Empirical Likelihood Estimators for Robust and Causal Learning

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Heiner Stephan Kremer

aus Bad Nauheim

Tübingen

2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	10.03.2026
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Bernhard Schölkopf
2. Berichterstatter:	Prof. Dr. Zeynep Akata
3. Berichterstatter:	Prof. Dr. Jilles Vreeken

Abstract

Some of the central problems in robust and causal machine learning, including learning under covariate shifts and instrumental variable regression, can be expressed as *conditional moment restrictions* (CMR). By restricting the conditional expectation of a signed error metric, models identified via CMR exhibit robustness against shifts in the distribution of the conditioning variable. In practice, this generally results in an ill-posed problem, as it requires the solution of an over-identified infinite-dimensional system of equations. For the unconditional case, *empirical likelihood* estimators have emerged as general and powerful tools to address over-identified moment restriction problems. These methods learn a model along with an approximation of the population distribution by means of minimizing a φ-divergence constrained by the moment restrictions. The main goal of this work is to advance the state-of-the art in CMR estimation by extending and refining the idea of empirical likelihood estimation in several directions. First, we generalize the classical framework to *conditional* moment restrictions using a functional formulation, that leverages modern machine learning models. Then, we extend the principle to alternative distributional distance notions based on kernel methods and optimal transport. The resulting estimators exhibit superior small sample properties and robustness against data corruptions at training time and adversarial attacks at test time, respectively. Finally, drawing inspiration from the close relation between empirical likelihood estimation and *distributionally robust optimization* (DRO), we provide an application of kernel-based DRO on chance-constrained programming.

Zusammenfassung

Einige der zentralen Probleme im Bereich des robusten und kausalen maschinellen Lernens können als *bedingte Momentenbeschränkungen* (engl. conditional moment restrictions (CMR)) formuliert werden. Bekannte Beispiele für solche Anwendungen sind beispielsweise Regression mit instrumentellen Variablen und das Lernen unter Änderungen der marginalen Wahrscheinlichkeitsverteilungen zwischen Test- und Trainingsdaten. Durch die Fixierung des bedingten Erwartungswertes eines vorzeichenbehafteten Fehlermaßes, zeigen Modelle, die durch CMR identifiziert sind, Robustheit gegenüber Änderungen in der Verteilung der bedingenden Variable. In der Praxis stellen bedingte Momentenbeschränkungen, im Allgemeinen, ein nicht-wohlgestelltes Problem dar, da sie die Lösung eines überidentifizierten, unendlich dimensionalen Gleichungssystems erfordern. Für den *unbedingten* Fall haben sich *empirische likelihood* (EL) Schätzer als breit anwendbare und effektive Werkzeuge zur Bewältigung von überidentifizierten Momentenbeschränkungsproblemen herausgestellt. Diese Methoden basieren darauf, ein Modell, zusammen mit einer Approximation der wahren Wahrscheinlichkeitsverteilung der Daten, durch Minimierung einer φ-Divergenz unter den Nebenbedingungen der Momentenbeschränkungen, zu lernen. Das Hauptziel dieser Arbeit ist, den Forschungsstand zum Lernen mit CMR voranzubringen, indem wir die Idee der empirischen likelihood Methode in mehrerer Hinsicht generalisieren. Zunächst präsentieren wir eine Erweiterung des Prinzips auf *bedingte* Momentenbeschränkungen, durch Verwendung einer funktionalen Formulierung, die in der Lage ist, moderne Modelle aus dem Feld des maschinellen Lernens zu inkorporieren. Dann erweitern wir die Methode auf alternative Abstandsbegriffe auf dem Raum der Wahrscheinlichkeitsverteilungen basierend auf Kernmethoden und optimalem Transport. Die resultierenden Schätzer zeigen verbesserte Eigenschaften bei kleinen Stichproben und erhöhte Robustheit gegenüber verfälschter Trainingsdaten beziehungsweise aktiv manipulierter Testdaten. Inspiriert durch die enge Relation zwischen empirischer likelihood Schätzung und der *verteilungsrobusten Optimierung* (engl. distributionally robust optimization (DRO)), präsentieren wir zum Abschluss eine Anwendung von Kernmethoden-basierter DRO auf Optimierungsprobleme mit stochastischen Nebenbedingungen.

Acknowledgements

First of all, I would like to thank Bernhard for giving me the opportunity to do a PhD in his amazing department. Thank you for your support, feedback and advice over the years and for encouraging me to pursue research in many different interesting directions. It has been a wonderful experience working alongside such a large group of motivated and talented people in the unparalleled environment you created.

I would also like to thank my second examiner Zeynep Akata and the rest of my thesis committee, Georg Martius and Jakob Macke. Thanks for completing this journey with me and for your unconditional dedication towards science.

Over the years I had the chance to work with an amazing set of collaborators, in particular I'd like to thank JJ, Manuel, Krikamol, Lars, Kai, Max-O and Yassine.

A special thanks I would like to extend to JJ for the great collaboration, mentor- and friendship that he provided during the second half of my PhD.

Another special thanks goes to Manuel Gomez-Rodriguez who taught me many important things in the first half of my PhD.

I am very grateful for all the great people and friends I met in the department that made the PhD an unforgettable experience, among many more and in random order, Luigi, Jun, Yassine, Giovanni, Alex, Julius, Yucen, Frederik, Max, Armin, Jonas W, Simon, Timmi, Nassim, Lennart and Jonas K.

Equally important for the completion of my PhD were all the friends I made in Tübingen outside the institute. In particular, Roza, who has been a great a friend for exciting activities and trips as much as for shared naps on the couch. Yassine, with whom I share so many memories in Tübingen and random places in the world, Jannis, Antonia, Leo, Anna and Fabian. You guys really helped me through the tougher parts of this journey.

Also outside of Tübingen, I could rely on unfailing support and friendship from a great number of people. Most importantly Max and the FRA4 community, Tim, my sisters Susanna and Katharina and, very importantly, my parents. Thanks for keeping me in your circle and for never stopping to count on me despite the countless things I missed!

Finally, I'd like to thank my old physics teacher Roland Pfeifer whose inspiring lessons caused my initial interest in science which eventually led me here.

Preface

The main body of this thesis is composed of four self-contained chapters based on the following publications:

Chapter 2
Functional Generalized Empirical Likelihood Estimation for Conditional Moment Restrictions
Heiner Kremer, Jia-Jie Zhu, Krikamol Muandet, Bernhard Schölkopf
International Conference for Machine Learning (ICML) 2022

Chapter 3
Estimation Beyond Data Reweighting: Kernel Method of Moments
Heiner Kremer, Yassine Nemmour, Bernhard Schölkopf, Jia-Jie Zhu
International Conference for Machine Learning (ICML) 2023

Chapter 4
Geometry-Aware Instrumental Variable Regression
Heiner Kremer, Bernhard Schölkopf
International Conference for Machine Learning (ICML) 2024

Chapter 5
Maximum Mean Discrepancy Distributionally Robust Nonlinear Chance-Constrained Optimization with Finite-Sample Guarantee
Yassine Nemmour*, **Heiner Kremer***, Bernhard Schölkopf, Jia-Jie Zhu
IEEE Conference for Decision and Control (CDC) 2022
*equal contribution

Every chapter provides the required background material and relevant literature. Proofs and additional information can be found in the corresponding appendices.

In the course of my PhD studies I have contributed to the following publications whose material is not included in this thesis:

Compact Holographic Sound Fields Enable Rapid One-step Assembly of Matter in 3D
K. Melde, **H. Kremer**, M. Shi, S. Seneca, C. Frey, I. Platzman, C. Degel, D. Schmitt, B. Schölkopf, P. Fischer
Science Advances (2023)

Non-Blind Glare-Removal for HDR Photography
M. van Bastelaer, **H. Kremer**, V. Volchkov, B. Schölkopf
International Conference on Computational Photography (ICCP) 2023

Listening to Bluetooth Beacons for Epidemic Risk Mitigation
G. Barthe, R. De Viti, P. Druschel, D. Garg, M. Gomez-Rodriguez, P. Ingo, **H. Kremer**, M. Lentz, L. Lorch, A. Mehta, B. Schölkopf (order alphabetical)
Scientific Reports (2022)

Quantifying the effects of contact tracing, testing, and containment measures in the presence of infection hotspots
L. Lorch, **H. Kremer**, W. Trouleau, S. Tsirtsis, A. Szanto, B. Schölkopf, M. Gomez-Rodriguez
ACM Transactions on Spatial Systems and Algorithms (2022)

Contents

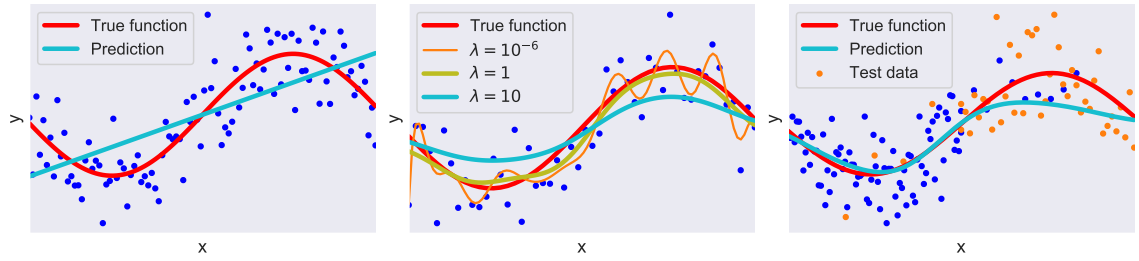
List of Figures

List of Tables

Chapter 1

Introduction

Randomness is an inherent property of nature and in fact most observable quantities in the world are derived from some kind of stochastic process [125]. The mathematical framework that describes how measurements are generated from such processes is called probability theory [55]. Given a process, it allows to argue about the probabilities and statistics of observations. In contrast, machine learning [167] can be seen as the reverse problem: Given observations generated from an unknown stochastic process, the goal is to learn a model that allows to reason about future observations. In many cases this comes in form of a function f that predicts a target Y given an input T. If the underlying data generating mechanism remains invariant across train and test time, an effective strategy to learn a model f that exhibits small average error on the test set is to find a model within a suitable hypothesis class $\mathcal{F}$ that minimizes the average error on the training set. This is the principle of empirical risk minimization (ERM) [167], one of the main concepts of machine learning. However, if the underlying process changes between the test and training datasets, or the *average* performance on the test set is not the main objective for the task at hand, empirical risk minimization might not yield the desired outcome and alternative learning paradigms have to be explored. For example, this is particularly relevant in the context of fairness. In the presence of training data with underrepresented groups of people one might want to avoid sacrificing performance on these groups for a slightly improved average performance. The property describing how well a model trained on a training dataset performs on a test dataset is called generalization. Generalization to test data can be hindered by many different factors some of which we will discuss in the following (see Figure 1.1). Firstly, in order to allow a model to approximate the true data generating process, the hypothesis class $\mathcal{F}$ needs to be sufficiently rich such that the true function or a sufficiently close approximation thereof is contained. Otherwise the model cannot accurately represent the relations in the data and thus cannot generalize to new data (Fig. 1.1 (a)). Secondly, with access only to a finite sample from the training distribution, flexible models might overfit to the noise in the training data, achieving small training error but large error on a different sample from the same distribution (Fig. 1.1 (b)). Avoiding this, is the purpose of regularization which restricts the complexity of the hypothesis class $\mathcal{F}$. Usually the trade-off between over- and underfitting, or in other words the bias-variance trade-off, is governed by a number of parameters which have to be determined as hyperparameters of the algorithm, e.g., by evaluating



(a) Model mis-specification (b) Finite data (c) Distribution shifts

Fig. 1.1 Visualization of different sources of generalization errors. Figure (a) shows the result of fitting non-linear data with a misspecified linear model class. Figure (b) demonstrates how with finite data the trade-off between over- and underfitting is controlled by regularization. The model is a kernel ridge regressor with different ridge penalties λ. Figure (c) demonstrates a scenario in which distributions of the train and test data do not agree. In this case, even with infinite training data, minimizing the average error will not yield the true function. This is the domain of robustness and causality and the focus of the present work.

the model on a separate validation dataset. Lastly, even with infinite training data and sufficiently expressive hypothesis classes, models trained to minimize the error on a sample from the training distribution might not generalize to test data if the test and training distributions differ from each other (Fig. 1.1 (c)). This is the problem of out-of-distribution (OOD) generalization and one of the application domains of the present work.

Achieving generalization to unknown test distributions without additional information is generally an impossible task. Making it tractable requires assumptions on either the expected shift, or the structure of the data generating process. The former corresponds to a purely statistical point of view and can be seen as the domain of distributionally robust optimization (DRO), the latter can be interpreted as the domain of causal inference. In the following we will introduce the relevant background and demonstrate how both viewpoints give rise to a problem formulation in terms of so-called *conditional moment restrictions* (CMR), the central problem setup considered in this work.

Distributionally Robust Optimization The idea behind distributionally robust optimization [12, 132] is to quantify the expected distribution shift by defining a so-called ambiguity set $\mathcal{P}$ of distributions in which one believes the test distribution to be contained. Often this is constructed as a metric ball around the empirical distribution of the training data. Instead of optimizing the average performance over the training data, DRO then optimizes the performance with respect to the worst-case distribution within the ambiguity set. If the test distribution is indeed contained in $\mathcal{P}$, the test error is upper bounded by the error on the worst case distribution. However, the construction of ambiguity sets as balls around the empirical distribution, bears several limitations. Firstly, the radius of the ball, which determines the magnitude of the distribution shifts the model will be robust against, usually has to be set as a hyperparameter of the training algorithm. The size has to be sufficiently large such that the test distribution is contained in $\mathcal{P}$. At the same time, if it is chosen too large, the solution becomes overly conservative as one robustifies against arbitrarily large shifts. In the extreme case of the radius going to infinity, DRO reduces to the robust optimization paradigm [24], which entirely

discards the information in the training data and optimizes the model with respect to the performance on the worst case realization of the random variables. As these values might occur with arbitrarily low probability, this might not lead to a satisfying average performance on the test distribution. Secondly, DRO methods generally account for shifts in the *joint* distribution over the inputs or *covariates* T and targets Y, i.e., they take into account any kind of changes between the train and test distributions up to some magnitude. Allowing arbitrary changes comes at the cost of having to restrict the ambiguity set to small shifts as otherwise the information in the training data gets quickly discarded. Oftentimes, however, the expected shift might be substantial in magnitude but one has prior knowledge about its structure. For example, a covariate shift describes the scenario in which the marginal distribution over the covariate T is changed between training and test time but the conditional distribution of the target Y given the covariate remains intact. With the notable exception of some works using φ-divergence based ambiguity sets [53], it is generally not clear how to incorporate such knowledge into the DRO framework. As a consequence, DRO is often used merely to robustify against finite sample errors in the independent identically distributed (IID) setting where train and test data are sampled from the same distribution [12, 54]. This domain, however, is already addressed by regularization and in fact, it has been shown that for certain ambiguity sets *joint* DRO is (asymptotically) equivalent to explicit regularization imposed on the model [52, 56]. Later, we will see that the concept of robustness against *structured* distribution shifts, e.g., covariate shifts, has a natural formulation in terms of conditional moment restriction.

Causal Inference Going beyond purely statistical arguments, causal inference [126, 79, 128] takes a different approach by imposing strong assumptions on the structure of the data generating process. In most cases, this comes in the form of a causal graph which entails the causal relations between the random variables (see e.g., Fig. 1.2). Additionally, often restrictive assumptions on the complexity of the involved function classes have to be imposed in order to grant identifiability of the true relations. In contrast to statistical learning, the goal of causal inference is not just to describe the statistical relations between variables but to infer their *causal* relations, i.e., the true data generating process. If this process is known entirely, distributional robustness, i.e., robustness against distribution shifts, is automatically entailed as long as the relevant causal mechanisms remain invariant across test and training environments. In this sense, distributionally robust optimization and, more broadly, OOD generalization can be seen as a subdomain of the even harder problem of causal inference [126].

The gold standard for inferring causal relations from an environment is the randomized controlled trial (RCT) [126]. In a RCT, one explores the effect of a treatment T on an outcome Y by assigning different values of T to statistically identical groups of individuals and observing the corresponding outcome Y. However, for a number of reasons including moral, financial or simply feasibility arguments it is often not possible to conduct an RCT in a given environment. For example, if one is interested in inferring the effect of smoking on a patient's health, it would be highly unethical to conduct a study where patients are assigned to take up smoking at random. In other situations one might have access to observational data but has no further access to the environment to conduct interventions by assigning treatments. The general goal of causal inference can be described as simulating an RCT based on observational data and assumptions on the data generating process [126].

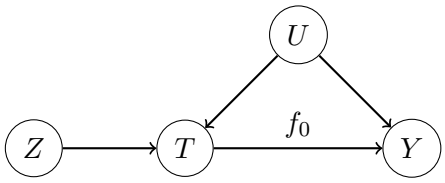


Fig. 1.2 Causal diagram for instrumental variable regression. Treatment T and outcome Y are affected by an unobserved confounder U. An instrumental variable Z, which i) causally affects T, ii) affects Y only through T, and, iii) is statistically independent of U, allows to learn the true causal effect f_0 independent of the confounder.

Instrumental Variable Regression One of the most successful practical applications of causal inference is instrumental variable (IV) regression [5, 4]. Suppose one is interested in inferring the effect of a treatment T on an outcome Y, where both treatment and outcome are affected by an unobserved confounder U. Consider the previous example in which one is interested in inferring the effect of smoking on a patient's health. In this case, one could imagine that there was a correlation between smoking and other lifestyle choices like consumption of alcohol/sugar or the level of physical activity, which might affect a person's health independent of their smoking behavior. Then collecting data $\{t_i, y_i\}_{i=1}^n$ and regressing Y on T, the learned function will be affected by the statistics of the lifestyle U and thus not represent the true causal effect of smoking T on health Y. While this is obviously unsatisfactory from the point of causal reasoning, even on a purely predictive level this may lead to inaccurate predictions if the distribution of U changes across test and training environments, e.g., if one employs the learned model in a place where alcohol consumption is strictly forbidden. In order to avoid incorporating the influence of the confounder into the learned model, one can take into account an additional random variable Z, a so-called instrument, which i) causally affects T, i.e., $Z \not\perp\!\!\!\perp T$, ii) only affects Y through T, i.e., $Z \perp\!\!\!\perp Y|T$ and iii) is statistically independent of U, i.e., $Z \perp\!\!\!\perp U$. In our example a valid instrument could be the price Z of tobacco products, which on average might affect a person's smoking behavior T but is independent of the lifestyle U and clearly does not affect a person's health Y other than through T. In mathematical terms, let f_0 describe the true unknown function describing the causal effect of T on Y and let the contribution of the confounder to Y be described by a zero mean random variable ϵ_U, then the outcomes Y are generated according to $Y = f_0(T) + \epsilon_U + \eta$, where η denotes zero mean Gaussian noise. Now, as $T \not\perp\!\!\!\perp U$ we have $f_0(t) \neq E[Y|T = t]$ and thus regressing Y on T leads to a biased estimate of f_0. However, for the instrumental variable Z we have $Z \perp\!\!\!\perp U$ and as $E[\epsilon_U] = 0$ it follows that

$$E[\epsilon_U|Z] = E[Y - f_0(T)|Z] = 0 \; P_Z\text{−a.s.}.$$

The last equation is to be understood as a probabilistic statement, where the condition needs to be satisfied almost surely (a.s.) with respect to the marginal distribution P_Z over the instruments Z. This means $E[Y - f_0(T)|Z = z] = 0$ needs to hold for all values of z that occur with non-zero probability. This is an example of a so-called *conditional moment restriction* (CMR) which will be the central problem formulation addressed in this work.

Conditional Moment Restrictions Conditional moment restrictions [165, 2] identify a model $f_0 \in \mathcal{F}$ by restricting the conditional expectation of a so-called moment function $\psi : \mathcal{T} \times \mathcal{Y} \times \mathcal{F} \to \mathbb{R}^m$, e.g., $\psi(T, Y; f) = Y - f(T)$ for IV regression, and take the form

$$\text{Find } \; f \in \mathcal{F} \quad \text{s.t.} \quad E[\psi(T, Y; f)|Z] = 0 \; P_Z\text{−a.s.}, \tag{1.1}$$

where the expectation is taken with respect to the (in practice unknown) population distribution over the random variables T and Y taking values in $\mathcal{T}$ and $\mathcal{Y}$ respectively. More generally, one can define a conditional moment model f_0 for a given error metric $\ell : \mathcal{T} \times \mathcal{Y} \times \mathcal{F} \to \mathbb{R}_+$ by simply choosing $\psi(t, y; f) = \nabla_f \ell(t, y; f)$. However, simpler choices like in the example above might be preferable as in the case of complex parametric models $\nabla\psi$ can be high-dimensional, which makes estimation harder in practice. Conditional moment restrictions provide a broad and versatile modelling framework which contains many problems as special cases. For example consider the trivial case $\{Z\} = \emptyset$ and a strictly convex loss function ℓ. Then the corresponding moment restrictions are exactly the first-order optimality conditions of the population risk, i.e.,

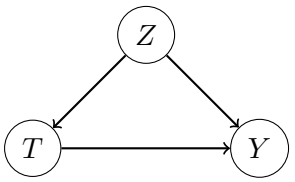
$$E[\nabla_f \ell(T, Y; f_0)] = 0 \quad \iff \quad f_0 = \operatorname*{arg\,min}_{f \in \mathcal{F}} E[\ell(T, Y; f)].$$

Generally, conditional moment restrictions enforce the moment function ψ to be zero across the distribution of the conditioning variable and in this sense can be used to enforce distributional robustness with respect to Z. This follows as for any model $f^* \in \mathcal{F}$ satisfying (1.1), we immediately have $E[\psi(T, Y; f^*)|Z] = 0 \; \tilde{P}_Z\text{−a.s.}$ for any shifted distribution $\tilde{P}_Z \ll P_Z$, where $\ll$ denotes absolute continuity. In words, this implies that a model trained on data from the training environment, will remain optimal in any test environment in which the distribution $\tilde{P}_Z$ of the variable Z does not exceed the support of the training distribution P_Z of Z. As an example, CMR can be used to enforce robustness against unsupervised covariate shifts. An unsupervised covariate shift describes the situation in which the distribution over the covariate T changes between train and test time in an unknown way, while leaving the conditional distribution of $Y|T$ unchanged. By conditioning on the covariate itself, i.e., using $Z = T$ and imposing $E[\nabla_f \ell(T, Y; f)|T] = 0 \; P_T\text{−a.s.}$, the model is forced to exhibit uniform performance over the values of the covariate which occur with non-zero probability. Therefore, shifts in the marginal distribution over T do not affect the performance as long as the support remains unchanged.

As demonstrated by these examples, the CMR formulation can be interpreted as an implicit stratification of the training data by a continuous variable Z. For discrete variables Z, stratification is a common strategy to induce robustness against shifts in the distribution of this variable. For example, suppose a treatment T has different effects on the outcome Y depending on a persons sex Z, and additionally there is an imbalance in the training data in the sense that it contains more male than female patients. Then, an effective way to force the model not to sacrifice performance on the female population in favor of an improved overall performance is to stratify the data by sex Z and minimize a composite objective which combines the average error over the male and female populations. In the context of causal inference this kind of stratification plays an important role to resolve contradictory conclusions like Simpson's paradox. Figure 1.3 demonstrates this phenomenon

	$T = A$	$T = B$
Small stones	**93%** $(81/87)$	87% $(234/270)$
Large stones	**73%** $(192/263)$	69% $(55/80)$
Total	78% $(273/350)$	**83%** $(289/350)$

(a) Treatment Success Rates



(b) Causal Graph

Fig. 1.3 Simpson's paradox in kidney stone surgery data adopted from Julious and Mullee [83]. Figure (a) shows the number of successful and total treatments per treatment and size of the kidney stones. Despite being superior for both, small and large stones, treatment A seems inferior if the data is pooled together and not stratified by the size of the stones. This is a consequence of the fact that treatments are not assigned at random but patients with more severe stones—and thus generally worse chances of successful treatment—are more often assigned the better treatment A. Figure (b) visualizes the corresponding data generating process.

using the example of kidney stone treatments adopted from Julious and Mullee [83]. In this example there are two treatments A and B and the patients are divided into groups of large and small kidney stones described by Z respectively. Figure 1.3 (a) visualizes the success rates per group and treatment. One observes that treatment A is superior for both, large and small kidney stones. However, if the patient data is pooled together, it seems like treatment B is more successful overall. This is due to the fact that treatments are not assigned at random but doctors take into account the size of the stones Z, as visualized in the causal diagram in Figure 1.3 (b). In the case of large stones either treatment is less likely to succeed—$247/343 \approx 72\%$ success rate compared to $315/357 \approx 88\%$ for small stones—but these severe cases are more often assigned the better treatment A and as a result it falsely seems like treatment A generally has lower success chances. This highlights the importance of stratifying with respect to relevant factors to avoid drawing incorrect conclusions from the data. In this example, stratification was possible due to the relevant factor Z being a categorical variable distinguishing between large and small stones. If, however, the variable Z is of continuous nature, e.g., some numerical health parameter, then it is a priori not clear how to perform such a stratification. One possibility to achieve this is to discretize the variable, which requires an arbitrary choice of discretization scheme. In contrast, CMR provides a natural way to enforce an optimality condition almost surely with respect to the distribution of Z, i.e., separately for each value of the confounding variable which occurs with non-zero probability and thus without resorting to discretization.

These examples demonstrate that conditional moment restrictions connect the fields of distributional robustness in causal inference. Through the lens of causal inference, the assumptions on the data generating process often bear a natural formulation in terms of conditional moment restrictions, e.g., in the case of instrumental variable regression. In contrast, through the lens of DRO, even without any knowledge about the data generating process, one can use CMR to enforce robustness with respect to shifts in the marginal distribution of some variable Z, e.g., in the case of covariate shifts.

Learning with Conditional Moment Restrictions Conditional moment restrictions phrase a problem in terms of a system of stochastic equations (1.1) involving a *conditional* expectation

over the unknown *population distribution* over (T, Y). In practice, however, one usually only has access to a *sample* from the *joint* distribution P_0 over (T, Y, Z) with empirical distribution $\hat{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(x_i,y_i,z_i)}$, where $\delta_{(x_i,y_i,z_i)}$ denotes a Dirac measure centered at (x_i, y_i, z_i). To address the conditional part without resorting to costly conditional density estimation, usually the first step is to write the CMR (1.1) in terms of an equivalent *variational* moment restriction [20],

$$E[\psi(T, Y; f)|Z] = 0 \;\; P_Z\text{−a.s.} \quad \iff \quad E[\psi(T, Y; f)^T h(Z)] = 0 \;\; \forall h \in \mathcal{H}, \tag{1.2}$$

where $\mathcal{H} \subseteq L^2(P_Z)$ is a sufficiently rich function space, e.g., the space $L^2(P_Z)$ of square integrable functions $\mathcal{Z} \to \mathbb{R}^m$ with measure P_Z itself, or the reproducing kernel Hilbert space of an integrally strictly positive definite universal kernel (e.g., Gaussian kernel) [92]. Rewriting the CMR as (1.2) transformed the intractable conditional expectation into a continuum of unconditional restrictions, one for each function h in the infinite dimensional function space $\mathcal{H}$. The second problem is the presence of the population expectation in (1.2) which has to be approximated with the sample. While by definition the population moment restrictions identify the function of interest f_0 and thus can be satisfied, this is not necessarily the case for the empirical counterpart of (1.2), i.e., there might exist $h \in \mathcal{H}$ for which $E_{\hat{P}_n}[\psi(T, Y; f_0)^T h(Z)] \neq 0$. Moreover, (1.2) involves an infinite number of constraints which might not be satisfied for any $f \in \mathcal{F}$, especially in the case of parametric function classes $\mathcal{F}$. In this so-called over-identified case, in which there are more moment restrictions than parameters, the most common approach is to resort to the generalized method of moments (GMM) estimator of Hansen [69]. Instead of enforcing the empirical version of the moment restrictions exactly, GMM relaxes the problem into a minimization of a quadratic form of the empirical moment restrictions. The original GMM estimator for unconditional moment restrictions takes the form

$$f^{\text{GMM}} = \operatorname*{arg\,min}_{f \in \mathcal{F}} E_{\hat{P}_n}[\psi(T, Y; f)] \left[\widehat{\Omega}(\tilde{f})\right]^{-1} E_{\hat{P}_n}[\psi(T, Y; f)], \tag{1.3}$$

where $\widehat{\Omega}(\tilde{f}) = E_{\hat{P}_n}[\psi(T, Y; \tilde{f})\psi(T, Y; \tilde{f})^T] \in \mathbb{R}^{m\times m}$ denotes the empirical covariance matrix of the moment function evaluated at an initial parameter estimate $\tilde{f} \in \mathcal{F}$. Generalizations of this approach to address conditional moment restrictions have been proposed in several works [26, 15, 101, 14].

Empirical Likelihood Estimation Empirical likelihood (EL) estimation [122, 121, 130] has emerged as a powerful alternative to GMM for addressing over-identified moment restrictions problems. Originally proposed by Owen [122, 121] as a tool to construct confidence intervals, it provides a non-parametric version of maximum likelihood estimation by means of minimizing a φ-divergence under the moment restrictions. The idea behind EL estimation is that while it may not be possible to satisfy the *empirical* moment restrictions exactly for any $f \in \mathcal{F}$, the true function f_0 satisfies the *population* moment restrictions by definition. As the empirical distribution $\hat{P}_n$ converges weakly to the population distribution P_0, the population distribution will be contained in a shrinking neighborhood of $\hat{P}_n$ as the sample size grows. Therefore, EL estimators approximate the population distribution P_0 by seeking the closest distribution to the empirical one, for which the moment restrictions can be satisfied. This concept is visualized in Figure 1.4. The (generalized) empirical likelihood estimator

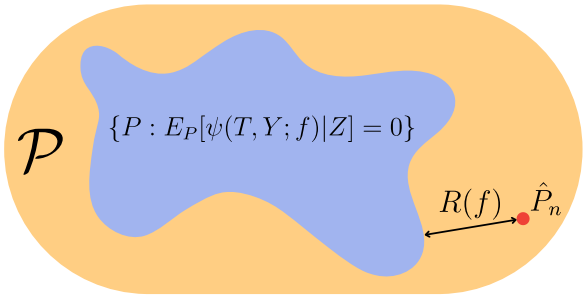


Fig. 1.4 Empirical Likelihood Estimation. For each f the likelihood profile $R(f)$ denotes the minimum distance between the empirical distribution $\hat{P}_n$ and the set of distributions for which the moment restrictions can be satisfied. The empirical likelihood solution is the minimizer of the likelihood profile over $f \in \mathcal{F}$.

takes the form

$$f^{\text{GEL}} = \arg\min_{f \in \mathcal{F}} R(f) := \left\{ \min_{P \in \mathcal{P}} D_\varphi(P||\hat{P}_n) \quad \text{s.t.} \quad E_P[\psi(T, Y; f)] = 0 \right\} \tag{1.4}$$

where $D_\varphi(P||\hat{P}_n) = \int \varphi\left(\frac{\mathrm{d}P}{\mathrm{d}\hat{P}_n}\right) \mathrm{d}\hat{P}_n$ denotes the φ-divergence between P and $\hat{P}_n$ and $\mathcal{P} = \{P \ll \hat{P}_n : E_P[1] = 1\}$ is the set of distributions absolutely continuous with respect to the empirical one. While the original EL estimator used $\varphi(\cdot) = \log(\cdot)$, generalizations to other φ-divergences have been proposed by Kitamura and Stutzer [87], Smith [151] and Imbens et al. [80], which are collectively referred to as generalized empirical likelihood (GEL) estimators. The method provides a broad framework which contains a variant of the generalized method of moments as a special case [118]. In fact, in the seminal paper by Newey and Smith [118] it was shown that alternative estimators from the GEL family admit smaller higher-order biases than GMM and thus might be preferable in practice. Extensions of GEL to the conditional case have been proposed by several authors from the econometrics community [88, 120, 30, 27], which usually rely on a hand-chosen set of instrument functions or approximations of the continuum of moment restrictions via basis function expansions of L^2. In contrast, despite its great potential, empirical likelihood has so far attracted comparatively little attention in the machine learning community, where there have been significant developments in terms of expressive and scalable models as well as efficient optimization tools. Introducing these new developments and ideas into the empirical likelihood framework will be our overarching goal.

Overview of the Thesis

The main focus of this thesis is to advance the state-of-the-art in conditional moment restriction estimation by extending the empirical likelihood framework along several dimensions. The methods we present are based on three different paradigms as visualized in Figure 1.5.

Chapter 2 sets the stage by formalizing an extension of the GEL framework to *conditional* moment restrictions, which provides the starting point for all subsequent methods. We present the Functional Generalized Empirical Likelihood (FGEL) estimator, which allows combining the classical GEL

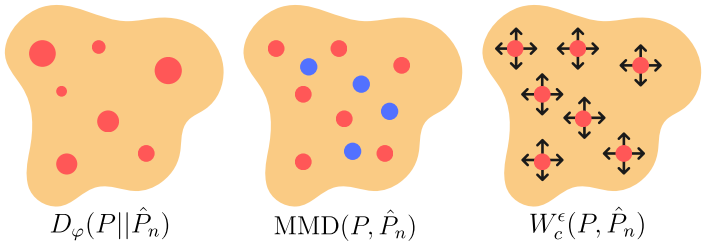


Fig. 1.5 Paradigms to approximate P_0 from samples in the GEL framework. The data is represented by the red dots with their size corresponding the respective weight. In Chapter 2 we introduce a φ-divergence-based estimators which approximates P_0 by reweighting the sample (left). In Chapter 3 we present an MMD-based estimator that allows for sampling additional data points (blue dots). Finally, Chapter 4 derives an optimal transport-based estimator which allows to move around the data points and thereby take into account the geometry of the data space (right).

method with modern machine learning approaches to train conditional moment models. We provide neural network and kernel-based implementations, which perform on par with related GMM-based methods. The GEL estimator minimizes the likelihood profile $R(f)$, i.e., the minimal distance between the empirical distribution and the set of distributions for which the moment restrictions can be satisfied. However, by measuring this distance with a φ-divergence one reduces the set of possible distributions to reweightings of the sample as $D_\varphi(P||\hat{P}_n) = \infty$, whenever $P \not\ll \hat{P}_n$. This is unsatisfactory in multiple ways. Firstly, while for large samples reweightings might provide enough flexibility to approximate the population distribution sufficiently well, in the small sample regime this might not be the case. Secondly, reweightings are blind towards the geometry of the data space, i.e., small perturbations of the data might lead to vastly different results but reweightings have no means of knowing which data points are vulnerable to such perturbations or which might have been actively poisoned. We address both these issues in Chapters 3 and 4 by generalizing the framework to alternative distance notions on the space of probability distributions.

Chapter 3 introduces the Kernel Method of Moments (KMM), an empirical likelihood-type estimator based on maximum-mean discrepancy (MMD) [65]. In contrast to φ-divergences, MMD is capable of comparing measures with different support and thus in principle allows for seeking approximations of P_0 within the whole space of (continuous) distributions. We utilize entropy regularization to arrive at a relaxed form of the dual MMD-profile (the MMD-equivalent of the likelihood profile $R(f)$) that can be computed with stochastic gradient methods. In practice, our formulation allows to sample and reweigh additional data points to find more fine-grained approximations of P_0 in a principled way. The resulting estimator shows favorable small sample properties compared to φ-divergence based methods.

Chapter 4 presents an empirical likelihood-type estimator based on a regularized optimal transport distance, which we term the Sinkhorn Method of Moments (SMM). Similar to KMM, the method allows to go beyond reweightings of the data but instead of sampling additional support points, it approximates P_0 by moving the training data in the data space. The estimator results from a leading order expansion of the dual Sinkhorn profile $R(f)$ and involves higher-order derivative information with respect to the data. This makes it aware of how the learning signal changes around the data

points and thus allows it to take into account the geometry of the data space. In particular, this property enables SMM to detect data points, which, when perturbed slightly, lead to drastically different outcomes, i.e., points vulnerable to adversarial attacks or data poisoning. Compared to related estimators, SMM provides stronger robustness against corrupted data at training time and adversarial attacks at inference time respectively, without sacrificing performance in standard IV settings.

Finally, branching out from empirical likelihood estimation, we demonstrate a more traditional application of distributionally robust optimization in Chapter 5. Empirical likelihood and distributionally robust optimization are closely related by a duality relationship and under certain conditions the Lagrangians of both problems agree up to a minus sign. Therefore, methods developed in one area can often be applied to the other. As an example, we show how, analogous to the Kernel Method of Moments, one can use MMD ambiguity sets for distributional robustness in chance-constrained programming.

Chapter 2

Functional Generalized Empirical Likelihood Estimation*

Important problems in causal inference, economics, and, more generally, robust machine learning can be expressed as conditional moment restrictions, but estimation becomes challenging as it requires solving a continuum of unconditional moment restrictions. Previous works addressed this problem by extending the generalized method of moments (GMM) to continuum moment restrictions. In contrast, generalized empirical likelihood (GEL) provides a more general framework and has been shown to enjoy favorable small-sample properties compared to GMM-based estimators. To benefit from recent developments in machine learning, we provide a functional reformulation of GEL in which arbitrary models can be leveraged. Motivated by a dual formulation of the resulting infinite dimensional optimization problem, we devise a practical method and explore its asymptotic properties. Finally, we provide kernel- and neural network-based implementations of the estimator, which achieve state-of-the-art empirical performance on two conditional moment restriction problems.

*Based on *Functional Generalized Empirical Likelihood Estimation for Conditional Moment Restrictions* [92]. Heiner Kremer, Jia-Jie Zhu, Krikamol Muandet, Bernhard Schölkopf. International Conference for Machine Learning (ICML) 2022

Declaration

This chapter is based on an updated version of the published manuscript:

> *Functional Generalized Empirical Likelihood Estimation for Conditional Moment Restrictions*
> Heiner Kremer, Jia-Jie Zhu, Krikamol Muandet, Bernhard Schölkopf
> International Conference for Machine Learning (ICML) 2022

Compared to the published version we provide an improved convergence rate of our estimator in explicit form and add a theorem on its efficiency.

Author contributions (CRediT)

Heiner Kremer: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization
Jia-Jie Zhu: Conceptualization, Writing - Review & Editing, Supervision
Krikamol Muandet: Conceptualization, Writing - Review & Editing, Supervision
Bernhard Schölkopf: Conceptualization, Writing - Review & Editing, Supervision, Resources

2.1 Introduction

Moment restrictions identify a parameter of interest by restricting the expectation value of so-called moment functions, which depend on the parameter and random variables representing the underlying noisy data generating process. Important problems in causal inference, economics, and generally robust machine learning can be cast in this form [116, 2, 14, 50]. Particularly challenging are problems formulated as *conditional* moment restrictions (CMR), which constrain the conditional expectation of the moment function. Such problems appear, e.g., in instrumental variable (IV) regression [117, 4], where the expectation of the residual of the prediction conditioned on so-called instruments is restricted to be zero. Other applications are policy learning [13] and off-policy evaluation in reinforcement learning [84, 16, 34] as well as double/debiased machine learning [36–38].

As conditional moment restrictions are difficult to handle directly, a common approach is to transform them into an infinite number of corresponding unconditional moment restrictions [20]. Generalizing the corresponding estimation methods from the finite-dimensional case to the infinite-dimensional case is an active area of research [26, 28, 30, 27, 109, 14, 179].

One of the most popular approaches to learning with moment restrictions is Hansen's celebrated generalized method of moments (GMM) [69]. In order to improve the small sample properties of GMM estimators, alternative methods have been proposed and are generally known as generalized empirical likelihood (GEL) estimators [151, 153, 118]. GEL generalizes the original empirical likelihood framework developed by Owen [122, 121] and Qin and Lawless [130] to different divergence functions and contains many related estimators as special cases. While closely related to GMM, the estimators from the GEL family have been shown theoretically to exhibit smaller higher-order-biases than those of GMM [118] and therefore promise to have favorable small sample properties. With increasing number of over-identifying restrictions, i.e., when the number of restrictions exceeds the number of parameters, this advantage has been shown theoretically to become more significant [115, 51]. Therefore, we expect the framework to be particularly suited for the case of infinitely many restrictions. We leverage this potential for conditional moment restrictions by developing the theoretical foundation for a GEL framework with continua of moment restrictions.

Our contributions First, we extend the GEL framework to conditional moment restrictions by generalizing it to *functional* moment restrictions. Second, building on a result from infinite optimization, we derive a dual form which allows us to employ modern machine learning models in the GEL context. This generalizes existing results not only to functional moment restrictions but also to general φ-divergences beyond the Cressie-Reed family. Third, we provide the asymptotic properties of our estimator, showing that it is consistent for functional moment restrictions. Then, we show how the result for the functional case can be translated to the conditional case yielding an efficient estimator for conditional moment restrictions. Finally, we discuss the relation to existing methods and provide experimental results.

Compared to previous extensions of GEL [88, 160, 30, 27], our approach combines the idea of a continuum generalization of GEL [30, 27] with the flexibility of machine learning models such as neural networks and kernel methods. Our general framework contains related estimators such as

a one-step/continuous updating version of the variational method of moments (VMM) estimator [14] as special cases. In contrast to VMM, our method allows the use of divergences other than the χ^2-divergence.

The remainder of this paper is organized as follows. Section 2.2 introduces the method of moments framework [67] and two popular relaxations. Section 2.3 presents our main contributions, the theoretical development of our FGEL estimator, followed by experimental results in Section 2.4. Finally, we discuss related works in Section 2.5.

Compared to the initial conference version, this paper has been updated using recent results of Kremer et al. [93] to derive an explicit convergence rate for the estimator, as well as to prove its semi-parametric efficiency.

2.2 Learning with Moment Restrictions

Let X be a random variable taking values in $\mathcal{X} \subseteq \mathbb{R}^r$ with distribution P_0 and let $\psi : \mathcal{X} \times \Theta \to \mathbb{R}^m$ denote a vector of m functions, the so-called moment functions, with parameters $\theta \in \Theta \subset \mathbb{R}^p$. We denote with $E_P[\cdot]$ the expectation over all random variables that are not conditioned on with respect to a distribution P and refer to the population distribution P_0 whenever we omit the subscript. Assume that there exists a unique parameter $\theta_0 \in \Theta$ such that $E[\psi(X;\theta_0)] = 0$. For instance, $E[X - \theta_0] = 0$ characterizes the mean of P_0. Our goal is to estimate θ_0 based on a sample $\{x_i\}_{i=1}^n$ from P_0. The corresponding empirical moment restrictions read

$$E_{\hat{P}_n}[\psi(X;\theta)] = 0, \quad \theta \in \Theta, \tag{2.1}$$

where $\hat{P}_n = \sum_{i=1}^n \frac{1}{n}\delta_{x_i}$ is the empirical distribution. This is a system of m estimating equations for p parameters which can be fulfilled exactly as long as $m \leq p$. For example, $E_{\hat{P}_n}[X - \theta] = 0$ gives $\theta = \frac{1}{n}\sum_{i=1}^n x_i$ as an empirical estimate of the mean of P_0. However, in the over-identified case, i.e., when the number of non-redundant moment restrictions exceeds the number of parameters ($m > p$), it is generally impossible to fulfill all moment restrictions (2.1) exactly. To obtain a feasible problem, the constraints (2.1) need to be relaxed. Below we discuss two popular approaches, namely, the generalized method of moments [69] and maximum (generalized) empirical likelihood estimation [122, 121, 130].

Generalized Method of Moments (GMM) The generalized method of moments relaxes the constraint (2.1) into a minimization of a quadratic form of the empirical expectation over the moment functions, i.e., $\theta_W^{\mathrm{GMM}} = \arg\min_{\theta\in\Theta} \hat{\psi}(\theta)^T W \hat{\psi}(\theta)$, where $\hat{\psi}(\theta) := E_{\hat{P}_n}[\psi(X;\theta)]$ and $W \in \mathbb{R}^{m\times m}$ denotes the so-called weighting matrix. Asymptotic normality theory shows that an efficient estimator, i.e., an estimator with minimal asymptotic variance among the class of GMM estimators, is obtained by choosing W as the inverse covariance matrix of the moment functions, $W = \widehat{\Omega}_\theta^{-1}$, where $\widehat{\Omega}_\theta := E_{\hat{P}_n}[\psi(X;\theta)\psi(X;\theta)^T]$, which itself a function of θ [69]. The resulting estimator, i.e.,

$$\theta^{\mathrm{CUE}} = \underset{\theta\in\Theta}{\arg\min}\ \hat{\psi}(\theta)^T \widehat{\Omega}_\theta^{-1} \hat{\psi}(\theta), \tag{2.2}$$

is the continuous updating estimator (CUE) of Hansen et al. [70] which results from a non-convex optimization problem and can exhibit unfavorable convergence properties if $\widehat{\Omega}_\theta$ is ill-conditioned [67]. Therefore, one often resorts to a 2-step procedure: first, an inefficient but consistent estimate $\tilde{\theta}$ of θ_0 is obtained, e.g., by setting $W = I$. Second, this estimate is used to compute $\widehat{\Omega}_{\tilde{\theta}}^{-1}$ which is kept fixed during the second optimization step. This yields the so-called optimally weighted GMM estimator [69],

$$\theta^{\mathrm{OWGMM}} = \operatorname*{arg\,min}_{\theta\in\Theta} \; \hat{\psi}(\theta)^T \widehat{\Omega}_{\tilde{\theta}}^{-1} \hat{\psi}(\theta). \tag{2.3}$$

A more in-depth exposition of the GMM framework can be found in Hall [67].

Generalized Empirical Likelihood (GEL) The empirical likelihood framework [122, 121, 130] relaxes the restrictions (2.1) by requiring $E_P[\psi(X;\theta)] = 0$ to be fulfilled exactly but allowing the distribution P to deviate from the empirical distribution $\hat{P}_n$. For a continuous function $\varphi : \mathbb{R} \to \mathbb{R}$ we define the φ-divergence between distributions P and Q as $D_\varphi(P||Q) = \int \varphi\big(\frac{\mathrm{d}P}{\mathrm{d}Q}\big)\mathrm{d}Q$, where $\frac{\mathrm{d}P}{\mathrm{d}Q}$ denotes the Radon-Nikodym derivative of P with respect to Q. Then, we can define the profile divergence with respect to this φ-divergence as

$$R(\theta) = \inf_{P \ll \hat{P}_n} D_\varphi(P||\hat{P}_n) \quad \text{s.t.} \quad E_P[\psi(X;\theta)] = 0, \quad E_P[1] = 1, \tag{2.4}$$

where $P \ll \hat{P}_n$ describes the set of positive measures P that are absolutely continuous with respect to the empirical distribution $\hat{P}_n$. In other words, P describes multinomial distributions on the sample, i.e., re-weightings of the data points. The maximum empirical likelihood estimator (MELE) for θ is then given by $\theta^{\mathrm{EL}} = \arg\min_{\theta\in\Theta} R(\theta)$. The framework, originally proposed as empirical likelihood by Owen for the case $\varphi(p) = -2\log(p)$, has been generalized to other divergence measures for which it is known as minimum discrepancy (MD) [41] or generalized empirical likelihood [151, 152]. The latter corresponds to its dual formulation. It contains many related estimators as special cases. For example, by choosing the function φ from the Cressie-Read family of non-parametric discrepancy measures [43],

$$\varphi_\gamma(p) = \frac{1}{\gamma(\gamma+1)}\left(p^{\gamma+1} - 1\right), \tag{2.5}$$

one retrieves the CUE for $\gamma = 1$ [118], the exponential tilting estimator for $\gamma \to 0$ [87] and finally the original empirical likelihood estimator for $\gamma \to -1$ [130]. Detailed exposition of the GEL framework can be found in Smith [151] and Owen [123].

2.3 Functional Generalized Empirical Likelihood

In this work, we are concerned with problems that can be expressed by infinitely many moment restrictions, especially those that arise from *conditional* moment restrictions (CMR) [116, 2] of the

form

$$E[\psi(X;\theta_0)|Z] = 0, \quad P_Z\text{-a.s.}, \tag{2.6}$$

where $Z \in \mathcal{Z}$ is an additional random variable with marginal distribution P_Z. By the law of iterated expectation, the CMR (2.6) can be expressed in terms of infinitely many unconditional moment restrictions [20]

$$E[\psi(X;\theta_0)^T h(Z)] = 0, \quad \forall h \in \mathcal{H}, \tag{2.7}$$

where $\mathcal{H}$ denotes a space of measurable functions $h : \mathcal{Z} \to \mathbb{R}^m$, i.e., the space of square integrable functions $L^2(P_Z)$. As (2.7) has to hold for all functions in $\mathcal{H}$, this implies an uncountable infinite number, i.e., a continuum, of moment restrictions ($m = \infty$). For example, the instrumental variable regression problem can be described by a CMR via $E[Y - f(X;\theta_0)|Z] = 0$ where Z is an instrumental variable and $\theta \in \Theta$ parameterizes a function $f : \mathcal{X} \to \mathcal{Y}$. Motivated by this example, in the following, we will refer to Z and h as instrument and instrument function, respectively, in the context of general CMR.

2.3.1 Our Method

Maximum empirical likelihood estimation is based on minimizing a profile divergence $R : \Theta \to \mathbb{R}$ over a parameter space Θ. Let $\mathcal{P} := \{P \ll \hat{P}_n : E_P[1] = 1\}$ denote the set of distributions that are absolutely continuous with respect to the empirical distribution. For conditional moment restrictions of the form (2.6), we can define the profile divergence as

$$R(\theta) := \min_{P \in \mathcal{P}} D_\varphi(P||\hat{P}_n) \quad \text{s.t.} \quad E_P[\psi(X;\theta)|Z] = 0 \quad P_Z\text{-a.s.}, \tag{2.8}$$

where D_φ is defined in terms of the φ-divergence (see Table 2.1). Let $\mathcal{H}$ be a sufficiently large Hilbert space of functions such that (2.7) implies (2.6) and let $\mathcal{H}_1 = \{h \in \mathcal{H} : \|h\| \leq 1\}$ denote the unit ball in $\mathcal{H}$. Let $\mathcal{H}^*$ be the corresponding dual space of functionals $\mathcal{H} \to \mathbb{R}$ equipped with the dual norm $\|\cdot\|_{\mathcal{H}^*}$ defined for $\Psi \in \mathcal{H}^*$ as $\|\Psi\|_{\mathcal{H}^*} = \sup_{h \in \mathcal{H}_1} \Psi(h)$. Then, we can define the *moment functional*, a statistical functional $\Psi(X, Z;\theta) \in \mathcal{H}^*$, as

$$\begin{aligned} \Psi(X, Z;\theta) : \ \mathcal{H} &\to \ \mathbb{R} \\ h &\mapsto \ \Psi(X, Z;\theta)(h) = \psi(X;\theta)^T h(Z), \end{aligned}$$

which can be seen as a weighted evaluation functional with respect to the conditioning variable Z. With this definition, we can express (2.7) as the functional constraint $\|E_{P_0}[\Psi(X, Z;\theta_0)]\|_{\mathcal{H}^*} = 0$. The computation of the profile likelihood thus becomes a *functionally-constrained* optimization problem

$$R(\theta) = \inf_{P \in \mathcal{P}} D_\varphi(P||\hat{P}_n) \quad \text{s.t.} \quad \|E_P[\Psi(X, Z;\theta_0)]\|_{\mathcal{H}^*} = 0. \tag{2.9}$$

The FGEL problem arises from the dual formulation of (2.9). For the case of finite dimensional moment restrictions, the duality relationship has been extensively explored by numerous works [151, 153, 88, 118]. However, as shown by Borwein [23] these duality results do not carry over to infinite dimensional restrictions. Following the approach of Borwein [23] and Carrasco and Kotchoni [27], we define a relaxed version of the functionally constrained profile likelihood (2.9) with relaxation parameter $\lambda > 0$ as

$$R_\lambda(\theta) := \inf_{P \in \mathcal{P}} D_\varphi(P||\hat{P}_n) \quad \text{s.t.} \quad \|E_P[\Psi(X, Z; \theta)]\|_{\mathcal{H}^*} \leq \lambda. \tag{2.10}$$

With this relaxation, a constraint qualification condition holds and (2.10) admits a strongly dual form as formalized in the following theorem.

Theorem 2.1. *Let $\varphi^*(v) = \sup_{p \in \mathbb{R}^n} \langle v, p \rangle - \varphi(p)$ denote the Legendre-Fenchel conjugate function of a strongly convex function φ. Then the problem*

$$R_\lambda(\theta) = \inf_{p \in \mathbb{R}^n} \sum_{i=1}^n \frac{1}{n} f(np_i) \quad \text{s.t.} \quad \left\| \frac{1}{n} \sum_{i=1}^n p_i \Psi(x_i, z_i; \theta) \right\|_{\mathcal{H}^*} \leq \lambda, \ \sum_{i=1}^n p_i = 1$$

admits the dual form

$$R_\lambda(\theta) = \sup_{\substack{h \in \mathcal{H} \\ \mu \in \mathbb{R}}} \mu - \frac{1}{n} \sum_{i=1}^n \varphi^* \left(\Psi(x_i, z_i; \theta)(h) + \mu \right) - \lambda \|h\|_{\mathcal{H}} \tag{2.11}$$

and strong duality holds between these formulations. Moreover, the unique minimizer of the primal problem is given by

$$p_i = \left(\frac{d}{dv} \varphi^* \right) \left(\Psi(x_i, z_i; \theta)(\hat{h}) + \hat{\mu} \right),$$

where $\hat{h}$, $\hat{\mu}$ are any solutions of the dual problem. Moreover, as $\lambda \to 0$, $R_\lambda(\theta) \to R(\theta)$.

Remark 2.2. *Theorem 2.1 can be seen as a generalization of the duality result of Newey and Smith [118] not only to functional-valued moment restrictions but also to general strongly convex divergence functions beyond the Cressie-Reed family.*

Equation (2.11) provides a regularized functional generalization of the profile divergence. Based on this result, we define our functional generalized empirical likelihood estimator by making two modifications: first, we substitute the norm term in (2.11) for a differentiable quadratic version. This modification is solely to simplify the analysis. Later we will choose the regularization parameter to be $\lambda_n = o_p(1)$ and find that $\|h\|_{\mathcal{H}} = o_p(1)$. Hence we can always find a $\chi' > 0$ and $\lambda'_n = O_p(n^{-\chi'})$ such that $\lambda_n/2\|h\|^2 \to 0$ and $\lambda'_n\|h\| \to 0$ at the same rate which implies that the formulations are asymptotically equivalent. Second, we drop the Lagrange parameter μ corresponding to the normalization constraint $\sum_{i=1}^n p_i = 1$. This is motivated by several observations. From a theoretical point of view, it simplifies the problem at no cost as we will show later in Theorems 2.8-2.10 that setting $\mu = 0$ still yields a consistent and efficient estimator. From a practical aspect it facilitates the implementation of the estimator using stochastic gradient methods. To see this, consider exemplarily

Table 2.1 Common choices for the φ-divergence and the corresponding convex conjugate $\varphi^*(v) = \sup_p p^T v - \varphi(p)$ and its domain. A GEL function ϕ can be defined for each φ-divergence as $\phi(v) = -\varphi^*(v)$.

$\varphi(p)$	$\varphi^*(v)$	$\mathrm{dom}(\varphi^*)$
$\frac{1}{2}(p-1)^2$	$\frac{1}{2}(1+v)^2$	$\mathbb{R}$
$-\log(p)$	$\log(1-v)$	$\big(-\infty, 1-\frac{1}{n}\big]$
$p\log(p)$	e^v	$\mathbb{R}$

the FGEL estimator with χ^2-divergence. Then we can carry out the supremum over $\mu \in \mathbb{R}$ in closed form and obtain

$$R_\lambda^{\chi^2}(\theta) = \sup_h E_{\hat{P}_n}[\psi(X;\theta)^T h(Z)] - \frac{1}{2} E_{\hat{P}_n}\left[\Big(\psi(X;\theta)^T h(Z) - E_{\hat{P}_n}[\psi(X;\theta)^T h(Z)] - 1\Big)^2\right] - \lambda\|h\|_{\mathcal{H}}.$$

This contains two expectation operators combined in a non-linear way, which generally leads to biased gradient estimates and thus renders mini-batch stochastic gradient descent optimization complicated. In addition, by setting $\mu = 0$, our estimator has the same form as the finite dimensional GEL estimator proposed by Smith [151] and Kitamura and Stutzer [87] with the difference of involving an optimization over functions $h \in \mathcal{H}$ and an additional regularization term for h. With these modifications, we define our FGEL estimator as follows.

Definition 2.3. *Let $V \subseteq \mathbb{R}$ be an open interval containing zero and $\phi : V \to \mathbb{R}$ be a twice differentiable concave function with first and second derivatives $\phi_1(0) \neq 0$ and $\phi_2(0) < 0$. Then we define the empirical FGEL objective $G : \Theta \times \widehat{\mathcal{H}}(\theta) \to \mathbb{R}$ as*

$$G_{\lambda_n}(\theta, h) := \frac{1}{n}\sum_{i=1}^{n} \phi\left(\Psi(x_i, z_i; \theta)(h)\right) - \frac{\lambda_n}{2}\|h\|_{\mathcal{H}}^2, \tag{2.12}$$

where $\Psi(x_i, z_i; \theta)(h) = \psi(x_i;\theta)^T h(z_i)$ and $\widehat{\mathcal{H}}(\theta) := \{h \in \mathcal{H} : \psi(x_i;\theta)^T h(z_i) \in \mathrm{dom}(\phi),\ 1 \leq i \leq n\}$. The FGEL estimate $\hat{\theta}$ of θ_0 results from a saddle point of $G_{\lambda_n}(\theta, h)$

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \sup_{h\in\widehat{\mathcal{H}}(\theta)} G_{\lambda_n}(\theta, h). \tag{2.13}$$

Remark 2.4. *While the choice of ϕ can be motivated by the duality relationship of Theorem 2.1, i.e., $\phi = -\varphi^*$ with $\varphi^*(v) = \sup_p v^T p - \varphi(p)$ for φ defined via some φ-divergence, GEL estimators can be defined for any function ϕ which fulfills the corresponding conditions of Definition 2.3.*

Within the FGEL framework, the regularization term is responsible for regularizing an originally ill-posed operator estimation problem, which results from the optimization of $G_{\lambda_n}(\theta, h)$ over the instrument functions $h \in \mathcal{H}$. We will demonstrate this here exemplarily for the χ^2-divergence, which admits a closed form solution. Note that a similar argument has been provided earlier by Carrasco and

Kotchoni [27]. Let $\Psi_i(\theta) := \Psi(x_i, z_i; \theta) \in \mathcal{H}^*$ and $\Psi_i^* \in \mathcal{H}$ denote its dual which can be identified with a function in $\mathcal{H}$ by the self-duality property of Hilbert spaces [176]. Then the first order condition for h reads

$$0 = -\frac{1}{n}\sum_{i=1}^{n}\left[\Psi_i^*(\theta) - \left(\Psi_i^*(\theta)\Psi_i(\theta) + \lambda_n I \otimes I\right)(h)\right]$$
$$\Rightarrow h = -\left(\widehat{\Omega}_\theta + \lambda_n I \otimes I\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\Psi_i^*(\theta),$$

where $\widehat{\Omega}_\theta = \frac{1}{n}\sum_{i=1}^{n}\Psi_i^*(\theta)\Psi_i(\theta)$ denotes the empirical covariance operator of the moment functional. For any $i = 1, \dots, n$, the operator $\Psi_i^*(\theta)\Psi_i(\theta)$ has at most rank 1 and thus as the sum of n such operators $\widehat{\Omega}_\theta$ can have most rank n. As $\widehat{\Omega}_\theta$ is an infinite dimensional linear operator $\mathcal{H} \to \mathcal{H}$ it thus must be rank deficient for any finite n and therefore singular and non-invertible. This highlights the fact that the regularization parameter in the FGEL framework is not merely an artefact of the restoration of the strong duality between the primal and dual GEL problems, but a fundamental requirement for any definition of a functional/continuum GEL extension. Note that by the uniform weak law of large number and the continuous mapping theorem we have for $\Psi(\theta)$ continuous in θ and $\hat{\theta} \xrightarrow{p} \theta_0$ that $\widehat{\Omega}(\hat{\theta}) \xrightarrow{p} \Omega_0 = E[\Psi(X, Z; \theta_0)^*\Psi(X, Z; \theta_0)]$. In Theorem 2.8 we show that for the conditional case under the mild assumption that $V(Z) := E[\psi(X; \theta_0)\psi(X; \theta_0)^T|Z]$ is non-singular with probability 1 together with assumptions on the instrument functions class $\mathcal{H}$ it follows that Ω_0 is non-singular and thus in the limit $n \to \infty$ the first order conditions for h remain well posed as $\lambda_n \to 0$.

The general formulation (2.13) allows us to employ a wide range of function classes $\mathcal{H}$ and generally for finite samples, the choice of $\mathcal{H}$ will influence the obtained estimator. Building on recent developments in machine learning, we can represent h by a flexible deep neural network [72, 101] or a random forest model [10], for example. In this work, we mainly focus our discussion on instrument functions from reproducing kernel Hilbert spaces for their favorable theoretical properties but also consider neural network function classes.

Remark 2.5. *The FGEL framework admits an interesting relation to* distributionally robust optimization *and as such can be used for (distributionally) robust learning. Refer to Section A.1.1 of the appendix for a more detailed account of this connection.*

2.3.2 Asymptotic Properties

In this section, we establish asymptotic properties of our estimator given in (2.13). The proofs generalize the ones of Newey and Smith [118] for the GEL estimator with finite dimensional moment restrictions to our regularized problem with functional-valued moment restrictions. Let in the following $\mathcal{H}_1$ denote the unit ball in $\mathcal{H}$.

Theorem 2.6 (Consistency)**.** *Assume that a) $\theta_0 \in \Theta$ is the unique solution to $\|E[\Psi(X, Z; \theta)]\|_{\mathcal{H}^*} = 0$; b) Θ, $\mathcal{X}$, $\mathcal{Z}$ are compact; c) $\Psi(x, z; \theta)$ is continuous in x, z and θ everywhere;*
d) $E[(\sup_{\theta\in\Theta}\|\Psi(X, Z; \theta)\|_{\mathcal{H}^*})^\nu] < \infty$ for some $\nu > 2$; e) $\Omega_0 = E[\Psi(X, Z; \theta_0) \otimes \Psi(X, Z; \theta_0)]$

is non-singular; f) $\lambda_n = O_p(n^{-\xi})$ *with* $0 < \xi < 1/2$*; g)* ϕ *is twice continuously differentiable in a neighborhood of zero and* $\phi_1(0) \neq 0$*,* $\phi_2(0) < 0$*; h) the class of functions* $\{\Psi(\cdot;\theta)(h) : \theta \in \Theta,\ h \in \mathcal{H}_1\}$ *is* P_0*-Donsker. Let* $\hat{\theta}$ *denote the FGEL estimator for* θ_0*, then* $\hat{\theta} \xrightarrow{p} \theta_0$ *and* $\|E[\Psi(X, Z; \hat{\theta})]\|_{\mathcal{H}^*} = O_p(n^{-1/2})$.

If additionally i) $\theta_0 \in \operatorname{int}(\Theta)$*; j)* $\Psi(x, z; \theta)$ *is continuously differentiable in a neighborhood* $\bar{\Theta}$ *of* θ_0 *and* $E[\sup_{\theta\in\bar{\Theta}} \|\nabla_\theta \Psi(X, Z; \theta)\|^2_{\mathcal{H}^*}] < \infty$*; and k)* $\Sigma_0 := \langle E[\nabla_\theta \Psi(X, Z; \theta_0)], E[\nabla_{\theta^T} \Psi(X, Z; \theta_0)]\rangle_{\mathcal{H}^*}$ *is a non-singular matrix in* $\mathbb{R}^{p\times p}$*, we have* $\|\hat{\theta} - \theta_0\|_2^2 = O_p(n^{-1/2})$.

The following theorem shows that the limiting distributions of the variables follow a normal distribution N with covariance matrix Σ_θ and Gaussian process $\mathcal{N}$ with kernel Σ_h respectively.

Theorem 2.7 (Asymptotic normality)**.** *Let the assumptions of Theorem 2.6 be satisfied and define* $\nabla_\theta \Psi_0 := E[\nabla_\theta \Psi(X, Z; \theta_0)]$*. Then,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Sigma_\theta), \quad \sqrt{n}(\hat{h} - h) \xrightarrow{d} \mathcal{N}(0, \Sigma_h),$$

where $\Sigma_\theta = ((\nabla_\theta \Psi_0)\Omega_0^{-1}(\nabla_{\theta^T} \Psi_0^*))^{-1}$ *and* $\Sigma_h = \Omega_0^{-1} - \Omega_0^{-1}(\nabla_{\theta^T} \Psi_0^*)\Sigma_\theta(\nabla_\theta \Psi_0)\Omega_0^{-1}$.

2.3.3 From Functional to Conditional Moment Restrictions

Theorems 2.6 and 2.7 show that FGEL provides a $n^{-1/2}$-consistent and asymptotically normal estimator for *functional* moment restrictions of the form $\|E[\Psi(X, Z; \theta_0)]\|_{\mathcal{H}^*} = 0$. While this is of independent interest, often functional moment restrictions are merely a way to express a set of corresponding conditional moment restrictions $E[\psi(X; \theta_0)|Z] = 0\ P_Z-$a.s. using a sufficiently expressive function space $\mathcal{H}$ in the corresponding variational form (2.7). The following theorems show how the results on the functional formulation translate to the case of conditional moment restrictions. These results were not contained in the conference version of this manuscript and are originally due to Kremer et al. [93].

Theorem 2.8 (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* (2.6) *and* (2.7) *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)* $\Theta \subset \mathbb{R}^p$ *as well as* $\mathcal{X} \times \mathcal{Z}$ *are compact; c)* $\psi(x; \theta)$ *is continuous in* x *and* θ *everywhere; 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; and f)* $\lambda_n = O_p(n^{-\xi})$ *with* $0 < \xi < 1/2$*; and g)* ϕ *is twice continuously differentiable in a neighborhood of zero and* $\phi_1(0) \neq 0$*,* $\phi_2(0) < 0$ *and h) the function classes* $\{\psi(\cdot;\theta) : \theta \in \Theta\}$ *and* $\mathcal{H}_1$ *are* P_0*-Donsker. Then for the FGEL estimator* $\hat{\theta}$ *we have* $\hat{\theta} \xrightarrow{p} \theta_0$.

If additionally i) $\theta_0 \in \operatorname{int}(\Theta)$*; j)* $\psi(x; \theta)$ *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 k)* $\operatorname{rank}(E[\nabla_\theta \psi(X; \theta_0)|Z]) = p$ *w.p.1, we have* $\|\hat{\theta} - \theta_0\| = O_p(n^{-1/2})$.

Theorem 2.9 (Asymptotic Normality)**.** *Let the assumptions of Theorem 2.8 be satisfied. Then, for the FGEL estimator $\hat{\theta}$ we have*

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

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 2.9 agrees with the semi-parametric efficiency bound of Chamberlain [29]. This implies that FGEL provides an efficient estimator for conditional moment restrictions.

Corollary 2.10 (Efficiency)**.** *Let the assumptions of Theorem 2.8 be satisfied. Then the FGEL estimate $\hat{\theta}$ is an efficient estimator 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..*

In order to translate the results for the functional estimator to conditional moment restrictions by applying Theorems 2.8-2.10 one needs to choose a space of instrument functions $\mathcal{H}$ which fulfills the corresponding conditions. In the following we show that the reproducing kernel Hilbert space of certain kind of kernel fulfills these assumptions.

2.3.4 Kernel FGEL

The definition of our FGEL estimator contains a supremum over a function space $\mathcal{H}$. In order to address the conditional moment restriction problem, the function space must be expressive enough to exhibit an equivalent unconditional formulation. At the same time, optimization over function spaces is generally intractable and thus requires approximations. Selecting instrument functions from a reproducing kernel Hilbert space, one obtains a computationally efficient formulation involving finite dimensional parameters.

Reproducing kernel Hilbert spaces Let $\mathcal{X}$ be a non-empty set and $\mathcal{H}$ a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. Let $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\| \cdot \|_{\mathcal{H}}$ denote the inner product and norm on $\mathcal{H}$ respectively. Then $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS) if there exists a symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $k(x, \cdot) \in \mathcal{H}$ for all $x \in \mathcal{X}$ and $\langle f, k(x, \cdot) \rangle_{\mathcal{H}} = f(x)$ for all $f \in \mathcal{H}$ and $x \in \mathcal{X}$. Every positive (semi-)definite kernel is the unique reproducing kernel of an RKHS. We call a reproducing kernel k *integrally strictly positive definite* (ISPD) if additionally for any $f \in \mathcal{H}$ with $0 < \|f\|_2^2 < \infty$ we have $\int_{\mathcal{X}} f(x) k(x, x') f(x') \, \mathrm{d}x \, \mathrm{d}x' > 0$. See, e.g., Schölkopf and Smola [140] for a comprehensive introduction.

Let $\mathcal{H} = \bigoplus_{i=1}^m \mathcal{H}_i$ denote the direct sum of m RKHS of universal kernels k_i [106]. The following theorem which is based on Theorem 3.2 of Muandet et al. [109] shows that the RKHS corresponding to a universal ISPD kernel (e.g., Gaussian kernel) is expressive enough to represent the conditional moment restriction (2.6) in terms of a continuum of unconditional restrictions.

Theorem 2.11. *Let $\mathcal{H} = \bigoplus_{i=1}^m \mathcal{H}_i$ denote the direct sum of m RKHS unit balls $\mathcal{H}_i$ corresponding to ISPD kernels k_i, $i = 1, \ldots, m$. Let P denote a distribution over random variables $X \in \mathcal{X}$ and*

$Z \in \mathcal{Z}$ with marginal distributions P_X and P_Z. Then

$$E_P[\psi(X;\theta)|Z] = 0 \;\; P_Z\text{-a.s.}, \tag{2.14}$$

if and only if

$$E_P[\psi(X;\theta)^T h(Z)] = 0 \;\; \forall h \in \mathcal{H}. \tag{2.15}$$

With this result at hand, we can apply Theorems 2.8-2.10 to show that FGEL combined with such an RKHS as instrument functions space, which we term Kernel FGEL, provides an efficient estimator for CMR problems as formalized in the following corollary.

Corollary 2.12. *The RKHS of a universal ISPD kernel satisfies the assumptions of Theorems 2.8-2.10. Thus, under Assumptions a)-j) of Theorem 2.8 Kernel FGEL provides a $\frac{1}{\sqrt{n}}$-consistent, asymptotically normal and semi-parametrically efficient estimator for conditional moment restriction problems.*

Applying the representer theorem [142] to the supremum over the instrument functions h in equation (2.13) allows us to represent the RKHS function in terms of finite dimensional parameters $\alpha_r \in \mathbb{R}^n$, $r = 1, \ldots, m$, and yields a finite dimensional and convex optimization problem as formalized by the following lemma.

Lemma 2.13. *Let $\mathcal{H} = \bigoplus_{i=1}^m \mathcal{H}_i$ be an RKHS corresponding to m universal kernels k_i, $i = 1, \ldots, m$. Let $K_r \in \mathbb{R}^{n \times n}$, $r = 1, \ldots, m$ denote the kernel matrices and let $\alpha = \{\alpha_r\}_{r=1}^m$ with $\alpha_r \in \mathbb{R}^n$. Then the maximization over the instrument functions in the FGEL objective* (2.13) *can be expressed as*

$$R_{\lambda_n}(\theta) := \max_{\alpha \in \widehat{A}_\theta} \left\{ \frac{1}{n} \sum_{i=1}^n \phi\left(v_i(\theta, \alpha)\right) - \frac{\lambda_n}{2} \sum_{r=1}^m \alpha_r^T K_r \alpha_r \right\},$$

with $v_i = \sum_{r=1}^m (\alpha_r^T K_r)_i \psi_r(x_i;\theta)$ and $\widehat{A}_\theta = \{\alpha : v_i \in \mathrm{dom}(\phi), 1 \leq i \leq n\}$. The Kernel FGEL estimator is then defined as the solution of $\hat{\theta} = \arg\min_{\theta \in \Theta} R_{\lambda_n}(\theta)$.

We provide details on the optimization algorithm in Section A.1.2 of the appendix.

2.3.5 Neural FGEL

As expressed by universal approximation theorems (e.g., [174]), neural networks can represent arbitrarily large function classes and have shown state-of-the-art performance on related tasks [72, 101, 15]. As such, they provide a particularly interesting choice of instrument function class. Let $h_\omega : \mathcal{Z} \to \mathbb{R}^m$ denote a feed-forward neural network with parameters ω. Then we can define the Neural FGEL estimator as a saddle point of

$$G_{\lambda_n}(\theta, \omega) := \frac{1}{n} \sum_{i=1}^n \phi\left(\psi(x_i;\theta)^T h_\omega(z_i)\right) - \frac{\lambda_n}{2n} \sum_{i=1}^n \|h_\omega(z_i)\|_{\mathbb{R}^m}^2,$$

where the regularization term penalizes the magnitude of the output as in Dikkala et al. [50] and Bennett and Kallus [14]. We leave the theoretical analysis of the Neural FGEL estimator for future work.

2.3.6 Other Instrument Function Classes

FGEL estimators can be defined for arbitrary instrument function classes $\mathcal{H}$ under mild conditions: Let P denote a reference measure over $X \in \mathcal{X}$, then we can place a class $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}^m$ into the Hilbert space of square-integrable functions $L^2(\mathcal{H}, P)$ as long as any $f \in \mathcal{H}$ is bounded on any set with non-zero measure, which is a realistic assumption for many model classes. The corresponding norm with respect to the empirical measure is then given via $\|h\|^2_{\mathcal{H}} = \frac{1}{n}\sum_{i=1}^n \|h(z_i)\|^2_{\mathbb{R}^m}$. If the underlying problem of interest is a conditional moment restriction (instead of a general functional moment restriction), $\mathcal{H}$ additionally must be expressive enough such that an equivalence between the conditional (2.6) and unconditional (2.7) formulations holds.

2.3.7 Choice of Divergence Function

In this section, we discuss various choices of divergences and establish connections to existing methods. In the finite dimensional case, it is well known that for any quadratic discrepancy function the GEL estimator coincides with the continuous updating GMM (CUE) estimator [118]. An interesting special choice of divergence function is given below.

Proposition 2.14. *Choosing the GEL function as $\phi(v) = -(1 \pm \frac{v}{2})^2$ and rescaling the regularization parameter $\tilde{\lambda}_n = 2\lambda_n$, the FGEL estimator becomes equivalent to the solution of the optimization problem*

$$\min_{\theta\in\Theta} \sup_{h\in\mathcal{H}} \Big\{ E_{\hat{P}_n}[\psi(X;\theta)^T h(X)] - \frac{1}{4} E_{\hat{P}_n}\Big[\big(\psi(X;\theta)^T h(X)\big)^2\Big] - \frac{\tilde{\lambda}_n}{4}\|h\|^2_{\mathcal{H}} \Big\}.$$

This resembles the objective of the VMM estimator of Bennett and Kallus [14] with the only difference that the covariance term contains the decision variable θ instead of a first-stage estimate $\tilde{\theta}$. In this sense, with this special choice of divergence function our FGEL estimator and the VMM estimator are related in the same way as the continuous updating estimator (CUE) (2.2) and the optimally weighted 2-step GMM estimator (2.3). With the kernel version of our FGEL estimator, we can carry out the optimization over $h \in \mathcal{H}$ in closed form and similarly obtain a continuous updating version of the Kernel VMM estimator.

A functional generalization of the original empirical likelihood estimator is retrieved by setting $\phi(v) = -\log(1 - v)$. The empirical likelihood estimator has many desirable properties. It has been shown by Newey and Smith [118] that the ordinary EL estimator has the smallest higher order bias among the family of GEL estimators (including GMM). Further, Corcoran [41] shows that confidence intervals constructed from the EL-based profile likelihood admit a Bartlett correction which by a simple subtraction allows to reduce the coverage error from $O(n^{-1})$ to $O(n^{-2})$. This property of the EL framework is unique among the family of GEL estimators [41].

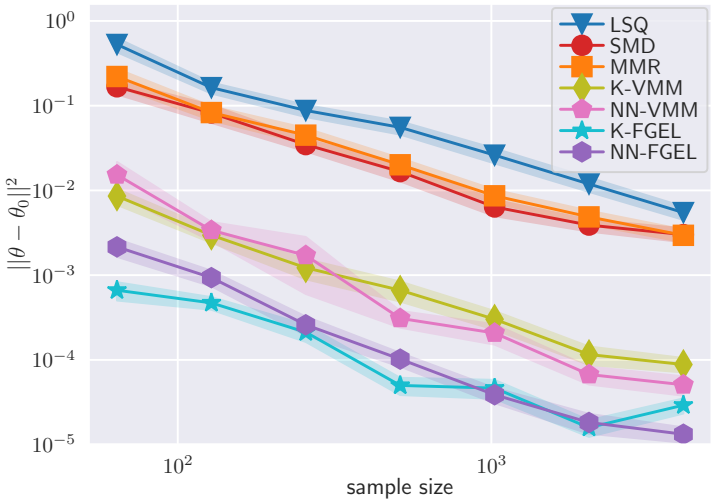


Fig. 2.1 Estimation error over sample size for the heteroskedastic regression experiment. Lines and shaded regions represent the MSE of the estimated parameters and the standard error averaged over 70 runs respectively.

Using the GEL function corresponding to the Kullback-Leibler (KL) divergence $\phi(v) = -e^v$ one obtains a functional generalization of the exponential tilting estimator of Kitamura and Stutzer [87] and Imbens et al. [80] which shows good empirical performance on many tasks [80]. In contrast to the χ^2-divergence, the KL-divergence enjoys great popularity as a distributional divergence measure in machine learning [22]. Therefore, a functional moment restriction estimator based on the KL-divergence instead of the dominating χ^2-divergence (GMM) could be of particular interest.

2.4 Experiments

For all experiments we use radial basis function kernels $k_i(x, x') = \exp(-\gamma\|x - x'\|^2)$, $i = 1, \ldots, m$ and set the bandwidth parameter γ via the common median heuristic [140, 57]. If not stated otherwise, we tune the remaining hyperparameters of all methods by evaluating the MMR objective $\ell(\theta) = 1/n^2 \sum_{i,j=1}^{n} \psi(x_i; \theta)^T K_{ij} \psi(x_j; \theta)$ [179] on a validation set of the same size as the training set (refer to Section A.1.3 of the appendix for details). We compare the performance of our kernel- and neural network-based methods with ordinary least-squares (LSQ), sieve minimum distance (SMD) [2], kernel maximum moment restrictions (MMR) [179] and the kernel- and neural network versions of the variational method of moments (K-VMM and NN-VMM) [14, 15] on two conditional moment restriction problems. Code for reproducing the experimental results is available at https://github.com/HeinerKremer/Functional-GEL.

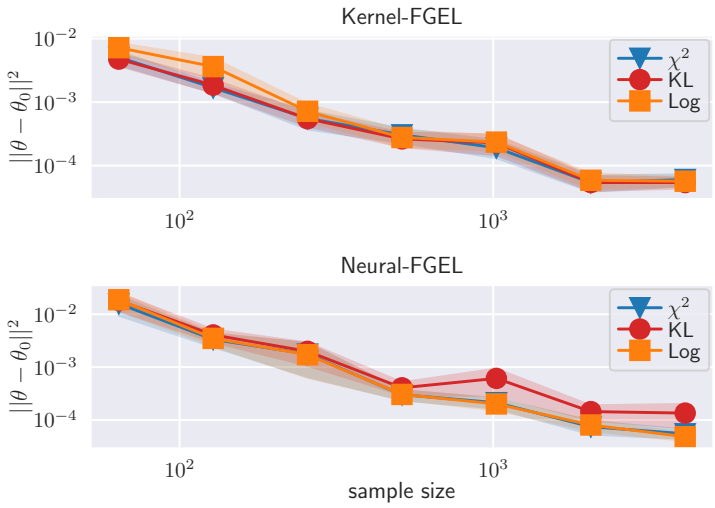


Fig. 2.2 Comparison of different divergence functions. Lines and shaded regions represent the MSE of the estimated parameters and the standard error averaged over 70 runs respectively.

2.4.1 Linear Regression under Heteroskedastic Noise

We define a simple data generating process for a one-dimensional estimation problem. Let $\theta = 1.7 \in \mathbb{R}$ and

$$y = x^T\theta + \varepsilon, \quad x \sim \text{Uniform}([-1.5, 1.5]),$$

where ε describes heteroskedastic noise such that $\varepsilon|x \sim \mathcal{N}(0, \sigma = 5x^2)$. We can formulate the regression task as the conditional moment restriction $E[Y - X^T\theta|X] = 0$ P_X-a.s.. As ε is a mean zero random variable, here, we can simply use the mean squared error on a validation set as an unbiased validation metric to tune the hyperparameters of all methods.

Figure 2.1 shows the mean-squared error (MSE) of the estimated parameters using different versions of FGEL and other state-of-the-art estimators for conditional moment restrictions in dependence on the sample size. Here we treat the choice of divergence as an additional hyperparameter. We observe that both our methods yield the lowest parameter MSE and even slightly outperform the recently proposed state-of-the-art VMM estimator [14]. In Figure 2.2 we evaluate the effect of the divergence function. We observe that while the average performance is largely independent of the choice of divergence function, a comparison with the results shown in Figure 2.1 reveals that for any fixed sample the different divergences yield estimators of different quality. Thus, treating the divergence as hyperparameter and choosing the estimator with the lowest validation loss, allows us to exceed the performance of the FGEL estimator with fixed divergence. As GMM-based methods implicitly build on the χ^2-divergence, this highlights an advantage of our method which can leverage any φ-divergences. Note that for any *fixed* divergence function the performance of our FGEL estimators are roughly on par with the corresponding VMM estimators.

Table 2.2 Prediction MSE for the instrumental variable task. Mean and standard deviation of the mean are computed over 50 random runs and multiplied by 10 for ease of presentation.

	LSQ	SMD	MMR	K-VMM	NN-VMM	K-FGEL	NN-FGEL
abs	3.72 ± 0.30	2.97 ± 0.97	2.78 ± 0.60	0.43 ± 0.15	0.45 ± 0.10	$\mathbf{0.17 \pm 0.01}$	0.23 ± 0.06
step	3.03 ± 0.03	0.37 ± 0.04	0.71 ± 0.03	$\mathbf{0.31 \pm 0.01}$	0.41 ± 0.01	0.41 ± 0.03	0.34 ± 0.01
sin	3.28 ± 0.04	$\mathbf{1.01 \pm 0.06}$	3.61 ± 0.07	1.55 ± 0.12	1.72 ± 0.11	1.97 ± 0.16	1.66 ± 0.12
linear	2.76 ± 0.06	0.97 ± 0.72	1.98 ± 0.38	0.31 ± 0.06	0.34 ± 0.05	0.32 ± 0.05	$\mathbf{0.20 \pm 0.03}$

2.4.2 Instrumental Variable Regression

We adopt a slightly modified version of the IV regression experiment of Lewis and Syrgkanis [101], which has also been used by Bennett et al. [15] and Zhang et al. [179]. Let the data generating process be given by

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

where f_0 is picked from the following simple functions

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

We approximate f_0 by a shallow neural network $f_\theta(x)$ with 2 layers of $[20, 3]$ units and leaky ReLU activation functions and base the estimation on the conditional moment restrictions $E[Y - f_\theta(X)|Z] = 0 \;\; P_Z$-a.s.. As generally the true model is not contained in this model class, this provides a typical case of model misspecification and theoretical properties of the our method (and equally all baseline methods) for this setting have yet to be developed (see Dikkala et al. [50] for recent progress in this direction). We use training and validation sets of size $n = 2000$ and evaluate the prediction error on a test set of 20000 samples. The results are visualized in Table 2.2. We observe that with the exception of one task the FGEL and VMM estimators outperform all other baselines. Compared to each other NN-FGEL seems to be preferable over NN-VMM but the kernel versions of both methods exhibit similar performance without showing a clear advantage of one over the other for this task.

Our experiments show that the FGEL estimator is a viable alternative to previously proposed continuum method of moments estimators for conditional moment restrictions and can surpass the previous state-of-the-art on some tasks. However, further empirical evidence needs to be collected to verify its predicted superior finite sample properties for infinitely many moment restrictions. We leave a comprehensive experimental evaluation to future work.

2.5 Related Work

Learning with conditional or infinite dimensional moment restrictions respectively has been an active field of research in econometrics and more recently in machine learning. In the former context, seminal work on extending the generalized method of moments to continua of moment restrictions has been carried out by Carrasco and Florens [26], Carrasco et al. [28] by placing the constraints in an RKHS. In the machine learning community, GMM-related estimators have been developed by casting the infinite dimensional moment restriction problem as a minimax game and representing the adversarial player by an RKHS function [179, 14] or a flexible neural network [72, 101, 50, 15]. While the neural network-based methods often achieve good performance in practice, they generally are computationally more expensive and lack the theoretical properties of traditional GMM estimators. In contrast, Bennett and Kallus [14]'s Kernel VMM estimator comes with strong theoretical guarantees but results from a 2-step procedure and thus depends on an initial parameter estimate. As discussed in Section 2.3.7, our framework contains a continuous updating version of VMM as a special case but allows for using alternative φ-divergence functions.

As an alternative to GMM estimation, sieve-based methods [117, 51, 2, 32] address conditional moment restrictions by growing the number of unconditional restrictions with the sample size by manually selecting an increasing number of basis functions. While these often come with desirable efficiency results, in practice they can be hard to tune and computationally demanding [14]. Another line of work implicitly estimates optimal instrument functions via a kernel-smoothed localized empirical likelihood function [165, 88]. Their use of kernels is different from our approach as we do not smooth the profile divergence but use RKHS functions (and other function classes) as instrument functions.

Several works extended the generalized empirical likelihood framework to handle infinite dimensional moment restrictions and thus conditional moment restrictions [51, 30, 27]. The GEL estimator of Chaussé [30] is based on approximately imposing a continuum of moment restrictions using a parameterized basis of functions and solving a regularized version of the GEL first order conditions. While it is theoretically closely related to our method, the regularization scheme and computational approach differs from ours. Similarly, closely related to our method is the regularized GEL estimator of Carrasco and Kotchoni [27], which is defined via a set of optimality conditions and solved using a procedure motivated by the Three-Steps Euclidean Likelihood procedure of Antoine et al. [6]. In contrast to these methods, our estimator is defined as a saddle point of an objective function and thus benefits from recent advances in mini-max optimization [46, 102]. To the best of our knowledge, our work is the first to combine GEL estimation with modern machine learning and in particular kernel methods and neural networks.

2.6 Conclusion

Several long-established problems in machine learning can naturally be expressed as a risk minimization problem. On the other hand, emerging areas such as causal inference, algorithmic decision making, and robust learning often involve problems that are formulated as (potentially infinite) mo-

ment restrictions and require different algorithmic frameworks for estimation and inference. Recent works have advanced this development by combining classical techniques from econometrics such as generalized method of moments (GMM) with modern machine learning models such as deep neural networks and kernel machines. Likewise, our work contributes to this endeavour by equipping the more general generalized empirical likelihood (GEL) framework with such powerful models. While the econometrics community enjoys the new class of algorithms, we believe the machine learning community will likewise benefit from new perspectives on causal inference and robust learning which will be explored in future works.

This paper laid the theoretical foundation of the functional GEL framework, but there remain open questions that impede real-world applications. Firstly, more efficient optimization procedures need to be developed that allow for large scale applications. Secondly, theoretical properties of the framework with specific function classes need to be explored. Lastly, the framework needs to be tested for the training of more complex models for real-world applications (e.g. robust learning). Our goal is to address some of these problems in future work.

Chapter 3

Estimation Beyond Data Reweighting: Kernel Method of Moments*

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.

*Based on *Estimation Beyond Data Reweighting: Kernel Method of Moments* [93]. Heiner Kremer, Yassine Nemmour, Bernhard Schölkopf, Jia-Jie Zhu. International Conference for Machine Learning 2023

Declaration

This chapter is based in parts or in full on the published manuscript:

> *Estimation Beyond Reweighting: Kernel Method of Moments*
> Heiner Kremer, Yassine Nemmour, Bernhard Schölkopf, Jia-Jie Zhu
> International Conference for Machine Learning (ICML) 2023

Author contributions

Heiner Kremer: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization
Yassine Nemmour: Methodology, Software, Validation, Investigation, Writing - Review & Editing, Visualization
Bernhard Schölkopf: Conceptualization, Writing - Review & Editing, Supervision, Resources
Jia-Jie Zhu: Conceptualization, Methodology, Investigation, Writing - Review & Editing, Supervision

3.1 Introduction

Many problems in machine learning, statistics, causal inference and economics can be formulated as (conditional) moment restrictions [117, 4]. 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 [20]. A prominent CMR problem is instrumental variable (IV) regression [117], 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 [4]. Other examples of CMR problems include variants of double machine learning [36–38] and off-policy evaluation in reinforcement learning [172, 34, 13, 16]. Perhaps the most popular approach to learning with moment restrictions is the generalized method of moments (GMM) of Hansen [69], which recently gained popularity in machine learning [101, 14]. GMM belongs to the wider family of generalized empirical likelihood (GEL) estimators of Owen [122, 121], Qin and Lawless [130], Smith [151]. 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 [70] 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 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 [29].
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 3.2 gives an overview of method of moments estimation for conditional and unconditional moment restrictions. Section 3.3 introduces our estimator, and provides duality results as well as asymptotic properties and practical considerations. Section 3.4 provides an empirical evaluation of our estimators on various conditional moment restriction tasks. Section 3.5 discusses connections to related methods and Section 3.6 concludes.

3.2 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 [69]. 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 [122, 121, 130]. In an attempt to improve the finite sample properties of GMM, alternative estimators from this family have been proposed [151, 118]. 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

$$R(\theta) = \inf_{P\ll\hat{P}_n} D_\varphi(P||\hat{P}_n) \quad \text{s.t.} \quad E_P[\psi(X;\theta)] = 0, \quad E_P[1] = 1, \tag{3.1}$$

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 $\theta \in \Theta$,

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

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 and Smith [118] (Theorem 2.1).

Proposition 3.1. *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

$$E[\psi(X;\theta_0)|Z] = 0,\ P_Z\text{-a.s.}, \tag{3.2}$$

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 [20] of the form

$$E[\psi(X;\theta)^T h(Z)] = 0\ \ \forall h \in \mathcal{H}, \tag{3.3}$$

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 [106]. Both the GMM and the GEL framework have been extended to conditional moment restrictions in multiple ways, building on basis function expansions of $\mathcal{H}$ [26, 165, 2, 30, 27] as well as modern machine learning models [72, 101, 15, 92].

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) [65] is defined as $\mathrm{MMD}(P, Q; \mathcal{F}) := \|\mu_P - \mu_Q\|_{\mathcal{F}}$. Refer to, e.g., Schölkopf and Smola [140], Berlinet and Thomas-Agnan [18], Steinwart and Christmann [158] for comprehensive introductions to kernel methods for machine learning.

3.3 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 B.5.

3.3.1 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

$$R(\theta) = \inf_{P\in\mathcal{P}} \frac{1}{2}\,\mathrm{MMD}(P, \hat{P}_n; \mathcal{F})^2 \quad \text{s.t.} \quad E_P[\psi(X;\theta)] = 0, \quad E_P[1] = 1, \tag{3.4}$$

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 (3.4) as $n \to \infty$. Using Lagrange duality we can derive the corresponding dual problem as formalized in the following.

Theorem 3.2. *The* MMD profile (3.4) *has the strongly dual form*

$$\begin{aligned} R(\theta) = \sup_{\substack{\eta\in\mathbb{R}, f\in\mathcal{F}, \\ h\in\mathbb{R}^m}} & \frac{1}{n}\sum_{i=1}^{n} f(x_i) + \eta - \frac{1}{2}\|f\|_{\mathcal{F}}^2 \\ \text{s.t.} \quad & f(x) + \eta \le \psi(x;\theta)^T h \quad \forall x \in \mathcal{X}. \end{aligned} \tag{3.5}$$

Note that the structure and derivation of (3.5) resembles recent reformulation techniques for MMD-based distributionally robust optimization (DRO) by Zhu et al. [183] and in fact GEL estimation can be seen as a dual problem to DRO [95].

While MMD enjoys many favorable properties, the MMD-profile (3.5), 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.

3.3.2 Entropy Regularization

Inspired by the interior point method for convex optimization in finite-dimensions [114], we define an entropy-regularized version of the MMD profile (3.5). This allows us to translate the semi-infinite constraint in (3.5) 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 [44]. 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

$$D_\varphi(P||\omega) = \int_{\mathcal{X}} \varphi(p(x))\omega(\mathrm{d}x). \tag{3.6}$$

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

$$R_\epsilon^\varphi(\theta) = \inf_{P \ll \omega} \frac{1}{2}\,\mathrm{MMD}(P, \hat{P}_n; \mathcal{F})^2 + \epsilon D_\varphi(P||\omega) \quad \text{s.t.} \quad E_P[\psi(X;\theta)] = 0, \quad E_P[1] = 1. \quad (3.7)$$

In contrast to the classical φ-divergence-based profile divergence (3.1), 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 B.4.3). 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 (3.7) as formalized in the following theorem.

Theorem 3.3 (KMM Duality)**.** *The entropy-regularized MMD profile* (3.7) *has the strongly dual form*

$$R_\epsilon^\varphi(\theta) = \sup_{\substack{\eta\in\mathbb{R}, f\in\mathcal{F}, \\ h\in\mathbb{R}^m}} \frac{1}{n}\sum_{i=1}^n f(x_i) + \eta - \frac{1}{2}\|f\|_{\mathcal{F}}^2 - \epsilon \int_{\mathcal{X}} \varphi^*\left(\frac{f(x) + \eta - \psi(x;\theta)^T h}{\epsilon}\right)\omega(\mathrm{d}x). \quad (3.8)$$

where $\varphi^*(t) := \sup_s \langle t, s\rangle - \varphi(s)$ *denotes the convex conjugate of* φ*. The optimization problem in* (3.8) *is jointly convex in the dual variables* (η, f, h)*.*

As opposed to the unregularized version (3.5) the entropy-regularized MMD profile (3.7) 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

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

3.3.3 Choices of Entropy Regularizers

Different choices of φ-divergences in (3.6) correspond to different relaxations of the generally intractable semi-infinite constraint in (3.5). 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 (3.5) 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

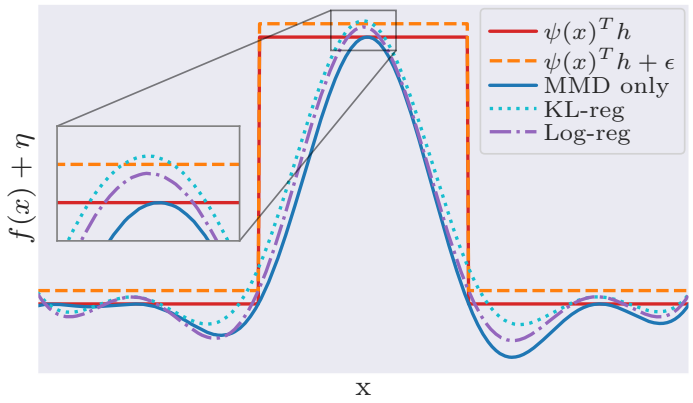


Fig. 3.1 Effect of Entropy Regularization. The red and orange lines correspond to an exemplary function $\psi(x;\theta)^T h$ and its relaxation $\psi(x;\theta)^T h + \epsilon$. The blue line shows the strictly minorizing RKHS function resulting from enforcing the constraint in (3.5) exactly. The cyan 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 (3.5) by up to ϵ. The KL-divergence yields a soft constraint by penalizing violations exponentially.

leads to a dual regularized MMD profile with a log-barrier

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

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 (3.5). Figure 3.1 provides a visualization of the different regularization schemes. Refer to Section B.4 for additional details.

3.3.4 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. [92].

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 (3.2) as a continuum of unconditional restrictions (3.3). Let $\mathcal{H}_1 = \{h \in \mathcal{H} : \|h\| \leq 1\}$ denote the unit ball in $\mathcal{H}$ and let $\mathcal{H}^*$ denote the dual space of functionals $\Psi : \mathcal{H} \to \mathbb{R}$, equipped with the dual norm $\|\Psi\|_{\mathcal{H}^*} = \sup_{h\in\mathcal{H}_1} \Psi(h)$. 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 (3.3) can thus be written as

$$E[\Psi(X, Z; \theta)] = 0 \in \mathcal{H}^*, \tag{3.9}$$

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 [176], 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

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

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. [92]. With this at hand, we can define the functional version of the entropy-regularized MMD profile for conditional moment restrictions (refer to Section B.1.1 for details on the duality relationship).

Definition 3.4 (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* (3.2) *and* (3.3) *holds. Then the entropy-regularized MMD profile for conditional moment restrictions is given as*

$$\begin{aligned} R^{\varphi}_{\epsilon,\lambda_n}(\theta) = \sup_{\substack{\eta \in \mathbb{R}, f \in \mathcal{F}, \\ h \in \mathcal{H}}} & \frac{1}{n}\sum_{i=1}^{n} f(x_i, z_i) + \eta - \frac{1}{2}\|f\|^2_{\mathcal{F}} - \frac{\lambda_n}{2}\|h\|^2_{\mathcal{H}} \\ & - \epsilon \int_{\mathcal{X}\times\mathcal{Z}} \varphi^*\left(\frac{f(x, z) + \eta - \Psi(x, z; \theta)(h)}{\epsilon}\right) \omega(\mathrm{d}x \otimes \mathrm{d}z). \end{aligned} \tag{3.10}$$

From the proof of Theorem 3.3 it directly follows that the optimization problem in (3.10) 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 [147] to guarantee the consistency of the solution for the conditional moment restrictions [92]. In practice, alternative choices, e.g., neural networks which lack these theoretical guarantees have proven successful and often preferable [14, 92].

3.3.5 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 B.1.2) 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.2.

Theorem 3.5 (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* (3.2) *and* (3.3) *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) $\Theta \subset \mathbb{R}^p$ and $\mathcal{X} \times \mathcal{Z}$ are compact; c) $\psi(x;\theta)$ is continuous in x and θ everywhere; 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; g) $\lambda_n = O_p(n^{-\xi})$ with $0 < \xi < 1/2$; and h) The function classes $\{\psi(\cdot;\theta) : \theta \in \Theta\}$ and $\mathcal{H}_1$ are P_0-Donsker. Then for the KMM estimator $\hat{\theta}$ we have $\hat{\theta} \xrightarrow{p} \theta_0$.*

If additionally i) $\theta_0 \in \operatorname{int}(\Theta)$; j) $\psi(x;\theta)$ 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 k) $\operatorname{rank}(E[\nabla_\theta \psi(X;\theta_0)|Z]) = p$ w.p.1, we have $\|\hat{\theta} - \theta_0\| = O_p(n^{-1/2})$.

Remark 3.6. *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 $\operatorname{supp}(\hat{P}_n) \subseteq \operatorname{supp}(Q)$, the form of $\omega = (1-\alpha)\hat{P}_n + \alpha Q$ does not restrict the set of candidate distributions P in* (3.10) *further than to distributions that admit a density w.r.t. Q.*

Remark 3.7. *A sufficiently rich function space for Theorem 3.5 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. [92]). Moreover, based on universal approximation theorems [77], $\mathcal{H}$ 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 3.8 (Asymptotic Normality)**.** *Let the assumptions of Theorem 3.5 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 3.8 agrees with the semi-parametric efficiency bound of Chamberlain [29]. This implies that the KMM estimator achieves the lowest possible asymptotic variance among any estimators based on the CMR (3.2) as formalized in the following corollary.

Corollary 3.9 (Efficiency)**.** *Let the assumptions of Theorem 3.5 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 mini-max approaches to CMR estimation [101, 15, 179] and which is matched only by the kernel VMM estimator of Bennett and Kallus [14] and the more traditional sieve-based approaches of Ai and Chen [2] and Chen and Pouzo [31]. Note that while this shows that our estimator is asymptotically first-order equivalent to these methods, the finite sample properties can be vastly different.

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 $\{(x_i, z_i)\}_{i=1}^{n_1} \sim \hat{P}_n$, $\{(x_j^\omega, z_j^\omega)\}_{j=1}^{n_2} \sim \omega$
  $G \leftarrow \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} G_{\epsilon,\lambda}(\theta, \beta; (x_i, z_i), (x_j^\omega, z_j^\omega))$
  $\beta \leftarrow \mathrm{AscentStep}(\beta, \nabla_\beta G)$
  $\theta \leftarrow \mathrm{DescentStep}(\theta, \nabla_\theta G)$
end while
Output: Parameter estimate θ

3.3.6 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 (3.10) 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 (3.10) 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 (3.10) 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

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \sup_{\beta \in \mathcal{M}} \widehat{G}_{\epsilon,\lambda_n}(\theta, \beta), \tag{3.11}$$

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 (3.11). 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 $\hat{P}_n$ 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 (3.11) 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 (3.11) 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 [131]. 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.

3.4 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) [179], sieve minimum distance (SMD) [2], DeepIV [72], DeepGMM [15] and the neural network version of functional GEL (FGEL) [92]. 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 [57]. The hyperparameters of all methods are set by evaluating the Hilbert-Schmidt independence criterion (HSIC) [63] between the residues and the conditioning variable $\mathrm{HSIC}(Y - g(T), Z)$ 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. [108] and Saengkyongam et al. [137] and we empirically find it to yield better estimates than MMR [109, 179] which has been used for as a validation metric in other works (see Section B.3.2). For the variational approaches we use a batchsize of $n_1 = 200$. Additionally, for our KMM estimator we represent the reference distribution ω by a kernel density estimator (KDE) trained on the empirical sample (see Section B.4.3) from which we sample mini-batches of size $n_2 = 200$. Refer to Section B.3 for additional details. An implementation of our estimator and code to reproduce our results is available at https://github.com/HeinerKremer/conditional-moment-restrictions.

Table 3.1 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.

n	OLS	MMR	DeepIV	DeepGMM	FGEL	KMM
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
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
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
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
10000	1.99 ± 0.04	N/A	1.99 ± 0.04	0.02 ± 0.01	0.03 ± 0.01	0.01 ± 0.00

Heteroskedastic Instrumental Variable Regression We adopt the HeteroskedasticIV experiment of Bennett and Kallus [14]. 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 3.1 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 and Syrgkanis [101] which has also been used by Bennett et al. [15], Zhang et al. [179] and Kremer et al. [92]. Consider a data generating process given by

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

where the function g_0 is chosen from

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

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

Table 3.2 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

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 3.2 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.

3.5 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 (3.2) and continuum (3.3) formulations. Seminal work in this direction has been carried out by [26, 28] and Ai and Chen [2], 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 [72, 101, 15, 14, 50]. 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 and Kitamura [165], Kitamura et al. [88], Chaussé [30], Carrasco and Kotchoni [27] 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 [92]. 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) [95]. In this context, it has been used to investigate the statistical properties of DRO [54, 95] and to calibrate the size of the distributional ambiguity set used in the DRO framework [96, 98, 21, 73]. With the notable exception of Blanchet et al. [21] these works build on the standard φ-divergence-based GEL framework. While Blanchet et al. [21] provide a GEL framework based on optimal transport distances,

their goal is to calibrate an ambiguity 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 (3.5). Classical approaches to handling such constraints using polynomial sum-of-squares (SOS) [99] do not apply here since we have a general moment function class outside the polynomials. Furthermore, both classical and infinite-dimensional SOS techniques [105] 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. [82] as a time-discretization scheme of PDEs. In recent literature related to machine learning, Arbel et al. [7] studied the variational structure of MMD as energy in the Wasserstein geometry. Chizat [40] 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. [183], 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.

3.6 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.

Chapter 4

Geometry-Aware Instrumental Variable Regression*

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

*Based on *Geometry-Aware Instrumental Variable Regression* [91]. Heiner Kremer and Bernhard Schölkopf. International Conference for Machine Learning (ICML) 2024.

Declaration

This chapter is based in parts or in full on the published manuscript:

> *Geometry-Aware Instrumental Variable Regression*
> Heiner Kremer, Bernhard Schölkopf
> International Conference for Machine Learning (ICML) 2024

Author contributions (CRediT)

Heiner Kremer: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization
Bernhard Schölkopf: Writing - Review & Editing, Supervision, Resources

4.1 Introduction

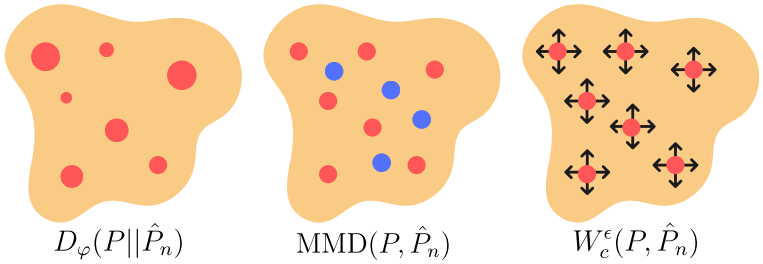


Fig. 4.1 Paradigms to approximate P_0 from data (red dots) in the GEL framework. φ-divergence-based estimators (left) approximate P_0 by reweighting (weight $\hat{=}$ size) the sample (e.g., [2, 14]. MMD-based estimators (middle) allow for sampling additional data points (blue dots) [93]. In contrast, optimal transport-based estimators (right) allow to move around the data points (present work).

Instrumental variable regression is one of the most widespread approaches for learning in presence of confounding [4]. It is applicable in situation where one is interested in inferring the outcome Y of some treatment T, where both, treatment and outcome, are affected by a so-called unobserved confounder U. To eliminate the confounding bias, one can take into account an instrumental variable Z, which i) affects the treatment T, ii) affects the outcome Y only through its effect on T, and, iii) is independent of the confounder U. While traditionally the problem has been addressed through the 2-stage least squares approach [4], in recent years the formulation in terms of conditional moment restrictions (CMR) has gained popularity for its potential to benefit from advances in machine learning models [15, 14, 50, 179, 110, 92, 93]. The CMR formulation of IV regression is based on restricting the expectation of the prediction residual $Y - f(T)$ conditioned on the instruments Z, where f denotes a causal relation from T to Y that one wants to infer. In general, this leads to a zero-sum game in which one minimizes an objective with respect to the model parameters and maximizes it with respect to an adversary function that detects the moment violations [15, 50]. One of the most general frameworks for learning with moment restrictions is the family of generalized empirical likelihood (GEL) estimators [123, 130, 80, 87], which includes the prominent generalized method of moments [69, 70, 67]. The idea behind empirical likelihood is to learn a model via maximum likelihood estimation without specifying a parametric form of the data distribution [123]. In practice, this is realized by learning a non-parametric approximation of the population data distribution P_0 along with the model f by means of minimizing a φ-divergence under the moment restrictions. However, by relying on φ-divergences one effectively restricts the estimator of the population distribution to reweightings of the sample. The reweighting assumption has recently been lifted by Kremer et al. [93] by introducing an estimator based on maximum mean discrepancy [65]. Their estimator allows for more fine-grained approximations of P_0 by sampling additional data points from a generative model. While reweightings of the present data or sampling of additional points might be suitable to find sufficiently close approximations of the population distribution in some cases, in presence of highly complex data manifolds, e.g., image spaces, they might become ineffective as they are blind towards the geometry of the data space. This is particularly relevant in the presence of poisoned [33] or adversarial [60] data points, i.e., data that has been corrupted with small perturbations which lead to vastly inaccurate predictions. The key to robustness against such perturbations is to look

at how the learning signal changes around the empirical data points, i.e., to take into account the geometry of the signal with respect to the data manifold. We implement the idea of a geometry-aware learning with conditional moment restrictions by proposing an empirical likelihood-type estimator based on a regularized optimal transport distance, which we call the Sinkhorn Method of Moments (SMM). Figure 4.1 schematically compares our method to previous approaches to empirical likelihood estimation.

Our contributions

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

Section 4.2 introduces empirical likelihood estimation for conditional moment restrictions, followed by the derivation of our method in Section 4.3. Empirical results are provided in Section 4.4 and related work is discussed in Section 4.5.

4.2 Empirical Likelihood Estimation for CMR

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

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

$$E[\psi(T, Y; f_0)|Z] = 0 \; P_Z\text{−a.s.} \tag{4.1}$$

The most prominent example of this problem is instrumental variable (IV) regression, where the moment function is given by the prediction residual $\psi(t, y; f) = y - f(t)$ and the conditioning variable Z denotes the instrument. IV regression is one of the major practical approaches to deal

with endogenous variables [126] and has been largely adopted by the causal machine learning community [72, 172, 148, 179, 137].

Learning with conditional moment restrictions is challenging mostly due to two factors. The first one is that equation (4.1) contains a *conditional* expectation over the treatments T and outcomes Y, while one generally has access to a sample from the *joint* distribution over $(T, Y, Z) \sim P_0$. For a sufficiently complex data generating process the accurate estimation of a conditional distribution from the corresponding joint distribution can require large amounts of data [68]. This can be avoided by rewriting the CMR (4.1) in terms of an equivalent *variational* formulation [20]

$$E[\psi(T, Y; f_0)^T h(Z)] = 0 \; \forall h \in \mathcal{H}, \tag{4.2}$$

where $\mathcal{H}$ is a sufficiently rich function space, e.g., the space of square-integrable functions [20] or the reproducing kernel Hilbert space of a certain kind of kernel [92]. While (4.2) avoids the conditional expectation operator, it involves an infinite-dimensional over-determined system of equations. The second difficulty is the fact that the moment restrictions identify the function of interest f_0 via the *population* distribution P_0 of the data, about which one usually only has partial information in terms of a sample $\mathcal{D} = \{(t_i, y_i, z_i)\}_{i=1}^n$ with empirical distribution $\hat{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{(t_i,y_i,z_i)}$, where $\delta_{(t_i,y_i,z_i)}$ denotes a point mass centered at (t_i, y_i, z_i). While the true function f_0 is identified by the population moment restrictions (4.2), it might not satisfy the empirical counterpart of (4.2) and thus one might not retrieve f_0 by enforcing it. Empirical likelihood estimation [122, 121, 130] has been proposed as a flexible tool to solve over-determined moment restriction problems with access to only a finite sample. The idea is based on approximating the population distribution by seeking a distribution with minimal distance to the empirical one for which the moment restrictions can be fulfilled. We visualize this approach in Figure 4.2. The standard generalized empirical likelihood estimator [130] with the extension to conditional moment restrictions of Kremer et al. [92] takes the form $f^{\mathrm{FGEL}} = \arg\min_{f \in \mathcal{F}} R(f)$ with

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

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

4.3 Sinkhorn Method of Moments

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

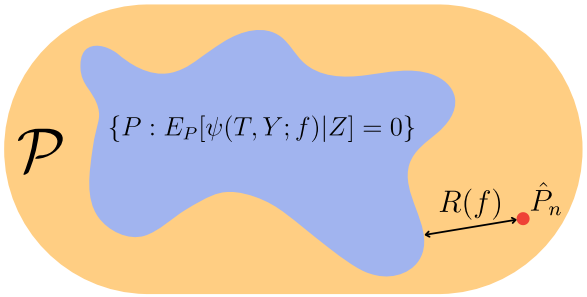


Fig. 4.2 Sinkhorn profile. For every $f \in \mathcal{F}$, the Sinkhorn profile $R(f)$, (4.3), is the minimal distance between the empirical distribution $\hat{P}_n$ and the set of distributions satisfying the CMR (4.1).

Optimal Transport Optimal transport provides an intuitive way of comparing two distributions by means of measuring the minimum effort of transforming one to another by moving probability mass at a certain cost. Let $P \in \mathcal{P}(\Xi)$ and $Q \in \mathcal{P}(\Xi)$ denote two probability distributions over Ξ with densities or probability mass functions (pmf) p and q respectively. Let $\Pi(P, Q) \subset \mathcal{P}(\Xi \times \Xi)$ denote the space of joint probability distributions over the product space $\Xi \times \Xi$ with marginals P and Q. Define the projection operators $\mathbb{P}_1$ and $\mathbb{P}_2$ with $\mathbb{P}_1(x, y) = x$ and $\mathbb{P}_2(x, y) = y$ and their push-forward operation $\mathbb{P}_{i\#}$ such that for any element of $\Pi(P, Q)$, with density (or pmf) π we have $\mathbb{P}_{1\#}\pi = \int \pi(\xi, \xi')\mathrm{d}\xi' = p(\xi)$ and $\mathbb{P}_{2\#}\pi = \int \pi(\xi, \xi')\mathrm{d}\xi = q(\xi')$. Then, for a cost function $c : \Xi \times \Xi \to \mathbb{R}$ we can define the Wasserstein distance between P and Q in the Kantorovich formulation as $W_c(P, Q) := \min_{\pi \in \Pi(P,Q)} \int c(\xi, \xi')\mathrm{d}\pi(\xi, \xi')$. Computation of the Wasserstein distance requires the solution of an infinite-dimensional linear program. In order to enhance its computational efficiency, Cuturi [44] proposed to modify the distance with an entropy regularization penalty,

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

where the relative entropy between π and a reference measure $\mu \otimes \nu \in \mathcal{P}(\Xi \times \Xi)$ is defined as

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

The resulting distance can be efficiently computed with the matrix scaling algorithm of Sinkhorn and Knopp [150], from where it derives its name, Sinkhorn distance. We refer to Peyré et al. [129] for a comprehensive introduction to computational optimal transport for machine learning.

In order to define an estimator for the conditional moment restriction problem (4.1), first, we resort to the functional formulation of Kremer et al. [92]. Let $\mathcal{H}$ denote a sufficiently rich space of functions such that equivalence between (4.1) and (4.2) holds. Then we define the moment functional $\Psi : \mathcal{T} \times \mathcal{Y} \times \mathcal{Z} \times \mathcal{F} \to \mathbb{R}^m$ via its action on $h \in \mathcal{H}$ as $\Psi(t, y, z; f)(h) = \psi(t, y; f)^T h(z)$. This lets us express the CMR (4.1) in its equivalent functional form, $\|E[\Psi(T, Y, Z; f)]\|_{\mathcal{H}^*} = 0$, where $\|\cdot\|_{\mathcal{H}^*}$ denotes the norm in the dual space $\mathcal{H}^*$ of $\mathcal{H}$.

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

$$\begin{aligned} R_\epsilon(f) := \min_{P \in \mathcal{P}} & W_c^\epsilon(P, \hat{P}_n) \\ \text{s.t.} \quad & \|E_P[\Psi(T, Y, Z; f)]\|_{\mathcal{H}^*} = 0. \end{aligned} \tag{4.3}$$

Using Lagrangian duality we can go over to the dual formulation of (4.3) as formalized by the following theorem whose proof is inspired by the mathematically closely related Sinkhorn Distributionally Robust Optimization (DRO) method of Wang et al. [169].

Theorem 4.1 (Duality)**.** *Consider the Sinkhorn profile* (4.3) *with reference measure* $\mu \otimes \nu \in \mathcal{P}(\Xi \times \Xi)$*. Then* (4.3) *has the strongly dual form* $R_\epsilon(f) = \sup_{h \in \mathcal{H}} D(f, h)$ *where*

$$D(f, h) := \underset{\xi' \sim \nu}{E}\left[-\epsilon \log \underset{\xi \sim \mu}{E}\left[e^{-\Psi(\xi; f)(h) - c(\xi, \xi')/\epsilon}\right]\right]. \tag{4.4}$$

In contrast to its original purpose, in our application, the goal of the entropic regularization penalty is not to make computation of the distance more efficient but rather to arrive at a relaxed dual problem (4.4). The dual Sinkhorn profile (4.4) contains expectation operators with respect to the reference distributions μ and ν combined in a non-linear way. This casts optimization of the objective difficult as stochastic gradient estimates will be biased. One way to proceed is to resort to de-biasing techniques as discussed by Wang et al. [169] for their related DRO objective. However, on top of the problem of gradient estimation, computation of (4.4) requires sampling from two reference distributions μ and ν such that accurate gradient estimation becomes costly.

To avoid these issues, we propose an alternative solution for a special choice of reference measures and cost function. Cuturi [44] chooses the reference measure as the product of the marginals of the coupling distribution π. For $W_c^\epsilon(P, Q)$ this corresponds to the choice $\mu \otimes \nu = P \otimes Q$. The choice of μ and ν can be interpreted as a prior for distributions P and Q respectively. Motivated by this, we choose $\nu = \hat{P}_n$ and in order not to restrict the form of P we use an uninformative prior and choose μ as the Lebesgue measure.

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

$$\begin{aligned} c(\xi, \xi') &:= \frac{1}{2}(\xi - \xi')^T \Gamma (\xi - \xi') \\ &= \frac{1}{2} \sum_{w \in \{t, y, z\}} \gamma_w \|w - w'\|_2^2, \end{aligned} \tag{4.5}$$

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

matrix in $\mathbb{R}^{d_i}$. With these choices, the objective (4.4) becomes

$$D(f,h) = E_{\xi' \sim \hat{P}_n}\left[-\epsilon \log E_{\xi \sim \mathcal{N}(\xi', \epsilon\Gamma^{-1})}\left[e^{-\Psi(\xi;f)(h)}\right]\right], \tag{4.6}$$

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

Theorem 4.2. *Let the moment functional $\Psi(\cdot; f) : \Xi \to \mathcal{H}^*$ be continuously differentiable everywhere for any $f \in \mathcal{F}$. Consider the SMM estimator with transport cost function* (4.5) *and reference measure $\hat{P}_n \otimes L$, where L denotes the Lebesgue measure over Ξ. Then, for ϵ/γ_i, $i \in [t, y, z]$, sufficiently small, up to constants and rescalings the objective of the dual Sinkhorn profile* (4.4) *takes the form*

$$\begin{aligned} D(f,h) =& E_{\xi\sim\hat{P}_n}\left[\left(I + \frac{\epsilon}{2}\Delta_\xi\right)\Psi(\xi; f)(h)\right] \\ &- \frac{\epsilon}{2}E_{\xi\sim\hat{P}_n}\left[\|\nabla_\xi\Psi(\xi; f)(h)\|_\Gamma^2\right] + O(\epsilon^{3/2}). \end{aligned} \tag{4.7}$$

Motivated by the classical 2-stage generalized method of moments (GMM) estimator [69] we define the Sinkhorn Method of Moments by substituting the instrument function in the second term in (4.7) by a first stage estimate $\tilde{f}$. We will show below that this does not harm the consistency and convergence properties of our method. Additionally, we add regularization on the instrument function $-\frac{\lambda}{2}\|h\|_{\mathcal{H}}^2$ to ensure that the optimization over h is well behaved on finite samples.

Definition 4.3 (SMM)**.** *Let $\tilde{f} \in \mathcal{F}$ denote a first-stage estimate of $f_0 \in \mathcal{F}$, then we define the Sinkhorn Method of Moments (SMM) estimator as the solution of the saddle-point problem*

$$f^{\mathrm{SMM}} = \arg\min_{f\in\mathcal{F}} \max_{h\in\mathcal{H}} M(f,h) - \epsilon\mathcal{R}(\tilde{f}, h) \tag{4.8}$$

with

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

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

By using the 2-stage GMM-style estimator we shift most of the computational complexity into the optimization of the instrument function $h \in \mathcal{H}$. The optimization over the possibly high-dimensional model remains simple and even is a convex program, whenever f has a convexity preserving parameterization, e.g., for linear models. In practice, if (4.8) is optimized with stochastic

gradient methods, one can dynamically update the first stage estimate $\tilde{f}$ using the result from the previous iteration. In the context of CMR estimation this GMM-inspired two stage procedure is a popular approach to stabilize the training [15, 101, 14]. Note that without the 2-stage adaptation we would obtain an estimator similar in spirit to the continuous updating GMM estimator of Hansen et al. [70] or the FGEL estimator of Kremer et al. [92], which can be harder to train in practice [67].

The objective (4.8) involves a gradient and a Laplacian with respect to the data, which allows the method to take into account the geometry of the moment violation with respect to the data manifold. As we maximize the objective over $h \in \mathcal{H}$, we promote instrument functions which correspond to local minima of the moment violation $\psi(t, y; f)^T h(z)$ with respect to the data. Generally for CMR estimators the instrument function is responsible for translating the data into a learning signal for the model f. Choosing h in a local minimum w.r.t. the data means that we attribute less importance to data points that lead to large increases in the moment violation when perturbed slightly. This makes the model less vulnerable to poisoned data and adversarial attacks. SMM's property to take into account how the learning signal changes in proximity of the data is unique compared to related estimators which are blind towards the geometry of the data manifold as they are based on reweighting the existing data [15, 101, 50, 14, 92] or sampling additional data points [93] respectively.

4.3.1 Consistency

The following assumptions allow us to guarantee consistency and derive a convergence rate of our 2-stage estimator (4.8) in the parametric, uniquely identified setting. Suppose there exists a unique parameter $\theta_0 \in \Theta \subseteq \mathbb{R}^p$ for which $E[\psi(T, Y; \theta_0)|Z] = 0$ P_Z-a.s.. In the following, let $x \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ denote the concatenation of $(t, y) \in \mathcal{T} \times \mathcal{Y}$ and let $i \in [m]$ be a shorthand for $i \in \{1, \dots, m\}$. Further, we define the Jacobian of a vector-valued function $\psi : \mathcal{X} \times \Theta \to \mathbb{R}^m$ as $J_x\psi(x; \theta) \in \mathbb{R}^{m \times d_x}$.

Assumption 4.1 (Identifiability)**.** *$\theta_0 \in \Theta$ is the unique solution to $E[\psi(X; \theta)|Z] = 0$ P_Z-a.s.; Θ is compact; $\psi(X; \theta)$ is continuous in θ everywhere w.p.1.*

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

Assumption 4.2 (Data regularity)**.** *The space $\Xi = \mathcal{T} \times \mathcal{Y} \times \mathcal{Z} \subset \mathbb{R}^{d_\xi}$ is compact.*

Assumption 4.3 (Smoothness w.r.t. data)**.** *The moment function $\psi(\cdot; \theta) : \mathcal{T} \times \mathcal{Y} \to \mathbb{R}^m$ is C^∞-smooth in the data for every $\theta \in \Theta$. Further the sets of functions $\{\psi(\cdot; \theta)_l : \theta \in \Theta\}$ and $\{(J_x\psi(\cdot; \theta))_{lr} : \theta \in \Theta\}$, are P_0-Donsker for every $l \in [m]$ and $r \in [d_x]$.*

Assumption 4.2 and 4.3 ensure that the moment function and its derivatives are well-behaved with respect to the data. While the compactness of the data space might be violated in practice, usually one can construct a sufficiently large compact set that contains the data with high probability.

Assumption 4.4. *The matrix $V(Z; \theta) \in \mathbb{R}^{m \times m}$ defined as*

$$V(Z; \theta) = E[J_x\psi(X; \theta)\Gamma^{-1}J_x\psi(X; \theta)^T|Z] \tag{4.9}$$

is non-singular for $\theta \in \{\theta_0, \bar{\theta}\}$ w.p.1, where $\bar{\theta}$ is an initial parameter estimate defined in Assumption 4.6.

This corresponds to the common assumption of a non-singular covariance matrix required by related estimators [118, 14, 92], but, here, imposed on the covariance of the data-Jacobian.

Assumption 4.5 (Instrument function)**.** *$\mathcal{H} = \bigoplus_{l=1}^{m} \mathcal{H}_l$ is a a sufficiently rich space of vector-valued functions such that equivalence between* (4.1) *and* (4.2) *holds. Further for $l \in [m]$, $h \in \mathcal{H}_l$ is C^∞-smooth and the unit ball $\mathcal{H}_{l,1} := \{h \in \mathcal{H}_l : \|h\|_{\mathcal{H}_l} \leq 1\}$ as well as $\{J_z h : h \in \mathcal{H}_{l,1}\}$ are P_0-Donsker.*

This is fulfilled, for example, by choosing each $\mathcal{H}_l$ as the RKHS of a universal, integrally stricly positive definite kernel, e.g., the Gaussian kernel, which we will formalize later. For neural network instrument function classes, equivalence between the variational and conditional formulations can be shown on basis of universal approximation theorems [174, 175]. In this case $\mathcal{H}_{l,1}$, C^∞-smoothness can be realized by using smooth activation functions.

Assumption 4.6 (Regularization)**.** *There is a first-stage parameter estimate $\bar{\theta}_n \xrightarrow{p} \bar{\theta}$ for which $E\left[\|\psi(X;\bar{\theta}_n) - \psi(X;\bar{\theta})\|_\infty\right] = O_p(n^{-\zeta})$ and $E\left[\|J_x\psi(X;\bar{\theta}_n) - J_x\psi(X;\bar{\theta})\|_\infty\right] = O_p(n^{-\zeta})$ with $0 < \zeta \leq 1/2$. Choose $\lambda_n = O_p(n^{-\rho})$ with $0 < \rho < \zeta$.*

For linear IV regression this implies $\|\bar{\theta}_n - \bar{\theta}\|_\infty = O_p(n^{-\zeta})$, which means $\bar{\theta}_n$ has to be a $n^{-\zeta}$-consistent estimator for $\bar{\theta}$, which can be any parameter for which (4.9) is non-singular, e.g., the true parameter θ_0.

Assumption 4.7 (Smoothness w.r.t. θ)**.** *$\theta_0 \in \mathrm{int}(\Theta)$; $\psi(x;\theta)$ is continuously differentiable in a neighborhood $\bar{\Theta}$ of θ_0; and $E[\sup_{\theta\in\bar{\Theta}} \|J_\theta\psi(X;\theta)\|^2|Z] < \infty$ w.p.1; $\mathrm{rank}\left(E[J_\theta\psi(X;\theta_0)|Z]\right) = p$, w.p.1.*

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

With that, we are ready to state the consistency theorem for our estimator.

Theorem 4.4 (Consistency)**.** *Let Assumptions 4.1-4.6 be satisfied. For any $0 < \epsilon_1 < \epsilon_2$, choose $\epsilon \sim \mathrm{Uniform}([\epsilon_1, \epsilon_2])$. Then the SMM estimator $\hat{\theta}$ converges to the true parameter θ_0 in probability $\hat{\theta} \xrightarrow{p} \theta_0$.*

If additionally Assumption 4.7 is satisfied, then $\|\hat{\theta} - \theta_0\| = O_p(n^{-1/2})$.

The consistency result is independent of the choice of instrument function space $\mathcal{H}$ as long as it fulfills Assumption 4.5. We now discuss two different implementations based on kernel methods and neural networks.

4.3.2 Kernel-SMM

Choosing $\mathcal{H}$ as the RKHS of a suitable kernel, we can guarantee equivalence between the conditional and variational moment restrictions formulations (4.1) and (4.2). On top of that, for RKHS instrument functions we can employ a representer theorem and carry out the optimization over the instrument

function $h \in \mathcal{H}$ in closed form. The resulting estimator can be obtained as the solution of a simple minimization problem bearing close resemblance to the optimally weighted 2-stage GMM estimator but taking into account the geometry of the moment violation with respect to the data. Before deriving the result we provide the necessary background on reproducing kernel Hilbert spaces (RKHS).

Reproducing kernel Hilbert space An RKHS $\mathcal{H}$ is a Hilbert space of functions $h : \mathcal{Z} \to \mathbb{R}$ in which point evaluation is a bounded functional. With every RKHS one can associate a positive semi-definite kernel $k(\cdot, \cdot) : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ with the reproducing property, i.e., for any $h \in \mathcal{H}$ we have $h(z) = \langle h, k(z, \cdot)\rangle_{\mathcal{H}}$. A kernel is called universal if its RKHS is dense in the set of all continuous real-valued functions [106]. Further, a kernel is called integrally strictly positive definite (ISPD) if for any $h \in \mathcal{H}$ with $0 < \|h\|^2_{\mathcal{H}} < \infty$, we have $\int_{\mathcal{Z}} h(z)k(z, z')h(z')\mathrm{d}z\mathrm{d}z' > 0$. Refer to, e.g., Schölkopf and Smola [140] and Berlinet and Thomas-Agnan [18] for comprehensive introductions.

The following proposition specifies the properties of an RKHS for which Assumption 4.5 is satisfied.

Proposition 4.5. *Let $\mathcal{Z} \subset \mathbb{R}^{d_z}$ be compact. Then, the instrument function space $\mathcal{H} = \bigoplus_{l=1}^m \mathcal{H}_l$, where each $\mathcal{H}_l$ corresponds to the RKHS of universal, integrally strictly positive definite kernel k_l, $l \in [m]$ fulfills Assumption 4.5.*

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

Theorem 4.6 (Kernel-SMM)**.** *Let $\mathcal{H} = \bigoplus_{l=1}^m \mathcal{H}_l$ be the direct sum of m reproducing kernel Hilbert spaces with kernels $k_l : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$. Let $\tilde{f} \in \mathcal{F}$ denote a first stage estimate of f_0 and let $\gamma_z = \infty$. Define $\psi_\Delta(f) \in \mathbb{R}^{nm}$, $L \in \mathbb{R}^{nm\times nm}$ and $Q(f) \in \mathbb{R}^{nm\times nm}$ with entries*

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

Then the Sinkhorn profile is given by

$$R_{Q(\tilde{f})}(f) = \frac{1}{2n^2}\psi_\Delta(f)^T L\left(Q(\tilde{f}) + \frac{\lambda}{\epsilon}L\right)^{-1} L\psi_\Delta(f). \tag{4.10}$$

Compared to the general saddle point formulation (4.8) the kernelized version (4.10) has the significant advantage that it only involves a minimization over the model parameters and thus avoids the

Algorithm 2 n-stage Kernel-SMM

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

difficulties of mini-max optimization [46]. Algorithm 2 details the implementation of the multi-stage Kernel-SMM approach. In order to minimize the number of hyperparameters, we implement the gradient descent step with the limited memory BFGS method [103]. We empirically observed that the n-step estimator effectively converges with the second iteration.

4.3.3 Neural-SMM

A particularly interesting alternative choice of instrument function space are neural network classes, as they can represent highly flexible functions while allowing for optimization via mini-batch stochastic gradient methods. As demonstrated by related works [101, 15, 92], such neural network-based approaches can lead to powerful and scalable estimators that may outperform the corresponding kernel method on large samples. On the downside, they tend to be difficult to train due to the instability and hyperparameter sensitivity of mini-max optimization. This is particularly problematic for IV regression, as in contrast to standard supervised learning, it is non-trivial to define suitable validation metrics to set these hyperparameters. As a result, compared to (4.10), those estimators require more attention and careful evaluation which makes them less suitable as plug-and-play IV estimators for practitioners. As the primary focus of this work is to introduce a new geometry-aware learning paradigm for IV regression independent of the instrument function class, we consider the simpler kernel version in the following and defer results for the Neural-SMM estimator to Appendix C.2.

4.4 Experimental Results

We benchmark the kernel version of our method against a selection of plug-and-play IV estimators including maximum moment restrictions (MMR) [179], sieve minimum distance (SMD) [2] as well as the kernel variational method of moments (VMM) [14]. Results for the neural network version and related estimators can be found in Appendix C.2. For all kernel methods we choose a radial basis function kernel $k(z, z') = \exp(-\eta \|z - z'\|_2^2)$, where we set η according to the median heuristic [57]. The remaining hyperparameters of all methods are set by using the MMR objective on a validation data set (see Appendix C.1). In all experiments we consider perturbations in the treatment variable t

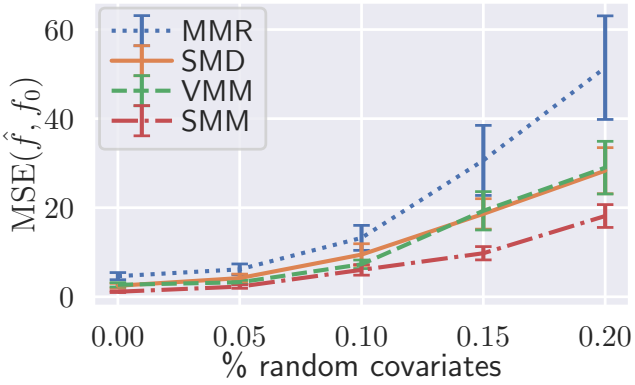


Fig. 4.3 Robustness against corrupted data. We generate 1000 points from the process (4.11) and substitute in a proportion of the data the treatment variable T for a random value sampled uniformly over the domain. Lines and error bars correspond to the mean and standard error computed over 20 training datasets.

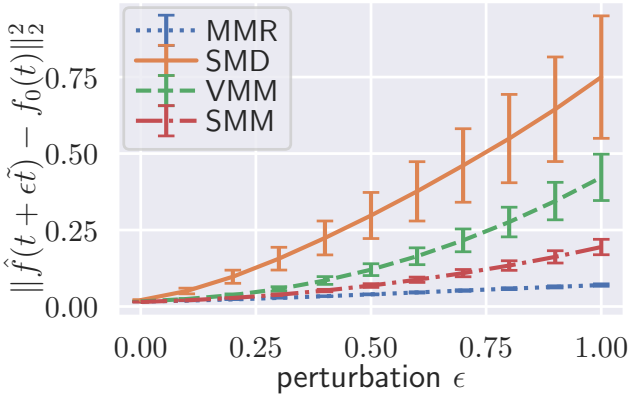
and fix the other variables at their empirical values by setting $\gamma_y, \gamma_z = \infty$ for SMM. Code will be released upon publication.

IV Regression with Corrupted Data We consider the SimpleIV experiment of Bennett and Kallus [14] with the following data generating process,

$$\begin{aligned} Z &= \sin(\pi Z_0/10) \\ T &= -0.75Z_0 + 3.5H + 0.14\eta - 0.6 \\ Y &= f(T;\theta_0) - 10U + 0.1\eta_2 \end{aligned} \tag{4.11}$$

where $\eta_1, \eta_2, U \sim N(0, I)$ and $Z_0 \sim \mathrm{Uniform}([-5, 5])$. The model is given by $f(t;\theta) = \theta^1 t^2 + \theta^2 t + \theta^3$ with $\theta_0 = [3.0, -0.5, 0.5]$. This is a typical IV problem, where the unobserved confounder U induces a non-causal dependence between T and Y. To investigate the robustness against corrupted data, we sample training sets of 1000 points and exchange a proportion of the covariates T by random values generated according to $\mathrm{Uniform}([t_{\min}, t_{\max}])$. Figure 4.3 shows the mean-squared error of the models trained with different methods over the proportion of random covariates in the training data. We observe that for no data corruption, all estimators perform similarly, with SMM providing a small advantage. With increasing proportion of corrupted data, SMM scales favorably compared to the baselines. We provide more details and a hyperparameter sensitivity analysis in Appendix C.1.

Adversarially Robust IV Regression We test the adversarial robustness of different IV estimators in the following setting. Define $C = 0.2I \in \mathbb{R}^{5\times 1}$, as well as $B \in \mathbb{R}^{5\times 1}$, with fixed entries sampled



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

Fig. 4.4 Adversarial robustness of IV estimators. We use a training set of size $n = 1000$ and evaluate the learned models over FGSM attacks with increasing strength ϵ. Lines and error bars show the mean and standard error over 20 random training datasets. The table contains the MSE in the perturbation-free case.

from $\mathrm{Uniform}([0.1, 0.3])$. Consider the non-linear data generating process,

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

with $U \sim N(0, 1)$, $\eta_1, \eta_2 \sim N(0, 0.1)$ and $f_0(t) = 1.5\cos(At) + 0.1At$, where $A \in \mathbb{R}^{1\times 5}$ with fixed entries sampled from $\mathrm{Uniform}([-1.5, 1.5])$. We approximate f_0 with a feed-forward neural network with $[20, 20, 3]$ hidden units and leaky ReLU activation functions. We train the network using different plug-and-play IV estimators and evaluate the adversarial robustness by running FGSM attacks [60] in directions $\tilde{t}$ with strength $\epsilon \in [0, 1.0]$. Figure 4.4 shows that all IV estimators yield comparable mean-squared errors for $\epsilon = 0$, clearly improving over the non-causal least squares (LSQ) solution (table). Moreover, for increasing attack strengths ϵ, we see that SMM demonstrates stronger adversarial robustness than the SMD and VMM estimators. Interestingly, here, the MMR estimator which performed worse in the first experiments exhibits the least sensitivity towards adversarial perturbations. This might be understood by the fact that the MMR estimator corresponds to the limit case of SMM and VMM for $\lambda \to \infty$. Generally, strong regularization promotes flat functions which are less sensitive to the inputs, which might explain MMR's superior robustness here.

In Appendix C.2 we provide results on a common modern IV benchmark that provides further evidence that SMM performs on par with state-of-the art estimators in standard IV settings. In this context, we also provide results for a Neural-SMM estimator, which proves to be competitive with state-of-the art deep learning approaches [15, 92].

4.5 Related Work

Instrumental variable regression has traditionally been addressed via the 2-stage least squares (2SLS) method, which limits both regression stages to linear models [4]. Extensions to non-linear models have been provided by multiple works [3], recently based on density estimators [72, 148] and deep features [172]. As an alternative to 2SLS, estimators based on the conditional moment restriction formulation have been used based on either basis function expansions of L^2 [2, 26, 28, 120] or machine learning models [15, 50, 110, 92, 93, 14]. Related to our Kernel-SMM estimator, multiple works have used RKHS functions as instrument models [26, 148, 14, 179], leading to similar formulations as our (4.10). However, in contrast to ours, none of them take into account the geometry of the moment violation with respect to the data.

Optimization over measure spaces by means of minimizing some notion of distributional distance between the optimization variable and an empirical distribution has recently attracted significant attention in the context of distributionally robust optimization [54, 149, 107, 52, 95, 53]. On a higher level, one can distinguish between three types of approaches based on the respective distance notion (cf. Figure 4.1): φ-divergences restrict the optimization variable to a finite dimensional vector of weights attributed to the data points and thus find optimal reweightings of the sample. Methods based on maximum-mean discrepancy [64] and the Fisher-Rao metric [11], allow for creation and annihilation of probability mass [183, 93, 173]. Finally, methods based on optimal transport distances effectively allow to move around the data points in the data space [107, 149]. While CMR estimation has been based on the previous two paradigms, to the best of our knowledge, our Sinkhorn Method of Moments is the first estimator based on the latter category.

In a different context, empirical likelihood has previously been combined with Wasserstein distances to calibrate the radius of ambiguity sets in distributionally robust optimization (DRO) [21]. However, their method does not extend to CMR estimation and neither does it make use of a regularized duality structure. From a mathematical perspective the derivation of our first duality result (Theorem 4.1) closely resembles the derivation of the dual Sinkhorn DRO estimator of Wang et al. [169], which, nevertheless, addresses an entirely different problem. In addition, Wang et al. [169] relies on debiasing techniques to optimize their objective, whereas we provided a form that can be directly optimized via stochastic gradient methods.

4.6 Conclusion

Instrumental variable regression is an important concept in the field of causal inference, which motivates the development of estimators adapted to the intricacies of real-world datasets. Notwithstanding recent mini-max estimators based on neural network instrument function classes showing convincing performance on benchmarks [50, 15, 92, 93], there remains a need for simple plug-and-play estimators that can be trained by practitioners without deep technical knowledge and with a manageable set of hyperparameters. We have extended the repertoire of such estimators by a method whose learning signal arises from an optimal transport geometry in the data space. We showed that our estimator exhibits favorable properties in presence of corrupted data or adversarial examples while maintaining

performance competitive with state-of-the art approaches on standard benchmarks. The simplicity of our plug-and-play estimator partially results from its kernel-based implementation which limits the scalability to large sample sizes. To address this, we provide a neural network-based implementation in the appendix, whose detailed analysis is left for future work.

Chapter 5

Robust Chance-Constrained Optimization Using Reproducing Kernel Hilbert Spaces*

Chance-constrained programming has recently been combined with distributional robustness based on the popular Wasserstein ambiguity sets. However, the involved computational techniques typically place restrictive assumptions on the constraint functions. Moreover, the size of the Wasserstein ambiguity sets is often set using costly cross-validation (CV) procedures or conservative measure concentration bounds. To address these shortcomings, we propose a practical distributionally robust chance constraint programming (DRCCP) algorithm using kernel *maximum mean discrepancy* (MMD) ambiguity sets, which we term MMD-DRCCP, to treat general nonlinear constraints without using ad-hoc reformulation techniques. MMD-DRCCP can handle general nonlinear and non-convex constraints with a proven finite-sample constraint satisfaction guarantee of a dimension-independent $1/\sqrt{n}$ rate, achievable by a practical algorithm. We further propose an efficient bootstrap scheme for constructing sharp MMD ambiguity sets in practice without resorting to CV. Our algorithm is validated numerically on a portfolio optimization problem and a tube-based distributionally robust model predictive control problem with non-convex constraints.

*Based on *Maximum Mean Discrepancy Distributionally Robust Nonlinear Chance-Constrained Optimization with Finite-Sample Guarantee* [113]. Yassine Nemmour*, Heiner Kremer*, Bernhard Schölkopf, Jia-Jie Zhu. IEEE Conference for Decision and Control (CDC) 2022; *equal contribution

Declaration

This chapter is based in parts or in full on the published manuscript:

> *Maximum Mean Discrepancy Distributionally Robust Nonlinear Chance-Constrained Optimization with Finite-Sample Guarantee*
> Yassine Nemmour*, Heiner Kremer*, Bernhard Schölkopf, Jia-Jie Zhu
> IEEE Conference for Decision and Control (CDC) 2022
> *equal contribution

Author contributions

Yassine Nemmour: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization
Heiner Kremer: Methodology, Software, Validation, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization
Bernhard Schölkopf: Supervision, Resources
Jia-Jie Zhu: Conceptualization, Methodology, Formal Analysis, Investigation, Writing - Review & Editing, Supervision

Table 5.1 Comparison of recent DRCCP works using Wasserstein ambiguity sets and our work.

Constraint function w.r.t. ξ	Approach
Affine	MIP: [171, 76, 35, 81], CVaR: [171, 78, 66]
Quadratic	CVaR: [66]
Concave	Cutting-plane algorithm: [78]
Convex (known Lipschitz constant for every x)	Convex inner approximation: [78]
General nonlinear	**This work**

5.1 Introduction

Chance-constrained programs (CCP) are optimization problems involving uncertainty via probabilistic constraints. Compared to their robust optimization counterparts which enforce a set of constraints to hold for any realization of the uncertainty, CCPs seek solutions that satisfy the constraints only with high probability. This allows to find less conservative solutions and thus improved objective values. Computing an exact solution of a CCP requires knowledge of the underlying probability distribution of the uncertain variables, which is generally unavailable and can only be estimated from data. To account for estimation errors, chance constraint programming has recently been combined with distributionally robust optimization (DRO) [47]. DRO addresses uncertain optimization problems by finding a solution for the worst-case distribution from a set of distributions, the so-called ambiguity set. Recent works construct the ambiguity set as balls around the empirical distribution in the Wasserstein distance [171, 78]. Within the Wasserstein ambiguity set DRCCP framework, with the notable exceptions of Hota et al. [78], Gu and Wang [66], most works limit the class of constraint functions to be affine in the uncertain variable [171, 35, 76], which severely restricts their application in practice. In addition, these methods usually set the radius of the ambiguity set via cross-validation (CV) procedures which is computationally expensive and thus rarely used in large-scale learning tasks, such as deep learning and problems involving complex simulations.

In order to address some of these shortcomings, we propose a DRCCP framework for general nonlinear constraints based on ambiguity sets constructed using maximum mean discrepancy (MMD), which we term (MMD-DRCCP). Different from DRCCP based on Wasserstein distances, our methodology does not rely on ad-hoc reformulation techniques that are highly dependent on the constraint function classes and their specific closed-form support functions, if they exist. Instead, we use Hilbert spaces generated using expressive kernels as universal function approximators to treat general nonlinear constraints with a single unified reformulation technique. In Table 5.1 we provide an overview of recent DRCCP approaches and their respective limitations.

The rest of the paper is structured as follows: First we define the MMD-DRCCP framework in Section 5.2 and show how in contrast to Wasserstein based approaches one can construct tight MMD ambiguity sets using a bootstrap approach. In Sections 5.2.4 and 5.2.5 we provide an exact reformulation of the DRCCP as well as a Conditional Value-at-Risk (CVaR) formulation that allows for arbitrary constraint classes. For the CVaR formulation we then derive a finite sample constraint

satisfaction guarantee with a dimension independent error rate of order $O(n^{-1/2})$ in Section 5.3. Finally we provide numerical results in Section 5.4.

5.2 MMD-DRCCP

In this section, we introduce DRCCPs with MMD ambiguity sets. We discuss how, in contrast to the Wasserstein ambiguity sets, we can construct MMD ambiguity sets in a principled way using a practical bootstrap approach. We derive an exact reformulation of our problem based on the strong duality result of Zhu et al. [183] and provide a tractable CVaR relaxation. Before formally introducing the problem, we provide necessary background material on reproducing kernel Hilbert spaces.

5.2.1 Reproducing Kernel Hilbert Spaces

A kernel is a similarity measure defined by a symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and said to be positive definite (PD) if $\sum_{i=1}^{n} \sum_{j=1}^{n} a_j a_i k(x_i, x_j) \geq 0$ for any $n \in \mathbb{N}$, $\{x_i\}_{i=1}^{n} \subset \mathcal{X}$ and $\{a_i\}_{i=1}^{n} \subset \mathbb{R}$. For every PD kernel there exists a feature map $\phi : \mathcal{X} \to \mathcal{H}$ taking values in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ such that $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes an inner product on $\mathcal{H}$. The inner product induces a norm via $\|f\|_{\mathcal{H}} \coloneqq \sqrt{\langle f, f \rangle_{\mathcal{H}}}$. For a given distribution P one denotes the kernel mean embedding (KME) as $\mu_P \coloneqq \int k(x, \cdot) dP$. Equipped with these tools, one can define a metric between two distributions P, Q as $\|\mu_P - \mu_Q\|_{\mathcal{H}}$, which corresponds to the maximum mean discrepancy (MMD) [65]. Using the reproducing property of an RKHS one can express this metric as $\mathrm{MMD}^2(P, Q; \mathcal{H}) = \mathbb{E}_{x,x' \sim P} k(x, x') + \mathbb{E}_{y,y' \sim Q} k(y, y') - 2\mathbb{E}_{x \sim P, y \sim Q} k(x, y)$. Having access to this closed-form expression is a major advantage of MMD since computing Wasserstein distances requires solving an infinite-dimensional linear program. Refer to [18, 140, 170, 154, 158] for comprehensive introductions to kernel methods.

5.2.2 Chance Constraint Programs with MMD Ambiguity Sets

Next, we formally introduce CCPs and DRCCPs. For simplicity, we restrict ourselves to scalar-valued constraint functions. Let $f : \mathcal{X} \subset \mathbb{R}^n \times \Xi \subseteq \mathbb{R}^m \to \mathbb{R}$ denote a function that defines an uncertain inequality constraint $f(x, \xi) \leq 0$ depending on a random variable $\xi \in \Xi$. In particular, we do not exclude the possibility that f may be nonlinear, non-convex, or semi-continuous. Without loss of generality we consider the linear cost function $c^T x$, with $c \in \mathbb{R}^n$. A CCP with risk level $\alpha \geq 0$ is then defined as

$$\begin{aligned} &\min_{x \in \mathcal{X}} c^T x \\ &\text{subject to } P_0[f(x, \xi) \leq 0] \geq 1 - \alpha. \end{aligned} \tag{5.1}$$

This has the interpretation that the inequality constraints can be violated with probability of at most α. Since generally the underlying data distribution P_0 is unknown and one often only has access to a sample from it, we expand this formulation to its distributionally robust counterpart. We consider a worst-case distribution within a set of plausible distributions, the so-called ambiguity set, and define

the DRCCP as

$$\begin{aligned} &\min_{x\in\mathcal{X}} c^T x \\ &\text{subject to } \inf_{P\in\mathcal{P}} P[f(x,\xi)\leq 0] \geq 1-\alpha. \end{aligned} \tag{5.2}$$

We construct the maximum mean discrepancy (MMD)-based ambiguity set as a ball of radius ε centered at the empirical distribution $\hat{P}_n$. Given samples $\{\xi\}_{i=1}^N$ of the true distribution P, the empirical distribution is then given by $\hat{P}_n = \sum_{i=1}^N \delta_{\xi_i}$. As the sample size goes to infinity, the empirical distribution $\hat{P}_n$ of the sample converges to the true distribution by the weak law of large numbers. Thus, the radius ε should be chosen in a data-driven way that reflects the confidence according to the sample size N. In the rest of the paper, we denote the MMD ambiguity set by

$$\mathcal{P} := \{P : \mathrm{MMD}(P, \hat{P}_n; \mathcal{H}) \leq \varepsilon\}. \tag{5.3}$$

If ε is chosen large enough such that the true distribution is contained in $\mathcal{P}$, then the solution of the DRCCP (5.2) will also satisfy the constraint of the original CCP (5.1). Notably, MMD ambiguity sets enable us to set the ambiguity radius a priori in a few simple-yet-principled ways not available for the Wasserstein counterpart. Using the estimation error bound of MMD estimators [163, 65], we have that, with probability $1-\delta$, the population distribution P_0 is contained in an MMD ball around the empirical distribution of radius

$$\mathrm{MMD}(P_0, \hat{P}_n; \mathcal{H}) \leq \sqrt{\frac{C}{N}} + \sqrt{\frac{2C\log(1/\delta)}{N}}, \tag{5.4}$$

where C is a constant such that $\sup_x k(x,x) \leq C < \infty$. For the common Gaussian kernel, we have $C = 1$. Note that (5.4) is dimension-free. With this result, we can simply set the radius of the MMD ambiguity set to the RHS of (5.4). In practice, however, concentration bounds such as (5.4) are often overly conservative [65]. Therefore, instead of relying on a fixed concentration rate we propose an MMD bootstrap scheme to obtain tighter confidence intervals.

5.2.3 Bootstrap Construction of MMD Ambiguity Sets

We propose to construct the bootstrap MMD ambiguity set in a similar fashion as for the two-sample test of Gretton et al. [65], based on the results of Arcones and Gine [8] for general degenerate V-statistics. Let $\{\tilde{\xi}_i\}_{i=1}^N$ denote a bootstrap sample of $\hat{P}_n$ with distribution $\tilde{P}$, i.e., drawn with replacement from $\{\xi_i\}_{i=1}^N$. Define the (biased) MMD estimator as $\widehat{\mathrm{MMD}}^2(\tilde{P}, \hat{P}_N; \mathcal{H}) = \sum_{i,j=1}^N k(\xi_i,\xi_j) + k(\tilde{\xi}_i,\tilde{\xi}_j) - 2k(\xi_i,\tilde{\xi}_j)$. Then by the weak law of large numbers and the bootstrap result for V-statistics of Arcones and Gine [8] we have that $\widehat{\mathrm{MMD}}(\tilde{P}, \hat{P}_n; \mathcal{H}) \xrightarrow{d} \mathrm{MMD}(P_0, \hat{P}_n; \mathcal{H})$ as $N \to \infty$. For a fixed confidence level β, this lets us determine the radius ε of the uncertainty set as the β-quantile of the bootstrap distribution $\widehat{\mathrm{MMD}}(\tilde{P}, \hat{P}_n; \mathcal{H})$ (see Figure 5.1). Details on the procedure can be found in Algorithm 3. We emphasize that bootstrap techniques for MMD have been used in large-scale machine learning tasks for high-dimensional data and are not available for Wasserstein ambiguity sets due to the lack of closed-form estimators.

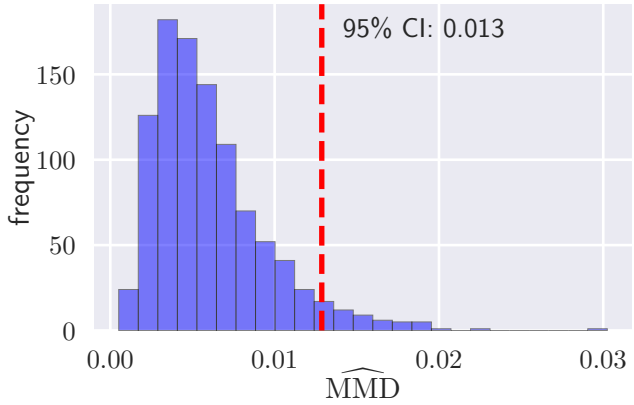


Fig. 5.1 Bootstrap construction of the MMD ambiguity set. We exemplarily sample $N = 100$ points from a standard normal distribution and compute bootstrap estimates of $\mathrm{MMD}(P_0, \hat{P}_n)$ over $B = 1000$ bootstrap samples using Algorithm 3. We set the radius of the ambiguity set to the $\beta = 95\%$ quantile $\varepsilon = 0.013$. For comparison, using instead 10000 additional samples from P_0 to estimate $\mathrm{MMD}(P_0, \hat{P}_n)$ yields $\varepsilon = 0.010$. We conclude that in this case with high probability the true distribution is contained in our bootstrap MMD ambiguity set.

Algorithm 3 Bootstrap MMD ambiguity set

Input: SampleS $\{\xi\}_{i=1}^N$, Number of bootstrap samples B, Confidence level β
$K \leftarrow \mathrm{kernel}(\xi, \xi)$
for $m = 1, \ldots, B$ **do**
  Draw a set I of N numbers from $\{1, \ldots, N\}$ with replacement
  $K_x \leftarrow \sum_{i,j=1}^N K_{ij}, \qquad K_y \leftarrow \sum_{i,j \in I} K_{ij}$
  $K_{xy} \leftarrow \sum_{j \in I} \sum_{i=1}^N K_{ij}$
  $\mathrm{MMD}[m] \leftarrow \frac{1}{N}\sqrt{K_x + K_y - 2K_{xy}}$
end for
$\mathrm{MMD} \leftarrow sort(\mathrm{MMD})$
$\varepsilon \leftarrow \mathrm{MMD}[ceil(B\beta)]$
Output: Radius of MMD ambiguity set ε

5.2.4 Exact Reformulation

Given the MMD ambiguity set $\mathcal{P}$, we denote the feasible set Z of the DRCCP (5.2) as

$$Z := \big\{x \in \mathcal{X} : \inf_{P \in \mathcal{P}} P\{f(x, \xi) \leq 0\} \geq 1 - \alpha\big\}, \tag{5.5}$$

where $\mathcal{P}$ is the MMD ambiguity set defined in (5.3). Since for MMD ambiguity sets, the infimum in (5.5) is challenging to compute, we first summon the strong duality result proved by Zhu et al. [183] to embed the chance constraint into an RKHS. With the dual form of the constraint, the DRCCP becomes a kernel machine learning problem in finding an RKHS function that majorizes $1(f(x, \xi) \leq 0)$, where 1 denotes the indicator function. We visualize this idea in Figure 5.2 and formalize it in the following proposition.

Proposition 5.1. *The feasible set* (5.5) *of the MMD-DRCCP* (5.2) *has the dual form*

$$Z := \left\{ x \in \mathcal{X} : \begin{array}{ll} g_0 + \frac{1}{N}\sum_{i=0}^{N} g(\xi_i) + \varepsilon\|g\|_{\mathcal{H}} \leq \alpha & (5.6a) \\ \mathbb{1}(f(x,\xi) > 0) \leq g(\xi) + g_0 \; \forall \xi \in \Xi & (5.6b) \\ g \in \mathcal{H}, g_0 \in \mathbb{R} & (5.6c) \end{array} \right\}$$

Proof. Note that we can replace $\inf_{P\in\mathcal{P}} P[f(x,\xi) \leq 0] \geq 1-\alpha$ with the equivalent $\sup_{P\in\mathcal{P}} P[f(x,\xi) > 0] \leq \alpha$. Later, we will introduce an inner approximation scheme based on this result.

Next, we can rewrite the probability with an indicator function

$$P[f(x,\xi) > 0] = \mathbb{E}_P[\mathbb{1}\big(f(x,\xi) > 0\big)].$$

The indicator function fulfills the assumptions on the constraint function of the strong duality result of Zhu et al. [183], with the Slater condition trivially satisfied. Using Theorem 3.1 of Zhu et al. [183] we can rewrite $\sup_{P\in\mathcal{P}} \mathbb{E}_P[\mathbb{1}\big(f(x,\xi) > 0\big)]$ as

$$\min_{g\in\mathbb{R}, g\in\mathcal{H}} \quad g_0 + \frac{1}{N}\sum_{i=0}^{N} g(\xi_i) + \varepsilon\|g\|_{\mathcal{H}} \tag{5.7}$$

$$\text{subject to} \quad \mathbb{1}\big(f(x,\xi) > 0\big) \leq g_0 + g(\xi) \quad \forall \xi \in \Xi. \tag{5.8}$$

The result follows by plugging this expression into (5.5). □

Solving MMD-DRCCP using this exact reformulation is intractable in practice. As a consequence, we will investigate a convex CVaR approximation, which can be solved with off-the-shelf solvers. Note that, unlike the approximations in Hota et al. [78], Chen et al. [35], Xie [171], our approximation still holds for general nonlinear $f(x,\xi)$ and only requires mild assumptions on the dependence on the uncertainty ξ. For more technical details, we refer to Zhu et al. [183, Theorem 3.1].

5.2.5 CVaR Approximation

In this section we present a convex inner approximation of the feasible set (5.6) based on the Conditional Value-at-Risk. First, note that we can rewrite the chance constraint in (5.1) equivalently in terms of the Value-at-Risk (VaR) which is defined as

$$\mathrm{VaR}^{P_0}_{1-\alpha}[f(x,\xi)] = \inf\{t \in \mathbb{R} : P_0[f(x,\xi) \leq t] \geq 1-\alpha\},$$

where $f(x,\xi)$ is to interpreted as a random variable. With this definition, it is straightforward to observe

$$\mathrm{VaR}^{P_0}_{1-\alpha}[f(x,\xi)] \leq 0 \iff P_0[f(x,\xi) \leq 0] \geq 1-\alpha. \tag{5.9}$$

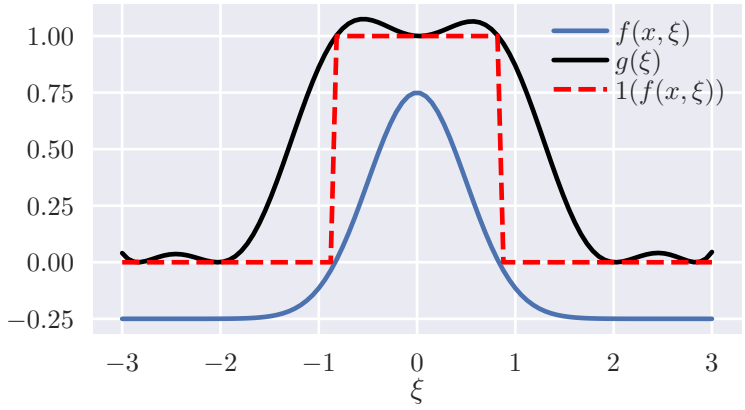


Fig. 5.2 Visualization of the RKHS-function $g(\xi)$ as a majorant of $\mathbb{1}(f(x,\xi) > 0)$ exemplary for a Gaussian constraint function f and fixed x.

While generally the VaR constraint is non-convex even for convex constraint functions $f(x,\xi)$, it is has been shown by Nemirovski and Shapiro [111], building on the idea of Rockafellar and Uryasev [133], that the tightest conservative convex approximation of VaR is given by the conditional value-at-risk (CVaR) defined as

$$\mathrm{CVaR}^{P_0}_{1-\alpha}[f(x,\xi)] = \inf_{t\in\mathbb{R}} \mathbb{E}_{P_0}[[f(x,\xi)+t]_+ - t\alpha], \tag{5.10}$$

where $[\cdot]_+ = \max[0,\cdot]$ denotes the maximum operator. Using this result we can express the tightest convex conservative approximation of our distributionally robust constraint (5.2) as

$$\sup_{P\in\mathcal{P}} \mathrm{CVaR}^{P}_{1-\alpha}[f(x,\xi)] = \sup_{P\in\mathcal{P}} \inf_{t\in\mathbb{R}} [\mathbb{E}_P[[f(x,\xi)+t]_+] - t\alpha] \leq 0 \tag{5.11}$$

The following Lemma is based on the stochastic min-max equality theorem of Shapiro and Kleywegt [145] and shows that we can exchange the supremum and infimum in (5.11).

Lemma 5.2. *Let $\Xi \subset \mathbb{R}^m$, and $f : \mathcal{X} \times \mathbb{R}^m \to \mathbb{R}$ such that $\xi \mapsto f(x,\xi)$ is bounded on Ξ, $\forall x \in \mathcal{X}$. Then,*

$$\sup_{P\in\mathcal{P}} \inf_{t\in\mathbb{R}} \mathbb{E}_P[[f(x,\xi)+t]_+] - t\alpha] = \inf_{t\in\mathbb{R}} \sup_{P\in\mathcal{P}} \mathbb{E}_P[[f(x,\xi)+t]_+] - t\alpha]. \tag{5.12}$$

Proof. The proof is identical to the one of Lemma IV.2. of Hota et al. [78] for Wasserstein ambiguity sets, using that the MMD-based uncertainty set $\mathcal{P}$ is also weakly compact [183]. □

Applying the generalized duality result of Zhu et al. [183] to the supremum in the RHS of (5.12) yields the following result which provides an expression for the feasible set of the CVaR approximation of MMD-DRCCP.

Proposition 5.3. *Let the conditions of Lemma 5.2 be fulfilled. Then the distributionally robust CVaR (DR-CVaR) constraint* $\sup_{P\in\mathcal{P}} \mathrm{CVaR}^P_{1-\alpha}[f(x,\xi)] \leq 0$ *is equivalent to* $x \in Z_{\mathrm{CVaR}}$*, where*

$$Z_{\mathrm{CVaR}} := \left\{ x \in \mathcal{X} : \begin{array}{ll} g_0 + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} g(\xi_i) + \varepsilon\|g\|_{\mathcal{H}} \leq t\alpha & \text{(5.13a)} \\ [f(x,\xi)+t]_+ \leq g_0 + g(\xi), \ \ \forall \xi \in \Xi & \text{(5.13b)} \\ g \in \mathcal{H}, \quad t \in \mathbb{R} & \text{(5.13c)} \end{array} \right\}$$

Proof. The proof follows directly from Lemma 5.2 and the generalized duality result of Zhu et al. [183] applied to the inner supremum in (5.12)

$$\begin{aligned} &\sup_{P\in\mathcal{P}} \mathrm{CVaR}^P_{1-\alpha}[f(x,\xi)] \\ &= \inf_{t\in\mathbb{R}} \sup_{P\in\mathcal{P}} \mathbb{E}_P[[f(x,\xi)+t]_+ - t\alpha] \\ &= \inf_{g_0, t\in\mathbb{R}, g\in\mathcal{H}} g_0 + \frac{1}{N}\sum_{i=1}^{N} g(\xi_i) + \varepsilon\|g\|_{\mathcal{H}} - t\alpha \\ &\quad \text{s.t. } [f(x,\xi)+t]_+ \leq g_0 + g(\xi) \ \ \forall \xi \in \Xi. \end{aligned}$$

The requirement of the infimum over (t, g_0, g) to be ≤ 0 is equivalent to requiring that there exists any $t, g_0 \in \mathbb{R}$ and $g \in \mathcal{H}$ such that $g_0 + 1/N \sum_{i=1}^{N} g(\xi_i) + \varepsilon\|g\|_{\mathcal{H}} - t\alpha \leq 0$ and thus the result follows. □

Following Zhu et al. [183], we use a constraint sampling approximation to the infinite constraint (5.13b) by replacing it with its empirical version

$$[f(x,\xi_i)+t]_+ \leq g_0 + g(\xi_i), \quad i = 1,\ldots,N. \tag{5.14}$$

However, unlike the convergence analysis for semi-infinite program discretizations such as in Royset and Pee [136], we later provide a finite-sample guarantee with a convergence rate of (at least) $\mathcal{O}(\frac{1}{\sqrt{N}})$ independent of the dimensionality of the uncertain variable ξ.

Using the robust representer theorem of Zhu et al. [183], we can express the RKHS function g in terms of finite dimensional parameters $\gamma \in \mathbb{R}^N$. Let K denote the kernel Gram matrix with $K_{i,j} = k(\xi_i, \xi_j)$, used to express $E_{\hat{P}_n}[g(\xi)]$ and $\|g\|_{\mathcal{H}}$, then we can write the sample approximation of (5.13) as

$$\hat{Z}_{\mathrm{CVaR}} := \left\{ x \in \mathcal{X} : \begin{array}{ll} g_0 + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} (K\gamma)_i + \varepsilon\sqrt{\gamma^T K \gamma} \leq t\alpha & \text{(5.15a)} \\ [f(x,\xi_i)+t]_+ \leq g_0 + (K\gamma)_i, & \text{(5.15b)} \\ i = 1,\ldots,N, & \text{(5.15c)} \\ g_0 \in \mathbb{R}, \quad \gamma \in \mathbb{R}^N, \quad t \in \mathbb{R} & \text{(5.15d)} \end{array} \right\}$$

In contrast to DRCCP using Wasserstein ambiguity sets that requires ad-hoc reformulation techniques depending on the form of f (cf. Table 5.1 and [78]), we simply use (5.15) to treat all forms of f. This generality is due to the fact that RKHS functions are universal function approximators [158, 170]. When the function f is convex in the decision variable x, the problem (5.15) is a convex kernel approximation problem and can be solved with an off-the-shelf convex optimization solver.

Remark 5.4. *Similar to DRCCP with Wasserstein ambiguity sets, within our MMD-DRCCP framework, one can also obtain so-called exact reformulations using certain kernel choices. For example, for piece-wise linear constraints supported on a closed convex cone, one can derive tractable CVaR approximations by choosing linear kernels. However, such reformulations only apply to ad-hoc f classes and restrict kernel choices to less expressive kernels, e.g., linear kernel. In those cases, MMD is not a metric since it only detects differences in the first moment. Therefore, we favor the single general approximation* (5.15) *over such exact reformulations.*

Remark 5.5. *Note that the convexity of the inner CVaR approximation is inherited from the convexity of the function f with respect to x, i.e., our reformulation technique is* convexity-preserving *but can also treat non-convex f in practice. We later demonstrate this in an optimal control problem with non-convex constraints.*

5.3 Finite Sample Guarantee for Constraint Satisfaction

Since we rely on the approximate formulation (5.15) to treat general nonlinear constraints, an important task is to quantify the approximation error. Furthermore, we are interested in finite-sample analysis instead of asymptotic consistency results such as the ones by Cherukuri and Hota [39] for Wasserstein ambiguity sets, since the former is more informative for quantifying robustness against estimation error. Our result in this section shows that, using (5.15), we can solve the MMD-DRCCP with a finite-sample guarantee for constraint satisfaction at a $\mathcal{O}(\frac{1}{\sqrt{N}})$ rate independent of dimensions. This concentration result give the foundation for the distributional robustness of (5.15).

Existing analysis for MMD ambiguity sets (see e.g., Lam and Zeng [97] for references) only concerns the exact solution to (5.2), which is unavailable in practice. To date, the only available algorithm to solve DRO problems with MMD ambiguity sets is that of Zhu et al. [183], whose statistical guarantee is not yet established. In contrast, we now show the first finite-sample guarantee of constraint satisfaction for the proposed practical algorithm. Furthermore, the finite-sample analysis in this section only assumes mild boundedness for the constraint function f, which is significantly less restrictive than the assumptions required for Wasserstein ambiguity sets.

In the proofs, we make the mild assumptions that the function $f(x,\cdot)$ and g are bounded in infinity norm, i.e., $\exists M > 0 : \forall \xi, |f(x,\xi)| \leq M/2, |g(\xi)| \leq M/2$. For conciseness, we simple write g instead of $g_0 + g$ since one can re-define a new RKHS to include the constant g_0 term.

Proposition 5.6 (Finite-sample guarantee for MMD-DRCCP constraint satisfaction)**.** *Let $(\hat{x}, \hat{g})$ be a pair of the solution to the CVaR approximation* (5.15)*. Suppose the kernel is bounded in the sense that $\exists C > 0, \sup_x |k(x,x)| \leq C$, and the radius of the MMD ambiguity set satisfies*

$\rho \geq \sqrt{\frac{C}{N}} + \sqrt{\frac{2C\log(1/\delta)}{N}}$ *(see* (5.4)*). Then,* $\hat{x}$ *is an* $M\sqrt{\frac{2\log(1/\delta)}{N}}$*-approximate feasible solution to the MMD-CVaR approximation* (5.13) *with probability at least* $1-\delta$*, i.e.,*

$$\mathrm{CVaR}^{P_0}_{1-\alpha}[f(\hat{x},\xi)] \leq M\sqrt{\frac{2\log(1/\delta)}{N}}. \tag{5.16}$$

Proof. We first expand the left-hand-side of (5.16),

$$\mathrm{CVaR}^{P_0}_{1-\alpha}[f(\hat{x},\xi)] = \inf_t \frac{1}{\alpha} E_{P_0}[f(\hat{x},\xi)+t]_+ - t. \tag{5.17}$$

Note that the expectation term in CVaR can be written as

$$E_{P_0}[(f(\hat{x},\xi)+t)_+] = E_{P_0}[\hat{g}(\xi] + E_{P_0}[(f(\hat{x},\xi)+t)_+ - \hat{g}(\xi)] \tag{5.18}$$

For the first term in (5.18), we have

$$\begin{aligned} E_{P_0}[\hat{g}(\xi)] &= E_{\hat{P}_n}[\hat{g}(\xi)] + \int \hat{g}(\xi)\,\mathrm{d}(P_0 - \hat{P}_n) && (5.19)\\ &\leq E_{\hat{P}_n}[\hat{g}(\xi)] + \mathrm{MMD}(P_0, \hat{P}_n; \mathcal{H}) \cdot \|\hat{g}\|_{\mathcal{H}} && (5.20)\\ &\leq E_{\hat{P}_n}[\hat{g}(\xi)] + \rho\,\|\hat{g}\|_{\mathcal{H}}\,, && (5.21) \end{aligned}$$

where the first inequality is simply Cauchy-Schwarz and the second inequality is due to the condition that $\rho \geq \mathrm{MMD}_{\mathcal{H}}(P_0, \hat{P}_n; \mathcal{H})$ with high probability as noted in (5.4).

For ease of notation, let

$$\mathcal{E}_n := M\sqrt{\frac{2\log(1/\delta)}{N}}.$$

For the second term in (5.18), we apply McDiarmid's inequality

$$E_{P_0}[(f(\hat{x},\xi)+t)_+ - \hat{g}(\xi)] \leq E_{\hat{P}_n}[(f(\hat{x},\xi)+t)_+ - \hat{g}(\xi)] + \mathcal{E}_n \leq 0 + \mathcal{E}_n \tag{5.22}$$

For the last inequality above, we exploited the relationship $[f(\hat{x},\xi)+t]_+ \leq \hat{g}$ that holds at the empricial sample ξ_i due to the constraints in (5.14). Plugging both inequalities back into (5.18), we obtain

$$E_{P_0}[(f(\hat{x},\xi)+t)_+] \leq E_{\hat{P}_n}[\hat{g}(\xi)] + \rho\,\|\hat{g}\|_{\mathcal{H}} + \mathcal{E}_n. \tag{5.23}$$

Combining the result above with the relationship in (5.13a) of the CVaR approximation of the MMD-DRCCP, we arrive at the proposition statement. □

Using the relationship between CVaR and chance constraints, this result directly translates to the following.

Corollary 5.7. *Under the same assumption as Proposition 5.6, with probability at least* $1-\delta$*,*

$$P\left(f(\hat{x},\xi) \leq M\sqrt{\frac{2\log(1/\delta)}{N}}\right) \geq 1-\alpha \tag{5.24}$$

Our guarantees above state that the MMD-DRCCP solution approximately satisfy the DRCCP constraint with a rate of $\mathcal{O}(\frac{1}{\sqrt{N}})$ independent of dimensions. This can further motivate a constraint back-off design that adds to the left-hand-side of (5.13a) and (5.15a) by the $M\sqrt{\frac{2\log(1/\delta)}{N}}$ term to guarantee safety. That way, the $M\sqrt{\frac{2\log(1/\delta)}{N}}$ term will no longer appear in our guarantee statements (5.16) and (5.24). However, due to the general conservatism of DR-CVaR approximations and ambiguity set sizes, we observe the current MMD-DRCCP alone is sufficient in practice.

Remark 5.8. *The convergence guarantees presented above can be further made uniform with respect to the decision variables* (x, g) *using uniform convergence results for empirical processes [166]. In addition, we leave further refinement of the convergence rate beyond* $\mathcal{O}(\frac{1}{\sqrt{N}})$ *for future work.*

5.4 Numerical Examples

In the following, we present numerical results for our MMD-DRCCP algorithms. Within a chance constrained portfolio optimization problem, we provide empirical evidence to support our theoretical finite sample guarantee (Proposition 5.6). Moreover we show that the bootstrap construction improves on the MMD-rate-based ambiguity set in terms of providing a less conservative ambiguity size.

Enabled by our theory, we simply use a Gaussian kernel $k(x,y) = \exp(-\frac{1}{2\sigma^2}\|x-y\|_2^2)$ with the bandwidth σ set via the median heuristic [57] for all experiments. We further emphasize that we do not exploit any ad-hoc transformations such as convex conjugate of certain specific f functions. We simply use our general approximation scheme (5.15) as a universal technique across all function classes.

5.4.1 Chance-Constrained Portfolio Optimization

We consider a chance-constraint portfolio optimization problem, where we want to optimally allocate resources $x \in \Delta := \{x \in \mathbb{R}^3_{0,+} : \sum_{i=1}^3 x_i \leq 1\}$ to investments with returns $c = (1, 1.5, 2)^T \in \mathbb{R}^3$ in presence of a chance constraint depending on uncertain variables $\xi \sim \mathcal{N}(0, \mathrm{diag}[0.5, 1, 1.5])$:

$$\max_{x\in\Delta} c^T x \quad \text{s.t.} \quad P[f(x,\xi) \leq 0] \geq 1-\alpha, \tag{5.25}$$

where the constraint function is given by the non-linear function $f(x,\xi) = (\xi^T x)^2 - 1$. The CVaR approximation of the constraint can be written as $\mathrm{CVaR}^{\hat{P}_n}_{1-\alpha}[f(x;\xi)] \leq 0$ and the corresponding MMD-DRCCP is then given by

$$\max_{x\in\Delta} c^T x \quad \text{s.t.} \quad \sup_{P\in\mathcal{P}} \mathrm{CVaR}^{P}_{1-\alpha}[f(x;\xi)] \leq 0. \tag{5.26}$$

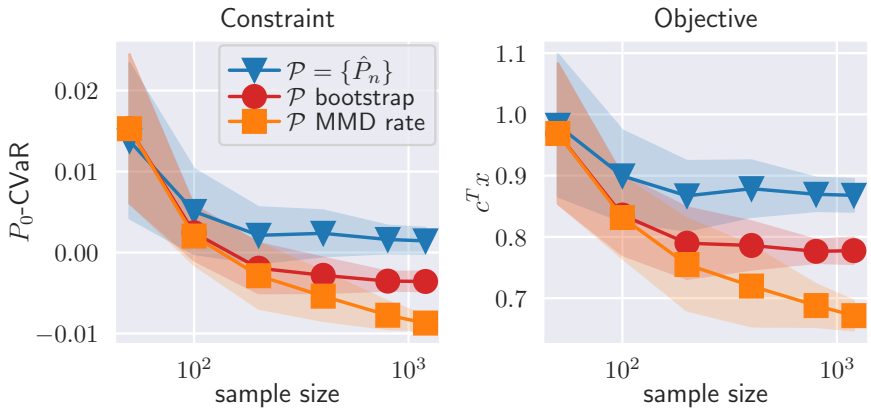


Fig. 5.3 CVaR relaxation for chance constrained portfolio optimization. Lines and shaded regions denote the mean and standard deviation over 16 runs respectively.

We construct MMD ambiguity sets using the MMD convergence rate (5.4) as well as our bootstrap construction (see Algorithm 3). We solve the problem for different sample sizes via the convex reformulation (5.15a)-(5.15d) using CVXPY [49] and compare the results to a (non-robust) CVaR approximation of (5.25) (equivalent to DR-CVaR with ambiguity set $\mathcal{P} = \{\hat{P}_n\}$). At test time we sample 10^6 data points from the true distribution in order to estimate $\mathrm{CVaR}^{P_0}_{1-\alpha}[f(\hat{x}, \xi)]$. We observe in Figure 5.3 that, while the non-robust solution fails to fulfill the CVaR constraint for the population distribution across all numbers of training samples, both MMD-DRCCP solutions fulfill the constraint for training sample sizes ≥ 100. Moreover, we observe that the bootstrap version yields a tighter estimate of the ambiguity set and thus a less conservative solution which allows for larger objective values as observed in the right panel of Figure 5.3.

5.4.2 Distributionally Robust Stochastic MPC with Nonlinear Constraints

In this example, we highlight a tube-based MPC problem with linear dynamics but nonlinear non-convex constraints. For a detailed explanation of the application of distributionally robust CC to tube-based MPC, we refer to Nemmour et al. [112]. We consider the problem of controlling a double-integrator system $f_{dyn}(x, u)$ with additive noise subject to a constraint given in the form of a non-convex SVM classifier. The optimal control problem (OCP) with horizon H and quadratic cost $J(\mathbf{x}, \mathbf{u})$ is subject to the constraints

$$x^l_{k+1} = f_{dyn}(x^l_k, u^l_k), \quad \sup_{\mathrm{MMD}(P,\hat{P}_n)\leq\varepsilon} \mathrm{CVaR}^{P}_{1-\alpha}[h(x^l_k)] \leq 0,$$

where l denotes the current iteration in the MPC-loop, $\mathbf{u}^l = [u^l_0, \ldots, u^l_{H-1}]$ the actions, $\mathbf{x}^l = [x^l_0, \ldots, x^l_{H-1}]$ the states, and $k = 0, \ldots, H-1$. We solve the OCP using MMD-DRCCP with bootstrap ambiguity sets and visualize the resulting closed-loop trajectories with high constraint satisfaction in Figure 5.4. Note that related methods for MPC with Wasserstein ambiguity sets [180] are restricted to affine constraints and thus not applicable to this problem, which highlights the greater generality of our approach.

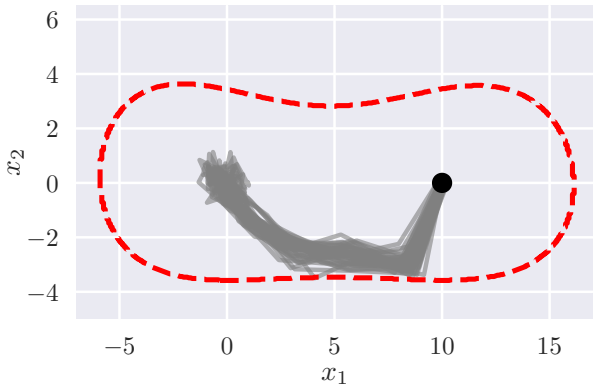


Fig. 5.4 Tube-based MPC with 40 samples of the additive disturbance $\mathcal{N}(0, 0.2)$. We add small uniform noise on the initial state $(10, 0)$ and plot the 30 different resulting trajectories in grey, visualizing the dynamics tube resulting from the MMD-DRCCP.

5.5 Further Related Work

DRCCPs have recently attracted significant attention in the stochastic programming and control community [42, 180]. A significant focus has been laid on Wasserstein ambiguity sets [171, 35, 78, 81, 76], for which the strong duality result of Gao and Kleywegt [56] plays a fundamental role. As chance constraints are generally non-convex even for convex constraints [111], a common approach to the problem relies on approximating the chance constraint via the conditional Value-at-Risk (CVaR) [133], which has been shown to provide the tightest conservative convex approximation [111]. As the dual formulation of the Wasserstein DRCCP contains a constraint involving a supremum over the uncertain variable, many works restrict their scope to constraint functions affine in the uncertain variable for which the supremum can be carried out in closed form. A notable exception is given by Hota et al. [78], which only assumes the constraints to be concave in the uncertainty and proposes solving the resulting semi-infinite program with a cutting-plane algorithm to compute an approximate solution.

Statistical guarantees for Wasserstein DRO [107] provide a basis for consistency results for DRCCP by Cherukuri and Hota [39]. They show that the DRCCP with Wasserstein ambiguity set converges to the CCP as the samples size goes to infinity. However, unlike the finite-sample analysis in the present work, those consistency results do not provide error bounds for solutions computed using finitely many samples, which is the key to certifying robustness. They also assume access to the optimal solution to the original program, which may not be available depending on the constraint function f.

In the context of CCPs, kernel methods have been used previously to estimate the unknown distribution over the uncertainty variables via kernel density estimation to obtain a deterministic problem [25, 85]. A number of works in stochastic control have considered mean embeddings of the chance constraints to obtain approximations [162, 181, 182, 61]. However, none of those works provide algorithms that can solve MMD-constrained DRCCP or provide finite-sample guarantees. To the best of our knowledge, this work is the first to solve DRCCP with MMD ambiguity sets in a principled way, thus

utilizing the unique advantage of MMD over Wasserstein-based approaches, e.g., bootstrap MMD ambiguity sets, approximating nonlinear constraints, and favorable finite-sample guarantees.

5.6 Conclusion

In this work, we presented distributionally robust chance-constrained optimization with MMD ambiguity sets. Leveraging recent results in kernel methods for robust machine learning, we provided a practical algorithm for distributionally robust Conditional Value-at-Risk constraints with finite-sample constraint satisfaction guarantees. Different from methods based on Wasserstein ambiguity sets, we give a practical bootstrap scheme that enables a priori computation of suitably sized ambiguity sets. Moreover, our method can be applied to general nonlinear constraint functions and thus parts with the strong assumptions of other recently proposed frameworks.

For future work, we plan to derive sharper finite-sample guarantees for constraint satisfaction and optimality and explore more applications to robust nonlinear control problems.

Chapter 6

Conclusion

Recap and Broader Relevance

Conditional moment restrictions provide an intuitive and versatile modelling framework that can be used to address some of the major tasks related to distributionally robust optimization and causal inference. The main goal of this thesis was to develop theoretically well-founded and practical estimators for this problem. As a learning paradigm, we built our research on the empirical likelihood framework [122, 121, 130], first by extending it from the unconditional to the conditional case using flexible machine learning models (FGEL, Chapter 2 [92]), and then from φ-divergences to more expressive distributional distance notions, namely maximum mean discrepancy (KMM, Chapter 3 [93]) and Wasserstein distances (SMM, Chapter 4). The different estimators come with certain advantages and drawbacks. While the FGEL estimator, just like the original GEL estimator, is restricted to an efficient reweighting of the training data, the Kernel Method of Moments approximates the population distribution with a class of general distributions. This comes at the price of requiring an additional sampling scheme which limits its scalability. Similarly, the Sinkhorn Method of Moments provides additional robustness by taking into account the geometry of the data manifold but requires computation of higher order derivatives.

A key difference between estimators for conditional moment restrictions and the dominating learning paradigm of empirical risk minimization is how the learning signal arises from the data and the model. While ERM derives the gradient information directly from the error of the model evaluated on the data, CMR estimators involve an additional step in which the instrument function determines how the error on the training data is translated into gradient information to improve the model. It is the flexibility of this additional step that allows CMR estimators to be used for more intricate learning tasks like robustness against covariate shifts or instrumental variable regression. From a different perspective, the instrument function allows to induce a relaxed notion of statistical independence between the moment function ψ and conditioning variable Z (e.g. the instrument). To see this, note that two zero-mean random variables ψ and Z are said to be *uncorrelated* if $E[\psi \otimes Z] = 0$, i.e., if they are statistically *linearly* independent. In contrast, the variables are said to be *independent*, i.e., $\psi \perp\!\!\!\perp Z$ if $E[f(\psi)h(Z)] = 0 \; \forall f \in \mathcal{F}$ and $\forall h \in \mathcal{H}$ where $\mathcal{F}$ and $\mathcal{H}$ are sufficiently rich

spaces of square-integrable functions.. Therefore, the concept of statistical correlation is strictly entailed by statistical independence as it corresponds to restricting the functions spaces $\mathcal{F}$ and $\mathcal{H}$ to linear functions. Conditional moment restrictions bridge these notions as $E[\psi^T h(Z)] = 0 \; \forall h \in \mathcal{H}$ corresponds to restricting only one of the function spaces to linear functions. Therefore CMR can be used to induce a notion of unrelatedness in between the concepts of uncorrelation and independence, which in practice is easier to enforce than strict independence [64]. The examples in this work demonstrated that this notion is sufficient to identify the true function in problems like instrumental variable regression. However, as CMR does not impose strict statistical independence, there might be applications where stronger notions of independence are required for which CMR might not be a suitable framework.

At the core of empirical likelihood estimation is the optimization over an infinite-dimensional space of distributions. In the classical GEL framework much of the complexity is circumvented as the use of φ-divergences effectively restricts the decision variable to finite dimensional weights attributed to the empirical data points. However, in the case of more complex distributional distance measures this is not the case anymore and one generally has to optimize over the entire infinite-dimensional space of distributions. Optimization over measure spaces is an active area of research which forms the basis of state-of-the-art deep generative modeling approaches like score-based or diffusion models [155, 156, 75] and more general energy based models [157]. These methods generate new samples from a distribution by starting with random samples from a base distribution and (approximately) following the Langevin dynamics [82]. The Langevin dynamics describe the discretized Wasserstein gradient flow of the Kullback-Leibler divergence between the base distribution and the population distribution of the data. While these so-called particle approaches, which discretize the infinite-dimensional measure into particles and optimize their positions and weights, have shown tremendous success in above mentioned applications, we empirically found them not to be effective for solving the optimization problem over the measure in the primal empirical likelihood formulation. Instead, we resorted to the duality relation between measures and functions. In this context, the choice of distributional distance metric in the primal formulation determines the function space of the dual variable. For example, the use of maximum mean discrepancy allowed us to transform the optimization over an infinite-dimensional measure space into an optimization over a reproducing kernel Hilbert space. In contrast to optimization in measure spaces, optimization in function spaces, especially RKHS, has been extensively explored and is at the heart of the success of kernel methods. In the context of the Kernel Method of Moments an important contribution of our work was the relaxation of the dual problem by means of imposing entropic regularization on the primal problem. In the derivation of the method, the duality relation allowed us to swap the optimization over a measure for an optimization over an RKHS function. However, it came at the cost of introducing a semi-infinite constraint in the dual problem, which generally requires approximations or relaxations. To address this, we proposed entropy regularization and showed that by adding a small φ-divergence between the decision variable and an arbitrary reference distribution, the semi-infinite constraint gets relaxed into a soft-constraint or log-barrier evaluated over the reference distribution. Entropy regularization has been used before in the context of computing optimal transport distances [44] and studying gradient flows in the MMD geometry [40]. However, using entropy regularization to derive relaxations of the dual problem is a rather new concept which could be of independent interest. More generally,

addressing optimization problems over measure spaces via their dual formulation as optimization over function spaces is a promising approach that might be applicable to a large variety of problems.

Limitations and Outlook

One limitation of our approach is that the majority of our estimators, and likewise related state-of-the-art estimators [15, 101, 14], result from saddle point optimization problems, where the objective is minimized with respect to the model parameters and maximized with respect to an instrument function. While there has been significant progress in this domain motivated by the development of Generative Adversarial Networks [59], mini-max optimization remains a difficult problem which usually exhibits instabilities and is generally sensitive to the choice of training hyperparameters. This is particularly problematic, as in several examples, including instrumental variable (IV) regression, it is not clear how to define a suitable validation metric to be used for hyperparameter tuning and early stopping. For example, in IV regression, a model that has small mean-squared error on a test set does not necessarily provide an accurate representation of the causal effect of the treatment on the outcome but might involve the effect of confounding variables to an arbitrary extent. In this context, we explored two kernel-based choices of causal validation metrics namely maximum moment restrictions (MMR) [179] and the Hilbert-Schmidt Independence Criterion (HSIC) [65] which exhibited vastly different behavior. While HSIC led to strongly improved results over MMR on some tasks [93] it did not yield satisfactory results on other tasks. In addition, both these metrics themselves involve hyperparmeters, i.e., the choice of kernel and e.g., the kernel bandwidth, which are usually set via common heuristics that might fail to capture the relevant relations in the data. Therefore, the development of reliable and hyperparameter-free validation metrics for instrumental variable regression and related problems remains an important topic for future research. Putting this further into perspective, concepts like IV regression are not of a purely theoretical nature, but they are encountered in practical problems on a frequent basis [4]. In order for methods to be accepted by the broader scientific community, they need to be applicable by practitioners who might not have expert knowledge about the intricacies of mini-max optimization neither have access to computer clusters to run extensive hyperparameter searches. This casts simple estimators that can be used in a plug-and-play fashion with a minimal number of hyperparameters exceedingly important. This insight motivated the development of our Kernel-SMM estimator, which apart from the choice of two internal hyperparameters does not involve any training parameters. While pushing the state-of-the-art in IV regression on artificial benchmarks with neural network based methods is definitely an interesting direction in its own right, we believe that the greater impact can be achieved by converging back to the needs of practitioners and developing estimators that can be used out of the box.

From a theoretical point of view, the asymptotic theory of the estimators developed in this work focuses on the uniquely identified case, i.e., it assumes there exists a unique parametric model $f_0 \in \mathcal{F}$ that fulfills the conditional moment restrictions. In the language of machine learning this corresponds to the assumption of a parametric model and a strictly convex loss function ℓ (as we can always use $\psi = \nabla\ell$). This is in agreement with most of the related methods [2, 15, 14, 179] and results beyond the uniquely identified case have only been developed recently [17]. While consistency in the uniquely identified setting is a necessary condition for a valid statistical estimator,

in many modern cases of interest, the model class $\mathcal{F}$ is described by highly flexible non-parametric models or over-parameterized neural networks. In this case there might be multiple functions that fulfill the population moment restrictions and thus the concept of parametric consistency becomes inapplicable. The asymptotic theory of conditional moment restriction estimators for this case and the characterization of the asymptotic convergence properties e.g., in terms of L^2 consistency largely remains an open problem and an important future direction.

Despite their great flexibility, conditional moment restrictions have been mostly used for instrumental variable regression and there is little work on their application to more general robust learning tasks. A potential reason that impeded their broader adoption is that by allowing robustifications against shifts in a *continuous* variable, they have a different application domain than many related methods that often assume access to *discrete* and in particular *unordered* group labels Z, for example gender, race or nationality. Examples of methods that fall into this category are group DRO [138], invariant risk minimization [9] and many other methods [146, 71, 1, 58, 177, 139, 104, 178]. As CMR estimators generally learn a smooth instrument function $h : \mathcal{Z} \to \mathbb{R}^m$ over the group label Z, they are able to incorporate relations between different values of Z and thus require $\mathcal{Z}$ to be equipped with some notion of distance, i.e., to be a metric space. This is in contrast to the aforementioned methods that are often applied in situations where the group labels denote abstract entities, e.g., gender, for which a notion of distance can hardly be defined. As a consequence popular benchmarks for distribution shifts exist for the discrete unordered case [89] but very little examples for shifts in a continuous variable taking values in a metric space. Developing benchmark datasets for this new class of problems and exploring the applicability of CMR estimators in this context seems a promising direction for future research.

Finally, while conditional moment restrictions bear a great potential for general robust learning tasks, the focus of this thesis has been on the methodological side, i.e., on developing and investigating new estimators for this problem. Therefore, the evaluation of our methods was restricted to relatively simple toy examples. This is mostly in line with related methods in causal machine learning [127, 128]. While in recent years, there has been significant progress in combining causal principles with more general machine learning approaches under the name of causal representation learning [143, 144], due to the difficulty of the problem, many tools are still restricted to rather low dimensional settings and have yet to catch up with the scale required by many modern applications. In contrast, state-of-the-art methods for large scale tasks like language and image generation, that drive the current trend towards general artificial intelligence, are mostly obtained by a small set of different deep learning architectures trained with plain empirical risk minimization [119, 161, 164]. Given this unprecedented performance, one might even wonder if causal learning, in the form of explicit inductive biases, is needed at all or standard empirical risk minimization, i.e., learning solely from statistical correlation and the implicitly contained causal information, is actually all one needs. Related to this question, the renowned reinforcement learning pioneer Richard Sutton hypothesized in his frequently quoted article "the bitter lesson" [159], that in the long term, significant progress in artificial intelligence is not due to developments in the methodology but rather just due to the increase of scale. While this largely seems to be in line with the current observations in deep learning research, there are examples in history where progress alternated between scale and structure. Although the first neural network architectures,

building on the perceptron of Rosenblatt [134], were already invented in the late 1950s, it took more than 50 years until the deep learning revolution took off propelled by the success of AlexNet [94] on the ImageNet challenge [48]. The retarding factor over this time was most likely the lack of training data and compute resources. While neural networks enjoyed some popularity in the second half of the 20th century, by the end of the millennium the emerging area of kernel methods [141, 140, 18] had revolutionized the field by trading in the largely unexploited scalability of neural networks for more structure which allowed better utilization of the available data and compute resources. Eventually with the growing availability of data as well as exponentially improved computer hardware, scale caught up and deep learning took over the field [100]. Nowadays, at the latest since the introduction of the highly scalable general purpose architecture of the transformer [168], scale has become the main driver of progress in the direction of artificial general intelligence in recent years [119, 161, 164]. This is supported by the fact that novel state-of-the art models are mostly still based on minor modifications of the original transformer architecture proposed in the seminal paper by Vaswani et al. [168]. Being able to incorporate the information from enormous amounts of data, current models show unprecedented performance in language and image processing tasks and empirical risk minimization is simply the most efficient way to incorporate such data. However, as current state-of-the-art models are already trained on large proportions of the entirety of available text on the internet [119, 161, 164], it seems inevitable that again scale is about to reach its limits. Once the seemingly endless reservoir of data is exhausted, scale cannot remain the main driver of progress and structure will be required for further advancements. Therefore, alternative training paradigms that allow for incorporation of additional, e.g., causal, information on top of the vast amount of statistical correlation in the data might rise to renewed importance. The field of causal machine learning is still in its early stages and there seems to be a long way to go before it becomes broadly applicable to practical large scale deep learning tasks. Nevertheless, we believe that eventually it will provide an important piece in the puzzle of artificial general intelligence and we hope that some of the ideas discussed in this thesis might contribute a tiny step towards this goal.

References

[1] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR, 2018.

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

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

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

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

[6] B. Antoine, H. Bonnal, and E. Renault. On the efficient use of the informational content of estimating equations: Implied probabilities and euclidean empirical likelihood. *Journal of Econometrics*, 138(2):461–487, 2007.

[7] M. Arbel, A. Korba, A. Salim, and A. Gretton. Maximum mean discrepancy gradient flow. *Advances in Neural Information Processing Systems*, 32, 2019.

[8] M. A. Arcones and E. Gine. On the bootstrap of u and v statistics. *The Annals of Statistics*, 20 (2):655–674, 1992.

[9] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[10] S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *The Annals of Statistics*, 47 (2):1148–1178, 04 2019.

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

[12] A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2): 341–357, 2013.

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

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

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

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

[17] A. Bennett, N. Kallus, X. Mao, W. Newey, V. Syrgkanis, and M. Uehara. Minimax instrumental variable regression and l_2 convergence guarantees without identification or closedness. In G. Neu and L. Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2291–2318. PMLR, 12–15 Jul 2023.

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

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

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

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

[22] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, Apr 2017.

[23] J. M. Borwein. On the failure of maximum entropy reconstruction for fredholm equations and other infinite systems. *Mathematical programming*, 61(1):251–261, 1993.

[24] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[25] J.-B. Caillau, M. Cerf, A. Sassi, E. Trélat, and H. Zidani. Solving chance constrained optimal control problems in aerospace via kernel density estimation. *Optimal Control Applications and Methods*, 39(5):1833–1858, Sept. 2018.

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

[27] M. Carrasco and R. Kotchoni. Regularized generalized empirical likelihood estimators. Technical report, Technical report, 2017.

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

[29] G. Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334, 1987.

[30] P. Chaussé. Generalized empirical likelihood for a continuum of moment conditions. *Technical Report*, 2012.

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

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

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

[34] Y. Chen, L. Xu, C. Gulcehre, T. L. Paine, A. Gretton, N. De Freitas, and A. Doucet. On instrumental variable regression for deep offline policy evaluation. *The Journal of Machine Learning Research*, 23(1):13635–13674, 2022.

[35] Z. Chen, D. Kuhn, and W. Wiesemann. Data-driven chance constrained programs over wasserstein balls. *Operations Research*, 2022.

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

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

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

[39] A. Cherukuri and A. R. Hota. Consistency of distributionally robust risk- and chance-constrained optimization under wasserstein ambiguity sets. *IEEE Control Systems Letters*, 5 (5):1729–1734, 2021.

[40] L. Chizat. Mean-field langevin dynamics: Exponential convergence and annealing. *arXiv preprint arXiv:2202.01009*, 2022.

[41] S. A. Corcoran. Bartlett adjustment of empirical discrepancy statistics. *Biometrika*, 85(4): 967–972, 12 1998.

[42] J. Coulson, J. Lygeros, and F. Dorfler. Distributionally robust chance constrained data-enabled predictive control. *IEEE Transactions on Automatic Control*, 2021.

[43] N. Cressie and T. R. C. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(3):440–464, 1984.

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

[45] J. M. Danskin. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4):641–664, 1966.

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

[47] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[49] S. Diamond and S. Boyd. Cvxpy: A python-embedded modeling language for convex optimization. *J. Mach. Learn. Res.*, 17(1):2909–2913, jan 2016.

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

[51] S. Donald, G. Imbens, and W. Newey. Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics*, 117:55–93, 11 2003.

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

[53] J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

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

[55] R. Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[56] R. Gao and A. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

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

[58] K. Goel, A. Gu, Y. Li, and C. Ré. Model patching: Closing the subgroup performance gap with data augmentation. *arXiv preprint arXiv:2008.06775*, 2020.

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

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

[61] B. Gopalakrishnan, A. K. Singh, K. M. Krishna, and D. Manocha. Solving Chance-Constrained Optimization Under Nonparametric Uncertainty Through Hilbert Space Embedding. *IEEE Transactions on Control Systems Technology*, pages 1–16, 2021.

[62] D. Greenfeld and U. Shalit. Robust learning with the hilbert-schmidt independence criterion. In *International Conference on Machine Learning*, pages 3759–3768. PMLR, 2020.

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

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

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

[66] Y. Gu and Y. Wang. Distributionally Robust Chance-Constrained Programmings for Non-Linear Uncertainties with Wasserstein Distance. *arXiv:2103.04790 [math]*, Nov. 2021.

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

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

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

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

[71] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

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

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

[74] C. Heinze-Deml and N. Meinshausen. Conditional variance penalties and domain shift robustness. *Machine Learning*, 110(2):303–348, 2021.

[75] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[76] N. Ho-Nguyen, F. Kılınç-Karzan, S. Küçükyavuz, and D. Lee. Distributionally Robust Chance-Constrained Programs with Right-Hand Side Uncertainty under Wasserstein Ambiguity. *arXiv:2003.12685 [math]*, Dec. 2020.

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

[78] A. R. Hota, A. Cherukuri, and J. Lygeros. Data-driven chance constrained optimization under wasserstein ambiguity sets. In *2019 American Control Conference (ACC)*, pages 1501–1506. IEEE, 2019.

[79] G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

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

[81] R. Ji and M. A. Lejeune. Data-driven distributionally robust chance-constrained optimization with wasserstein metric. *Journal of Global Optimization*, 79(4):779–811, 2021.

[82] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[83] S. A. Julious and M. A. Mullee. Confounding and simpson's paradox. *Bmj*, 309(6967): 1480–1481, 1994.

[84] N. Kallus and M. Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *J. Mach. Learn. Res.*, 21(167):1–63, 2020.

[85] R. E. Keil, A. Miller, M. Kumar, and A. V. Rao. Biased Kernel Density Estimators for Chance Constrained Optimal Control Problems. In *2020 American Control Conference (ACC)*, pages 2820–2825, July 2020.

[86] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

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

[89] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

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

[91] H. Kremer and B. Schölkopf. Geometry-aware instrumental variable regression. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 25560–25582. PMLR, 21–27 Jul 2024.

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

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

[94] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

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

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

[97] H. Lam and Y. Zeng. Complexity-free generalization via distributionally robust optimization. *arXiv e-prints*, pages arXiv–2106, 2021.

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

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

[100] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[101] G. Lewis and V. Syrgkanis. Adversarial generalized method of moments. *arXiv preprint arXiv:1803.07164*, 2018.

[102] T. Lin, C. Jin, and M. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020.

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

[104] E. Z. Liu, B. Haghgoo, A. S. Chen, A. Raghunathan, P. W. Koh, S. Sagawa, P. Liang, and C. Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021.

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

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

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

[108] J. Mooij, D. Janzing, J. Peters, and B. Schölkopf. Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the 26th annual international conference on machine learning*, pages 745–752, 2009.

[109] K. Muandet, W. Jitkrittum, and J. Kübler. Kernel conditional moment test via maximum moment restriction. In *Conference on Uncertainty in Artificial Intelligence*, pages 41–50. PMLR, 2020.

[110] K. Muandet, A. Mehrjou, S. K. Lee, and A. Raj. Dual instrumental variable regression. *Advances in Neural Information Processing Systems*, 33:2710–2721, 2020.

[111] A. Nemirovski and A. Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2007.

[112] Y. Nemmour, B. Schölkopf, and J.-J. Zhu. Approximate Distributionally Robust Nonlinear Optimization with Application to Model Predictive Control: A Functional Approach. In *Learning for Dynamics and Control*, pages 1255–1269. PMLR, May 2021.

[113] Y. Nemmour, H. Kremer, B. Schölkopf, and J.-J. Zhu. Maximum mean discrepancy distributionally robust nonlinear chance-constrained optimization with finite-sample guarantee. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 5660–5667, 2022.

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

[115] W. Newey and R. Smith. Asymptotic bias and equivalence of gmm and gel estimators. *Econometrica*, 72, 10 2002.

[116] W. K. Newey. Efficient estimation of models with conditional moment restrictions. In *Econometrics*, volume 11 of *Handbook of Statistics*, pages 419–454. Elsevier, 1993.

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

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

[119] OpenAI. GPT-4 Technical Report, 2023.

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

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

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

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

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

[125] E. Parzen. *Stochastic processes*. SIAM, 1999.

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

[127] J. Pearl, M. Glymour, and N. P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[128] J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

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

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

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

[132] H. Rahimian and S. Mehrotra. Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3:1–85, July 2022.

[133] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2(3):21–41, 2000.

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

[135] D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):215–246, 2021.

[136] J. O. Royset and E. Y. Pee. Rate of convergence analysis of discretization and smoothing algorithms for semiinfinite minimax problems. *Journal of Optimization Theory and Applications*, 155:855–882, 2012.

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

[138] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[139] S. Sagawa, A. Raghunathan, P. W. Koh, and P. Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020.

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

[141] B. Schölkopf, C. J. Burges, and A. J. Smola. *Advances in kernel methods: support vector learning*. MIT press, 1999.

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

[143] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[144] B. Schölkopf. *Causality for Machine Learning*, page 765–804. ACM, 2022.

[145] A. Shapiro and A. Kleywegt. Minimax analysis of stochastic problems. *Optimization Methods and Software*, 17(3):523–542, Jan. 2002.

[146] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

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

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

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

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

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

[152] R. J. Smith. Local GEL methods for conditional moment restrictions. Technical report, cemmap working paper, 2005.

[153] R. J. Smith. GEL criteria for moment condition models. *Econometric Theory*, 27(6):1192–1235, 2011.

[154] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

[155] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[156] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[157] Y. Song and D. P. Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

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

[159] R. Sutton. The bitter lesson. http://www.incompleteideas.net/IncIdeas/BitterLesson.html, 2019. Accessed: 2024-01-02.

[160] C. Y. Tang and C. Leng. Penalized high-dimensional empirical likelihood. *Biometrika*, 97(4): 905–920, 2010.

[161] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[162] A. Thorpe, T. Lew, M. Oishi, and M. Pavone. Data-driven chance constrained control using kernel distribution embeddings. In *Learning for Dynamics and Control Conference*, pages 790–802. PMLR, 2022.

[163] I. Tolstikhin, B. K. Sriperumbudur, and K. Muandet. Minimax Estimation of Kernel Mean Embeddings. *Journal of Machine Learning Research*, 18:1–47, 2017.

[164] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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

[166] A. W. van der Vaart and J. A. Wellner. Weak Convergence. In A. W. van der Vaart and J. A. Wellner, editors, *Weak Convergence and Empirical Processes: With Applications to Statistics*, pages 16–28. Springer, 1996.

[167] V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[168] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[169] J. Wang, R. Gao, and Y. Xie. Sinkhorn distributionally robust optimization. *arXiv preprint arXiv:2109.11926*, 2021.

[170] H. Wendland. *Scattered Data Approximation*. Cambridge University Press, 2004. ISBN 978-1-139-45665-4.

[171] W. Xie. On distributionally robust chance constrained programs with wasserstein distance. *Mathematical Programming*, 186(1-2):115–155, 2021.

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

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

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

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

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

[177] J. Zhang, A. Menon, A. Veit, S. Bhojanapalli, S. Kumar, and S. Sra. Coping with label shift via distributionally robust optimisation. *arXiv preprint arXiv:2010.12230*, 2020.

[178] M. Zhang and C. Ré. Contrastive adapters for foundation model group robustness. *Advances in Neural Information Processing Systems*, 35:21682–21697, 2022.

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

[180] Z. Zhong, E. A. del Rio-Chanona, and P. Petsagkourakis. Data-driven distributionally robust MPC using the Wasserstein metric. *arXiv:2105.08414 [cs, eess, math]*, May 2021.

[181] J.-J. Zhu, W. Jitkrittum, M. Diehl, and B. Schölkopf. Worst-case risk quantification under distributional ambiguity using kernel mean embedding in moment problem. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 3457–3463. IEEE, 2020.

[182] J.-J. Zhu, B. Schoelkopf, and M. Diehl. A Kernel Mean Embedding Approach to Reducing Conservativeness in Stochastic Programming and Control. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, pages 915–923. PMLR, 2020.

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

Appendix A

Functional Generalized Empirical Likelihood

A.1 Additional Information

A.1.1 Distributional Robustness of FGEL

It is well-known that the profile divergence is a dual formulation to the distributionally robust optimization (DRO) formulation [95, 54]. In the context of this paper, one can show that $R_{\lambda_n}(\theta) \leq \rho$ if and only if

$$\lambda_n \geq \inf_{P \in \mathcal{P}} \|E_P[\Psi(X, Z; \theta)]\|_{\mathcal{H}^*} \text{ s.t. } D_\varphi(P||\hat{P}_n) \leq \rho.$$

However, we do not simply rely on the divergence-ball centered at the empirical data distribution $\{P|\ D_\varphi(P||\hat{P}_n) \leq \rho\}$ (referred to as an ambiguity set in the DRO literature) for robustness. Since that robustness is often used to account for the statistical error due to finite samples. Instead, we are concerned with a second and stronger layer of robustness.

First, note that the quantity $E_P[\Psi(X, Z; \theta)]$ is used to approximate the conditional moment constraint in our original formulation (2.8). Since the goal of FGEL is to satisfy the the conditional moment restrictions $E_P[\psi(X; \theta) \mid Z] = 0$ almost everywhere in the domain, in robust optimization terms, we are robustifying against the instrument Z. The instrument Z can create much stronger distribution shifts in the data-generating process than the mere statistical fluctuation described by divergence-ball-based DRO works following Ben-Tal et al. [12] and Duchi et al. [54]. We leave an alternative DRO algorithm against such strong distribution shifts for future work.

From another perspective, our method can also be seen as enforcing independence between Z and the moment restriction, e.g., for IV regression the residual $\psi(X; \theta) = Y - f_\theta(X)$. Intuitively, we want the residual $Y - f_\theta(X)$ to be small and invariant to transformations of the Z variable (marginal shift). This kind of robust learning strategy has also been studied in works by Greenfeld and Shalit [62], Rothenhäusler et al. [135], Heinze-Deml and Meinshausen [74].

Algorithm 4 Kernel-FGEL

Input: data (x_i, y_i, z_i), hyperparameter λ
while not converged **do**
  while not converged **do**
    $\alpha \leftarrow \mathrm{LBFGS}(G_\lambda(\theta, h_\alpha))$
  end while
  $\theta \leftarrow \mathrm{LBFGS}(G_\lambda(\theta, h_\alpha))$
end while
Output: Parameter estimate θ

Algorithm 5 Neural-FGEL

Input: data (x_i, y_i, z_i), hyperparameter λ
while not converged **do**
  $\alpha \leftarrow \mathrm{OAdam}(G_\lambda(\theta, h_\alpha))$
  $\theta \leftarrow \mathrm{OAdam}(G_\lambda(\theta, h_\alpha))$
end while
Output: Parameter estimate θ

A.1.2 Computing the FGEL Estimator

Problem (2.13) is generally a non-convex-convex min-max problem in the parameter θ and function h. Let $h = h_\alpha$ be described by a finite dimensional set of parameters $\alpha \in A$, which is the case, e.g., for neural network function classes or RKHS after using a representer theorem. Furthermore let the set of parameters A be compact. If additionally the parameterization leaves the convexity of the inner problem intact (e.g., in the case of kernel-FGEL) we can use a simplified version of Danskin's theorem [45] to compute gradients of $R_{\lambda_n}(\theta) := \sup_{h \in \widehat{\mathcal{H}}_\theta} G_{\lambda_n}(\theta, h)$ in a principled way.

Lemma A.1 (Danskin)**.** *Let $\hat{h}(\theta)$ denote the solution of the inner convex optimization over $h \in \mathcal{H}_\theta$ such that $\hat{h}(\theta) = \arg\max_{h \in \widehat{\mathcal{H}}_\theta} G_{\lambda_n}(\theta, h)$. Then the gradient of the profile divergence $R_{\lambda_n}(\theta)$ with respect to the parameters $\theta \in \Theta$ is given by*

$$\nabla R_{\lambda_n}(\theta) = \nabla G_{\lambda_n}(\theta, \hat{h}(\theta)).$$

Therefore, we can adopt a gradient-based strategy for the outer optimization problem over θ using in each step the gradient estimate obtained from the solution of the inner maximization over h. Depending on the GEL function ϕ the optimization of both, the outer and inner problem can then be solved efficiently with an off-the-shelf solver e.g. using LBFGS (cf. Algorithm 4).

For the case of a neural network instrument function classes, we build on the recent progress in mini-max optimization and employ the optimistic Adam optimizer [46] which has been developed to solve similar saddle point problems for training generative adversarial networks [59] (cf. Algorithm 5). Implementations of both approaches are available under https://github.com/HeinerKremer/Functional-GEL.

A.1.3 Hyperparameter selection

Tuning the hyperparameter of our method, i.e., the regularization parameter λ_n (and, e.g., learning rates) requires a data-driven performance measure of the obtained model parameters. We know that for the true distribution P_0 and true parameter θ_0 we obtain $\|E_{P_0}[\Psi(X, Z; \theta_0)]\|^2_{\mathcal{H}^*} = 0$. Let β denote the set of hyperparameters and $\hat{\theta}(\beta)$ the corresponding solution to (2.13). Then we can define a performance measure of the solution candidate $\hat{\theta}(\beta)$ as $\ell(\beta) = \|E_{P_0}[\Psi(X, Z; \hat{\theta}(\beta))]\|^2_{\mathcal{H}^*}$. As we do not have access to the true distribution P_0 we can define a natural surrogate loss $\hat{\ell}$ using a validation set with empirical distribution $\hat{P}_{\text{val}}$ as

$$\hat{\ell}(\beta) = \|E_{\hat{P}_{\text{val}}}[\Psi(X, Z; \hat{\theta}(\beta))]\|^2_{\mathcal{H}^*} \tag{A.1}$$

Choosing $\mathcal{H}$ as an RKHS, this can be expressed as the kernel maximum of moment restriction objective of Muandet et al. [109] and [179] evaluated on the validation data as shown by the following lemma.

Lemma A.2. *Let $\{x_i, z_i\}_{i=1}^n$ denote the validation data and define $\boldsymbol{\psi}_j(\boldsymbol{x}; \theta) = \text{vec}(\{\psi_j(x_i; \theta)\}_{i=1}^n)$. Let K_j denote the kernel Gram matrix with entries $(K_j)_{pq} = k_j(z_p, z_q)$, $p, q = 1, \ldots, n$, $j = 1, \ldots, m$. Then we can express* (A.1) *as*

$$\hat{\ell}(\beta) = \frac{1}{n^2} \sum_{j=1}^{m} \boldsymbol{\psi}_j(\boldsymbol{x}; \theta)^T K_j \boldsymbol{\psi}_j(\boldsymbol{x}; \theta).$$

Here we assume that possible hyperparameters of the kernel are already set via commonly employed heuristics like the median heuristic [140, 57] for the kernel bandwidth and only tune the remaining parameters of our method.

A.2 Proofs

A.2.1 Preliminaries

For ease of notation we define some expressions first. Define $\Psi_i(\theta) := \Psi(x_i, z_i; \theta)$ and denote $\phi_i(v) = \frac{d^i}{(dv)^i}\phi(v)$ and $\phi_i = \phi_i(0)$. Without loss of generality we assume that $\phi_1(0) = \phi_2(0) = -1$, as any ϕ with $\phi_1 \neq 0$ and $\phi_2 < 0$ can be rescaled to achieve this (see Newey and Smith [118]). Define the empirical objective as $\widehat{G}_{\lambda_n}(\theta, h) = \sum_{i=1}^n \phi(\Psi(x_i, z_i; \theta)(h)) - \lambda\|h\|^2_{\mathcal{H}}$ and the empirical constraint set as $\widehat{\mathcal{H}}_n(\theta) = \{h \in \mathcal{H} : \Psi(x_i, z_i; \theta)(h) \in \text{dom}(\phi)\ \forall (x_i, z_i),\ i = 1, \ldots, n\}$. Throughout the proofs we will make use of functional derivatives and a functional version of Taylor's theorem with Lagrange remainder, which we define and state next, respectively.

Definition A.3 (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 $h, \tilde{h} \in \mathcal{H}$, we define the derivative operator $D_h G(h)[\tilde{h}] =$*

$\frac{d}{dt}G(h+t\tilde{h})\Big|_{t=0}$. *Likewise, we define*

$$D_h^k G(h)\,[h_1,\ldots,h_k] = \frac{\partial^k}{\partial t_1 \ldots \partial t_k} G\left(h + t_1 h_1 + \ldots + t_k h_k\right)\Big|_{t_1=\cdots=t_k=0}.$$

Similarly, when considering a function of a vector-valued parameter, $G : \Theta \subseteq \mathbb{R}^p \to \mathbb{R}$, we denote the k-th standard directional derivative at $\theta \in \Theta$ as $D_\theta^k G(\theta)(\theta_1,\ldots,\theta_k)$. Alternatively we use $\nabla_\theta G(\theta)$ to denote a row vector in $\mathbb{R}^p$ (the gradient) and $\nabla_{\theta^T} G(\theta)$ to denote the corresponding column vector in the dual space.

Proposition A.4 (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*

$$G\left(h'\right) = G(h) + \sum_{i=1}^{k} \frac{1}{i!} D_h^i G(h)[\underbrace{h'-h,\ldots,h'-h}_{i \text{ times}}] + \frac{1}{(k+1)!} D_h^{k+1} G(\bar{h})[\underbrace{h'-h,\ldots,h'-h}_{k+1 \text{ times}}].$$

Equally, using the notation of Definition A.3 the same result holds for functions of vector-valued parameters $G : \Theta \subseteq \mathbb{R}^p \to \mathbb{R}$.

Our duality result builds on Theorem 3.1 of Borwein [23]. For completeness, we will state it here adapted to our notation. Note that while the theorem is already closely related to our result, a direct application of the theorem to our case is impeded as we additionally need to take into account the normalization constraint for p, i.e., $\sum_{i=1}^n p_i = 1$.

Proposition A.5 (Borwein's theorem)**.** *For the problem*

$$P = \inf_{p \in \mathbb{R}^n} \sum_{i=1}^n \frac{1}{n} f(np_i) \quad \text{s.t.} \quad \|\frac{1}{n}\sum_{i=1}^n p_i \Psi(x_i, z_i; \theta)\|_{\mathcal{H}^*} = 0 \tag{A.2}$$

where we assume the infimum is attained (when finite), consider for $\lambda > 0$ the relaxed problem

$$P_\lambda = \min_{p \in \mathbb{R}^n} \sum_{i=1}^n \frac{1}{n} f(np_i) \quad \text{s.t.} \quad \|\frac{1}{n}\sum_{i=1}^n p_i \Psi(x_i, z_i; \theta)\|_{\mathcal{H}^*} \leq \lambda. \tag{A.3}$$

Then the value P_λ equals the value of the dual program

$$D_\lambda = \max_{h \in \mathcal{H}} -\frac{1}{n}\sum_{i=1}^n f^*(\Psi(x_i, z_i; \theta)(h)) - \lambda\|h\|_{\mathcal{H}}, \tag{A.4}$$

and the unique optimal solution of (A.3) *is given by*

$$(p_\lambda)_i = \left(\frac{d}{dv} f^*\right)\left(\Psi(x_i, z_i; \theta)(\hat{h})\right), \quad i = 1, \ldots, n,$$

where $\hat{h}$ *is any solution of* (A.4)*. Moreover, as* $\lambda \to 0$*,* p_λ *converges in mean to the unique solution of* (A.2) *and* $P_\lambda \to P$*.*

For the proofs of the asymptotic properties of our estimator, we need the following results.

Lemma A.6 (Corollary 9.31, Kosorok [90])**.** *Let* $\mathcal{F}$ *and* $\mathcal{G}$ *be Donsker classes of functions. Then* $\mathcal{F} + \mathcal{G}$ *is Donsker. Further if additionally* $\mathcal{F}$ *and* $\mathcal{G}$ *are uniformly bounded, then* $\mathcal{F} \cdot \mathcal{G}$ *is Donsker.*

Lemma A.7 (Lemma 18, Bennett and Kallus [14])**.** *Suppose that* $\mathcal{G}$ *is a class of functions of the form* $g : \Xi \to \mathbb{R}$*, and that* $\mathcal{G}$ *is* P*-Donsker in the sense of Kosorok [90]. Then we have*

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

Proof of Theorem 2.1

Proof. The proof follows almost directly from application of Proposition A.5 by taking into account the additional constraint $\sum_{i=1}^n p_i = 1$.

The dual problem can be derived by introducing Lagrange parameters $\nu > 0$ and $\mu \in \mathbb{R}$ and defining the Lagrangian

$$L(\theta, p, \mu, \nu) = \sum_{i=1}^n \frac{1}{n} f(np_i) - \mu\left(\sum_{i=1}^n p_i - 1\right) + \nu\left(\|\sum_{i=1}^n p_i \Psi(x_i, z_i; \theta)\|_{\mathcal{H}^*} - \lambda\right).$$

Using the definition of the dual norm and the fact that trivially $\lambda = \max_{\|h\|=1} \|h\|\lambda$, we have

$$L = \sum_{i=1}^n \frac{1}{n} f(np_i) - \mu\left(\sum_{i=1}^n p_i - 1\right) + \sup_{\|\tilde{h}\|_{\mathcal{H}}=1}\left(\sum_{i=1}^n \langle \nu\tilde{h}, p_i \Psi(x_i, z_i; \theta)\rangle - \|\nu\tilde{h}\|_{\mathcal{H}}\lambda\right).$$

By defining new dual Lagrange parameters $h = \nu\tilde{h} \in \mathcal{H}$, we thus obtain

$$L(\theta, p, \mu, h) = \sum_{i=1}^n \frac{1}{n} f(np_i) - \mu\left(\sum_{i=1}^n p_i - 1\right) + \sum_{i=1}^n \langle h, p_i \Psi(x_i, z_i; \theta)\rangle - \|h\|_{\mathcal{H}}\lambda.$$

Now, redefining $p_i \to np_i$ and optimizing the Lagrangian with respect to p we get

$$\begin{aligned}
&\min_p \left\{ \mu - \frac{1}{n} \sum_{i=1}^{n} \left[(\mu - \Psi(x_i, z_i; \theta)(h)) p_i - f(p_i) \right] - \lambda \|h\|_{\mathcal{H}} \right\} \\
&= \mu - \frac{1}{n} \sum_{i=1}^{n} \max_{p_i} \left\{ (\mu - \Psi(x_i, z_i; \theta)(h)) p_i - f(p_i) \right\} - \lambda \|h\|_{\mathcal{H}} \\
&= \mu - \frac{1}{n} \sum_{i=1}^{n} f^*(\mu - \Psi(x_i, z_i; \theta)(h)) - \lambda \|h\|_{\mathcal{H}},
\end{aligned}$$

where we used the definition of the Legrendre-Fenchel (convex) conjugate function $f^*(v) = \sup_x \langle v, x \rangle - f(x)$. As for any $h \in \mathcal{H}$, $-h \in \mathcal{H}$, we can redefine $h \to -h$ and finally obtain the result. Finally from Proposition A.5 it follows that strong duality holds and the unique minimizer of the primal problem is given by

$$p_i = \left(\frac{d}{dv} f^* \right) \left(\Psi(x_i, z_i; \theta)(\hat{h}) + \hat{\mu} \right), \quad i = 1, \ldots, n,$$

where $\hat{h}$, $\hat{\mu}$ are any solutions of the dual problem. □

A.2.2 Asymptotic Properties of FGEL

Proof of Theorem 2.6 (Consistency)

For the proof of Lemma A.9 we will need the following result whose proof closely follows a similar result for vector-valued moment restrictions of Owen [121] and Kitamura et al. [88] (Lemma D.2):

Lemma A.8. *Let X be a RV taking values in $\mathcal{X}$, for a bounded functional $\Psi : \mathcal{X} \times \Theta \times \mathcal{H} \to \mathbb{R}$ with $E\left[(\sup_{\theta \in \Theta} \|\Psi(X; \theta)\|_{\mathcal{H}^*})^m \right] < \infty$, it follows that $\max_{1 \le j \le n} \sup_{\theta \in \Theta} \|\Psi(x_j; \theta)\|_{\mathcal{H}^*} = o\left(n^{1/m}\right)$ with probability 1.*

Proof. For ease of notation define the random variable $Y := \sup_{\theta \in \Theta} \|\Psi(X; \theta)\|_{\mathcal{H}^*}$ and let for $i \in \mathbb{N}$, Y_i denote independent copies of Y. Then as $E[Y^m] < \infty$, we must have that $\sum_{i=1}^{\infty} P(Y_i^m > n) < \infty$ or equivalently $\sum_{i=1}^{\infty} P(Y_i > n^{1/m}) < \infty$. Hence by the Borel-Cantelli Lemma the event $Y_i > n^{1/m}$ happens only finitely often with probability 1 which likewise implies $Z_n := \max_{1 \le i \le n} Y_i > n^{1/m}$ happens only finitely often with probability 1. By the same argument the event $Z_n > \epsilon n^{1/m}$ happens only finitely often for any $\epsilon > 0$ and thus

$$\limsup Z_n / n^{1/m} \le \epsilon$$

with probability 1 and thus $Z_n = o(n^{1/m})$ with probability 1. □

The following Lemma shows that if we constrain the space of the dual parameter to a ball of radius ζ with $1/\nu < \zeta < 1/2 - \xi$, the largest value the empirical moment functional evaluated on the dual

parameter can take converges to zero in probability. Furthermore any such ball is contained in the empirical constraint set $\widehat{\mathcal{H}}(\theta) = \{h \in \mathcal{H} : \ \psi(x_i;\theta)^T h(z_i) \in \mathrm{dom}(\phi),\ 1 \leq i \leq n\}$.

Lemma A.9. *Let the assumptions of Theorem 2.6 be satisfied, then for any ζ with $1/\nu < \zeta < 1/2$ define $\mathcal{H}_n = \{h \in \mathcal{H} : \|h\|_{\mathcal{H}} \leq n^{-\zeta}\}$. Then $\sup_{\theta\in\Theta, h\in\mathcal{H}_n, 1\leq i\leq n} |\Psi(x_i, z_i;\theta)(h)| \xrightarrow{p} 0$ and w.p.a.1, $\mathcal{H}_n \subseteq \widehat{\mathcal{H}}(\theta)$ for all $\theta \in \Theta$.*

Proof. Using the Cauchy-Schwarz inequality together with Lemma A.8 we have

$$\begin{aligned}
&\sup_{\theta\in\Theta, h\in\mathcal{H}_n, 1\leq i\leq n} |\Psi(x_i, z_i;\theta)(h)| \\
&\leq \sup_{\theta\in\Theta, h\in\mathcal{H}_n, 1\leq i\leq n} (\|h\|_{\mathcal{H}} \cdot \|\Psi(x_i, z_i;\theta)\|_{\mathcal{H}^*}) \\
&\leq n^{-\zeta} \sup_{\theta\in\Theta, 1\leq i\leq n} \|\Psi(x_i, z_i;\theta)\|_{\mathcal{H}^*} \\
&= O_p(n^{-\zeta+1/\nu}) \xrightarrow{p} 0.
\end{aligned}$$

As $V = \mathrm{dom}(\phi)$ is an open interval containing zero it follows that $\Psi(x_i, z_i;\theta)(h) \in \mathrm{dom}(\phi)$ w.p.a.1 for all $\theta \in \Theta$ and $h \in \mathcal{H}_n$. □

Lemma A.10. *Let the assumptions of Theorem 2.6 be satisfied and assume $\bar{\theta} \xrightarrow{p} \theta_0$ with $E[\|\psi(X;\bar{\theta}) - \psi(X;\theta_0)\|_\infty] = O_p(n^{-\rho})$ with $0 < \rho < 1/2$. Define the operators*

$$\begin{aligned}
\Omega(\theta) &= E[\Psi(X, Z;\theta) \otimes \Psi(X, Z;\theta)] \\
\widehat{\Omega}(\theta) &= E_{\hat{P}_n}[\Psi(X, Z;\theta) \otimes \Psi(X, Z;\theta)].
\end{aligned}$$

Then we have $\|\widehat{\Omega}(\bar{\theta}) - \Omega(\theta_0)\| = O_p(n^{-\rho})$.

Proof. The proof follows the proof of Lemma 20 of Bennett and Kallus [14]. Using the triangle inequality we have,

$$\|\widehat{\Omega}(\bar{\theta}) - \Omega_0\| \leq \|\widehat{\Omega}(\bar{\theta}) - \Omega(\bar{\theta})\| + \|\Omega(\bar{\theta}) - \Omega(\theta_0)\|.$$

For the first term we have

$$\begin{aligned}
\|\widehat{\Omega}(\bar{\theta}) - \Omega(\bar{\theta})\| &= \sup_{h,h'\in\mathcal{H}_1} E_{\hat{P}_n}[h(Z)^T\psi(X;\bar{\theta})\psi(X;\bar{\theta})^T h'(Z)] - E[h(Z)^T\psi(X;\bar{\theta})\psi(X;\bar{\theta})^T h'(Z)] \\
&= \sup_{g\in\mathcal{G}^2} E_{\hat{P}_n}[g(X, Z)] - E[g(X, Z)],
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{G} &= \{g : g(x, z) = h(z)^T\psi(x;\bar{\theta}),\ h \in \mathcal{H}_1\}, \\
\mathcal{G}^2 &= \{g : g(x, z) = g_1(x, z)g_2(x, z),\ g_1, g_2 \in \mathcal{G}\}.
\end{aligned}$$

Now by Assumption h), $\mathcal{G}$ is P_0-Donsker and uniformly bounded by continuity of Ψ and compactness of its domain. Therefore $\mathcal{G}^2$ is P_0-Donsker by Lemma A.6 which lets us employ Lemma A.7 to conclude that $\|\widehat{\Omega}(\bar{\theta}) - \Omega(\bar{\theta})\| = O_p(n^{-1/2})$.

For the second term we have

$$\begin{aligned}
\|\Omega(\bar{\theta}) - \Omega(\theta_0)\| = &\sup_{h,h' \in \mathcal{H}} E[h(Z)^T \left(\psi(X;\bar{\theta})\psi(X;\bar{\theta})^T - \psi(X;\theta_0)\psi(X;\theta_0)^T\right) h'(Z)] \\
\leq &\sum_{i,j=1}^{m} \sup_{h,h' \in \mathcal{H}} E[h_i(Z)h'_j(Z)\psi_i(X;\bar{\theta}) \left(\psi_j(X;\bar{\theta}) - \psi_j(X;\theta_0)\right)] \\
&+ \sum_{i,j=1}^{m} \sup_{h,h' \in \mathcal{H}} E[h_i(Z)h'_j(Z)\psi_j(X;\theta_0) \left(\psi_i(X;\bar{\theta}) - \psi_i(X;\theta_0)\right)] \\
\leq &2m^2 C_h^2 C_\psi E[\|\psi(X;\bar{\theta}) - \psi(X;\theta_0)\|_\infty] \\
\leq &O_p(n^{-\rho}).
\end{aligned}$$

Putting things together we get $\|\widehat{\Omega}(\bar{\theta}) - \Omega_0\| = O_p(n^{-\rho})$. □

Lemma A.11. *Let the assumptions of Theorem 2.6 be satisfied and assume $\bar{\theta} \xrightarrow{p} \theta_0$ with $E[\|\psi(X;\bar{\theta}) - \psi(X;\theta_0)\|_\infty] = O_p(n^{-\rho})$. Then for $\lambda_n = O_p(n^{-\xi})$ with $0 < \xi < \rho$ and any $\dot{h} \in \operatorname{conv}(\{0, \bar{h}\})$ with $\bar{h} \in \mathcal{H}_n$, as defined in Lemma A.9, the operator*

$$\widehat{\Omega}_{\lambda_n}(\dot{h}, \bar{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} \phi_2(\bar{\Psi}_i(\dot{h}))(\bar{\Psi}_i \otimes \bar{\Psi}_i) + \lambda_n I \otimes I$$

is non-singular w.p.a.1 with largest eigenvalue $C < \infty$.

Proof. Define $\widehat{\Omega}(\bar{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} \phi_2(\Psi(x_i, z_i; \bar{\theta})(\dot{h})) \left(\Psi(x_i, z_i; \bar{\theta}) \otimes \Psi(x_i, z_i; \bar{\theta})\right)$. As $\dot{h}$ lies in between 0 and $\bar{h}$ and $\bar{h} \in \mathcal{H}_n$, we have $\dot{h} \in \mathcal{H}_n$ and thus by Lemma A.9 we have $\sup_{\theta \in \Theta} \max_{1 \leq i \leq n} |\phi_2(\Psi_i(\theta)(\dot{h})) + 1| \xrightarrow{p} 0$ and we have $\widehat{\Omega}(\bar{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \Psi(x_i, z_i; \bar{\theta}) \otimes \Psi(x_i, z_i; \bar{\theta})$ w.p.a.1. By Lemma A.10, $\widehat{\Omega}(\bar{\theta})$ converges to the non-singular operator Ω_0 at rate $O_p(n^{-\rho})$ and as $\lambda_n = O_p(n^{-\xi})$ with $0 < \xi < \rho$, we have that $\widehat{\Omega}_{\lambda_n}(\dot{h}, \bar{\theta})$ is non-singular w.p.a.1.

To bound the largest eigenvalue of $\widehat{\Omega}_{\lambda_n} := \widehat{\Omega}_{\lambda_n}(\dot{h}, \bar{\theta})$ consider any $h \in \mathcal{H}$ and

$$\langle h, \widehat{\Omega}_{\lambda_n} h \rangle_{\mathcal{H}} = \frac{1}{n}\sum_{i=1}^{n} \|\Psi(x_i, z_i; \bar{\theta})(h)\|^2 \leq \frac{1}{n}\sum_{i=1}^{n} \|\Psi(x_i, z_i; \bar{\theta})\|_{\mathcal{H}^*}^2 \|h\|_{\mathcal{H}}^2 = C_\psi^2 \|h\|_{\mathcal{H}}^2,$$

where we used that by assumption for any $\theta \in \Theta$ and $h \in \mathcal{H}_1$, $(x, z) \mapsto \psi(x;\theta)^T h(z)$ is a continuous function on a compact domain and therefore bounded. Thus the largest eigenvalue of $\widehat{\Omega}_{\lambda_n}$ must be bounded by some constant $C < \infty$. □

The following lemma generalizes Lemma A2 of Newey and Smith [118] to our regularized continuum formulation. It is different from a similar proof of Chaussé [30] (Lemma 2), as our regularization procedure differs from theirs.

Lemma A.12. *Let the assumptions of Theorem 2.6 be satisfied. Additionally let $\bar{\theta} \in \Theta$, $\bar{\theta} \xrightarrow{p} \theta_0$, with $\|E_{\hat{P}_n}[\Psi(X, Z; \bar{\theta})]\|_{\mathcal{H}^*} = O_p\left(n^{-1/2}\right)$ as well as $E[\|\psi(X; \bar{\theta}) - \psi(X; \theta_0)\|_\infty] = O_p(n^{-1/2})$. Further let $\lambda_n = O_p(n^{-\xi})$ where $0 < \xi < 1/2 - 1/\nu$. Then $\bar{h} = \arg\max_{h \in \widehat{\mathcal{H}}(\bar{\theta})} \widehat{G}_{\lambda_n}(\bar{\theta}, h)$ exists w.p.a.1, $\|\bar{h}\|_{\mathcal{H}} = O_p\left(n^{-1/2}\right)$, and $\widehat{G}_{\lambda_n}(\bar{\theta}, \bar{h}) \leq \phi(0) + O_p\left(n^{-1}\right)$.*

Proof. Let $\bar{\Psi}_i := \Psi_i(\bar{\theta}) := \Psi(x_i, z_i; \bar{\theta})$ and $\bar{\Psi} = \frac{1}{n}\sum_{i=1}^n \bar{\Psi}_i$. By Lemma A.9 and twice continuous differentiability of $\phi(v)$ in a neighborhood of zero, $\widehat{G}_{\lambda_n}(\bar{\theta}, h)$ is twice continuously differentiable on $\mathcal{H}_n = \{h : \|h\|_{\mathcal{H}} \leq n^{-\zeta}\}$ w.p.a.1. Then $\tilde{h} = \arg\max_{h \in \mathcal{H}_n} \widehat{G}_{\lambda_n}(\bar{\theta}, h)$ exists w.p.a.1. Using Taylor's theorem (Proposition A.4) we can expand the regularized GEL objective about $h = 0$ and obtain

$$\begin{aligned}
\phi_0 &= \widehat{G}_{\lambda_n}(\bar{\theta}, 0) \\
&\leq \widehat{G}_{\lambda_n}(\bar{\theta}, \tilde{h}) \\
&= \phi_0 - \bar{\Psi}(\tilde{h}) - \frac{1}{2}\underbrace{\left[-\frac{1}{n}\sum_{i=1}^n \phi_2(\bar{\Psi}_i(\dot{h}))(\bar{\Psi}_i \otimes \bar{\Psi}_i) + \lambda_n I \otimes I\right]}_{=:\widehat{\Omega}_{\lambda_n}(\dot{h}, \bar{\theta})}(\tilde{h}, \tilde{h})
\end{aligned}$$

for some $\dot{h}$ on the line between 0 and $\tilde{h}$. Now, by Lemma A.11 the regularized covariance operator $\widehat{\Omega}_{\lambda_n}(\dot{h}, \bar{\theta})$ is positive definite with smallest eigenvalue $C > 0$ bounded away from zero w.p.a.1. Using this and subtracting ϕ_0 on both sides yields

$$0 = \bar{\Psi}(\tilde{h}) - \frac{1}{2}\langle \tilde{h}, \widehat{\Omega}_{\lambda_n}(\bar{\theta})\tilde{h}\rangle_{\mathcal{H}} \leq \|\bar{\Psi}\|_{\mathcal{H}^*}\|\tilde{h}\|_{\mathcal{H}} - \frac{1}{2}C\|\tilde{h}\|_{\mathcal{H}}^2,$$

where in the second line we used the Cauchy-Schwarz inequality for the first term. This means we have $\frac{1}{2}C\|\tilde{h}\|_{\mathcal{H}} \leq \|\bar{\Psi}\|_{\mathcal{H}^*}$ w.p.a.1. As by assumption $\|\bar{\Psi}\| = O_p(n^{-1/2})$ it follows that $\|\tilde{h}\| = O_p(n^{-1/2})$. Now, as $n^{-1/2} \leq n^{-\zeta}$ we have $\tilde{h} \in \operatorname{int}(\mathcal{H}_n)$ w.p.a.1. Then, as $\tilde{h}$ is a maximizer contained in the interior of the domain $\mathcal{H}_n$, it must correspond to a stationary point of $\widehat{G}_{\lambda_n}$, i.e., $\left(\partial \widehat{G}_{\lambda_n}/\partial h\right)(\bar{\theta}, \tilde{h}) = 0$. However, from Lemma A.9 it follows that w.p.a.1 $\tilde{h} \in \widehat{\mathcal{H}}(\bar{\theta})$ and as $\widehat{G}_{\lambda_n}(\bar{\theta}, h)$ is concave and $\widehat{\mathcal{H}}(\bar{\theta})$ is convex we must have $\widehat{G}_{\lambda_n}(\bar{\theta}, \tilde{h}) = \sup_{h \in \widehat{\mathcal{H}}(\bar{\theta})} \widehat{G}_{\lambda_n}(\bar{\theta}, h)$, which directly implies $\bar{h} = \tilde{h}$ and proves the first conclusion. The second conclusion follows directly as $\bar{h} \in \operatorname{int}(\mathcal{H}_n)$ and thus $\bar{h} = O_p(n^{-\zeta})$. Finally as $\|\bar{\Psi}\|_{\mathcal{H}^*} = O_p(n^{-1/2})$ by assumption, we have $\widehat{G}_{\lambda_n}(\bar{\theta}, \bar{h}) \leq \phi_0 + \|\bar{\Psi}\|_{\mathcal{H}^*}\|\bar{h}\|_{\mathcal{H}} - \frac{1}{2}C\|\bar{h}\|_{\mathcal{H}}^2 = \phi_0 + O_p\left(n^{-1}\right)$, which completes the proof. □

The following lemma uses Lemmas A.9 and A.12 to show that the empirical moment functional $E_{\hat{P}_n}[\Psi(X, Z; \hat{\theta})]$ evaluated at the FGEL estimator $\hat{\theta}$ converges to zero in the dual norm. The proof closely follows the proof of Lemma A3 of Newey and Smith [118].

Lemma A.13. *Let the assumptions of Theorem 2.6 be satisfied and denote $\hat{\theta}$ the corresponding FGEL estimator $\hat{\theta} = \arg\min_{\theta \in \Theta} \sup_{h \in \widehat{\mathcal{H}}(\theta)} \widehat{G}_{\lambda_n}(\theta, h)$. Then $\|E_{\hat{P}_n}[\Psi(X, Z; \hat{\theta})]\|_{\mathcal{H}^*} = O_p\left(n^{-1/2}\right)$.*

Proof. Define $\hat{\Psi}_i := \Psi_i(\hat{\theta}) := \Psi(x_i, z_i; \hat{\theta})$ and $\hat{\Psi} = \frac{1}{n}\sum_{i=1}^n \hat{\Psi}_i$. Let $\mu(\hat{\Psi})$ be the Riesz representer of $\hat{\Psi} \in \mathcal{H}^*$ in $\mathcal{H}$. Further, let $1/\nu < \zeta < 1/2$ be defined as in Lemma A.9 and consider $\tilde{h} = -n^{-\zeta}\mu(\hat{\Psi})/\|\mu(\hat{\Psi})\|_{\mathcal{H}}$, which implies $\tilde{h} \in \mathcal{H}_n$ and therefore by Lemma A.9 $\max_{1 \leq i \leq n} \|\hat{\Psi}_i(\tilde{h})\| \xrightarrow{p} 0$

and $\tilde{h} \in \widehat{\mathcal{H}}(\hat{\theta})$ w.p.a.1. Using the same steps as in the proof of Lemma A.12 we can Taylor expand the empirical FGEL objective about $h = 0$,

$$\widehat{G}_{\lambda_n}(\hat{\theta}, \tilde{h}) = \phi(0) - \hat{\Psi}(\tilde{h}) - \frac{1}{2}\langle \tilde{h}, \widehat{\Omega}_{\lambda_n}(\dot{h}, \hat{\theta})\tilde{h}\rangle_{\mathcal{H}},$$

for some $\dot{h}$ on the line between 0 and $\tilde{h}$. Now by Lemma A.11 the largest eigenvalue of the regularized covariance operator $\widehat{\Omega}_{\lambda_n}(\dot{h}, \hat{\theta})$ can be bounded by a constant $\tilde{C} > 0$ w.p.a.1. Therefore, we have w.p.a.1,

$$\widehat{G}_{\lambda_n}(\hat{\theta}, \tilde{h}) \geq \phi(0) + n^{-\zeta}\|\hat{\Psi}\|_{\mathcal{H}^*} - Cn^{-2\zeta},$$

where for the second term we have used the definition of $\tilde{h}$ and the notation $C = \frac{1}{2}\tilde{C}$. Consider $\bar{\theta} = \theta_0$ in Lemma A.12, for which the requirements are fulfilled as $\|E_{\hat{P}_n}[\Psi(X, Z; \theta_0)]\|_{\mathcal{H}^*} = O_p\left(n^{-1/2}\right)$ by Lemma A.7 using Assumption h). Moreover, being the solution to the mini-max problem, $(\hat{\theta}, \hat{h})$ correspond to a saddle point of the *empirical* FGEL objective $\widehat{G}_{\lambda_n}$. Using this and Lemma A.12 we have

$$\phi(0) + n^{-\zeta}\|\hat{\Psi}\|_{\mathcal{H}^*} - Cn^{-2\zeta} \leq \widehat{G}_{\lambda_n}(\hat{\theta}, \tilde{h}) \leq \widehat{G}_{\lambda_n}(\hat{\theta}, \hat{h}) \leq \sup_{h \in \widehat{\mathcal{H}}(\theta_0)} \widehat{G}_{\lambda_n}(\theta_0, h) \leq \phi(0) + O_p(n^{-1}).$$

Now, subtracting $\phi(0)$ on both sides and solving for $\|\hat{\Psi}\|_{\mathcal{H}^*}$, we obtain

$$\|\hat{\Psi}\|_{\mathcal{H}^*} \leq O_p(n^{\zeta-1}) + Cn^{-\zeta} = O_p(n^{-\zeta}), \tag{A.5}$$

which follows as $1/\nu < \zeta < 1/2$ and therefore $\zeta - 1 < -1/2 < -\zeta$. Consider any $\epsilon_n \to 0$ and let $\tilde{h} = -\epsilon_n \mu(\hat{\Psi})$. Then by (A.5), $\tilde{h} = o_p(n^{-\zeta})$ and therefore $\tilde{h} \in \mathcal{H}_n$ w.p.a.1. Then as previously we have

$$\phi(0) - \hat{\Psi}(\tilde{h}) - C\|h\|_{\mathcal{H}}^2 = \phi(0) + \epsilon_n\|\hat{\Psi}\|_{\mathcal{H}^*}^2 - C\epsilon_n^2\|\hat{\Psi}\|_{\mathcal{H}^*}^2 \leq \phi(0) + O_p(n^{-1}).$$

As $1 - \epsilon_n C$ is bounded away from zero, for all n large enough, we have $\epsilon_n\|\hat{\Psi}\|_{\mathcal{H}^*}^2 = O_p(n^{-1})$. As this holds for all $\epsilon_n \to 0$, it follows that $\|\hat{\Psi}\|_{\mathcal{H}^*} = O_p(n^{-1/2})$. □

Proof of Theorem 2.6

Proof. Define $\hat{\Psi}_i = \Psi(x_i, z_i; \hat{\theta})$ and $\hat{\Psi} = \frac{1}{n}\sum_{i=1}^n \hat{\Psi}_i$. By Assumption h) and Lemma A.7, we have $\|\hat{\Psi}(\theta) - E[\Psi(X, Z; \theta)]\|_{\mathcal{H}^*} = O_p(n^{-1/2})$ for any $\theta \in \Theta$. From Lemma A.13 we also have $\|\hat{\Psi}\|_{\mathcal{H}^*} = O_p(n^{-1/2})$ and thus using the triangle inequality we get

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

As by Assumption a) θ_0 is the unique parameter for which $\|E[\Psi(X,Z;\theta)]\|_{\mathcal{H}^*} = 0$ it follows that $\hat{\theta} \xrightarrow{p} \theta_0$.

Following the proof of Theorem A.1 of Kremer et al. [93] we can use this result to translate the convergence rate of the moment functional to a convergence rate of the FGEL estimator $\hat{\theta}$ using Assumptions i)-k). The proof is identical to the one provided by Kremer et al. [93] and we state it here merely for completeness.

By the mean value theorem, there exists $\bar{\theta} \in \operatorname{conv}(\{\theta_0, \hat{\theta}\})$ such that

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

Using this we have

$$\begin{aligned}
\|E[\Psi(X,Z;\hat{\theta})]\|^2_{\mathcal{H}^*} &= \| \underbrace{E[\Psi(X,Z;\theta_0)]}_{=0} + (\hat{\theta}-\theta_0)^T E[\nabla_\theta \Psi(X,Z;\bar{\theta})]\|^2_{\mathcal{H}^*} \\
&= \left\langle (\hat{\theta}-\theta_0)^T E[\nabla_\theta \Psi(X,Z;\bar{\theta})], (\hat{\theta}-\theta_0)^T E[\nabla_\theta \Psi(X,Z;\bar{\theta})] \right\rangle_{\mathcal{H}^*} \\
&= (\hat{\theta}-\theta_0)^T \underbrace{\left\langle E[\nabla_\theta \Psi(X,Z;\bar{\theta})], E[\nabla_{\theta^T} \Psi(X,Z;\bar{\theta})] \right\rangle_{\mathcal{H}^*}}_{=:\Sigma(\bar{\theta})} (\hat{\theta}-\theta_0) \\
&\geq \lambda_{\min}\left(\Sigma(\bar{\theta})\right) \|\hat{\theta}-\theta_0\|_2^2
\end{aligned}$$

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 k), thus the smallest eigenvalue of $\Sigma(\bar{\theta})$, $\lambda_{\min}(\Sigma(\bar{\theta}))$, is positive and bounded away from zero w.p.a.1. Finally as $\|E[\Psi(X,Z;\hat{\theta})]\| = O_p(n^{-1/2})$ taking the square-root on both sides we have $\|\hat{\theta}-\theta_0\| = O_p(n^{-1/2})$.

□

Proof of Theorem 2.7 (Asymptotic Normality)

Lemma A.14. *Let* $\Omega_0 = E[\Psi(X,Z;\theta_0)^* \Psi(X,Z;\theta_0)] \in \mathcal{H} \times \mathcal{H}$ *and* $\Sigma_0 = \langle E[\nabla_\theta \Psi(X,Z;\theta_0)], E[\nabla_{\theta^T} \Psi(X,Z;\theta_0)] \rangle_{\mathcal{H}^*} \in \mathbb{R}^{p \times p}$ *be non-singular. Then the matrix*

$$M = -\begin{pmatrix} 0 & \nabla_\theta \Psi_0 \\ \nabla_{\theta^T} \Psi_0^* & \Omega_0 \end{pmatrix}.$$

is invertible and its inverse is given by

$$M^{-1} = \begin{pmatrix} \Xi & -\Xi \left(\nabla_\theta \Psi_0\right) \Omega_0^{-1} \\ -\Omega_0^{-1} \left(\nabla_{\theta^T} \Psi_0^*\right) \Xi & \Omega_0^{-1} + \Omega_0^{-1} \left(\nabla_{\theta^T} \Psi_0^*\right) \Xi \left(\nabla_\theta \Psi_0\right) \Omega_0^{-1} \end{pmatrix}$$

with $\Xi = \left(\left(\nabla_\theta \Psi_0\right) \Omega_0^{-1} \left(\nabla_{\theta^T} \Psi_0^*\right)\right)^{-1} \in \mathbb{R}^{p \times p}$.

Proof. In order to find the inverse of M we resort to standard blockmatrix algebra. Note that the inverse of a matrix

$$P = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

can be derived as

$$P^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{pmatrix}$$

whenever D and its Schur complement in P, $P/D = A - BD^{-1}C$ are invertible (see e.g., Bernstein [19]).

The Schur complement of Ω_0 in M is a matrix $\Gamma \in \mathbb{R}^{p \times p}$ defined as

$$\Gamma := M/\Omega_0 = -\left(\nabla_\theta \Psi_0\right) \Omega_0^{-1} \left(\nabla_{\theta^T} \Psi_0^*\right).$$

Now, as Ω_0 is a positive definite operator by assumption, its smallest eigenvalue is bounded away from zero. It immediately follows that the smallest eigenvalue of Ω_0^{-1}, $\lambda_{\min}(\Omega_0^{-1}) > 0$ is bounded away from zero and thus we have for any $\theta \in \Theta$ with $\|\theta\|_2 > 0$,

$$\begin{aligned} \theta^T \Gamma \theta &= -\theta^T \left(\nabla_\theta \Psi_0\right) \Omega_0^{-1} \left(\nabla_{\theta_T} \Psi_0^*\right) \theta \\ &\leq -\lambda_{\min}(\Omega_0^{-1}) \theta^T \underbrace{\langle E[\nabla_\theta \Psi(X, Z; \theta_0)], E[\nabla_{\theta^T} \Psi(X, Z; \theta_0)] \rangle_{\mathcal{H}^*}}_{=\Sigma_0} \theta \\ &\leq -\lambda_{\min}(\Omega_0^{-1}) \lambda_{\min}(\Sigma_0) \|\theta\|_2^2 \\ &< 0, \end{aligned}$$

where we used that Σ_0 is symmetric by construction and non-singular by assumption and therefore positive definite so its smallest eigenvalue $\lambda_{\min}(\Sigma_0) > 0$. Now as $\Gamma \in \mathbb{R}^{p \times p}$ is a strictly negative definite matrix it is non-singular and thus invertible. Finally, as Ω_0 and its Schur complement in M, Γ, are invertible, we can employ the standard blockmatrix inversion formula to arrive at the result. □

Proof of Theorem 2.7 The proof generalizes Theorem 3.2 of Newey and Smith [118] to our regularized continuum estimator.

Proof. Define $\Psi_i(\theta) := \Psi_i(x_i, z_i; \theta)$ and $\hat{\Psi} = \frac{1}{n} \sum_{i=1}^n \Psi_i(\theta)$ and analogous $\Psi_i^*(\theta) = \Psi_i^*(x_i, z_i; \theta)$ and $\Psi^*(\theta) = \frac{1}{n} \sum_{i=1}^n \Psi_i^*(\theta)$. Let $\hat{\theta}, \hat{h}$ denote the FGEL estimates of the parameters θ and Lagrange multiplier function h. The first order optimality conditions of the saddle point objective (2.12) are

given by

$$D_h \widehat{G}_{\lambda_n}(\hat{\theta}, \hat{h}) = \frac{1}{n} \sum_{i=1}^{n} \phi_1(\Psi_i(\hat{\theta})(\hat{h})) \Psi_i^*(\hat{\theta}) - \lambda_n \hat{h} = 0 \tag{A.6}$$

$$\nabla_\theta \widehat{G}_{\lambda_n}(\hat{\theta}, \hat{h}) = \frac{1}{n} \sum_{i=1}^{n} \phi_1(\Psi_i(\hat{\theta})(\hat{h}))(\nabla_\theta \Psi_i(\hat{\theta}))(\hat{h}) = 0, \tag{A.7}$$

where $\nabla_\theta \Psi_i(\theta) \in \Theta \times \mathcal{H}^*$ is the gradient of the function $\theta \mapsto \Psi_i(\theta)$ w.r.t. θ. Define $\beta = (\theta, h)$ then using Taylor's theorem (Proposition A.4) we can linearize the first order conditions about the true parameters $\beta_0 = (\theta_0, 0)$ which yields for the first condition (A.6)

$$\begin{aligned}
0 = &- \frac{1}{n} \sum_{i=1}^{n} \Psi_i^*(\theta_0) + \frac{1}{n} \sum_{i=1}^{n} \phi_2(\Psi_i(\dot{\theta})(\dot{h})) \Psi_i^*(\dot{\theta}) \Psi_i(\dot{\theta})(\hat{h}) \\
&- \lambda_n \hat{h} + \frac{1}{n} \sum_{i=1}^{n} \phi_1(\Psi_i(\dot{\theta})(\dot{h})) \nabla_{\theta^T} \Psi_i^*(\dot{\theta})(\hat{\theta} - \theta_0) \\
&+ \frac{1}{n} \sum_{i=1}^{n} \phi_2(\Psi_i(\dot{\theta})(\dot{h})) \Psi_i^*(\dot{\theta})(\nabla_{\theta^T} \Psi_i(\dot{\theta}))(\dot{h})(\hat{\theta} - \theta_0),
\end{aligned}$$

where $(\dot{\theta}, \dot{h})$ lies on the line between $(\hat{\theta}, \hat{h})$ and $(\theta_0, 0)$. For the second condition (A.7) we obtain

$$\begin{aligned}
0 = &\frac{1}{n} \sum_{i=1}^{n} \phi_1(\Psi_i(\bar{\theta})(\bar{h}))(\nabla_\theta \Psi_i(\bar{\theta}))(\hat{h}) \\
&+ \frac{1}{n} \sum_{i=1}^{n} \phi_2(\Psi_i(\bar{\theta})(\bar{h}))(\nabla_\theta \Psi_i(\bar{\theta}))(\bar{h}) \Psi_i(\bar{\theta})(\hat{h}) \\
&+ \Bigg\{ \frac{1}{n} \sum_{i=1}^{n} \phi_2(\Psi_i(\bar{\theta})(\bar{h}))(\nabla_\theta \Psi_i(\bar{\theta}))(\bar{h})(\nabla_\theta \Psi_i(\bar{\theta}))(\bar{h}) \\
&+ \frac{1}{n} \sum_{i=1}^{n} \phi_1(\Psi_i(\bar{\theta})(\bar{h})) D_\theta^2(\Psi_i(\bar{\theta}))(\bar{h}) \Bigg\} (\hat{\theta} - \theta_0),
\end{aligned}$$

where again $(\bar{\theta}, \bar{h})$ lies on the line between $(\hat{\theta}, \hat{h})$ and $(\theta_0, 0)$. Now as $\hat{h} = o_p(1)$ we have $\bar{h} = o_p(1)$ and $\dot{h} = o_p(1)$. Therefore for $n \to \infty$ most terms go to zero and we are left with the first condition (A.6) reducing to

$$\begin{aligned}
0 = &- \frac{1}{n} \sum_{i=1}^{n} \Psi_i^*(\theta_0) + \frac{1}{n} \sum_{i=1}^{n} \phi_2(\Psi_i(\dot{\theta})(\dot{h})) \Psi_i^*(\dot{\theta}) \Psi_i(\dot{\theta})(\hat{h}) \\
&- \lambda_n \hat{h} + \frac{1}{n} \sum_{i=1}^{n} \phi_1(\Psi_i(\dot{\theta})(\dot{h})) \nabla_{\theta^T} \Psi_i^*(\dot{\theta})(\hat{\theta} - \theta_0) \\
&+ o_p(1)
\end{aligned}$$

and the second (A.7) to

$$0 = \frac{1}{n}\sum_{i=1}^{n} \phi_1(\Psi_i(\bar{\theta})(\bar{h}))(\nabla_\theta \Psi_i(\bar{\theta}))(\hat{h}) + o_p(1).$$

As $\hat{h} = O_p(n^{-1/2})$ and $\bar{h}, \dot{h}$ lie between $\hat{h}$ and 0, the conditions of Lemma A.9 are fulfilled, i.e., $\bar{h}, \dot{h} \in \mathcal{H}_n$, and hence $\max_{1\le i\le n} |\Psi_i(\bar{\theta})(\bar{h})| \xrightarrow{p} 0$ and $\max_{1\le i\le n} |\phi_1(\Psi_i(\bar{\theta})(\bar{h})) + 1| \xrightarrow{p} 0$ as well as $\max_{1\le i\le n} |\phi_2(\Psi_i(\bar{\theta})(\bar{h})) + 1| \xrightarrow{p} 0$ and the same equivalently holds for $\dot{\theta}$ and $\dot{h}$.

Define $\hat{\beta} = (\hat{\theta}, \hat{h})$ and $\beta_0 = (\theta_0, 0)$, then, we can write the conditions in matrix form as

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{1}{n}\sum_{i=1}^{n} \Psi_i^*(\theta_0) \end{pmatrix} + \underbrace{\begin{pmatrix} 0 & -\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \Psi_i(\dot{\theta}) \\ -\frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta^T} \Psi_i^*(\bar{\theta}) & -\widehat{\Omega}_{\lambda_n}(\dot{h}, \dot{\theta}) \end{pmatrix}}_{=:M_n} (\hat{\beta} - \beta_0) + o_p(1) \tag{A.8}$$

Now by the weak law of large numbers and the continuous mapping theorem, we have $\widehat{\Omega}_{\lambda_n}(\dot{h}, \dot{\theta}) \xrightarrow{p} \Omega(\theta_0) =: \Omega_0$, which is a non-singular and thus invertible operator. Further for the off-diagonal elements, as $\hat{\theta} \xrightarrow{p} \theta_0$ and by continuity of $\theta \mapsto \nabla_\theta \Psi(x, z; \theta)$ we have by the continuous mapping theorem and the weak law of large numbers $-\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \Psi_i(\dot{\theta}) \xrightarrow{p} -E[\nabla_\theta \Psi(X, Z; \theta_0)] =: -\nabla_\theta \Psi_0$. Therefore, as $n \to \infty$ we have $M_n \to M$ with

$$M = -\begin{pmatrix} 0 & \nabla_\theta \Psi_0 \\ \nabla_{\theta^T} \Psi_0^* & \Omega_0 \end{pmatrix}.$$

Using Assumptions e) and k) of Theorem 2.6, it follows from Lemma A.14 that the blockoperator M is non-singular and thus invertible with inverse

$$M^{-1} = -\begin{pmatrix} -B & C \\ C^* & D \end{pmatrix}$$

with $B = ((\nabla_\theta \Psi_0)\Omega_0^{-1}(\nabla_{\theta^T} \Psi_0^*))^{-1}$, $C = B(\nabla_\theta \Psi_0)\Omega_0^{-1}$ and $D = \Omega_0^{-1} - \Omega_0^{-1}(\nabla_{\theta^T} \Psi_0^*)B(\nabla_\theta \Psi_0)\Omega_0^{-1}$.

With this at hand we can solve (A.8) for $\hat{\beta} - \beta_0$ which yields

$$\sqrt{n}(\hat{\theta} - \theta) = -\sqrt{n}C\left(\frac{1}{n}\sum_{i=1}^{n} \Psi_i^*(\theta_0)\right) + o_p(1)$$
$$\sqrt{n}(\hat{h} - h) = -\sqrt{n}D\left(\frac{1}{n}\sum_{i=1}^{n} \Psi_i^*(\theta_0)\right) + o_p(1)$$

Finally by the Donsker property of Ψ we have $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\Psi(X,Z;\theta_0)\right) \sim \mathcal{N}(0,\Omega_0)$ and thus we get

$$\sqrt{n}(\hat{\theta}-\theta_0) \sim \mathcal{N}(0, C\Omega_0 C^*)$$

where

$$\begin{aligned} C\Omega_0 C^* &= ((\nabla_\theta \Psi_0)\Omega_0^{-1}(\nabla_{\theta^T}\Psi_0^*))^{-1}(\nabla_\theta\Psi_0)\Omega_0^{-1}\Omega_0\Omega_0^{-1}(\nabla_{\theta^T}\Psi_0^*)((\nabla_\theta\Psi_0)\Omega_0^{-1}(\nabla_{\theta^T}\Psi_0^*))^{-1} \\ &= (\nabla_\theta\Psi_0)\Omega_0^{-1}(\nabla_{\theta^T}\Psi_0^*))^{-1} \end{aligned}$$

□

Asymptotic Properties for CMR

The asymptotic properties of the FGEL estimator for conditional moment restrictions follow by expressing the conditional moment restrictions (2.6) in terms of the equivalent variational/functional formulation (2.7) and translating the assumptions of Theorems 2.8 and 2.9 into the conditions required for Theorems 2.6 and 2.7 respectively. The proofs are almost identical to the ones provided by Kremer et al. [93] for translating the results on their KMM estimator for functional moment restrictions to the conditional moment restriction case. Therefore, we only state the differences here and refer to the proofs of Theorems 3.4-3.6 of Kremer et al. [93] for details.

Proof of Theorem 2.8 The proof follows directly from the proof of Theorem 3.4 of Kremer et al. [93], ignoring their assumption f) which is not required here and instead using that the additional Assumption g) of the FGEL estimator is fulfilled by the identical Assumption g) in Theorem 2.6. Further the Donsker property h) of Theorem 2.6 is fulfilled by the corresponding Assumption h) and Lemma A.6 using that both ψ and h are uniformly bounded as continuous functions on compact domains.

Proof of Theorem 2.9 The proof is identical to the one provided by Kremer et al. [93] for their Theorem 3.5.

Proof of Theorem 2.10 Efficiency follows immediately from Theorem 2.9 as the asymptotic variance of the FGEL estimator agrees with the semi-parametric efficiency bound for CMR estimators by Chamberlain [29].

A.2.3 Kernel FGEL

Proof of Theorem 2.11

Proof. The proof follows the proof of Theorem 3.2 in Muandet et al. [109]. Equation (2.15) follows from (2.14) directly by the law of iterated expectation. To see this, assume $E[\psi(X;\theta)|Z] = 0$ P_Z-a.s.,

then $\forall h \in \mathcal{H}$

$$\begin{aligned} E[\Psi(X, Z; \theta)(h)] &= E[\psi(X; \theta)h(Z)] \\ &= E[E[\psi(X; \theta)h(Z)|Z]] \\ &= E[E[\psi(X; \theta)|Z]h(Z)] \\ &= 0. \end{aligned}$$

For the other direction, note that $E[\Psi(X, Z; \theta)(h)] = 0 \ \forall h \in \mathcal{H}$ implies $\sup_{h\in\mathcal{H}} E[\Psi(X, Z; \theta)(h)] = 0$ and thus

$$\begin{aligned} 0 &= \sup_{h\in\mathcal{H}} E[\Psi(X, Z; \theta)(h)] \\ &= \sum_{j=1}^{m} \sup_{\|h_j\|_{\mathcal{H}}\leq 1} E[\psi_j(X; \theta)h_j(Z)] \\ &= \sum_{j=1}^{m} \sup_{\|h_j\|_{\mathcal{H}}\leq 1} \langle E[\psi_j(X; \theta)k_j(Z, \cdot)], h_j\rangle \\ &= \sum_{j=1}^{m} \|E[\psi_j(X; \theta)k_j(Z, \cdot)]\|_{\mathcal{H}} \\ &= \sum_{j=1}^{m} \|E_Z[\underbrace{E_X[\psi_j(X; \theta)|Z]}_{:=\xi_j(Z)} k_j(Z, \cdot)]\|_{\mathcal{H}} \\ &= \sum_{j=1}^{m} \| \int_{\mathcal{Z}} \xi_j(z)k_j(z, \cdot)p(z)dz\|_{\mathcal{H}} \end{aligned}$$

As each element of the sum is non-negative, we must have for $j = 1, \ldots, m$,

$$\begin{aligned} 0 &= \| \int_{\mathcal{Z}} \xi_j(z)k_j(z, \cdot)p(z)\mathrm{d}z\|_{\mathcal{H}} \\ &= \| \int_{\mathcal{Z}} \xi_j(z)k_j(z, \cdot)p(z)\mathrm{d}z\|^2_{\mathcal{H}} \\ &= \int_{\mathcal{Z}\times\mathcal{Z}} \xi_j(z)\langle k_j(z, \cdot), k_j(z', \cdot)\rangle_{\mathcal{H}}\xi_j(z')p(z)p(z')\mathrm{d}z\mathrm{d}z' \\ &= \int_{\mathcal{Z}\times\mathcal{Z}} \xi_j(z)k_j(z, z')\xi_j(z')p(z)p(z')\mathrm{d}z\mathrm{d}z'. \end{aligned}$$

By definition of ISPD kernels (see Section 2.3) this directly implies $\|\xi_j(z)p(z)\|_2^2 = 0$. It follows that $\xi_j(z) = 0$ a.e. on the support of $p(z)$ and thus $P_Z(\{z \in \mathcal{Z} : \xi_j(z) = 0\}) = 1$. Finally this implies

$$\xi_j(Z) = E[\psi_j(X; \theta)|Z] = 0 \ \ P_Z\text{-a.s.}, \ \ j = 1, \ldots, m,$$

which completes the equivalence between (2.14) and (2.15). □

Proof of Corollary 2.12

Proof. Equivalence between (2.6) and (2.7) holds for any RKHS corresponding to a universal ISPD kernels by Theorem 2.11. Moreover, the local Lipschitz property is fulfilled as in any RKHS the evaluation functional is bounded, which implies that for any $h \in \mathcal{H}$

$$\begin{aligned}\|h(z_1) - h(z_2)\| &= \|\langle h, k(z_1, \cdot)\rangle - \langle h, k(z_2, \cdot)\rangle\| \\ &\leq \|h\|\|k(z_1, \cdot) - k(z_2, \cdot)\| \\ &\leq C\left(\|k(z_1, \cdot)\|_{\mathcal{H}} + \|k(z_2, \cdot)\|_{\mathcal{H}}\right) \\ &\leq L\end{aligned}$$

where we used the Cauchy-Schwarz and triangle inequalities. Finally the Donsker property of $\mathcal{H}_1$ follows from Lemma 17 of Bennett and Kallus [14].

□

Proof of Lemma 2.13

Proof. The profile divergence can be written as

$$R_{\lambda_n}(\theta) = \inf_{h \in \widehat{\mathcal{H}}} -\sum_{i=1}^{n} \phi(\Psi(x_i, z_i; \theta)(h) + \frac{\lambda_n}{2}\|h\|_{\mathcal{H}}.$$

As $-\phi$ is a convex function and $\widehat{\mathcal{H}}$ is convex it follows that this is a convex optimization problem. Therefore, we can employ the representer theorem Schölkopf et al. [142] and express each component r of the m-dimensional vector of RKHS functions as $h_r(\cdot) = \sum_{i=1}^{n}(\alpha_r)_i k_r(z_i, \cdot)$, with $\alpha_r \in \mathbb{R}^n$. Therefore, we get

$$\begin{aligned}\Psi(x_i, z_i; \theta)(h) &= \sum_{r=1}^{m}\sum_{i,j=1}^{n} (\alpha_r)_j \left(K_r\right)_{ji} \psi_r(x_i, z_i; \theta) \\ \|h\|_{\mathcal{H}}^2 &= \sum_{r=1}^{m}\sum_{i,j=1}^{n} (\alpha_r)_i \langle k_r(z_i, \cdot), k_r(z_j, \cdot)\rangle (\alpha_r)_j = \sum_{r=1}^{m} \alpha_r^T K_r \alpha_r\end{aligned}$$

Inserting this back into $R_{\lambda_n}(\theta)$ yields the result. □

A.2.4 Additional Proofs

Proof of Proposition 2.14

Proof. The result follows directly by inserting $\phi(v) = \left(1 \pm \frac{v}{2}\right)^2$ into (2.12) and using that as $\mathcal{H}$ is a vector space, for every $h \in \mathcal{H}$, its negative $-h$ is also contained in $\mathcal{H}$. Therefore the first order conditions agree for the positive and negative sign in ϕ. □

Appendix B

Kernel Method of Moments

B.1 KMM for Functional Moment Restrictions

B.1.1 Duality

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

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

where we relaxed the moment restrictions to hold only exactly for $n \to \infty$. Note that to be precise, the dual of (B.1), which can be obtained following the proof of Theorem 3.3, contains a regularization term $-\lambda_n \|h\|_{\mathcal{H}}$ instead of $-\frac{1}{2}\|h\|_{\mathcal{H}}^2$ as in our Definition 3.4. 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 (B.1) 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. [92], 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. [92] where the relaxation scheme is a fundamental necessity to restore strong duality, here the regularization parameter can be seen merely as a computational tool.

B.1.2 Asymptotic Properties

For the KMM estimator for functional moment restrictions (FMR) of the form (3.9) based on the functional MMD profile (3.10) we have the following properties.

Theorem B.1 (Consistency for FMR)**.** *Assume that $\mathcal{H}$ is a space of continuous functions and a) $\theta_0 \in \Theta$ is the unique solution to $\|E[\Psi(X, Z; \theta)]\|_{\mathcal{H}^*} = 0$; b) $\Theta \subset \mathbb{R}^p$ and $\mathcal{X} \times \mathcal{Z}$ are compact; c) $\Psi(x, z; \theta)$ is continuous in x, z and θ everywhere; d) $E[\sup_{\theta \in \Theta} \|\Psi(X, Z; \theta)\|^2_{\mathcal{H}^*}] < \infty$; 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 $E_Q[\sup_{\theta \in \Theta} \|\Psi(X, Z; \theta)\|^2_{\mathcal{H}^*}] < \infty$; g) $\lambda_n = O_p(n^{-\xi})$ with $0 < \xi < 1/2$; and h) the function class $\{\Psi(\cdot; \theta)(h) : \theta \in \Theta, h \in \mathcal{H}_1\}$ is P_0-Donsker. Let $\hat{\theta}$ denote the functional KMM estimator for θ_0, then $\hat{\theta} \xrightarrow{p} \theta_0$.*

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

Theorem B.2 (Asymptotic Normality for FMR)**.** *Let the Assumptions of Theorem B.1 be satisfied. Then,*

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

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

B.2 Asymptotic Properties of the Finite-Dimensional KMM Estimator

For the KMM estimator for finite dimensional moment restrictions based on (3.8) we have the following results.

Theorem B.3 (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) $\theta_0 \in \operatorname{int}(\Theta)$; 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) $\operatorname{rank}\left(E[\nabla_\theta \psi(X; \theta_0)]\right) = p$, we have $\|\hat{\theta} - \theta_0\| = O_p(n^{-1/2})$.

Theorem B.4 (Asymptotic Normality for MR)**.** *Let Assumptions a)-i) of Theorem B.3 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.5. *The asymptotic variance* Ξ_0 *of the KMM estimator agrees with the one of the optimally weighted GMM estimator [69], thus for finite dimensional moment restrictions KMM and OW-GMM are asymptotically first-order equivalent.*

B.3 Additional Experimental Details

B.3.1 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 $\mathrm{HSIC}(\psi(X;\theta), Z)$ 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 [46] 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 $\varphi \in [\mathrm{KL}, \log, \chi^2]$.

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 3.3.3. In agreement with the observations of Kremer et al. [92] we noticed experimentally that the choice of φ-divergence has only a minor effect on the obtained estimator.

B.3.2 Choice of Validation Metric and Failure of MMR

The computation of modern CMR estimators including DeepGMM [15], Functional GEL [92] 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

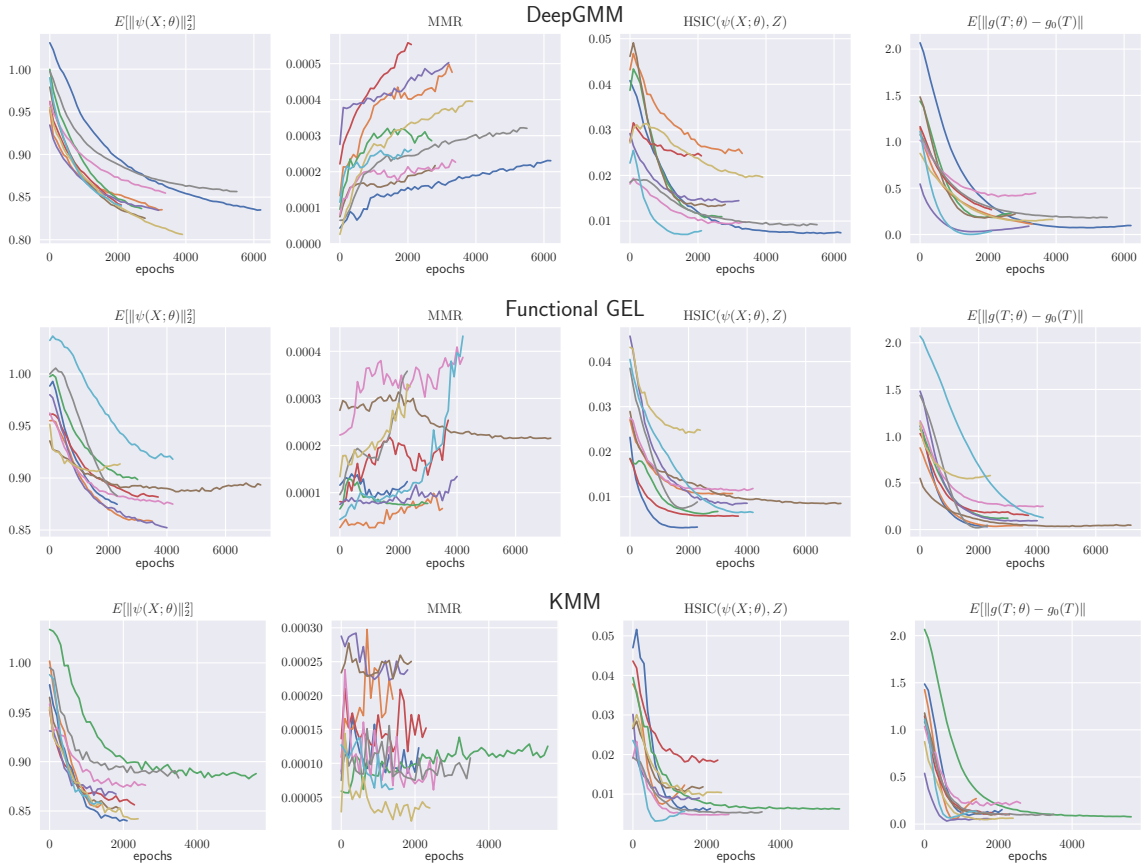


Fig. B.1 Effects of Validation Metrics for Early Stopping. Visualization of different validation losses for 10 training samples and different estimators. Goal of the estimation is to minimize the error with respect to the true function g_0 shown on the right which is unknown in practice. We observe that 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.

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. [15] and Kremer et al. [92] used the maximum moment restriction objective [179] $\text{MMR}(\theta) = E_{\hat{P}_n}[\psi(X;\theta)K(Z,Z')\psi(X';\theta)]$ 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. [179] 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. [137] 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) [63].

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 B.1. 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.

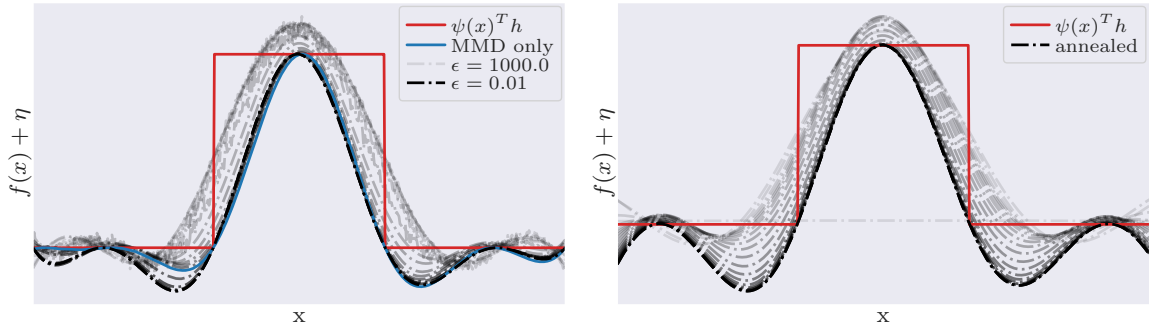
B.4 Entropy Regularization

B.4.1 Effect of the Regularization Parameter

As discussed in Section 3.3.2, 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 [114]. 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 (3.8) while keeping h and θ fixed, which preserves the convex structure of the problem. In Figure B.2(a) we observe that for smaller ϵ we get closer and closer to the original *MMD-profile*, which we obtain from equation (3.5) by using a sample approximation of the semi-infinite constraint in a convex solver.

B.4.2 Annealing of Entropy Regularization

Instead of keeping the regularization parameter ϵ fixed during the optimization procedure as in Figure B.2(a), we study an annealing schedule in which it is gradually decreased, similar to actual interior-point methods. Chizat [40] 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



(a) Regularization Parameters

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

Fig. B.2 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.

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 B.2(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.

B.4.3 Choice of Reference Measure

The KMM estimator based on (3.8) and (3.10) 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 (3.8) with the information contained in the empirical distribution, one can represent the reference distribution Q by a kernel density estimator (KDE) [134] 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 [59], variational auto-encoders [86], normalizing flows [124] or diffusion models [75]. 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 (3.7) as

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

Carrying out the optimization over $\mu \in \mathcal{F}$ and defining $\beta = (\eta, f, h) \in \mathcal{M}$, the Lagrangian of the MMD profile (3.7) can be cast in the (semi-dual) form,

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

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

$$P_{k+1} \in \arg\min_P \int \left(-f(x) - \eta + \psi(x;\theta)^T h\right) dP(x) + \epsilon D(P\|P_k). \tag{B.2}$$

and subsequently update the dual variables β, where we solve

$$\beta_{k+1} \in \arg\max_{\beta \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} f(x_i) + \eta - \frac{1}{2}\|f\|_{\mathcal{H}}^2 + \int \left(-f(x) - \eta + \psi(x;\theta)^T h\right) dP_{k+1}(x) \qquad \text{(B.3)}$$
$$+ \frac{1}{2t}\|\beta - \beta_k\|_{\mathcal{M}}^2.$$

The update rule (B.3) is a standard convex optimization problem. We leave a specific implementation of this approach to future work.

B.5 Proofs

B.5.1 Definitions and Preliminaries

To simplify notation and analysis we first provide a compact formulation of the functional KMM objective (3.10) given by

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

where $\beta = (\eta, f, h)$. 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 $k((x,z),\cdot):\mathcal{F}\to\mathbb{R}$, 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 $i=1,\dots,n$,

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

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 (3.10) can be written in the compact form

$$G_{\epsilon,\lambda_n}(\theta,\beta) = \frac{1}{n}\sum_{i=1}^{n} b_i^T\beta - \epsilon\int_{\mathcal{X}\times\mathcal{Z}}\varphi^*\left(\frac{1}{\epsilon}a(x,z;\theta)^T\beta\right)\omega(\mathrm{d}x\otimes\mathrm{d}z) - \frac{1}{2}\beta^T R_{\lambda_n}\beta. \tag{B.4}$$

Analogously for the objective of the finite dimensional KMM estimator (3.8) we have $\mathcal{H}=\mathbb{R}^m$ and $\mathcal{M}=\mathbb{R}\times\mathcal{F}\times\mathbb{R}^m$ and further define for $i=1,\dots,n$,

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

Then the unconditional KMM objective (3.8) can be written in the compact form

$$\widehat{G}_{\epsilon}(\theta,\beta) = \frac{1}{n}\sum_{i=1}^{n} b_i^T\beta - \epsilon\int_{\mathcal{X}}\varphi^*\left(\frac{1}{\epsilon}a(x;\theta)^T\beta\right)\omega(\mathrm{d}x) - \frac{1}{2}\beta^T R\beta. \tag{B.5}$$

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 B.6 (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 $\frac{\partial}{\partial h} G(h_0)(h_1) = \frac{d}{dt} G(h_0 + t h_1)\big|_{t=0}$. Likewise, we define the k-th functional derivative $\frac{\partial^k}{(\partial h)^k} G(h_0)$ at h_0 via*

$$\frac{\partial^k}{(\partial h)^k} G(h_0)\,(h_1, \ldots, h_k) = \frac{\partial^k}{\partial t_1 \ldots \partial t_k} G\left(h_0 + t_1 h_1 + \ldots + t_k h_k\right)\Big|_{t_1 = \cdots = t_k = 0}.$$

Moreover, we write $\frac{\partial^k}{(\partial h)^k} G(h_0) = 0$ as a shorthand for $\frac{\partial^k}{(\partial h)^k} G(h_0)\,(h_1, \ldots, h_k) = 0$ for all $h_1, \ldots, h_k \in \mathcal{H}$. 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 $\frac{\partial}{(\partial \theta)} G(\theta_0) = (\nabla_\theta G)\,(\theta_0) =: \nabla_\theta G(\theta_0)$.

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

Proposition B.7 (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*

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

For the consistency proofs we will require the following result.

Lemma B.8 (Lemma 18, Bennett and Kallus [14])**.** *Suppose that $\mathcal{G}$ is a class of functions of the form $g : \Xi \to \mathbb{R}$, and that $\mathcal{G}$ is P-Donsker in the sense of Kosorok [90]. Then we have*

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

B.5.2 Duality Results

Proof of Theorem 3.2

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

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

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

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

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

$$\begin{aligned} \sup_{f_0,f,h} \inf_{\mu\in\mathcal{F}} & \frac{1}{2}\|\mu - \mu_{\hat{P}_n}\|_{\mathcal{F}}^2 + \langle f,\mu\rangle_{\mathcal{F}} + f_0 \\ \text{s.t.} \quad & \langle h, \psi(x;\theta)\rangle_{\mathbb{R}^m} \geq f(x) + f_0 \quad \forall x \in \mathcal{X}. \end{aligned}$$

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

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

and reinserting yields the final dual problem

$$\begin{aligned} \sup_{f_0,f,h} & \frac{1}{n}\sum_{i=1}^n f(x_i) - \frac{1}{2}\|f\|_{\mathcal{H}}^2 + f_0 \\ \text{s.t.} \quad & \langle h, \psi(x;\theta)\rangle_{\mathbb{R}^m} \geq f(x) + f_0 \quad \forall x \in \mathcal{X}. \end{aligned}$$

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

Proof of Theorem 3.3

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 3.2 we can write the entropy regularized MMD profile

as

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

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

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

Now, collecting terms containing p we get

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

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

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

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)

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

Inserting this back into the Lagrangian we get

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

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 (B.5),

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

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

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

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

$$\begin{aligned}
\frac{1}{2}\|\mu - \mu_{\hat{P}_n}\|_{\mathcal{F}}^2 + \langle f, \mu\rangle_{\mathcal{F}} =& \frac{1}{2}\|\mu - \mu_{\hat{P}_n}\|_{\mathcal{F}}^2 + \langle f, \mu - \mu_{\hat{P}_n}\rangle_{\mathcal{F}} + \frac{1}{2}\|f\|_{\mathcal{F}}^2 + \langle f, \mu_{\hat{P}_n}\rangle_{\mathcal{F}} - \frac{1}{2}\|f\|_{\mathcal{F}}^2 \\
=& \frac{1}{2}\|\mu - \mu_{\hat{P}_n} + f\|_{\mathcal{F}}^2 + \langle f, \mu_{\hat{P}_n}\rangle_{\mathcal{F}} - \frac{1}{2}\|f\|_{\mathcal{F}}^2 \\
\geq& \langle f, \mu_{\hat{P}_n}\rangle_{\mathcal{F}} - \frac{1}{2}\|f\|_{\mathcal{F}}^2,
\end{aligned}$$

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

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

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

B.5.3 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 (3.2) as functional moment restrictions of the form (3.3) over a sufficiently rich Hilbert space of functions. In the following we show that the assumptions of Theorem 3.5 suffice to fulfill the assumptions of the theorems for the functional KMM estimator (Theorems B.1 and B.2) from which the results follow. The proofs for the functional case are deferred to Section B.5.4.

Proof of Theorem 3.5 (Consistency for CMR)

Lemma B.9. *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*

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

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

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

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

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

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

Lemma B.10. *Let the assumptions of Theorem 3.5 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

$$\begin{aligned}
\theta^T \Sigma_0 \theta &= \left\langle E[\theta^T \nabla_\theta \Psi(X, Z; \theta_0)], E[\theta^T \nabla_\theta \Psi(X, Z; \theta_0)] \right\rangle_{\mathcal{H}^*} \\
&= \left\langle E[\theta^T \nabla_\theta \Psi(X, Z; \theta_0)], \left(\sum_{i=1}^m \sum_{j=1}^\infty h_j^i \left(h_j^i\right)^* \right) E[\theta^T \nabla_\theta \Psi(X, Z; \theta_0)] \right\rangle_{\mathcal{H}^*} \\
&= \sum_{i=1}^m \sum_{j=1}^\infty \left(E[\theta^T \nabla_\theta \psi_i(X; \theta_0) h_j^i(Z)] \right)^2 \\
&= \sum_{i=1}^m \sum_{j=1}^\infty \left(E[\theta^T D_0^i(Z) h_j^i(Z)] \right)^2, \qquad \text{(B.6)}
\end{aligned}$$

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 k), 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 (B.6) we get

$$\begin{aligned}
\theta^T \Sigma_0 \theta &= \sum_{i=1}^m \sum_{j=1}^\infty \left(E[\theta^T D_0^i(Z) h_j^i(Z)] \right)^2 \\
&\geq \left(E[\theta^T D_0^s(Z) h_r^s(Z)] \right)^2 > 0.
\end{aligned}$$

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

Proof of Theorem 3.5

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

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

It remains to be shown that the assumptions imposed on ψ are sufficient for Ψ to fulfill the conditions of Theorem B.1. Assumptions a) and b) directly translate to the corresponding assumptions in Theorem B.1. Assumption c) of Theorem 3.5 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 B.1 follows as

$$
\begin{aligned}
&E[\sup_{\theta\in\Theta} \|\Psi(X, Z; \theta)\|^2_{\mathcal{H}^*}] \\
&= E[\sup_{\theta\in\Theta} \sup_{h\in\mathcal{H},\|h\|\leq 1} \|\psi(X; \theta)^T h(Z)\|^2 \\
&\leq E[E[\sup_{\theta\in\Theta} \|\psi(X; \theta)\|_2^2 \,|Z] \sup_{h\in\mathcal{H},\|h\|\leq 1} \|h(Z)\|_2^2] \\
&\leq C \int_{\mathcal{Z}} \sup_{h\in\mathcal{H},\|h\|\leq 1} \|h(z)\|_2^2 \, P_0(\mathrm{d}z)
\end{aligned}
$$

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 $z \in \mathrm{supp}(P_0)$. 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

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

Assumption e) of Theorem B.1 follows from Assumption e) and Lemma B.9. Assumption f) is identical to the corresponding Assumption f) in Theorem B.1 using the same argument as for Assumption d) for the integrability condition. Assumption g) of Theorem B.1 is identical to Assumption g). Finally, Assumption h) of Theorem B.1 follows from Assumption h) and the fact that $\psi(\cdot; \theta)$ and h are uniformly bounded as continuous functions on compact domains. Therefore Assumptions a)-h) of Theorem B.1 are fulfilled and it follows that $\hat{\theta} \xrightarrow{p} \theta_0$.

Now further, Assumption i) of Theorem B.1 is identical with Assumption i). Assumption j) of Theorem B.1 follow from Assumption j) by the same argument presented earlier for Assumption c) and d) of Theorem B.1. Finally Assumption k) of Theorem B.1 follows from Assumption k) by Lemma B.10. Therefore, Assumptions i)-k) of Theorem B.1 are fulfilled and we have $\|\hat{\theta} - \theta_0\| = O_p(n^{-1/2})$. □

Proof of Theorem 3.8 (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 B.2 using that by Theorem 3.5 the assumptions of Theorem 3.5 are sufficient to satisfy the assumptions of Theorem B.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 and Kallus [14].

Lemma B.11. *Let the assumptions of Theorem B.1 be fulfilled. Then we have*

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

Proof. By Lemma 14 of Bennett and Kallus [14] we have for any Hilbert space $\mathcal{H}$, and element $h \in \mathcal{H}$, that

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

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

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

where we used that $\mathrm{Range}(\Omega_0^{-1/2}) = \mathcal{H}^*$. 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 $\mathrm{Range}(\Omega_0^{-1/2}) = \mathcal{H}^*$. 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 and Kallus [14].

Lemma B.12. *Let the assumptions of Theorem B.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*

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

Proof. Using Lemma B.11 we can write

$$\begin{aligned} &E[\nabla_\theta \Psi(X, Z; \theta_0) \Omega_0^{-1} \Psi(X, Z; \theta_0)] \\ =&\sup_{h \in \mathcal{H}} E[\nabla_\theta \psi(X; \theta_0)^T h(Z)] - \frac{1}{4} E\left[\left(\psi(X; \theta_0)^T h(Z)\right)^2\right] =: L(h) \end{aligned} \tag{B.7}$$

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

$$\begin{aligned} \left(\frac{\partial}{\partial h} L(h^*)\right)(\epsilon) &= E[\nabla_\theta \psi(X; \theta_0)^T \epsilon(Z)] - \frac{1}{2} E[\epsilon(Z)^T \psi(X; \theta_0) \psi(X; \theta_0)^T h^*(Z)] \\ &= E\left[\epsilon(Z)^T \left(\nabla_\theta \psi(X; \theta_0) - \frac{1}{2} \psi(X; \theta_0) \psi(X; \theta_0)^T h^*(Z)\right)\right] \\ &= E\left[\epsilon(Z)^T \left(E[\nabla_\theta \psi(X; \theta_0) | Z] - \frac{1}{2} V_0(Z) h^*(Z)\right)\right] \end{aligned}$$

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

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

Inserting back into equation (B.7) and denoting $D_0(z) := E[\nabla_\theta \psi(X; \theta_0) | Z = z]$ we have

$$\begin{aligned} &E[\nabla_\theta \Psi(X, Z; \theta_0) \Omega_0^{-1} \Psi(X, Z; \theta_0)] \\ =&2E[D_0(Z)^T V_0(Z)^{-1} D_0(Z)] - E[D_0(Z)^T V_0(Z)^{-1} V_0(Z) V_0(Z)^{-1} D_0(Z)] \\ =&E[D_0(Z)^T V_0(Z)^{-1} D_0(Z)]. \end{aligned}$$

□

Proof of Theorem 3.8

Proof. The conditions of Theorem B.2 are fulfilled by the conditions of Theorem 3.8 by the proof of Theorem 3.5. We can translate the expression for the asymptotic variance in terms of the moment functional into the conditional counterpart by applying Lemma B.12 whose conditions are fulfilled by Assumption e) of Theorem 3.5. □

Proof of Corollary 3.9 (Efficiency for CMR)

Proof. This is a direct implication of Theorem 3.8 as the asymptotic variance of the KMM estimator achieves the semi-parametric efficiency bound of Chamberlain [29]. □

B.5.4 Asymptotic Properties of KMM for Functional Moment Restrictions

The consistency proofs roughly follow the general idea laid out in the seminal paper by Newey and Smith [118] with the adaption to functional moment restrictions by Kremer et al. [92]. 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.

Proof of Theorem B.1

Lemma B.13. *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 B.14. *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,

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

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 B.15. *Let the assumptions of Theorem B.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 $n \to \infty$*,*

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

Proof. Using the Cauchy-Schwarz inequality and Assumption d) with Lemma B.13,

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

Now, by Assumptions d) and f) of Theorem B.1 and Lemma B.13 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

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

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 B.16. *Under the assumptions of Theorem B.1 we have for any* $\theta \in \Theta$*,*

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

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

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

Proof. For the first statement note that

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

Now $a(x,z;\theta_0) = (1, k((x,z),\cdot), \Psi(x,z;\theta))^T$ 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 B.1. Moreover, by Assumption d) of Theorem B.1 and Lemma B.13 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

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

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

$$\begin{aligned}\int_{\mathcal{X}\times\mathcal{Z}} a(x,z;\theta)a(x,z;\theta)^T\beta\,\omega(\mathrm{d}x\otimes\mathrm{d}z) =& \frac{1}{n}\sum_{i=1}^{n} a(x_i,z_i;\theta)a(x_i,z_i;\theta)^T\beta \\ &+ \alpha\int_{\mathcal{X}\times\mathcal{Z}} a(x,z;\theta)a(x,z;\theta)^T\beta\left(\mathrm{d}Q - \mathrm{d}\hat{P}_n\right).\end{aligned} \tag{B.8}$$

Now for the second term we have

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

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 B.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 B.1 and $\hat{P}_n \to P_0$ weakly. In conclusion the norm of the integral in equation (B.8) is bounded by some constant C and as $\alpha = O_p(n^{-1})$ the statement follows. □

Lemma B.17. *Let the assumptions of Theorem B.1 be satisfied and consider $\bar{\theta} \in \Theta$ such that $\bar{\theta} \xrightarrow{p} \theta_0$ and $E[\|\psi(X;\bar{\theta}) - \psi(X;\theta_0)\|_\infty] = O_p(n^{-1/2})$. 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*

$$\Lambda_n(\beta,\theta) := \int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon} a(x,z;\theta)a(x,z;\theta)^T \varphi_2^*\left(\frac{1}{\epsilon}a(x,z;\theta)^T\beta\right)\omega(\mathrm{d}x\otimes\mathrm{d}z) + R_{\lambda_n}. \tag{B.9}$$

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 B.15 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

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

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

$$\begin{aligned}
\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} =& \Lambda_n(\bar{\beta}, \bar{\theta}) \begin{pmatrix} \eta \\ 0 \\ h \end{pmatrix} \\
&= \int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon} a(x,z;\bar{\theta}) a(x,z;\bar{\theta})^T \begin{pmatrix} \eta \\ 0 \\ h \end{pmatrix} \omega(\mathrm{d}x \otimes \mathrm{d}z) + \begin{pmatrix} 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & \lambda_n I \end{pmatrix} \begin{pmatrix} \eta \\ 0 \\ h \end{pmatrix} \\
&= \int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon} \begin{pmatrix} \eta - \Psi(x,z;\bar{\theta})(h) \\ k((x,z),\cdot)\eta - k((x,z),\cdot)\Psi(x,z;\bar{\theta})(h) \\ -\eta\Psi(x,z;\bar{\theta}) + \Psi(x,z;\bar{\theta})\Psi(x,z;\bar{\theta})(h) + \lambda_n h \end{pmatrix} \omega(\mathrm{d}x \otimes \mathrm{d}z) \\
&= \frac{1}{\epsilon}\frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} \eta - \Psi(x_i,z_i;\bar{\theta})(h) \\ k((x_i,z_i),\cdot)\eta - k((x_i,z_i),\cdot)\Psi(x_i,z_i;\bar{\theta})(h) \\ -\eta\Psi(x_i,z_i;\bar{\theta}) + \Psi(x_i,z_i;\bar{\theta})\Psi(x_i,z_i;\bar{\theta})(h) + \lambda_n h \end{pmatrix} + O_p(n^{-1}),
\end{aligned}$$

where we used Lemma B.16 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

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

The centered empirical covariance operator $\widehat{\Omega}_c(\bar{\theta})$ converges to the population covariance operator $\Omega_0 = E[\Psi(X,Z;\theta_0) \otimes \Psi(X,Z;\theta_0)]$ at rate $O_p(n^{-1/2})$, which follows as

$$\begin{aligned}
\|\widehat{\Omega}_c(\bar{\theta}) - \Omega_0\| &\leq \|\widehat{\Omega}(\bar{\theta}) - \Omega_0\| + \|E_{\hat{P}_n}[\Psi(X,Z;\bar{\theta})]\|_{\mathcal{H}^*} \\
&= O_p(n^{-1/2}) + O_p(n^{-1}) = O_p(n^{-1/2}),
\end{aligned}$$

where we used the result of Kremer et al. [92] (Lemma A.10 in this thesis) to bound the first term. As $\lambda_n = O_p(n^{-\xi})$ with $0 < \xi < 1/2$ we have that $\widehat{\Omega}_c(\bar{\theta}) + \lambda_n I$ is non-singular w.p.a.1 and thus 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

$$\begin{aligned}
\beta^T \Lambda_n \beta &= \int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon} \beta^T a(x,z;\theta) a(x,z;\theta)^T \beta \, \omega(\mathrm{d}x \otimes \mathrm{d}z) + \beta^T R_{\lambda_n} \beta \\
&\leq \int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon} \|a(x,z;\theta)^T \beta\|^2 \, \omega(\mathrm{d}x \otimes \mathrm{d}z) + \|\beta\|^2 \\
&= \int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon} \|\eta + \langle (k((x,z),\cdot), f\rangle_{\mathcal{F}} - \Psi(x,z;\theta)(h)\|^2 \, \omega(\mathrm{d}x \otimes \mathrm{d}z) + \|\beta\|^2 \\
&\leq \int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon} \left(\|\eta\|^2 + \|\langle (k((x,z),\cdot), f\rangle_{\mathcal{F}}\|^2 + \|\Psi(x,z;\theta)(h)\|^2 \right) \omega(\mathrm{d}x \otimes \mathrm{d}z) + \|\beta\|^2 \\
&\leq \int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon} \left(\|\eta\|^2 + \|f\|^2 \|k((x,z),\cdot)\|^2 + \|h\|_{\mathcal{H}}^2 \|\Psi(x,z;\theta)\|_{\mathcal{H}^*}^2 \right) \omega(\mathrm{d}x \otimes \mathrm{d}z) + \|\beta\|^2.
\end{aligned}$$

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 B.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

$$\begin{aligned}
\beta^T \Lambda_{\epsilon,\lambda_n} \beta &\leq \frac{1}{\epsilon} \left(\|\eta\|^2 + C_1 \|f\|^2 + C_2 \|h\|_{\mathcal{H}}^2 \right) + \|\beta\|^2 \\
&\leq \left(\frac{C_3}{\epsilon} + 1 \right) \|\beta\|^2,
\end{aligned}$$

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 B.18. *Let the assumptions of Theorem B.1 be satisfied. Additionally let $\bar{\theta} \in \Theta$, $\bar{\theta} \xrightarrow{p} \theta_0$, and $\|E_{\hat{P}_n}[\Psi(X,Z;\bar{\theta})]\|_{\mathcal{H}^*} = O_p\left(n^{-1/2}\right)$. 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

$$\begin{aligned}
\frac{\partial \widehat{G}}{\partial \beta}(\bar{\theta},\beta) &= \frac{1}{n}\sum_{i=1}^n b_i - \int_{\mathcal{X}\times\mathcal{Z}} a(x,z;\bar{\theta}) \varphi_1^* \left(\frac{1}{\epsilon} a(x,z;\bar{\theta})^T \beta \right) \omega(\mathrm{d}x \otimes \mathrm{d}z) - R_{\lambda_n}\beta \\
\frac{\partial^2 \widehat{G}}{(\partial\beta)^2}(\bar{\theta},\beta) &= -\int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon} a(x,z;\bar{\theta}) a(x,z;\bar{\theta})^T \varphi_2^* \left(\frac{1}{\epsilon} a(x,z;\bar{\theta})^T \beta \right) \omega(\mathrm{d}x \otimes \mathrm{d}z) - R_{\lambda_n}.
\end{aligned}$$

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 $\beta = 0$,

$$\begin{aligned}
\widehat{G}(\bar{\theta}, \beta_\zeta) =& \widehat{G}(\bar{\theta}, 0) + \frac{\partial \widehat{G}}{\partial \beta}(\bar{\theta}, 0)\beta_\zeta + \frac{1}{2}\beta_\zeta^T \frac{\partial^2 \widehat{G}}{(\partial \beta)^2}(\bar{\theta}, \dot{\beta})\beta_\zeta \qquad \text{(B.10)} \\
=& -\epsilon\varphi^*(0) + \int_{\mathcal{X}\times\mathcal{Z}} \Psi(x, z, \bar{\theta})(h_\zeta)\, \omega(\mathrm{d}x \otimes \mathrm{d}z) \\
&+ \frac{1}{n}\sum_{i=1}^n f_\zeta(x_i, z_i) - \int_{\mathcal{X}\times\mathcal{Z}} f_\zeta(x, z)\, \omega(\mathrm{d}x \otimes \mathrm{d}z) \\
&- \frac{1}{2}\beta_\zeta^T \underbrace{\left(\int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon} a(x, z; \bar{\theta}) a(x, z; \bar{\theta})^T \varphi_2^*\left(\frac{1}{\epsilon} a(x, z; \bar{\theta})^T \dot{\beta} \right) \omega(\mathrm{d}x \otimes \mathrm{d}z) + R_{\lambda_n} \right)}_{:= \Lambda_n(\dot{\beta}, \bar{\theta})} \beta_\zeta
\end{aligned}$$

for some $\dot{\beta} \in \mathrm{conv}(\{0, \beta_\zeta\})$. 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) \\
&+ \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) \\
&+ \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) \\
\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}^*} \\
&+ \|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}^*} \\
&+ \|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) \\
\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}^*} \\
&+ \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) \\
&+ \|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) \\
\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}} \\
&+ \alpha \|h_\zeta\|_{\mathcal{H}} \left(C_Q + \|\bar{\Psi}\|_{\mathcal{H}^*} \right) + \|f_\zeta\|_{\mathcal{F}} \mathrm{MMD}(\hat{P}_n, \omega; \mathcal{F}) \\
\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}^*}) + \mathrm{MMD}(\hat{P}_n, \omega; \mathcal{F}) \right)
\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 B.1 and Lemma B.13.

Lemma B.17 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,

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

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

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

As $\|\bar{\Psi}\|_{\mathcal{H}^*} = O_p(n^{-1/2})$ by assumption and by Assumption f) $\alpha = O_p(n^{-1})$ as well as $\mathrm{MMD}(\hat{P}_n,\omega;\mathcal{F}) = O(n^{-1}) = o_p(n^{-1/2})$ by Lemma B.14, 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., $\beta_\zeta \in \mathrm{int}(\mathcal{M}_n)$. 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 $\widehat{G}(\bar{\theta},\bar{\beta}) \leq -\epsilon\varphi^*(0) + \left(\|\bar{\Psi}\|_{\mathcal{H}^*} + o_p(n^{-1/2})\right))\|\bar{\beta}\|_{\mathcal{M}} - C\|\bar{\beta}\|_{\mathcal{M}}^2 = -\epsilon\varphi^*(0) + O_p(n^{-1})$. □

Lemma B.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* $\|E_{\hat{P}_n}[\Psi(X,Z;\hat{\theta})]\|_{\mathcal{H}^*} = O_p\left(n^{-1/2}\right)$.

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 B.15. 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) \\
&+ \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) \\
=& -\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) \\
&+ \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) \\
&+ \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) \\
\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}) \\
&- \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) \\
\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}^*} \\
&- \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}),
\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 B.14 and Assumption f). Using Lemma B.17 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

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

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

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

where in the last line we used Lemma B.18 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 Assumption h) and Lemma B.8

$\|E_{\hat{P}_n}[\Psi(X,Z;\theta_0)]\|_{\mathcal{H}^*} = O_p(n^{-1/2})$. Adding $\epsilon\varphi^*(0)$ on both sides and solving for $\|\hat{\Psi}\|_{\mathcal{H}^*}$ yields

$$\|\hat{\Psi}\|_{\mathcal{H}^*} \leq O_p(n^{-1+\zeta}) + O_p(n^{-\zeta}) = O_p(n^{-\zeta}), \tag{B.11}$$

where the last step follows as $\zeta < 1/2$ by definition and thus $-1+\zeta < -\zeta$. Equation (B.11) 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

$$\begin{aligned}
-\epsilon\varphi^*(0) + \kappa_n\|\hat{\Psi}\|^2_{\mathcal{H}^*} - M\kappa_n^2\|\hat{\Psi}\|^2_{\mathcal{H}^*} + O_p(n^{-1-\zeta}) &\leq \widehat{G}(\hat{\theta}, \kappa_n\tilde{\beta}) \\
&\leq \widehat{G}(\hat{\theta}, \hat{\beta}) \\
&\leq \max_{\beta\in\mathcal{M}} \widehat{G}(\theta_0, \beta) \\
&\leq -\epsilon\varphi^*(0) + O_p(n^{-1}).
\end{aligned}$$

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 B.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 Assumption h) and Lemma B.8, we have $\|\hat{\Psi}(\theta) - E[\Psi(X,Z;\theta)]\|_{\mathcal{H}^*} = O_p(n^{-1/2})$ for any $\theta \in \Theta$. From Lemma B.19 we also have $\|\hat{\Psi}\|_{\mathcal{H}^*} = O_p(n^{-1/2})$ and thus using the triangle inequality we get

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

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

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

Using this we have

$$\begin{aligned}\|E[\Psi(X,Z;\hat{\theta})]\|_{\mathcal{H}^*}^2 &= \|\underbrace{E[\Psi(X,Z;\theta_0)]}_{=0} + (\hat{\theta}-\theta_0)^T E[\nabla_\theta \Psi(X,Z;\bar{\theta})]\|_{\mathcal{H}^*}^2 \\ &= \left\langle (\hat{\theta}-\theta_0)^T E[\nabla_\theta \Psi(X,Z;\bar{\theta})], (\hat{\theta}-\theta_0)^T E[\nabla_\theta \Psi(X,Z;\bar{\theta})] \right\rangle_{\mathcal{H}^*} \\ &= (\hat{\theta}-\theta_0)^T \underbrace{\left\langle E[\nabla_\theta \Psi(X,Z;\bar{\theta})], E[\nabla_\theta \Psi(X,Z;\bar{\theta})] \right\rangle_{\mathcal{H}^*}}_{=:\Sigma(\bar{\theta})} (\hat{\theta}-\theta_0) \\ &\geq \lambda_{\min}\left(\Sigma(\bar{\theta})\right) \|\hat{\theta}-\theta_0\|_2^2\end{aligned}$$

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 k), thus the smallest eigenvalue of $\Sigma(\bar{\theta})$, $\lambda_{\min}(\Sigma(\bar{\theta}))$, is positive and bounded away from zero w.p.a.1. Finally as $\|E[\Psi(X,Z;\hat{\theta})]\| = O_p(n^{-1/2})$ taking the square-root on both sides we have $\|\hat{\theta}-\theta_0\| = O_p(n^{-1/2})$. □

Proof of Theorem B.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 B.20. *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*

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

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

Proof. Let $\mathcal{H} = \bigoplus_{i=1}^m \mathcal{H}_i$ and for $i = 1, \ldots, m$ 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 $0 < \|\theta\| < \infty$,

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

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 B.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

$$0 = \frac{\partial \widehat{G}}{\partial \theta}(\hat{\theta}, \hat{\beta}) = -\int_{\mathcal{X} \times \mathcal{Z}} \varphi_1^* \left(\frac{1}{\epsilon} a(x, z; \hat{\theta})^T \hat{\beta} \right) \nabla_\theta \left(a(x, z; \hat{\theta})^T \hat{\beta} \right) \omega(\mathrm{d}x \otimes \mathrm{d}z) \tag{B.12}$$

$$0 = \frac{\partial \widehat{G}}{\partial \beta}(\hat{\theta}, \hat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} b_i - \int_{\mathcal{X} \times \mathcal{Z}} \varphi_1^* \left(\frac{1}{\epsilon} a(x, z; \hat{\theta})^T \hat{\beta} \right) a(x, z; \hat{\theta}) \, \omega(\mathrm{d}x \otimes \mathrm{d}z) - R_{\lambda_n} \beta \tag{B.13}$$

Linearizing the first condition (B.12) about the true parameters $(\theta_0, 0)$ yields

$$\begin{aligned}
0 =& \frac{\partial \widehat{G}}{\partial \theta}(\theta_0, 0) + \left(\frac{\partial^2 \widehat{G}}{\partial \theta\, \partial \theta}(\bar{\theta}, \bar{\beta})\right)\left(\hat{\theta} - \theta_0\right) + \left(\frac{\partial^2 \widehat{G}}{\partial \theta\, \partial \beta}(\bar{\theta}, \bar{\beta})\right)\hat{\beta} \\
=& -\left(\int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon}\varphi_2^*\left(\frac{1}{\epsilon}a(x,z;\bar{\theta})^T\bar{\beta}\right)\left(\nabla_\theta a(x,z;\bar{\theta})^T\bar{\beta}\right)\left(\nabla_\theta a(x,z;\bar{\theta})^T\bar{\beta}\right)^T \omega(\mathrm{d}x\otimes\mathrm{d}z)\right)\left(\hat{\theta}-\theta_0\right) \\
&-\left(\int_{\mathcal{X}\times\mathcal{Z}} \varphi_1^*\left(\frac{1}{\epsilon}a(x,z;\bar{\theta})^T\bar{\beta}\right)\left(\nabla_\theta^2 a(x,z;\bar{\theta})^T\bar{\beta}\right)\omega(\mathrm{d}x\otimes\mathrm{d}z)\right)\left(\hat{\theta}-\theta_0\right) \\
&-\left(\int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon}\varphi_2^*\left(\frac{1}{\epsilon}a(x,z;\bar{\theta})^T\bar{\beta}\right)\left(\nabla_\theta a(x,z;\bar{\theta})^T\bar{\beta}\right)a(x,z;\bar{\theta})^T\,\omega(\mathrm{d}x\otimes\mathrm{d}z)\right)\hat{\beta} \\
&-\left(\int_{\mathcal{X}\times\mathcal{Z}} \varphi_1^*\left(\frac{1}{\epsilon}a(x,z;\bar{\theta})^T\bar{\beta}\right)\nabla_\theta a(x,z;\bar{\theta})^T\,\omega(\mathrm{d}x\otimes\mathrm{d}z)\right)\hat{\beta}
\end{aligned}$$

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 (B.13) is given by

$$\begin{aligned}
0 =& \frac{\partial \widehat{G}}{\partial \beta}(\theta_0, 0) + \left(\frac{\partial^2 \widehat{G}}{\partial \theta\, \partial \beta}(\dot{\theta}, \dot{\beta})\right)\left(\hat{\theta} - \theta_0\right) + \left(\frac{\partial^2 \widehat{G}}{\partial \beta\, \partial \beta}(\dot{\theta}, \dot{\beta})\right)\hat{\beta} \\
=& -\frac{1}{n}\sum_{i=1}^n b_i + \int_{\mathcal{X}\times\mathcal{Z}} a(x,z;\theta_0)\,\omega(\mathrm{d}x\otimes\mathrm{d}z) \\
&-\left(\int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon}\varphi_2^*\left(\frac{1}{\epsilon}a(x,z;\dot{\theta})^T\dot{\beta}\right)a(x,z;\dot{\theta})\left(\nabla_\theta a(x,z;\dot{\theta})^T\dot{\beta}\right)\omega(\mathrm{d}x\otimes\mathrm{d}z)\right)\left(\hat{\theta}-\theta_0\right) \\
&-\left(\int_{\mathcal{X}\times\mathcal{Z}} \varphi_1^*\left(\frac{1}{\epsilon}a(x,z;\dot{\theta})^T\dot{\beta}\right)\nabla_\theta a(x,z;\dot{\theta})\,\omega(\mathrm{d}x\otimes\mathrm{d}z)\right)\left(\hat{\theta}-\theta_0\right) \\
&-\left(\int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon}\varphi_2^*\left(\frac{1}{\epsilon}a_i(\dot{\theta})^T\dot{\beta}\right)a(x,z;\dot{\theta})a(x,z;\dot{\theta})^T\,\omega(\mathrm{d}x\otimes\mathrm{d}z) + R_{\lambda_n}\right)\hat{\beta}
\end{aligned}$$

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 B.18, 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 B.15 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 (B.12) reduces to

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

For the second condition note that by Lemma B.16

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

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

$$\begin{aligned} 0 =& \frac{1}{n}\sum_{i=1}^{n}\left(b_i - a(x_i, z_i; \theta_0)\right) + \int_{\mathcal{X}\times\mathcal{Z}} \nabla_\theta a(x, z; \dot{\theta})^T \, \omega(\mathrm{d}x \otimes \mathrm{d}z) \left(\hat{\theta} - \theta_0\right) \\ &+ \underbrace{\left(\int_{\mathcal{X}\times\mathcal{Z}} \frac{1}{\epsilon} a(x, z; \dot{\theta}) a(x, z; \dot{\theta})^T \, \omega(\mathrm{d}x \otimes \mathrm{d}z) + R_{\lambda_n}\right)}_{=:\Lambda_n(\dot{\theta})} \hat{\beta} + O_p(n^{-1}), \end{aligned}$$

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

$$\begin{aligned} \begin{pmatrix} 0 \\ 0 \end{pmatrix} =& \begin{pmatrix} 0 \\ \frac{1}{n}\sum_{i=1}^{n}\left(b_i - a(x_i, z_i; \theta_0)\right) \end{pmatrix} \\ &+ \underbrace{\begin{pmatrix} 0 & \int_{\mathcal{X}\times\mathcal{Z}} \nabla_\theta a(x, z; \bar{\theta})^T \, \omega(\mathrm{d}x \otimes \mathrm{d}z) \\ \int_{\mathcal{X}\times\mathcal{Z}} \nabla_\theta a(x, z; \dot{\theta})^T \, \omega(\mathrm{d}x \otimes \mathrm{d}z) & \Lambda_n(\dot{\theta}) \end{pmatrix}}_{:=M_n} \begin{pmatrix} \hat{\theta} - \theta_0 \\ \hat{\beta} - \beta_0 \end{pmatrix}, \end{aligned} \tag{B.14}$$

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 j) 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 j) with Lemma B.13 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 B.17 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,

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

is invertible, then it follows from standard blockmatrix algebra (see e.g. Bernstein [19]), that the inverse of M is given by

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

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

$$\begin{aligned}\Lambda =& \frac{1}{\epsilon}\begin{pmatrix} 1 & E[1 \otimes k((X,Z),\cdot)] & -E[1 \otimes \Psi(X,Z;\theta_0)] \\ E[k((X,Z),\cdot) \otimes 1] & E[k((X,Z),\cdot) \otimes k((X,Z),\cdot)] + I & -E[k((X,Z),\cdot) \otimes \Psi(X,Z;\theta_0)] \\ -E[\Psi(X,Z;\theta_0) \otimes 1] & -E[\Psi(X,Z;\theta_0) \otimes k((X,Z),\cdot)] & E[\Psi(X,Z;\theta_0) \otimes \Psi(X,Z;\theta_0)] \end{pmatrix} \\ =& \frac{1}{\epsilon}\begin{pmatrix} 1 & E[1 \otimes k((X,Z),\cdot)] & 0 \\ E[k((X,Z),\cdot) \otimes 1] & E[k((X,Z),\cdot) \otimes k((X,Z),\cdot)] + I & 0 \\ 0 & 0 & \Omega_0 \end{pmatrix}\end{aligned}$$

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 $\|E[k((X,Z),\cdot) \otimes \Psi(X,Z;\theta_0)]\|_{\mathcal{F}\times\mathcal{H}^*} \leq C\|E[\Psi(X,Z;\theta_0)]\|_{\mathcal{H}^*} = 0$. Now Λ is of blockdiagonal form and for the upper block B we have

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

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 B.1 Ω_0 is invertible and thus we can conclude

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

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 B.20 we have that Γ is non-singular and invertible which legitimates the inversion of M.

With this at hand, we can solve equation (B.14) for $\hat{\theta} - \theta_0$ and obtain

$$\begin{aligned}\sqrt{n}\left(\hat{\theta} - \theta_0\right) &= \left(\Gamma^{-1}\nabla_\theta\Psi_0\left(\Lambda^{-1}\right)_{3,3}\right)\sqrt{n}\hat{\Psi}(\theta_0) \\ &= -\left(\left(\nabla_\theta\Psi_0\left(\Lambda^{-1}\right)_{3,3}\nabla_\theta\Psi_0\right)^{-1}\nabla_\theta\Psi_0\left(\Lambda^{-1}\right)_{3,3}\right)\sqrt{n}\hat{\Psi}(\theta_0). \qquad \text{(B.15)}\end{aligned}$$

By the Donsker property of Ψ 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 (B.15) we get

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

with

$$\begin{aligned}\Xi &= \left(\left(\left(\nabla_\theta \Psi_0\right) \Omega_0^{-1} \left(\nabla_\theta \Psi_0\right)\right)^{-1} \left(\nabla_\theta \Psi_0\right) \Omega_0^{-1}\right) \Omega_0 \left(\left(\left(\nabla_\theta \Psi_0\right) \Omega_0^{-1} \left(\nabla_\theta \Psi_0\right)\right)^{-1} \left(\nabla_\theta \Psi_0\right) \Omega_0^{-1}\right)^T \\ &= \left(\left(\nabla_\theta \Psi_0\right) \Omega_0^{-1} \left(\nabla_\theta \Psi_0\right)\right)^{-1}.\end{aligned}$$

□

B.5.5 Asymptotic Properties of KMM for Finite-Dimensional Moment Restrictions

Proof of Theorem B.3 (Consistency for MR)

The consistency for the finite dimensional case follows as a special case of the consistency result for the functional case (Theorem B.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 B.1 we need finite dimensional versions of Lemmas B.15-B.19, 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 B.1 for details.

Lemma B.21. *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 B.13 by exchanging $\Psi(X, Z; \theta) \to \psi(X; \theta)$ and $\mathcal{H} \to \mathbb{R}^m$. □

Lemma B.22. *Let the assumptions of Theorem B.3 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 $n \to \infty$,*

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

Proof. The proof follows immediately from the one for Lemma B.15 by exchanging $\Psi(X, Z; \theta) \to \psi(X; \theta)$ and $\mathcal{H} \to \mathbb{R}^m$ and using Lemma B.21 instead of Lemma B.13 to bound $\sup_{\theta\in\Theta} \|\psi(X;\theta)\|_2 \leq C$ ω-a.s.. □

Lemma B.23. *Let the assumptions of Theorem B.3 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*

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

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 B.17 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.3. □

Lemma B.24. *Let the assumptions of Theorem B.3 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 B.18) with the usual substitutions. □

Lemma B.25. *Let the assumptions of Theorem B.3 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 B.19) with the usual substitutions. □

Lemma B.26. *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 $\operatorname{rank}\left(E[\nabla_\theta \psi(X;\theta_0)]\right) = p$ 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 $\operatorname{rank}(E[\nabla_\theta \psi(X;\theta_0)] = p$, its rows are linearly independent and thus for any $\theta \in \Theta$ with $0 < \|\theta\| < \infty$ there exists $j \in \{1,\dots,m\}$ 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.3

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 B.25 we also have $\|\hat{\psi}\| = O_p(n^{-1/2})$ and thus using the triangle inequality we get

$$\begin{aligned}\left\|E[\psi(X;\hat{\theta})]\right\|_2 &= \left\|E[\psi(\hat{\theta})] - \hat{\psi} + \hat{\psi}\right\|_2 \\ &\leq \left\|E[\psi(X;\hat{\theta})] - \hat{\psi}\right\|_2 + \left\|\hat{\psi}\right\|_2 \\ &= O_p(n^{-1/2}) \xrightarrow{p} 0.\end{aligned}$$

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

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

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

$$\begin{aligned}
\|E[\psi(X;\hat{\theta})]\|_2^2 &= \|\underbrace{E[\psi(X;\theta_0)]}_{=0} + (\hat{\theta}-\theta_0)^T E[\nabla_\theta \psi(X;\bar{\theta})]\|_2^2 \\
&= \left\langle (\hat{\theta}-\theta_0)^T E[\nabla_\theta \psi(X;\bar{\theta})], (\hat{\theta}-\theta_0)^T E[\nabla_\theta \psi(X;\bar{\theta})] \right\rangle \\
&= (\hat{\theta}-\theta_0)^T \underbrace{E[\nabla_\theta \psi(X;\bar{\theta})] E[\nabla_\theta \psi(X;\bar{\theta})]^T}_{=:\Sigma(\bar{\theta})} (\hat{\theta}-\theta_0) \\
&\geq \lambda_{\min}\left(\Sigma(\bar{\theta})\right) \|\hat{\theta}-\theta_0\|_2^2
\end{aligned}$$

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.3 and Lemma B.26 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 $\|\hat{\theta}-\theta_0\| = O_p(n^{-1/2})$. □

Proof of Theorem B.4 (Asymptotic Normality for MR)

Proof of Theorem B.4

Proof. The proof follows directly from the one for functional moment restrictions Theorem B.2 by identifying $\Psi(x,z;\theta) = \psi(x;\theta)$, $\mathcal{H} = \mathbb{R}^m$, setting $\lambda_n = 0$ and using Lemma B.26 to translate the rank condition, Assumption i), into non-singularity of $\Sigma_0 = E[\nabla_\theta \psi(X;\theta_0)] E[\nabla_\theta \psi(X;\theta_0)]^T \in \mathbb{R}^{p\times p}$. □

Appendix C

Geometry-Aware Instrumental Variable Regression

C.1 Experimental Details

Hyperparameters For SMM we choose the hyperparameters from the grid defined by $\epsilon \in [10^{-6}, 10^{-4}, 10^{-2}]$ and $\lambda/\epsilon \in [10^{-6}, 10^{-4}, 10^{-2}, 1.0]$. Note that as ϵ and γ_t only appear as ϵ/γ_t, we absorb the factor γ_t into ϵ and consider $\gamma_t = 1$ everywhere. For VMM we choose the hyperparameters from $\lambda \in [10^{-6}, 10^{-4}, 10^{-2}, 1.0]$ as done by the authors of the method [14]. We pick the best hyperparameter configuration by evaluating the MMR objective [179] on a validation data set of the same size as the training set. We visualize the dependency on the hyperparameters for the first experiment without random covariates in Figure C.1. We observe that the method is rather insensitive to the choice of ϵ but admits a stronger dependence on the choice of the regularization parameter λ.

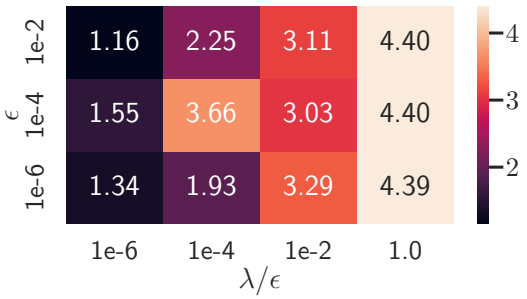


Fig. C.1 Kernel-SMM dependency on hyperparameters. We evaluate the SMM estimator on the first experiment without random covariates for different hyperparameter configurations. Values correspond to the mean of the prediction error $E[\|f(T;\hat{\theta}) - f(T;\theta_0)\|_2^2]$ averaged over models trained on 20 random training sets.

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

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

C.2 Additional Results

NetworkIV Here, we consider a common modern benchmark for IV regression in the standard setting without any data corruptions. Consider the following data generating process introduced by Bennett et al. [15] and subsequently used by many other works [179, 92, 93],

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

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

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

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

Neural Estimators We explore an alternative SMM implementation where we represent the instrument function $h \in \mathcal{H}$ as a neural network parameterized by $\omega \in \Omega$. With this choice, the estimator (4.8) takes the form

$$f^* = \arg\min_{f \in \mathcal{F}} \max_{\omega \in \Omega} E_{\hat{P}_n} \left[\left(I + \frac{\epsilon}{2} \Delta_\xi \right) \left(\psi(\cdot; f)^T h_\omega(\cdot) \right)(\xi) - \frac{\epsilon}{2} \| \nabla_\xi \left(\psi(\cdot; f)^T h_\omega(\cdot) \right)(\xi) \|_\Gamma^2 - \frac{\lambda}{2} \| h_\omega(Z) \|_2^2 \right]. \tag{C.1}$$

The Neural-SMM estimator can be trained in the same fashion as the DeepGMM [15] or FunctionalGEL [92] estimators by alternating mini-batch stochastic gradient descent steps in the the model parameters and the adversary parameters ω.

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

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

We benchmark the Neural-SMM estimator against DeepGMM [15] and FunctionalGEL [92] which achieved state-of-the-art results on several benchmarks including the NetworkIV experiment. For all methods we use the same instrument network architecture consisting of a feed-forward neural network with $[50, 20]$ hidden units and leaky ReLU activation functions. We optimize the objective by alternating steps with an optimistic Adam [46] optimizer with parameters $\beta = (0.5, 0.9)$. We tuned the learning rates, for the model and adversary by evaluating the DeepGMM estimator for different values and fix them both to $5e^{-4}$ for all methods. In the same way we fix the batch size to 200 and the number of epochs to 3000. For the FunctionalGEL estimator we use the Kullback-Leibler divergence version. For all methods we choose the regularization parameter λ from $[10^{-6}, 10^{-4}, 10^{-2}, 1.0]$ and for Neural-SMM we additionally choose ϵ from $[10^{-6}, 10^{-4}, 10^{-2}, 1.0]$ by using the MMR objective on a validation set of the same size as the training set.

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

Figure C.2 visualizes the dependence of Neural-SMM on its hyperparameters. We observe that for this experiment SMM requires either one or both parameters to be chosen large for optimal performance but the performance remains stable across a range of parameters.

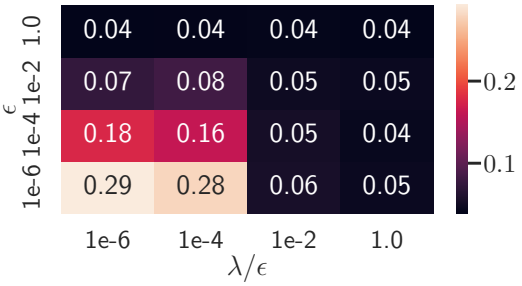


Fig. C.2 Neural-SMM dependency on hyperparameters. We evaluate the Neural-SMM estimator for different hyperparameter configurations exemplarily for the abs function in the network IV experiment. Values correspond to the mean of the prediction error $E[\|f(T;\hat{\theta}) - f(T;\theta_0)\|_2^2]$ averaged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

C.3 Proofs

C.3.1 Duality Results

Proof of Theorem 4.1

Proof. Introducing the Lagrange parameter $\rho \in \mathbb{R}$, the Lagrangian of (4.3) reads

$$L(P, \rho, f) = \min_{\pi \in \Pi(P, \hat{P}_n)} E_{(\xi,\xi')\sim\pi}\left[c(\xi, \xi') + \epsilon \log\left(\frac{\mathrm{d}\pi(\xi, \xi')}{\mathrm{d}\mu(\xi)\mathrm{d}\nu(\xi')}\right)\right] \tag{C.2}$$

$$+ \rho \sup_{h\in\mathcal{H}, \|h\|_{\mathcal{H}}=1} E_P[\Psi(\xi; f)(h)]. \tag{C.3}$$

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

$$L(P, h, f) = \min_{\pi \in \Pi(P, \hat{P}_n)} E_{(\xi,\xi')\sim\pi}\left[c(\xi, \xi') + \epsilon \log\left(\frac{\mathrm{d}\pi(\xi, \xi')}{\mathrm{d}\mu(\xi)\mathrm{d}\nu(\xi')}\right)\right] + E_P[\Psi(\xi; f)(h)]. \tag{C.4}$$

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

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

$$D(h, f) = \min_{\pi \in \Pi(\hat{P}_n)} E_{(\xi,\xi')\sim\pi}\left[c(\xi, \xi') + \epsilon \log\left(\frac{\mathrm{d}\pi(\xi, \xi')}{\mathrm{d}\mu(\xi)\mathrm{d}\nu(\xi')}\right)\right] + E_{\mathbb{P}_{1\#}\pi}[\Psi(\xi; f)(h)]. \tag{C.5}$$

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

$$E_{P_{1\#}\pi}[g(\xi)] = \int_\Xi g(\xi)(\mathbb{P}_{1\#}\pi)(\xi)\mathrm{d}\xi = \int_\Xi g(\xi)\sum_{i=1}^n \pi(\xi, \xi_i')\mathrm{d}\xi \tag{C.6}$$

$$= \int_\Xi g(\xi)\sum_{i=1}^n \pi(\xi|\xi_i')\hat{p}(\xi_i')\mathrm{d}\xi = \int_\Xi g(\xi)\frac{1}{n}\sum_{i=1}^n \pi(\xi|\xi_i')\mathrm{d}\xi \tag{C.7}$$

$$= E_{\xi'\sim\hat{P}_n}[E_{\xi\sim\pi|\xi'}[g(\xi)|\xi']]. \tag{C.8}$$

Therefore the optimization over $\pi \in \Pi(\hat{P}_n)$ is equivalent to a sequence of optimization problems over $\pi|\xi' \in \mathcal{P}(\Xi)$, one for each value of $\xi' \in \Xi$. With this we can express the dual problem (C.5) as

$$D(h,f) = E_{\xi' \sim \hat{P}_n}\left[\min_{\pi|\xi' \in \mathcal{P}(\Xi)} E_{\xi \sim \pi|\xi'}\left[c(\xi,\xi') + \epsilon \log\left(\frac{\mathrm{d}(\pi|\xi')(\xi)}{\mathrm{d}\mu(\xi)}\right) + \Psi(\xi;f)(h)\middle|\xi'\right]\right] \tag{C.9}$$

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

$$G(\xi';h,f) := \min_{\pi|\xi' \in \mathcal{P}(\Xi)} E_{\xi \sim \pi|\xi'}\left[c(\xi,\xi') + \epsilon \log\left(\frac{\mathrm{d}(\pi|\xi')(\xi)}{\mathrm{d}\mu(\xi)}\right) + \Psi(\xi;f)(h)\right]. \tag{C.10}$$

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

$$G(\xi';h,f) = \min_{r \in \mathcal{R}} E_{\xi \sim \mu}\left[r(\xi)c(\xi,\xi') + \epsilon r(\xi)\log\left(r(\xi)\right) + r(\xi)\Psi(\xi;f)(h)\right]. \tag{C.11}$$

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

$$G(\xi';h,f) = \sup_{\eta \in \mathbb{R}} \min_{r:\Xi \to \mathbb{R}_+} E_{\xi \sim \mu}[r(\xi)c(\xi,\xi') + \epsilon r(\xi)\log\left(r(\xi)\right) + r(\xi)\Psi(\xi;f)(h) + \eta(1-r(\xi))] \tag{C.12}$$

$$= \sup_{\eta \in \mathbb{R}} \eta - \epsilon E_{\xi \sim \mu}\left[\sup_{t \geq 0} t\frac{\eta - c(\xi,\xi') - \Psi(\xi;f)(h)}{\epsilon} - t\log t\right] \tag{C.13}$$

$$= \sup_{\eta \in \mathbb{R}} \eta - \epsilon E_{\xi \sim \mu}\left[\exp\left(\frac{\eta - c(\xi,\xi') - \Psi(\xi;f)(h)}{\epsilon} - 1\right)\right], \tag{C.14}$$

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

$$0 = 1 - e^{\eta/\epsilon - 1} E_{X \sim \mu}\left[\exp\left(\frac{-\Psi(\xi;f)(h) - c(\xi,\xi')}{\epsilon}\right)\right], \tag{C.15}$$

which yields

$$\eta = \epsilon - \epsilon \log E_{X \sim \mu}\left[\exp\left(\frac{-\Psi(\xi;f)(h) - c(\xi,\xi')}{\epsilon}\right)\right]. \tag{C.16}$$

Inserting back into (C.14), we obtain for each $\xi' \in \Xi$

$$G(\xi';h,f) = -\epsilon \log E_{\xi \sim \mu}\left[\exp\left(\frac{-\Psi(\xi;f)(h) - c(\xi,\xi')}{\epsilon}\right)\right]. \tag{C.17}$$

and the result follows by inserting into (C.9) and redefining $h/\epsilon \to h$. □

Proof of Theorem 4.2

Proof. Using the assumptions on the reference measure and cost function, we can write the objective in the form (4.6), where the inner expectation is given as

$$E_{\xi\sim\mathcal{N}(\xi',\epsilon\Gamma^{-1})}\left[e^{-\Psi(\xi;f)(h)}\right] = \int_{\Xi} e^{-\Psi(\xi;f)} e^{-\frac{1}{2\epsilon}\|\xi-\xi'\|^2_{\Gamma^{-1}}}\mathrm{d}\xi. \tag{C.18}$$

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

$$\begin{aligned}\Psi(\xi;f)(h) =&\Psi(\xi';f)(h) + (\xi-\xi')^T\nabla_\xi\Psi(\xi';f)(h) &\text{(C.19)}\\ &+\frac{1}{2}(\xi-\xi')^T\nabla^2_\xi\Psi(\xi';f)(h)(\xi-\xi') + O(\|\xi-\xi'\|^3). &\text{(C.20)}\end{aligned}$$

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

$$E_{\xi\sim\mathcal{N}(\xi',\gamma)}\left[e^{-\Psi(\xi;f)(h)}\right] = e^{-\Psi(\xi';f)(h)}\int_\Xi \exp\left(-\frac{1}{2\epsilon}\left(2\epsilon\delta^T G(\xi') + \epsilon\delta^T H(\xi')\delta + \delta^T\Gamma\delta\right)\right)\mathrm{d}\delta + O(\epsilon^{3/2}). \tag{C.21}$$

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

$$\begin{aligned}&E_{\xi\sim\mathcal{N}(\xi',\gamma)}\left[e^{-\Psi(\xi;f)(h)}\right] &\text{(C.22)}\\ =&e^{-\Psi(\xi';f)(h)}\int\frac{1}{\left|\det\Omega_\epsilon^{1/2}\right|}\times\exp\left(-\frac{1}{2\epsilon}\left(\omega^T\omega + 2\epsilon\omega^T\Omega_\epsilon^{-1/2}G(\xi')\right)\right)\mathrm{d}\omega + O(\epsilon^{3/2}). &\text{(C.23)}\end{aligned}$$

Now, completing the square we obtain

$$\begin{aligned}&E_{\xi\sim\mathcal{N}(\xi',\gamma)}\left[e^{-\Psi(\xi;f)(h)}\right] &\text{(C.24)}\\ =&e^{-\Psi(\xi';f)(h)}e^{\frac{\epsilon}{2}G(\xi')^T\Omega_\epsilon^{-1}G(\xi')}\int\frac{1}{\left|\det\Omega_\epsilon^{1/2}\right|}\exp\left(-\frac{1}{2\epsilon}\left(\omega+\epsilon\Omega_\epsilon^{-1/2}G(\xi')^2\right)\right)\mathrm{d}\omega + O(\epsilon^{3/2}) &\text{(C.25)}\\ =&\left(\frac{2\pi}{\epsilon}\right)^{d_\xi/2}\left|\det\Omega_\epsilon^{1/2}\right|^{-1}e^{-\Psi(\xi';f)(h)}e^{\frac{\epsilon}{2}G(\xi')^T\Omega_\epsilon^{-1}G(\xi')} + O(\epsilon^{3/2}). &\text{(C.26)}\end{aligned}$$

Finally inserting back into (4.6) we get

$$D(f,h) = E_{\xi' \sim \hat{P}_n}\left[-\epsilon \log E_{\xi \sim \mathcal{N}(\xi',\gamma)}\left[e^{\Psi(\xi;f)(h)}\right]\right] \tag{C.27}$$

$$= E_{\xi' \sim \hat{P}_n}\left[\epsilon\Psi(\xi';f)(h) - \frac{\epsilon^2}{2}G(\xi')^T\Omega_\epsilon^{-1}G(\xi') + \frac{\epsilon}{2}\log|\det\Omega_\epsilon|\right] - \frac{\epsilon d_\xi}{2}\log\frac{2\pi}{\epsilon} + O(\epsilon^{5/2}). \tag{C.28}$$

Dividing by ϵ and neglecting constant terms we get

$$D(f,h) = E_{\xi' \sim \hat{P}_n}\left[\Psi(\xi';f)(h) - \frac{\epsilon}{2}G(\xi')^T\Omega_\epsilon^{-1}G(\xi') + \frac{1}{2}\log|\det\Omega_\epsilon|\right] + O(\epsilon^{3/2}). \tag{C.29}$$

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

$$\Omega_\epsilon^{-1} = \left(\Gamma + \epsilon H(\xi')\right)^{-1} \tag{C.30}$$

$$= \Gamma^{-1}\left(I + \epsilon\Gamma^{-1}H\right)^{-1} \tag{C.31}$$

$$= \Gamma^{-1}\left(I - \epsilon\Gamma^{-1}H\right) + O(\epsilon^2) \tag{C.32}$$

$$= \Gamma^{-1} + O(\epsilon). \tag{C.33}$$

Similarly we have

$$\log|\det\Omega_\epsilon| = \log|\det\left(\Gamma + \epsilon H\right)| \tag{C.34}$$

$$= \log|\det\Gamma| + \log\left|\det\left(I + \epsilon\Gamma^{-1}H\right)\right| \tag{C.35}$$

$$= \underbrace{\left(\sum_{x\in\{t,y,z\}} d_x \log\gamma_x\right)}_{=:C} + \log\det\left(I + \epsilon\Gamma^{-1}H\right) \tag{C.36}$$

$$= C + \mathrm{Tr}\log\left(I + \epsilon\Gamma^{-1}H\right) \tag{C.37}$$

$$= C + \mathrm{Tr}\left(\epsilon\Gamma^{-1}H + O(\epsilon^2)\right) \tag{C.38}$$

$$= C + \epsilon\sum_{x\in\{t,y,z\}}\frac{1}{\gamma_x}\Delta_x\Psi(\xi';f)(h) + O(\epsilon^2). \tag{C.39}$$

So we finally obtain

$$D(f,h) = E_{\hat{P}_n}\left[\Psi(\xi;f)(h) - \frac{\epsilon}{2}\sum_{x\in\{t,y,z\}}\frac{1}{\gamma_x}\left(\|\nabla_x\Psi(\xi;f)(h)\|_2^2 - \Delta_x\Psi(\xi;f)(h)\right)\right] + O(\epsilon^{3/2}). \tag{C.40}$$

□

C.3.2 Proof of Theorem 4.4 (Consistency)

The objective of the SMM estimator (4.8) can be written as

$$\widehat{D}(h,\theta) = \left(I + \frac{\epsilon}{2}\Delta_\xi\right) E_{\hat{P}_n}[\Psi(\xi; f)(h)] - \frac{\epsilon}{2}\langle h, \widehat{\Omega}_{\lambda_n}(\bar{\theta}_n)h\rangle_{\mathcal{H}}, \tag{C.41}$$

where we defined the linear operator $\widehat{\Omega}_{\lambda_n}(\bar{\theta}_n) : \mathcal{H} \to \mathcal{H}$ as $\widehat{\Omega}_{\lambda_n}(\bar{\theta}_n) = E_{\hat{P}_n}\left[\left(\nabla_\xi \Psi(\xi;\bar{\theta}_n)\right)^T \Gamma^{-1} \nabla_\xi \Psi(\xi;\bar{\theta}_n)\right] + \lambda_n I \otimes I$. Our proof of Theorem 4.4 uses properties of the spectrum of $\widehat{\Omega}_{\lambda_n}(\bar{\theta}_n)$ which we will derive in the following.

Previous Results

Lemma C.1 (Corollary 9.31, Kosorok [90])**.** *Let $\mathcal{F}$ and $\mathcal{G}$ be Donsker classes of functions. Then $\mathcal{F} + \mathcal{G}$ is Donsker. Further if additionally $\mathcal{F}$ and $\mathcal{G}$ are uniformly bounded, then $\mathcal{F} \cdot \mathcal{G}$ is Donsker.*

Lemma C.2 (Lemma 18, Bennett and Kallus [14])**.** *Suppose that $\mathcal{G}$ is a class of functions of the form $g : \Xi \to \mathbb{R}$, and that $\mathcal{G}$ is P-Donsker in the sense of Kosorok [90]. Then we have*

$$\sup_{g\in\mathcal{G}} E_{\hat{P}_n}[g(\xi)] - E[g(\xi)] = O_p(n^{-1/2}). \tag{C.42}$$

Lemma C.3 (Lemma E.4, Kremer et al. [93])**.** *Let Assumptions 4.1-4.7 be satisfied. Then the matrix*

$$\Sigma(\theta_0) = \langle E[\nabla_\theta \Psi(\xi;\theta_0)], E[\nabla_{\theta^T}\Psi(\xi;\theta_0)]\rangle_{\mathcal{H}^*} \tag{C.43}$$

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

Spectrum of $\widehat{\Omega}$

Lemma C.4. *Let Assumptions 4.2 and 4.3 be satisfied. Then we have*

$$\sup_{\theta\in\Theta, x\in\mathcal{T}\times\mathcal{Y}} \|\psi(x;\theta)\|_\infty \le C_\psi < \infty \tag{C.44}$$

$$\sup_{\theta\in\Theta, x\in\mathcal{T}\times\mathcal{Y}} \|J_x(\psi)(x;\theta)\|_\infty \le L_\psi < \infty \tag{C.45}$$

$$\sup_{\theta\in\Theta, x\in\mathcal{T}\times\mathcal{Y}} \|\Delta_x\psi(x;\theta)\|_\infty \le D_\psi < \infty \tag{C.46}$$

$$\sup_{\theta\in\Theta, z\in\mathcal{Z}} \|h(z)\|_\infty \le C_h < \infty \tag{C.47}$$

$$\sup_{\theta\in\Theta, z\in\mathcal{Z}} \|J_z h(z)\|_\infty \le L_h < \infty \tag{C.48}$$

$$\sup_{\theta\in\Theta, z\in\mathcal{Z}} \|\Delta_z h(z)\|_\infty \le D_h < \infty, \tag{C.49}$$

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

Proof. The proof follows directly from the fact that a continuous function on a compact domain is bounded and both $\psi(\cdot;\theta)$ and h are C^∞-smooth by Assumptions 4.3 and 4.5. □

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

$$\Omega(\theta) = E\left[\left(\nabla_\xi \Psi(\xi;\theta)\right)^T \Gamma^{-1} \nabla_\xi \Psi(\xi;\theta)\right] \tag{C.50}$$

is non-singular.

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

$$\begin{aligned}
\langle h, \Omega(\theta)h\rangle_{\mathcal{H}} =& E[h(Z)^T J_x(\psi)(X;\theta)\Gamma^{-1}J_x(\psi)(X;\theta)h(Z)] && \text{(C.51)}\\
=& E[h(Z)^T E[J_x(\psi)(X;\theta)\Gamma^{-1}J_x(\psi)(X;\theta)|Z]h(Z)] && \text{(C.52)}\\
=& E[h(Z)^T V_0(Z;\theta)h(Z)] && \text{(C.53)}\\
=& CE[\|h(Z)\|_2^2] && \text{(C.54)}\\
=& C\|h\|^2_{L^2(\mathcal{H},P_0)} > 0 && \text{(C.55)}
\end{aligned}$$

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

Lemma C.6 (Spectrum of $\widehat{\Omega}$)**.** *Let the assumptions of Theorem 4.4 be satisfied. Then for $\bar{\theta} \in \Theta$ with $\bar{\theta}_n \to \bar{\theta}$, the empirical gradient covariance operator*

$$\widehat{\Omega}_{\lambda_n}(\bar{\theta}_n) = E_{\hat{P}_n}\left[\left(\nabla_\xi \Psi(\xi;\bar{\theta}_n)\right)^T \Gamma^{-1} \nabla_\xi \Psi(\xi;\bar{\theta}_n)\right] + \lambda_n I \otimes I \tag{C.56}$$

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

Proof. Let in the following $\widehat{\Omega}(\theta) = \widehat{\Omega}_{\lambda_n=0}(\theta)$. With Assumption 4.4 it follows from Lemma C.5 that the operator $\Omega(\bar{\theta}) := E\left[\left(\nabla_\xi \Psi(\xi;\bar{\theta})\right)^T \Gamma^{-1} \nabla_\xi \Psi(\xi;\bar{\theta})\right]$ is non-singular and thus its smallest eigenvalue bounded away from zero. In the following we show that $\widehat{\Omega}(\bar{\theta}_n) \xrightarrow{p} \Omega(\bar{\theta})$, where the convergence rate in operator norm is $O_p(n^{-\zeta})$. Therefore, by adding the identity operator with regularization parameter λ_n that goes to zero slower than $O_p(n^{-\zeta})$ we ensure that $\widehat{\Omega}_{\lambda_n}(\bar{\theta}_n)$ remains positive definite w.p.a.1. The derivation of this result follows the proof of Lemma 20 of Bennett and Kallus [14]. By the triangle inequality we have

$$\|\widehat{\Omega}(\bar{\theta}_n) - \Omega(\bar{\theta})\|_{\text{op}} \leq \|\widehat{\Omega}(\bar{\theta}_n) - \Omega(\bar{\theta}_n)\| + \|\Omega(\bar{\theta}_n) - \Omega(\bar{\theta})\|. \tag{C.57}$$

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

$\|h\|_{\mathcal{H}} \leq 1\}$ denote the unit ball in $\mathcal{H}$, then

$$\|\widehat{\Omega}(\bar{\theta}_n) - \Omega(\bar{\theta}_n)\| = \sup_{h,h' \in \mathcal{H}_1} \langle h', \widehat{\Omega}(\bar{\theta}_n) - \Omega(\bar{\theta}_n) h \rangle_{\mathcal{H}} \tag{C.58}$$

$$= \sup_{h,h' \in \mathcal{H}_1} \Bigg\{ E_{\hat{P}_n} \left[h(Z)^T J_\psi(X;\bar{\theta}_n) \Gamma_x^{-1} J_\psi(X;\bar{\theta}_n)^T h'(Z) \right] \tag{C.59}$$

$$- E \left[h(Z)^T J_\psi(X;\bar{\theta}_n) \Gamma_x^{-1} J_\psi(X;\bar{\theta}_n)^T h'(Z) \right] \tag{C.60}$$

$$+ \frac{1}{\gamma_z} E_{\hat{P}_n} \left[\psi(X;\bar{\theta}_n)^T J_h(Z) J_{h'}(Z)^T \psi(X;\bar{\theta}_n) \right] \tag{C.61}$$

$$- \frac{1}{\gamma_z} E \left[\psi(X;\bar{\theta}_n)^T J_h(Z) J_{h'}(Z)^T \psi(X;\bar{\theta}_n) \right] \Bigg\} \tag{C.62}$$

$$\leq \sup_{g \in \mathcal{G}^2} \left\{ E_{\hat{P}_n}[g(\xi)] - E[g(\xi)] \right\} + \frac{1}{\gamma_z} \sup_{s \in \mathcal{S}^2} \left\{ E_{\hat{P}_n}[s(\xi)] - E[s(\xi)] \right\} \tag{C.63}$$

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

$$\mathcal{G}_i = \{ g_i : g_i(\xi) = \sum_{j=1}^{m} h_j(z) \left(J_\psi(x;\theta) \right)_{ji} \Gamma_{ii}^{-1/2},\ h \in \mathcal{H}_{i,1},\ \theta \in \Theta \} \tag{C.64}$$

$$\mathcal{G}^2 = \{ g : g(\xi) = \sum_{i \in [d_x]} g_i(\xi) g_i'(\xi),\ g_i, g_i' \in \mathcal{G}_i \} \tag{C.65}$$

$$\mathcal{S}_i = \{ s_i : s_i(\xi) = \sum_{j=1}^{m} \psi_j(x;\theta) \left(J_h(z) \right)_{ji},\ h \in \mathcal{H}_{i,1},\ \theta \in \Theta \} \tag{C.66}$$

$$\mathcal{S}^2 = \{ s_i : s_i(\xi) = \sum_{i \in [d_z]} s_i(\xi) s_i'(\xi),\ s_i, s_i' \in \mathcal{S}_i \} \tag{C.67}$$

Now for the first term, we have that each $h_j \in \mathcal{H}_{i,1}$ is P_0-Donsker by Assumption 4.5 and uniformly bounded by Lemma C.4. Similarly each entry of the Jacobian $J_\psi(\cdot;\theta)$ is P_0-Donsker by Assumption 4.3 and uniformly bounded by Lemma C.4. With that we can employ Lemma C.1 to conclude that $\mathcal{G}_i$ is P_0-Donsker and thus using Lemma C.1 again it follows that $\mathcal{G}^2$ is P_0-Donsker. Therefore we can use Lemma C.2 to obtain $\sup_{g \in \mathcal{G}^2} \left\{ E_{\hat{P}_n}[g(\xi)] - E[g(\xi)] \right\} = O_p(n^{-1/2})$.

For the second term in (C.63) we have that each $\psi_j(\cdot;\theta)$ is P_0-Donsker by Assumption 4.3 and uniformly bounded by Lemma C.4. Similarly each entry of the Jacobian $J_z h$ is P_0-Donsker by Assumption 4.5 and uniformly bounded by Lemma C.4. With that, again, we can employ Lemma C.1 to conclude that $\mathcal{S}_i$ is P_0-Donsker and thus using Lemma C.1 again it follows that $\mathcal{S}^2$ is P_0-Donsker. Therefore we can use Lemma C.2 to obtain $\frac{1}{\gamma_z} \sup_{s \in \mathcal{S}^2} \left\{ E_{\hat{P}_n}[s(\xi)] - E[s(\xi)] \right\} = O_p(n^{-1/2})$.

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

For the second term in (C.57) we have

$$\|\Omega(\bar{\theta}_n) - \Omega(\bar{\theta})\| = \sup_{h,h' \in \mathcal{H}_1} \langle h', \Omega(\bar{\theta}_n) - \Omega(\bar{\theta}) h \rangle_{\mathcal{H}} \leq C_x + \frac{1}{\gamma_z} C_z \tag{C.68}$$

where

$$C_x = \sup_{h,h' \in \mathcal{H}_1} E\Big[h'(Z)^T \Big(J_\psi(X;\bar{\theta}_n)\Gamma_x^{-1} J_\psi(X;\bar{\theta}_n)^T \tag{C.69}$$

$$- J_\psi(X;\bar{\theta})\Gamma_x^{-1} J_\psi(X;\bar{\theta})^T \Big) h(Z)\Big] \tag{C.70}$$

$$= \sup_{h,h' \in \mathcal{H}_1} E\Big[h'(Z)^T \Big(J_\psi(X;\bar{\theta}_n)\Gamma_x^{-1} J_\psi(X;\bar{\theta}_n) - J_\psi(X;\bar{\theta})\Gamma_x^{-1} J_\psi(X;\bar{\theta}) \Big) h(Z)\Big] \tag{C.71}$$

$$= \sup_{h,h' \in H_1} E\Big[h'(Z)^T J_\psi(X;\bar{\theta}_n)\Gamma_x^{-1} \left(J_\psi(X;\bar{\theta}_n) - J_\psi(X;\bar{\theta}) \right)^T h(Z) \tag{C.72}$$

$$+ h'(Z)^T J_\psi(X;\bar{\theta})\Gamma_x^{-1} \left(J_\psi(X;\bar{\theta}_n) - J_\psi(X;\bar{\theta}) \right)^T h(Z)\Big] \tag{C.73}$$

$$\leq \frac{2}{\min(\{\gamma_t, \gamma_y\})} m^2 C_h^2 L_\psi E\left[\|J_\psi(X;\bar{\theta}_n) - J_\psi(X;\bar{\theta})\|_\infty\right] \tag{C.74}$$

$$= O_p(n^{-\zeta}) \tag{C.75}$$

where we used that by Lemma C.4, $\sup_{\theta\in\Theta, x\in\mathcal{T}\times\mathcal{Y}} \|J_\psi(x;\theta)\|_\infty \leq L_\psi$ and $\sup_{h\in\mathcal{H}_1, z\in\mathcal{Z}} |h(z)| \leq C_h$ as well as by Assumption 4.6 $E\left[\|J_\psi(X;\bar{\theta}_n) - J_\psi(X;\bar{\theta})\|_\infty\right] = O_p(n^{-\zeta})$.

Now similarly for the second term in (C.68) we have

$$C_z = \sup_{h,h' \in \mathcal{H}_1} E\Big[\operatorname{Tr}\left(J_{h'}(Z)^T \psi(X;\bar{\theta}_n)\psi(X;\bar{\theta}_n)^T J_h(Z) \right) \tag{C.76}$$

$$- \operatorname{Tr}\left(J_{h'}(Z)^T \psi(X;\bar{\theta})\psi(X;\bar{\theta})^T J_h(Z) \right) \Big] \tag{C.77}$$

$$\leq L_h^2 E[\mathbf{1}^T \left(\psi(X;\bar{\theta}_n)\psi(X;\bar{\theta}_n)^T - \psi(X;\bar{\theta})\psi(X;\bar{\theta})^T \right) \mathbf{1}] \tag{C.78}$$

$$= L_h^2 E\left[\mathbf{1}^T \psi(X;\bar{\theta}_n) \left(\psi(X;\bar{\theta}_n) - \psi(X;\bar{\theta}) \right)^T \mathbf{1}\right] \tag{C.79}$$

$$+ L_h^2 E\left[\mathbf{1}^T \psi(X;\bar{\theta}) \left(\psi(X;\bar{\theta}_n) - \psi(X;\bar{\theta}) \right)^T \mathbf{1}\right] \tag{C.80}$$

$$\leq 2m^2 L_h^2 C_\psi E\left[\|\psi(X;\bar{\theta}_n) - \psi(X;\bar{\theta})\|_\infty\right] \tag{C.81}$$

$$= O_p(n^{-\zeta}), \tag{C.82}$$

where again we used Lemma C.4 and Assumption 4.6. Combining both results we obtain $\|\Omega(\bar{\theta}_n) - \Omega(\bar{\theta})\| = \leq C_x + \frac{1}{\gamma_z} C_z \leq O_p(n^{-\zeta})$.

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

$$\|\widehat{\Omega}(\bar{\theta}_n) - \Omega(\bar{\theta})\| \leq \|\widehat{\Omega}(\bar{\theta}_n) - \Omega(\bar{\theta}_n)\| + \|\Omega(\bar{\theta}_n) - \Omega(\bar{\theta})\| \tag{C.83}$$

$$\leq O_p(n^{-1/2}) + O_p(n^{-\zeta}) = O_p(n^{-\zeta}). \tag{C.84}$$

In conclusion we have shown that $\widehat{\Omega}(\bar{\theta}_n)$ converges to the non-singular operator $\Omega(\bar{\theta})$ at rate $O_p(n^{-\zeta})$ and by Assumption 4.6 we have $\lambda_n = O_p(n^{-\rho})$ with $0 < \rho < \zeta$, therefore the operator $\widehat{\Omega}_{\lambda_n}(\bar{\theta}_n) = \widehat{\Omega}(\bar{\theta}_n) + \lambda_n I$ is non-singular with smallest eigenvalue bounded away from zero w.p.a.1.

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

$$\begin{aligned}\langle h, \widehat{\Omega}(\bar{\theta}_n)h\rangle &= E_{\hat{P}_n}[h(Z)^T J_\psi(X;\bar{\theta}_n) J_\psi(X;\bar{\theta}_n)^T h(Z)] && \text{(C.85)}\\ &\leq E[J_\psi(X;\bar{\theta}_n)\|_\infty^2] E[\|h(Z)\|_\infty^2] && \text{(C.86)}\\ &\leq L_\psi^2 C_h^2 < \infty. && \text{(C.87)}\end{aligned}$$

□

Proof of Theorem 4.4

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

$$\|E_{\hat{P}_n}[\Psi(\xi;\theta)] - E[\Psi(\xi;\theta)]\|_{\mathcal{H}^*} = O_p(n^{-1/2}). \tag{C.88}$$

Proof.

$$\begin{aligned}\|E_{\hat{P}_n}[\Psi(\xi;\theta)] - E[\Psi(\xi;\theta)]\|_{\mathcal{H}^*} &= \sup_{h\in\mathcal{H}_1} E_{\hat{P}_n}[\psi(X;\theta)^T h(Z)] - E[\psi(X;\theta)^T h(Z)] && \text{(C.89)}\\ &= \sup_{g\in\mathcal{G}} E_{\hat{P}_n}[g(\xi)] - E[g(\xi)] && \text{(C.90)}\end{aligned}$$

where

$$\mathcal{G} = \{g : g(\xi) = \sum_{i=1}^m \psi_i(x;\theta) h_i(z),\ h_i \in \mathcal{H}_{i,1}, \theta \in \Theta\}. \tag{C.91}$$

Now as each h_i and $\psi_i(\cdot;\theta)$ are P_0-Donsker by Assumption 4.5 and 4.3 respectively and uniformly bounded by Lemma C.4, we can employ Lemma C.1 to conclude that $\mathcal{G}$ is P_0-Donsker. From this, the result follows by application of Lemma C.2. □

Lemma C.8 (Convergence of $\widehat{D}$)**.** *Let the assumptions of Theorem 4.4 be satisfied. Additionally let $\tilde{\theta} \in \Theta$ be a consistent estimator for θ_0, i.e., $\tilde{\theta} \xrightarrow{p} \theta_0$ with $\|E_{\hat{P}_n}[\Psi(\xi;\tilde{\theta})]\|_{\mathcal{H}^*} = O_p\left(n^{-1/2}\right)$. Then for $\tilde{h} = \arg\max_{h\in\mathcal{H}} D(\tilde{\theta}, h)$ we have $\|\tilde{h}\|_{\mathcal{H}} = O_p\left(n^{-1/2}\right)$ and $\widehat{D}(\tilde{\theta}, \tilde{h}) \leq O_p\left(n^{-1}\right)$.*

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

$$
\begin{aligned}
0 &= \widehat{D}(\tilde{\theta}, 0) && \text{(C.92)}\\
&\leq \arg\max_{h\in\mathcal{H}} D(\tilde{\theta}, \tilde{h}) && \text{(C.93)}\\
&= \left(I + \frac{\epsilon}{2}\Delta_\xi\right)\tilde{\Psi}(\tilde{h}) - \frac{\epsilon}{2}\langle \tilde{h}, \widehat{\Omega}_{\lambda_n}(\bar{\theta}_n)\tilde{h}\rangle_{\mathcal{H}} && \text{(C.94)}\\
&\leq \|I + \frac{\epsilon}{2}\Delta_\xi\|_{\mathrm{op}}\|\tilde{\Psi}\|_{\mathcal{H}^*}\|\tilde{h}\|_{\mathcal{H}} - \frac{\epsilon}{2}\lambda_{\min}\left(\widehat{\Omega}_{\lambda_n}(\bar{\theta}_n)\right)\left\|\tilde{h}\right\|_{\mathcal{H}}^2 && \text{(C.95)}\\
&\leq \left(1 + \frac{\epsilon}{2}\|\Delta_\xi\|\right)\|\tilde{\Psi}\|_{\mathcal{H}^*}\|\tilde{h}\|_{\mathcal{H}} - \frac{\epsilon}{2}\lambda_{\min}\left(\widehat{\Omega}_{\lambda_n}(\bar{\theta}_n)\right)\left\|\tilde{h}\right\|_{\mathcal{H}}^2 && \text{(C.96)}
\end{aligned}
$$

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

Lemma C.9 (Convergence of $\|\hat{\Psi}\|_{\mathcal{H}^*}$)**.** *Let the assumptions of Theorem 4.4 be satisfied. Let* $\hat{\theta} = \arg\min_{\theta\in\Theta}\sup_{h\in\mathcal{H}}\widehat{D}(\theta, h)$ *denote the SMM estimator for* θ_0*. Then* $\left\|E_{\hat{P}_n}[\Psi(\xi;\hat{\theta})]\right\|_{\mathcal{H}^*} = O_p(n^{-1/2})$.

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

$$
\begin{aligned}
\widehat{D}(\hat{\theta}, h_{\hat{\Psi}}) &= \left(I + \frac{\epsilon}{2}\Delta_\xi\right)\hat{\Psi}(h_{\hat{\Psi}}) - \frac{\epsilon}{2}\langle h_{\hat{\Psi}}, \widehat{\Omega}(\bar{\theta}_n)h_{\hat{\Psi}}\rangle_{\mathcal{H}} && \text{(C.97)}\\
&\geq \left(1 + \frac{\epsilon}{2}\lambda_{\min}(\Delta_\xi)\right)\hat{\Psi}(h_{\hat{\Psi}}) - \frac{\epsilon}{2}C\|h_{\hat{\Psi}}\|_{\mathcal{H}}^2 && \text{(C.98)}\\
&\geq C'\sigma_n\|\hat{\Psi}\|_{\mathcal{H}^*}^2 - \frac{C\epsilon}{2}\sigma_n^2\|\hat{\Psi}\|_{\mathcal{H}^*}^2, && \text{(C.99)}
\end{aligned}
$$

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

$$
C'\sigma_n\|\hat{\Psi}\|_{\mathcal{H}^*}^2 - \frac{C\epsilon}{2}\sigma_n^2\|\hat{\Psi}\|_{\mathcal{H}^*}^2 \leq \widehat{D}(\hat{\theta}, h_{\hat{\Psi}}) \leq \widehat{D}(\hat{\theta}, \hat{h}) \leq \max_{h\in\mathcal{H}}\widehat{D}(\theta_0, h) \leq O(n^{-1}), \qquad \text{(C.100)}
$$

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

Proof of Theorem 4.4 Using the result of Lemma C.9 for the convergence rate of the empirical moment functional, the proof of the consistency of our SMM estimator is identical to the ones

provided by Kremer et al. [92] and Kremer et al. [93] for their estimtators. We provide it here for completeness.

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

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

As by Assumption 4.1, θ_0 is the unique parameter for which $E[\psi(T,Y;\theta)|Z] = 0$ P_z-a.s. and by Assumption 4.5 this is fulfilled if and only if $\|E[\psi(\xi;\theta)]\|_{\mathcal{H}^*} = 0$, it follows that $\hat{\theta} \xrightarrow{p} \theta_0$.

Under the additional Assumption 4.7 we can use this result to translate the convergence rate of the moment functional to a convergence rate of the estimator $\hat{\theta}$.

As $\Psi(\xi;\theta)$ is continuously differentiable in its second argument which follows immediately from Assumption 4.7 and the definition of Ψ, we can use the mean value theorem to expand $\Psi(\xi,\hat{\theta})$ about θ_0, i.e., there exists $\bar{\theta} \in \operatorname{conv}(\{\theta_0,\hat{\theta}\})$ such that

$$\Psi(\xi;\hat{\theta}) = \Psi(\xi;\theta_0) + (\hat{\theta}-\theta_0)^T \nabla_\theta \Psi(\xi;\bar{\theta}). \tag{C.101}$$

Using this we have

$$\begin{aligned}\|E[\Psi(\xi;\hat{\theta})]\|^2_{\mathcal{H}^*} &= \|\underbrace{E[\Psi(\xi;\theta_0)]}_{=0} + (\hat{\theta}-\theta_0)^T E[\nabla_\theta \Psi(\xi;\bar{\theta})]\|^2_{\mathcal{H}^*} && \text{(C.102)}\\ &= \left\langle (\hat{\theta}-\theta_0)^T E[\nabla_\theta \Psi(\xi;\bar{\theta})], (\hat{\theta}-\theta_0)^T E[\nabla_\theta \Psi(\xi;\bar{\theta})]\right\rangle_{\mathcal{H}^*} && \text{(C.103)}\\ &= (\hat{\theta}-\theta_0)^T \underbrace{\left\langle E[\nabla_\theta \Psi(\xi;\bar{\theta})], E[\nabla_{\theta^T} \Psi(\xi;\bar{\theta})]\right\rangle_{\mathcal{H}^*}}_{=:\Sigma(\bar{\theta})} (\hat{\theta}-\theta_0) && \text{(C.104)}\\ &\geq \lambda_{\min}\left(\Sigma(\bar{\theta})\right) \|\hat{\theta}-\theta_0\|_2^2. && \text{(C.105)}\end{aligned}$$

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 Lemma C.3, thus the smallest eigenvalue of $\Sigma(\bar{\theta})$, $\lambda_{\min}(\Sigma(\bar{\theta}))$, is positive and bounded away from zero w.p.a.1. Finally as $\|E[\Psi(X,Z;\hat{\theta})]\| = O_p(n^{-1/2})$ taking the square-root on both sides we have $\|\hat{\theta}-\theta_0\| = O_p(n^{-1/2})$. □

Proof of Proposition 4.5

Proof. For a universal ISPD kernel, equivalence of the conditional and the variational moment restrictions (4.1) and (4.2) follows by Theorem 3.9 of Kremer et al. [92]. The Donsker property of the

unit ball in an RKHS of a smooth universal kernel with compact domain follows from Lemma 17 of Bennett and Kallus [14]. Finally, the Donsker property of the Jacobian $J_z h$ of h follows by the same argument as Lemma 17 of Bennett et al. [15] using C^∞ smoothness of h and boundedness of $J_z h$. □

Proof of Theorem 4.6

Proof. Under the assumptions the Sinkhorn profile is given as

$$R_\lambda(f) = \sup_{h\in\mathcal{H}} \Bigg\{ E_{\hat{P}_n}\left[h(Z)^T\left(I + \frac{\epsilon}{2}\Delta_x\right)\psi(X;f)\right] \tag{C.106}$$

$$- \epsilon E_{\hat{P}_n}\left[h(Z)^T J_\psi(X;\tilde{f})\Gamma_x^{-1} J_\psi(X;\tilde{f})^T h(Z)\right] - \frac{\lambda}{2}\|h\|_{\mathcal{H}}^2 \Bigg\} \tag{C.107}$$

which as the unconstrained maximization of a concave objective is a convex optimization problem. Moreover, the conditions of the classical representer theorem [142] are fulfilled and thus the maximizer of (C.107) is given as $h_l = \sum_{i=1}^n \alpha_i^l k_l(z_i,\cdot)$ with $\alpha^l \in \mathbb{R}^n$. Inserting this into (C.107) and defining the kernel Gram matrices $K_l \in \mathbb{R}^{n\times n}$ with entries $(K_l)_{ij} = k_l(z_i,z_j)$ we obtain

$$R_\lambda(f) = \sup_{\alpha\in\mathbb{R}^{nm}} \frac{1}{n}\sum_{i,j=1}^n\sum_{l=1}^m \alpha_i^l (K_l)_{ij}\left(I + \frac{\epsilon}{2}\Delta_x\right)\psi_l(x_j;f) - \frac{\lambda}{2}\sum_{l=1}^m (\alpha^l)^T K_l(\alpha^l) \tag{C.108}$$

$$- \frac{\epsilon}{2n}\sum_{i,j,k=1}^n\sum_{l,r=1}^m \alpha_i^l (K_l)_{ij}\nabla_x\psi_l(x_j;\tilde{f})^T\Gamma_x^{-1}\nabla_x\psi_r(x_j;\tilde{f})(K_r)_{jk}\alpha_r^k \tag{C.109}$$

$$= \sup_{\alpha\in\mathbb{R}^{nm}} \frac{1}{n}\alpha^T L\psi_\Delta - \frac{1}{2}\alpha^T\left(\epsilon Q(\tilde{f}) + \lambda L\right)\alpha \tag{C.110}$$

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

$$\psi_\Delta(f)_{i\cdot l} = \left(I + \frac{\epsilon}{2}\Delta_x\right)\psi_l(x_i;f) \tag{C.111}$$

$$L_{(i\cdot l),(j\cdot r)} = \delta_{lr} k_l(z_i,z_j) \tag{C.112}$$

$$Q(f)_{(i\cdot l),(j\cdot r)} = \frac{1}{n}\sum_{k=1}^n\sum_{s=1}^{d_x} k_l(z_i,z_k)\nabla_{x_s}\psi_l(x_k;f)(\Gamma_x^{-1})_{ss}\nabla_{x_s}\psi_r(x_k;f)k_r(z_k,z_j). \tag{C.113}$$

The first order optimality conditions for α read

$$0 = \frac{1}{n}L\psi_\Delta(f) - \left(\epsilon Q(\tilde{f}) + \lambda L\right)\alpha, \tag{C.114}$$

which immediately gives

$$\alpha = \left(\epsilon Q(\tilde{f}) + \lambda L\right)^{-1}\frac{1}{n}L\psi_\Delta(f). \tag{C.115}$$

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

$$R_\lambda(f) = \frac{1}{2n^2}\psi_\Delta(f)^T L \left(Q(\tilde{f}) + \frac{\lambda}{\epsilon} L \right)^{-1} L\psi_\Delta(f). \tag{C.116}$$

□