Geometry-Aware Instrumental Variable Regression

Heiner Kremer [1] **Bernhard Schölkopf** [1]

Abstract

Instrumental variable (IV) regression can be approached through its formulation in terms of conditional moment restrictions (CMR). Building on variants of the generalized method of moments, most CMR estimators are implicitly based on approximating the population data distribution via reweightings of the empirical sample. While for large sample sizes, in the independent identically distributed (IID) setting, reweightings can provide sufficient flexibility, they might fail to capture the relevant information in presence of corrupted data or data prone to adversarial attacks. To address these shortcomings, we propose the Sinkhorn Method of Moments, an optimal transport-based IV estimator that takes into account the geometry of the data manifold through data-derivative information. We provide a simple plug-and-play implementation of our method that performs on par with related estimators in standard settings but improves robustness against data corruption and adversarial attacks.

1. Introduction

Instrumental variable regression is one of the most widespread approaches for learning in presence of confounding (Angrist & Pischke, 2008). It is applicable in situation where one is interested in inferring the outcome Y of some treatment T, where both, treatment and outcome, are affected by a so-called unobserved confounder U. To eliminate the confounding bias, one can take into account an instrumental variable Z, which i) affects the treatment T, ii) affects the outcome Y only through its effect on T, and, iii) is independent of the confounder U. While traditionally the problem has been addressed through the 2-stage least squares approach (Angrist & Pischke, 2008), in recent years the formulation in terms of conditional moment restrictions (CMR) has gained popularity for its potential to benefit from advances in machine learning models (Bennett et al., 2019; Dikkala et al., 2020; Muandet et al., 2020; Kremer et al., 2022; 2023; Bennett & Kallus, 2023; Zhang et al., 2023). The CMR formulation of IV regression is based on restricting the expectation of the prediction residual $Y - f(T)$ conditioned on the instruments Z, where f denotes a causal relation from T to Y that one wants to infer. In general, this leads to a zero-sum game in which one minimizes an objective with respect to the model parameters and maximizes it with respect to an adversary function that detects the moment violations (Bennett et al., 2019; Dikkala et al., 2020). One of the most general frameworks for learning with moment restrictions is the family of generalized empirical likelihood (GEL) estimators (Owen, 1988; Qin & Lawless, 1994; Kitamura & Stutzer, 1997; Imbens et al., 1998; Owen, 2001), which includes the prominent generalized method of moments (Hansen, 1982; Hansen et al., 1996; Hall, 2004). The idea behind empirical likelihood is to learn a model via maximum likelihood estimation without specifying a parametric form of the data distribution (Owen, 2001). In practice, this is realized by learning a non-parametric approximation of the population data distribution P_0 along with the model f by means of minimizing a φ-divergence under the moment restrictions. However, by relying on φ-divergences one effectively restricts the estimator of the population distribution to reweightings of the sample. The reweighting assumption has recently been lifted by Kremer et al. (2023) by introducing an estimator based on maximum mean discrepancy (Gretton et al., 2012).

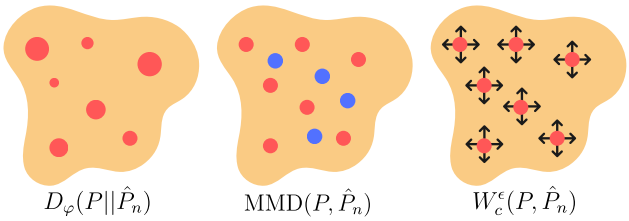


Figure 1. Paradigms to approximate P_0 from data (red dots) in the GEL framework. φ-divergence-based estimators (left) approximate P_0 by reweighting (weight $\hat{=}$ size) the sample (e.g., (Ai & Chen, 2003; Bennett & Kallus, 2023). MMD-based estimators (middle) allow to sample additional data points (blue dots) (Kremer et al., 2023). In contrast, optimal transport-based estimators (right) allow to move around the data points (present work).

[1]Max Planck Institute for Intelligent Systems. Correspondence to: Heiner Kremer <heiner.kremer@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Their estimator allows for more fine-grained approximations of P_0 by sampling additional data points from a generative model. While reweightings of the present data or sampling of additional points might be suitable to find sufficiently close approximations of the population distribution in some cases, in presence of highly complex data manifolds, e.g., image spaces, they might become ineffective as they are blind towards the geometry of the data space. This is particularly relevant in the presence of poisoned (Chen et al., 2017) or adversarial (Goodfellow et al., 2014) data points, i.e., data that has been corrupted with small perturbations which lead to vastly inaccurate predictions. The key to robustness against such perturbations is to look at how the learning signal changes around the empirical data points, i.e., to take into account the geometry of the signal with respect to the data manifold. We implement the idea of a geometry-aware learning with conditional moment restrictions by proposing an empirical likelihood-type estimator based on a regularized optimal transport distance, which we call the Sinkhorn Method of Moments (SMM). Figure 1 schematically compares our method to previous approaches to empirical likelihood estimation.

Our Contributions

- We propose the Sinkhorn Method of Moments (SMM), the first geometry-aware approach to IV regression resulting from an empirical likelihood-type estimator based on the Sinkhorn distance.
- We derive the dual form of our estimator and a leading order expansion that lets us compute our estimator with stochastic gradient methods.
- We show that under standard assumptions, our method is consistent for models identified via conditional moment restrictions.
- We derive a kernel-based implementation of our method that can be interpreted as a geometry-aware variant of a 2-stage generalized method of moments estimator for conditional moment restrictions.
- Our experiments demonstrate that SMM is competitive with state-of-the-art IV estimators in standard settings and can provide an improvement in presence of corrupted data and adversarial examples.

The remainder of the paper is structured as follows. Section 2 introduces empirical likelihood estimation for conditional moment restrictions, followed by the derivation of our method and its theoretical properties in Section 3. Empirical results are provided in Section 4 and related work is discussed in Section 5.

2. Empirical Likelihood Estimation for CMR

In the following let T, Y and Z denote random variables taking values in $\mathcal{T} \subseteq \mathbb{R}^{d_t}$, $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ and $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$ respectively. We denote by $E_P[\cdot]$ the expectation operator with respect to a distribution P and drop the subscript whenever we refer to the population distribution P_0.

Conditional moment restrictions identify a function of interest $f_0 \in \mathcal{F}$ by restricting the conditional expectation of a so-called moment function $\psi : \mathcal{T} \times \mathcal{Y} \times \mathcal{F} \to \mathbb{R}^m$,

$$E[\psi(T, Y; f_0)|Z] = 0 \; P_Z\text{−a.s.} \tag{1}$$

The most prominent example of this problem is instrumental variable (IV) regression, where the moment function is given by the prediction residual $\psi(t, y; f) = y - f(t)$ and the conditioning variable Z denotes the instrument. IV regression is one of the major practical approaches to deal with endogenous variables (Pearl, 2000) and has been largely adopted by the causal machine learning community (Hartford et al., 2017; Singh et al., 2019; Xu et al., 2021; Saengkyongam et al., 2022; Zhang et al., 2023).

Learning with conditional moment restrictions is challenging mostly due to two factors. The first one is that equation (1) contains a *conditional* expectation over the treatments T and outcomes Y, while one generally has access to a sample from the *joint* distribution over $(T, Y, Z) \sim P_0$. For a sufficiently complex data generating process the accurate estimation of a conditional distribution from the corresponding joint distribution can require large amounts of data (Hall et al., 1999). This can be avoided by rewriting the CMR (1) in terms of an equivalent *variational* formulation (Bierens, 1982)

$$E[\psi(T, Y; f_0)^T h(Z)] = 0 \; \forall h \in \mathcal{H}, \tag{2}$$

where $\mathcal{H}$ is a sufficiently rich function space, e.g., the space of square-integrable functions (Bierens, 1982) or the reproducing kernel Hilbert space of a certain kind of kernel (Kremer et al., 2022). While (2) avoids the conditional expectation operator, it involves an infinite-dimensional over-determined system of equations. The second difficulty is the fact that the moment restrictions identify the function of interest f_0 via the *population* distribution P_0 of the data, about which one usually only has partial information in terms of a sample $\mathcal{D} = \{(t_i, y_i, z_i)\}_{i=1}^n$ with empirical distribution $\hat{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{(t_i,y_i,z_i)}$, where $\delta_{(t_i,y_i,z_i)}$ denotes a point mass centered at (t_i, y_i, z_i). While the true function f_0 is identified by the population moment restrictions (2), it might not satisfy the empirical counterpart of (2) and thus one might not retrieve f_0 by enforcing it. Empirical likelihood estimation (Owen, 1988; 1990; Qin & Lawless, 1994) has been proposed as a flexible tool to solve over-determined moment restriction problems with access to only a finite sample. The idea is based on approximating

the population distribution by seeking a distribution with minimal distance to the empirical one for which the moment restrictions can be fulfilled. We visualize this approach in Figure 2. The standard generalized empirical likelihood estimator (Qin & Lawless, 1994) with the extension to conditional moment restrictions of Kremer et al. (2022) takes the form $f^{\mathrm{FGEL}} = \arg\min_{f\in\mathcal{F}} R(f)$ with

$$R(f) = \min_{P\in\mathcal{P}_n} D_\varphi(P||\hat{P}_n) \quad \text{s.t.} \quad E_P[\psi(T,Y;f)^T h(Z)] = 0 \ \ \forall h \in \mathcal{H},$$

where the distance $D_\varphi(P||Q) = \int \varphi\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}Q$ denotes the φ-divergence between distributions P and Q and the set $\mathcal{P}_n := \{P \ll \hat{P}_n : E_P[1] = 1\}$ contains all distributions which are absolutely continuous with respect to the empirical one, i.e., reweightings of the data points.

3. Sinkhorn Method of Moments

The goal of this work is to extend the idea of empirical likelihood estimation to optimal transport distances. Before deriving the method, we provide a brief introduction to optimal transport. Consider the random variable $\xi := (T, Y, Z)$ taking values in $\Xi := \mathcal{T}\times\mathcal{Y}\times\mathcal{Z} \subseteq \mathbb{R}^{d_\xi}$, with $d_\xi = d_t + d_y + d_z$, and let $\mathcal{P}(\Xi)$ denote the space of probability distributions over Ξ.

Optimal Transport Optimal transport provides an intuitive way of comparing two distributions by means of measuring the minimum effort of transforming one to another by moving probability mass at a certain cost. Let $P \in \mathcal{P}(\Xi)$ and $Q \in \mathcal{P}(\Xi)$ denote two probability distributions over Ξ with densities or probability mass functions (pmf) p and q respectively. Let $\Pi(P,Q) \subset \mathcal{P}(\Xi\times\Xi)$ denote the space of joint probability distributions over the product space $\Xi\times\Xi$ with marginals P and Q. Define the projection operators $\mathbb{P}_1$ and $\mathbb{P}_2$ with $\mathbb{P}_1(x,y) = x$ and $\mathbb{P}_2(x,y) = y$ and their push-forward operation $\mathbb{P}_{i\sharp}$ such that for any element of $\Pi(P,Q)$, with density (or pmf) π we have $\mathbb{P}_{1\sharp}\pi = \int \pi(\xi,\xi')\mathrm{d}\xi' = p(\xi)$ and $\mathbb{P}_{2\sharp}\pi = \int \pi(\xi,\xi')\mathrm{d}\xi = q(\xi')$. Then, for a cost function $c : \Xi\times\Xi \to \mathbb{R}$ we can define the Wasserstein distance between P and Q in the Kantorovich formulation as $W_c(P,Q) := \min_{\pi\in\Pi(P,Q)} \int c(\xi,\xi')\mathrm{d}\pi(\xi,\xi')$. Computation of the Wasserstein distance requires the solution of an infinite-dimensional linear program. In order to enhance its computational efficiency, Cuturi (2013) proposed to regularize the distance by penalizing the relative entropy, i.e., the Kullback-Leibler divergence, between the coupling distribution π and a reference measure $\mu\otimes\nu \in \mathcal{P}(\Xi\times\Xi)$,

$$W_c^\epsilon(P,Q) = \min_{\pi\in\Pi(P,Q)} \int c(\xi,\xi')\mathrm{d}\pi(\xi,\xi') + \epsilon H(\pi|\mu\otimes\nu),$$

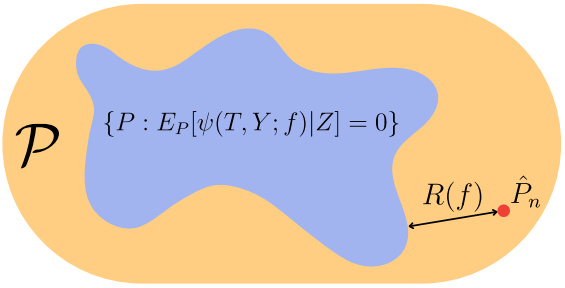


Figure 2. Sinkhorn profile. For every $f \in \mathcal{F}$, the Sinkhorn profile $R(f)$, (3), is the minimal distance between the empirical distribution $\hat{P}_n$ and the set of distributions satisfying the CMR (1).

where the relative entropy is defined as

$$H(\pi|\mu\otimes\nu) = \int_{\Xi\times\Xi} \log\left(\frac{\mathrm{d}\pi(\xi,\xi')}{\mathrm{d}\mu(\xi)\mathrm{d}\nu(\xi')}\right)\mathrm{d}\pi(\xi,\xi').$$

The resulting distance can be efficiently computed with the matrix scaling algorithm of Sinkhorn & Knopp (1967), from where it derives its name, Sinkhorn distance. We refer to Peyré et al. (2019) for a comprehensive introduction to computational optimal transport for machine learning.

In order to define an estimator for the conditional moment restriction problem (1), first, we resort to the functional formulation of Kremer et al. (2022). Let $\mathcal{H}$ denote a sufficiently rich space of functions such that equivalence between (1) and (2) holds. Then we define the moment functional $\Psi : \mathcal{T}\times\mathcal{Y}\times\mathcal{Z}\times\mathcal{F} \to \mathbb{R}^m$ via its action on $h \in \mathcal{H}$ as $\Psi(t,y,z;f)(h) = \psi(t,y;f)^T h(z)$. This lets us express the CMR (1) in its equivalent functional form, $\|E[\Psi(T,Y,Z;f)]\|_{\mathcal{H}^*} = 0$, where $\|\cdot\|_{\mathcal{H}^*}$ denotes the norm in the dual space $\mathcal{H}^*$ of $\mathcal{H}$.

With this at hand, we can define the primal problem of the Sinkhorn Method of Moments estimator for conditional moment restrictions as the minimizer of the *Sinkhorn profile* R_ϵ defined as

$$R_\epsilon(f) := \min_{P\in\mathcal{P}} W_c^\epsilon(P,\hat{P}_n) \quad \text{s.t.} \quad \|E_P[\Psi(T,Y,Z;f)]\|_{\mathcal{H}^*} = 0. \tag{3}$$

Using Lagrangian duality we can go over to the dual formulation of (3) as formalized by the following theorem whose proof is inspired by the mathematically closely related Sinkhorn Distributionally Robust Optimization (DRO) method of Wang et al. (2023).

Theorem 3.1 (Duality)**.** *Consider the Sinkhorn profile* (3) *with reference measure* $\mu\otimes\nu \in \mathcal{P}(\Xi\times\Xi)$*. Then* (3) *has the strongly dual form* $R_\epsilon(f) = \sup_{h\in\mathcal{H}} D(f,h)$*, where*

$$D(f,h) := E_{\xi'\sim\nu}\left[-\epsilon\log E_{\xi\sim\mu}\left[e^{-\Psi(\xi;f)(h)-c(\xi,\xi')/\epsilon}\right]\right]. \tag{4}$$

In contrast to its original purpose, in our application, the goal of the entropic regularization penalty is not to make computation of the distance more efficient but rather to arrive at a relaxed dual problem (4). The dual Sinkhorn profile (4) contains expectation operators with respect to the reference distributions μ and ν combined in a non-linear way. This casts optimization of the objective difficult as stochastic gradient estimates will be biased. One way to proceed is to resort to de-biasing techniques as discussed by Wang et al. (2023) for their related DRO objective. However, on top of the problem of gradient estimation, computation of (4) requires sampling from two reference distributions μ and ν such that accurate gradient estimation becomes costly.

To avoid these issues, we propose an alternative solution for a special choice of reference measures and cost function. Cuturi (2013) chooses the reference measure as the product of the marginals of the coupling distribution π. For $W_c^\epsilon(P, Q)$ this corresponds to the choice $\mu \otimes \nu = P \otimes Q$. The choice of μ and ν can be interpreted as a prior for distributions P and Q respectively. Motivated by this, we choose $\nu = \hat{P}_n$ and in order not to restrict the form of P we use an uninformative prior and choose μ as the Lebesgue measure.

The second modeling choice is the transport cost function c. Here, we use a weighted Euclidean norm,

$$\begin{aligned} c(\xi, \xi') &:= \frac{1}{2}(\xi - \xi')^T \Gamma (\xi - \xi') \\ &= \frac{1}{2} \sum_{w \in \{t,y,z\}} \gamma_w \|w - w'\|_2^2, \end{aligned} \tag{5}$$

where the factors $\gamma_w > 0$ determine the transport cost in the spaces $\mathcal{T}$, $\mathcal{Y}$ and $\mathcal{Z}$ and we defined the block diagonal matrix $\Gamma := \operatorname{diag}(\{\gamma_t I_{d_t}, \gamma_y I_{d_y}, \gamma_z I_{d_z}\}) \in \mathbb{R}^{d_\xi \times d_\xi}$, with I_{d_i} denoting the identity matrix in $\mathbb{R}^{d_i}$. With these choices, the objective (4) becomes

$$D(f, h) = E_{\xi' \sim \hat{P}_n} \left[-\epsilon \log E_{\xi \sim \mathcal{N}(\xi', \epsilon\Gamma^{-1})} \left[e^{-\Psi(\xi; f)(h)} \right] \right], \tag{6}$$

where $\mathcal{N}(\xi', \epsilon\Gamma^{-1})$ denotes a multivariate Gaussian centered at $\xi' = (t', y', z')$ with diagonal covariance $\epsilon\Gamma^{-1}$. Thus, for each value of ξ' we need to carry out an expectation over the moment violation $\exp(-\Psi(\xi; f)(h))$ with respect to a narrow Gaussian distribution centered at ξ'. Now, as ϵ is a small regularization parameter, the integrand will only provide relevant contributions in a neighborhood of ξ' and thus, for a sufficiently smooth moment function ψ and instrument function h, we can employ a Taylor expansion and carry out the Gaussian expectation over ξ in closed form. In the following, we define the weighted Laplacian $\Delta_\xi = \nabla_\xi \cdot \left(\Gamma^{-1} \nabla_\xi\right) = \sum_{w \in \{t,y,z\}} \frac{1}{\gamma_w} \Delta_w$ and the weighted l_2-norm $\|\cdot\|_\Gamma$ as $\|v\|_\Gamma^2 = v^T \Gamma^{-1} v$ for $v \in \mathbb{R}^{d_\xi}$.

Theorem 3.2. *Let the moment functional $\Psi(\cdot; f) : \Xi \to \mathcal{H}^*$ be continuously differentiable everywhere for any $f \in \mathcal{F}$. Consider the SMM estimator with transport cost function (5) and reference measure $\hat{P}_n \otimes L$, where L denotes the Lebesgue measure over Ξ. Then, for ϵ/γ_i, $i \in [t, y, z]$, sufficiently small, up to constants and rescalings the objective of the dual Sinkhorn profile (4) takes the form*

$$\begin{aligned} D(f, h) =& E_{\xi \sim \hat{P}_n} \left[\left(I + \frac{\epsilon}{2} \Delta_\xi \right) \Psi(\xi; f)(h) \right] \\ &- \frac{\epsilon}{2} E_{\xi \sim \hat{P}_n} \left[\|\nabla_\xi \Psi(\xi; f)(h)\|_\Gamma^2 \right] + O(\epsilon^{3/2}). \end{aligned} \tag{7}$$

Motivated by the classical 2-stage generalized method of moments (GMM) estimator (Hansen, 1982) we define the Sinkhorn Method of Moments by substituting the instrument function in the second term in (7) by a first stage estimate $\tilde{f}$. We will show below that this does not harm the consistency and convergence properties of our method. Additionally, we add regularization on the instrument function $-\frac{\lambda}{2}\|h\|_{\mathcal{H}}^2$ to ensure that the optimization over h is well behaved on finite samples.

Definition 3.3 (SMM)**.** Let $\tilde{f} \in \mathcal{F}$ denote a first-stage estimate of $f_0 \in \mathcal{F}$, then we define the Sinkhorn Method of Moments (SMM) estimator as the solution of the saddle-point problem

$$f^{\text{SMM}} = \arg\min_{f \in \mathcal{F}} \max_{h \in \mathcal{H}} M(f, h) - \epsilon \mathcal{R}(\tilde{f}, h) \tag{8}$$

with

$$\begin{aligned} M(f, h) &= E_{\hat{P}_n} \left[\left(I + \frac{\epsilon}{2} \Delta_\xi \right) \Psi(\xi; f)(h) \right] \\ \mathcal{R}(\tilde{f}, h) &= \frac{1}{2} E_{\hat{P}_n} \left[\|\nabla_\xi \Psi(\xi; \tilde{f})(h)\|_\Gamma^2 \right] + \frac{\lambda}{2\epsilon} \|h\|_{\mathcal{H}}^2, \end{aligned}$$

where as before $\Psi(\xi; f)(h) = \psi(t, y; f)^T h(z)$.

By using the 2-stage GMM-style estimator we shift most of the computational complexity into the optimization of the instrument function $h \in \mathcal{H}$. The optimization over the possibly high-dimensional model remains simple and even is a convex program, whenever f has a convexity preserving parameterization, e.g., for linear models. In practice, if (8) is optimized with stochastic gradient methods, one can dynamically update the first stage estimate $\tilde{f}$ using the result from the previous iteration. In the context of CMR estimation this GMM-inspired two stage procedure is a popular approach to stabilize the training (Lewis & Syrgkanis, 2018; Bennett et al., 2019; Bennett & Kallus, 2023). Note that without the 2-stage adaptation we would obtain an estimator similar in spirit to the continuous updating GMM estimator of Hansen et al. (1996) or the FGEL estimator of Kremer et al. (2022), which can be harder to train in practice (Hall, 2004).

The objective (8) involves a gradient and a Laplacian with respect to the data, which allows the method to take into account the geometry of the moment violation with respect to the data manifold. As we maximize the objective over $h \in \mathcal{H}$, we promote instrument functions which correspond to local minima of the moment violation $\psi(t, y; f)^T h(z)$ with respect to the data. Generally for CMR estimators the instrument function is responsible for translating the data into a learning signal for the model f. Choosing h in a local minimum w.r.t. the data means that we attribute less importance to data points that lead to large increases in the moment violation when perturbed slightly. This makes the model less vulnerable to poisoned data and adversarial attacks. SMM's property to take into account how the learning signal changes in proximity of the data is unique compared to related estimators which are blind towards the geometry of the data manifold as they are based on reweighting the existing data (Lewis & Syrgkanis, 2018; Bennett et al., 2019; Dikkala et al., 2020; Kremer et al., 2022; Bennett & Kallus, 2023) or sampling additional data points (Kremer et al., 2023) respectively.

3.1. Consistency

The following assumptions allow us to guarantee consistency and derive a convergence rate of our 2-stage estimator (8) in the parametric, uniquely identified setting. Suppose there exists a unique parameter $\theta_0 \in \Theta \subseteq \mathbb{R}^p$ for which $E[\psi(T, Y; \theta_0)|Z] = 0$ P_Z-a.s.. In the following, let $x \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ denote the concatenation of $(t, y) \in \mathcal{T} \times \mathcal{Y}$ and let $i \in [m]$ be a shorthand for $i \in \{1, \dots, m\}$. Further, we define the Jacobian of a vector-valued function $\psi : \mathcal{X} \times \Theta \to \mathbb{R}^m$ as $J_x\psi(x; \theta) \in \mathbb{R}^{m \times d_x}$.

Assumption 1 (Identifiability)**.** $\theta_0 \in \Theta$ is the unique solution to $E[\psi(X; \theta)|Z] = 0$ P_Z-a.s.; Θ is compact; $\psi(X; \theta)$ is continuous in θ everywhere w.p.1.

This is a standard assumption in IV regression that provides identifiability of the true parameter θ_0.

Assumption 2 (Data regularity)**.** The space $\Xi = \mathcal{T} \times \mathcal{Y} \times \mathcal{Z} \subset \mathbb{R}^{d_\xi}$ is compact.

Assumption 3 (Smoothness w.r.t. data)**.** The moment function $\psi(\cdot; \theta) : \mathcal{T} \times \mathcal{Y} \to \mathbb{R}^m$ is C^∞-smooth in the data for every $\theta \in \Theta$. Further the sets of functions $\{\psi(\cdot; \theta)_l : \theta \in \Theta\}$ and $\{(J_x\psi(\cdot; \theta))_{lr} : \theta \in \Theta\}$, are P_0-Donsker for every $l \in [m]$ and $r \in [d_x]$.

Assumption 2 and 3 ensure that the moment function and its derivatives are well-behaved with respect to the data. While the compactness of the data space might be violated in practice, usually one can construct a sufficiently large compact set that contains the data with high probability.

Assumption 4. The matrix $V(Z; \theta) \in \mathbb{R}^{m \times m}$ defined as

$$V(Z; \theta) = E[J_x\psi(X; \theta)\Gamma^{-1}J_x\psi(X; \theta)^T|Z] \quad (9)$$

is non-singular for $\theta \in \{\theta_0, \bar{\theta}\}$ w.p.1, where $\bar{\theta}$ is an initial parameter estimate defined in Assumption 6.

This corresponds to the common assumption of a non-singular covariance matrix required by related estimators (Newey & Smith, 2004; Kremer et al., 2022; Bennett & Kallus, 2023), but, here, imposed on the covariance of the data-Jacobian.

Assumption 5 (Instrument function)**.** $\mathcal{H} = \bigoplus_{l=1}^m \mathcal{H}_l$ is a a sufficiently rich space of vector-valued functions such that equivalence between (1) and (2) holds. Further for $l \in [m]$, $h \in \mathcal{H}_l$ is C^∞-smooth and the unit ball $\mathcal{H}_{l,1} := \{h \in \mathcal{H}_l : \|h\|_{\mathcal{H}_l} \leq 1\}$ as well as $\{J_z h : h \in \mathcal{H}_{l,1}\}$ are P_0-Donsker.

This is fulfilled, for example, by choosing each $\mathcal{H}_l$ as the RKHS of a universal, integrally stricly positive definite kernel, e.g., the Gaussian kernel, which we will formalize later. For neural network instrument function classes, equivalence between the variational and conditional formulations can be shown on basis of universal approximation theorems (Yarotsky, 2017; 2018). In this case $\mathcal{H}_{l,1}$, C^∞-smoothness can be realized by using smooth activation functions.

Assumption 6 (Regularization)**.** There is a first-stage parameter estimate $\bar{\theta}_n \xrightarrow{p} \bar{\theta}$ for which $E\left[\|\psi(X; \bar{\theta}_n) - \psi(X; \bar{\theta})\|_\infty\right] = O_p(n^{-\zeta})$ and $E\left[\|J_x\psi(X; \bar{\theta}_n) - J_x\psi(X; \bar{\theta})\|_\infty\right] = O_p(n^{-\zeta})$ with $0 < \zeta \leq 1/2$. Choose $\lambda_n = O_p(n^{-\rho})$ with $0 < \rho < \zeta$.

For linear IV regression this implies $\|\bar{\theta}_n - \bar{\theta}\|_\infty = O_p(n^{-\zeta})$, which means $\bar{\theta}_n$ has to be a $n^{-\zeta}$-consistent estimator for $\bar{\theta}$, which can be any parameter for which (9) is non-singular, e.g., the true parameter θ_0.

Assumption 7 (Smoothness w.r.t. θ)**.** $\theta_0 \in \text{int}(\Theta)$; $\psi(x; \theta)$ is continuously differentiable in a neighborhood $\bar{\Theta}$ of θ_0; and $E[\sup_{\theta \in \bar{\Theta}} \|J_\theta\psi(X; \theta)\|^2|Z] < \infty$ w.p.1; $\text{rank}\left(E[J_\theta\psi(X; \theta_0)|Z]\right) = p$, w.p.1.

This allows us to translate the convergence rate of the moment functional into a rate for the parameter estimate.

With these assumptions, consistency of SMM follows:

Theorem 3.4 (Consistency)**.** *Let Assumptions 1-6 be satisfied. For any $0 < \epsilon_1 < \epsilon_2$, choose $\epsilon \sim \text{Uniform}([\epsilon_1, \epsilon_2])$. Then the SMM estimator $\hat{\theta}$ converges to the true parameter θ_0 in probability $\hat{\theta} \xrightarrow{p} \theta_0$.*

If additionally Assumption 7 is satisfied, then $\|\hat{\theta} - \theta_0\| = O_p(n^{-1/2})$.

The consistency result is independent of the choice of instrument function space $\mathcal{H}$ as long as it fulfills Assumption 5. Next, we discuss two different implementations of $\mathcal{H}$ based on kernel methods and neural networks.

Algorithm 1 n-stage Kernel-SMM

Input: Initial function $\tilde{f}$, hyperparameters ϵ, λ, γ_x
for $i = 1, \dots, n$ **do**
  Compute $Q(\tilde{f})$
  while not converged **do**
    $f \leftarrow \text{GradientDescent}(f, \nabla_f R_{Q(\tilde{f})}(f))$
  end while
  $\tilde{f} \leftarrow f$
end for
Output: Function estimate f

3.2. Kernel-SMM

Choosing $\mathcal{H}$ as the RKHS of a suitable kernel, we can guarantee equivalence between the conditional and variational moment restrictions formulations (1) and (2). On top of that, for RKHS instrument functions we can employ a representer theorem and carry out the optimization over the instrument function $h \in \mathcal{H}$ in closed form. The resulting estimator can be obtained as the solution of a simple minimization problem bearing close resemblance to the optimally weighted 2-stage GMM estimator but taking into account the geometry of the moment violation with respect to the data. Before deriving the result we provide the necessary background on reproducing kernel Hilbert spaces (RKHS).

Reproducing Kernel Hilbert Space An RKHS $\mathcal{H}$ is a Hilbert space of functions $h : \mathcal{Z} \to \mathbb{R}$ in which point evaluation is a bounded functional. With every RKHS one can associate a positive semi-definite kernel $k(\cdot, \cdot) : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ with the reproducing property, i.e., for any $h \in \mathcal{H}$ we have $h(z) = \langle h, k(z, \cdot)\rangle_{\mathcal{H}}$. A kernel is called universal if its RKHS is dense in the set of all continuous real-valued functions (Micchelli et al., 2006). Further, a kernel is called integrally strictly positive definite (ISPD) if for any $h \in \mathcal{H}$ with $0 < \|h\|^2_{\mathcal{H}} < \infty$, we have $\int_{\mathcal{Z}} h(z)k(z, z')h(z')\mathrm{d}z\mathrm{d}z' > 0$. Refer to, e.g., Schölkopf & Smola (2002) and Berlinet & Thomas-Agnan (2011) for comprehensive introductions.

The following proposition specifies the properties of an RKHS for which Assumption 5 is satisfied.

Proposition 3.5. *Let $\mathcal{Z} \subset \mathbb{R}^{d_z}$ be compact. Then, the instrument function space $\mathcal{H} = \bigoplus_{l=1}^m \mathcal{H}_l$, where each $\mathcal{H}_l$ corresponds to the RKHS of universal, integrally strictly positive definite kernel k_l, $l \in [m]$ fulfills Assumption 5.*

Now, for a representer theorem to hold, in the following, we place infinite cost $\gamma_z = \infty$ on the transport of $z \in \mathcal{Z}$, i.e., we fix the instruments at their empirical locations. As long as $\gamma_t, \gamma_y < \infty$ this still allows for varying the functional relation between Z and T as well as T and Y in the training data. In the following, define the block-diagonal matrix $\Gamma_x := \text{diag}(\{\gamma_t I_{d_t}, \gamma_y I_{d_y}\}) \in \mathbb{R}^{d_x}$ and the weighted Laplace operator $\Delta_x = \nabla_x \cdot \left(\Gamma_x^{-1} \nabla_x\right)$.

Theorem 3.6 (Kernel-SMM). *Let $\mathcal{H} = \bigoplus_{l=1}^m \mathcal{H}_l$ be the direct sum of m reproducing kernel Hilbert spaces with kernels $k_l : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$. Let $\tilde{f} \in \mathcal{F}$ denote a first stage estimate of f_0 and let $\gamma_z = \infty$. Define $\psi_\Delta(f) \in \mathbb{R}^{nm}$, $L \in \mathbb{R}^{nm \times nm}$ and $Q(f) \in \mathbb{R}^{nm \times nm}$ with entries*

$$\begin{aligned}
\psi_\Delta(f)_{i\cdot l} =& \left(I + \frac{\epsilon}{2}\Delta_x\right) \psi_l(x_i; f) \\
L_{(i\cdot l),(j\cdot r)} =& \delta_{lr} k_l(z_i, z_j) \\
Q(f)_{(i\cdot l),(j\cdot r)} =& \frac{1}{n} \sum_{k=1}^{n} \sum_{s=1}^{d_x} \Big\{ k_l(z_i, z_k) \nabla_{x_s} \psi_l(x_k; f) \\
& \times \left(\Gamma_x^{-1}\right)_{ss} \nabla_{x_s} \psi_r(x_k; f) k_r(z_k, z_j) \Big\}.
\end{aligned}$$

Then the Sinkhorn profile is given by

$$R_{Q(\tilde{f})}(f) = \frac{1}{2n^2} \psi_\Delta(f)^T L \left(Q(\tilde{f}) + \frac{\lambda}{\epsilon} L\right)^{-1} L \psi_\Delta(f). \tag{10}$$

Compared to the general saddle point formulation (8) the kernelized version (10) has the significant advantage that it only involves a minimization over the model parameters and thus avoids the difficulties of mini-max optimization (Daskalakis et al., 2017). Algorithm 1 details the implementation of the multi-stage Kernel-SMM approach. In order to minimize the number of hyperparameters, we implement the gradient descent step with the limited memory BFGS method (Liu & Nocedal, 1989). We empirically observed that the n-step estimator effectively converges with the second iteration.

3.3. Neural-SMM

A particularly interesting alternative choice of instrument function space are neural network classes, as they can represent highly flexible functions while allowing for optimization via mini-batch stochastic gradient methods. As demonstrated by related works (Lewis & Syrgkanis, 2018; Bennett et al., 2019; Kremer et al., 2022), such neural network-based approaches can lead to powerful and scalable estimators that may outperform the corresponding kernel method on large samples. On the downside, they tend to be difficult to train due to the instability and hyperparameter sensitivity of mini-max optimization. This is particularly problematic for IV regression, as in contrast to standard supervised learning, it is non-trivial to define suitable validation metrics to set these hyperparameters. As a result, compared to (10), those estimators require more attention and careful evaluation which makes them less suitable as plug-and-play IV estimators for practitioners. As the primary focus of this work is to introduce a new geometry-aware learning paradigm for IV regression independent of the instrument function class, we consider the simpler kernel version in the following and defer results for the Neural-SMM estimator to Appendix B.

4. Experimental Results

We benchmark the kernel version of our method against a selection of plug-and-play IV estimators including maximum moment restrictions (MMR) (Zhang et al., 2023), sieve minimum distance (SMD) (Ai & Chen, 2003) as well as the kernel variational method of moments (VMM) (Bennett & Kallus, 2023). Results for the neural network version and related estimators can be found in Appendix B. For all kernel methods we choose a radial basis function kernel $k(z, z') = \exp(-\eta\|z - z'\|_2^2)$, where we set η according to the median heuristic (Garreau et al., 2017). The remaining hyperparameters of all methods are set by using the MMR objective on a validation data set (see Appendix A). In all experiments we consider perturbations in the treatment variable t and fix the other variables at their empirical values by setting $\gamma_y, \gamma_z = \infty$ for SMM. Implementations of our estimators are available at `https://github.com/HeinerKremer/sinkhorn-iv/`.

IV Regression with Corrupted Data We consider the SimpleIV experiment of Bennett & Kallus (2023) with the following data generating process,

$$\begin{aligned} Z &= \sin(\pi Z_0/10) \\ T &= -0.75Z_0 + 3.5H + 0.14\eta - 0.6 \\ Y &= f(T;\theta_0) - 10U + 0.1\eta_2 \end{aligned} \tag{11}$$

where $\eta_1, \eta_2, U \sim N(0, I)$ and $Z_0 \sim \text{Uniform}([-5, 5])$. The model is given by $f(t;\theta) = \theta^1 t^2 + \theta^2 t + \theta^3$ with $\theta_0 = [3.0, -0.5, 0.5]$. This is a typical IV problem, where the unobserved confounder U induces a non-causal dependence between T and Y. To investigate the robustness against corrupted data, we sample training sets of 1000 points and exchange a proportion of the covariates T by random values generated according to $\text{Uniform}([t_{\min}, t_{\max}])$. Figure 3 shows the mean-squared error of the models trained with different methods over the proportion of random covariates in the training data. We observe that for no data corruption, all estimators perform similarly, with SMM providing a small advantage. With increasing proportion of corrupted data, SMM scales favorably compared to the baselines. We provide more details and a hyperparameter sensitivity analysis in Appendix A.

Adversarially Robust IV Regression We test the adversarial robustness of different IV estimators in the following setting. Define $C = 0.2I \in \mathbb{R}^{5\times1}$, as well as $B \in \mathbb{R}^{5\times1}$, with fixed entries sampled from $\text{Uniform}([0.1, 0.3])$. Consider the non-linear data generating process,

$$\begin{aligned} Z &\sim \text{Uniform}([-3, 3]) \\ T &= BZ + CU + \eta_1 \\ Y &= f_0(T) + U + \eta_2 \end{aligned}$$

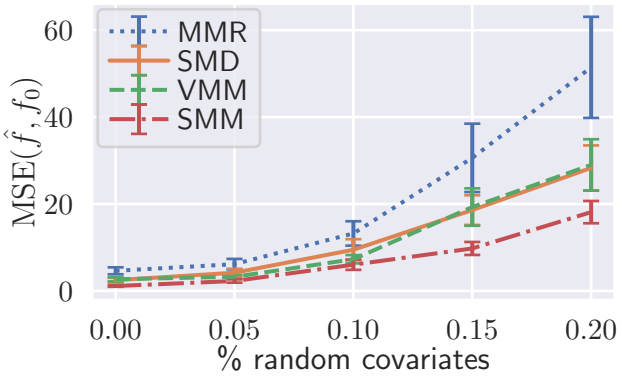


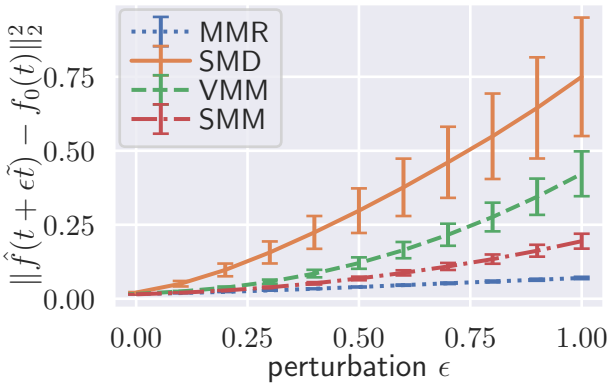
Figure 3. Robustness against corrupted data. We generate 1000 data points from the process (11) and substitute in a proportion of the data the treatment variable T for a random value sampled uniformly over the domain. Lines and error bars correspond to the mean and standard error computed over 20 training datasets.

with $U \sim N(0, 1)$, $\eta_1, \eta_2 \sim N(0, 0.1)$ and $f_0(t) = 1.5\cos(At) + 0.1At$, where $A \in \mathbb{R}^{1\times5}$ with fixed entries sampled from $\text{Uniform}([-1.5, 1.5])$. We approximate f_0 with a feed-forward neural network with $[20, 20, 3]$ hidden units and leaky ReLU activation functions. We train the network using different plug-and-play IV estimators and evaluate the adversarial robustness by running FGSM attacks (Goodfellow et al., 2014) in directions $\tilde{t}$ with strength $\epsilon \in [0, 1.0]$. Figure 4 shows that all IV estimators yield comparable mean-squared errors for $\epsilon = 0$, clearly improving over the non-causal least squares (LSQ) solution (table). Moreover, for increasing attack strengths ϵ, we see that SMM demonstrates stronger adversarial robustness than the SMD and VMM estimators. Interestingly, here, the MMR estimator which performed worse in the first experiments exhibits the least sensitivity towards adversarial perturbations. This might be understood by the fact that the MMR estimator corresponds to the limit case of SMM and VMM for $\lambda \to \infty$. Generally, strong regularization promotes flat functions which are less sensitive to the inputs, which could explain MMR's superior robustness here.

In Appendix B we provide results on a common modern IV benchmark that provides further evidence that SMM performs on par with state-of-the art estimators in standard IV settings. In this context, we also provide results for a Neural-SMM estimator, which proves to be competitive with state-of-the art deep learning approaches (Bennett et al., 2019; Kremer et al., 2022).

5. Related Work

Instrumental variable regression has traditionally been addressed via the 2-stage least squares (2SLS) method, which limits both regression stages to linear models (Angrist & Pis-



	LSQ	MMR	SMD	VMM	SMM
MSE ($\epsilon = 0$)	0.45	0.014	0.018	0.012	0.012

Figure 4. Adversarial robustness of IV estimators. We use a training set of size $n = 1000$ and evaluate the learned models over FGSM attacks with increasing strength ϵ. Lines and error bars show the mean and standard error over 20 random training datasets. The table contains the MSE in the perturbation-free case.

chke, 2008). Extensions to non-linear models have been provided by multiple works (Amemiya, 1974), recently based on density estimators (Hartford et al., 2017; Singh et al., 2019) and deep features (Xu et al., 2021). As an alternative to 2SLS, estimators based on the conditional moment restriction formulation have been used based on either basis function expansions of L^2 (Carrasco & Florens, 2000; Ai & Chen, 2003; Carrasco et al., 2007; Otsu, 2011) or machine learning models (Bennett et al., 2019; Dikkala et al., 2020; Muandet et al., 2020; Kremer et al., 2022; 2023; Bennett & Kallus, 2023). Related to our Kernel-SMM estimator, multiple works have used RKHS functions as instrument models (Carrasco & Florens, 2000; Singh et al., 2019; Bennett & Kallus, 2023; Zhang et al., 2023), leading to similar formulations as our (10). However, in contrast to ours, none of them take into account the geometry of the moment violation with respect to the data.

Optimization over measure spaces by means of minimizing some notion of distributional distance between the optimization variable and an empirical distribution has recently attracted significant attention in the context of distributionally robust optimization (Duchi & Namkoong, 2017; Sinha et al., 2018; Mohajerin Esfahani & Kuhn, 2018; Lam, 2019; Duchi & Namkoong, 2020; Duchi et al., 2021). On a higher level, one can distinguish between three types of approaches based on the respective distance notion (cf. Figure 1): φ-divergences restrict the optimization variable to a finite dimensional vector of weights attributed to the data points and thus find optimal reweightings of the sample. Methods based on maximum-mean discrepancy (Gretton et al., 2007) and the Fisher-Rao metric (Bauer et al., 2016), allow for creation and annihilation of probability mass (Zhu et al., 2021; Kremer et al., 2023; Yan et al., 2023). Finally, methods based on optimal transport distances effectively allow to move around the data points in the data space (Mohajerin Esfahani & Kuhn, 2018; Sinha et al., 2018). While CMR estimation has been based on the previous two paradigms, to the best of our knowledge, our Sinkhorn Method of Moments is the first estimator based on the latter category.

In a different context, empirical likelihood has previously been combined with Wasserstein distances to calibrate the radius of ambiguity sets in distributionally robust optimization (DRO) (Blanchet et al., 2019). However, their method does not extend to CMR estimation and neither does it make use of a regularized duality structure. From a mathematical perspective the derivation of our first duality result (Theorem 3.1) closely resembles the derivation of the dual Sinkhorn DRO estimator of Wang et al. (2023), which, nevertheless, addresses an entirely different problem. In addition, Wang et al. (2023) relies on de-biasing techniques to optimize their objective, whereas we provided a form that can be directly optimized via stochastic gradient methods.

6. Conclusion

Instrumental variable regression is an important concept in the field of causal inference, which motivates the development of estimators adapted to the intricacies of real-world datasets. Notwithstanding recent mini-max estimators based on neural network instrument function classes showing convincing performance on benchmarks (Bennett et al., 2019; Dikkala et al., 2020; Kremer et al., 2022; 2023), there remains a need for simple plug-and-play estimators that can be trained by practitioners without deep technical knowledge and with a manageable set of hyperparameters. We have extended the repertoire of such estimators by a method whose learning signal arises from an optimal transport geometry in the data space. We showed that our estimator exhibits favorable properties in presence of corrupted data or adversarial examples while maintaining performance competitive with state-of-the art approaches on standard benchmarks. The simplicity of our plug-and-play estimator partially results from its kernel-based implementation which limits the scalability to large sample sizes. To address this, we provide a neural network-based implementation in the appendix, whose detailed analysis is left for future work.

Acknowledgements

We thank Yassine Nemmour and Frederike Lübeck for helpful initial discussions on the project as well as Frederik Träuble for insisting on using certain matrix identities.

Impact Statement

This paper presents work whose goal is to conceptually advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here. Causal analysis of systems has the potential to advance our understanding of a system's response under interventions, which may lead to more rational decision making.

References

Ai, C. and Chen, X. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

Amemiya, T. The nonlinear two-stage least-squares estimator. *Journal of Econometrics*, 2(2):105–110, 1974.

Angrist, J. D. and Pischke, J.-S. *Mostly harmless econometrics*. Princeton university press, 2008.

Bauer, M., Bruveris, M., and Michor, P. W. Uniqueness of the fisher–rao metric on the space of smooth densities. *Bulletin of the London Mathematical Society*, 48(3):499–506, 2016.

Bennett, A. and Kallus, N. The variational method of moments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):810–841, 2023.

Bennett, A., Kallus, N., and Schnabel, T. Deep generalized method of moments for instrumental variable analysis. *Advances in neural information processing systems*, 32, 2019.

Berlinet, A. and Thomas-Agnan, C. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Bierens, H. J. Consistent model specification tests. *Journal of Econometrics*, 20(1):105–134, 1982.

Blanchet, J., Kang, Y., and Murthy, K. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

Carrasco, M. and Florens, J.-P. Generalization of gmm to a continuum of moment conditions. *Econometric Theory*, 16(6):797–834, 2000. ISSN 02664666, 14694360.

Carrasco, M., Chernov, M., Florens, J.-P., and Ghysels, E. Efficient estimation of general dynamic models with a continuum of moment conditions. *Journal of econometrics*, 140(2):529–573, 2007.

Chen, X., Liu, C., Li, B., Lu, K., and Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Dikkala, N., Lewis, G., Mackey, L., and Syrgkanis, V. Minimax estimation of conditional moment models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12248–12262. Curran Associates, Inc., 2020.

Duchi, J. and Namkoong, H. Variance-based regularization with convex objectives. *Advances in neural information processing systems*, 30, 2017.

Duchi, J. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization, 2020.

Duchi, J. C., Glynn, P. W., and Namkoong, H. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3): 946–969, 2021.

Garreau, D., Jitkrittum, W., and Kanagawa, M. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

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

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Hall, A. *Generalized method of moments*. Wiley Online Library, 2004.

Hall, P., Wolff, R. C., and Yao, Q. Methods for estimating a conditional distribution function. *Journal of the American Statistical association*, 94(445):154–163, 1999.

Hansen, L. P. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4): 1029–1054, 1982. ISSN 00129682, 14680262.

Hansen, L. P., Heaton, J., and Yaron, A. Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, 14(3):262–280, 1996. ISSN 07350015.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pp. 1414–1423. PMLR, 2017.

Imbens, G. W., Spady, R. H., and Johnson, P. Information theoretic approaches to inference in moment condition models. *Econometrica*, 66(2):333–357, 1998. ISSN 00129682, 14680262.

Kitamura, Y. and Stutzer, M. An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65(4):861–874, 1997. ISSN 00129682, 14680262.

Kosorok, M. R. *Introduction to empirical processes and semiparametric inference*, volume 61. Springer, 2008.

Kremer, H., Zhu, J.-J., Muandet, K., and Schölkopf, B. Functional generalized empirical likelihood estimation for conditional moment restrictions. In *International Conference on Machine Learning*, pp. 11665–11682. PMLR, 2022.

Kremer, H., Nemmour, Y., Schölkopf, B., and Zhu, J.-J. Estimation beyond data reweighting: Kernel method of moments. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17745–17783. PMLR, 23–29 Jul 2023.

Lam, H. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105, 2019.

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

Liu, D. C. and Nocedal, J. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

Micchelli, C., Xu, Y., and Zhang, H. Universal kernels. *Mathematics*, 7, 12 2006.

Mohajerin Esfahani, P. and Kuhn, D. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

Muandet, K., Mehrjou, A., Lee, S. K., and Raj, A. Dual instrumental variable regression, 2020.

Newey, W. K. and Smith, R. J. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004. ISSN 00129682, 14680262.

Otsu, T. Empirical likelihood estimation of conditional moment restriction models with unknown functions. *Econometric Theory*, 27(1):8–46, 2011.

Owen, A. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120, 1990. ISSN 00905364.

Owen, A. B. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988. ISSN 00063444.

Owen, A. B. *Empirical likelihood*. Chapman and Hall/CRC, 2001.

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2000.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Qin, J. and Lawless, J. Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22(1): 300–325, 1994. ISSN 00905364.

Saengkyongam, S., Henckel, L., Pfister, N., and Peters, J. Exploiting independent instruments: Identification and distribution generalization. In *International Conference on Machine Learning*, pp. 18935–18958. PMLR, 2022.

Schölkopf, B. and Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Schölkopf, B., Herbrich, R., and Smola, A. J. A generalized representer theorem. In *Computational Learning Theory*, pp. 416–426, 2001.

Singh, R., Sahani, M., and Gretton, A. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Sinha, A., Namkoong, H., and Duchi, J. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Sinkhorn, R. and Knopp, P. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Wang, J., Gao, R., and Xie, Y. Sinkhorn distributionally robust optimization, 2023.

Xu, L., Chen, Y., Srinivasan, S., de Freitas, N., Doucet, A., and Gretton, A. Learning deep features in instrumental variable regression. In *International Conference on Learning Representations*, 2021.

Yan, Y., Wang, K., and Rigollet, P. Learning gaussian mixtures using the wasserstein-fisher-rao gradient flow. *arXiv preprint arXiv:2301.01766*, 2023.

Yarotsky, D. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

Yarotsky, D. Optimal approximation of continuous functions by very deep relu networks. In *Conference on learning theory*, pp. 639–649. PMLR, 2018.

Zhang, R., Imaizumi, M., Schölkopf, B., and Muandet, K. Instrumental variable regression via kernel maximum moment loss. *Journal of Causal Inference*, 11(1):20220073, 2023.

Zhu, J.-J., Jitkrittum, W., Diehl, M., and Schölkopf, B. Kernel distributionally robust optimization: Generalized duality theorem and stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 280–288. PMLR, 2021.

A. Experimental Details

Hyperparameters For SMM we choose the hyperparameters from the grid defined by $\epsilon \in [10^{-6}, 10^{-4}, 10^{-2}]$ and $\lambda/\epsilon \in [10^{-6}, 10^{-4}, 10^{-2}, 1.0]$. Note that as ϵ and γ_t only appear as ϵ/γ_t, we absorb the factor γ_t into ϵ and consider $\gamma_t = 1$ everywhere. For VMM we choose the hyperparameters from $\lambda \in [10^{-6}, 10^{-4}, 10^{-2}, 1.0]$ as done by the authors of the method (Bennett & Kallus, 2023). We pick the best hyperparameter configuration by evaluating the MMR objective (Zhang et al., 2023) on a validation data set of the same size as the training set. We visualize the dependency on the hyperparameters for the first experiment without random covariates in Figure 5. We observe that the method is rather insensitive to the choice of ϵ but admits a stronger dependence on the choice of the regularization parameter λ.

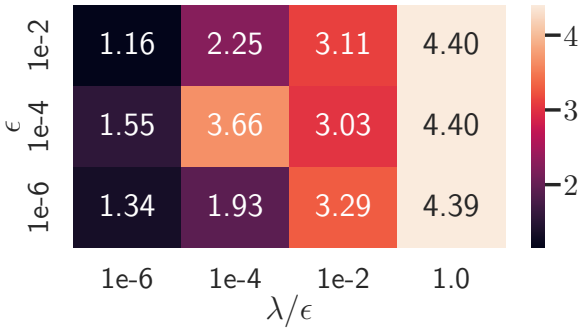


Figure 5. Kernel-SMM dependency on hyperparameters. We evaluate the SMM estimator on the first experiment without random covariates for different hyperparameter configurations. Values correspond to the mean of the prediction error $E[\|f(T;\hat{\theta}) - f(T;\theta_0)\|_2^2]$ averaged over models trained on 20 random training sets.

B. Additional Results

NetworkIV Here, we consider a common modern benchmark for IV regression in the standard setting without any data corruptions. Consider the following data generating process introduced by Bennett et al. (2019) and subsequently used by many other works (Zhang et al., 2023; Kremer et al., 2022; 2023),

$$\begin{aligned} &y = f_0(t) + e + \delta, && t = z + e + \gamma, \\ &z \sim \text{Uniform}([-3, 3]), \\ &e \sim N(0, 1), && \gamma, \delta \sim N(0, 0.1), \end{aligned}$$

where the function f_0 is chosen from the set of simple functions

$$\begin{aligned} &\text{sin: } f_0(t) = \sin(t), \quad && \text{abs: } f_0(t) = |t|, \\ &\text{linear: } f_0(t) = t, \quad && \text{step: } f_0(t) = I_{\{t \geq 0\}}. \end{aligned}$$

We learn a neural network f_θ with two layers of $[20, 3]$ hidden units and leaky ReLU activation functions to approximate the function f_0 by imposing the conditional moment restriction $E[Y - f_\theta(T)|Z] = 0$ P_Z-a.s.. Table 1 contains the results of different plug-and-play IV estimators trained on a dataset of 1000 points and averaged over 20 random training datasets. We observe that SMD, VMM and SMM perform roughly on par whereas MMR only improves in one of the settings over the non-causal least squares solution (LSQ) which ignores the instruments entirely.

Neural Estimators We explore an alternative SMM implementation where we represent the instrument function $h \in \mathcal{H}$ as a neural network parameterized by $\omega \in \Omega$. With this choice, the estimator (8) takes the form

$$f^* = \arg\min_{f \in \mathcal{F}} \max_{\omega \in \Omega} E_{\hat{P}_n} \left[\left(I + \frac{\epsilon}{2} \Delta_\xi \right) \left(\psi(\cdot; f)^T h_\omega(\cdot) \right)(\xi) - \frac{\epsilon}{2} \|\nabla_\xi \left(\psi(\cdot; f)^T h_\omega(\cdot) \right)(\xi)\|_\Gamma^2 - \frac{\lambda}{2} \|h_\omega(Z)\|_2^2 \right]. \tag{12}$$

The Neural-SMM estimator can be trained in the same fashion as the DeepGMM (Bennett et al., 2019) or Functional-GEL (Kremer et al., 2022) estimators by alternating mini-batch stochastic gradient descent steps in the the model parameters and the adversary parameters ω.

Table 1. NetworkIV experiment. Results represent the mean and standard error of the prediction error $E[\|f(T;\hat{\theta}) - f(T;\theta_0)\|_2^2]$ resulting from 20 random training datasets.

	LSQ	MMR	SMD	VMM	SMM
sin	0.36 ± 0.03	0.40 ± 0.02	0.12 ± 0.01	0.17 ± 0.02	0.15 ± 0.01
abs	1.94 ± 1.48	0.61 ± 0.28	0.20 ± 0.08	0.09 ± 0.04	0.12 ± 0.04
step	0.35 ± 0.04	> 100	0.04 ± 0.01	0.05 ± 0.01	0.04 ± 0.00
linear	0.36 ± 0.05	0.36 ± 0.09	0.07 ± 0.04	0.03 ± 0.01	0.07 ± 0.03

Table 2. Neural CMR estimators. Results represent the mean and standard error of the prediction error $E[\|f(T;\hat{\theta}) - f(T;\theta_0)\|_2^2]$ resulting from 20 random runs of the NetworkIV experiment.

	DeepGMM	NeuralFGEL	NeuralSMM
sin	0.08 ± 0.01	0.10 ± 0.01	0.07 ± 0.01
abs	0.04 ± 0.01	0.04 ± 0.01	0.04 ± 0.01
step	0.07 ± 0.01	0.08 ± 0.01	0.07 ± 0.01
linear	0.05 ± 0.01	0.06 ± 0.01	0.05 ± 0.01

We benchmark the Neural-SMM estimator against DeepGMM (Bennett et al., 2019) and FunctionalGEL (Kremer et al., 2022) which achieved state-of-the-art results on several benchmarks including the NetworkIV experiment. For all methods we use the same instrument network architecture consisting of a feed-forward neural network with $[50, 20]$ hidden units and leaky ReLU activation functions. We optimize the objective by alternating steps with an optimistic Adam (Daskalakis et al., 2017) optimizer with parameters $\beta = (0.5, 0.9)$. We tuned the learning rates, for the model and adversary by evaluating the DeepGMM estimator for different values and fix them both to $5e^{-4}$ for all methods. In the same way we fix the batch size to 200 and the number of epochs to 3000. For the FunctionalGEL estimator we use the Kullback-Leibler divergence version. For all methods we choose the regularization parameter λ from $[10^{-6}, 10^{-4}, 10^{-2}, 1.0]$ and for Neural-SMM we additionally choose ϵ from $[10^{-6}, 10^{-4}, 10^{-2}, 1.0]$ by using the MMR objective on a validation set of the same size as the training set.

We observe in Table 2 that Neural-SMM performs on par with these SOTA estimators on all variants of the NetworkIV experiment, suggesting that the geometry-awareness and additional robustness of our estimator does not come at the price of reduced performance in standard settings. It does, however, come at the price of increased computation due to the presence of the gradient and Laplace operators with respect to the data in the objective.

Figure 6 visualizes the dependence of Neural-SMM on its hyperparameters. We observe that for this experiment SMM requires either one or both parameters to be chosen large for optimal performance but the performance remains stable across a range of parameters.

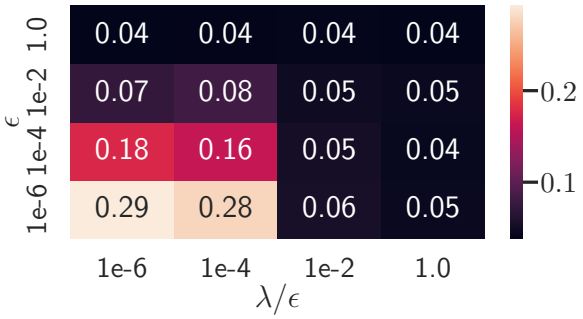


Figure 6. Neural-SMM dependency on hyperparameters. We evaluate the Neural-SMM estimator for different hyperparameter configurations exemplarily for the abs function in the network IV experiment. Values correspond to the mean of the prediction error $E[\|f(T;\hat{\theta}) - f(T;\theta_0)\|_2^2]$ averaged over models trained on 20 random training sets.

C. Proofs

C.1. Duality Results

Proof of Theorem 3.1

Proof. Introducing the Lagrange parameter $\rho \in \mathbb{R}$, the Lagrangian of (3) reads

$$L(P, \rho, f) = \min_{\pi \in \Pi(P, \hat{P}_n)} E_{(\xi,\xi')\sim\pi}\left[c(\xi, \xi') + \epsilon \log\left(\frac{\mathrm{d}\pi(\xi, \xi')}{\mathrm{d}\mu(\xi)\mathrm{d}\nu(\xi')}\right)\right] \tag{13}$$

$$+ \rho \sup_{h\in\mathcal{H}, \|h\|_{\mathcal{H}}=1} E_P[\Psi(\xi; f)(h)]. \tag{14}$$

As eventually the Lagrangian will be maximized with respect to ρ, we can merge it with the optimization over the unit ball in $\mathcal{H}$ to obtain a Lagrangian with an unrestricted parameter $h \in \mathcal{H}$,

$$L(P, h, f) = \min_{\pi \in \Pi(P, \hat{P}_n)} E_{(\xi,\xi')\sim\pi}\left[c(\xi, \xi') + \epsilon \log\left(\frac{\mathrm{d}\pi(\xi, \xi')}{\mathrm{d}\mu(\xi)\mathrm{d}\nu(\xi')}\right)\right] + E_P[\Psi(\xi; f)(h)]. \tag{15}$$

Note that the Wasserstein distance is mass preserving, i.e., we do not need to explicitly impose the constraint $E_P[1] = 1$ as this is implied directly by normalization of the empirical distribution, i.e., let p and $\hat{p}$ denote the density and probability mass functions of P and $\hat{P}_n$ respectively, then $E_P[1] = \int_\Xi p(\xi)\mathrm{d}\xi = \int_\Xi \sum_{i=1}^n \pi(\xi, \xi_i')\mathrm{d}\xi = \sum_{i=1}^n \hat{p}(\xi_i') = \sum_{i=1}^n \frac{1}{n} = 1$.

To derive the dual problem we need to minimize the Lagrangian over the primal variable P. By definition of the coupling distribution π we have $p = \mathbb{P}_{1\sharp}\pi$ and thus we can collapse the minimizations over the π and P into a single minimization over $\pi \in \Pi(\hat{P}_n) := \{\mathcal{P}(\Xi \times \Xi) : \mathbb{P}_{2\sharp}\pi = \hat{P}_n\}$,

$$D(h, f) = \min_{\pi \in \Pi(\hat{P}_n)} E_{(\xi,\xi')\sim\pi}\left[c(\xi, \xi') + \epsilon \log\left(\frac{\mathrm{d}\pi(\xi, \xi')}{\mathrm{d}\mu(\xi)\mathrm{d}\nu(\xi')}\right)\right] + E_{\mathbb{P}_{1\sharp}\pi}[\Psi(\xi; f)(h)]. \tag{16}$$

Now to extract the relevant degree of freedom we can write all expectation operators as combinations of the empirical expectation and conditional expectation over $\pi(\xi, \xi')$ given its second argument $\xi' \in \Xi$. To see this, note that by the product rule we have $\pi(\xi, \xi') =: \pi(\xi|\xi')\hat{p}(\xi')$ and by the law of iterated expectation we have for any function $g : \Xi \times \Xi \to \mathbb{R}$, $E_\pi[g(\xi, \xi')] = E_{\xi'\sim\hat{P}_n}[E_{\xi\sim\pi|\xi'}[g(\xi, \xi')|\xi']]$, where we defined $\pi|\xi'$ as the conditional distribution of ξ given ξ', with density $\pi(\xi|\xi')$. Similarly we have for any function $g : \Xi \to \mathbb{R}$,

$$E_{P_{1\sharp}\pi}[g(\xi)] = \int_\Xi g(\xi)(\mathbb{P}_{1\sharp}\pi)(\xi)\mathrm{d}\xi = \int_\Xi g(\xi)\sum_{i=1}^n \pi(\xi, \xi_i')\mathrm{d}\xi \tag{17}$$

$$= \int_\Xi g(\xi)\sum_{i=1}^n \pi(\xi|\xi_i')\hat{p}(\xi_i')\mathrm{d}\xi = \int_\Xi g(\xi)\frac{1}{n}\sum_{i=1}^n \pi(\xi|\xi_i')\mathrm{d}\xi \tag{18}$$

$$= E_{\xi'\sim\hat{P}_n}[E_{\xi\sim\pi|\xi'}[g(\xi)|\xi']]. \tag{19}$$

Therefore the optimization over $\pi \in \Pi(\hat{P}_n)$ is equivalent to a sequence of optimization problems over $\pi|\xi' \in \mathcal{P}(\Xi)$, one for each value of $\xi' \in \Xi$. With this we can express the dual problem (16) as

$$D(h, f) = E_{\xi'\sim\hat{P}_n}\left[\min_{\pi|\xi'\in\mathcal{P}(\Xi)} E_{\xi\sim\pi|\xi'}\left[c(\xi, \xi') + \epsilon \log\left(\frac{\mathrm{d}(\pi|\xi')(\xi)}{\mathrm{d}\mu(\xi)}\right) + \Psi(\xi; f)(h)\middle|\xi'\right]\right] \tag{20}$$

Now for each $\xi' \in \Xi$ consider the inner optimization problem

$$G(\xi'; h, f) := \min_{\pi|\xi'\in\mathcal{P}(\Xi)} E_{\xi\sim\pi|\xi'}\left[c(\xi, \xi') + \epsilon \log\left(\frac{\mathrm{d}(\pi|\xi')(\xi)}{\mathrm{d}\mu(\xi)}\right) + \Psi(\xi; f)(h)\right]. \tag{21}$$

Define the density of $\pi|\xi' \in \mathcal{P}(\Xi)$ with respect to the reference measure $\mu \in \mathcal{P}(\Xi)$ as $r(\xi) = \frac{\mathrm{d}(\pi|\xi')(\xi)}{\mathrm{d}\mu(\xi)}$, then we can rewrite the optimization problem as an optimization over $r \in \mathcal{R} := \{r : \Xi \to \mathbb{R}_+ : E_\mu[r(\xi)] = 1\}$,

$$G(\xi'; h, f) = \min_{r\in\mathcal{R}} E_{\xi\sim\mu}\left[r(\xi)c(\xi, \xi') + \epsilon r(\xi)\log\left(r(\xi)\right) + r(\xi)\Psi(\xi; f)(h)\right]. \tag{22}$$

Now introducing Lagrange parameter $\eta \in \mathbb{R}$ and using Lagrangian duality we get

$$G(\xi'; h, f) = \sup_{\eta \in \mathbb{R}} \min_{r:\Xi \to \mathbb{R}_+} E_{\xi \sim \mu}[r(\xi)c(\xi, \xi') + \epsilon r(\xi) \log(r(\xi)) + r(\xi)\Psi(\xi; f)(h) + \eta(1 - r(\xi))] \tag{23}$$

$$= \sup_{\eta \in \mathbb{R}} \eta - \epsilon E_{\xi \sim \mu}\left[\sup_{t \geq 0} t \frac{\eta - c(\xi, \xi') - \Psi(\xi; f)(h)}{\epsilon} - t \log t\right] \tag{24}$$

$$= \sup_{\eta \in \mathbb{R}} \eta - \epsilon E_{\xi \sim \mu}\left[\exp\left(\frac{\eta - c(\xi, \xi') - \Psi(\xi; f)(h)}{\epsilon} - 1\right)\right], \tag{25}$$

where we used that the Fenchel conjugate of the Kullback Leibler divergence $t \log t$ is $\sup_t \langle p, t\rangle - t \log t = e^{p-1}$. We can eliminate the dual normalization variable $\eta \in \mathbb{R}$ from the problem by solving the corresponding first order optimality condition

$$0 = 1 - e^{\eta/\epsilon - 1} E_{X \sim \mu}\left[\exp\left(\frac{-\Psi(\xi; f)(h) - c(\xi, \xi')}{\epsilon}\right)\right], \tag{26}$$

which yields

$$\eta = \epsilon - \epsilon \log E_{X \sim \mu}\left[\exp\left(\frac{-\Psi(\xi; f)(h) - c(\xi, \xi')}{\epsilon}\right)\right]. \tag{27}$$

Inserting back into (25), we obtain for each $\xi' \in \Xi$

$$G(\xi'; h, f) = -\epsilon \log E_{\xi \sim \mu}\left[\exp\left(\frac{-\Psi(\xi; f)(h) - c(\xi, \xi')}{\epsilon}\right)\right]. \tag{28}$$

and the result follows by inserting into (20) and redefining $h/\epsilon \to h$. □

Proof of Theorem 3.2

Proof. Using the assumptions on the reference measure and cost function, we can write the objective in the form (6), where the inner expectation is given as

$$E_{\xi \sim \mathcal{N}(\xi', \epsilon\Gamma^{-1})}\left[e^{-\Psi(\xi; f)(h)}\right] = \int_\Xi e^{-\Psi(\xi; f)} e^{-\frac{1}{2\epsilon}\|\xi - \xi'\|^2_{\Gamma^{-1}}} \mathrm{d}\xi. \tag{29}$$

As for small ϵ the integrand only provides a finite contribution in a neighborhood of ξ', we can use that Ψ is continuously differentiable everywhere and employ a Taylor expansion,

$$\Psi(\xi; f)(h) = \Psi(\xi'; f)(h) + (\xi - \xi')^T \nabla_\xi \Psi(\xi'; f)(h) \tag{30}$$

$$+ \frac{1}{2}(\xi - \xi')^T \nabla^2_\xi \Psi(\xi'; f)(h)(\xi - \xi') + O(\|\xi - \xi'\|^3). \tag{31}$$

Note that due to the Gaussian measure under the integral we have $\|\xi - \xi'\| = O(\epsilon^{1/2})$. Now defining $\delta := \xi - \xi' \in \Xi$ as well as the gradient $G(\xi') := \nabla_\xi \Psi(\xi'; f)(h)$ and Hessian $H(\xi') := \nabla^2_\xi \Psi(\xi'; f)(h)$ of the evaluated moment functional we can insert back and get

$$E_{\xi \sim \mathcal{N}(\xi', \gamma)}\left[e^{-\Psi(\xi; f)(h)}\right] = e^{-\Psi(\xi'; f)(h)} \int_\Xi \exp\left(-\frac{1}{2\epsilon}\left(2\epsilon\delta^T G(\xi') + \epsilon\delta^T H(\xi')\delta + \delta^T \Gamma \delta\right)\right) \mathrm{d}\delta + O(\epsilon^{3/2}). \tag{32}$$

Define the regularized Hessian $\Omega_\epsilon := \Omega_\epsilon(\xi') := \Gamma + \epsilon H(\xi')$, which is invertible w.p.1, as for sufficiently small ϵ/γ we have $\lambda_{\min}(\Gamma) = \min_{w \in \{t,y,z\}} \gamma_w > \epsilon \lambda_{\min}(H(\xi'))$ w.p.1 and thus Ω_ϵ is strictly positive definite w.p.1. Then we can employ a change of variables by defining $\omega := \Omega_\epsilon^{1/2}\delta$ and obtain

$$E_{\xi \sim \mathcal{N}(\xi', \gamma)}\left[e^{-\Psi(\xi; f)(h)}\right] \tag{33}$$

$$= e^{-\Psi(\xi'; f)(h)} \int \frac{1}{\left|\det \Omega_\epsilon^{1/2}\right|} \times \exp\left(-\frac{1}{2\epsilon}\left(\omega^T\omega + 2\epsilon\omega^T \Omega_\epsilon^{-1/2} G(\xi')\right)\right) \mathrm{d}\omega + O(\epsilon^{3/2}). \tag{34}$$

Now, completing the square we obtain

$$E_{\xi\sim\mathcal{N}(\xi',\gamma)}\left[e^{-\Psi(\xi;f)(h)}\right] \tag{35}$$

$$=e^{-\Psi(\xi';f)(h)}e^{\frac{\epsilon}{2}G(\xi')^T\Omega_\epsilon^{-1}G(\xi')}\int\frac{1}{\left|\det\Omega_\epsilon^{1/2}\right|}\exp\left(-\frac{1}{2\epsilon}\left(\omega+\epsilon\Omega_\epsilon^{-1/2}G(\xi')^2\right)\right)\mathrm{d}\omega+O(\epsilon^{3/2}) \tag{36}$$

$$=\left(\frac{2\pi}{\epsilon}\right)^{d_\xi/2}\left|\det\Omega_\epsilon^{1/2}\right|^{-1}e^{-\Psi(\xi';f)(h)}e^{\frac{\epsilon}{2}G(\xi')^T\Omega_\epsilon^{-1}G(\xi')}+O(\epsilon^{3/2}). \tag{37}$$

Finally inserting back into (6) we get

$$D(f,h)=E_{\xi'\sim\hat{P}_n}\left[-\epsilon\log E_{\xi\sim\mathcal{N}(\xi',\gamma)}\left[e^{\Psi(\xi;f)(h)}\right]\right] \tag{38}$$

$$=E_{\xi'\sim\hat{P}_n}\left[\epsilon\Psi(\xi';f)(h)-\frac{\epsilon^2}{2}G(\xi')^T\Omega_\epsilon^{-1}G(\xi')+\frac{\epsilon}{2}\log\left|\det\Omega_\epsilon\right|\right]-\frac{\epsilon d_\xi}{2}\log\frac{2\pi}{\epsilon}+O(\epsilon^{5/2}). \tag{39}$$

Dividing by ϵ and neglecting constant terms we get

$$D(f,h)=E_{\xi'\sim\hat{P}_n}\left[\Psi(\xi';f)(h)-\frac{\epsilon}{2}G(\xi')^T\Omega_\epsilon^{-1}G(\xi')+\frac{1}{2}\log\left|\det\Omega_\epsilon\right|\right]+O(\epsilon^{3/2}). \tag{40}$$

Now, for small ϵ we can Taylor expand Ω_ϵ^{-1} as

$$\Omega_\epsilon^{-1}=\left(\Gamma+\epsilon H(\xi')\right)^{-1} \tag{41}$$

$$=\Gamma^{-1}\left(I+\epsilon\Gamma^{-1}H\right)^{-1} \tag{42}$$

$$=\Gamma^{-1}\left(I-\epsilon\Gamma^{-1}H\right)+O(\epsilon^2) \tag{43}$$

$$=\Gamma^{-1}+O(\epsilon). \tag{44}$$

Similarly we have

$$\log\left|\det\Omega_\epsilon\right|=\log\left|\det\left(\Gamma+\epsilon H\right)\right| \tag{45}$$

$$=\log\left|\det\Gamma\right|+\log\left|\det\left(I+\epsilon\Gamma^{-1}H\right)\right| \tag{46}$$

$$=\underbrace{\left(\sum_{x\in\{t,y,z\}}d_x\log\gamma_x\right)}_{=:C}+\log\det\left(I+\epsilon\Gamma^{-1}H\right) \tag{47}$$

$$=C+\mathrm{Tr}\log\left(I+\epsilon\Gamma^{-1}H\right) \tag{48}$$

$$=C+\mathrm{Tr}\left(\epsilon\Gamma^{-1}H+O(\epsilon^2)\right) \tag{49}$$

$$=C+\epsilon\sum_{x\in\{t,y,z\}}\frac{1}{\gamma_x}\Delta_x\Psi(\xi';f)(h)+O(\epsilon^2). \tag{50}$$

So we finally obtain

$$D(f,h)=E_{\hat{P}_n}\left[\Psi(\xi;f)(h)-\frac{\epsilon}{2}\sum_{x\in\{t,y,z\}}\frac{1}{\gamma_x}\left(\|\nabla_x\Psi(\xi;f)(h)\|_2^2-\Delta_x\Psi(\xi;f)(h)\right)\right]+O(\epsilon^{3/2}). \tag{51}$$

□

C.2. Proof of Theorem 3.4 (Consistency)

The objective of the SMM estimator (8) can be written as

$$\widehat{D}(h,\theta)=\left(I+\frac{\epsilon}{2}\Delta_\xi\right)E_{\hat{P}_n}[\Psi(\xi;f)(h)]-\frac{\epsilon}{2}\langle h,\widehat{\Omega}_{\lambda_n}(\bar{\theta}_n)h\rangle_{\mathcal{H}}, \tag{52}$$

where we defined the linear operator $\widehat{\Omega}_{\lambda_n}(\bar{\theta}_n):\mathcal{H}\to\mathcal{H}$ as $\widehat{\Omega}_{\lambda_n}(\bar{\theta}_n)=E_{\hat{P}_n}\left[\left(\nabla_\xi\Psi(\xi;\bar{\theta}_n)\right)^T\Gamma^{-1}\nabla_\xi\Psi(\xi;\bar{\theta}_n)\right]+\lambda_n I\otimes I$. Our proof of Theorem 3.4 uses properties of the spectrum of $\widehat{\Omega}_{\lambda_n}(\bar{\theta}_n)$ which we will derive in the following.

C.2.1. Previous Results

Lemma C.1 (Corollary 9.31, Kosorok (2008)). *Let $\mathcal{F}$ and $\mathcal{G}$ be Donsker classes of functions. Then $\mathcal{F}+\mathcal{G}$ is Donsker. Further if additionally $\mathcal{F}$ and $\mathcal{G}$ are uniformly bounded, then $\mathcal{F}\cdot\mathcal{G}$ is Donsker.*

Lemma C.2 (Lemma 18, Bennett & Kallus (2023)). *Suppose that $\mathcal{G}$ is a class of functions of the form $g:\Xi\to\mathbb{R}$, and that $\mathcal{G}$ is P-Donsker in the sense of Kosorok (2008). Then we have*

$$\sup_{g\in\mathcal{G}} E_{\hat{P}_n}[g(\xi)] - E[g(\xi)] = O_p(n^{-1/2}). \tag{53}$$

Lemma C.3 (Lemma E.4, Kremer et al. (2023)). *Let Assumptions 1-7 be satisfied. Then the matrix*

$$\Sigma(\theta_0) = \langle E[\nabla_\theta \Psi(\xi;\theta_0)], E[\nabla_{\theta^T}\Psi(\xi;\theta_0)]\rangle_{\mathcal{H}^*} \tag{54}$$

is strictly positive definite and non-singular with smallest eigenvalue bounded away from zero.

C.2.2. Spectrum of $\widehat{\Omega}$

Lemma C.4. *Let Assumptions 2 and 3 be satisfied. Then we have*

$$\sup_{\theta\in\Theta, x\in\mathcal{T}\times\mathcal{Y}} \|\psi(x;\theta)\|_\infty \le C_\psi < \infty \tag{55}$$

$$\sup_{\theta\in\Theta, x\in\mathcal{T}\times\mathcal{Y}} \|J_x(\psi)(x;\theta)\|_\infty \le L_\psi < \infty \tag{56}$$

$$\sup_{\theta\in\Theta, x\in\mathcal{T}\times\mathcal{Y}} \|\Delta_x\psi(x;\theta)\|_\infty \le D_\psi < \infty \tag{57}$$

$$\sup_{\theta\in\Theta, z\in\mathcal{Z}} \|h(z)\|_\infty \le C_h < \infty \tag{58}$$

$$\sup_{\theta\in\Theta, z\in\mathcal{Z}} \|J_z h(z)\|_\infty \le L_h < \infty \tag{59}$$

$$\sup_{\theta\in\Theta, z\in\mathcal{Z}} \|\Delta_z h(z)\|_\infty \le D_h < \infty, \tag{60}$$

which directly implies $\|\Delta_\xi\|_{\mathrm{op}} < \infty$ on $\mathcal{H}^*$.

Proof. The proof follows directly from the fact that a continuous function on a compact domain is bounded and both $\psi(\cdot;\theta)$ and h are C^∞-smooth by Assumptions 3 and 5. □

Lemma C.5. *Let $V(Z;\theta) = E[J_x(\psi)(X;\theta)\Gamma^{-1}J_x(\psi)(X;\theta)^T|Z]$ be non-singular with probability 1. Then the linear operator $\Omega(\theta):\mathcal{H}\to\mathcal{H}$ defined as*

$$\Omega(\theta) = E\left[\left(\nabla_\xi\Psi(\xi;\theta)\right)^T\Gamma^{-1}\nabla_\xi\Psi(\xi;\theta)\right] \tag{61}$$

is non-singular.

Proof. We derive the result by showing that the smallest eigenvalue of $\Omega(\theta)$ is positive. Consider any $h\in\mathcal{H}$ with $\|h\|_{L^2(\mathcal{H},P_0)} > 0$ then we have

$$\begin{aligned}
\langle h, \Omega(\theta)h\rangle_{\mathcal{H}} =& E[h(Z)^T J_x(\psi)(X;\theta)\Gamma^{-1}J_x(\psi)(X;\theta)h(Z)] && (62)\\
=& E[h(Z)^T E[J_x(\psi)(X;\theta)\Gamma^{-1}J_x(\psi)(X;\theta)|Z]h(Z)] && (63)\\
=& E[h(Z)^T V_0(Z;\theta)h(Z)] && (64)\\
=& CE[\|h(Z)\|_2^2] && (65)\\
=& C\|h\|^2_{L^2(\mathcal{H},P_0)} > 0 && (66)
\end{aligned}$$

where we used that by assumption $V(Z;\theta)$ is non-singular and thus its smallest eigenvalue C bounded away from zero w.p.1. □

Lemma C.6 (Spectrum of $\widehat{\Omega}$)**.** *Let the assumptions of Theorem 3.4 be satisfied. Then for $\bar{\theta} \in \Theta$ with $\bar{\theta}_n \to \bar{\theta}$, the empirical gradient covariance operator*

$$\widehat{\Omega}_{\lambda_n}(\bar{\theta}_n) = E_{\hat{P}_n}\left[\left(\nabla_\xi \Psi(\xi;\bar{\theta}_n)\right)^T \Gamma^{-1} \nabla_\xi \Psi(\xi;\bar{\theta}_n)\right] + \lambda_n I \otimes I \tag{67}$$

is a positive definite operator with smallest eigenvalue $\lambda_{\min}(\widehat{\Omega})$ bounded away from zero and largest eigenvalue $\lambda_{\max}(\widehat{\Omega}) < C < \infty$ bounded from above w.p.a.1.

Proof. Let in the following $\widehat{\Omega}(\theta) = \widehat{\Omega}_{\lambda_n=0}(\theta)$. With Assumption 4 it follows from Lemma C.5 that the operator $\Omega(\bar{\theta}) := E\left[\left(\nabla_\xi \Psi(\xi;\bar{\theta})\right)^T \Gamma^{-1} \nabla_\xi \Psi(\xi;\bar{\theta})\right]$ is non-singular and thus its smallest eigenvalue bounded away from zero. In the following we show that $\widehat{\Omega}(\bar{\theta}_n) \xrightarrow{p} \Omega(\bar{\theta})$, where the convergence rate in operator norm is $O_p(n^{-\zeta})$. Therefore, by adding the identity operator with regularization parameter λ_n that goes to zero slower than $O_p(n^{-\zeta})$ we ensure that $\widehat{\Omega}_{\lambda_n}(\bar{\theta}_n)$ remains positive definite w.p.a.1. The derivation of this result follows the proof of Lemma 20 of Bennett & Kallus (2023). By the triangle inequality we have

$$\|\widehat{\Omega}(\bar{\theta}_n) - \Omega(\bar{\theta})\|_{\text{op}} \leq \|\widehat{\Omega}(\bar{\theta}_n) - \Omega(\bar{\theta}_n)\| + \|\Omega(\bar{\theta}_n) - \Omega(\bar{\theta})\|. \tag{68}$$

The first term we can estimate using standard results from empirical process theory. Define $\|h\|_{\mathcal{H}}^2 = \frac{1}{m}\sum_{i=1}^m \|h_i\|_{\mathcal{H}_i}^2$ as well as $J_\psi(X;\theta) = J_x\psi(X;\theta)$ and $J_h(Z) = J_z h(Z)$. Let $\mathcal{H}_1 = \{h \in \mathcal{H} : \|h\|_{\mathcal{H}} \leq 1\}$ denote the unit ball in $\mathcal{H}$, then

$$\begin{aligned}
\|\widehat{\Omega}(\bar{\theta}_n) - \Omega(\bar{\theta}_n)\| &= \sup_{h,h' \in \mathcal{H}_1} \langle h', \widehat{\Omega}(\bar{\theta}_n) - \Omega(\bar{\theta}_n) h\rangle_{\mathcal{H}} && (69)\\
&= \sup_{h,h' \in \mathcal{H}_1} \Bigg\{ E_{\hat{P}_n}\left[h(Z)^T J_\psi(X;\bar{\theta}_n)\Gamma_x^{-1} J_\psi(X;\bar{\theta}_n)^T h'(Z)\right] && (70)\\
&\quad - E\left[h(Z)^T J_\psi(X;\bar{\theta}_n)\Gamma_x^{-1} J_\psi(X;\bar{\theta}_n)^T h'(Z)\right] && (71)\\
&\quad + \frac{1}{\gamma_z} E_{\hat{P}_n}\left[\psi(X;\bar{\theta}_n)^T J_h(Z) J_{h'}(Z)^T \psi(X;\bar{\theta}_n)\right] && (72)\\
&\quad - \frac{1}{\gamma_z} E\left[\psi(X;\bar{\theta}_n)^T J_h(Z) J_{h'}(Z)^T \psi(X;\bar{\theta}_n)\right] \Bigg\} && (73)\\
&\leq \sup_{g \in \mathcal{G}^2} \left\{ E_{\hat{P}_n}[g(\xi)] - E[g(\xi)] \right\} + \frac{1}{\gamma_z} \sup_{s \in \mathcal{S}^2} \left\{ E_{\hat{P}_n}[s(\xi)] - E[s(\xi)] \right\} && (74)
\end{aligned}$$

where for $i \in [d_\xi]$ we define

$$\mathcal{G}_i = \{g_i : g_i(\xi) = \sum_{j=1}^m h_j(z) \left(J_\psi(x;\theta)\right)_{ji} \Gamma_{ii}^{-1/2},\ h \in \mathcal{H}_{i,1},\ \theta \in \Theta\} \tag{75}$$

$$\mathcal{G}^2 = \{g : g(\xi) = \sum_{i \in [d_x]} g_i(\xi) g_i'(\xi),\ g_i, g_i' \in \mathcal{G}_i\} \tag{76}$$

$$\mathcal{S}_i = \{s_i : s_i(\xi) = \sum_{j=1}^m \psi_j(x;\theta) \left(J_h(z)\right)_{ji},\ h \in \mathcal{H}_{i,1},\ \theta \in \Theta\} \tag{77}$$

$$\mathcal{S}^2 = \{s_i : s_i(\xi) = \sum_{i \in [d_z]} s_i(\xi) s_i'(\xi),\ s_i, s_i' \in \mathcal{S}_i\} \tag{78}$$

Now for the first term, we have that each $h_j \in \mathcal{H}_{i,1}$ is P_0-Donsker by Assumption 5 and uniformly bounded by Lemma C.4. Similarly each entry of the Jacobian $J_\psi(\cdot;\theta)$ is P_0-Donsker by Assumption 3 and uniformly bounded by Lemma C.4. With that we can employ Lemma C.1 to conclude that $\mathcal{G}_i$ is P_0-Donsker and thus using Lemma C.1 again it follows that $\mathcal{G}^2$ is P_0-Donsker. Therefore we can use Lemma C.2 to obtain $\sup_{g \in \mathcal{G}^2} \left\{ E_{\hat{P}_n}[g(\xi)] - E[g(\xi)] \right\} = O_p(n^{-1/2})$.

For the second term in (74) we have that each $\psi_j(\cdot;\theta)$ is P_0-Donsker by Assumption 3 and uniformly bounded by Lemma C.4. Similarly each entry of the Jacobian $J_z h$ is P_0-Donsker by Assumption 5 and uniformly bounded by Lemma C.4. With that,

again, we can employ Lemma C.1 to conclude that $\mathcal{S}_i$ is P_0-Donsker and thus using Lemma C.1 again it follows that $\mathcal{S}^2$ is P_0-Donsker. Therefore we can use Lemma C.2 to obtain $\frac{1}{\gamma_z}\sup_{s\in\mathcal{S}^2}\left\{E_{\hat{P}_n}[s(\xi)]-E[s(\xi)]\right\}=O_p(n^{-1/2})$.

Putting these results together we finally obtain $\|\widehat{\Omega}(\bar{\theta}_n)-\Omega(\bar{\theta}_n)\|\leq O_p(n^{-1/2})$.

For the second term in (68) we have

$$\|\Omega(\bar{\theta}_n)-\Omega(\bar{\theta})\|=\sup_{h,h'\in\mathcal{H}_1}\langle h',\Omega(\bar{\theta}_n)-\Omega(\bar{\theta})h\rangle_{\mathcal{H}}\leq C_x+\frac{1}{\gamma_z}C_z \tag{79}$$

where

$$\begin{aligned}
C_x=&\sup_{h,h'\in\mathcal{H}_1}E\Big[h'(Z)^T\Big(J_\psi(X;\bar{\theta}_n)\Gamma_x^{-1}J_\psi(X;\bar{\theta}_n)^T &(80)\\
&-J_\psi(X;\bar{\theta})\Gamma_x^{-1}J_\psi(X;\bar{\theta})^T\Big)h(Z)\Big] &(81)\\
=&\sup_{h,h'\in\mathcal{H}_1}E\Big[h'(Z)^T\Big(J_\psi(X;\bar{\theta}_n)\Gamma_x^{-1}J_\psi(X;\bar{\theta}_n)-J_\psi(X;\bar{\theta})\Gamma_x^{-1}J_\psi(X;\bar{\theta})\Big)h(Z)\Big] &(82)\\
=&\sup_{h,h'\in H_1}E\Big[h'(Z)^TJ_\psi(X;\bar{\theta}_n)\Gamma_x^{-1}\left(J_\psi(X;\bar{\theta}_n)-J_\psi(X;\bar{\theta})\right)^Th(Z) &(83)\\
&+h'(Z)^TJ_\psi(X;\bar{\theta})\Gamma_x^{-1}\left(J_\psi(X;\bar{\theta}_n)-J_\psi(X;\bar{\theta})\right)^Th(Z)\Big] &(84)\\
\leq&\frac{2}{\min(\{\gamma_t,\gamma_y\})}m^2C_h^2L_\psi E\left[\|J_\psi(X;\bar{\theta}_n)-J_\psi(X;\bar{\theta})\|_\infty\right] &(85)\\
=&O_p(n^{-\zeta}) &(86)
\end{aligned}$$

where we used that by Lemma C.4, $\sup_{\theta\in\Theta,x\in\mathcal{T}\times\mathcal{Y}}\|J_\psi(x;\theta)\|_\infty\leq L_\psi$ and $\sup_{h\in\mathcal{H}_1,z\in\mathcal{Z}}|h(z)|\leq C_h$ as well as by Assumption 6 $E\left[\|J_\psi(X;\bar{\theta}_n)-J_\psi(X;\bar{\theta})\|_\infty\right]=O_p(n^{-\zeta})$.

Now similarly for the second term in (79) we have

$$\begin{aligned}
C_z=&\sup_{h,h'\in\mathcal{H}_1}E\Big[\operatorname{Tr}\left(J_{h'}(Z)^T\psi(X;\bar{\theta}_n)\psi(X;\bar{\theta}_n)^TJ_h(Z)\right) &(87)\\
&-\operatorname{Tr}\left(J_{h'}(Z)^T\psi(X;\bar{\theta})\psi(X;\bar{\theta})^TJ_h(Z)\right)\Big] &(88)\\
\leq&L_h^2E[\mathbf{1}^T\left(\psi(X;\bar{\theta}_n)\psi(X;\bar{\theta}_n)^T-\psi(X;\bar{\theta})\psi(X;\bar{\theta})^T\right)\mathbf{1}] &(89)\\
=&L_h^2E\left[\mathbf{1}^T\psi(X;\bar{\theta}_n)\left(\psi(X;\bar{\theta}_n)-\psi(X;\bar{\theta})\right)^T\mathbf{1}\right] &(90)\\
&+L_h^2E\left[\mathbf{1}^T\psi(X;\bar{\theta})\left(\psi(X;\bar{\theta}_n)-\psi(X;\bar{\theta})\right)^T\mathbf{1}\right] &(91)\\
\leq&2m^2L_h^2C_\psi E\left[\|\psi(X;\bar{\theta}_n)-\psi(X;\bar{\theta})\|_\infty\right] &(92)\\
=&O_p(n^{-\zeta}), &(93)
\end{aligned}$$

where again we used Lemma C.4 and Assumption 6. Combining both results we obtain $\|\Omega(\bar{\theta}_n)-\Omega(\bar{\theta})\|=\leq C_x+\frac{1}{\gamma_z}C_z\leq O_p(n^{-\zeta})$.

Finally as $0<\zeta\leq 1/2$ it follows that

$$\begin{aligned}
\|\widehat{\Omega}(\bar{\theta}_n)-\Omega(\bar{\theta})\|&\leq\|\widehat{\Omega}(\bar{\theta}_n)-\Omega(\bar{\theta}_n)\|+\|\Omega(\bar{\theta}_n)-\Omega(\bar{\theta})\| &(94)\\
&\leq O_p(n^{-1/2})+O_p(n^{-\zeta})=O_p(n^{-\zeta}). &(95)
\end{aligned}$$

In conclusion we have shown that $\widehat{\Omega}(\bar{\theta}_n)$ converges to the non-singular operator $\Omega(\bar{\theta})$ at rate $O_p(n^{-\zeta})$ and by Assumption 6 we have $\lambda_n=O_p(n^{-\rho})$ with $0<\rho<\zeta$, therefore the operator $\widehat{\Omega}_{\lambda_n}(\bar{\theta}_n)=\widehat{\Omega}(\bar{\theta}_n)+\lambda_nI$ is non-singular with smallest eigenvalue bounded away from zero w.p.a.1.

It remains to be shown that the largest eigenvalue of $\widehat{\Omega}(\bar{\theta}_n)$ is bounded. This is a direct consequence of Lemma C.4. Consider any $h \in \mathcal{H}$ with $\|h\|_{\mathcal{H}} > 0$ and

$$\begin{aligned}\langle h, \widehat{\Omega}(\bar{\theta}_n)h\rangle &= E_{\hat{P}_n}[h(Z)^T J_\psi(X;\bar{\theta}_n)J_\psi(X;\bar{\theta}_n)^T h(Z)] && (96)\\ &\leq E[J_\psi(X;\bar{\theta}_n)\|_\infty^2]E[\|h(Z)\|_\infty^2] && (97)\\ &\leq L_\psi^2 C_h^2 < \infty. && (98)\end{aligned}$$

□

C.2.3. Proof of Theorem 3.4

Lemma C.7. *Let the sets of functions $\{\psi(\cdot;\theta)_l : \theta \in \Theta, l \in [m]\}$ and H_1 be P_0-Donsker. Then we have for any $\theta \in \Theta$*

$$\|E_{\hat{P}_n}[\Psi(\xi;\theta)] - E[\Psi(\xi;\theta)]\|_{\mathcal{H}^*} = O_p(n^{-1/2}). \tag{99}$$

Proof.

$$\begin{aligned}\|E_{\hat{P}_n}[\Psi(\xi;\theta)] - E[\Psi(\xi;\theta)]\|_{\mathcal{H}^*} &= \sup_{h\in\mathcal{H}_1} E_{\hat{P}_n}[\psi(X;\theta)^T h(Z)] - E[\psi(X;\theta)^T h(Z)] && (100)\\ &= \sup_{g\in\mathcal{G}} E_{\hat{P}_n}[g(\xi)] - E[g(\xi)] && (101)\end{aligned}$$

where

$$\mathcal{G} = \{g : g(\xi) = \sum_{i=1}^m \psi_i(x;\theta)h_i(z),\ h_i \in \mathcal{H}_{i,1}, \theta \in \Theta\}. \tag{102}$$

Now as each h_i and $\psi_i(\cdot;\theta)$ are P_0-Donsker by Assumption 5 and 3 respectively and uniformly bounded by Lemma C.4, we can employ Lemma C.1 to conclude that $\mathcal{G}$ is P_0-Donsker. From this, the result follows by application of Lemma C.2. □

Lemma C.8 (Convergence of $\widehat{D}$)**.** *Let the assumptions of Theorem 3.4 be satisfied. Additionally let $\tilde{\theta} \in \Theta$ be a consistent estimator for θ_0, i.e., $\tilde{\theta} \xrightarrow{p} \theta_0$ with $\|E_{\hat{P}_n}[\Psi(\xi;\tilde{\theta})]\|_{\mathcal{H}^*} = O_p\left(n^{-1/2}\right)$. Then for $\tilde{h} = \arg\max_{h\in\mathcal{H}} D(\tilde{\theta}, h)$ we have $\|\tilde{h}\|_{\mathcal{H}} = O_p\left(n^{-1/2}\right)$ and $\widehat{D}(\tilde{\theta}, \tilde{h}) \leq O_p\left(n^{-1}\right)$.*

Proof. Let $\tilde{\Psi} := \frac{1}{n}\sum_{i=1}^n \Psi(\xi_i, \tilde{\theta})$. Then we have

$$\begin{aligned}0 &= \widehat{D}(\tilde{\theta}, 0) && (103)\\ &\leq \arg\max_{h\in\mathcal{H}} D(\tilde{\theta}, \tilde{h}) && (104)\\ &= \left(I + \frac{\epsilon}{2}\Delta_\xi\right)\tilde{\Psi}(\tilde{h}) - \frac{\epsilon}{2}\langle\tilde{h}, \widehat{\Omega}_{\lambda_n}(\bar{\theta}_n)\tilde{h}\rangle_{\mathcal{H}} && (105)\\ &\leq \|I + \frac{\epsilon}{2}\Delta_\xi\|_{\text{op}}\|\tilde{\Psi}\|_{\mathcal{H}^*}\|\tilde{h}\|_{\mathcal{H}} - \frac{\epsilon}{2}\lambda_{\min}\left(\widehat{\Omega}_{\lambda_n}(\bar{\theta}_n)\right)\left\|\tilde{h}\right\|_{\mathcal{H}}^2 && (106)\\ &\leq \left(1 + \frac{\epsilon}{2}\|\Delta_\xi\|\right)\|\tilde{\Psi}\|_{\mathcal{H}^*}\|\tilde{h}\|_{\mathcal{H}} - \frac{\epsilon}{2}\lambda_{\min}\left(\widehat{\Omega}_{\lambda_n}(\bar{\theta}_n)\right)\left\|\tilde{h}\right\|_{\mathcal{H}}^2 && (107)\end{aligned}$$

Using that $\|\Delta_\xi\| < \infty$ by Lemma C.4 and moreover $\lambda_{\min}\left(\widehat{\Omega}_{\lambda_n}(\bar{\theta}_n)\right) > 0$ by Lemma C.6, we get $\|\tilde{h}\|_{\mathcal{H}} \leq C\|\tilde{\Psi}\|_{\mathcal{H}^*}$ and thus $\|\tilde{h}\|_{\mathcal{H}} = O_p\left(n^{-1/2}\right)$. Now inserting back into $\widehat{D}$ we get $\widehat{D}(\tilde{\theta}, \tilde{h}) \leq O_p\left(n^{-1}\right)$. □

Lemma C.9 (Convergence of $\|\hat{\Psi}\|_{\mathcal{H}^*}$)**.** *Let the assumptions of Theorem 3.4 be satisfied. Let $\hat{\theta} = \arg\min_{\theta\in\Theta}\sup_{h\in\mathcal{H}}\widehat{D}(\theta, h)$ denote the SMM estimator for θ_0. Then $\left\|E_{\hat{P}_n}[\Psi(\xi;\hat{\theta})]\right\|_{\mathcal{H}^*} = O_p(n^{-1/2})$.*

Proof. Let $\hat{\Psi} = \frac{1}{n}\sum_{i=1}^{n} \Psi(\xi, \hat{\theta})$. Let $\phi(\hat{\Psi}) \in \mathcal{H}$ denote the Riesz representer of $\hat{\Psi} \in \mathcal{H}^*$. Consider any $\sigma_n \to 0$ and define $h_{\hat{\Psi}} = \sigma_n \phi(\hat{\Psi})$. Using that the eigenvalues of the Laplacian Δ_ξ are bounded by Lemma C.4 and the largest eigenvalue of $\widehat{\Omega}(\bar{\theta}_n)$ is bounded by a constant C by Lemma C.6, we have

$$\widehat{D}(\hat{\theta}, h_{\hat{\Psi}}) = \left(I + \frac{\epsilon}{2}\Delta_\xi\right)\hat{\Psi}(h_{\hat{\Psi}}) - \frac{\epsilon}{2}\langle h_{\hat{\Psi}}, \widehat{\Omega}(\bar{\theta}_n) h_{\hat{\Psi}}\rangle_{\mathcal{H}} \tag{108}$$

$$\geq \left(1 + \frac{\epsilon}{2}\lambda_{\min}(\Delta_\xi)\right)\hat{\Psi}(h_{\hat{\Psi}}) - \frac{\epsilon}{2} C \|h_{\hat{\Psi}}\|^2_{\mathcal{H}} \tag{109}$$

$$\geq C'\sigma_n \|\hat{\Psi}\|^2_{\mathcal{H}^*} - \frac{C\epsilon}{2}\sigma_n^2 \|\hat{\Psi}\|^2_{\mathcal{H}^*}, \tag{110}$$

where by assumption on ϵ we have $C' = 1 + \frac{\epsilon}{2}\lambda_{\min}(\Delta_\xi) \neq 0$ w.p.1. Now, as $\hat{\theta}$ is the minimizer of the Sinkhorn profile $R(\theta) = \max_{h\in\mathcal{H}} \widehat{D}(\theta, h)$ we have

$$C'\sigma_n \|\hat{\Psi}\|^2_{\mathcal{H}^*} - \frac{C\epsilon}{2}\sigma_n^2 \|\hat{\Psi}\|^2_{\mathcal{H}^*} \leq \widehat{D}(\hat{\theta}, h_{\hat{\Psi}}) \leq \widehat{D}(\hat{\theta}, \hat{h}) \leq \max_{h\in\mathcal{H}} \widehat{D}(\theta_0, h) \leq O(n^{-1}), \tag{111}$$

where in the last step we used that $\|E_{\hat{P}_n}[\Psi(\xi;\theta_0)]\|_{\mathcal{H}^*} = O_p(n^{-1/2})$ by Lemma C.7 and thus the assumptions of Lemma C.8 are fulfilled and we get $\max_{h\in\mathcal{H}} \widehat{D}(\theta_0, h) \leq O(n^{-1})$. Thus we have $\sigma_n(C' - \frac{C\epsilon}{2}\sigma_n)\|\hat{\Psi}\|^2_{\mathcal{H}^*} = O_p(n^{-1})$ and as $(C' - \frac{C\epsilon}{2}\sigma_n)$ is bounded away from zero for all n large enough, we have $\sigma_n \|\hat{\Psi}\|^2_{\mathcal{H}^*} \leq O_p(n^{-1})$. As this holds for any $\sigma_n \xrightarrow{p} 0$ we finally have $\|\hat{\Psi}\|_{\mathcal{H}^*} = O_p(n^{-1/2})$. □

Proof of Theorem 3.4 Using the result of Lemma C.9 for the convergence rate of the empirical moment functional, the proof of the consistency of our SMM estimator is identical to the ones provided by Kremer et al. (2022) and Kremer et al. (2023) for their estimtators. We provide it here for completeness.

Proof. From Lemma C.7 it follows that $\|E_{\hat{P}_n}[\Psi(\xi;\theta)] - E[\Psi(\xi;\theta)]\|_{\mathcal{H}^*} = O_p(n^{-1/2})$ for any $\theta \in \Theta$. By Lemma C.9 we have $\|E_{\hat{P}_n}[\Psi(\xi;\hat{\theta})]\|_{\mathcal{H}^*} = O_p(n^{-1/2})$ and thus using the triangle inequality we get

$$\begin{aligned}
\left\|E[\Psi(\xi;\hat{\theta})]\right\|_{\mathcal{H}^*} &= \left\|E[\Psi(\xi;\hat{\theta})] - E_{\hat{P}_n}[\Psi(\xi;\hat{\theta})] + E_{\hat{P}_n}[\Psi(\xi;\hat{\theta})]\right\|_{\mathcal{H}^*} \\
&\leq \left\|E[\Psi(\xi;\hat{\theta})] - E_{\hat{P}_n}[\Psi(\xi;\hat{\theta})]\right\|_{\mathcal{H}^*} + \left\|E_{\hat{P}_n}[\Psi(\xi;\hat{\theta})]\right\|_{\mathcal{H}^*} \\
&= O_p(n^{-1/2}) \xrightarrow{p} 0.
\end{aligned}$$

As by Assumption 1, θ_0 is the unique parameter for which $E[\psi(T, Y; \theta)|Z] = 0$ P_z-a.s. and by Assumption 5 this is fulfilled if and only if $\|E[\psi(\xi;\theta)]\|_{\mathcal{H}^*} = 0$, it follows that $\hat{\theta} \xrightarrow{p} \theta_0$.

Under the additional Assumption 7 we can use this result to translate the convergence rate of the moment functional to a convergence rate of the estimator $\hat{\theta}$.

As $\Psi(\xi;\theta)$ is continuously differentiable in its second argument which follows immediately from Assumption 7 and the definition of Ψ, we can use the mean value theorem to expand $\Psi(\xi, \hat{\theta})$ about θ_0, i.e., there exists $\bar{\theta} \in \mathrm{conv}(\{\theta_0, \hat{\theta}\})$ such that

$$\Psi(\xi;\hat{\theta}) = \Psi(\xi;\theta_0) + (\hat{\theta} - \theta_0)^T \nabla_\theta \Psi(\xi;\bar{\theta}). \tag{112}$$

Using this we have

$$\|E[\Psi(\xi;\hat{\theta})]\|^2_{\mathcal{H}^*} = \|\underbrace{E[\Psi(\xi;\theta_0)]}_{=0} + (\hat{\theta} - \theta_0)^T E[\nabla_\theta \Psi(\xi;\bar{\theta})]\|^2_{\mathcal{H}^*} \tag{113}$$

$$= \left\langle (\hat{\theta} - \theta_0)^T E[\nabla_\theta \Psi(\xi;\bar{\theta})], (\hat{\theta} - \theta_0)^T E[\nabla_\theta \Psi(\xi;\bar{\theta})] \right\rangle_{\mathcal{H}^*} \tag{114}$$

$$= (\hat{\theta} - \theta_0)^T \underbrace{\left\langle E[\nabla_\theta \Psi(\xi;\bar{\theta})], E[\nabla_{\theta^T} \Psi(\xi;\bar{\theta})]\right\rangle_{\mathcal{H}^*}}_{=:\Sigma(\bar{\theta})} (\hat{\theta} - \theta_0) \tag{115}$$

$$\geq \lambda_{\min}\left(\Sigma(\bar{\theta})\right) \|\hat{\theta} - \theta_0\|_2^2. \tag{116}$$

Now as $\hat{\theta} \xrightarrow{p} \theta_0$ and $\bar{\theta} \in \text{conv}(\{\theta_0, \hat{\theta}\})$ we have $\bar{\theta} \xrightarrow{p} \theta_0$ and thus $\Sigma(\bar{\theta}) \xrightarrow{p} \Sigma(\theta_0) =: \Sigma_0$ by the continuous mapping theorem. By the non-negativity of the norm Σ_0 is positive-semi definite and non-singular by Lemma C.3, thus the smallest eigenvalue of $\Sigma(\bar{\theta})$, $\lambda_{\min}(\Sigma(\bar{\theta}))$, is positive and bounded away from zero w.p.a.1. Finally as $\|E[\Psi(X, Z; \hat{\theta})]\| = O_p(n^{-1/2})$ taking the square-root on both sides we have $\|\hat{\theta} - \theta_0\| = O_p(n^{-1/2})$. □

Proof of Proposition 3.5

Proof. For a universal ISPD kernel, equivalence of the conditional and the variational moment restrictions (1) and (2) follows by Theorem 3.9 of Kremer et al. (2022). The Donsker property of the unit ball in an RKHS of a smooth universal kernel with compact domain follows from Lemma 17 of Bennett & Kallus (2023). Finally, the Donsker property of the Jacobian $J_z h$ of h follows by the same argument as Lemma 17 of Bennett et al. (2019) using C^∞ smoothness of h and boundedness of $J_z h$. □

Proof of Theorem 3.6

Proof. Under the assumptions the Sinkhorn profile is given as

$$R_\lambda(f) = \sup_{h \in \mathcal{H}} \Bigg\{ E_{\hat{P}_n}\left[h(Z)^T \left(I + \frac{\epsilon}{2}\Delta_x\right) \psi(X; f)\right] \tag{117}$$

$$- \epsilon E_{\hat{P}_n}\left[h(Z)^T J_\psi(X; \tilde{f}) \Gamma_x^{-1} J_\psi(X; \tilde{f})^T h(Z)\right] - \frac{\lambda}{2}\|h\|_{\mathcal{H}}^2 \Bigg\} \tag{118}$$

which as the unconstrained maximization of a concave objective is a convex optimization problem. Moreover, the conditions of the classical representer theorem (Schölkopf et al., 2001) are fulfilled and thus the maximizer of (118) is given as $h_l = \sum_{i=1}^n \alpha_i^l k_l(z_i, \cdot)$ with $\alpha^l \in \mathbb{R}^n$. Inserting this into (118) and defining the kernel Gram matrices $K_l \in \mathbb{R}^{n \times n}$ with entries $(K_l)_{ij} = k_l(z_i, z_j)$ we obtain

$$R_\lambda(f) = \sup_{\alpha \in \mathbb{R}^{nm}} \frac{1}{n} \sum_{i,j=1}^n \sum_{l=1}^m \alpha_i^l (K_l)_{ij} \left(I + \frac{\epsilon}{2}\Delta_x\right) \psi_l(x_j; f) - \frac{\lambda}{2} \sum_{l=1}^m (\alpha^l)^T K_l (\alpha^l) \tag{119}$$

$$- \frac{\epsilon}{2n} \sum_{i,j,k=1}^n \sum_{l,r=1}^m \alpha_i^l (K_l)_{ij} \nabla_x \psi_l(x_j; \tilde{f})^T \Gamma_x^{-1} \nabla_x \psi_r(x_j; \tilde{f}) (K_r)_{jk} \alpha_r^k \tag{120}$$

$$= \sup_{\alpha \in \mathbb{R}^{nm}} \frac{1}{n} \alpha^T L \psi_\Delta - \frac{1}{2} \alpha^T \left(\epsilon Q(\tilde{f}) + \lambda L\right) \alpha \tag{121}$$

where we defined $\psi_\Delta(f) \in \mathbb{R}^{nm}$, $L \in \mathbb{R}^{nm \times nm}$ and $Q(f) \in \mathbb{R}^{nm \times nm}$ with entries

$$\psi_\Delta(f)_{i \cdot l} = \left(I + \frac{\epsilon}{2}\Delta_x\right) \psi_l(x_i; f) \tag{122}$$

$$L_{(i \cdot l),(j \cdot r)} = \delta_{lr} k_l(z_i, z_j) \tag{123}$$

$$Q(f)_{(i \cdot l),(j \cdot r)} = \frac{1}{n} \sum_{k=1}^n \sum_{s=1}^{d_x} k_l(z_i, z_k) \nabla_{x_s} \psi_l(x_k; f) (\Gamma_x^{-1})_{ss} \nabla_{x_s} \psi_r(x_k; f) k_r(z_k, z_j). \tag{124}$$

The first order optimality conditions for α read

$$0 = \frac{1}{n} L \psi_\Delta(f) - \left(\epsilon Q(\tilde{f}) + \lambda L\right) \alpha, \tag{125}$$

which immediately gives

$$\alpha = \left(\epsilon Q(\tilde{f}) + \lambda L\right)^{-1} \frac{1}{n} L \psi_\Delta(f). \tag{126}$$

Inserting back into $R_\lambda(f)$ and multiplying by $\epsilon > 0$ we obtain

$$R_\lambda(f) = \frac{1}{2n^2}\psi_\Delta(f)^T L \left(Q(\tilde{f}) + \frac{\lambda}{\epsilon}L\right)^{-1} L\psi_\Delta(f). \tag{127}$$

□