM		3.1 ± 5.3	2.3 ± 4.1	1.7 ± 2.0	.85 ± 1.0	.33 ± .42	.20 ± .24
NCB		6.2 ± 1.3	6.0 ± .71	5.8 ± .45	5.8 ± .47	5.8 ± .25	5.8 ± .20
(b) HeteroskedasticIV							
K-VMM	$\alpha_n = 0$	>100	3.8 ± 5.5	>100	.63 ± 1.4	.24 ± .29	.09 ± .18
	$\alpha_n = 10^{-8}$	>100	>100	1.3 ± 2.2	.63 ± 2.0	.21 ± .23	.06 ± .05
	$\alpha_n = 10^{-6}$	8.7 ± 22.9	2.0 ± 2.6	.78 ± .98	.35 ± .50	.22 ± .27	.06 ± .05
	$\alpha_n = 10^{-4}$	9.9 ± 27.6	1.9 ± 2.2	.79 ± .96	.35 ± .45	.21 ± .26	.05 ± .05
	$\alpha_n = 10^{-2}$	9.1 ± 19.7	2.6 ± 3.6	1.1 ± 1.3	.40 ± .49	.21 ± .23	.06 ± .06
	$\alpha_n = 1$	10.1 ± 15.5	5.2 ± 7.0	3.5 ± 5.8	2.5 ± 4.7	1.6 ± 1.5	1.4 ± 1.5
N-VMM	$\lambda_n = 0$	9.3 ± 3.7	5.3 ± 2.8	2.8 ± 1.6	1.9 ± 1.3	1.2 ± .84	.68 ± .64
	$\lambda_n = 10^{-4}$	8.2 ± 4.0	5.4 ± 2.5	2.9 ± 1.7	1.7 ± 1.3	1.1 ± .80	.71 ± .68
	$\lambda_n = 1$	7.3 ± 2.7	4.9 ± 2.1	2.7 ± 1.9	2.0 ± 1.3	1.1 ± .84	.67 ± .68
SMD	Identity	>100	>100	>100	>100	>100	>100
	Homo	>100	>100	>100	>100	>100	>100
	Hetero	>100	>100	>100	>100	>100	>100
MMR		10.3 ± 1.9	10.2 ± 1.2	9.7 ± 1.2	9.8 ± .85	9.7 ± .70	9.6 ± .60
OWGMM		>100	>100	>100	>100	>100	>100
NCB		9.1 ± 6.7	8.8 ± 5.1	7.6 ± 3.0	7.9 ± 2.4	7.7 ± 1.2	7.4 ± .89
(c) PolicyLearning							
K-VMM	$\alpha_n = 0$	>100	>100	>100	1.6 ± 1.1	>100	>100
	$\alpha_n = 10^{-8}$	>100	>100	>100	>100	>100	>100
	$\alpha_n = 10^{-6}$	10.7 ± 13.6	1.8 ± 1.7	1.6 ± 1.0	1.6 ± .78	1.9 ± .57	2.1 ± .47
	$\alpha_n = 10^{-4}$	6.9 ± 7.4	2.1 ± 1.3	2.4 ± 1.4	2.4 ± .93	2.6 ± .65	2.8 ± .53
	$\alpha_n = 10^{-2}$	4.2 ± 3.5	3.9 ± 1.9	4.3 ± 1.6	4.1 ± 1.0	4.6 ± .74	4.8 ± .71
	$\alpha_n = 1$	6.9 ± 4.8	8.2 ± 2.8	8.7 ± 2.1	8.4 ± 1.8	8.6 ± 1.1	8.6 ± .99
N-VMM	$\lambda_n = 0$	>100	53.4 ± 88.6	6.6 ± 12.0	1.1 ± .77	.50 ± .33	.92 ± .39
	$\lambda_n = 10^{-4}$	>100	>100	7.9 ± 18.0	1.1 ± .70	.54 ± .46	.92 ± .47

(continued)

Table 1. Continued

Method	n						
	200	500	1,000	2,000	5,000	10,000	
SMD	$\lambda_n = 1$	>100	5.7 ± 8.7	1.0 ± .74	.93 ± .45	1.7 ± .67	2.0 ± .40
	Identity	>100	>100	43.3 ± 97.6	22.0 ± 21.0	>100	>100
	Homo	24.5 ± 17.8	43.1 ± 83.0	>100	26.4 ± 48.3	28.4 ± 40.5	32.5 ± 63.9
	Hetero	>100	>100	>100	>100	>100	>100
MMR		9.2 ± 6.0	10.7 ± 4.1	11.6 ± 3.5	12.3 ± 3.3	13.2 ± 2.7	13.4 ± 2.3
OWGMM		>100	10.8 ± 38.0	>100	20.7 ± 78.7	25.9 ± 87.7	5.2 ± 7.0
NCB		>100	>100	>100	>100	>100	>100

Note. We write >100 whenever the MSE or standard error was greater than 100.

suggests that VMM can generally do very well with generic choices for all hyperparameters and is not very sensitive to these choices as long as they do not take extreme values.

In the SimpleIV scenario, where $\mathbb{E}[\rho(X; \theta) | Z]$ is very simple and easy to fit uniformly over Θ , the SMD and OWGMM baselines performed competitively with our VMM algorithms. However, in the other more challenging scenarios, their behaviour was generally inconsistent and poor. We note that although the average squared error obtained by these methods was extremely high, this seems to be mostly dominated by some outliers, and the typical performance was much more reasonable. For example, in the HeteroskedasticIV scenario when $n = 10,000$, the median squared error of the Identity, Homoskedastic, and Heteroskedastic versions of SMD were 48.3, 10.7, and 0.22, respectively, which is much less bad than the average squared error. This is also evident, for example, from the separate bias and standard deviation results in the [Online Supplementary Material](#). We also note that for both SMD and OWGMM algorithms, we experimented with a wide range of choices for the underlying sieve basis sets, including the number of knots and polynomial degree for the B-splines that we used, as well as ridge-regularization values, and the results presented are for the least-bad choices. We speculate that the superior performance of our approach is due to the kernel-based regularization of the critic class, which in practice is better able to approximate the efficient instruments with good accuracy and stability. Indeed, it is plausible that sieve-based approaches could also achieve competitive performance using better choices of sieves, with appropriate regularization. In general, however, the use of such sieve spaces, rather than simple linear sieves with optional ridge regularization, as we experimented with, is either intractable or redundant. In the case of SMD, the corresponding sieve estimates $q_n(z; \theta)$ for $\mathbb{E}[\rho(X; \theta) | Z = z]$ would no longer have closed-form solutions in θ in general, and we would somehow have to solve a bi-level optimization problem. On the other hand, if we were to introduce such regularization to the sieve space that implicitly arises from the variational reformulation of OWGMM, we would just end up with our VMM approach as in Equation (2).

On the other hand, we see that the MMR baseline performed in a way that was relatively very stable, but consistently sub-optimal. The results of MMR were in general similar to kernel VMM with the largest choices for α_n , which is expected given that it is equivalent to kernel VMM with $\alpha_n \rightarrow \infty$. In addition, as expected, the non-causal baseline is consistently biased with very poor performance.

Finally, we provide a breakdown of these mean squared error results in terms of bias and variance in the [Online Supplementary Material](#). One interesting observation there is that some methods, in particular our neural VMM algorithm and the OWGMM baseline, do not display the expected behaviour of bias vanishing at a more rapid rate than standard deviation; rather, even though both shrink, their ratio often remains approximately constant. This could be explained by a couple of factors. First, in the case of neural VMM we are not exactly optimizing the minimax optimization problem; rather, we are trying to approximate this using an alternating gradient ascent/descent approach. Therefore, such discrepancies may be explained by this deviation from theory in the practical implementation of the algorithm. Note that issue does not exist for kernel

VMM, which performs the optimization over \mathcal{F}_n analytically. Second, in the case of OWGMM, this discrepancy seems to be explained by the instability and poor performance described above. This could be interpreted as ‘finite sample’ behaviour, reflecting the fact that we are not yet in the asymptotic regime for this method. Alternatively, it may reflect intractable bias due to approximation errors of the sieve basis for the efficient instruments.

7.2 Inference experiments

7.2.1 Inference scenarios

SimpleIV. Our first considered scenario for our inference experiments is based on the same SimpleIV scenario as in our estimation experiments. For this scenario here, our target for inference is the instantaneous treatment effect at $T=0$; that is, we wish to estimate $\psi(\theta_0)$ where $\psi(\theta) = \frac{\partial}{\partial t} g(t; \theta)|_{t=0} = \theta_2$.

HeteroskedasticIV. For our second inference scenario, we consider again the same HeteroskedasticIV scenario from our prior estimation experiments. Here, our target for inference is the change in slope in the true hinge function $g(\cdot; \theta_0)$. This corresponds to the function $\psi(\theta) = \theta_4 - \theta_3$.

7.2.2 Inference methods

Kernel inference. For our kernel inference method, we implemented the algorithm described by Theorem 5. We used the same kernel function as for our Kernel VMM estimation algorithm in our prior estimation experiments, and we present results for a variety of values of α_n .

Neural inference. For our neural inference method, we solved the game objective described by Equation (9). We used the same choice of \mathcal{F}_n and a similar alternating SGD optimization procedure as for our NeuralVMM estimation method. We provide additional details in the [Online Supplementary Material](#). As in our estimation experiments, we used Frobenius norm regularization, and we present results for varying values of λ_n .

7.2.3 Inference results

For each scenario and each $n \in \{200, 2,000\}$, we repeated the following procedure 200 times: (1) we drew a training set of n random iid data points using the respective scenario’s data generating process; (2) we estimated $\hat{\theta}_n$ using each of our VMM methods; and (3) we estimate the efficient asymptotic variance using each of our inference methods and each of the estimated $\hat{\theta}_n$ as plug-ins. That is, for each random draw of data, we estimate the efficient asymptotic variance using each combination of VMM estimation method and inference method. In all cases, we compute an estimated 95% confidence interval as where \hat{v} is the estimated asymptotic variance of $\psi(\hat{\theta}_n)$ via the delta method, which we computed using the corresponding inference method as detailed in Section 5. In addition, for each combination of n , scenario, estimation method, and inference method, we computed the following summary statistics: (1) the coverage rate of our estimated confidence intervals; (2) the corresponding coverage when we adjust the confidence intervals by subtracting the bias of $\psi(\hat{\theta}_n)$ (which we estimated by $\frac{1}{200} \sum_{i=1}^{200} \psi(\hat{\theta}_n^{(i)}) - \psi(\theta_0)$, where $\hat{\theta}_n^{(i)}$ denotes the estimate from the i th replication); and (3) the 5%, 50%, and 95% percentiles of the estimated standard deviation of $\psi(\hat{\theta}_n)$ (given by $\sqrt{\hat{v}/n}$) across the 200 replications.

Given our previous results that kernel VMM performed very consistently with α_n in the range of 10^{-2} – 10^{-6} , for brevity, we only present results in the main paper for when the estimation method is kernel VMM with $\alpha_n = 10^{-4}$. However, we present additional results using other estimation methods in the Appendix. We summarize the results from this procedure in [Table 2](#).

Overall, we see that in both scenarios, the results are very good when $n = 2,000$, with very accurate estimates of the standard deviation of $\psi(\hat{\theta}_n)$, and high coverage. For the HeteroskedasticIV scenario, all inference methods produce almost perfect (95%) coverage when $n = 2,000$, and for the SimpleIV scenario the coverage is only slightly lower, and becomes very close to 95% when bias of $\hat{\theta}_n$ is taken into account.

Table 2. Inference results using kernel VMM estimation with $\alpha_n = 10^{-4}$ and different inference methods

n	Method		Cov	CovBC	PredSD(.05)	PredSD(.5)	PredSD(.95)
(a) SimpleIV; the true standard deviation over the 200 replications was 0.34 for $n = 200$ and 0.23 for $n = 2000$.							
200	Kernel	$\alpha_n = 0$	83.0	94.0	.21	.32	.52
		$\alpha_n = 10^{-8}$	83.0	94.0	.21	.32	.51
		$\alpha_n = 10^{-6}$	83.0	94.5	.21	.32	.53
		$\alpha_n = 10^{-4}$	84.5	95.5	.22	.33	.56
		$\alpha_n = 10^{-2}$	86.5	95.5	.23	.35	.62
		$\alpha_n = 1$	91.0	96.0	.25	.39	.72
	Neural	$\lambda_n = 0$	82.0	94.0	.21	.31	.48
		$\lambda_n = 10^{-4}$	81.5	94.0	.21	.31	.49
		$\lambda_n = 1$	82.5	93.5	.21	.30	.49
2000	Kernel	$\alpha_n = 0$	91.5	93.5	.19	.21	.24
		$\alpha_n = 10^{-8}$	92.0	94.0	.19	.22	.25
		$\alpha_n = 10^{-6}$	92.5	94.0	.20	.22	.24
		$\alpha_n = 10^{-4}$	92.5	94.5	.20	.22	.25
		$\alpha_n = 10^{-2}$	95.0	96.0	.21	.23	.28
		$\alpha_n = 1$	100.0	100.0	.48	.55	.87
	Neural	$\lambda_n = 0$	90.0	92.5	.19	.21	.23
		$\lambda_n = 10^{-4}$	90.5	92.5	.19	.21	.22
		$\lambda_n = 1$	90.0	92.5	.19	.21	.22
(b) HeteroskedasticIV; the true standard deviation over the 200 replications was 1.6 for $n = 200$ and 0.21 for $n = 2000$.							
200	Kernel	$\alpha_n = 0$	84.5	85.0	.35	.54	1.9
		$\alpha_n = 10^{-8}$	83.5	83.5	.37	.55	2.1
		$\alpha_n = 10^{-6}$	87.5	88.5	.42	.58	2.3
		$\alpha_n = 10^{-4}$	91.5	92.5	.49	.66	2.8
		$\alpha_n = 10^{-2}$	95.0	98.0	.59	.87	4.6
		$\alpha_n = 1$	100.0	100.0	1.4	2.5	13.3
	Neural	$\lambda_n = 0$	70.5	65.5	.24	.40	.84
		$\lambda_n = 10^{-4}$	71.5	68.0	.25	.43	.84
		$\lambda_n = 1$	70.0	66.0	.25	.42	.84
2000	Kernel	$\alpha_n = 0$	95.5	97.5	.20	.21	.24
		$\alpha_n = 10^{-8}$	95.5	97.5	.19	.21	.24
		$\alpha_n = 10^{-6}$	95.5	97.5	.20	.21	.24
		$\alpha_n = 10^{-4}$	96.0	97.5	.20	.22	.25
		$\alpha_n = 10^{-2}$	97.5	98.5	.21	.23	.27
		$\alpha_n = 1$	100.0	100.0	.47	.55	.88
	Neural	$\lambda_n = 0$	95.0	95.5	.19	.21	.22
		$\lambda_n = 10^{-4}$	94.5	95.5	.19	.21	.22
		$\lambda_n = 1$	94.5	95.5	.20	.21	.22

Note. For each inference method and value of n , we list: Cov the coverage of the respective 95% confidence intervals; CovBC the corresponding bias-corrected coverage, by subtracting the bias of $\psi(\hat{\theta}_n)$ from the confidence intervals; and PredSD(q) the q 'th percentile of the estimated standard deviation of $\psi(\hat{\theta}_n)$, for $q \in \{5, 50, 95\}$.

When $n = 200$, our inference results are slightly poorer. This likely reflects several distinct issues when n is small: the bias of $\hat{\theta}_n$ may be significant, the variance may not be well characterized by the asymptotic variance and the tails by normal tails, and the estimates of the asymptotic variance of $\psi(\hat{\theta}_n)$ may be poor. Any of these issues may lead to invalid confidence intervals and lower than expected coverage. Indeed, we can see some or all of these issues at play in our results. In [Table 2a](#) we see that coverage is very good when we account for bias, and that the range of the predicted standard deviation of $\hat{\theta}_n$ is reasonably close to the empirically observed standard deviation of 0.34, which suggests we are suffering from the first issue. Conversely, in [Table 2b](#) we see that, even accounting for bias, the coverage is lower than expected when $n = 200$, and that the range of predicted standard deviations of $\psi(\hat{\theta}_n)$ is low compared to the empirically observed standard deviation of 1.9, which suggests that we are suffering from the second and/or third issues.

Regarding the difference in performances between our inference methods, we observe that, as expected given [Theorem 5](#), larger values of a_n for our kernel method lead to wider confidence intervals. For $n = 2,000$, where our asymptotic theory seems to be more relevant, we see very overly wide confidence intervals for our kernel method when a_n is very large, with typically good results when a_n takes the same range of values that worked well for estimation in our prior experiments (i.e., in the range of 10^{-6} – 10^{-2}). This suggests that we can tune a_n for estimation, and use similar values for inference, and also that we can err on the side of caution and wider confidence intervals by using larger values of a_n . Conversely, we found our neural inference method to be very insensitive to λ_n , and in general we found that it produced relatively narrow confidence intervals, with widths similar to those from our kernel method using the smallest values of a_n .

Finally, we make a note to emphasize the fact that biased-corrected coverage values are listed merely so we can analyze, in cases where coverage is poor, to what extent this is due to bias in the estimate $\hat{\theta}_n$, versus due to poor estimates of the standard deviation of $\hat{\theta}_n$. Indeed, the bias-correction we perform is *not* something that can be done in practice, and these bias-corrected coverages should not be interpreted as actual coverages that can be obtained.

8 Related work

8.1 Methods for solving conditional moment problems

For the general conditional moment problem, one classical approach is to solve [Equation \(3\)](#) using a growing sieve basis expansion for $\{f_1, \dots, f_k\}$ based on, e.g., splines, Fourier series, or power series ([Chamberlain, 1987](#)). It would be expected, however, that such methods would suffer from curse of dimension issues and therefore their application would be limited to low-dimensional settings. Furthermore, it has been observed in past work ([Bennett & Kallus, 2020](#); [Bennett et al., 2019](#)) that methods of this kind can be very unstable and perform very poorly in comparison to VMM estimators.

A very similar method to this is the SMD approach of [Ai and Chen \(2003\)](#), which instead uses a growing sieve basis expansion to approximate the conditional function $\mathbb{E}[\rho(X; \theta) | Z = z]$ for every $\theta \in \Theta$. They propose to minimize a loss of the form $J_n(\theta) = \mathbb{E}_n[\hat{q}(Z; \theta)^\top \Gamma(Z)^- \hat{q}(Z; \theta)]$, where $\hat{q}(z; \theta)$ is the sieve estimate for $\mathbb{E}[\rho(X; \theta) | Z = z]$ and $\Gamma(z)$ is some consistent estimate of $\mathbb{E}[\rho(X; \theta_0)\rho(X; \theta_0)^\top | Z = z]$. One nice feature of this kind of approach is that it can readily handle infinite-dimensional nuisance components. In the case that θ can be partitioned as $\theta = (\beta, \gamma)$, where β is a finite-dimensional parameter of interest and γ is an infinite-dimensional functional nuisance component, [Ai and Chen \(2003\)](#) propose to model γ using a second growing sieve basis expansion, and minimize $J_n(\theta)$ over both β and the sieve coefficients for γ . There is a long line of work on the theoretical efficiency of this kind of approach, even in the presence of infinite-dimensional nuisance components ([Ai & Chen, 2003](#); [Chen & Pouzo, 2009, 2012](#)), which is something that our theory does not address. However, these methods have similar practical drawbacks to using a sieve basis expansion for OWGMM, which seems to particularly be the case when the conditional expectation function $q(z; \theta) = \mathbb{E}[\rho(X; \theta) | Z = z]$ is complex, as highlighted by the experimental results in this paper. A very similar approach was also proposed concurrently by [Newey and Powell \(2003\)](#), however their approach has the same drawbacks, and furthermore they do not address efficiency.

Another related classical approach is to solve Equation (3) using estimates of the *efficient instruments*, which are the set of b functions $\{f_1^*, \dots, f_b^*\}$ mapping \mathcal{Z} to \mathbb{R}^b , given by $f_i^*(z)_j = F^*(z)_{ij}$, where

$$F^*(z) = V(z; \theta_0)^{-1} \mathbb{E}[D(X; \theta_0) \mid Z = z].$$

Past work such as Newey (1990, 1993) provide sufficient conditions for such estimators to be efficient. However, since θ_0 is unknown, such methods require some other method for first-stage estimation of θ_0 , and are likely sensitive to the quality of this method; indeed, if the estimates of f_i^* are heavily biased due to poor first-stage estimation, it is unclear whether the corresponding moments will be sufficient for identification, let alone efficiency. By contrast, our method is guaranteed to be well behaved as long as our regularized critic class \mathcal{F}_n can approximate the optimal instruments, regardless of the quality of our first-stage estimate. Furthermore, estimators that have been previously proposed based on this approach (Newey, 1990, 1993) employ nearest neighbour or sieve methods with similar weaknesses as discussed above.

The *continuum GMM* estimators of Carrasco and Florens (2000) are theoretically closely related to our proposed kernel VMM estimators. However, the form of their proposed estimators is very different. Suppose that we define some set of functions $\{f(\cdot; t) : t \in T\}$ of the form $\mathcal{Z} \mapsto \mathbb{R}^m$ indexed by set T , and we let \mathcal{H}_T be some Hilbert space of functions in the form $T \mapsto \mathbb{R}$. In addition, define $b'_n(\theta) \in \mathcal{H}$ according to $b'_n(\theta)(t) = \mathbb{E}_n[f(Z; t)^\top \rho(X; \theta)]$, and the linear operator $C'_n : \mathcal{H}_T \mapsto \mathcal{H}_T$ according to

$$(C'_n b)(t) = \langle k_n(t, \cdot), b \rangle_{\mathcal{H}_T}, \quad \text{where} \quad k_n(t, s) = \mathbb{E}_n[f(Z; s)^\top \rho(X; \tilde{\theta}_n) \rho(X; \tilde{\theta}_n)^\top f(Z; t)],$$

and $\tilde{\theta}_n$ is some prior estimate for θ_0 . Then, Carrasco and Florens (2000) study estimators of the form $\arg \min_{\theta \in \Theta} \|((C'_n)^2 + \alpha_n I)^{-1/2} (C'_n)^{1/2} b'_n(\theta)\|_{\mathcal{H}_T}^2$. In that case, we choose T to be an RKHS class \mathcal{F} , with functions indexed by themselves, and \mathcal{H} chosen as the dual of this RKHS, then it easily follows that the terms C'_n and b'_n defined here are equivalent to the terms C_n and b_n defined in Section 3. However, the form of Tikhonov regularization applied in the inversion of $C_n^{1/2}$ is slightly different; by Lemma 2 we regularize using $(C_n + \alpha_n I)^{-1/2}$, whereas they regularized using $(C_n^2 + \alpha_n I)^{-1/2} C_n^{1/2}$. This difference is significant, since our form of regularization gives rise to the simple minimax VMM-style interpretation, whereas theirs does not. Furthermore, their proposed estimators use the index set $T = [0, t_{\max}]$ for some $t_{\max} > 0$, with \mathcal{H}_T chosen as the L_2 space on T . This choice is much less flexible than ours of using a function class as the index set and makes it more difficult to guarantee that θ_0 is uniquely identified or to guarantee semiparametric efficiency, which they do not. More concretely, the main efficiency claim they provide is that their estimator is efficient compared to other estimators of the form $\sup_{\theta \in \Theta} \|B'_n b_n(\theta)\|^2$, for any choice of bounded linear operator B'_n . Finally, they propose to solve their optimization problem by computing an explicit rank- n eigenvalue, eigenvector decomposition of C'_n , and constructing a cost function to minimize based on this decomposition. In particular, if we define $g_i(\theta) \in \mathcal{H}_T$ according to $g_i(\theta)(t) = f(Z_i; t)^\top \rho(X_i; \theta)$ for each $i \in [n]$ and $\theta \in \Theta$, then their objective function is given by a quadratic form on all terms of the form $\langle g_i(\theta), g_j(\theta) \rangle_{\mathcal{H}_T}$ for $i, j \in [n]$. This involves n^4 terms in total, and is therefore very computationally expensive to compute for large n . In comparison, the cost function in θ implied by our kernel VMM estimator could be calculated analytically as a quadratic form in n^2 terms based on the representer theorem. Furthermore, our variational reformulation allows for estimators based on alternating stochastic gradient descent, which may be more practical in some situations, for example when n is large.

Another recently proposed and related class of estimators are given by the *adversarial GMM* estimators of Lewis and Syrgkanis (2018), which were recently extended to the more general class of *minimax GMM* estimators by Dikkala et al. (2020). In general, these estimators are defined according to $\arg \min_{\theta \in \Theta} \sup_{f \in \mathcal{F}} \mathbb{E}_n[f(Z)^\top \rho(X; \theta)] + R_n(f) - \Psi_n(\theta)$, where R_n is some regularizer on f , and Ψ_n is some regularizer on θ . In particular, Dikkala et al. (2020) analyze estimators where \mathcal{F}

and Θ are both normed function spaces, and the regularizers take the form $R_n(f) = \alpha_n \|f\|_{\mathcal{F}}^2 + \lambda_n \sum_{i=1}^n \|f(Z_i)\|^2$ and $\Psi_n(\theta) = \mu_n \|\theta\|_{\Theta}^2$. On the theoretical side, they provide general results bounding the L_2 distance between $\mathbb{E}[\rho(X; \hat{\theta}_n) | Z]$ and $\mathbb{E}[\rho(X; \theta_0) | Z]$ for this form of estimator. Furthermore, they propose various specific estimators of this kind, for example with \mathcal{F} chosen as a RKHS or a class of neural networks. We note that these are similar to our proposed kernel and neural VMM estimators, with the difference that they do not include the $-(1/4)\mathbb{E}_n[(f(Z)^\top \rho(X; \hat{\theta}_n))^2]$ term motivated by optimal weighting, and that they explicitly regularize θ . In a sense, the focus of their estimators and theory is very different than ours; we focus on the question of efficiency, and provide theoretical guarantees of efficiency when Θ is finite-dimensional, whereas they focus on the case where Θ is a function space, but restrict their analysis to providing finite-sample bounds rather than addressing efficiency. We speculate that the benefits of both kinds of approaches could be combined, and by using both the optimal weighting-based term and regularizing θ one could construct estimators that are semiparametrically efficient when θ_0 is finite-dimensional, and have explicit risk guarantees in the more general setting. However, we leave this question to future work.

8.2 Methods for solving the instrumental variable regression problem

Recall that for the instrumental variable regression problem we have $X = (Z, T, Y)$, where T is the treatment we are regressing on, Y is the outcome, and Z is the instrumental variable, and $\rho(X; \theta) = Y - g(T; \theta)$, for some regression function g parameterized by θ . In this setup, Θ may either be a finite-dimensional parameter space, which corresponds to having a parametric model for g , or alternatively we may allow Θ to be some infinite-dimensional function space and simply define $g(z; \theta) = \theta(z)$, which corresponds to performing nonparametric regression.

Perhaps, the most classic method for instrumental variable regression is two-stage least squares (2SLS). First, we perform least-squares linear regression of $\phi(T)$ on $\psi(Z)$, where ϕ and ψ are finite-dimensional feature maps on T and Z , respectively. That is, we learn some linear model $b(\cdot; \hat{\gamma}_n)$, where $b(z; \gamma) = \gamma^\top \psi(z)$, and $\hat{\gamma}_n = \arg \min_{\gamma} \sum_{i=1}^n \|\phi(T_i) - \gamma^\top \psi(Z_i)\|^2$. Then, we again perform least squares linear regression, this time of Y on $b(\psi(Z); \hat{\gamma}_n)$. That is, we learn a linear model $g(\cdot; \hat{\theta}_n)$, where $g(t; \theta) = \theta^\top \phi(t)$, and $\hat{\theta}_n = \arg \min_{\theta} \sum_{i=1}^n (Y_i - \theta^\top \phi(T_i; \hat{\gamma}_n))^2$. Under the assumption that these linear models are correctly specified, then the resulting 2SLS estimator is known to be consistent for θ_0 (Angrist & Pischke, 2008, Section 4.1.1). However, such estimators are limited in that they require finding some finite-dimensional feature map ϕ such that the linear model given above is well-specified, which in practice may be infeasible. The sieve methods of Newey and Powell (2003), Ai and Chen (2003) discussed in Section 8.1 applied specifically to the instrumental variable regression problem could be viewed as similar approaches, but using growing sieve basis expansions for ϕ and ψ . However, as discussed already these methods may be problematic in practice.

Alternatively, a couple of recent works propose extending the 2SLS method in the case where both stages are performed using infinite-dimensional feature maps and ridge regularization; i.e., both stages are performed using kernel ridge regression. The *Kernel IV* method of Singh et al. (2019) proposes to do this in a very direct way, by regressing $\phi(T)$ on $\psi(Z)$, and then regressing Y on $b(\phi(Z))$, where both the feature maps ϕ and ψ are infinite dimensional, and implicitly defined by some kernels K_Z and K_T under Mercer's theorem. In the case of learning b , this corresponds to solving for a linear operator between two RKHSs and in general is ill-posed, so this regression is performed using Tikhonov regularization. Then, the second-stage problem corresponds to performing RKHS regression using some implicit kernel depending on b , and is performed again using Tikhonov regularization. Ultimately, however, by appealing to the representer theorem the regressions don't need to be performed separately, and there is a simple closed form solution. Similarly, the *Dual IV* method of Muandet et al. (2019) considers 2SLS using RKHSs for each stage and formulates this as a minimax problem of the form $\arg \min_{\theta \in \Theta} \sup_{f \in \mathcal{F}} \mathbb{E}_n[(g(T; \theta) - Y) f(Y, Z)] - (1/2)\mathbb{E}_n[f(Y, Z)^2]$. Ultimately, both this work and that of Singh et al. (2019) propose closed-form estimators that are superficially similar to ours in Lemma 7, but without any terms corresponding to optimal weighting. However, their focus is slightly different to ours; their theoretical analysis where present is in terms of consistency or regret, whereas the focus of our theoretical analysis is semiparametric efficiency.

The recent *Deep IV* method of [Hartford et al. \(2017\)](#) proposes to extend 2SLS using deep learning. Specifically, they propose in the first stage to fit the conditional distribution of X given Z , for example using a mixture of Gaussians parametrized by neural networks, or by fitting a generative model using some other methodology such as generative adversarial networks or variational autoencoders. Then, in the second stage, they propose to minimize $(1/n) \sum_{i=1}^n (Y_i - \hat{\mathbb{E}}[g(X; \theta) | Z_i])^2$, where the conditional expectation $\hat{\mathbb{E}}[\cdot | z]$ is estimated using the model from the first stage, and g is parameterized using some neural network architecture. This approach has the advantage of being flexible and building on recent advances in deep learning, however they do not provide any concrete theoretical characterizations, and since the first stage is bound to be imperfectly specified this can suffer from the ‘forbidden regression’ issue ([Angrist & Pischke, 2008](#), Section 4.6.1).

[Zhang et al. \(2020\)](#) recently proposed the *maximum moment restriction instrumental variable* algorithm. They present multiple estimators for approximately solving $\arg \min_{\theta \in \Theta} \sup_{f: \|f\|_{\mathcal{F}} \leq 1} \mathbb{E}[f(Z)(W - g(T; \theta))] - \Psi_n(\theta)$, where \mathcal{F} is an RKHS, and Ψ_n is an optional regularizer on θ in the case that it is infinite-dimensional (however, they also analyze case where θ is finite-dimensional.) Of particular note, the ‘V-statistic’ version of their algorithm is equivalent to minimizing $J_n^{\text{MMR}}(\theta) = (1/n^2)\rho(\theta)^\top L\rho(\theta) + \Psi_n(\theta)$, where $\rho(\theta)$ and L are defined as in Lemma 6. Letting $J_n^{\text{K-VMM}}(\theta; \alpha)$ denote our kernel VMM objective with regularization strength α , and assuming $\Psi_n(\theta) = 0$, Lemma 6 immediately implies that $\alpha J_n^{\text{K-VMM}}(\theta; \alpha) \rightarrow J_n^{\text{MMR}}(\theta)$ as $\alpha \rightarrow \infty$. In other words, there is an equivalence between MMR and kernel VMM with infinite regularization. [Zhang et al. \(2020\)](#) provide theory showing that their estimators are consistent and asymptotically normal under various assumptions. However, unlike us, they do *not* establish efficiency.

8.3 Applications of VMM estimators

Finally, we discuss some past work where VMM estimators have been applied. The original such work was by [Bennett et al. \(2019\)](#), who proposed the *DeepGMM* estimator for the problem of instrumental variable regression. Specifically, the proposed estimator takes the form $\arg \min_{\theta} \sup_{f \in \mathcal{F}} \mathbb{E}_n[f(Z)^\top (Y - g(T; \theta))] - (1/4)\mathbb{E}_n[f(Z)^2 (Y - g(T; \tilde{\theta}_n))^2]$, where $\{g(\cdot; \theta) : \theta \in \Theta\}$ and \mathcal{F} are both given by neural network function classes. That is, the DeepGMM estimator can be interpreted as a neural VMM estimator for the instrumental variable problem in the form of equation (6) with $R_n(f) = 0$ and fixed \mathcal{F}_n that does not grow with n . In their experiments DeepGMM consistently outperformed other recently proposed methods ([Hartford et al., 2017](#); [Lewis & Syrgkanis, 2018](#)) across a variety of simple low-dimensional scenarios, and it was the only method to continue working when using high-dimensional data where the treatments and instruments were images. In addition, DeepGMM has continued to perform competitively in more recent experimental comparisons ([Muandet et al., 2019](#); [Singh et al., 2019](#)). [Bennett et al. \(2019, Theorem 2\)](#) provided conditions under which DeepGM is consistent. In addition, we could also justify that it is asymptotically normal and semiparametrically efficient by Theorem 4, under some additional assumptions and by introducing kernel-based regularization.

In addition, this style of estimator was applied to the problem of policy learning from convex surrogate loss reductions by [Bennett and Kallus \(2020\)](#). A common approach for optimizing binary treatment decision policies from logged cross-sectional data is to construct a surrogate cost function to minimize of the form $\mathbb{E}_n[|\psi|l(g(X; \theta), \text{sign}(\psi))]$, where X denotes observed pre-treatment information about the individual, ψ is some weighting variable depending on all observed pre- and post-treatment information about the individual, the function $g(\cdot; \theta)$ encodes the policy we are optimizing which we assume is parameterized by $\theta \in \Theta$, and l is some smooth convex loss function such as logistic regression loss. [Bennett and Kallus \(2020\)](#) showed that the model where this surrogate loss is correctly specified is given by the conditional moment problem $\mathbb{E}[|\psi|l'(g(X; \theta), \text{sign}(\psi)) | X] = 0$, where l' is the derivative of l with respect to its first argument. Consequently, they proposed the empirical surrogate loss policy risk minimization (ESPRM) estimator, according to $\arg \min_{\theta \in \Theta} \sup_{f \in \mathcal{F}} \mathbb{E}_n[f(X)\rho(X, \psi; \theta)] - (1/4)\mathbb{E}_n[f(X)^2\rho(X, \psi; \tilde{\theta}_n)]$, where $\rho(X, \psi; \theta) = |\psi|l'(g(X; \theta), \text{sign}(\psi))$, and \mathcal{F} is a neural network function classes. That is, again this estimator can be interpreted as a neural VMM estimator as in equation (6), with $R_n(f) = 0$. Not only did the authors demonstrate that this algorithm led to consistently improved empirical performance over the standard approach of empirical risk minimization using the surrogate loss,

but they proved that if the resulting estimator $\hat{\theta}_n$ is semiparametrically efficient, then this implies optimal asymptotic regret for the learnt policy compared with *any* policy identified by the model given by correct specification. We note that, although the authors did not address the question of how to guarantee such efficiency for $\hat{\theta}_n$, we could guarantee it by Theorem 4 under some additional assumptions and kernel-based regularization, or under Theorem 3 by instead using a kernel VMM estimator.

Finally, Bennett et al. (2021) applied this style of estimator to the problem of reinforcement learning using offline data logged from some fixed behaviour policy, also known as the problem of off policy evaluation (OPE). They proposed an algorithm for the OPE problem under unmeasured confounding, which requires as an input an estimate of the state density ratio d between the behaviour policy and the target policy they are evaluating. As stated in Section 1, d can be identified by a conditional moment problem, up to a constant factor, with the normalization constraint $\mathbb{E}[d(S)] = 1$. Bennett et al. (2021) proposed a VMM-style estimator for d , using both the conditional moment condition $\mathbb{E}[d(S)\beta(A, S) - d(S') | S'] = 0$ and the marginal moment condition $\mathbb{E}[d(S) - 1] = 0$, based on a slightly more general form of Lemma 1 where the vector of conditional moment restrictions can depend on different random variables to be conditioned on. That is, the more general problem is given by the m moment conditions $\mathbb{E}[\rho_i(X; \theta_0) | Z_i] = 0$ for $i \in [m]$, for some set of random variables $Z_1 \in \mathcal{Z}_1, \dots, Z_m \in \mathcal{Z}_m$. Specifically, they propose a kernel VMM-style estimator, where both d and f are optimized over balls in RKHSs. In practice, by successively applying the representer theorem to this two-stage optimization problem, they presented a closed-form solution for the estimate \hat{d}_n (in a similar vein to Lemma 7). Note that since their kernel VMM estimator is based on a slightly more intricate conditional moment formulation than we considered, with varying conditioning sets, our theoretical analysis may not apply to it. We leave the question of extending our theoretical analysis to this more general problem to future work.

9 Conclusion

In this paper, we presented a detailed theoretical analysis for the class of VMM estimators, which are motivated by a variational reformulation of the optimally weighted generalized method of moments and which encompass several recently proposed estimators for solving conditional moment problems. We studied multiple varieties of these estimators based on kernel methods or deep learning, and provided appropriate conditions under which these estimators are consistent, asymptotically normal, and semiparametrically efficient. This is in contrast to other recently proposed approaches for solving conditional moment problems using machine learning tools, which do not provide any results regarding efficiency. In addition, we proposed inference algorithms based on the same kind of variational reformulation, again with specific algorithms based on both kernel methods and deep learning. Finally, we demonstrated in a detailed series of experiments that our VMM estimators achieve very strong estimation performance in comparison to relevant baselines and that the confidence intervals we generate are reliable.

Our paper suggests a few immediate directions for future work. First, unlike, e.g., the sieve minimum distance approaches of Ai and Chen (2003); Chen and Pouzo (2009, 2012), our efficiency theory when θ_0 is finite-dimensional does not accommodate possible infinite-dimensional nuisance components. Furthermore, as discussed in Section 3, the latter two works allow for weaker assumptions on the smoothness and complexity of $\rho(X; \theta)$. We suspect that our theory could be extended accordingly without fundamentally changing the VMM algorithm, but this is left to future work.

Second, we only consider conditional moment restrictions using a single conditioning variable Z . In some settings, such as longitudinal studies or the RL application discussed in Section 8.3, one faces conditional moment problems with different, nested conditioning variables for each conditional moment restriction, and our current theory does not accommodate such formulations. Again, we believe that our theory could naturally be extended to this kind of setting.

Third, we only present theory for neural VMM estimators using a kernel-based regularizer, yet we see compelling empirical results for simpler regularizers. We speculate that under appropriate conditions on the neural net classes \mathcal{F}_n , our efficiency result in Theorem 4 could be extended to neural VMM estimators with such regularizers.

Next, an important further direction is the automatic selection of the hyperparameter α_n for our kernel VMM method and corresponding inference algorithm. We speculate, for instance, that it may be possible to approximate the resulting bias and variance for different values of α_n and optimize a bias-variance trade-off. At the same time, work on approximating the bias of our estimator could be helpful for improving the quality of confidence intervals from our proposed inference algorithm, as we observed that in many cases coverage of our confidence intervals significantly improved when they were corrected for bias. Similarly, it is known that continuously updating GMM can have lower bias than k -step GMM algorithms (Hansen et al., 1996), which suggests that we may be able to reduce bias using a continuously updating VMM where instead of using a prior estimate $\hat{\theta}_n$ in the second term of the game objective we use the same θ that we are optimizing over.

Finally, we hope that this work will help motivate the construction of efficient VMM estimators for other conditional moment problems.

Supplementary material

Supplementary material are available at *Journal of the Royal Statistical Society: Series B* online.

Conflict of interest: None declared.

Funding

This material is based upon work supported by the National Science Foundation under Grant No. 1846210.

Data availability

The data underlying this article were synthetically generated, and can be reproduced using the code in our repository at <https://github.com/CausalML/VMM>.

References

- Ai C., & Chen X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6), 1795–1843. <https://doi.org/10.1111/1468-0262.00470>
- Angrist J. D., & Pischke J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.
- Bennett A., & Kallus N. (2020). Efficient policy learning from surrogate-loss classification reductions. In *International Conference on Machine Learning* (pp. 788–798). PMLR.
- Bennett A., Kallus N., Li L., & Mousavi A. (2021). Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. In *International Conference on Artificial Intelligence and Statistics* (pp. 1999–2007). PMLR.
- Bennett A., Kallus N., & Schnabel T. (2019). Deep generalized method of moments for instrumental variable analysis. In *Advances in Neural Information Processing Systems* (pp. 3559–3569). Curran Associates, Inc.
- Berry S., Levinsohn J., & Pakes A. (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, 63(4), 841–890. <https://doi.org/10.2307/2171802>
- Carrasco M., & Florens J.-P. (2000). Generalization of GMM to a continuum of moment conditions. *Econometric Theory*, 16(6), 797–834. <https://doi.org/10.1017/S0266466600166010>
- Chamberlain G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3), 305–334. [https://doi.org/10.1016/0304-4076\(87\)90015-7](https://doi.org/10.1016/0304-4076(87)90015-7)
- Chen X., & Pouzo D. (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*, 152(1), 46–60. <https://doi.org/10.1016/j.jeconom.2009.02.002>
- Chen X., & Pouzo D. (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1), 277–321. <https://doi.org/10.3982/ECTA7888>
- Chernozhukov V., Imbens G. W., & Newey W. K. (2007). Instrumental variable estimation of nonseparable models. *Journal of Econometrics*, 139(1), 4–14. <https://doi.org/10.1016/j.jeconom.2006.06.002>
- Cucker F., & Smale S. (2002). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1), 1–49. <https://doi.org/10.1090/S0273-0979-01-00923-5>
- Daskalakis C., Ilyas A., Syrgkanis V., & Zeng H. (2017). ‘Training GANS with optimism’, arXiv, arXiv:1711.00141, preprint: ICLR. <https://openreview.net/forum?id=SJJySbbAZ>

- Dikkala N., Lewis G., Mackey L., & Syrgkanis V. (2020). Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 33, 12248–12262.
- Fiez T., Chasnov B., & Ratliff L. (2020). Implicit learning dynamics in stackelberg games: Equilibria characterization, convergence analysis, and empirical study. In *International Conference on Machine Learning* (pp. 3133–3144). PMLR.
- Gidel G., Hemmat R. A., Pezeshki M., Le Priol R., Huang G., Lacoste-Julien S., & Mitliagkas I. (2019). Negative momentum for improved game dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics* (pp. 1802–1811). PMLR.
- Hansen L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4), 1029–1054. <https://doi.org/10.2307/1912775>
- Hansen L. P., Heaton J., & Yaron A. (1996). Finite-sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics*, 14(3), 262–280. <https://doi.org/10.2307/1392442>
- Hartford J., Lewis G., Leyton-Brown K., & Taddy M. (2017). Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (pp. 1414–1423). JMLR.org.
- Horowitz J. L., & Lee S. (2007). Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica*, 75(4), 1191–1208. <https://doi.org/10.1111/j.1468-0262.2007.00786.x>
- Kallus N., Mao X., & Uehara M. (2021). ‘Causal inference under unmeasured confounding with negative controls: A minimax learning approach’, arXiv, arXiv:2103.14029, preprint: not peer reviewed.
- Kallus N., & Uehara M. (2022). Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research*, 70(6), 3282–3302. <https://doi.org/10.1287/opre.2021.2249>
- Kosorok M. R. (2007). *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media.
- Lewis G., & Syrgkanis V. (2018). ‘Adversarial generalized method of moments’, arXiv, arXiv:1803.07164, preprint: not peer reviewed.
- Lin T., Jin C., & Jordan M. I. (2020a). Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory* (pp. 2738–2779). PMLR.
- Lin T., Jin C., & Jordan M. I. (2020b). On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning* (pp. 6083–6093). PMLR.
- Liu Q., Li L., Tang Z., & Zhou D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems* (pp. 5356–5366). Curran Associates, Inc.
- Loizou N., Berard H., Jolicoeur-Martineau A., Vincent P., Lacoste-Julien S., & Mitliagkas I. (2020). Stochastic Hamiltonian gradient methods for smooth games. In *International Conference on Machine Learning* (pp. 6370–6381) PMLR.
- Muandet K., Mehrjou A., Lee S. K., & Raj A. (2019). ‘Dual IV: A single stage instrumental variable regression’, arXiv, arXiv:1910.12358, preprint: NeurIPS: <https://proceedings.neurips.cc/paper/2020/hash/1c383cd30b7c298ab50293adfecb7b18-Abstract.html>
- Newey W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica: Journal of the Econometric Society*, 58(4), 809–837. <https://doi.org/10.2307/2938351>
- Newey W. K. (1993). Efficient estimation of models with conditional moment restrictions. In *Econometrics of Handbook of Statistics*, (Vol. 11), chapter 16 (pp. 419–454). Elsevier.
- Newey W. K., & Powell J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, 55(4), 819–847. <https://doi.org/10.2307/1911031>
- Newey W. K., & Powell J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5), 1565–1578. <https://doi.org/10.1111/1468-0262.00459>
- Shankar V., Fang A., Guo W., Fridovich-Keil S., Ragan-Kelley J., Schmidt L., & Recht B. (2020). Neural kernels without tangents. In *International Conference on Machine Learning* (pp. 8614–8623). PMLR.
- Singh R., Sahani M., & Gretton A. (2019). Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems* (pp. 4593–4605). Curran Associates, Inc.
- Sobotka F., Radice R., Marra G., & Kneib T. (2013). Estimating the relationship between women’s education and fertility in botswana by using an instrumental variable approach to semiparametric expectile regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1), 25–45.
- Sriperumbudur B. K., Fukumizu K., & Lanckriet G. R. (2011). Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(7), 2389–2410.
- Thekumparampil K. K., Jain P., Netrapalli P., & Oh S. (2019). Efficient algorithms for smooth minimax optimization. *Advances in Neural Information Processing Systems*, 32, 12680–12691.
- Uehara M., Imaizumi M., Jiang N., Kallus N., Sun W., & Xie T. (2021). ‘Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency’, arXiv, arXiv:2102.02981, preprint: not peer reviewed.
- Van der Vaart A. W. (2000). *Asymptotic statistics*. (Vol. 3). Cambridge University Press.

- Yarotsky D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94, 103–114. <https://doi.org/10.1016/j.neunet.2017.07.002>
- Yarotsky D. (2018). ‘Optimal approximation of continuous functions by very deep ReLU networks’, arXiv, arXiv:1802.03620, preprint: COLT: <https://proceedings.mlr.press/v75/yarotsky18a.html>
- Zhang R., Imaizumi M., Schölkopf B., & Muandet K. (2020). ‘Maximum moment restriction for instrumental variable regression’, arXiv, arXiv:2010.07684, preprint: not peer reviewed.