Estimation Beyond Data Reweighting: Kernel Method of Moments

Heiner Kremer [1] **Yassine Nemmour** [1] **Bernhard Schölkopf** [1 2] **Jia-Jie Zhu** [3]

Abstract

Moment restrictions and their conditional counterparts emerge in many areas of machine learning and statistics ranging from causal inference to reinforcement learning. Estimators for these tasks, generally called *methods of moments*, include the prominent *generalized method of moments* (GMM) which has recently gained attention in causal inference. GMM is a special case of the broader family of *empirical likelihood estimators* which are based on approximating a population distribution by means of minimizing a φ-divergence to an empirical distribution. However, the use of φ-divergences effectively limits the candidate distributions to reweightings of the data samples. We lift this long-standing limitation and provide a method of moments that goes beyond data reweighting. This is achieved by defining an empirical likelihood estimator based on maximum mean discrepancy which we term the *kernel method of moments* (KMM). We provide a variant of our estimator for conditional moment restrictions and show that it is asymptotically first-order optimal for such problems. Finally, we show that our method achieves competitive performance on several conditional moment restriction tasks.

1. Introduction

Many problems in machine learning, statistics, causal inference and economics can be formulated as (conditional) moment restrictions (Newey & Powell, 2003; Angrist & Pischke, 2008). Moment restrictions (MR) identify a parameter of interest by restricting the expectation over a so-called moment function to a fixed value. From a machine learning perspective, moment restrictions subsume empirical risk minimization since the corresponding first order conditions imply that the expectation of the gradient of the loss function vanishes. A significantly harder problem is posed by conditional moment restrictions (CMR), which restrict the *conditional* expectation of the moment function. In this case estimation effectively requires solving a continuum of unconditional moment restrictions (Bierens, 1982). A prominent CMR problem is instrumental variable (IV) regression (Newey & Powell, 2003), where the expectation of the prediction residual conditioned on the instruments is required to be zero. The CMR formulation of IV regression is a powerful way to define estimators that avoid two-step procedures as, e.g., in the common two-stage least squares method (Angrist & Pischke, 2008). Other examples of CMR problems include variants of double machine learning (Chernozhukov et al., 2016; 2017; 2018) and off-policy evaluation in reinforcement learning (Xu et al., 2021; Chen et al., 2021; Bennett & Kallus, 2020a; Bennett et al., 2021). Perhaps the most popular approach to learning with moment restrictions is the generalized method of moments (GMM) of Hansen (1982), which recently gained popularity in machine learning (Lewis & Syrgkanis, 2018; Bennett & Kallus, 2020b). GMM belongs to the wider family of generalized empirical likelihood (GEL) estimators of Owen (1988; 1990); Qin & Lawless (1994); Smith (1997). While moment restrictions are imposed with respect to a population distribution, in practice one usually only has access to an empirical sample from this distribution. GEL estimators are based on simultaneously finding the model parameters and an approximation of the population distribution by considering distributions with minimal distance to the empirical distribution for which the moment restrictions can be fulfilled exactly. The various GEL estimators differ in the choice of φ-divergence used to define this distance. In this context, the continuous updating version of GMM (Hansen et al., 1996) can be interpreted as a GEL estimator with χ^2-divergence. However, the use of φ-divergences effectively restricts the set of candidate distributions to multinomial distributions on the empirical sample, i.e., reweightings of the data, which can be a crude approximation especially in the low sample regime. In the present work, we define the first method of moments estimator that parts with this limitation by defining a GEL framework based on a fundamentally different notion of distributional distance, namely the maximum mean discrepancy (MMD). This allows us to consider arbitrary candidate distributions

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]Eidgenössische Technische Hochschule Zürich, Switzerland [3]Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany. Correspondence to: Heiner Kremer <hkremer@tue.mpg.de>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

with support different from the empirical distribution. As in many cases the population distribution is continuous, this bears the potential to find better approximations thereof. In principle, our flexible framework even allows to evolve the class of candidate distributions over the course of the optimization and thus might benefit from developments in gradient flows and optimal transport. The practical benefit of our approach is demonstrated by competitive empirical performance.

Our Contributions

1. We propose the first method of moments estimator without the limitation to data reweightings by extending the GEL framework to MMD. We derive the dual problem of the resulting inner optimization problem which is a semi-infinitely constrained convex program.
2. To overcome computational challenges, we introduce entropy regularization and show that the dual of the inner problem gives rise to an unconstrained convex program, turning a semi-infinite formulation into either a soft-constraint or log-barrier setting.
3. We provide the first order asymptotics and demonstrate that our estimator is asymptotically optimal for CMR estimation in the sense that it achieves the semi-parametric efficiency bound of Chamberlain (1987).
4. We provide details on the practical implementation and empirically demonstrate state-of-the-art performance of our method on several CMR problems.
5. We release an implementation of our method as part of a software package for (conditional) moment restriction estimation.

The remainder of the paper is structured as follows. Section 2 gives an overview of method of moments estimation for conditional and unconditional moment restrictions. Section 3 introduces our estimator, and provides duality results as well as asymptotic properties and practical considerations. Section 4 provides an empirical evaluation of our estimators on various conditional moment restriction tasks. Section 5 discusses connections to related methods and Section 6 concludes.

2. Background

Method of Moments Let X be a random variable taking values in $\mathcal{X} \subseteq \mathbb{R}^r$ distributed according to P_0. In the following we will denote the expectation with respect to a distribution P by $E_P[\cdot]$ and drop the subscript whenever we refer to the population distribution P_0. Moment restrictions identify a parameter of interest $\theta_0 \in \Theta \subseteq \mathbb{R}^p$ by restricting the expectation of a so-called moment function $\psi : \mathcal{X} \times \Theta \to \mathbb{R}^m$, such that

$$E[\psi(X; \theta_0)] = 0 \in \mathbb{R}^m.$$

In practice, the true distribution P_0 is generally unknown and one only has access to a sample $\{x_i\}_{i=1}^n$ with empirical distribution $\hat{P}_n = \sum_{i=1}^n \frac{1}{n} \delta_{x_i}$, where δ_{x_i} denotes a Dirac measure centered at x_i. The corresponding *empirical* moment restrictions can be defined as

$$E_{\hat{P}_n}[\psi(X; \theta)] = 0, \quad \theta \in \Theta.$$

If the number of restrictions m does not exceed the number of parameters p, these can often be solved exactly. For example, suppose we are interested in estimating the mean θ of a distribution. Then, solving the empirical moment restrictions for the moment function $\psi(X; \theta) = X - \theta$ yields the maximum likelihood estimate $\theta = \frac{1}{n} \sum_{i=1}^n x_i$. However, in the so-called *overidentified* case with $m > p$, the system of equations is generally over-determined and the empirical moment restrictions cannot be satisfied exactly. This is the domain of the celebrated generalized method of moments (GMM) of Hansen (1982). Instead of trying to satisfy the moment restrictions exactly, GMM relaxes the problem into a minimization of a quadratic form,

$$\theta^{\text{GMM}} = \arg\min_{\theta \in \Theta} E_{\hat{P}_n}[\psi(X; \theta)]^T \left(\widehat{\Omega}(\tilde{\theta}) \right)^{-1} E_{\hat{P}_n}[\psi(X; \theta)],$$

where $\widehat{\Omega}(\tilde{\theta}) = E_{\hat{P}_n}[\psi(X; \tilde{\theta}) \psi(X; \tilde{\theta})^T] \in \mathbb{R}^{m \times m}$ denotes the empirical covariance matrix evaluated at a first stage estimate $\tilde{\theta}$.

Empirical Likelihood Estimation GMM is a special case of the wider family of generalized empirical likelihood estimators (Owen, 1988; 1990; Qin & Lawless, 1994). In an attempt to improve the finite sample properties of GMM, alternative estimators from this family have been proposed (Smith, 1997; Newey & Smith, 2004). GEL estimation is based on the idea that while it might not be possible to satisfy the moment restrictions with respect to the empirical distribution $\hat{P}_n$, the population distribution P_0, for which the moment restrictions hold at the true parameter θ_0, will be in a shrinking neighbourhood of $\hat{P}_n$ as the number of samples n grows. Therefore GEL seeks to find a parameter θ and a distribution P for which the moment restrictions hold exactly while staying as close as possible to the empirical distribution. For a convex function $\varphi : [0, \infty) \to (-\infty, \infty]$ define the φ-divergence between distributions P and Q as $D_\varphi(P||Q) = \int \varphi\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q$, where $\frac{\mathrm{d}P}{\mathrm{d}Q}$ denotes the Radon-Nikodym derivative of P with respect to Q. Define the *profile divergence* with respect to a φ-divergence as

$$R(\theta) = \inf_{P \ll \hat{P}_n} D_\varphi(P||\hat{P}_n) \quad \text{s.t.} \quad E_P[\psi(X; \theta)] = 0, \quad E_P[1] = 1, \tag{1}$$

where $P \ll \hat{P}_n$ is the set of positive measures P that are absolutely continuous w.r.t. the empirical distribution $\hat{P}_n$. The GEL estimator then results from minimizing the profile divergence over $\theta \in \Theta$,

$$\theta^{\mathrm{GEL}} = \arg\min_{\theta \in \Theta} R(\theta).$$

Due to the absolute continuity assumption, the distributions considered by GEL are reweightings of the empirical data. Being a special case of GEL, GMM therefore also implicitly corresponds to reweightings of the data as formalized by the following proposition which follows directly from the equivalence result of Newey & Smith (2004) (Theorem 2.1).

Proposition 2.1. *The first order optimality conditions for the continuous updating GMM estimator and the GEL estimator with χ^2-divergence coincide. As the optimal distribution of the latter is given by $P^* = \sum_{i=1}^n p_i \delta_{x_i}$ for some $p \in \mathbb{R}^n$ with $\sum_{i=1}^n p_i = 1$, in consequence, GMM implicitly corresponds to a reweighting of the data.*

Conditional Moment Restrictions In practice, many interesting problems can be formulated as *conditional* moment restrictions, where the estimating equations are given by a conditional expectation over the moment function. Let Z be an additional random variable taking values in $\mathcal{Z}$, then conditional moment restrictions take the form

$$E[\psi(X; \theta_0)|Z] = 0, \ P_Z\text{-a.s.}, \tag{2}$$

where the restrictions need to hold almost surely (a.s.) with respect to the marginal distribution P_Z over Z corresponding to P_0. As conditional moment restrictions are difficult to handle in practice, many proposed estimators rely on transforming them into a corresponding continuum of unconditional restrictions (Bierens, 1982) of the form

$$E[\psi(X; \theta)^T h(Z)] = 0 \ \ \forall h \in \mathcal{H}, \tag{3}$$

where the expectation is taken over the joint distribution of X and Z and $\mathcal{H}$ is a sufficiently rich function space. Examples of such spaces are the Hilbert space of square integrable functions or the reproducing kernel Hilbert space of a universal kernel (Micchelli et al., 2006). Both the GMM and the GEL framework have been extended to conditional moment restrictions in multiple ways, building on basis function expansions of $\mathcal{H}$ (Carrasco & Florens, 2000; Tripathi & Kitamura, 2003; Ai & Chen, 2003; Chaussé, 2012; Carrasco & Kotchoni, 2017) as well as modern machine learning models (Hartford et al., 2017; Lewis & Syrgkanis, 2018; Bennett et al., 2019; Kremer et al., 2022).

Reproducing Kernel Hilbert Spaces A reproducing kernel Hilbert space (RKHS) $\mathcal{F}$ is a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ in which all point evaluation functionals are bounded. Let $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ denote the inner product on $\mathcal{F}$ and define the RKHS norm as the induced norm $\|f\|_{\mathcal{F}} = \sqrt{\langle f, f \rangle_{\mathcal{F}}}$. With every RKHS one can associate a unique kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the reproducing property $\langle k(x, \cdot), f \rangle_{\mathcal{F}} = f(x)$ for any $f \in \mathcal{F}$ and $x \in \mathcal{X}$. A kernel is called integrally strictly positive definite (ISPD) if for any $f \in \mathcal{F}$ with $0 < \|f\|_2^2 < \infty$ we have $\int_{\mathcal{X}} f(x) k(x, x') f(x') \, \mathrm{d}x \, \mathrm{d}x' > 0$. Let $\mathcal{P}$ denote a space of probability distributions, then we define the kernel mean embedding of $P \in \mathcal{P}$ as $\mu_P = E_P[k(X, \cdot)] \in \mathcal{F}$, which has the property that $\langle \mu_P, f \rangle_{\mathcal{F}} = E_P[f(X)] \, \forall f \in \mathcal{F}$. This can be used to define a metric on a space of probability distributions $\mathcal{P}$. For $P, Q \in \mathcal{P}$ the maximum mean discrepancy (MMD) (Gretton et al., 2012) is defined as $\mathrm{MMD}(P, Q; \mathcal{F}) := \|\mu_P - \mu_Q\|_{\mathcal{F}}$. Refer to, e.g., Schölkopf & Smola (2002); Berlinet & Thomas-Agnan (2011); Steinwart & Christmann (2008) for comprehensive introductions to kernel methods for machine learning.

3. Kernel Method of Moments

In this section we derive the KMM estimator for unconditional and conditional moment restrictions and explore its properties. We first derive an exact MMD-based GEL estimator that leads to a difficult semi-infinitely constrained optimization problem for the *MMD profile* $R(\theta)$. We show that an entropy regularized version of our estimator leads to an unconstrained convex dual program which can be readily solved with, e.g., first order optimization methods. We show that our estimator is consistent and optimal for (conditional) moment restriction problems in the sense that it achieves the lowest possible asymptotic variance among all estimators based solely on the CMR. Finally we provide details on the computational procedure. All proofs are deferred to Section E.

3.1. Our Method

Our goal is to define a profile function $R(\theta)$ based on maximum mean discrepancy instead of φ-divergences, such that the KMM estimator can be obtained as $\hat{\theta} = \arg\min_{\theta \in \Theta} R(\theta)$. Let $\psi : \mathcal{X} \times \Theta \to \mathbb{R}^m$ denote a moment function and let $\mathcal{P}$ denote the space of positive measures. Further let $\mathcal{F}$ be an RKHS corresponding to a universal kernel. Then we can define the *MMD-profile* as

$$\begin{aligned} R(\theta) = \inf_{P \in \mathcal{P}} \frac{1}{2} \mathrm{MMD}(P, \hat{P}_n; \mathcal{F})^2 \\ \text{s.t.} \quad E_P[\psi(X; \theta)] = 0, \quad E_P[1] = 1, \end{aligned} \tag{4}$$

where we set $R(\theta) = \infty$ whenever the optimization problem is infeasible. The restriction to RKHS $\mathcal{F}$ corresponding to universal kernels ensures that $\mathrm{MMD}(P, Q; \mathcal{F}) = 0$ if and only if $P = Q$ and thus ensures uniqueness of the infimum in (4) as $n \to \infty$. Using Lagrange duality we can derive the corresponding dual problem as formalized in the following.

Theorem 3.1. *The* MMD profile (4) *has the strongly dual form*

$$R(\theta) = \sup_{\substack{\eta\in\mathbb{R}, f\in\mathcal{F},\\ h\in\mathbb{R}^m}} \frac{1}{n}\sum_{i=1}^{n} f(x_i) + \eta - \frac{1}{2}\|f\|_{\mathcal{F}}^2 \qquad (5)$$
$$\text{s.t.} \quad f(x) + \eta \le \psi(x;\theta)^T h \quad \forall x \in \mathcal{X}.$$

Note that the structure and derivation of (5) resembles recent reformulation techniques for MMD-based distributionally robust optimization (DRO) by Zhu et al. (2021) and in fact GEL estimation can be seen as a dual problem to DRO (Lam, 2019).

While MMD enjoys many favorable properties, the MMD-profile (5), involves a (semi-)infinite constraint which is difficult to handle in practice, especially when combined with stochastic-gradient-type algorithms in machine learning. In the following section, we will show that these limitations can be lifted by introducing entropy regularization.

3.2. Entropy Regularization

Inspired by the interior point method for convex optimization in finite-dimensions (Nesterov & Nemirovskii, 1994), we define an entropy-regularized version of the MMD profile (5). This allows us to translate the semi-infinite constraint in (5) into an additional term in the objective of an unconstrained optimization problem. Note that entropy regularization has been used before in the context of the computation of optimal transport distances (Cuturi, 2013). Our use here is different as we do not regularize a distance computation but instead regularize the duality structure to handle the semi-infinite constraint. To the best of our knowledge entropy regularization has not been combined with MMD in this context.

For a convex function $\varphi : [0,\infty) \to (-\infty,\infty]$ define the relative entropy (or φ-divergence) between a reference distribution ω and a distribution P, which admits a density p w.r.t. ω, as

$$D_\varphi(P||\omega) = \int_{\mathcal{X}} \varphi(p(x))\omega(\mathrm{d}x). \qquad (6)$$

Then for any $\epsilon > 0$ we define the *entropy-regularized MMD profile* for a moment restriction of the form $E[\psi(X;\theta_0)] = 0$ as

$$R_\epsilon^\varphi(\theta) = \inf_{P\ll\omega} \frac{1}{2}\mathrm{MMD}(P,\hat{P}_n;\mathcal{F})^2 + \epsilon D_\varphi(P||\omega) \qquad (7)$$
$$\text{s.t.} \quad E_P[\psi(X;\theta)] = 0, \quad E_P[1] = 1.$$

In contrast to the classical φ-divergence-based profile divergence (1), the regularized MMD profile does not require $P \ll \hat{P}_n$. Instead, absolute continuity is only imposed with respect to an arbitrary (potentially continuous) reference distribution ω which can be constructed in a data-driven way (cf. Section D.3). In practice, by sampling from ω, this allows us to approximate the population distribution P_0 in an arbitrarily fine-grained way instead of using mere reweightings of the training data as in GEL/GMM, which can be a rough approximation especially in the low data regime.

Using Lagrangian duality we can derive the dual problem of (7) as formalized in the following theorem.

Theorem 3.2 (KMM Duality)**.** *The entropy-regularized MMD profile* (7) *has the strongly dual form*

$$R_\epsilon^\varphi(\theta) = \sup_{\substack{\eta\in\mathbb{R}, f\in\mathcal{F},\\ h\in\mathbb{R}^m}} \frac{1}{n}\sum_{i=1}^{n} f(x_i) + \eta - \frac{1}{2}\|f\|_{\mathcal{F}}^2 \qquad (8)$$
$$- \epsilon\int_{\mathcal{X}} \varphi^*\left(\frac{f(x)+\eta-\psi(x;\theta)^T h}{\epsilon}\right)\omega(\mathrm{d}x).$$

where $\varphi^*(t) := \sup_s \langle t, s\rangle - \varphi(s)$ *denotes the convex conjugate of* φ*. The optimization problem in* (8) *is jointly convex in the dual variables* (η, f, h)*.*

As opposed to the unregularized version (5) the entropy-regularized MMD profile (7) is a jointly convex, unconstrained optimization problem over the dual variables. Finally the KMM estimator can be obtained as the minimizer of the entropy-regularized MMD profile

$$\hat{\theta} = \arg\min_{\theta\in\Theta} R_\epsilon^\varphi(\theta).$$

3.3. Choices of Entropy Regularizers

Different choices of φ-divergences in (6) correspond to different relaxations of the generally intractable semi-infinite constraint in (5). Choosing the φ-divergence as the Kullback-Leibler divergence, i.e., $\varphi(p) = p\log(p) - p + 1$ we obtain

$$R_\epsilon^{\mathrm{KL}}(\theta) = \sup_{\substack{\eta\in\mathbb{R}, f\in\mathcal{F},\\ h\in\mathbb{R}^m}} \frac{1}{n}\sum_{i=1}^{n} f(x_i) + \eta - \frac{1}{2}\|f\|_{\mathcal{F}}^2$$
$$- \epsilon\int_{\mathcal{X}} \exp\left(\frac{f(x)+\eta-\psi(x;\theta)^T h}{\epsilon}\right)\omega(\mathrm{d}x).$$

This corresponds to relaxing the constraint in (5) into a soft version, such that violations can occur but are exponentially penalized. Another particularly interesting example is obtained by choosing the φ-divergence to be the backward KL-divergence or Burg's entropy $\varphi(p) = -\log p + p - 1$, which leads to a dual regularized MMD profile with a log-barrier

$$R_\epsilon^{\log}(\theta) = \sup_{\substack{\eta\in\mathbb{R}, f\in\mathcal{F},\\ h\in\mathbb{R}^m}} \frac{1}{n}\sum_{i=1}^{n} f(x_i) + \eta - \frac{1}{2}\|f\|_{\mathcal{F}}^2$$
$$+\epsilon\int_{\mathcal{X}} \log\left(1 - \frac{f(x)+\eta-\psi(x;\theta)^T h}{\epsilon}\right)\omega(\mathrm{d}x).$$

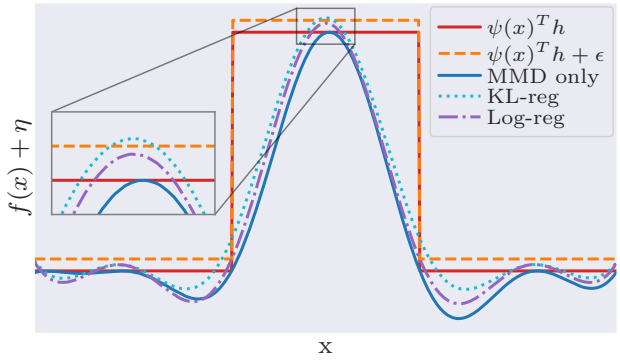


Figure 1. Effect of Entropy Regularization. The red and orange lines correspond to an exemplary function $\psi(x;\theta)^T h$ and its relaxation $\psi(x;\theta)^T h + \epsilon$. The blue line shows the strictly minorizing RKHS function resulting from enforcing the constraint in (5) exactly. The cyan and purple lines correspond to the φ-divergence regularized problem. The log-divergence works as a barrier-function which allows to violate the constraint in (5) by up to ϵ. The KL-divergence yields a soft constraint by penalizing violations exponentially.

Numerically, the log-barrier enforces the solution to lie in the interior of the constraint set, i.e.,

$$f(x_i) + \eta - \psi(x_i;\theta)^T h < \epsilon.$$

Therefore, this can be seen as an interior-point method for handling the infinite constraint in (5). Figure 1 provides a visualization of the different regularization schemes. Refer to Section D for additional details.

3.4. KMM for Conditional Moment Restrictions

The KMM estimator can be generalized to conditional moment restrictions via a functional formulation following the approach of Kremer et al. (2022).

Suppose we have a sufficiently rich Hilbert space $\mathcal{H}$ of functions $h : \mathcal{Z} \to \mathbb{R}^m$, such that we can write conditional moment restrictions of the form (2) as a continuum of unconditional restrictions (3). Let $\mathcal{H}^*$ denote the dual space of functionals $\Psi : \mathcal{H} \to \mathbb{R}$, equipped with the dual norm $\|\Psi\|_{\mathcal{H}^*} = \sup_{\|h\|_{\mathcal{H}}=1} \Psi(h)$. Then for any $(x, z, \theta) \in \mathcal{X} \times \mathcal{Y} \times \Theta$ we define the *moment functional* $\Psi(x, z; \theta) \in \mathcal{H}^*$ such that

$$\Psi(x, z; \theta)(h) = \psi(x;\theta)^T h(z).$$

The continuum of moment restrictions (3) can thus be written as

$$E[\Psi(X, Z; \theta)] = 0 \in \mathcal{H}^*, \tag{9}$$

where $0 \in \mathcal{H}^*$ describes the functional that maps each element in $\mathcal{H}$ to zero, which is equivalent to requiring $\|E[\Psi(X, Z; \theta)]\|_{\mathcal{H}^*} = 0$. By the Riesz representation theorem (Zeidler, 2012), we can identify each element $\Psi \in \mathcal{H}^*$ with an element $\phi(\Psi) \in \mathcal{H}$ such that $\Psi(h) = \langle \phi(\Psi), h \rangle_{\mathcal{H}}$ $\forall h \in \mathcal{H}$ and $\|\Psi\|_{\mathcal{H}^*} = \|\phi(\Psi)\|_{\mathcal{H}}$. Generalizing the KMM estimator to conditional moment restrictions then just becomes a matter of substituting

$$h \in \mathbb{R}^m \to h \in \mathcal{H}$$
$$\psi(x;\theta)^T h \to \Psi(x, z; \theta)(h)$$

and adding a regularization term $-\frac{1}{2}\|h\|^2_{\mathcal{H}}$ for the Lagrange parameter $h \in \mathcal{H}$, which regularizes the first order conditions for h as argued by Kremer et al. (2022). With this at hand, we can define the functional version of the entropy-regularized MMD profile for conditional moment restrictions (refer to Section A.1 for details on the duality relationship).

Definition 3.3 (Functional KMM)**.** Let $\mathcal{H} \subseteq L^2(\mathcal{Z}, \mathbb{R}^m, P_Z)$ be a sufficiently rich Hilbert space of functions $\mathcal{Z} \to \mathbb{R}^m$ such that equivalence between (2) and (3) holds. Then the entropy-regularized MMD profile for conditional moment restrictions is given as

$$R^{\varphi}_{\epsilon,\lambda_n}(\theta) \tag{10}$$
$$= \sup_{\substack{\eta\in\mathbb{R}, f\in\mathcal{F}, \\ h\in\mathcal{H}}} \frac{1}{n}\sum_{i=1}^{n} f(x_i, z_i) + \eta - \frac{1}{2}\|f\|^2_{\mathcal{F}} - \frac{\lambda_n}{2}\|h\|^2_{\mathcal{H}}$$
$$- \epsilon \int_{\mathcal{X}\times\mathcal{Z}} \varphi^*\left(\frac{f(x,z) + \eta - \Psi(x,z;\theta)(h)}{\epsilon}\right) \omega(\mathrm{d}x \otimes \mathrm{d}z).$$

From the proof of Theorem 3.2 it directly follows that the optimization problem in (10) is jointly convex in the dual variables. The space $\mathcal{H}$ can be chosen, e.g., as the RKHS of a universal integrally strictly positive definite (ISPD) kernel (Simon-Gabriel & Schölkopf, 2018) to guarantee the consistency of the solution for the conditional moment restrictions (Kremer et al., 2022). In practice, alternative choices, e.g., neural networks which lack these theoretical guarantees have proven successful and often preferable (Bennett & Kallus, 2020b; Kremer et al., 2022).

3.5. Asymptotic Properties

Consider the regularized KMM estimator with any φ-divergence such that $\frac{d}{dt}\varphi^*(t)|_{t=0} =: \varphi_1^*(0) = 1$ and $\frac{d^2}{(dt)^2}\varphi^*(t)|_{t=0} =: \varphi_2^*(0) = 1$, which is fulfilled for, e.g., the forward and backward KL divergence.

For space reasons, here we focus on the estimator for *conditional* moment restrictions by combining the theory for the functional KMM estimator (see Section A.2) with a sufficiently rich space of locally Lipschitz functions $\mathcal{H}$. The properties of the unconditional/finite-dimensional KMM estimator are provided in Section B.

Theorem 3.4 (Consistency)**.** *Let $\mathcal{H} \subseteq L^2(\mathcal{Z}, \mathbb{R}^m, P_Z)$ be a Hilbert space of locally Lipschitz functions which is sufficiently rich such that equivalence between (2) and (3) holds. Further assume that a) $\theta_0 \in \Theta$ is the unique solution to $E[\psi(X;\theta)|Z] = 0$ P_Z-a.s.; b) $\Theta \subset \mathbb{R}^p$ is compact; c) $\psi(X;\theta)$ is continuous at each $\theta \in \Theta$ w.p.1; d) $E[\sup_{\theta\in\Theta} \|\psi(X;\theta)\|_2^2|Z] < \infty$ w.p.1; e) $V_0(Z) := E[\psi(X;\theta_0)\psi(X;\theta_0)|Z]$ is non-singular w.p.1; f) $\omega = (1-\alpha)\hat{P}_n + \alpha Q$ for $\alpha = O_p(n^{-1})$ and any distribution Q such that $E_Q[\sup_{\theta\in\Theta} \|\psi(X;\theta)\|_2^2|Z] < \infty$ w.p.1; and g) $\lambda_n = O_p(n^{-\xi})$ with $0 < \xi < 1/2$. Then for the KMM estimator $\hat{\theta}$ we have $\hat{\theta} \xrightarrow{p} \theta_0$.*

If additionally h) $\theta_0 \in \text{int}(\Theta)$; i) $\psi(x;\theta)$ is continuously differentiable in a neighborhood $\bar{\Theta}$ of θ_0 and $E[\sup_{\theta\in\bar{\Theta}} \|\nabla_\theta \psi(X;\theta)\|^2|Z] < \infty$ w.p.1; as well as j) $\text{rank}\,(E[\nabla_\theta \psi(X;\theta_0)|Z]) = p$ w.p.1, we have $\|\hat{\theta} - \theta_0\| = O_p(n^{-1/2})$.

Remark 3.5. Assumption f) implies that asymptotically the reference distribution ω is required to converge weakly to the population distribution P_0. However, as Q can be chosen as an arbitrary (continuous) distribution, as long as $\text{supp}(\hat{P}_n) \subseteq \text{supp}(Q)$, the form of $\omega = (1-\alpha)\hat{P}_n + \alpha Q$ does not restrict the set of candidate distributions P in (10) further than to distributions that admit a density w.r.t. Q.

Remark 3.6. A sufficiently rich function space for Theorem 3.4 is for example given by the RKHS of an integrally strictly positive definite universal kernel (e.g., Gaussian kernel; see Theorem 3.9 of Kremer et al. (2022)). Moreover, based on universal approximation theorems (Hornik et al., 1989), $\mathcal{H}$ can be represented by neural networks of asymptotically growing width/depth. In this case the local Lipschitz property can be ensured by restricting the weights to a compact domain (e.g., via weight clipping).

Theorem 3.7 (Asymptotic Normality)**.** *Let the assumptions of Theorem 3.4 be satisfied. Then,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Xi_0)$$

where

$$\Xi_0 = E\left[E[\nabla_\theta \psi(X;\theta_0)|Z]\, V_0^{-1}(Z)\, E[\nabla_\theta \psi(X;\theta_0)|Z]\right]^{-1}.$$

The asymptotic variance in Theorem 3.7 agrees with the semi-parametric efficiency bound of Chamberlain (1987). This implies that the KMM estimator achieves the lowest possible asymptotic variance among any estimators based on the CMR (2) as formalized in the following corollary.

Corollary 3.8 (Efficiency)**.** *Let the assumptions of Theorem 3.4 be satisfied. Then the KMM estimator $\hat{\theta}$ is efficient for θ_0, i.e., it has the smallest asymptotic variance among all estimators based solely on the conditional moment restrictions $E[\psi(X;\theta_0)|Z] = 0$ P_Z-a.s..*

This is a particularly strong result, setting our estimator apart from a number of recently proposed modern minimax approaches to CMR estimation (Lewis & Syrgkanis, 2018; Bennett et al., 2019; Zhang et al., 2021) and which is matched only by the kernel VMM estimator of Bennett & Kallus (2020b) and the more traditional sieve-based approaches of Ai & Chen (2003) and Chen & Pouzo (2009). Note that while this shows that our estimator is asymptotically first-order equivalent to these methods, the finite sample properties can be vastly different.

3.6. Computing KMM Estimators

In the following we restrict the discussion to the conditional version of the KMM estimator. The version for unconditional MR follows directly by setting $\lambda_n = 0$, $\mathcal{H} = \mathbb{R}^m$ and $\Psi(x, z; \theta) = \psi(x; \theta)$.

The functional entropy-regularized MMD profile (10) is a convex optimization problem over function-valued dual parameters $f \in \mathcal{F}$ and $h \in \mathcal{H}$ as well as $\eta \in \mathbb{R}$. The dual formulation (10) allows us to base our method on a dual functional gradient ascent algorithm. This is in contrast to particle gradient descent methods that address the primal problem by relying on discretizing measures which is commonly used in the gradient flow literature. Define the saddle point objective from (10) as

$$\begin{aligned}
&\widehat{G}_{\epsilon,\lambda_n}(\theta, \eta, f, h) \\
&= \frac{1}{n}\sum_{i=1}^{n} f(x_i, z_i) + \eta - \frac{1}{2}\|f\|_{\mathcal{F}}^2 - \frac{\lambda_n}{2}\|h\|_{\mathcal{H}}^2 \\
&- \epsilon \int_{\mathcal{X}\times\mathcal{Z}} \varphi^*\left(\frac{f(x,z) + \eta - \Psi(x,z;\theta)(h)}{\epsilon}\right) \omega(\mathrm{d}x \otimes \mathrm{d}z).
\end{aligned}$$

Then the KMM estimator is given as the solution to the problem

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \sup_{\beta\in\mathcal{M}} \widehat{G}_{\epsilon,\lambda_n}(\theta, \beta), \tag{11}$$

where we defined $\beta := (\eta, f, h) \in \mathcal{M} = \mathbb{R} \times \mathcal{F} \times \mathcal{H}$.

Stochastic Approximation In order to evaluate the KMM objective we use a stochastic approximation of the integral term in (11). Let the random variables (X, Z) and (X^ω, Z^ω) be distributed according to P_0 and ω respectively and define the random variable

$$\begin{aligned}
&G_{\epsilon,\lambda_n}(\theta, \beta; (X, Z), (X^\omega, Z^\omega)) \\
=&f(X, Z) + \eta - \frac{1}{2}\|f\|_{\mathcal{F}}^2 - \frac{\lambda_n}{2}\|h\|_{\mathcal{H}}^2 \\
&- \epsilon\varphi^*\left(\frac{f(X^\omega, Z^\omega) + \eta - \Psi(X^\omega, Z^\omega; \theta)(h)}{\epsilon}\right).
\end{aligned}$$

Then we can express the KMM objective as an expectation with respect to the empirical and reference distributions $\hat{P}_n$

Algorithm 1 Gradient Descent Ascent for KMM

Input: empirical distribution $\hat{P}_n$, reference distribution ω, hyperparameters ϵ, λ, batchsizes n_1, n_2
while not converged **do**
  Sample $\{(x_i, z_i)\}_{i=1}^{n_1} \sim \hat{P}_n$, $\{(x_j^\omega, z_j^\omega)\}_{j=1}^{n_2} \sim \omega$
  $G \leftarrow \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} G_{\epsilon,\lambda}(\theta, \beta; (x_i, z_i), (x_j^\omega, z_j^\omega))$
  $\beta \leftarrow \text{AscentStep}(\beta, \nabla_\beta G)$
  $\theta \leftarrow \text{DescentStep}(\theta, \nabla_\theta G)$
end while
Output: Parameter estimate θ

and ω respectively,

$$\widehat{G}_{\epsilon,\lambda_n}(\theta, \beta) = E_{\hat{P}_n} E_\omega [G_{\epsilon,\lambda_n}(\theta, \beta; (X, Z), (X^\omega, Z^\omega))].$$

This has the form required for mini-batch stochastic gradient descent (SGD) optimization. We solve problem (11) by alternating between functional SGD steps in the dual variables β and SGD steps in θ. Our approach is detailed in Algorithm 1.

Random Feature Approximation The supremum over β in (11) involves the optimization over the function space $\mathcal{F}$ which is generally intractable. To provide a scalable estimator that can be optimized with variants of stochastic gradient descent (SGD), we resort to a random Fourier feature approximation of the RKHS function f (Rahimi & Recht, 2007). For $\alpha \in \mathbb{R}^d$, let $f(x, z) = \alpha^T \phi(x, z)$ define the random feature approximation of f with random Fourier features $\phi(x, z) \in \mathbb{R}^d$, where d is the number of random features. The KMM objective then becomes

$$\begin{aligned}
&\widehat{G}_{\epsilon,\lambda_n}(\theta, \eta, \alpha, h) \\
&= \frac{1}{n} \sum_{i=1}^{n} \alpha^T \phi(x_i, z_i) + \eta - \frac{1}{2} \|\alpha\|_2^2 - \frac{\lambda_n}{2} \|h\|_{\mathcal{H}}^2 \\
&- \epsilon \int_{\mathcal{X} \times \mathcal{Z}} \varphi^* \left(\frac{\alpha^T \phi(x, z) + \eta - \Psi(x, z; \theta)(h)}{\epsilon} \right) \omega(\mathrm{d}x \otimes \mathrm{d}z).
\end{aligned}$$

Combined with the stochastic approximation and a scalable instrument function h (e.g., neural network or RKHS function with RF approximation) this allows our method to scale to large sample sizes.

4. Empirical Results

We benchmark our estimator on two different tasks against state-of-the-art estimators for conditional moment restrictions, including maximum moment restrictions (MMR) (Zhang et al., 2021), sieve minimum distance (SMD) (Ai & Chen, 2003), DeepIV (Hartford et al., 2017), DeepGMM (Bennett et al., 2019) and the neural network version of functional GEL (FGEL) (Kremer et al., 2022). As an additional baseline we compare to ordinary least squares (OLS) which ignores the conditioning variable and minimizes $\frac{1}{n} \sum_{i=1}^{n} \|\psi(x_i; \theta)\|_2^2$. For all methods involving kernels we use a radial basis function (RBF) kernel, whose bandwidth is set via the median heuristic (Garreau et al., 2018). The hyperparameters of all methods are set by evaluating the Hilbert-Schmidt independence criterion (HSIC) (Gretton et al., 2005) between the residues and the conditioning variable $\text{HSIC}(Y - g(T), Z)$ over a validation set of the size of the training set. HSIC has been proposed as an objective for CMR problems by Mooij et al. (2009) and Saengkyongam et al. (2022) and we empirically find it to yield better estimates than MMR (Muandet et al., 2020; Zhang et al., 2021) which has been used for as a validation metric in other works (see Section C.2). For the variational approaches we use a batchsize of $n_1 = 200$. Additionally, for our KMM esimator we represent the reference distribution ω by a kernel density estimator (KDE) trained on the empirical sample (see Section D.3) from which we sample mini-batches of size $n_2 = 200$. Refer to Section C for additional details. An implementation of our estimator and code to reproduce our results is available at https://github.com/HeinerKremer/conditional-moment-restrictions.

Heteroskedastic Instrumental Variable Regression We adopt the HeteroskedasticIV experiment of Bennett & Kallus (2020b). Let the data-generating process be given by

$$\begin{aligned}
&Z \sim \text{Uniform}([-5, 5]^2), \qquad H, \eta, \varepsilon \sim \mathcal{N}(0, 1) \\
&T_{\text{exo}} = Z_1 + |Z_2|, \qquad T_{\text{endo}} = 5H + 0.2\eta \\
&T = 0.75 T_{\text{exo}} + 0.25 T_{\text{endo}}, \quad S = 0.1 \log(1 + \exp(T_{\text{exo}})) \\
&Y = g(T; \theta_0) + 5H + S\varepsilon,
\end{aligned}$$

where the parameter of interest $\theta_0 = [2.0, 3.0, -0.5, 3.0] \in \mathbb{R}^4$ enters the process via the function

$$g(t; \theta) = \theta_2 + \theta_3 (t - \theta_1) + \frac{\theta_4 - \theta_3}{2} \log(1 + e^{2(t - \theta_1)}).$$

This task is particularly challenging as it involves heteroskedastic noise on the instruments. The true parameter θ_0 is identified by imposing the CMR $E[Y - g(T; \theta)|Z] = 0$ P_Z-a.s.. Table 1 shows the mean squared error (MSE) of the parameter estimate for different methods and sample sizes. Our method provides a significantly lower MSE for small sample sizes and approaches the results of DeepGMM and FGEL for larger samples.

Neural Network Instrumental Variable Regression To explore the viability of our estimator in the non-uniquely identified setting, we adopt the non-parametric instrumental variable regression experiment of Lewis & Syrgkanis (2018) which has also been used by Bennett et al. (2019), Zhang et al. (2021) and Kremer et al. (2022). Consider a data

Table 1. Instrumental Variable Regression with Heteroskedastic Instrument Noise. Mean of the parameter MSE $\|\theta - \theta_0\|^2$ and its standard error are computed over 20 random runs.

	OLS	MMR	DeepIV	DeepGMM	FGEL	KMM
n=500	1.78 ± 0.21	1.73 ± 0.22	2.57 ± 0.06	1.03 ± 0.17	1.02 ± 0.19	0.40 ± 0.13
n=1000	2.27 ± 0.18	1.97 ± 0.23	2.53 ± 0.08	0.70 ± 0.16	0.45 ± 0.11	0.16 ± 0.04
n=2000	1.79 ± 0.10	2.11 ± 0.20	2.43 ± 0.06	0.15 ± 0.04	0.14 ± 0.03	0.10 ± 0.02
n=4000	1.92 ± 0.06	1.65 ± 0.15	2.41 ± 0.04	0.07 ± 0.02	0.05 ± 0.01	0.07 ± 0.02
n=10000	1.99 ± 0.04	N/A	1.99 ± 0.04	0.02 ± 0.01	0.03 ± 0.01	0.01 ± 0.00

Table 2. Neural Network Instrumental Variable Regression. Mean of the prediction MSE $E[\|g_\theta(T) - g_0(T)\|^2]$ and its standard error are computed over 30 random runs and scaled by a factor of ten for ease of presentation.

	OLS	SMD	MMR	DeepIV	DeepGMM	FGEL	KMM
abs	3.21 ± 0.14	1.15 ± 0.53	1.41 ± 0.48	2.25 ± 0.68	0.42 ± 0.04	0.37 ± 0.05	0.32 ± 0.06
step	3.16 ± 0.05	0.54 ± 0.06	0.58 ± 0.03	0.74 ± 0.04	0.43 ± 0.04	0.40 ± 0.04	0.35 ± 0.02
sin	3.33 ± 0.06	1.31 ± 0.08	2.67 ± 0.13	3.75 ± 0.15	0.64 ± 0.05	0.62 ± 0.04	0.88 ± 0.10
linear	2.95 ± 0.08	0.47 ± 0.11	0.96 ± 0.20	1.66 ± 0.50	0.49 ± 0.05	0.95 ± 0.27	0.43 ± 0.11

generating process given by

$$y = g_0(t) + e + \delta, \qquad t = z + e + \gamma,$$
$$z \sim \text{Uniform}([-3, 3]),$$
$$e \sim N(0, 1), \qquad \gamma, \delta \sim N(0, 0.1),$$

where the function g_0 is chosen from

$$\text{sin: } g_0(t) = \sin(t), \quad \text{abs: } g_0(t) = |t|,$$
$$\text{linear: } g_0(t) = t, \quad \text{step: } g_0(t) = I_{\{t \geq 0\}}.$$

We try to learn an approximation of g_0 represented by a shallow neural network g_θ with 2 layers of $[20, 3]$ units and leaky ReLU activation functions. We identify g_θ by imposing the conditional moment restrictions $E[Y - g_\theta(T)|Z] = 0$ P_Z-a.s.. We use training and validation sets of size $n = 1000$ and evaluate the prediction error on a test set of size 20000. Table 2 shows the MSE of the predicted models trained with different CMR estimation methods. We observe that our estimator consistently shows competitive performance and slightly outperforms the baselines on three out of four tasks.

5. Related Work

Conditional moment restrictions have been addressed in multiple ways by extending the GMM to continua of moment restrictions building on the equivalence between the conditional (2) and continuum (3) formulations. Seminal work in this direction has been carried out by (Carrasco & Florens, 2000; Carrasco et al., 2007) and Ai & Chen (2003), which approximate the continuum of MR by a basis function expansion. Recently, the problem gained popularity in the machine learning community as many problems in causal inference can be formulated as CMR, most prominently instrumental variable regression. These modern approaches represent the continuum of MR via machine learning models, i.e., RKHS functions or neural networks and solve a mini-max formulation (Hartford et al., 2017; Lewis & Syrgkanis, 2018; Bennett et al., 2019; Bennett & Kallus, 2020b; Dikkala et al., 2020). Other GEL methods have historically played a less prominent role for CMR estimation, most likely due to their more complex mini-max structure compared to the simple minimization of traditional GMM-based methods. However, generalizations of GEL to continua of MR have been developed by Tripathi & Kitamura (2003); Kitamura et al. (2004); Chaussé (2012); Carrasco & Kotchoni (2017) building on basis function expansions, which empirically have been competitive with their GMM-counterparts. Recently the problem has been addressed via modern machine learning models (Kremer et al., 2022). All the aforementioned methods have in common that they either explicitly (GEL) or implicitly (GMM) optimize a φ-divergence between the candidate distributions and the empirical distribution and thus only allow for reweightings of the data. To the best of our knowledge, we provide the first method of moments estimator that lifts this restriction and allows for arbitrary candidate distributions.

The GEL framework bears a close duality relation to distributionally robust optimization (DRO) (Lam, 2019). In this context, it has been used to investigate the statistical properties of DRO (Duchi et al., 2018; Lam, 2019) and to calibrate the size of the distributional ambiguity set used in the DRO framework (Lam & Qian, 2017; Lam & Zhou, 2017; Blanchet et al., 2019; He & Lam, 2021). With the notable exception of Blanchet et al. (2019) these works build on the standard φ-divergence-based GEL framework. While Blanchet et al. (2019) provide a GEL framework based on optimal transport distances, their goal is to calibrate an am-

biguity set for DRO and they do not provide an estimator for moment restriction problems.

Computationally, an important contribution of this paper is handling the (semi)-infinite constraint in (5). Classical approaches to handling such constraints using polynomial sum-of-squares (SOS) (Lasserre, 2001) do not apply here since we have a general moment function class outside the polynomials. Furthermore, both classical and infinite-dimensional SOS techniques (Marteau-Ferey et al., 2020) suffer from scalability issues in high dimensions and large data sizes. Compared to those, our entropy-regularization approach can be implemented with general nonlinear problems and stochastic-gradient-type algorithms.

The objective of our inner optimization is a variational problem in the measure of the form $\min_P \{\mathrm{F}(P) + \epsilon \mathrm{H}(P, Q)\}$, where F is some energy functional and H is some metric or divergence measure. This was notably studied in the seminal work of Jordan et al. (1998) as a time-discretization scheme of PDEs. In recent literature related to machine learning, Arbel et al. (2019) studied the variational structure of MMD as energy in the Wasserstein geometry. Chizat (2022) applied noisy particle gradient descent to an energy objective similar to ours, i.e., $\mathrm{MMD} + \epsilon D_{\mathrm{KL}}$. Compared with that work, our goal is to train a model θ by minimizing this objective over θ. We also do not rely on gradient descent on the particles obtained from the discretization of the measure but adopt a dual functional gradient ascent scheme. Our reformulation technique for the MMD-profile is similar to that of Zhu et al. (2021), who solved a similar variational problem involving MMD for DRO. Different from their method, our goal is to provide an estimator for CMR problems and we introduce entropy regularization as an interior point method for handling the constraint.

6. Conclusion

The emergence of conditional moment restrictions in areas such as causal inference and robust machine learning has created the need for effective and robust estimation methods. Existing method of moments estimators (implicitly) rely on approximating the population distribution by reweighting a discrete empirical distribution. Our KMM estimator parts with this restrictive assumption and allows considering arbitrary (continuous) distributions as candidates for the population distribution. As in many cases the population distribution is in fact continuous, this has the potential to find more accurate estimates especially in the low sample regime where reweightings can provide crude approximations. Our estimator comes with strong theoretical guarantees showing that it is first order efficient with respect to any estimator based on CMR and its competitive practical performance is demonstrated on several CMR tasks.

This paper laid the foundation of the KMM framework, which can inspire future work in multiple ways. Such work could include the development of more sophisticated and adaptive reference measures for the regularization scheme, e.g., by evolving the reference measure over the course of the optimization. Another important direction would be a statistical learning theory analysis to provide theoretical properties of our estimator in the non-uniquely identified case. Other possibilities are extensions of the framework beyond estimation to construct confidence intervals for the estimates. Lastly, more efficient and tailored optimization methods can be developed to facilitate the application at larger scales.

References

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

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

Arbel, M., Korba, A., Salim, A., and Gretton, A. Maximum Mean Discrepancy Gradient Flow. *arXiv:1906.04370 [cs, stat]*, December 2019. arXiv: 1906.04370.

Bennett, A. and Kallus, N. Efficient policy learning from surrogate-loss classification reductions. In *International Conference on Machine Learning*, pp. 788–798. PMLR, 2020a.

Bennett, A. and Kallus, N. The variational method of moments, 2020b.

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

Bennett, A., Kallus, N., Li, L., and Mousavi, A. Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. In *International Conference on Artificial Intelligence and Statistics*, pp. 1999–2007. PMLR, 2021.

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

Bernstein, D. S. Matrix mathematics. In *Matrix Mathematics*. Princeton university press, 2009.

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. and Kotchoni, R. Regularized generalized empirical likelihood estimators. Technical report, Technical report, 2017.

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.

Chamberlain, G. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334, 1987. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(87)90015-7.

Chaussé, P. Generalized empirical likelihood for a continuum of moment conditions. 2012.

Chen, X. and Pouzo, D. Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*, 152(1):46–60, 2009.

Chen, Y., Xu, L., Gulcehre, C., Paine, T. L., Gretton, A., de Freitas, N., and Doucet, A. On instrumental variable regression for deep offline policy evaluation. *arXiv preprint arXiv:2105.10148*, 2021.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*, 2016.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters, 2018.

Chizat, L. Mean-Field Langevin Dynamics: Exponential Convergence and Annealing, August 2022. arXiv:2202.01009 [math].

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, 2018.

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., Glynn, P., and Namkoong, H. Statistics of robust optimization: A generalized empirical likelihood approach, 2018.

Garreau, D., Jitkrittum, W., and Kanagawa, M. Large sample analysis of the median heuristic, 2018.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic Learning Theory: 16th International Conference, ALT 2005, Singapore, October 8-11, 2005. Proceedings 16*, pp. 63–77. Springer, 2005.

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.

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.

He, S. and Lam, H. Higher-order expansion and bartlett correctability of distributionally robust optimization. *arXiv preprint arXiv:2108.05908*, 2021.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models, 2020.

Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Jordan, R., Kinderlehrer, D., and Otto, F. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998. Publisher: SIAM.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes, 2022.

Kitamura, Y., Tripathi, G., and Ahn, H. Empirical likelihood-based inference in conditional moment restriction models. *Econometrica*, 72(6):1667–1714, 2004. ISSN 00129682, 14680262.

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.

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

Lam, H. and Qian, H. Optimization-based quantification of simulation input uncertainty via empirical likelihood. *arXiv preprint arXiv:1707.05917*, 2017.

Lam, H. and Zhou, E. The empirical likelihood approach to quantifying uncertainty in sample average approximation. *Operations Research Letters*, 45(4):301–307, 2017.

Lasserre, J. B. Global optimization with polynomials and the problem of moments. *SIAM Journal on optimization*, 11(3):796–817, 2001.

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

Marteau-Ferey, U., Bach, F., and Rudi, A. Non-parametric models for non-negative functions. *Advances in neural information processing systems*, 33:12816–12826, 2020.

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

Mooij, J., Janzing, D., Peters, J., and Schölkopf, B. Regression by dependence minimization and its application to causal inference. pp. 94, 06 2009. doi: 10.1145/1553374.1553470.

Muandet, K., Jitkrittum, W., and Kübler, J. Kernel conditional moment test via maximum moment restriction, 2020.

Nesterov, Y. and Nemirovskii, A. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

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

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.

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.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680, 2021.

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

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In Platt, J., Koller, D., Singer, Y., and Roweis, S. (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Rosenblatt, M. Remarks on some nonparametric estimates of a density function. *The annals of mathematical statistics*, pp. 832–837, 1956.

Saengkyongam, S., Henckel, L., Pfister, N., and Peters, J. Exploiting independent instruments: Identification and distribution generalization, 2022.

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

Simon-Gabriel, C.-J. and Schölkopf, B. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 19(44):1–29, 2018. URL http://jmlr.org/papers/v19/16-291.html.

Smith, R. J. Alternative semi-parametric likelihood approaches to generalised method of moments estimation. *The Economic Journal*, 107(441):503–519, 1997.

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

Tripathi, G. and Kitamura, Y. Testing conditional moment restrictions. *The Annals of Statistics*, 31(6):2059–2095, 2003.

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.

Zeidler, E. *Applied functional analysis: applications to mathematical physics*, volume 108. Springer Science & Business Media, 2012.

Zhang, R., Imaizumi, M., Schölkopf, B., and Muandet, K. Maximum moment restriction for instrumental variable regression, 2021. arXiv 2010.07684.

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. KMM for Functional Moment Restrictions

A.1. Duality

The primal problem of the entropy regularized KMM estimator for functional moment restrictions is given by

$$R_\epsilon^\varphi(\theta) = \inf_{P \in \mathcal{P}} \frac{1}{2} \mathrm{MMD}(P, \hat{P}_n; \mathcal{F})^2 + \epsilon D_\varphi(P||\omega) \quad \text{s.t.} \quad \|E_P[\Psi(X, Z; \theta)]\|_{\mathcal{H}^*} \leq \lambda_n, \quad E_P[1] = 1, \tag{12}$$

where we relaxed the moment restrictions to hold only exactly for $n \to \infty$. Note that to be precise, the dual of (12), which can be obtained following the proof of Theorem 3.2, contains a regularization term $-\lambda_n \|h\|_{\mathcal{H}}$ instead of $-\frac{1}{2}\|h\|_{\mathcal{H}}^2$ as in our Definition 3.3. However, by Lagrangian duality the regularizer $-\lambda_n \|h\|_{\mathcal{H}}$ corresponds to restricting $\mathcal{H}$ to a norm ball of some radius ρ, and equally $-\frac{1}{2}\|h\|_{\mathcal{H}}^2$ corresponds to a restriction to a norm ball of different radius ρ'. Therefore both formulations are practically equivalent and we use the squared version for its greater smoothness and facilitated theoretical analysis. Note that in this context that a theoretical analysis of the $-\lambda_n \|h\|_{\mathcal{H}}$ version would be possible by resorting to the variational formulation of the norm $\|h\|_{\mathcal{H}} = \sup_{h' \in \mathcal{H}, \|h'\| \leq 1} \langle h', h \rangle_{\mathcal{H}}$. For an appropriate choice of reference distribution ω a solution to the KMM problem (12) at the true parameter $\theta_0 \in \Theta$ always exists as the true distribution P_0 is contained in an MMD ball around the empirical distribution with probability 1. This is in stark contrast to the φ-divergence based FGEL estimator of Kremer et al. (2022), as a φ-divergence ball around the empirical distribution $\hat{P}_n$ generally contains the corresponding continuous true distribution with probability 0, as the φ divergence between a discrete distribution $\hat{P}_n$ and continuous distribution P_0 diverges. However, at different parameters $\theta \in \Theta$ existence of a distribution $P \in \mathcal{P}$ for which the functional moment restrictions hold exactly cannot be guaranteed which implies $R(\theta) = \infty$ and thus gradient-based optimization over $\theta \in \Theta$ can become difficult. Therefore the role of the relaxation parameter λ_n here is to smooth the MMD profile such that $R(\theta) < \infty$ in a neighbourhood of the true parameter to facilitate gradient-based optimization over θ. Note that even for fixed values of λ_n, i.e., $\lambda_n = O_p(1)$, as $n \to \infty$ the objective has its global minimum of 0 at $P = P_0$ as $\hat{P}_n \xrightarrow{p} P_0$ and $\omega \xrightarrow{p} P_0$ weakly and thus we will retrieve the true solution θ_0. Therefore, compared to Kremer et al. (2022) where the relaxation scheme is a fundamental necessity to restore strong duality, here the regularization parameter can be seen merely as a computational tool.

A.2. Asymptotic Properties

For the KMM estimator for functional moment restrictions (FMR) of the form (9) based on the functional MMD profile (10) we have the following properties.

Theorem A.1 (Consistency for FMR)**.** *Assume that a) $\theta_0 \in \Theta$ is the unique solution to $E[\Psi(X, Z; \theta)] = 0 \in \mathcal{H}^*$; b) $\Theta \subset \mathbb{R}^p$ is compact; c) $\Psi(X, Z; \theta)$ is continuous in θ at any $\theta \in \Theta$ with probability one; d) $E[\sup_{\theta \in \Theta} \|\Psi(X, Z; \theta)\|_{\mathcal{H}^*}^2] < \infty$; e) $\Omega_0 = E[\Psi(X, Z; \theta_0) \otimes \Psi(X, Z; \theta_0)]$ is non-singular; f) $\omega = (1 - \alpha)\hat{P}_n + \alpha Q$ for $\alpha = O_p(n^{-1})$ and any distribution Q such that $E_Q[\sup_{\theta \in \Theta} \|\Psi(X, Z; \theta)\|_{\mathcal{H}^*}^2] < \infty$; and g) $\lambda_n = O_p(n^{-\xi})$ with $0 < \xi < 1/2$. Let $\hat{\theta}$ denote the functional KMM estimator for θ_0, then $\hat{\theta} \xrightarrow{p} \theta_0$.*

If additionally h) $\theta_0 \in \mathrm{int}(\Theta)$; i) $\Psi(x, z; \theta)$ is continuously differentiable in a neighborhood $\bar{\Theta}$ of θ_0 and $E[\sup_{\theta \in \bar{\Theta}} \|\nabla_\theta \Psi(X, Z; \theta)\|_{\mathcal{H}^*}^2] < \infty$; as well as j) $\Sigma_0 = \langle E[\nabla_\theta \Psi(X, Z; \theta_0)], E[\nabla_\theta \Psi(X, Z; \theta_0)] \rangle_{\mathcal{H}^*} \in \mathbb{R}^{p \times p}$ is non-singular, we have $\|\hat{\theta} - \theta_0\| = O_p(n^{-1/2})$.

Theorem A.2 (Asymptotic Normality for FMR)**.** *Let Assumptions a)-j) of Theorem A.1 be satisfied. Then,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Xi_0)$$

where $\Xi_0 = \left(E[\nabla_\theta \Psi(X, Z; \theta_0)] \Omega_0^{-1} E[\nabla_\theta \Psi(X, Z; \theta_0)]\right)^{-1}$.

B. Asymptotic Properties of the Finite-Dimensional KMM Estimator

For the KMM estimator for finite dimensional moment restrictions based on (8) we have the following results.

Theorem B.1 (Consistency for MR)**.** *Assume that a) $\theta_0 \in \Theta$ is the unique solution to $E[\psi(X; \theta)] = 0 \in \mathbb{R}^m$; b) $\Theta \subset \mathbb{R}^p$ is compact; c) $\psi(X; \theta)$ is continuous at each $\theta \in \Theta$ with probability one; d) $E[\sup_{\theta \in \Theta} \|\psi(X; \theta)\|_2^2] < \infty$; e) The covariance*

matrix $\Omega_0 := E\left[\psi(X,\theta_0)\psi(X;\theta_0)^T\right]$ *is non-singular; and f)* $\omega = (1-\alpha)\hat{P}_n + \alpha Q$ *for* $\alpha = O_p(n^{-1})$ *and any distribution* Q *such that* $E_Q[\sup_{\theta\in\Theta} \|\psi(X;\theta)\|_2^2] < \infty$. *Let* $\hat{\theta}$ *denote the KMM estimator for* θ_0, *then* $\hat{\theta} \xrightarrow{p} \theta_0$.

If additionally g) $\theta_0 \in \text{int}(\Theta)$; *h)* $\psi(x;\theta)$ *is continuously differentiable in a neighborhood* $\bar{\Theta}$ *of* θ_0 *and* $E[\sup_{\theta\in\bar{\Theta}} \|\nabla_\theta \psi(X;\theta)\|^2] < \infty$ *w.p.1 as well as i)* $\text{rank}\left(E[\nabla_\theta \psi(X;\theta_0)]\right) = p$, *we have* $\|\hat{\theta} - \theta_0\| = O_p(n^{-1/2})$.

Theorem B.2 (Asymptotic Normality for MR)**.** *Let Assumptions a)-i) of Theorem B.1 be satisfied. Then,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Xi_0)$$

where $\Xi_0 = \left(E[\nabla_\theta \psi(X;\theta_0)]\, \Omega_0^{-1}\, E[\nabla_\theta \psi(X;\theta_0)]\right)^{-1}$.

Remark B.3. The asymptotic variance Ξ_0 of the KMM estimator agrees with the one of the optimally weighted GMM estimator (Hansen, 1982), thus for finite dimensional moment restrictions KMM and OW-GMM are asymptotically first-order equivalent.

C. Additional Experimental Details

C.1. Hyperparameter choices

For the KMM estimator and the baselines we set the hyperparameters within the values described below using the setting with the minimal value for $\text{HSIC}(\psi(X;\theta), Z)$ evaluated on a validation set of the same size as the training set. As for large samples the HSIC computation becomes increasingly expensive we partition the validation data into batches of size $n_b = 2000$ and average HSIC over the batches.

For the variational methods we use an optimistic Adam (Daskalakis et al., 2018) implementation with a mini-batch size of $n = 200$ and a learning rate of $\tau_\theta = 5 \cdot 10^{-4}$ for optimization over θ and $\tau_h = 2.5 \cdot 10^{-3}$ for optimization over h and $\beta = (\eta, f, h)$ respectively. The regularization parameter λ for the instrument function $h \in \mathcal{H}$ is picked from $\lambda \in [0, 10^{-4}, 10^{-2}, 1]$.

Specific to FGEL we treat the divergence φ as a hyperparameter which we pick from $\varphi \in [\text{KL}, \log, \chi^2]$.

Specific to KMM we use $n_{\text{RF}} = 2000$ random Fourier features and for every batch of size $n_{\text{batch}} = 200$ sampled from $\hat{P}_n$ we attach $n_{\text{reference}} = 200$ samples from a reference distribution Q which we represent by a kernel density estimator with Gaussian kernel and bandwidth of $\sigma = 0.1$ trained on $\hat{P}_n$. We observed that the results are largely insensitive to the choice of bandwidth parameter σ. The entropy regularization parameter ϵ is picked from $\epsilon \in [0.1, 1, 10]$. The entropy regularizer is chosen as the Kullback-Leibler divergence as in the first part of Section 3.3. In agreement with the observations of Kremer et al. (2022) we noticed experimentally that the choice of φ-divergence has only a minor effect on the obtained estimator.

C.2. Choice of Validation Metric and Failure of MMR

The computation of modern CMR estimators including DeepGMM (Bennett et al., 2019), Functional GEL (Kremer et al., 2022) and our KMM estimator generally requires solving a mini-max or saddle point problem where the minimization is with respect to the model parameters and the maximization with respect to the instrument function h (and the RKHS function f in the case of KMM). For such problems it is not obvious how to monitor the success of the training procedure as for conditional moment restriction problems it is not clear which validation objective is supposed to be optimized, which makes tuning of hyperparameters and early stopping cumbersome and ambiguous. This is in contrast to standard supervised learning via empirical risk minimization where training and target objectives are usually aligned and thus one can simply evaluate the loss function over a suitable validation set. There exist different approaches to quantify how well the restrictions $E[\psi(X;\theta)|Z] = 0$ P_Z-a.s. are satisfied. The authors of Bennett et al. (2019) and Kremer et al. (2022) used the maximum moment restriction objective (Zhang et al., 2021) $\text{MMR}(\theta) = E_{\hat{P}_n}[\psi(X;\theta)K(Z,Z')\psi(X';\theta)]$ which results from the variational formulation $\text{MMR}(\theta) = \sup_{h\in\mathcal{H}} \left(E_{\hat{P}_n}[\psi(X;\theta)^T h(Z)]\right)^2$ with $\mathcal{H}$ corresponding to a unit ball of a reproducing kernel Hilbert space. While Zhang et al. (2021) show that this leads to a consistent estimator for θ_0 when optimized over $\theta \in \Theta$ and thus quantifies the satisfaction of the CMR in a meaningful way, it is a priori not clear if it provides a suitable validation metric in finite samples.

As an alternative Saengkyongam et al. (2022) proposed to measure the satisfaction of $E[\psi(X;\theta)|Z] = 0$ P_Z-a.s. by quantifying the independence of the random variables $\psi(X;\theta)$ and Z via the Hilbert-Schmidt independence criterion (HSIC) (Gretton

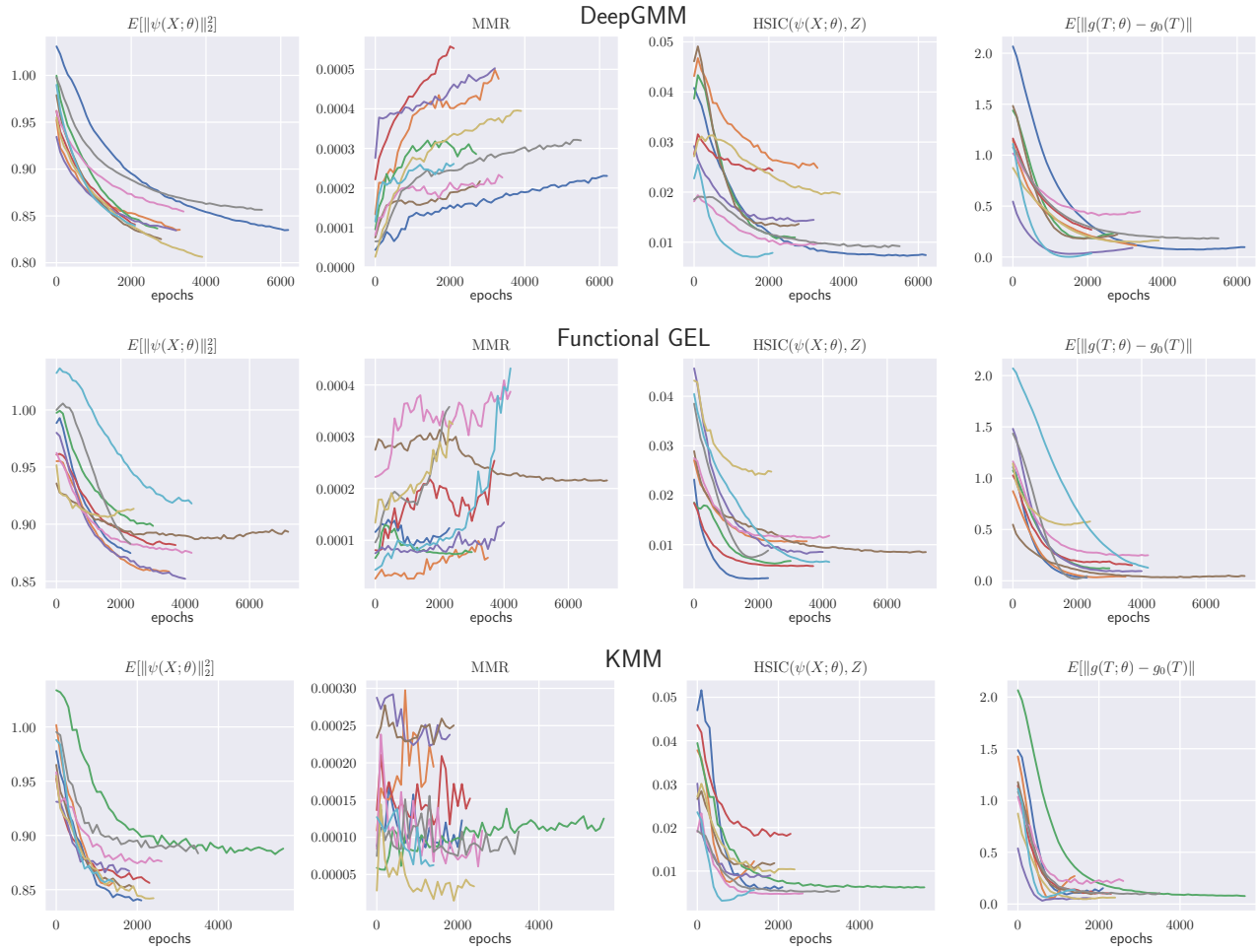
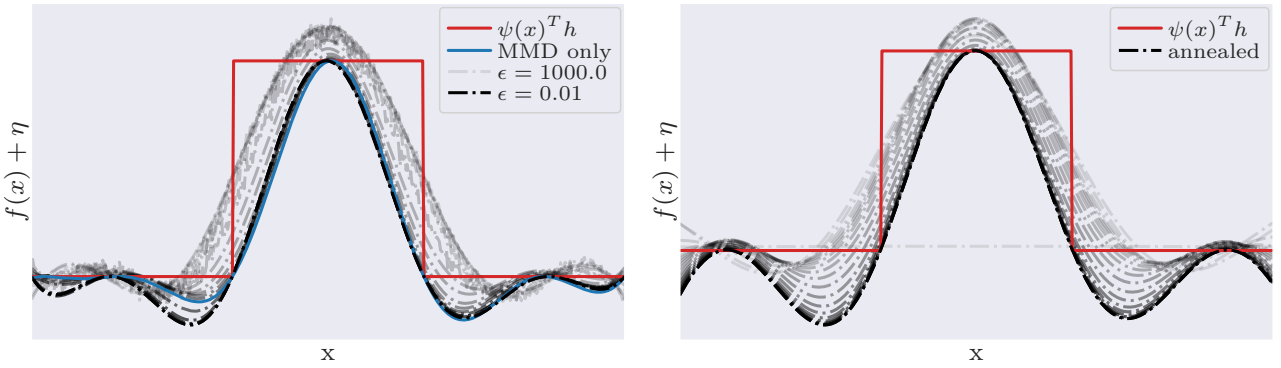


Figure 2. Effects of Validation Metrics for Early Stopping. Visualization of different validation losses for 10 training samples and different estimators. Goal of the estimation is to minimize the error with respect to the true function g_0 shown on the right which is unknown in practice. We observe that among the considered validation metrics, HSIC is the only one that approximately follows the behavior of the error with respect to the true function and thus allows for effective early stopping. The author's implementations of DeepGMM and FGEL use MMR as validation loss. Switching to HSIC allowed us to improve the performance of these baselines by a factor of 2-10.

et al., 2005).

We tested these two validation metrics for hyperparameter optimization and early stopping and observed that using HSIC instead of MMR as validation metric leads to improvements of the predictive MSE of the variational estimators (DeepGMM, FGEL, KMM) by a factor of 2-10, when keeping all other settings (i.e. hyperparameter grids) fixed.

We exemplarily visualize the effect of different validation metrics for early stopping for the heteroskedastic IV experiment in Figure 2. We train all estimators for 10 different random samples and use HSIC with a loose stopping criterion as validation metric in order to train beyond the optimal validation loss for visualization. The left column shows the prediction MSE of the learned function. While we aim to optimize this quantity, in practice we do not have access to it as the true function g_0 is unknown. The remaining columns show the different validation metrics over the course of the optimization. We observe that using the simplistic unconditional moment violation generally leads to overfitting as the estimator would be trained beyond the minimum of the true objective of interest. Interestingly, in most cases the MMR objective does not decrease over the course of the optimization procedure and thus any early stopping strategy based on it might stop the training at random. Of the three metrics, HSIC is the only one that approximately mimics the behavior of the true objective of interest and thus allows for an effective early stopping strategy to prevent overfitting and unnecessary long training.



(a) Regularization Parameters

(b) Annealing to Approximate the (Semi-)Infinite Constraint

Figure 3. Effect of Entropy Regularization. Figure a) shows the effect of entropy regularization for fixed parameters ϵ. The gray lines correspond to logarithmically decreasing values of ϵ between 1000 and 0.01. Figure b) shows the annealing procedure for entropy regularization, where the shaded curves show the intermediate progress of the optimization.

D. Entropy Regularization

D.1. Effect of the Regularization Parameter

As discussed in Section 3.2, for the case of the backward KL divergence or Burg entropy, our entropy regularization can be interpreted as a barrier function in an interior-point method, see (Nesterov & Nemirovskii, 1994). For decreasing values of ϵ, the *entropy-regularized MMD profile* approaches the unregularized *MMD-profile*. To validate this empirically we carry out the maximization over the dual parameters (η, f) in equation (8) while keeping h and θ fixed, which preserves the convex structure of the problem. In Figure 3(a) we observe that for smaller ϵ we get closer and closer to the original *MMD-profile*, which we obtain from equation (5) by using a sample approximation of the semi-infinite constraint in a convex solver.

D.2. Annealing of Entropy Regularization

Instead of keeping the regularization parameter ϵ fixed during the optimization procedure as in Figure 3(a), we study an annealing schedule in which it is gradually decreased, similar to actual interior-point methods. Chizat (2022) also studied effects of annealing in a setting where they use particle-gradient descent. While their work also builds on an energy functional consisting of a combination of MMD and KL-divergence, the dissipation is done in the Wasserstein geometry. In comparison, we do not carry out the optimization by moving in the Wasserstein space, but instead in the dual RKHS. To visualize the effect of annealing, we keep h and θ fixed and only maximize with respect to the remaining dual variables (η, f), while gradually decreasing ϵ with the number of iterations. We empirically observe that the annealing procedure eventually leads to a solution that satisfies the (semi-)infinite constraint (11). This is visualized in Figure 3(b) where the shaded black curves, corresponding to $f(x) + \eta$ at different iterations, are slowly pushed below the red curve in the course of the optimization.

D.3. Choice of Reference Measure

The KMM estimator based on (8) and (10) respectively requires a choice of distribution Q that enters the reference distribution $\omega = (1 - \alpha)\hat{P}_n + \alpha Q$ to define the entropy regularizer. As the candidate distributions are required to admit a density with respect to ω, the choice of Q directly determines the class of distributions considered in the minimization over P. Optimally Q should be chosen as close as possible to the population distribution P_0. As generally P_0 is unknown, in the following we discuss several (data-driven) choices for Q.

Lebesgue Measure Choosing Q as the Lebesgue measure or the uniform distribution over $\mathcal{X} \times \mathcal{Z}$ respectively, allows for considering arbitrary distributions on $\mathcal{X} \times \mathcal{Z}$. At the same time, this choice corresponds to an uninformative prior which discards the information contained in the sample and does not converge to the population distribution as $n \to \infty$. Empirically the Lebesgue measure did not show competitive performance.

Empirical Distribution The empirical distribution converges to the population distribution as $n \to \infty$, and therefore it provides a viable candidate as reference distribution. However, as the empirical distribution is a discrete distribution supported on the samples, the considered candidate distributions are again reweightings of the data and thus some of the competitive advantage of KMM over other method of moments estimators is lost. Note however, that MMD provides different gradient information compared to φ-divergences as used in GEL/GMM and therefore the obtained estimator will still be different.

Kernel Density Estimation In order to combine the strength of considering continuous candidate distributions in (8) with the information contained in the empirical distribution, one can represent the reference distribution Q by a kernel density estimator (KDE) (Rosenblatt, 1956) trained on the empirical sample. This allows us to sample from a continuous distribution in Algorithm 1 and thus taking into account candidate distributions with support different from the empirical distribution, while still converging to the population distribution as $n \to \infty$. Representing Q by a KDE proved to be the most effective choice in practice.

Modern Machine Learning Models As a straight-forward extension of the KDE approach, one can represent Q by any density estimator from which one can sample and can thus leverage the potential of modern machine learning approaches like generative adversarial networks (Goodfellow et al., 2014), variational auto-encoders (Kingma & Welling, 2022), normalizing flows (Papamakarios et al., 2021) or diffusion models (Ho et al., 2020). This seems particularly promising for complex high-dimensional data, where KDE estimators become increasingly inaccurate. Note however, that while better density estimators most likely improve the finite sample performance of our estimator, the role of Q is to define the class of (continuous) candidate distributions via its support. As long as $P_0 \ll Q$, we can find a P arbitrarily close to P_0 and better choices of Q (closer to P_0) mostly only facilitate finding these.

Time Evolution of Q via Primal-Dual Schemes Instead of using a fixed choice of reference distribution Q, one could choose the reference distribution adaptively over the course of the optimization via primal-dual schemes. To this aim, take $P^0 = \hat{P}_n$ and consider for timesteps $k = 1, \dots, T$ problem (7) as

$$R_\epsilon^\varphi(\theta) = \inf_{P \in \mathcal{P}} \frac{1}{2} \mathrm{MMD}(P, \hat{P}_n)^2 + \epsilon D(P||P_k) \quad \text{s.t.} \quad E_P[\psi(X;\theta)] = 0, \quad E_P[1] = 1. \tag{13}$$

Carrying out the optimization over $\mu \in \mathcal{F}$ and defining $\beta = (\eta, f, h) \in \mathcal{M}$, the Lagrangian of the MMD profile (7) can be cast in the (semi-dual) form,

$$L(P, \beta) = \frac{1}{n} \sum_{i=1}^n f(x_i) + \eta - \frac{1}{2}\|f\|_{\mathcal{H}}^2 + \int \left(-f(x) - \eta + \psi(x;\theta)^T h\right) dP(x) + \epsilon D_\varphi(P||P_k) \tag{14}$$

Now, instead of the dual approach used in our dual MMD profile, one could consider a primal dual update where one alternates between updates of the primal measure P via a proximal term using a Bregman divergence D

$$P_{k+1} \in \arg\min_P \int \left(-f(x) - \eta + \psi(x;\theta)^T h\right) dP(x) + \epsilon D(P\|P_k). \tag{15}$$

and subsequently update the dual variables β, where we solve

$$\beta_{k+1} \in \arg\max_{\beta \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^n f(x_i) + \eta - \frac{1}{2}\|f\|_{\mathcal{H}}^2 + \int \left(-f(x) - \eta + \psi(x;\theta)^T h\right) dP_{k+1}(x) + \frac{1}{2t}\|\beta - \beta_k\|_{\mathcal{M}}^2. \tag{16}$$

The update rule (16) is a standard convex optimization problem. We leave a specific implementation of this approach to future work.

E. Proofs

E.1. Definitions and Preliminaries

To simplify notation and analysis we first provide a compact formulation of the functional KMM objective (10) given by

$$\widehat{G}_{\epsilon,\lambda_n}(\theta,\eta,f,h) = \frac{1}{n}\sum_{i=1}^{n} f(x_i,z_i) + \eta - \frac{1}{2}\|f\|_{\mathcal{F}}^2 - \epsilon \int_{\mathcal{X}\times\mathcal{Z}} \varphi^*\left(\frac{f(x,z)+\eta-\langle\Psi(x,z;\theta),h\rangle_{\mathcal{H}}}{\epsilon}\right)\omega(\mathrm{d}x\otimes\mathrm{d}z) - \frac{\lambda_n}{2}\|h\|_{\mathcal{H}}^2.$$

As $\mathcal{F}$ is an RKHS of functions $\mathcal{X}\times\mathcal{Z}\to\mathbb{R}$, the evaluation functional in $\mathcal{F}$ is given by $k\left((x,z),\cdot\right):\mathcal{F}\to\mathbb{R}$, such that $\langle k((x,z),\cdot),f\rangle_{\mathcal{F}} = f(x,z)\ \forall f\in\mathcal{F}$ and $(x,z)\in\mathcal{X}\times\mathcal{Z}$. Let $\mathcal{M} := \mathbb{R}\times\mathcal{F}\times\mathcal{H}$. For $\beta=(\eta,f,h)\in\mathcal{M}$ define a norm on $\mathcal{M}$ as $\|\beta\|_{\mathcal{M}} = \sqrt{|\eta|^2+\|f\|_{\mathcal{F}}^2+\|h\|_{\mathcal{H}}^2}$. Define for $i=1,\dots,n$,

$$b_i = \begin{pmatrix}1\\ k((x_i,z_i),\cdot)\\ 0\end{pmatrix}\in\mathcal{M},\quad a(x,z;\theta) = \begin{pmatrix}1\\ k((x,z),\cdot)\\ -\Psi(x,z;\theta)\end{pmatrix}\in\mathcal{M},\quad R_\lambda = \begin{pmatrix}0 & & \\ & I & \\ & & \lambda I\end{pmatrix}\in\mathcal{M}\times\mathcal{M}. \tag{17}$$

where we used that we can identify $\mathcal{M}^*$ with $\mathcal{M}$ by the self-duality property of Hilbert spaces. Then the functional KMM objective (10) can be written in the compact form

$$G_{\epsilon,\lambda_n}(\theta,\beta) = \frac{1}{n}\sum_{i=1}^{n} b_i^T\beta - \epsilon\int_{\mathcal{X}\times\mathcal{Z}}\varphi^*\left(\frac{1}{\epsilon}a(x,z;\theta)^T\beta\right)\omega(\mathrm{d}x\otimes\mathrm{d}z) - \frac{1}{2}\beta^T R_{\lambda_n}\beta. \tag{18}$$

Analogously for the objective of the finite dimensional KMM estimator (8) we have $\mathcal{H}=\mathbb{R}^m$ and $\mathcal{M}=\mathbb{R}\times\mathcal{F}\times\mathbb{R}^m$ and further define for $i=1,\dots,n$,

$$b_i = \begin{pmatrix}1\\ k(x_i,\cdot)\\ 0\end{pmatrix}\in\mathcal{M},\quad a(x;\theta) = \begin{pmatrix}1\\ k(x,\cdot)\\ -\psi(x;\theta)\end{pmatrix}\in\mathcal{M},\quad R = \begin{pmatrix}0 & & \\ & I & \\ & & 0\end{pmatrix}\in\mathcal{M}\times\mathcal{M}. \tag{19}$$

Then the unconditional KMM objective (8) can be written in the compact form

$$\widehat{G}_{\epsilon}(\theta,\beta) = \frac{1}{n}\sum_{i=1}^{n} b_i^T\beta - \epsilon\int_{\mathcal{X}}\varphi^*\left(\frac{1}{\epsilon}a(x;\theta)^T\beta\right)\omega(\mathrm{d}x) - \frac{1}{2}\beta^T R\beta. \tag{20}$$

In the proofs we will consider derivatives of the KMM objective with respect to the dual parameters $\beta\in\mathcal{M}$, the second and the third component of which live in function spaces $\mathcal{F}$ and $\mathcal{H}$ respectively. We define the corresponding functional derivative as follows.

Definition E.1 (Functional Derivative)**.** Let $\mathcal{H}$ be a vector space of functions. For a functional $G:\mathcal{H}\to\mathbb{R}$ and a pair of functions $h_0,h_1\in\mathcal{H}$, we define the derivative operator $\frac{\partial}{\partial h}G(h_0)$ at h_0 via $\frac{\partial}{\partial h}G(h_0)(h_1) = \frac{d}{dt}G(h_0+th_1)\big|_{t=0}$. Likewise, we define the k-th functional derivative $\frac{\partial^k}{(\partial h)^k}G(h_0)$ at h_0 via

$$\frac{\partial^k}{(\partial h)^k}G(h_0)\left(h_1,\dots,h_k\right) = \left.\frac{\partial^k}{\partial t_1\dots\partial t_k}G\left(h_0+t_1h_1+\dots+t_kh_k\right)\right|_{t_1=\dots=t_k=0}. \tag{21}$$

Moreover, we write $\frac{\partial^k}{(\partial h)^k}G(h_0)=0$ as a shorthand for $\frac{\partial^k}{(\partial h)^k}G(h_0)\left(h_1,\dots,h_k\right)=0$ for all $h_1,\dots,h_k\in\mathcal{H}$. Similarly, when considering a vector-valued function of a vector-valued parameter, $G:\Theta\subseteq\mathbb{R}^p\to\mathbb{R}^m$, we denote the k-th standard directional derivative at $\theta_0\in\Theta$ as $\frac{\partial^k}{(\partial\theta)^k}G(\theta_0)\in\mathbb{R}^{p\times m}$ and in the case $k=1$ we write the Jacobian as $\frac{\partial}{(\partial\theta)}G(\theta_0) = \left(\nabla_\theta G\right)(\theta_0) =: \nabla_\theta G(\theta_0)$.

Additionally we will make use of the functional version of Taylor's theorem with Lagrange remainder, which we state here for completeness.

Proposition E.2 (Taylor's Theorem)**.** *Let $G : \mathcal{H} \to \mathbb{R}$, where $\mathcal{H}$ is a vector space of functions. For any $h, h' \in \mathcal{H}$, if $t \mapsto G\left(th + (1-t)h'\right)$ is $(k+1)$-times differentiable over an open interval containing $[0,1]$, then there exists $\bar{h} \in \operatorname{conv}\left(\{h, h'\}\right)$ such that*

$$G\left(h'\right) = G(h) + \sum_{i=1}^{k} \frac{1}{i!} \frac{\partial^i}{(\partial h)^i} G(h)(\underbrace{h' - h, \ldots, h' - h}_{i \text{ times}}) \tag{22}$$

$$+ \frac{1}{(k+1)!} \frac{\partial^{k+1}}{(\partial h)^{k+1}} G(\bar{h})(\underbrace{h' - h, \ldots, h' - h}_{k+1 \text{ times}}). \tag{23}$$

E.2. Duality Results

E.2.1. Proof of Theorem 3.1

Proof. Let $\mu_{\hat{P}_n} = E_{\hat{P}_n}[k(X,\cdot)] = \frac{1}{n}\sum_{i=1}^n k(x_i,\cdot)$ denote the kernel mean embedding of the empirical distribution. Instead of working with the measure P directly, we introduce an auxiliary variable $\mu \in \mathcal{F}$, which serves as the kernel mean embedding of P, so we can write the MMD profile as

$$R(\theta) = \inf_{P\in\mathcal{P},\mu\in\mathcal{H}} \frac{1}{2}\|\mu - \mu_{\hat{P}_n}\|_{\mathcal{F}}^2 \quad \text{s.t.} \quad \int k(x,\cdot)dP(x) = \mu, \quad \int dP(x) = 1, \quad \int \psi(x;\theta)dP(x) = 0. \tag{24}$$

Introducing Lagrange parameters $\eta \in \mathbb{R}$, $f \in \mathcal{F}$ and $h \in \mathbb{R}^m$ we can define the Lagrangian as

$$L(P,\mu,\eta,f,h) = \frac{1}{2}\|\mu - \mu_{\hat{P}_n}\|_{\mathcal{H}}^2 + \langle f, \mu - \int k(x,\cdot)dP(x)\rangle_{\mathcal{F}} + \eta\left(1 - \int dP(x)\right) + \langle h, \int \psi(x;\theta)dP(x)\rangle_{\mathbb{R}^m} \tag{25}$$

$$= \frac{1}{2}\|\mu - \mu_{\hat{P}_n}\|_{\mathcal{F}}^2 + \int \left(-f(x) - \eta + \langle h, \psi(x;\theta)\rangle_{\mathbb{R}^m}\right) dP(x) + \langle f,\mu\rangle_{\mathcal{F}} + \eta. \tag{26}$$

Now as we minimize the Lagrangian with respect to all positive measures P, this only yields a finite expression as long as $-f(x) - f_0 + \langle h, \psi(x;\theta)\rangle_{\mathbb{R}^m} \geq 0 \; \forall x \in \mathcal{X}$. This directly translates into a semi-infinite constraint and the problem becomes

$$\sup_{f_0,f,h} \inf_{\mu\in\mathcal{F}} \frac{1}{2}\|\mu - \mu_{\hat{P}_n}\|_{\mathcal{F}}^2 + \langle f,\mu\rangle_{\mathcal{F}} + f_0 \tag{27}$$

$$\text{s.t.} \quad \langle h, \psi(x;\theta)\rangle_{\mathbb{R}^m} \geq f(x) + f_0 \quad \forall x \in \mathcal{X}. \tag{28}$$

Now, the first order optimality conditions for μ yield (see subsection Optimality Condition in μ in the following for details)

$$\mu = \mu_{\hat{P}_n} - f, \tag{29}$$

and reinserting yields the final dual problem

$$\sup_{f_0,f,h} \frac{1}{n}\sum_{i=1}^n f(x_i) - \frac{1}{2}\|f\|_{\mathcal{H}}^2 + f_0 \tag{30}$$

$$\text{s.t.} \quad \langle h, \psi(x;\theta)\rangle_{\mathbb{R}^m} \geq f(x) + f_0 \quad \forall x \in \mathcal{X}. \tag{31}$$

Strong duality holds trivially as the primal problem only contains equality constraints. □

E.2.2. Proof of Theorem 3.2

Proof. Let again $\mu_{\hat{P}_n} = E_{\hat{P}_n}[k(X,\cdot)] = \frac{1}{n}\sum_{i=1}^n k(x_i,\cdot)$ denote the KME of the empirical distribution. Following the proof of Theorem 3.1 we can write the entropy regularized MMD profile as

$$R_\epsilon(\theta) = \inf_{P\ll\omega,\mu\in\mathcal{H}} \frac{1}{2}\|\mu - \mu_{\hat{P}_n}\|_{\mathcal{F}}^2 + \epsilon D_\varphi(P||\omega) \quad \text{s.t.} \quad \int k(x,\cdot)dP(x) = \mu, \quad \int dP(x) = 1, \quad \int \psi(x;\theta)dP(x) = 0.$$

Let $p(x)$ denote the density of P with respect to the reference measure ω. Introducing dual variables $\eta \in \mathbb{R}$, $f \in \mathcal{F}$ and $h \in \mathbb{R}^m$, the Lagrangian of the problem can be obtained as

$$L(p,\mu,f,h) = \frac{1}{2}\|\mu - \mu_{\hat{P}_n}\|_{\mathcal{F}}^2 + \epsilon\int_{\mathcal{X}} \varphi(p(x))\omega(\mathrm{d}x) + \langle f, \mu - \int_{\mathcal{X}} k(x,\cdot)p(x)\omega(\mathrm{d}x)\rangle_{\mathcal{F}} \tag{32}$$

$$+ \eta\left(1 - \int_{\mathcal{X}} p(x)\omega(\mathrm{d}x)\right) + \langle h, \int_{\mathcal{X}} \psi(x;\theta)p(x)\omega(\mathrm{d}x)\rangle_{\mathbb{R}^m}. \tag{33}$$

Now, collecting terms containing p we get

$$L(p,\mu,f,h) = \frac{1}{2}\|\mu - \mu_{\hat{P}_n}\|_{\mathcal{F}}^2 + \langle f,\mu\rangle_{\mathcal{F}} + \eta - \epsilon\int_{\mathcal{X}} \left(\frac{f(x) + \eta - \langle h, \psi(x;\theta)\rangle_{\mathbb{R}^m}}{\epsilon}p(x) - \varphi(p(x))\right)\omega(\mathrm{d}x). \tag{34}$$

The dual formulation follows from minimizing the Lagrangian with respect to the primal variables $\mu \in \mathcal{F}$ and $p \in \Pi(\omega)$, where $\Pi(\omega)$ denotes the set of all densities with respect to ω. Taking the infimum of L with respect to p we obtain

$$\inf_{p\in\Pi(\omega)} \frac{1}{2}\|\mu - \mu_{\hat{P}_n}\|_{\mathcal{F}}^2 + \langle f, \mu\rangle_{\mathcal{F}} + \eta - \epsilon \int_{\mathcal{X}} \left(\frac{f(x) + \eta - \langle h, \psi(x;\theta)\rangle_{\mathbb{R}^m}}{\epsilon} p(x) - \varphi(p(x)) \right) \omega(\mathrm{d}x) \tag{35}$$

$$= \frac{1}{2}\|\mu - \mu_{\hat{P}_n}\|_{\mathcal{F}}^2 + \langle f, \mu\rangle_{\mathcal{F}} + \eta - \epsilon \sup_{p\in\Pi(\omega)} \int_{\mathcal{X}} \left(\frac{f(x) + \eta - \langle h, \psi(x;\theta)\rangle_{\mathbb{R}^m}}{\epsilon} p(x) - \varphi(p(x)) \right) \omega(\mathrm{d}x) \tag{36}$$

$$= \frac{1}{2}\|\mu - \mu_{\hat{P}_n}\|_{\mathcal{F}}^2 + \langle f, \mu\rangle_{\mathcal{F}} + \eta - \epsilon \int_{\mathcal{X}} \sup_{t\in\mathbb{R}_+} \left(\frac{f(x) + \eta - \langle h, \psi(x;\theta)\rangle_{\mathbb{R}^m}}{\epsilon} t - \varphi(t) \right) \omega(\mathrm{d}x) \tag{37}$$

$$= \frac{1}{2}\|\mu - \mu_{\hat{P}_n}\|_{\mathcal{F}}^2 + \langle f, \mu\rangle_{\mathcal{F}} + \eta - \epsilon \int_{\mathcal{X}} \varphi^* \left(\frac{f(x) + \eta - \langle h, \psi(x;\theta)\rangle_{\mathbb{R}^m}}{\epsilon} \right) \omega(\mathrm{d}x), \tag{38}$$

where we used the definition of the Fenchel conjugate function $\varphi^*(q) = \sup_p \langle q, p\rangle - \varphi(p)$. In the third line we used that as $p : \mathcal{X} \to \mathbb{R}_+$ is an arbitrary function, we can swap the supremum outside the integral for a pointwise supremum over $t := p(x)$ for each $x \in \mathcal{X}$ in the integral. Now, the first order optimality conditions for the function μ yield (refer to the following subsection for details)

$$\mu = \mu_{\hat{P}_n} - f. \tag{39}$$

Inserting this back into the Lagrangian we get

$$L(f, \eta, h) = \frac{1}{n}\sum_{i=1}^n f(x_i) + \eta - \frac{1}{2}\|f\|_{\mathcal{H}}^2 - \epsilon \int_{\mathcal{X}} \varphi^* \left(\frac{f(x) + \eta - \langle h, \psi(x;\theta)\rangle_{\mathbb{R}^m}}{\epsilon} \right) \omega(\mathrm{d}x), \tag{40}$$

from which the dual program follows. Strong duality follows trivially as the primal problem only contains equality constraints. In order to show convexity, consider the compact notation (20),

$$G_\epsilon(\theta, \beta) = \frac{1}{n}\sum_{i=1}^n b_i^T \beta - \epsilon \int_{\mathcal{X}} \varphi^* \left(\frac{1}{\epsilon} a(x)^T \beta \right) \omega(\mathrm{d}x) - \frac{1}{2}\beta^T R \beta. \tag{41}$$

The first term is linear in β and thus trivially concave. The second term is concave as by definition the Fenchel conjugate of any function is convex (and thus its negative concave) and the composition of a concave function with a linear function yields a concave function. Finally the third term is a negative semi-definite quadratic form for any $\lambda_n \geq 0$ and thus concave. As an unconstrained maximization over a jointly concave objective the optimization over the dual parameters is a convex program. □

Optimality condition in μ It is easily verified that the functional

$$\frac{1}{2}\|\mu - \mu_{\hat{P}_n}\|_{\mathcal{F}}^2 + \langle f, \mu\rangle_{\mathcal{F}} \tag{42}$$

is (strongly) convex in μ. In fact, its minimizer can be seen by a straightforward manipulation of the terms

$$\frac{1}{2}\|\mu - \mu_{\hat{P}_n}\|_{\mathcal{F}}^2 + \langle f, \mu\rangle_{\mathcal{F}} = \frac{1}{2}\|\mu - \mu_{\hat{P}_n}\|_{\mathcal{F}}^2 + \langle f, \mu - \mu_{\hat{P}_n}\rangle_{\mathcal{F}} + \frac{1}{2}\|f\|_{\mathcal{F}}^2 + \langle f, \mu_{\hat{P}_n}\rangle_{\mathcal{F}} - \frac{1}{2}\|f\|_{\mathcal{F}}^2 \tag{43}$$

$$= \frac{1}{2}\|\mu - \mu_{\hat{P}_n} + f\|_{\mathcal{F}}^2 + \langle f, \mu_{\hat{P}_n}\rangle_{\mathcal{F}} - \frac{1}{2}\|f\|_{\mathcal{F}}^2 \tag{44}$$

$$\geq \langle f, \mu_{\hat{P}_n}\rangle_{\mathcal{F}} - \frac{1}{2}\|f\|_{\mathcal{F}}^2, \tag{45}$$

where the optimum is attained at $\mu = \mu_{\hat{P}_n} - f$.

Alternatively, we can also characterize the optimality condition via the differentiability structure. Since $\mathcal{F}$ is a normed space, we use $\nabla G(\mu)$ to denote the Fréchet derivative of a functional G. Suppose μ is a minimizer of the problem

$$\min_{\mu'} \left\{ G(\mu') := \langle f, \mu'\rangle_{\mathcal{F}} + \frac{1}{2}\|\mu' - \mu_{\hat{P}_n}\|_{\mathcal{F}}^2 \right\}. \tag{46}$$

Then $\nabla G(\mu) = 0$ and a straightforward calculation yields $\mu = \mu_{\hat{P}_n} - f$.

E.3. Asymptotic Properties of KMM for Conditional Moment Restrictions

The asymptotic properties of the KMM estimator for conditional moment restrictions follow from phrasing the conditional moment restrictions (2) as functional moment restrictions of the form (3) over a sufficiently rich Hilbert space of functions. In the following we show that the assumptions of Theorem 3.4 suffice to fulfill the assumptions of the theorems for the functional KMM estimator (Theorems A.1 and A.2) from which the results follow. The proofs for the functional case are deferred to Section E.4.

E.3.1. Proof of Theorem 3.4 (Consistency for CMR)

Lemma E.3. *For a moment function $\psi(x;\theta)$ taking values in $\mathbb{R}^m$ define the* conditional covariance matrix $V_0(Z) = E[\psi(X;\theta_0)\psi(X;\theta_0)^T|Z]$ *as a function of the conditioning random variable Z taking values in $\mathcal{Z}$. Let $\mathcal{H}$ be a Hilbert space of square integrable functions equipped with the norm $h \mapsto \|h\|_{L^2(\mathcal{H},P_0)} = \left(\int_{\mathcal{Z}} \|h(z)\|_2^2 \, P_0(\mathrm{d}z)\right)^{1/2}$. Define the moment functional $\Psi(x,z;\theta) : \mathcal{H} \to \mathbb{R}$ such that $\Psi(x,z;\theta)(h) = \psi(x;\theta)^T h(z)$ for any $(x,z) \in \mathcal{X} \times \mathcal{Z}$, $\theta \in \Theta$ and $h \in \mathcal{H}$. Then the covariance operator $\Omega_0 : \mathcal{H} \to \mathcal{H}$ defined as*

$$\Omega_0 = E[\Psi(X,Z;\theta_0) \otimes \Psi(X,Z;\theta_0)] \tag{47}$$

is non-singular if $V_0(Z)$ is non-singular with probability 1.

Proof. Note that Ω_0 is non-singular if $\|\Omega_0 h\|_{L^2(\mathcal{H},P_0)} > 0$ for any $h \in \mathcal{H}$ with $\|h\|_{L^2(\mathcal{H},P_0)} > 0$, or equivalently if $\langle h, \Omega_0 h\rangle \neq 0$. Consider any $h \in \mathcal{H}$ with $\|h\|_{L^2(\mathcal{H},P_0)} > 0$, then by the law of iterated expectation we have

$$\begin{aligned}
\langle h, \Omega_0 h\rangle_{\mathcal{H}} &= E\left[\langle \Psi(X,Z;\theta_0)(h), \Psi(X,Z;\theta_0)(h)\rangle\right] && (48)\\
&= E\left[\left(\psi(X;\theta_0)^T h(Z)\right)^T \left(\psi(X;\theta_0)^T h(Z)\right)\right] && (49)\\
&= E\left[h(Z)^T E[\psi(X;\theta_0)\psi(X;\theta_0)^T|Z] h(Z)\right] && (50)\\
&= \int h(z)^T V_0(z) h(z) \mathrm{d}P_0(z) && (51)
\end{aligned}$$

Now, $V_0(Z)$ is a positive-semi definite matrix by construction and non-singular P_0-a.s. by assumption and thus its smallest eigenvalue C is bounded away from zero. Therefore we have

$$\langle h, \Omega_0 h\rangle_{\mathcal{H}} \geq C \int \|h(z)\|_2^2 \mathrm{d}P_0(z) = C\|h\|^2_{L^2(\mathcal{H},P_0)} > 0 \tag{52}$$

and thus Ω_0 is non-singular with smallest eigenvalue bounded away from zero. □

Lemma E.4. *Let the assumptions of Theorem 3.4 be satisfied and define for any $(x,z,\theta) \in \mathcal{X} \times \mathcal{Z} \times \Theta$ the moment functional $\Psi(x,z;\theta) : \mathcal{H} \to \mathbb{R}$ with $\Psi(x,z;\theta)(h) = \psi(x;\theta)^T h(z)$. Then the matrix $\Sigma_0 = \langle E[\nabla_\theta \Psi(X,Z;\theta_0)], E[\nabla_\theta \Psi(X,Z;\theta_0)]\rangle_{\mathcal{H}^*} \in \mathbb{R}^{p\times p}$ is strictly positive definite and non-singular with smallest eigenvalue bounded away from zero.*

Proof. By definition we have $\mathcal{H} = \bigoplus_{i=1}^m \mathcal{H}_i$ and thus $\mathcal{H}^* = \bigoplus_{i=1}^m \mathcal{H}_i^*$. For each $i \in \{1,\dots,m\}$ let $\{h_j^i\}_{j=1}^\infty$ denote an orthonormal basis of $\mathcal{H}_i^*$ such that $\langle h_i^k, h_j^l\rangle = \delta_{ij}\delta_{kl}$. Then the identity operator in $\mathcal{H}^*$ can be expressed as $I_{\mathcal{H}^*} = \sum_{i=1}^m \sum_{j=1}^\infty h_j^i \left(h_j^i\right)^*$, where $\left(h_j^i\right)^* \in \mathcal{H}^{**}$ can be uniquely identified with an element in $\mathcal{H}$ by the property of Hilbert spaces. In the following, we overload notation and denote with h_j^i also the Riesz representer of $h_j^i \in \mathcal{H}^*$ in $\mathcal{H}$ which is

uniquely identified by the self-duality property of Hilbert spaces. Consider any $\theta \in \Theta$ with $0 < \|\theta\| < \infty$ then

$$\theta^T \Sigma_0 \theta = \left\langle E[\theta^T \nabla_\theta \Psi(X, Z; \theta_0)], E[\theta^T \nabla_\theta \Psi(X, Z; \theta_0)] \right\rangle_{\mathcal{H}^*} \tag{53}$$

$$= \left\langle E[\theta^T \nabla_\theta \Psi(X, Z; \theta_0)], \left(\sum_{i=1}^{m} \sum_{j=1}^{\infty} h_j^i \left(h_j^i\right)^* \right) E[\theta^T \nabla_\theta \Psi(X, Z; \theta_0)] \right\rangle_{\mathcal{H}^*} \tag{54}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{\infty} \left(E[\theta^T \nabla_\theta \psi_i(X; \theta_0) h_j^i(Z)] \right)^2 \tag{55}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{\infty} \left(E[\theta^T D_0^i(Z) h_j^i(Z)] \right)^2, \tag{56}$$

where $D_0^i(z) = E[\nabla_\theta \psi_i(X; \theta_0) | Z = z] \in \mathbb{R}^p$ denotes the columns of $D_0(z) = E[\nabla_\theta \psi(X; \theta_0) | Z = z] \in \mathbb{R}^{p \times m}$. Now as $\operatorname{rank}(D_0(Z)) = p$ w.p.1 by Assumption j), the p rows of $D_0(Z)$ are linearly independent w.p.1 which means that for any $\theta \in \Theta$ with $0 < \|\theta\| < \infty$ there exists $s \in \{1, \dots, m\}$ such that $\theta^T D_0^s(Z) \neq 0$ w.p.1. Now, by assumption the function space $\mathcal{H}$ is chosen such that we have equivalence between the conditional and variational/functional forms of the moment restrictions, i.e., for any continuous function ρ we have $E[\rho(X; \theta)^T h(Z)] = 0 \; \forall h \in \mathcal{H}$ if and only if $E[\rho(X; \theta) | Z] = 0$ w.p.1. In particular this implies $E[\theta^T \nabla_\theta \psi_s(X; \theta) h_s(Z)] = 0 \; \forall h_s \in \mathcal{H}_s$ if and only if $E[\theta^T \nabla_\theta \psi_s(X; \theta) | Z] = \theta^T D_0^s(Z) = 0$ w.p.1. As $\theta^T D_0^s(Z) \neq 0$ w.p.1 this means there must exist $h^s \in \mathcal{H}_s$ such that $E[\theta^T D_0^s(Z) h^s(Z)] \neq 0$. As we can expand any $h^s \in \mathcal{H}_s$ in terms of an orthonormal basis $\{h_k^s\}_{k=1}^{\infty}$ of $\mathcal{H}_s$ as $h = \sum_{k=1}^{\infty} \alpha_k h_k^s$, there must exist at least one $r \in \mathbb{N}$ with $\alpha_r \neq 0$ and $E[D_0^s(Z) h_r^s(Z)] \neq 0$. Inserting this back into (56) we get

$$\theta^T \Sigma_0 \theta = \sum_{i=1}^{m} \sum_{j=1}^{\infty} \left(E[\theta^T D_0^i(Z) h_j^i(Z)] \right)^2 \tag{57}$$

$$\geq \left(E[\theta^T D_0^s(Z) h_r^s(Z)] \right)^2 > 0. \tag{58}$$

From this it follows that Σ_0 is non-singular with probability 1. □

Proof of Theorem 3.4

Proof. By definition the function space $\mathcal{H}$ is expressive enough such that we can express the conditional moment restriction $E[\psi(X; \theta) | Z] = 0$ P_Z-a.s. in functional form as

$$E[\Psi(X, Z; \theta)] = 0 \in \mathcal{H}^*. \tag{59}$$

It remains to be shown that the assumptions imposed on ψ are sufficient for Ψ to fulfill the conditions of Theorem A.1. Assumptions a) and b) directly translate to the corresponding assumptions in Theorem A.1. Assumption c) of Theorem 3.4 follows directly from Assumption c) as $\Psi(X, Z; \theta)(h) = \psi(X; \theta)^T h(Z)$ is continuous in θ for any $\theta \in \Theta$ if $\psi(X; \theta)$ is continuous in θ for any $\theta \in \Theta$. As this holds for any $h \in \mathcal{H}$ continuity of $\Psi(X, Z; \theta)$ in θ follows. Assumption d) for Theorem A.1 follows as

$$E[\sup_{\theta \in \Theta} \|\Psi(X, Z; \theta)\|_{\mathcal{H}^*}^2] \tag{60}$$

$$= E[\sup_{\theta \in \Theta} \sup_{h \in \mathcal{H}, \|h\| \leq 1} \|\psi(X; \theta)^T h(Z)\|^2 \tag{61}$$

$$\leq E[E[\sup_{\theta \in \Theta} \|\psi(X; \theta)\|_2^2 \, | Z] \sup_{h \in \mathcal{H}, \|h\| \leq 1} \|h(Z)\|_2^2] \tag{62}$$

$$\leq C \int_{\mathcal{Z}} \sup_{h \in \mathcal{H}, \|h\| \leq 1} \|h(z)\|_2^2 \, P_0(\mathrm{d}z) \tag{63}$$

where we used that $E[\sup_{\theta \in \Theta} \|\psi(X; \theta)\|_2^2 | Z] \leq C$ with probability 1 by Assumption d). Now for any function $h \in \mathcal{H}$ with $\|h\|_{L^2(\mathcal{H}, P_0)} \leq 1$ we must have that $h(Z) < \infty$ w.p.1 and thus by the local Lipschitz property it follows $h(z) \leq M < \infty$

for any $z \in \text{supp}(P_0)$. As this holds for any $h \in \mathcal{H}$, in particular it also holds for the supremum over $\mathcal{H}$ and thus $\sup_{h\in\mathcal{H},\|h\|\leq 1} \|h(z)\|_2^2 \leq M \; \forall z \in \mathcal{Z}$. Therefore, we obtain

$$E[\sup_{\theta\in\Theta} \|\Psi(X,Z;\theta)\|_{\mathcal{H}^*}^2] \leq CM \int_{\mathcal{Z}} P_0(\mathrm{d}z) = CM < \infty. \tag{64}$$

Assumption e) of Theorem A.1 follows from Assumption e) and Lemma E.3. Assumption f) is identical to the corresponding Assumption f) in Theorem A.1 using the same argument as for Assumption d) for the integrability condition. Finally, Assumption g) of Theorem A.1 is identical to Assumption g). Therefore Assumptions a)-g) of Theorem A.1 are fulfilled and it follows that $\hat{\theta} \xrightarrow{p} \theta_0$.

Now further, Assumption h) of Theorem A.1 is identical with Assumption h). Assumption i) of Theorem A.1 follow from Assumption i) by the same argument presented earlier for Assumption c) and d) of Theorem A.1. Finally Assumption j) of Theorem A.1 follows from Assumption j) by Lemma E.4. Therefore, Assumptions h)-j) of Theorem A.1 are fulfilled and we have $\|\hat{\theta} - \theta_0\| = O_p(n^{-1/2})$. □

E.3.2. Proof of Theorem 3.7 (Asymptotic Normality for CMR)

The asymptotic normality of the KMM estimator for conditional moment restrictions follows directly from the result for functional moment restrictions Theorem A.2 using that by Theorem 3.4 the assumptions of Theorem 3.4 are sufficient to satisfy the assumptions of Theorem A.1. What remains to be shown is that we can translate the asymptotic covariance of the KMM estimator for functional moment restrictions into an expression containing the conditional quantities. To this aim, first, we show that we can express the asymptotic covariance of the KMM estimator for FMR in a variational form following Lemma 15 of Bennett & Kallus (2020b).

Lemma E.5. *Let the assumptions of Theorem A.1 be fulfilled. Then we have*

$$E[\nabla_\theta \Psi(X,Z;\theta_0)\Omega_0^{-1}\nabla_\theta \Psi(X,Z;\theta_0)] = \sup_{h\in\mathcal{H}} E[\nabla_\theta \psi(X;\theta_0)^T h(Z)] - \frac{1}{4}E\left[\left(\psi(X;\theta_0)^T h(Z)\right)^2\right] \tag{65}$$

Proof. By Lemma 14 of Bennett & Kallus (2020b) we have for any Hilbert space $\mathcal{H}$, and element $h \in \mathcal{H}$, that

$$\|h\|_{\mathcal{H}}^2 = \sup_{h'\in\mathcal{H}} \langle h, h'\rangle - \frac{1}{4}\|h'\|^2. \tag{66}$$

Moreover, as the dual space of $\mathcal{H}$, $\mathcal{H}^*$ is a Hilbert space itself, we can write for any $(x,z) \in \mathcal{X}\times\mathcal{Z}$,

$$\begin{aligned}
\nabla_\theta \Psi(x,z;\theta_0)\Omega_0^{-1}\nabla_\theta \Psi(x,z;\theta_0) =& \|\Omega_0^{-1/2}\nabla_\theta \Psi(x,z;\theta)\|_{\mathcal{H}^*}^2 && (67)\\
=& \sup_{h'\in\mathcal{H}^*} \langle \Omega_0^{-1/2}\nabla_\theta \Psi(x,z;\theta), h'\rangle_{\mathcal{H}^*} - \frac{1}{4}\|h'\|_{\mathcal{H}^*}^2 && (68)\\
=& \sup_{h'\in\mathcal{H}^*} \langle \nabla_\theta \Psi(x,z;\theta_0), \Omega_0^{-1/2}h'\rangle_{\mathcal{H}^*} - \frac{1}{4}\|h'\|_{\mathcal{H}^*}^2 && (69)\\
=& \sup_{h'\in\text{Range}(\Omega^{-1/2})} \langle \nabla_\theta \Psi(x,z;\theta_0), h'\rangle - \frac{1}{4}\langle \Omega^{1/2}h', \Omega^{1/2}h'\rangle_{\mathcal{H}^*} && (70)\\
=& \sup_{h'\in\mathcal{H}^*} \langle \nabla_\theta \Psi(x,z;\theta_0), h'\rangle - \frac{1}{4}\langle h', \Omega h'\rangle_{\mathcal{H}^*} && (71)\\
=& \sup_{h\in\mathcal{H}} \nabla_\theta \psi(x;\theta_0)^T h(z) - \frac{1}{4}h(z)^T\psi(x;\theta_0)\psi(x;\theta_0)^T h(z) && (72)
\end{aligned}$$

where we used that $\text{Range}(\Omega_0^{-1/2}) = \mathcal{H}^*$. This follows as Ω_0 is defined on all of $\mathcal{H}$ and invertible which immediately implies $\Omega_0^{1/2}$ is defined on all of $\mathcal{H}$ and invertible. This means that $\Omega_0^{1/2}$ is injective and thus $\text{Range}(\Omega_0^{-1/2}) = \mathcal{H}^*$. The result follows by taking the expectation over (x,z) on both sides. □

With the variational formulation at hand we can translate the expression of the covariance of the KMM estimator for FMR into an expression for CMR. The following result is a special case of Lemma 25 of Bennett & Kallus (2020b).

Lemma E.6. *Let the assumptions of Theorem A.1 be fulfilled. Then, if* $V_0(Z) = E[\psi(X;\theta_0)\psi(X;\theta_0)^T|Z]$ *is non-singular with probability* 1*, we have*

$$E[\nabla_\theta \Psi(X,Z;\theta_0)\Omega_0^{-1}\Psi(X,Z;\theta_0)] = E\left[E[\nabla_\theta\psi(X;\theta_0)|Z]\; V_0^{-1}(Z)\; E[\nabla_\theta\psi(X;\theta_0)|Z]\right]. \tag{73}$$

Proof. Using Lemma E.5 we can write

$$E[\nabla_\theta \Psi(X,Z;\theta_0)\Omega_0^{-1}\Psi(X,Z;\theta_0)] = \sup_{h\in\mathcal{H}} E[\nabla_\theta\psi(X;\theta_0)^T h(Z)] - \frac{1}{4}E\left[\left(\psi(X;\theta_0)^T h(Z)\right)^2\right] =: L(h) \tag{74}$$

The functional derivative of L at $h^* \in \mathcal{H}$ in direction $\epsilon \in \mathcal{H}$ is given by

$$\left(\frac{\partial}{\partial h}L(h^*)\right)(\epsilon) = E[\nabla_\theta\psi(X;\theta_0)^T\epsilon(Z)] - \frac{1}{2}E[\epsilon(Z)^T\psi(X;\theta_0)\psi(X;\theta_0)^T h^*(Z)] \tag{75}$$

$$= E\left[\epsilon(Z)^T\left(\nabla_\theta\psi(X;\theta_0) - \frac{1}{2}\psi(X;\theta_0)\psi(X;\theta_0)^T h^*(Z)\right)\right] \tag{76}$$

$$= E\left[\epsilon(Z)^T\left(E[\nabla_\theta\psi(X;\theta_0)|Z] - \frac{1}{2}V_0(Z)h^*(Z)\right)\right] \tag{77}$$

Now, by assumption $V_0(Z)$ is non-singular and thus invertible, moreover L is a concave functional in h and thus the global maximizer is given for any $z \in \mathcal{Z}$ by

$$h^*(z) = 2V_0(z)^{-1}E[\nabla_\theta\psi(X;\theta_0)|Z=z]. \tag{78}$$

Inserting back into equation (74) and denoting $D_0(z) := E[\nabla_\theta\psi(X;\theta_0)|Z=z]$ we have

$$E[\nabla_\theta \Psi(X,Z;\theta_0)\Omega_0^{-1}\Psi(X,Z;\theta_0)] = 2E[D_0(Z)^T V_0(Z)^{-1}D_0(Z)] - E[D_0(Z)^T V_0(Z)^{-1}V_0(Z)V_0(Z)^{-1}D_0(Z)] \tag{79}$$

$$= E[D_0(Z)^T V_0(Z)^{-1}D_0(Z)]. \tag{80}$$

□

Proof of Theorem 3.7

Proof. The conditions of Theorem A.2 are fulfilled by the conditions of Theorem 3.7 by the proof of Theorem 3.4. We can translate the expression for the asymptotic variance in terms of the moment functional into the conditional counterpart by applying Lemma E.6 whose conditions are fulfilled by Assumption e) of Theorem 3.4. □

E.3.3. Proof of Corollary 3.8 (Efficiency for CMR)

Proof. This is a direct implication of Theorem 3.7 as the asymptotic variance of the KMM estimator achieves the semi-parametric efficiency bound of Chamberlain (1987). □

E.4. Asymptotic Properties of KMM for Functional Moment Restrictions

The consistency proofs roughly follow the general idea laid out in the seminal paper by Newey & Smith (2004) with the adaption to functional moment restrictions by Kremer et al. (2022). The proof for the finite dimensional case is mostly a special case of the proof of the functional version. Therefore, we provide a detailed proof for the arguably more interesting functional case and a short version for the finite dimensional case, emphasizing the differences to the former.

E.4.1. Proof of Theorem A.1

Lemma E.7. *Let $\mathcal{A}$ denote a σ-algebra on $\mathcal{X} \times \mathcal{Z}$ and let $(\mathcal{X} \times \mathcal{Z}, \mathcal{A}, \omega)$ be a probability space with measure ω. For any functional $\Psi : (\mathcal{X} \times \mathcal{Z}) \times \Theta \times \mathcal{H} \to \mathbb{R}$ with $\int_{\mathcal{X}\times\mathcal{Z}} \sup_{\theta\in\Theta} \|\Psi(x,z;\theta)\|^2_{\mathcal{H}^*} \, \omega(\mathrm{d}x \otimes \mathrm{d}z) < \infty$, it follows that $\sup_{\theta\in\Theta} \|\Psi(X,Z;\theta)\|_{\mathcal{H}^*} \leq C$ ω-a.s. for some constant $C < \infty$.*

Proof. The proof is trivially implied by the definition of the almost surely property. If the event $\mathcal{E} = \{(x,z) \in \mathcal{X} \times \mathcal{Z} : \sup_{\theta\in\Theta} \|\Psi(x,z;\theta)\|_{\mathcal{H}^*} = \infty\}$ has non-zero measure, i.e., $\omega[\mathcal{E}] \neq 0$, then $\int_{\mathcal{X}\times\mathcal{Z}} \sup_{\theta\in\Theta} \|\Psi(x,z;\theta)\|_{\mathcal{H}^*} \, \omega(\mathrm{d}x \otimes \mathrm{d}z) = \infty$ and thus $\int_{\mathcal{X}\times\mathcal{Z}} \sup_{\theta\in\Theta} \|\Psi(x,z;\theta)\|^2_{\mathcal{H}^*} \, \omega(\mathrm{d}x \otimes \mathrm{d}z) = \infty$. Therefore we must have $\omega[\mathcal{E}] = 0$ and there exists some constant C such that $\sup_{\theta\in\Theta} \|\Psi(X,Z;\theta)\|_{\mathcal{H}^*} \leq C$ ω-a.s.. □

Lemma E.8. *For two distributions Q_1 and Q_2 on $\mathcal{X} \times \mathcal{Y}$ define the mixing distribution $\omega = (1-\alpha)Q_1 + \alpha Q_2$, with $\alpha = O_p(n^{-\zeta})$ and $\zeta > 0$. Then $\mathrm{MMD}(Q_1, \omega; \mathcal{F}) = O_p(n^{-\zeta})$ for any RKHS $\mathcal{F}$ of functions $\mathcal{X} \times \mathcal{Z} \to \mathbb{R}$. In particular it follows for any distribution Q and $\omega = (1-\alpha)\hat{P}_n + \alpha Q$ that $\mathrm{MMD}(\hat{P}_n, \omega; \mathcal{F}) = O_p(n^{-\zeta})$.*

Proof. The proof follows directly by using the definition of MMD,

$$\begin{aligned}
\mathrm{MMD}(Q_1, \omega; \mathcal{F}) =& \sup_{f\in\mathcal{F}, \|f\|_{\mathcal{F}}=1} \left(\int_{\mathcal{X}\times\mathcal{Z}} f(x,z) Q_1(\mathrm{d}x \otimes \mathrm{d}z) - \int_{\mathcal{X}\times\mathcal{Z}} f(x,z) \, \omega(\mathrm{d}x \otimes \mathrm{d}z) \right) && (81)\\
=& \alpha \sup_{f\in\mathcal{F}, \|f\|_{\mathcal{F}}=1} \left(\int_{\mathcal{X}\times\mathcal{Z}} f(x,z) Q_1(\mathrm{d}x \otimes \mathrm{d}z) - \int_{\mathcal{X}\times\mathcal{Z}} f(x,z) Q_2(\mathrm{d}x \otimes \mathrm{d}z) \right) && (82)\\
=& \alpha \, \mathrm{MMD}(Q_1, Q_2; \mathcal{F}) && (83)\\
=& \alpha C && (84)\\
=& O_p(n^{-\zeta}), && (85)
\end{aligned}$$

where we used that $\mathrm{MMD}(Q_1, Q_2; \mathcal{F})$ can be bounded by some positive constant C for any Q_1, Q_2 and $\mathcal{F}$ which directly follows from the fact that by definition of an RKHS the evaluation functional in $\mathcal{F}$ is bounded and Q_1, Q_2 are finite measures normalized to 1. The second statement is a direct application of the former. □

Lemma E.9. *Let the assumptions of Theorem A.1 be satisfied. For any ζ with $0 < \zeta < 1/2$ define the magnitude constrained set of dual variables $\mathcal{M}_n = \{\beta = (\eta, f, h) \in \mathcal{M} : \|\beta\|_{\mathcal{M}} \leq n^{-\zeta}\}$. Then as $n \to \infty$,*

$$\sup_{\theta\in\Theta, \ (\eta,f,h)\in\mathcal{M}_n} |\Psi(X,Z;\theta)(h)| = O_p(n^{-\zeta}) \ \omega\text{-a.s.}, \tag{86}$$

$$\sup_{\theta\in\Theta, \ \beta\in\mathcal{M}_n} |a(X,Z;\theta)^T \beta| = O_p(n^{-\zeta}) \ \omega\text{-a.s.}. \tag{87}$$

Proof. Using the Cauchy-Schwarz inequality and Assumption d) with Lemma E.7,

$$\begin{aligned}
& \sup_{\theta\in\Theta, \beta\in\mathcal{M}_n} |\Psi(X,Z;\theta)(h)| && (88)\\
\leq & \sup_{\theta\in\Theta, \beta\in\mathcal{M}_n} (\|h\|_{\mathcal{H}} \cdot \|\Psi(X,Z;\theta)\|_{\mathcal{H}^*}) && (89)\\
\leq & \sup_{\theta\in\Theta, \beta\in\mathcal{M}_n} (\|\beta\|_{\mathcal{M}} \cdot \|\Psi(X,Z;\theta)\|_{\mathcal{H}^*}) && (90)\\
\leq & n^{-\zeta} \sup_{\theta\in\Theta} \|\Psi(X,Z;\theta)\|_{\mathcal{H}^*}. && (91)
\end{aligned}$$

Now, by Assumptions d) and f) of Theorem A.1 and Lemma E.7 we have that $\sup_{\theta\in\Theta}\|\Psi(X,Z;\theta)\|_{\mathcal{H}^*} < C$ P_0-a.s. and $\sup_{\theta\in\Theta}\|\Psi(X,Z;\theta)\|_{\mathcal{H}^*} < C$ Q-a.s. respectively and as $\omega \to P_0$ weakly we have w.p.a.1 that $\sup_{\theta\in\Theta}\|\Psi(X,Z;\theta)\|_{\mathcal{H}^*} < C$ ω-a.s. and thus w.p.a.1 $\sup_{\theta\in\Theta,\beta\in\mathcal{M}_n}|\Psi(X,Z;\theta)(h)| = 0$ ω-a.s..

For the second part note that if $\|\beta\|_{\mathcal{M}} \leq n^{-\zeta}$, we must have that $|\eta|, \|f\|_{\mathcal{F}}, \|h\|_{\mathcal{H}} \leq n^{-\zeta}$. Then we have

$$
\begin{aligned}
&\sup_{\theta\in\Theta,\ \beta\in\mathcal{M}_n} |a(X,Z;\theta)^T\beta| && (92)\\
=&\sup_{\theta\in\Theta,\ (\eta,f,h)\in\mathcal{M}_n} |\eta + \langle k((X,Z),\cdot), f\rangle_{\mathcal{F}} + \Psi(X,Z;\theta)(h)| && (93)\\
\leq&\sup_{(\eta,f,h)\in\mathcal{M}_n} |\eta| + \sup_{(\eta,f,h)\in\mathcal{M}_n} |\langle k((X,Z),\cdot), f\rangle_{\mathcal{F}}| + \sup_{\theta\in\Theta,\beta\in\mathcal{M}_n} |\Psi(X,Z;\theta)(h)| && (94)\\
\leq&n^{-\zeta} + \sup_{(\eta,f,h)\in\mathcal{M}_n} \|f\|_{\mathcal{F}}\|k((X,Z),\cdot)\|_{\mathcal{F}} + \sup_{\theta\in\Theta,\beta\in\mathcal{M}_n} |\Psi(X,Z;\theta)(h)| && (95)\\
\leq&n^{-\zeta} + Cn^{-\zeta} + \sup_{\theta\in\Theta,\beta\in\mathcal{M}_n} |\Psi(X,Z;\theta)(h)| && (96)\\
\leq&\left(n^{-\zeta} + Cn^{-\zeta} + Cn^{-\zeta}\right)\ \omega\text{-a.s.} && (97)\\
\leq&O(n^{-\zeta})\ \omega\text{-a.s.}, && (98)
\end{aligned}
$$

where for the second term in the fourth line we applied the Cauchy-Schwarz inequality and used the fact that in an RKHS the evaluation functional is bounded by some constant $C > 0$. □

Lemma E.10. *Under the assumptions of Theorem A.1 we have for any $\theta \in \Theta$,*

$$\int_{\mathcal{X}\times\mathcal{Z}} a(x,z;\theta)\,\omega(\mathrm{d}x\otimes\mathrm{d}z) = \frac{1}{n}\sum_{i=1}^n a(x_i,z_i;\theta) + O_p(n^{-1}) \tag{99}$$

and for any $\beta \in \mathcal{M}$ with $\|\beta\|_{\mathcal{M}} < \infty$,

$$\int_{\mathcal{X}\times\mathcal{Z}} a(x,z;\theta)a(x,z;\theta)^T\beta\,\omega(\mathrm{d}x\otimes\mathrm{d}z) = \frac{1}{n}\sum_{i=1}^n a(x_i,z_i;\theta)a(x_i,z_i;\theta)^T\beta + O_p(n^{-1}). \tag{100}$$

Proof. For the first statement note that

$$
\begin{aligned}
\int_{\mathcal{X}\times\mathcal{Z}} a(x,z;\theta)\,\omega(\mathrm{d}x\otimes\mathrm{d}z) &= \int_{\mathcal{X}\times\mathcal{Z}} a(x,z;\theta)\,\omega(\mathrm{d}x\otimes\mathrm{d}z) + \int a(x,z;\theta)\mathrm{d}\hat{P}_n - \int a(x,z;\theta)\mathrm{d}\hat{P}_n && (101)\\
&= \frac{1}{n}\sum_{i=1}^n a(x_i,z_i;\theta) + \alpha\int a(x,z;\theta)(\mathrm{d}Q - \mathrm{d}\hat{P}_n) && (102)
\end{aligned}
$$

Now $a(x,z;\theta_0) = \left(1, k\left((x,z),\cdot\right), \Psi(x,z;\theta)\right)^T$ is trivially integrable in the first component and second component with respect to any probability distribution as the evaluation functional $k((x,z),\cdot)$ in $\mathcal{F}$ is bounded by definition of an RKHS. For the third component integrability with respect to Q follows by Assumption f) of Theorem A.1. Moreover, by Assumption d) of Theorem A.1 and Lemma E.7 we have $\sup_{\theta\in\Theta}\|\Psi(X,Z;\theta)\| \leq C$ w.p.1 with respect to P_0. And thus as $\hat{P}_n \xrightarrow{p} P_0$ weakly, we have $\int \sup_{\theta\in\Theta}\|\Psi(X,Z;\theta)\|\hat{P}_n(\mathrm{d}x\otimes\mathrm{d}z) < \infty$ w.p.a.1. In conclusion we have $\int a(x,z;\theta)(\mathrm{d}Q - \mathrm{d}\hat{P}_n) < \infty$ w.p.a.1 and as $\alpha = O_p(n^{-1})$ we finally get

$$\int_{\mathcal{X}\times\mathcal{Z}} a(x,z;\theta)\,\omega(\mathrm{d}x\otimes\mathrm{d}z) = \frac{1}{n}\sum_{i=1}^n a(x_i,z_i;\theta) + O_p(n^{-1}). \tag{103}$$

For the second statement consider any $\beta \in \mathcal{M}$ with $\|\beta\|_{\mathcal{M}} < \infty$,

$$
\begin{aligned}
\int_{\mathcal{X}\times\mathcal{Z}} a(x,z;\theta)a(x,z;\theta)^T\beta\,\omega(\mathrm{d}x\otimes\mathrm{d}z) =& \frac{1}{n}\sum_{i=1}^n a(x_i,z_i;\theta)a(x_i,z_i;\theta)^T\beta && (104)\\
&+ \alpha\int_{\mathcal{X}\times\mathcal{Z}} a(x,z;\theta)a(x,z;\theta)^T\beta\left(\mathrm{d}Q - \mathrm{d}\hat{P}_n\right). && (105)
\end{aligned}
$$

Now for the second term we have

$$\| \int_{\mathcal{X}\times\mathcal{Z}} a(x,z;\theta)a(x,z;\theta)^T\beta \left(\mathrm{d}Q - \mathrm{d}\hat{P}_n\right) \|^2_{\mathcal{M}^*} \tag{106}$$

$$\leq \int_{\mathcal{X}\times\mathcal{Z}} \|a(x,z;\theta)a(x,z;\theta)^T\beta\|^2 \left(\mathrm{d}Q + \mathrm{d}\hat{P}_n\right) \tag{107}$$

$$= \int_{\mathcal{X}\times\mathcal{Z}} |a(x,z;\theta)^T\beta|^2 \|a(x,z;\theta)\|^2 \left(\mathrm{d}Q + \mathrm{d}\hat{P}_n\right) \tag{108}$$

$$\leq \int_{\mathcal{X}\times\mathcal{Z}} |a(x,z;\theta)^T\beta|^2 \left(\mathrm{d}Q + \mathrm{d}\hat{P}_n\right) \int_{\mathcal{X}\times\mathcal{Z}} \|a(x,z;\theta)\|^2 \left(\mathrm{d}Q + \mathrm{d}\hat{P}_n\right) \tag{109}$$

$$\leq \|\beta\|^2_{\mathcal{M}} \left(\int_{\mathcal{X}\times\mathcal{Z}} \|a(x,z;\theta)\|^2 \left(\mathrm{d}Q + \mathrm{d}\hat{P}_n\right)\right)^2 \tag{110}$$

$$\leq \|\beta\|^2_{\mathcal{M}} \left(\int_{\mathcal{X}\times\mathcal{Z}} 1 + \|k((x,z),\cdot)\|^2_{\mathcal{F}} + \|\Psi(x,z;\theta)\|^2_{\mathcal{H}^*} \left(\mathrm{d}Q + \mathrm{d}\hat{P}_n\right)\right)^2. \tag{111}$$

The first term is trivially bounded, the second bounded as $\mathcal{F}$ is an RKHS and thus its evaluation functional $k((x,z),\cdot)$ is bounded. The third term is bounded as $E_Q[\sup_{\theta\in\Theta} \|\Psi(X,Z;\theta)\|^2_{\mathcal{H}^*} < \infty$ by Assumption f) of Theorem A.1 and $E_{\hat{P}_n}[\sup_{\theta\in\Theta} \|\Psi(X,Z;\theta)\|^2_{\mathcal{H}^*} < \infty$ w.p.a.1 as $E[\sup_{\theta\in\Theta} \|\Psi(X,Z;\theta)\|^2_{\mathcal{H}^*} < \infty$ by Assumption d) of Theorem A.1 and $\hat{P}_n \to P_0$ weakly. In conclusion the norm of the integral in equation (105) is bounded by some constant C and as $\alpha = O_p(n^{-1})$ the statement follows. □

Lemma E.11. *Let the assumptions of Theorem A.1 be satisfied and consider $\bar{\theta} \in \Theta$ such that $\bar{\theta} \xrightarrow{p} \theta_0$. Further let $\beta_\zeta := \arg\max_{\beta\in\mathcal{M}_n} \widehat{G}(\bar{\theta},\beta)$, where $\mathcal{M}_n = \{\beta \in \mathcal{M} : \|\beta\|_{\mathcal{M}} \leq n^{-\zeta}\}$ with $0 < \zeta < 1/2$. Define the operator $\Lambda_n(\beta,\theta) : \mathcal{M} \to \mathcal{M}$ as*

$$\Lambda_n(\beta,\theta) := \int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon} a(x,z;\theta)a(x,z;\theta)^T \varphi_2^*\left(\frac{1}{\epsilon} a(x,z;\theta)^T\beta\right) \omega(\mathrm{d}x\otimes\mathrm{d}z) + R_{\lambda_n}. \tag{112}$$

Then w.p.a.1 for any $\bar{\beta} \in \mathrm{conv}(\{0,\beta_\zeta\})$, $\Lambda_n(\bar{\beta},\bar{\theta})$ is strictly positive definite and its smallest eigenvalue is bounded away from zero. Moreover, for any $\theta \in \Theta$ the largest eigenvalue of $\Lambda_n(\bar{\beta},\theta)$ is bounded from above by a positive constant M.

Proof. As $\bar{\beta} \in \mathrm{conv}(\{0,\beta_\zeta\})$ we have $\bar{\beta} \in \mathcal{M}_n$, and hence Lemma E.9 implies that $\sup_{\theta\in\Theta} |a(X,Z;\theta)^T\bar{\beta}| \xrightarrow{n\to\infty} 0$ ω-a.s., which implies for every fixed value of $\epsilon > 0$, $\varphi_2^*\left(\frac{1}{\epsilon}a(X,Z;\theta)^T\bar{\beta}\right) \xrightarrow{n\to\infty} \varphi_2(0) = 1$ ω-a.s. by the continuous mapping theorem. This means that for every value of (x,z) that provides a non-vanishing contribution to the integral, we have $\varphi_2^*\left(\frac{1}{\epsilon}a(x,z;\theta)^T\bar{\beta}\right) \xrightarrow{n\to\infty} 1$ and so as $n \to \infty$ the first term is equivalent to $\int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon}a(x,z;\theta)a(x,z;\theta)^T\,\omega(\mathrm{d}x\otimes\mathrm{d}z)$ which clearly is a positive semi-definite operator. In the following we will show that its smallest eigenvalue is bounded away from zero w.p.a.1. First note that for any vector $\beta = (\eta,f,h) \in \mathcal{M}$ with $f \neq 0$ we have

$$\beta^T\Lambda_n\beta = \int_{\mathcal{X}\times\mathcal{Z}} \underbrace{(a(x,z;\theta)^T\beta)^2}_{\geq 0} \omega(\mathrm{d}x\otimes\mathrm{d}z) + \|f\|^2 + \lambda_n\|h\|^2 \tag{113}$$

$$\geq \|f\|^2 + \lambda_n\|h\|^2 \tag{114}$$

$$> 0, \tag{115}$$

and thus such vector cannot correspond to an eigenvalue of 0. Therefore consider any vector $\beta = (\eta,0,h) \in \mathcal{M}$, then if

such vector corresponds to an eigenvalue of zero we must have

$$\begin{pmatrix}0\\0\\0\end{pmatrix} =\Lambda_n(\bar{\beta},\bar{\theta}) \begin{pmatrix}\eta\\0\\h\end{pmatrix} \tag{116}$$

$$= \int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon} a(x,z;\bar{\theta})a(x,z;\bar{\theta})^T \begin{pmatrix}\eta\\0\\h\end{pmatrix} \omega(\mathrm{d}x\otimes\mathrm{d}z) + \begin{pmatrix}0&0&0\\0&I&0\\0&0&\lambda_n I\end{pmatrix}\begin{pmatrix}\eta\\0\\h\end{pmatrix} \tag{117}$$

$$= \int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon} \begin{pmatrix} \eta - \Psi(x,z;\bar{\theta})(h) \\ k((x,z),\cdot)\eta - k((x,z),\cdot)\Psi(x,z;\bar{\theta})(h) \\ -\eta\Psi(x,z;\bar{\theta}) + \Psi(x,z;\bar{\theta})\Psi(x,z;\bar{\theta})(h) + \lambda_n h \end{pmatrix} \omega(\mathrm{d}x\otimes\mathrm{d}z) \tag{118}$$

$$= \frac{1}{\epsilon}\frac{1}{n}\sum_{i=1}^n \begin{pmatrix} \eta - \Psi(x_i,z_i;\bar{\theta})(h) \\ k((x_i,z_i),\cdot)\eta - k((x_i,z_i),\cdot)\Psi(x_i,z_i;\bar{\theta})(h) \\ -\eta\Psi(x_i,z_i;\bar{\theta}) + \Psi(x_i,z_i;\bar{\theta})\Psi(x_i,z_i;\bar{\theta})(h) + \lambda_n h \end{pmatrix} + O_p(n^{-1}), \tag{119}$$

where we used Lemma E.10 to express the integral term in terms of the empirical average. Now the first row gives $\eta = E_{\hat{P}_n}[\Psi(X,Z;\bar{\theta})(h)] + O_p(n^{-1})$ which inserted in the last row gives

$$0 = \left(\underbrace{E_{\hat{P}_n}[\Psi(x_i,z_i;\bar{\theta})\otimes\Psi(x_i,z_i;\bar{\theta})] - E_{\hat{P}_n}[\Psi(x_i,z_i;\bar{\theta})]\otimes E_{\hat{P}_n}[\Psi(x_i,z_i;\bar{\theta})]}_{=:\hat{\Omega}(\bar{\theta})} + \lambda_n \right) h + O_p(n^{-1}). \tag{120}$$

As $n\to\infty$ we have $\bar{\theta}\to\theta_0$ and by the uniform weak law of large numbers and the continuous mapping theorem (as Ψ is continuous in θ by Assumption c) of Theorem A.1) $\hat{\Omega}(\bar{\theta}) \to E[\Psi(X,Z;\theta_0)\otimes\Psi(X,Z;\theta_0)] - E[\Psi(X,Z;\theta_0)]\otimes E[\Psi(X,Z;\theta_0)] = E[\Psi(X,Z;\theta_0)\otimes\Psi(X,Z;\theta_0)] = \Omega_0$. Thus as $n\to\infty$ we have

$$0 = (\Omega_0 + \lambda_n)\, h + O_p(n^{-1}) \tag{121}$$

From Assumption e) of Theorem A.1 it follows that Ω_0 is non-singular and thus 0 is not in its spectrum. Moreover, by Assumption g) of Theorem A.1 $\lambda_n = O_p(n^{-\xi})$ with $0<\xi<1/2$, so the RHS is $\neq 0$ w.p.a.1 and the eigenvalue equations can only be fulfilled with $h=0$ which implies $\eta=0$ and thus $\beta=0$. Therefore it follows that the smallest eigenvalue of $\Lambda_n(\bar{\beta},\bar{\theta})$ is bounded away from zero w.p.a.1.

In order to bound the largest eigenvalue of $\Lambda_n(\bar{\beta},\theta)$ for any $\theta\in\Theta$ recall that for the second term we have $\mathrm{eig}(R_{\lambda_n}) = \{0,1,\lambda_n\}$ where $\lambda_n\to 0$. Therefore, the boundedness depends on the eigenvalues of the first term. For any $\beta\in\mathcal{M}$ we have

$$\beta^T\Lambda_n\beta = \int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon}\beta^T a(x,z;\theta)a(x,z;\theta)^T\beta\, \omega(\mathrm{d}x\otimes\mathrm{d}z) + \beta^T R_{\lambda_n}\beta \tag{122}$$

$$\leq \int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon}\|a(x,z;\theta)^T\beta\|^2\, \omega(\mathrm{d}x\otimes\mathrm{d}z) + \|\beta\|^2 \tag{123}$$

$$= \int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon}\|\eta + \langle(k((x,z),\cdot),f\rangle_{\mathcal{F}} - \Psi(x,z;\theta)(h)\|^2\, \omega(\mathrm{d}x\otimes\mathrm{d}z) + \|\beta\|^2 \tag{124}$$

$$\leq \int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon}\left(\|\eta\|^2 + \|\langle(k((x,z),\cdot),f\rangle_{\mathcal{F}}\|^2 + \|\Psi(x,z;\theta)(h)\|^2\right)\omega(\mathrm{d}x\otimes\mathrm{d}z) + \|\beta\|^2 \tag{125}$$

$$\leq \int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon}\left(\|\eta\|^2 + \|f\|^2\|k((x,z),\cdot)\|^2 + \|h\|_{\mathcal{H}}^2\|\Psi(x,z;\theta)\|_{\mathcal{H}^*}^2\right)\omega(\mathrm{d}x\otimes\mathrm{d}z) + \|\beta\|^2. \tag{126}$$

Now, as $\mathcal{F}$ is an RKHS, the evaluation functional $k((x,z),\cdot)$ can be bounded by a constant C_1. Moreover, by Assumption d) and f) of Theorem A.1, we have $\int\sup_{\theta\in\Theta}\|\Psi(x,z;\theta)\|_{\mathcal{H}^*}^2\mathrm{d}P_0<\infty$ and $\int\sup_{\theta\in\Theta}\|\Psi(x,z;\theta)\|_{\mathcal{H}^*}^2\mathrm{d}Q<\infty$ and thus as $\omega = (1-\alpha)\hat{P}_n + \alpha Q \xrightarrow{p} (1-\alpha)P_0 + \alpha Q$ it follows $\sup_{\theta\in\Theta}\int\|\Psi(x,z;\theta)\|_{\mathcal{H}^*}^2\mathrm{d}\omega \leq \int\sup_{\theta\in\Theta}\|\Psi(x,z;\theta)\|_{\mathcal{H}^*}^2\mathrm{d}\omega < C_2$ for some $C_2>0$ w.p.a.1. Inserting this back we obtain

$$\beta^T\Lambda_{\epsilon,\lambda_n}\beta \leq \frac{1}{\epsilon}\left(\|\eta\|^2 + C_1\|f\|^2 + C_2\|h\|_{\mathcal{H}}^2\right) + \|\beta\|^2 \tag{127}$$

$$\leq \left(\frac{C_3}{\epsilon}+1\right)\|\beta\|^2, \tag{128}$$

where $C_3 = \max(1, C_1, C_2)$. It follows w.p.a.1 that the largest eigenvalue of Λ_n can be bounded by some constant $M = \frac{C_3}{\epsilon} + 1$ for any finite value of $\epsilon > 0$. □

Lemma E.12. *Let the assumptions of Theorem A.1 be satisfied. Additionally let* $\bar{\theta} \in \Theta$, $\bar{\theta} \xrightarrow{p} \theta_0$, *and* $\|E_{\hat{P}_n}[\Psi(X, Z; \bar{\theta})]\|_{\mathcal{H}^*} = O_p\left(n^{-1/2}\right)$. *Then for* $\bar{\beta} = \arg\max_{\beta \in \mathcal{M}} \widehat{G}_{\epsilon,\lambda_n}(\bar{\theta}, \beta)$ *we have* $\|\bar{\beta}\|_{\mathcal{M}} = O_p\left(n^{-1/2}\right)$, *and* $\widehat{G}_{\epsilon,\lambda_n}(\bar{\theta}, \bar{\beta}) \leq -\epsilon\varphi^*(0) + O_p\left(n^{-1}\right)$.

Proof. Define $\bar{\Psi}_i := \Psi(x_i, z_i; \bar{\theta})$ and $\bar{\Psi} = \frac{1}{n}\sum_{i=1}^n \bar{\Psi}_i$. For simplicity of notation let $\widehat{G}(\theta, \beta) := \widehat{G}_{\epsilon,\lambda_n}(\theta, \beta)$. The first and second derivative of $\widehat{G}(\bar{\theta}, \beta)$ with respect to β are given by

$$\frac{\partial \widehat{G}}{\partial \beta}(\bar{\theta}, \beta) = \frac{1}{n}\sum_{i=1}^n b_i - \int_{\mathcal{X}\times\mathcal{Z}} a(x, z; \bar{\theta})\varphi_1^*\left(\frac{1}{\epsilon}a(x, z; \bar{\theta})^T\beta\right)\omega(\mathrm{d}x \otimes \mathrm{d}z) - R_{\lambda_n}\beta \tag{129}$$

$$\frac{\partial^2 \widehat{G}}{(\partial \beta)^2}(\bar{\theta}, \beta) = -\int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon}a(x, z; \bar{\theta})a(x, z; \bar{\theta})^T\varphi_2^*\left(\frac{1}{\epsilon}a(x, z; \bar{\theta})^T\beta\right)\omega(\mathrm{d}x \otimes \mathrm{d}z) - R_{\lambda_n}. \tag{130}$$

Consider the optimal dual parameter within the magnitude constrained set $\mathcal{M}_n = \{\beta \in \mathcal{M} : \|\beta\|_{\mathcal{M}} \leq n^{-\zeta}\}$, i.e., $\beta_\zeta := \arg\max_{\beta \in \mathcal{M}_n} \widehat{G}(\bar{\theta}, \beta)$ with $\beta_\zeta = (\eta_\zeta, f_\zeta, h_\zeta)$. Later on, we will show that this maximizer can be identified with the maximizer over the original set $\mathcal{M}$. Using Taylor's theorem we can expand the empirical KMM objective about $\beta = 0$,

$$\widehat{G}(\bar{\theta}, \beta_\zeta) = \widehat{G}(\bar{\theta}, 0) + \frac{\partial \widehat{G}}{\partial \beta}(\bar{\theta}, 0)\beta_\zeta + \frac{1}{2}\beta_\zeta^T\frac{\partial^2 \widehat{G}}{(\partial \beta)^2}(\bar{\theta}, \dot{\beta})\beta_\zeta \tag{131}$$

$$= -\epsilon\varphi^*(0) + \int_{\mathcal{X}\times\mathcal{Z}} \Psi(x, z, \bar{\theta})(h_\zeta)\,\omega(\mathrm{d}x \otimes \mathrm{d}z) \tag{132}$$

$$+ \frac{1}{n}\sum_{i=1}^n f_\zeta(x_i, z_i) - \int_{\mathcal{X}\times\mathcal{Z}} f_\zeta(x, z)\,\omega(\mathrm{d}x \otimes \mathrm{d}z) \tag{133}$$

$$- \frac{1}{2}\beta_\zeta^T \underbrace{\left(\int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon}a(x, z; \bar{\theta})a(x, z; \bar{\theta})^T\varphi_2^*\left(\frac{1}{\epsilon}a(x, z; \bar{\theta})^T\dot{\beta}\right)\omega(\mathrm{d}x \otimes \mathrm{d}z) + R_{\lambda_n}\right)}_{:=\Lambda_n(\dot{\beta}, \bar{\theta})}\beta_\zeta \tag{134}$$

for some $\dot{\beta} \in \operatorname{conv}(\{0, \beta_\zeta\})$. Now adding and subtracting the empirical expectation of the moment functional Ψ we get

$$\begin{aligned}
\widehat{G}(\bar{\theta}, \beta_\zeta) =& -\epsilon\varphi^*(0) - \frac{1}{2}\beta_\zeta^T \Lambda_n(\dot{\beta}, \bar{\theta})\beta_\zeta + \int_{\mathcal{X}\times\mathcal{Z}} \Psi(x, z; \bar{\theta})(h_\zeta)\hat{P}_n(\mathrm{d}x \otimes \mathrm{d}z) && (135)\\
&+ \int_{\mathcal{X}\times\mathcal{Z}} \Psi(x, z; \bar{\theta})(h_\zeta)\, \omega(\mathrm{d}x \otimes \mathrm{d}z) - \int_{\mathcal{X}\times\mathcal{Z}} \Psi(x, z; \bar{\theta})(h_\zeta)\hat{P}_n(\mathrm{d}x \otimes \mathrm{d}z) && (136)\\
&+ \int_{\mathcal{X}\times\mathcal{Z}} f_\zeta(x, z)\hat{P}_n(\mathrm{d}x \otimes \mathrm{d}z) - \int_{\mathcal{X}\times\mathcal{Z}} f_\zeta(x, z)\, \omega(\mathrm{d}x \otimes \mathrm{d}z) && (137)\\
\leq& -\epsilon\varphi^*(0) - \frac{1}{2}\beta_\zeta^T \Lambda_n(\dot{\beta}, \bar{\theta})\beta_\zeta + \|h_\zeta\|_{\mathcal{H}} \left\| \int_{\mathcal{X}\times\mathcal{Z}} \Psi(x, z; \bar{\theta})\hat{P}_n(\mathrm{d}x \otimes \mathrm{d}z) \right\|_{\mathcal{H}^*} && (138)\\
&+ \|h_\zeta\|_{\mathcal{H}} \left\| \int_{\mathcal{X}\times\mathcal{Z}} \Psi(x, z; \bar{\theta}) \left(\hat{P}_n(\mathrm{d}x \otimes \mathrm{d}z) - \omega(\mathrm{d}x \otimes \mathrm{d}z)\right) \right\|_{\mathcal{H}^*} && (139)\\
&+ \|f_\zeta\|_{\mathcal{F}} \sup_{f\in\mathcal{F}, \|f\|_{\mathcal{F}}=1} \int_{\mathcal{X}\times\mathcal{Z}} f(x, z) \left(\hat{P}_n(\mathrm{d}x \otimes \mathrm{d}z) - \omega(\mathrm{d}x \otimes \mathrm{d}z)\right) && (140)\\
\leq& -\epsilon\varphi^*(0) - \frac{1}{2}\beta_\zeta^T \Lambda_n(\dot{\beta}, \bar{\theta})\beta_\zeta + \|h_\zeta\|_{\mathcal{H}} \|\bar{\Psi}\|_{\mathcal{H}^*} && (141)\\
&+ \alpha\|h_\zeta\|_{\mathcal{H}} \left(\int_{\mathcal{X}\times\mathcal{Z}} \left\|\Psi(x, z; \bar{\theta})\right\|_{\mathcal{H}^*} Q(\mathrm{d}x \otimes \mathrm{d}z) + \|\bar{\Psi}\|_{\mathcal{H}^*} \right) && (142)\\
&+ \|f_\zeta\|_{\mathcal{F}} \sup_{f\in\mathcal{F}, \|f\|_{\mathcal{F}}=1} \int_{\mathcal{X}\times\mathcal{Z}} f(x, z) \left(\hat{P}_n(\mathrm{d}x \otimes \mathrm{d}z) - \omega(\mathrm{d}x \otimes \mathrm{d}z)\right) && (143)\\
\leq& -\epsilon\varphi^*(0) - \frac{1}{2}\beta_\zeta^T \Lambda_n(\dot{\beta}, \bar{\theta})\beta_\zeta + \|\bar{\Psi}\|_{\mathcal{H}^*} \|h_\zeta\|_{\mathcal{H}} && (144)\\
&+ \alpha\|h_\zeta\|_{\mathcal{H}} \left(C_Q + \|\bar{\Psi}\|_{\mathcal{H}^*}\right) + \|f_\zeta\|_{\mathcal{F}} \operatorname{MMD}(\hat{P}_n, \omega; \mathcal{F}) && (145)\\
\leq& -\epsilon\varphi^*(0) - \frac{1}{2}\beta_\zeta^T \Lambda_n(\dot{\beta}, \bar{\theta})\beta_\zeta + \|\beta_\zeta\|_{\mathcal{M}} \left(\|\bar{\Psi}\|_{\mathcal{H}} + \alpha(C_Q + \|\bar{\Psi}\|_{\mathcal{H}^*}) + \operatorname{MMD}(\hat{P}_n, \omega; \mathcal{F}) \right) && (146)
\end{aligned}$$

where we repeatedly used the Cauchy-Schwarz inequality and the fact that $\int_{\mathcal{X}\times\mathcal{Z}} \left\|\Psi(x, z; \bar{\theta})\right\|_{\mathcal{H}^*} Q(\mathrm{d}x \otimes \mathrm{d}z) < \int_{\mathcal{X}\times\mathcal{Z}} \sup_{\theta\in\Theta} \|\Psi(x, z; \theta)\|_{\mathcal{H}^*} Q(\mathrm{d}x \otimes \mathrm{d}z) =: C_Q < \infty$ by Assumption f) of Theorem A.1 and Lemma E.7.

Lemma E.11 states that the smallest eigenvalue C of $\Lambda_n(\dot{\beta}, \bar{\theta})$ is bounded away from zero w.p.a.1. As β_ζ is a global maximizer of $\widehat{G}(\bar{\theta}, \beta)$ over $\mathcal{M}_n$ we have that $\widehat{G}(\bar{\theta}, \beta_\zeta) \geq \widehat{G}(\bar{\theta}, \beta)$ for any $\beta \in \mathcal{M}_n$ and therefore,

$$\begin{aligned}
-\epsilon\varphi^*(0) &= \widehat{G}(\bar{\theta}, 0) && (147)\\
&\leq \widehat{G}(\bar{\theta}, \beta_\zeta) && (148)\\
&\leq -\epsilon\varphi^*(0) - \frac{1}{2}\beta_\zeta^T \Lambda_n(\dot{\beta}, \bar{\theta})\beta_\zeta + \|\beta_\zeta\|_{\mathcal{M}} \left(\|\bar{\Psi}\|_{\mathcal{H}} + \alpha(C + \|\bar{\Psi}\|_{\mathcal{H}^*}) + \operatorname{MMD}(\hat{P}_n, \omega; \mathcal{F}) \right) && (149)\\
&\leq -\epsilon\varphi^*(0) - C\|\beta_\zeta\|_{\mathcal{M}}^2 + \|\beta_\zeta\|_{\mathcal{M}} \left(\|\bar{\Psi}\|_{\mathcal{H}} + \alpha(C + \|\bar{\Psi}\|_{\mathcal{H}^*}) + \operatorname{MMD}(\hat{P}_n, \omega; \mathcal{F}) \right) && (150)
\end{aligned}$$

Now, adding $-\epsilon\varphi^*(0)$ on both sides and dividing by $\|\beta_\zeta\|_{\mathcal{M}}$, we have

$$C\|\beta_\zeta\|_{\mathcal{M}} \leq \|\bar{\Psi}\|_{\mathcal{H}^*} + \alpha(C + \|\bar{\Psi}\|_{\mathcal{H}^*}) + \operatorname{MMD}(\hat{P}_n, \omega; \mathcal{F}). \tag{151}$$

As $\|\bar{\Psi}\|_{\mathcal{H}^*} = O_p(n^{-1/2})$ by assumption and by Assumption f) $\alpha = O_p(n^{-1})$ as well as $\operatorname{MMD}(\hat{P}_n, \omega; \mathcal{F}) = O(n^{-1}) = o_p(n^{-1/2})$ by Lemma E.8, we thus obtain $\|\beta_\zeta\|_{\mathcal{M}} = O_p(n^{-1/2})$.

So far, we have restricted the analysis to the maximizer β_ζ over the magnitude constrained set of dual variables $\mathcal{M}_n$. In the following we will show that this maximizer agrees with the maximizer over the unconstrained (original) set of dual variables $\mathcal{M}$. First, note that with $\|\beta_\zeta\|_{\mathcal{M}} = O_p(n^{-1/2})$ and $\zeta < 1/2$ we have that $n^{-\zeta} > n^{-1/2}$, which means that asymptotically β_ζ is contained in the interior of $\mathcal{M}_n$, i.e., $\beta_\zeta \in \operatorname{int}(\mathcal{M}_n)$. As β_ζ is a maximizer contained in the interior of the domain, it must correspond to a stationary point of $\widehat{G}$, i.e., $\frac{\partial \widehat{G}}{\partial \beta}(\bar{\theta}, \beta_\zeta) = 0$. Clearly $\mathcal{M}_n \subset \mathcal{M}$, so the stationary point is contained also in $\mathcal{M}$. As the empirical objective $\widehat{G}$ is concave with respect to β, this means we must have that $\widehat{G}(\bar{\theta}, \beta_\zeta) = \sup_{\beta\in\mathcal{M}} \widehat{G}(\bar{\theta}, \beta)$ and thus $\bar{\beta} = \beta_\zeta$, where again $\bar{\beta} = \arg\max_{\beta\in\mathcal{M}} \widehat{G}(\bar{\theta}, \beta)$.

As $\bar{\beta} = \beta_\zeta$, and $\|\beta_\zeta\|_{\mathcal{M}} = O_p(n^{-1/2})$ it directly follows that $\|\bar{\beta}\|_{\mathcal{M}} = O_p(n^{-1/2})$. Finally by assumption we have $\|\bar{\Psi}\|_{\mathcal{H}^*} = O_p(n^{-1/2})$ and thus $\widehat{G}(\bar{\theta}, \bar{\beta}) \leq -\epsilon\varphi^*(0) + \big(\|\bar{\Psi}\|_{\mathcal{H}^*} + o_p(n^{-1/2})\big)\|\bar{\beta}\|_{\mathcal{M}} - C\|\bar{\beta}\|_{\mathcal{M}}^2 = -\epsilon\varphi^*(0) + O_p(n^{-1})$. □

Lemma E.13. *Let the assumptions of Theorem A.1 be satisfied and denote the KMM estimator as* $\hat{\theta} = \arg\min_{\theta\in\Theta} \sup_{\beta\in\mathcal{M}} \widehat{G}_{\epsilon,\lambda_n}(\theta, \beta)$. *Then* $\|E_{\hat{P}_n}[\Psi(X, Z; \hat{\theta})]\|_{\mathcal{H}^*} = O_p\left(n^{-1/2}\right)$.

Proof. Define $\hat{\Psi}_i := \Psi(x_i, z_i; \hat{\theta})$ and $\hat{\Psi} = \frac{1}{n}\sum_{i=1}^n \hat{\Psi}_i$. For simplicity of notation let $\widehat{G}(\theta, \beta) := \widehat{G}_{\epsilon,\lambda_n}(\theta, \beta)$. For any $\eta \in \mathbb{R}$ and $f \in \mathcal{F}$ consider the dual variable $\bar{\beta} = (\eta, f, \phi(\hat{\Psi}))$ and its normalized version $\bar{\beta}_\zeta = n^{-\zeta}\bar{\beta}/\|\bar{\beta}\|$, where $\phi(\hat{\Psi})$ denotes the Riesz representer of $\hat{\Psi} \in \mathcal{H}^*$ in $\mathcal{H}$ and $0 < \zeta < 1/2$ as in Lemma E.9. Taylor expanding the KMM objective about $\beta = 0$ again yields

$$
\begin{aligned}
\widehat{G}(\hat{\theta}, \bar{\beta}_\zeta) =& -\epsilon\varphi^*(0) - \frac{1}{2}\bar{\beta}_\zeta^T \Lambda_n(\dot{\beta}, \hat{\theta})\bar{\beta}_\zeta + \frac{n^{-\zeta}}{\|\bar{\beta}\|}\int_{\mathcal{X}\times\mathcal{Z}} \Psi(x, z; \hat{\theta})(\phi(\Psi))\, \omega(\mathrm{d}x \otimes \mathrm{d}z) && (152)\\
&+ \frac{n^{-\zeta}}{\|\bar{\beta}\|}\int_{\mathcal{X}\times\mathcal{Z}} f(x, z)\hat{P}_n(\mathrm{d}x \otimes \mathrm{d}z) - \frac{n^{-\zeta}}{\|\bar{\beta}\|}\int_{\mathcal{X}\times\mathcal{Z}} f(x, z)\, \omega(\mathrm{d}x \otimes \mathrm{d}z) && (153)\\
=& -\epsilon\varphi^*(0) - \frac{1}{2}\bar{\beta}_\zeta^T \Lambda_n(\dot{\beta}, \hat{\theta})\bar{\beta}_\zeta + \frac{n^{-\zeta}}{\|\bar{\beta}\|}\int_{\mathcal{X}\times\mathcal{Z}} \Psi(x, z; \hat{\theta})(\phi(\hat{\Psi}))\hat{P}_n(\mathrm{d}x \otimes \mathrm{d}z) && (154)\\
&+ \frac{n^{-\zeta}}{\|\bar{\beta}\|}\int_{\mathcal{X}\times\mathcal{Z}} \Psi(x, z; \hat{\theta})(\phi(\hat{\Psi}))\, \omega(\mathrm{d}x \otimes \mathrm{d}z) - \frac{n^{-\zeta}}{\|\bar{\beta}\|}\int_{\mathcal{X}\times\mathcal{Z}} \Psi(x, z; \hat{\theta})(\phi(\hat{\Psi}))\hat{P}_n(\mathrm{d}x \otimes \mathrm{d}z) && (155)\\
&+ \frac{n^{-\zeta}}{\|\bar{\beta}\|}\int_{\mathcal{X}\times\mathcal{Z}} f(x, z)\hat{P}_n(\mathrm{d}x \otimes \mathrm{d}z) - \frac{n^{-\zeta}}{\|\bar{\beta}\|}\int_{\mathcal{X}\times\mathcal{Z}} f(x, z)\, \omega(\mathrm{d}x \otimes \mathrm{d}z) && (156)\\
\geq& -\epsilon\varphi^*(0) - \frac{1}{2}\bar{\beta}_\zeta^T \Lambda_n(\dot{\beta}, \hat{\theta})\bar{\beta}_\zeta + \frac{n^{-\zeta}}{\|\bar{\beta}\|}\|\hat{\Psi}\|_{\mathcal{H}^*}^2 - \frac{n^{-\zeta}}{\|\bar{\beta}\|}\|f\|_{\mathcal{F}}\,\mathrm{MMD}(\hat{P}_n, \omega; \mathcal{F}) && (157)\\
&- \frac{\alpha n^{-\zeta}}{\|\bar{\beta}\|}\left(\|\hat{\Psi}\|_{\mathcal{H}^*}\int_{\mathcal{X}\times\mathcal{Z}} \|\Psi(x, z; \hat{\theta})\|_{\mathcal{H}^*} Q(\mathrm{d}x \otimes \mathrm{d}z) + \|\hat{\Psi}\|_{\mathcal{H}^*}^2\right) && (158)\\
\geq& -\epsilon\varphi^*(0) - \frac{1}{2}\bar{\beta}_\zeta^T \Lambda_n(\dot{\beta}, \hat{\theta})\bar{\beta}_\zeta + C_\psi n^{-\zeta}\|\hat{\Psi}\|_{\mathcal{H}^*} && (159)\\
&- \alpha n^{-\zeta} C_\psi\left(C_Q + \|\hat{\Psi}\|_{\mathcal{H}^*}\right) - n^{-\zeta} C_f\,\mathrm{MMD}(\hat{P}_n, \omega; \mathcal{F}), && (160)
\end{aligned}
$$

where $C_\psi, C_f \in [0, 1]$ as $\|\hat{\Psi}\|_{\mathcal{H}^*}/\|\bar{\beta}\|_{\mathcal{M}} \leq 1$ and $\|f\|_{\mathcal{F}}/\|\bar{\beta}\|_{\mathcal{M}} \leq 1$ by definition of $\bar{\beta}$ and $\alpha = O_p(n^{-1})$ as well as $\mathrm{MMD}(\hat{P}_n, \omega; \mathcal{F}) = O_p(n^{-1})$ by Lemma E.8 and Assumption f). Using Lemma E.11 we can bound the largest eigenvalue of $\Lambda_n(\dot{\beta}, \hat{\theta})$ by some positive constant M which is independent of n, so we obtain

$$
\widehat{G}(\hat{\theta}, \bar{\beta}_\zeta) \geq -\epsilon\varphi^*(0) - Mn^{-2\zeta} + C_\psi n^{-\zeta}\|\hat{\Psi}\|_{\mathcal{H}^*} + O_p(n^{-1-\zeta}), \tag{161}
$$

Now as $(\hat{\theta}, \hat{\beta})$ is a saddle point of the *empirical* KMM objective, we have $\widehat{G}(\hat{\theta}, \bar{\beta}_\zeta) \leq \widehat{G}(\hat{\theta}, \hat{\beta}) \leq \max_{\beta\in\mathcal{M}} \widehat{G}(\theta_0, \beta)$. Putting this together with the previous inequality we have

$$
\begin{aligned}
-\epsilon\varphi^*(0) + C_\psi n^{-\zeta}\|\hat{\Psi}\|_{\mathcal{H}^*} - Mn^{-2\zeta} + O_p(n^{-1-\zeta}) &\leq \widehat{G}(\hat{\theta}, \bar{\beta}_\zeta) && (162)\\
&\leq \widehat{G}(\hat{\theta}, \hat{\beta}) && (163)\\
&\leq \max_{\beta\in\mathcal{M}} \widehat{G}(\theta_0, \beta) && (164)\\
&\leq -\epsilon\varphi^*(0) + O_p(n^{-1}), && (165)
\end{aligned}
$$

where in the last line we used Lemma E.12 with $\bar{\theta} = \theta_0$ which fulfills the corresponding conditions as $\|E[\Psi(X, Z; \theta_0)]\|_{\mathcal{H}^*} = 0$ by definition and thus by the central limit theorem $\|E_{\hat{P}_n}[\Psi(X, Z; \theta_0)]\|_{\mathcal{H}^*} = O_p(n^{-1/2})$. Adding $\epsilon\varphi^*(0)$ on both sides and solving for $\|\hat{\Psi}\|_{\mathcal{H}^*}$ yields

$$
\|\hat{\Psi}\|_{\mathcal{H}^*} \leq O_p(n^{-1+\zeta}) + O_p(n^{-\zeta}) = O_p(n^{-\zeta}), \tag{166}
$$

where the last step follows as $\zeta < 1/2$ by definition and thus $-1+\zeta < -\zeta$. Equation (166) provides an upper bound on the convergence rate for $\|\hat{\Psi}\|_{\mathcal{H}^*}$. To further refine this rate define $\tilde{\beta} := (0, 0, \phi(\hat{\Psi}))$ and for any sequence $\kappa_n \to 0$ consider $\kappa_n\tilde{\beta}$. Then as $\|\tilde{\beta}\|_{\mathcal{M}} = \|\hat{\Psi}\|_{\mathcal{H}^*} \leq O_p(n^{-\zeta})$ we immediately have $\|\kappa_n\tilde{\beta}\|_{\mathcal{M}} = o_p(n^{-\zeta})$ which implies $\kappa_n\tilde{\beta} \in \mathcal{M}_n$ w.p.a.1 and

$$-\epsilon\varphi^*(0) + \kappa_n\|\hat{\Psi}\|^2_{\mathcal{H}^*} - M\kappa_n^2\|\hat{\Psi}\|^2_{\mathcal{H}^*} + O_p(n^{-1-\zeta}) \leq \widehat{G}(\hat{\theta}, \kappa_n\tilde{\beta}) \tag{167}$$

$$\leq \widehat{G}(\hat{\theta}, \hat{\beta}) \tag{168}$$

$$\leq \max_{\beta\in\mathcal{M}} \widehat{G}(\theta_0, \beta) \tag{169}$$

$$\leq -\epsilon\varphi^*(0) + O_p(n^{-1}). \tag{170}$$

This implies $(1-\kappa_n M)\kappa_n\|\hat{\Psi}\|^2_{\mathcal{H}^*} \leq O_p(n^{-1})$ and as $(1-\kappa_n M)$ is bounded away from zero for all sufficiently large n, we get $\kappa_n\|\hat{\Psi}\|^2_{\mathcal{H}^*} = O_p(n^{-1})$. As this holds for all $\kappa_n \to 0$, we finally obtain $\|\hat{\Psi}\|_{\mathcal{H}^*} = O_p(n^{-1/2})$. □

Proof of Theorem A.1

Proof. Define $\hat{\Psi}_i = \Psi(x_i, z_i; \hat{\theta})$ and $\hat{\Psi} = \frac{1}{n}\sum_{i=1}^n \hat{\Psi}_i$. As $\hat{\Psi}$ is the average of n i.i.d. random variables $\hat{\Psi}_i$, by the central limit theorem and absolute homogeneity of the dual norm, we have $\|\hat{\Psi}(\theta) - E[\Psi(X, Z; \theta)]\|_{\mathcal{H}^*} = O_p(n^{-1/2})$ for any $\theta \in \Theta$. From Lemma E.13 we also have $\|\hat{\Psi}\|_{\mathcal{H}^*} = O_p(n^{-1/2})$ and thus using the triangle inequality we get

$$\left\|E[\Psi(X, Z; \hat{\theta})]\right\|_{\mathcal{H}^*} = \left\|E[\Psi(X, Z; \hat{\theta})] - \hat{\Psi} + \hat{\Psi}\right\|_{\mathcal{H}^*} \tag{171}$$

$$\leq \left\|E[\Psi(X, Z; \hat{\theta})] - \hat{\Psi}\right\|_{\mathcal{H}^*} + \left\|\hat{\Psi}\right\|_{\mathcal{H}^*} \tag{172}$$

$$= O_p(n^{-1/2}) \xrightarrow{p} 0. \tag{173}$$

As by assumption $\theta = \theta_0$ is the unique parameter for which $\theta \mapsto \|E[\Psi(X, Z; \theta)]\|_{\mathcal{H}^*} = 0$ it follows that $\hat{\theta} \xrightarrow{p} \theta_0$. To derive a convergence rate for $\hat{\theta}$ note that by the mean value theorem, there exists $\bar{\theta} \in \operatorname{conv}(\{\theta_0, \hat{\theta}\})$ such that

$$\Psi(X, Z; \hat{\theta}) = \Psi(X, Z; \theta_0) + (\hat{\theta} - \theta_0)^T\nabla_\theta\Psi(X, Z; \bar{\theta}). \tag{174}$$

Using this we have

$$\|E[\Psi(X, Z; \hat{\theta})]\|^2_{\mathcal{H}^*} = \|\underbrace{E[\Psi(X, Z; \theta_0)]}_{=0} + (\hat{\theta} - \theta_0)^T E[\nabla_\theta\Psi(X, Z; \bar{\theta})]\|^2_{\mathcal{H}^*} \tag{175}$$

$$= \left\langle (\hat{\theta} - \theta_0)^T E[\nabla_\theta\Psi(X, Z; \bar{\theta})], (\hat{\theta} - \theta_0)^T E[\nabla_\theta\Psi(X, Z; \bar{\theta})]\right\rangle_{\mathcal{H}^*} \tag{176}$$

$$= (\hat{\theta} - \theta_0)^T \underbrace{\left\langle E[\nabla_\theta\Psi(X, Z; \bar{\theta})], E[\nabla_\theta\Psi(X, Z; \bar{\theta})]\right\rangle_{\mathcal{H}^*}}_{=:\Sigma(\bar{\theta})} (\hat{\theta} - \theta_0) \tag{177}$$

$$\geq \lambda_{\min}\left(\Sigma(\bar{\theta})\right)\|\hat{\theta} - \theta_0\|_2^2 \tag{178}$$

Now as $\hat{\theta} \xrightarrow{p} \theta_0$ and $\bar{\theta} \in \operatorname{conv}(\{\theta_0, \hat{\theta}\})$ 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 Assumption j), 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 $\|\hat{\theta} - \theta_0\| = O_p(n^{-1/2})$. □

E.4.2. Proof of Theorem A.2

To show asymptotic normality we linearize the first order conditions for $(\hat{\theta}, \hat{\beta})$ about the true parameters $(\theta_0, 0)$ and solve for the KMM estimates. This involves the inversion of a blockmatrix for whose invertibility we require the following Lemma.

Lemma E.14. *For a moment functional $\Psi(X, Z; \theta) : \mathcal{H} \to \mathbb{R}$, continuously differentiable in θ, define the covariance operator $\Omega_0 := E[\Psi(X, Z; \theta) \otimes \Psi(X, Z; \theta)]$. Further define the matrix $\Sigma_0 := \langle E[\nabla_\theta\Psi(X, Z; \theta_0)], E[\nabla_\theta\Psi(X, Z; \theta_0)]\rangle_{\mathcal{H}^*} \in$*

$\mathbb{R}^{p\times p}$, *where the inner product is only taken with respect to the* $\mathcal{H}^*$ *index. If* Ω_0 *and* Σ_0 *are non-singular with smallest eigenvalue bounded away from zero, then the matrix*

$$\Gamma := -E[\nabla_\theta \Psi(X, Z; \theta_0)]\Omega_0^{-1} E[\nabla_\theta \Psi(X, Z; \theta_0)] \tag{179}$$

is non-singular with smallest eigenvalue bounded away from zero.

Proof. Let $\mathcal{H} = \bigoplus_{i=1}^m \mathcal{H}_i$ and for $i = 1, \dots, m$ let $\{h_j^i\}_{j=1}^\infty$ denote a orthonormal basis of $\mathcal{H}_i^*$ such that $\langle h_i^k, h_j^l\rangle = \delta_{ij}\delta_{kl}$. Then we can write the identity operator in $\mathcal{H}^*$ as $I_{\mathcal{H}^*} = \sum_{i=1}^m \sum_{j=1}^\infty h_j^i \left(h_j^i\right)^*$, where $\left(h_j^i\right)^*$ is the Riesz representer of $h_j^i \in \mathcal{H}^*$ in $\mathcal{H}^{**}$ which can be uniquely identified with an element in $\mathcal{H}$ by the property of Hilbert spaces. Further, let $\nabla_\theta \Psi_0 := E[\nabla_\theta \Psi(X, Z; \theta_0)]$. Then we can write for any $\theta \in \Theta$ with $0 < \|\theta\| < \infty$,

$$\begin{aligned}
-\theta^T \Gamma \theta &= \theta^T \nabla_\theta \Psi_0 \Omega_0^{-1} \nabla_\theta \Psi_0 \theta && (180)\\
&= \theta^T \nabla_\theta \Psi_0 \sum_{i=1}^m \sum_{j=1}^\infty h_j^i \left(h_j^i\right)^* \Omega_0^{-1} \sum_{k=1}^m \sum_{l=1}^\infty h_l^k \left(h_l^k\right)^* \left(\nabla_\theta \Psi_0\right)^T \theta && (181)\\
&= \theta^T \nabla_\theta \Psi_0 \left(\sum_{i,k=1}^m \sum_{j,l=1}^\infty h_j^i \, \langle h_j^i, \Omega_0^{-1} h_l^k \rangle_{\mathcal{H}^*} \, \left(h_l^k\right)^* \right) \left(\nabla_\theta \Psi_0\right)^T \theta && (182)\\
&\geq \lambda_{\min}(\Omega_0^{-1}) \theta^T \nabla_\theta \Psi_0 \left(\sum_{i=1}^m \sum_{j=1}^\infty h_j^i \left(h_j^i\right)^* \right) \left(\nabla_\theta \Psi_0\right)^T \theta && (183)\\
&= \lambda_{\min}(\Omega_0^{-1}) \theta^T \langle \nabla_\theta \Psi_0, \nabla_\theta \Psi_0 \rangle_{\mathcal{H}^*} \theta && (184)\\
&= \lambda_{\min}(\Omega_0^{-1}) \theta^T \Sigma_0 \theta && (185)\\
&\geq \lambda_{\min}(\Omega_0^{-1}) \lambda_{\min}(\Sigma_0) \|\theta\|^2 > 0, && (186)
\end{aligned}$$

where we used that Ω_0 is positive semi-definite by construction and non singular by Assumption e) and thus being the inverse of a strictly positive definite operator the smallest eigenvalue $\lambda_{\min}(\Omega_0^{-1})$ of Ω_0^{-1} is positive and bounded away from zero. Moreover, Σ_0 is positive semi-definite by construction and non singular by Assumptions j) and therefore its smallest eigenvalue $\lambda_{\min}(\Sigma_0)$ positive and bounded away from zero. From this it immediately follows that Γ is strictly negative definite and thus non-singular. □

Proof of Theorem A.2

Proof. The KMM estimator $\hat{\theta}$ and the optimal Lagrange parameter $\hat{\beta}$ are determined via the first order optimality conditions for (θ, β) which are given by

$$0 = \frac{\partial \widehat{G}}{\partial \theta}(\hat{\theta}, \hat{\beta}) = -\int_{\mathcal{X}\times\mathcal{Z}} \varphi_1^* \left(\frac{1}{\epsilon} a(x, z; \hat{\theta})^T \hat{\beta} \right) \nabla_\theta \left(a(x, z; \hat{\theta})^T \hat{\beta} \right) \omega(\mathrm{d}x \otimes \mathrm{d}z) \tag{187}$$

$$0 = \frac{\partial \widehat{G}}{\partial \beta}(\hat{\theta}, \hat{\beta}) = \frac{1}{n}\sum_{i=1}^n b_i - \int_{\mathcal{X}\times\mathcal{Z}} \varphi_1^* \left(\frac{1}{\epsilon} a(x, z; \hat{\theta})^T \hat{\beta} \right) a(x, z; \hat{\theta}) \, \omega(\mathrm{d}x \otimes \mathrm{d}z) - R_{\lambda_n} \beta \tag{188}$$

Linearizing the first condition (187) about the true parameters $(\theta_0, 0)$ yields

$$0 = \frac{\partial \widehat{G}}{\partial \theta}(\theta_0, 0) + \left(\frac{\partial^2 \widehat{G}}{\partial \theta\, \partial \theta}(\bar{\theta}, \bar{\beta})\right)\left(\hat{\theta} - \theta_0\right) + \left(\frac{\partial^2 \widehat{G}}{\partial \theta\, \partial \beta}(\bar{\theta}, \bar{\beta})\right)\hat{\beta} \tag{189}$$

$$= -\left(\int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon}\varphi_2^*\left(\frac{1}{\epsilon}a(x,z;\bar{\theta})^T\bar{\beta}\right)\left(\nabla_\theta a(x,z;\bar{\theta})^T\bar{\beta}\right)\left(\nabla_\theta a(x,z;\bar{\theta})^T\bar{\beta}\right)^T \omega(\mathrm{d}x\otimes\mathrm{d}z)\right)\left(\hat{\theta}-\theta_0\right) \tag{190}$$

$$-\left(\int_{\mathcal{X}\times\mathcal{Z}} \varphi_1^*\left(\frac{1}{\epsilon}a(x,z;\bar{\theta})^T\bar{\beta}\right)\left(\nabla_\theta^2 a(x,z;\bar{\theta})^T\bar{\beta}\right)\omega(\mathrm{d}x\otimes\mathrm{d}z)\right)\left(\hat{\theta}-\theta_0\right) \tag{191}$$

$$-\left(\int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon}\varphi_2^*\left(\frac{1}{\epsilon}a(x,z;\bar{\theta})^T\bar{\beta}\right)\left(\nabla_\theta a(x,z;\bar{\theta})^T\bar{\beta}\right)a(x,z;\bar{\theta})^T\,\omega(\mathrm{d}x\otimes\mathrm{d}z)\right)\hat{\beta} \tag{192}$$

$$-\left(\int_{\mathcal{X}\times\mathcal{Z}} \varphi_1^*\left(\frac{1}{\epsilon}a(x,z;\bar{\theta})^T\bar{\beta}\right)\nabla_\theta a(x,z;\bar{\theta})^T\,\omega(\mathrm{d}x\otimes\mathrm{d}z)\right)\hat{\beta} \tag{193}$$

for some $(\bar{\theta}, \bar{\beta})$ on the line between $(\hat{\theta}, \hat{\beta})$ and $(\theta_0, 0)$. Analogously the linearization of the second condition (188) is given by

$$0 = \frac{\partial \widehat{G}}{\partial \beta}(\theta_0, 0) + \left(\frac{\partial^2 \widehat{G}}{\partial \theta\, \partial \beta}(\dot{\theta}, \dot{\beta})\right)\left(\hat{\theta} - \theta_0\right) + \left(\frac{\partial^2 \widehat{G}}{\partial \beta\, \partial \beta}(\dot{\theta}, \dot{\beta})\right)\hat{\beta} \tag{194}$$

$$= -\frac{1}{n}\sum_{i=1}^n b_i + \int_{\mathcal{X}\times\mathcal{Z}} a(x,z;\theta_0)\,\omega(\mathrm{d}x\otimes\mathrm{d}z) \tag{195}$$

$$-\left(\int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon}\varphi_2^*\left(\frac{1}{\epsilon}a(x,z;\dot{\theta})^T\dot{\beta}\right)a(x,z;\dot{\theta})\left(\nabla_\theta a(x,z;\dot{\theta})^T\dot{\beta}\right)\omega(\mathrm{d}x\otimes\mathrm{d}z)\right)\left(\hat{\theta}-\theta_0\right) \tag{196}$$

$$-\left(\int_{\mathcal{X}\times\mathcal{Z}} \varphi_1^*\left(\frac{1}{\epsilon}a(x,z;\dot{\theta})^T\dot{\beta}\right)\nabla_\theta a(x,z;\dot{\theta})\,\omega(\mathrm{d}x\otimes\mathrm{d}z)\right)\left(\hat{\theta}-\theta_0\right) \tag{197}$$

$$-\left(\int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon}\varphi_2^*\left(\frac{1}{\epsilon}a_i(\dot{\theta})^T\dot{\beta}\right)a(x,z;\dot{\theta})a(x,z;\dot{\theta})^T\,\omega(\mathrm{d}x\otimes\mathrm{d}z) + R_{\lambda_n}\right)\hat{\beta} \tag{198}$$

Now, as $\bar{\beta}, \dot{\beta}$ are on the line between $\hat{\beta}$ and 0 and $\hat{\beta} = O_p(n^{-1/2})$ by Lemma E.12, we have that all derivative terms of $\widehat{G}$ involving $\bar{\beta}, \dot{\beta}$ linearly are $O_p(n^{-1})$, as each term additionally gets multiplied by $\hat{\theta} - \theta$ or $\hat{\beta}$ which both are $O_p(n^{-1/2})$ in the respective norms. Further it follows from Lemma E.9 that $\varphi_j(a(X,Z;\theta)^T\tilde{\beta}) \xrightarrow{p} 1$ ω-a.s. for any $\theta \in \Theta$, $\tilde{\beta} \in \{\bar{\beta}, \dot{\beta}\}$ and $j = 1, 2$. Therefore, the first linearized first order condition (187) reduces to

$$0 = \left(\int_{\mathcal{X}\times\mathcal{Z}} \nabla_\theta a(x,z;\bar{\theta})^T\,\omega(\mathrm{d}x\otimes\mathrm{d}z)\right)\hat{\beta} + O_p(n^{-1}). \tag{199}$$

For the second condition note that by Lemma E.10

$$\int_{\mathcal{X}\times\mathcal{Z}} a(x,z;\theta_0)\,\omega(\mathrm{d}x\otimes\mathrm{d}z) = \frac{1}{n}\sum_{i=1}^n a(x_i,z_i;\theta_0) + O_p(n^{-1}). \tag{200}$$

Now inserting this into the second linearized first order condition we obtain

$$0 = \frac{1}{n}\sum_{i=1}^n (b_i - a(x_i,z_i;\theta_0)) + \int_{\mathcal{X}\times\mathcal{Z}} \nabla_\theta a(x,z;\dot{\theta})^T\,\omega(\mathrm{d}x\otimes\mathrm{d}z)\left(\hat{\theta}-\theta_0\right) \tag{201}$$

$$+ \underbrace{\left(\int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon}a(x,z;\dot{\theta})a(x,z;\dot{\theta})^T\,\omega(\mathrm{d}x\otimes\mathrm{d}z) + R_{\lambda_n}\right)}_{=:\Lambda_n(\dot{\theta})}\hat{\beta} + O_p(n^{-1}), \tag{202}$$

where $\frac{1}{n}\sum_{i=1}^n (b_i - a(x_i,z_i;\theta_0)) = \left(0, 0, \hat{\Psi}(\theta_0)\right)^T$. Writing the two linearized first order conditions in matrix-vector form

we obtain

$$\begin{pmatrix}0\\0\end{pmatrix} = \begin{pmatrix}0\\ \frac{1}{n}\sum_{i=1}^n \left(b_i - a(x_i,z_i;\theta_0)\right)\end{pmatrix} \tag{203}$$

$$+ \underbrace{\begin{pmatrix}0 & \int_{\mathcal{X}\times\mathcal{Z}} \nabla_\theta a(x,z;\bar\theta)^T\, \omega(\mathrm{d}x\otimes\mathrm{d}z)\\ \int_{\mathcal{X}\times\mathcal{Z}} \nabla_\theta a(x,z;\dot\theta)^T\, \omega(\mathrm{d}x\otimes\mathrm{d}z) & \Lambda_n(\dot\theta)\end{pmatrix}}_{:=M_n} \begin{pmatrix}\hat\theta-\theta_0\\ \hat\beta-\beta_0\end{pmatrix}, \tag{204}$$

where $\beta_0 = 0$. Now $\hat\theta \xrightarrow{p} \theta_0$ and $\dot\theta$ and $\bar\theta$ are on the line between $\hat\theta$ and θ, and $\omega \xrightarrow{p} P_0$ weakly by definition. Moreover, $\nabla_\theta\Psi(X,Z;\bar\theta)$ is continuous for any $\bar\theta$ in a neighborhood $\bar\Theta$ of θ_0 by Assumption i) and thus by the continuous mapping theorem we have $\int_{\mathcal{X}\times\mathcal{Z}} \nabla_\theta a(x,z;\bar\theta)\, \omega(\mathrm{d}x\otimes\mathrm{d}z) \xrightarrow{p} E[\nabla_\theta a(X,Z;\theta_0)] =: \nabla_\theta a_0$ and the same holds for the other off-diagonal entry. In addition, the off-diagonal entries are bounded as Assumption i) with Lemma E.7 implies $E[\sup_{\theta\in\bar\Theta}\|\nabla_\theta\Psi(X,Z;\theta)\|] < \infty$. Finally, by the uniform weak law of large numbers and the continuous mapping theorem we have $\Lambda_n(\dot\theta) \xrightarrow{p} \Lambda(\theta_0) = E[a(X,Z;\theta_0)a(X,Z;\theta_0)^T] + R_{\lambda=0} =: \Lambda$. Let correspondingly M denote the limit operator for M_n, i.e., $M_n \xrightarrow{p} M$. From Lemma E.11 it follows that the smallest eigenvalue of Λ is bounded away from zero and thus it is invertible. Now suppose the Schur complement of Λ in M,

$$\Gamma := M/\Lambda = -\left(\nabla_\theta a_0^T\right)\Lambda^{-1}\left(\nabla_\theta a_0^T\right) = -\nabla_\theta\Psi_0\left(\Lambda^{-1}\right)_{3,3}\nabla_\theta\Psi_0 \tag{205}$$

is invertible, then it follows from standard blockmatrix algebra (see e.g. Bernstein (2009)), that the inverse of M is given by

$$M^{-1} = \begin{pmatrix}\Gamma^{-1} & \Gamma^{-1}\left(\nabla_\theta a_0\right)\Lambda^{-1}\\ \Lambda^{-1}\left(\nabla_\theta a_0\right)\Gamma^{-1} & \Lambda^{-1} + \Lambda^{-1}\left(\nabla_\theta a_0\right)\Gamma^{-1}\left(\nabla_\theta a_0\right)\Lambda^{-1}\end{pmatrix}. \tag{206}$$

Now it remains to find an explicit expression for $(\Lambda^{-1})_{3,3}$ and to show that Γ is indeed invertible. To this aim we write out the outer product over $\mathcal{M}\times\mathcal{M}$ in Λ which yields

$$\Lambda = \frac{1}{\epsilon}\begin{pmatrix}1 & E[1\otimes k((X,Z),\cdot)] & -E[1\otimes\Psi(X,Z;\theta_0)]\\ E[k((X,Z),\cdot)\otimes 1] & E[k((X,Z),\cdot)\otimes k((X,Z),\cdot)] + I & -E[k((X,Z),\cdot)\otimes\Psi(X,Z;\theta_0)]\\ -E[\Psi(X,Z;\theta_0)\otimes 1] & -E[\Psi(X,Z;\theta_0)\otimes k((X,Z),\cdot)] & E[\Psi(X,Z;\theta_0)\otimes\Psi(X,Z;\theta_0)]\end{pmatrix} \tag{207}$$

$$= \frac{1}{\epsilon}\begin{pmatrix}1 & E[1\otimes k((X,Z),\cdot)] & 0\\ E[k((X,Z),\cdot)\otimes 1] & E[k((X,Z),\cdot)\otimes k((X,Z),\cdot)] + I & 0\\ 0 & 0 & \Omega_0\end{pmatrix} \tag{208}$$

where we used that $\|E[\Psi(X,Z;\theta_0)]\|_{\mathcal{H}^*} = 0$ by definition and as $\mathcal{F}$ is an RKHS, the evaluation functional $k((x,z),\cdot)$ can be bounded by some constant C and thus we have $\|E[k((X,Z),\cdot)\otimes\Psi(X,Z;\theta_0)]\|_{\mathcal{F}\times\mathcal{H}^*} \le C\|E[\Psi(X,Z;\theta_0)]\|_{\mathcal{H}^*} = 0$. Now Λ is of blockdiagonal form and for the upper block B we have

$$B = \begin{pmatrix}1 & E[1\otimes k((X,Z),\cdot)]\\ E[k(X,Z),\cdot)\otimes 1] & E[k(X,Z),\cdot)\otimes k((X,Z),\cdot)]\end{pmatrix} + \begin{pmatrix}1 & 0\\ 0 & I\end{pmatrix}. \tag{209}$$

The first term is symmetric and thus positive semi-definite and the second term diagonal with positive entries, thus B is a strictly positive definite operator and thus invertible. Moreover, by Assumption e) of Theorem A.1 Ω_0 is invertible and thus we can conclude

$$\Lambda^{-1} = \begin{pmatrix}B^{-1} & 0\\ 0 & \Omega_0^{-1}\end{pmatrix} \tag{210}$$

and $\left(\Lambda^{-1}\right)_{3,3} = \Omega_0^{-1}$. Now, from invertibility of Ω_0 we directly obtain that Ω_0^{-1} is non-singular and thus by Assumption e) and Lemma E.14 we have that Γ is non-singular and invertible which legitimates the inversion of M.

With this at hand, we can solve equation (204) for $\hat\theta - \theta_0$ and obtain

$$\sqrt{n}\left(\hat\theta - \theta_0\right) = \left(\Gamma^{-1}\nabla_\theta\Psi_0\left(\Lambda^{-1}\right)_{3,3}\right)\sqrt{n}\hat\Psi(\theta_0) \tag{211}$$

$$= -\left(\left(\nabla_\theta\Psi_0\left(\Lambda^{-1}\right)_{3,3}\nabla_\theta\Psi_0\right)^{-1}\nabla_\theta\Psi_0\left(\Lambda^{-1}\right)_{3,3}\right)\sqrt{n}\hat\Psi(\theta_0). \tag{212}$$

By the central limit theorem we have $\sqrt{n}E_{\hat{P}_n}[\Psi(X,Z;\theta_0)] \sim \mathcal{N}(0,\Omega_0)$ where as before $\Omega_0 = E[\Psi(X,Z;\theta_0) \otimes \Psi(X,Z;\theta_0)]$ and thus inserting into equation (212) we get

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = -\left(\left(\left(\nabla_\theta \Psi_0\right)\Omega_0^{-1}\left(\nabla_\theta \Psi_0\right)\right)^{-1}\left(\nabla_\theta \Psi_0\right)\Omega_0^{-1}\right)\sqrt{n}\hat{\Psi}(\theta_0) \sim \mathcal{N}(0,\Xi) \tag{213}$$

with

$$\Xi = \left(\left(\left(\nabla_\theta \Psi_0\right)\Omega_0^{-1}\left(\nabla_\theta \Psi_0\right)\right)^{-1}\left(\nabla_\theta \Psi_0\right)\Omega_0^{-1}\right)\Omega_0\left(\left(\left(\nabla_\theta \Psi_0\right)\Omega_0^{-1}\left(\nabla_\theta \Psi_0\right)\right)^{-1}\left(\nabla_\theta \Psi_0\right)\Omega_0^{-1}\right)^T \tag{214}$$

$$= \left(\left(\nabla_\theta \Psi_0\right)\Omega_0^{-1}\left(\nabla_\theta \Psi_0\right)\right)^{-1}. \tag{215}$$

□

E.5. Asymptotic Properties of KMM for Finite-Dimensional Moment Restrictions

E.5.1. Proof of Theorem B.1 (Consistency for MR)

The consistency for the finite dimensional case follows as a special case of the consistency result for the functional case (Theorem A.1) by identifying $\mathcal{H} = \mathbb{R}^m$, $\Psi(x, z; \theta) = \psi(x; \theta) \in \mathbb{R}^m$ and $\lambda_n = 0$. For a finite dimensional version of Theorem A.1 we need finite dimensional versions of Lemmas E.9-E.13, which we will state in the following and describe the differences in the proofs compared to the functional case. Refer to the proof of Theorem A.1 for details.

Lemma E.15. *Let $\mathcal{A}$ denote a σ-algebra on $\mathcal{X}$ and let $(\mathcal{X}, \mathcal{A}, \omega)$ be a probability space with measure ω. For any function $\psi : \mathcal{X} \times \Theta \to \mathbb{R}^m$ with $\int_{\mathcal{X}} \sup_{\theta \in \Theta} \|\psi(x; \theta)\|_2^2 \, \omega(\mathrm{d}x) < \infty$, it follows that $\sup_{\theta \in \Theta} \|\psi(X; \theta)\|_2 \leq C$ ω-a.s. for some constant $C < \infty$.*

Proof. The proof follows immediately from the one for Lemma E.7 by exchanging $\Psi(X, Z; \theta) \to \psi(X; \theta)$ and $\mathcal{H} \to \mathbb{R}^m$. □

Lemma E.16. *Let the assumptions of Theorem B.1 be satisfied, then for any ζ with $0 < \zeta < 1/2$ define the magnitude constrained set of dual variables $\mathcal{M}_n = \{\beta = (\eta, f, h) \in \mathcal{M} : \|\beta\|_{\mathcal{M}} \leq n^{-\zeta}\}$. Then as $n \to \infty$,*

$$\sup_{\theta \in \Theta, \ (\eta, f, h) \in \mathcal{M}_n} |\psi(X; \theta)^T h| = O_p(n^{-\zeta}) \ \omega\text{-a.s.}, \tag{216}$$

$$\sup_{\theta \in \Theta, \ \beta \in \mathcal{M}_n} |a(X; \theta)^T \beta| = O_p(n^{-\zeta}) \ \omega\text{-a.s.}. \tag{217}$$

Proof. The proof follows immediately from the one for Lemma E.9 by exchanging $\Psi(X, Z; \theta) \to \psi(X; \theta)$ and $\mathcal{H} \to \mathbb{R}^m$ and using Lemma E.15 instead of Lemma E.7 to bound $\sup_{\theta \in \Theta} \|\psi(X; \theta)\|_2 \leq C$ ω-a.s.. □

Lemma E.17. *Let the assumptions of Theorem B.1 be satisfied and consider $\bar{\theta} \in \Theta$ such that $\bar{\theta} \xrightarrow{p} \theta_0$. Further let $\beta_\zeta := \arg\max_{\beta \in \mathcal{M}_n} \widehat{G}(\bar{\theta}, \beta)$, where $\mathcal{M}_n = \{\beta \in \mathcal{M} : \|\beta\|_{\mathcal{M}} \leq n^{-\zeta}\}$ with $0 < \zeta < 1/2$. Define the operator $\Lambda_n(\beta, \theta) : \mathcal{M} \to \mathcal{M}$ as*

$$\Lambda_n(\beta, \theta) := \int_{\mathcal{X}} \frac{1}{\epsilon} a(x; \theta) a(x; \theta)^T \varphi_2^* \left(\frac{1}{\epsilon} a(x; \theta)^T \beta \right) \omega(\mathrm{d}x) + R. \tag{218}$$

Then w.p.a.1 for any $\bar{\beta} \in \mathrm{conv}(\{0, \beta_\zeta\})$, $\Lambda_n(\bar{\beta}, \bar{\theta})$ is strictly positive definite and its smallest eigenvalue is bounded away from zero. Moreover, for any $\theta \in \Theta$ the largest eigenvalue of $\Lambda_n(\bar{\beta}, \theta)$ is bounded from above by a positive constant M.

Proof. The proof follows from the one for the functional case Lemma E.11 with the difference that we directly impose non-singularity of the covariance matrix $\Omega_0 = E[\psi(X; \theta_0)\psi(X; \theta_0)^T]$ by Assumption e) of Theorem B.1. □

Lemma E.18. *Let the assumptions of Theorem B.1 be satisfied. Additionally let $\bar{\theta} \in \Theta$, $\bar{\theta} \xrightarrow{p} \theta_0$, and $\|E_{\hat{P}_n}[\psi(X; \bar{\theta})]\|_2 = O_p\left(n^{-1/2}\right)$. Then for $\bar{\beta} = \arg\max_{\beta \in \mathcal{M}} \widehat{G}_{\epsilon, \lambda_n}(\bar{\theta}, \beta)$ we have $\|\bar{\beta}\|_{\mathcal{M}} = O_p\left(n^{-1/2}\right)$, and $\widehat{G}_{\epsilon, \lambda_n}(\bar{\theta}, \bar{\beta}) \leq -\epsilon\varphi^*(0) + O_p\left(n^{-1}\right)$.*

Proof. The proof follows immediately from the one for the functional case (Lemma E.12) with the usual substitutions. □

Lemma E.19. *Let the assumptions of Theorem B.1 be satisfied and denote the KMM estimator as $\hat{\theta} = \arg\min_{\theta \in \Theta} \sup_{\beta \in \mathcal{M}} \widehat{G}_{\epsilon, \lambda_n}(\theta, \beta)$. Then $\|E_{\hat{P}_n}[\psi(X; \hat{\theta})]\|_{\mathcal{H}^*} = O_p\left(n^{-1/2}\right)$.*

Proof. The proof follows immediately from the one for the functional case (Lemma E.13) with the usual substitutions. □

Lemma E.20. *Consider a moment function $\psi : \mathcal{X} \times \Theta \to \mathbb{R}^m$ with $\nabla_\theta \psi(x; \theta) \in \mathbb{R}^{p \times m}$. Then if $\mathrm{rank}\left(E[\nabla_\theta \psi(X; \theta_0)]\right) = p$ the matrix $\Sigma_0 = E[\nabla_\theta \psi(X; \theta_0)] E[\nabla_\theta \psi(X; \theta_0)]^T \in \mathbb{R}^{p \times p}$ is non-singular with smallest eigenvalue positive and bounded away from zero.*

Proof. As $\mathrm{rank}(E[\nabla_\theta \psi(X; \theta_0)] = p$, its rows are linearly independent and thus for any $\theta \in \Theta$ with $0 < \|\theta\| < \infty$ there exists $j \in \{1, \ldots, m\}$ such that $\theta^T E[\nabla_\theta \psi_j(X; \theta_0)] \neq 0$. This means that $\theta^T \Sigma_0 \theta = \sum_{i=1}^m \left(\theta^T E[\nabla_\theta \psi_i(X; \theta_0)]\right)^2 \geq \left(\theta^T E[\nabla_\theta \psi_j(X; \theta_0)]\right)^2 > 0$, which follows as any term in the sum is non-negative and there exists at least one term of index j which is positive. Therefore, the smallest eigenvalue of Σ_0 is positive and bounded away from zero. □

Proof of Theorem B.1

Proof. Define $\hat{\psi}_i = \psi(x_i; \hat{\theta})$ and $\hat{\psi} = \frac{1}{n}\sum_{i=1}^n \hat{\psi}_i$. As $\hat{\psi}$ is the average of n i.i.d. random variables $\hat{\psi}_i$, by the central limit theorem and absolute homogeneity of the dual norm, we have $\|\hat{\psi}(\theta) - E[\psi(X;\theta)]\|_2 = O_p(n^{-1/2})$ for any $\theta \in \Theta$. From Lemma E.19 we also have $\|\hat{\psi}\| = O_p(n^{-1/2})$ and thus using the triangle inequality we get

$$\left\|E[\psi(X;\hat{\theta})]\right\|_2 = \left\|E[\psi(\hat{\theta})] - \hat{\psi} + \hat{\psi}\right\|_2 \tag{219}$$

$$\leq \left\|E[\psi(X;\hat{\theta})] - \hat{\psi}\right\|_2 + \left\|\hat{\psi}\right\|_2 \tag{220}$$

$$= O_p(n^{-1/2}) \xrightarrow{p} 0. \tag{221}$$

As by assumption θ_0 is the unique parameter for which $E[\psi(X;\theta)] = 0$ it follows that $\hat{\theta} \xrightarrow{p} \theta_0$. To derive a convergence rate for $\hat{\theta}$ note that by the mean value theorem, there exists $\bar{\theta} \in \operatorname{conv}(\{\theta_0, \hat{\theta}\})$ such that

$$\psi(X;\hat{\theta}) = \psi(X;\theta_0) + (\hat{\theta} - \theta_0)^T \nabla_\theta \psi(X;\bar{\theta}), \tag{222}$$

where $\nabla_\theta \psi(x;\theta) \in \mathbb{R}^{p \times m}$. Using this we have

$$\|E[\psi(X;\hat{\theta})]\|_2^2 = \|\underbrace{E[\psi(X;\theta_0)]}_{=0} + (\hat{\theta} - \theta_0)^T E[\nabla_\theta \psi(X;\bar{\theta})]\|_2^2 \tag{223}$$

$$= \left\langle (\hat{\theta} - \theta_0)^T E[\nabla_\theta \psi(X;\bar{\theta})], (\hat{\theta} - \theta_0)^T E[\nabla_\theta \psi(X;\bar{\theta})] \right\rangle \tag{224}$$

$$= (\hat{\theta} - \theta_0)^T \underbrace{E[\nabla_\theta \psi(X;\bar{\theta})] E[\nabla_\theta \psi(X;\bar{\theta})]^T}_{=:\Sigma(\bar{\theta})} (\hat{\theta} - \theta_0) \tag{225}$$

$$\geq \lambda_{\min}\left(\Sigma(\bar{\theta})\right) \|\hat{\theta} - \theta_0\|_2^2 \tag{226}$$

Now as $\hat{\theta} \xrightarrow{p} \theta_0$ and $\bar{\theta} \in \operatorname{conv}(\{\theta_0, \hat{\theta}\})$ 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. Further by Assumption i) of Theorem B.1 and Lemma E.20 it follows that Σ_0 is positive definite and thus as $\Sigma(\bar{\theta}) \xrightarrow{p} \Sigma_0$ it follows that $\Sigma(\bar{\theta})$ is positive definite with smallest eigenvalue $\lambda_{\min}\left(\Sigma(\bar{\theta})\right)$ positive and bounded away from zero w.p.a.1. Finally as $\|E[\psi(X;\hat{\theta})]\| = O_p(n^{-1/2})$ taking the square-root on both sides we have $\|\hat{\theta} - \theta_0\| = O_p(n^{-1/2})$. □

E.5.2. Proof of Theorem B.2 (Asymptotic Normality for MR)

Proof of Theorem B.2

Proof. The proof follows directly from the one for functional moment restrictions Theorem A.2 by identifying $\Psi(x, z; \theta) = \psi(x;\theta)$, $\mathcal{H} = \mathbb{R}^m$, setting $\lambda_n = 0$ and using Lemma E.20 to translate the rank condition, Assumption i), into non-singularity of $\Sigma_0 = E[\nabla_\theta \psi(X;\theta_0)] E[\nabla_\theta \psi(X;\theta_0)]^T \in \mathbb{R}^{p \times p}$. □