A Measure-Theoretic Axiomatisation of Causality and Kernel Regression

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Junhyung Park
aus Seoul/Südkorea

Tübingen
2024

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

Tag der mündlichen Qualifikation:	22.07.2024
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Bernhard Schölkopf
2. Berichterstatter/-in:	Prof. Dr. Bob Williamson

A Measure-Theoretic Axiomatisation of Causality and Kernel Regression

Junhyung Park

2024

Abstract

This thesis is composed of two broad strands of research. The first part of the thesis will discuss *causality*, with focus on a novel, measure-theoretic axiomatisation thereof, and the second part of the thesis will tackle some problems in *regression*, with focus on kernel methods and infinite-dimensional output spaces. Even though the two topics are very distinct in nature, we tackle them through a shared principle that places emphasis on *theory*.

Causality is a topic that has recently garnered much interest among the *artificial* intelligence research community, but it has always been a centrepiece of *human* intelligence. Humans have always perceived that, in addition to *observing* how events unfold around them, they can also make *interventions* on the world that potentially change the course of events. In other words, interventions on the world (not necessarily by the observers themselves) can *cause* events to occur, change their chances of occurring, or prevent them from occurring. This notion of intervention is viewed by many as the essence behind the concept of causality, and this is the view that we take in this thesis.

Mathematical modelling aims to describe the world in an abstract way, using mathematical concepts and language. Therefore, to describe the world with causality in mind, such that interventions can be modelled, the development of an axiomatic mathematical framework that can encode such information is a necessity. In this thesis, we take the view that such an axiomatic framework has been developed and established for the concept of *uncertainty* (or *randomness*, or *stochasticity*), namely *probability theory*, but we argue that, despite many competing propositions, most notably the *structural causal models* and the *potential outcomes* frameworks, a universally agreed, axiomatic framework that plays the role of probability spaces in the study of uncertainty does not yet exist for the study of causality. It is clear that, since interventions on the world do not, in general, cause the world to behave in a deterministic way, but there is ensuing uncertainty following most interventions as to how events will subsequently unfold, probability theory will play a fundamental role in any theory of causality. Based on this standpoint, we propose an axiomatic framework of causality, called *causal spaces*, that is built directly on probability spaces.

The second part of the thesis will be concerned with several aspects of *kernel regression*. Regression is a concept that is ubiquitous in statistics and machine learning, and has an endless list of applications in a wide range of domains, and regression techniques based on kernels have been some of the most popular and influential in statistics and machine learning research. It is natural, then, that it has also received much attention from theoreticians regarding its properties, guarantees and limitations. In this thesis, we make modest contributions to several aspects of them. First, we discuss *kernel conditional mean embeddings*, which have been known to researchers for over a decade. Our contribution lies in the fact that we view them as Bochner conditional expectations, as opposed to operators between reproducing kernel Hilbert spaces (RKHSs) as had been prevalently done in the literature, and hence, their estimation is precisely a regression problem in which the output space is an RKHS. The hypothesis space in which this regression is carried out is itself a (vector-valued) RKHS, and such a technique is widely known as *kernel ridge regression.*

Next, we propose a particular form of kernel ridge regression called *U-statistic regression*, and apply this and the previously studied conditional mean embeddings to the study of *conditional distributional treatment effect* in the potential outcomes framework, which is widely used in the domains of medicine or social sciences. The thesis then takes a more theoretical turn to study learning-theoretic and empirical process-theoretic aspects of regression with infinite-dimensional output spaces, which can naturally occur if the outputs are themselves functions, and of which kernel conditional mean embeddings are a particular case. We extend the existing theory of *empirical processes*, an indispensable tool in statistical learning theory but that was previously only developed for classes of real-valued functions, to take into account classes of (possibly infinite-dimensional) vector-valued functions; in particular, we propose bounds on the metric entropy of classes of smooth vector-valued functions. We also take a look at the special case of vector-valued kernel ridge regression and prove a consistency result, based not on empirical process theory, but on the powerful integral operator techniques that are popular in the analysis of kernel ridge regression.

Zusammenfassung

Diese Arbeit besteht aus zwei großen Forschungssträngen. Der erste Teil der Arbeit befasst sich mit der Kausalität, wobei der Schwerpunkt auf einer neuartigen maßtheoretischen Axiomatisierung liegt, und der zweite Teil der Arbeit befasst sich mit einigen Problemen der Regression, wobei der Schwerpunkt auf Kernel-Methoden und unendlich-dimensionalen Ausgangsräumen liegt. Obwohl die beiden Themen sehr unterschiedlicher Natur sind, gehen wir sie nach einem gemeinsamen Prinzip an, das den Schwerpunkt auf die Theorie legt.

Kausalität ist ein Thema, das in jüngster Zeit in der Forschungsgemeinschaft der künstlichen Intelligenz auf großes Interesse gestoßen ist, das aber schon immer ein Kernstück der menschlichen Intelligenz war. Menschen haben schon immer erkannt, dass sie nicht nur beobachten können, wie sich die Ereignisse um sie herum entfalten, sondern dass sie auch in die Welt eingreifen können, um den Verlauf der Ereignisse zu verändern. Mit anderen Worten, Eingriffe in die Welt (nicht notwendigerweise durch den Beobachter selbst) können das Eintreten von Ereignissen bewirken, die Wahrscheinlichkeit ihres Eintretens verändern oder sie verhindern. Dieser Begriff des Eingreifens wird von vielen als der Kern des Kausalitätskonzepts angesehen, und dies ist auch die Auffassung, die wir in dieser Arbeit vertreten.

Die mathematische Modellierung zielt darauf ab, die Welt auf abstrakte Weise mit Hilfe mathematischer Konzepte und Sprache zu beschreiben. Um die Welt unter Berücksichtigung der Kausalität so zu beschreiben, dass Interventionen modelliert werden können, ist die Entwicklung eines axiomatischen mathematischen Rahmens, der solche Informationen kodieren kann, eine Notwendigkeit. In dieser Arbeit vertreten wir die Auffassung, dass ein solcher axiomatischer Rahmen für das Konzept der Ungewissheit (oder des Zufalls oder der Stochastik) entwickelt und etabliert wurde, nämlich die Wahrscheinlichkeitstheorie. Wir argumentieren jedoch, dass es trotz vieler konkurrierender Vorschläge, insbesondere der strukturellen Kausalmodelle und der Rahmen für potenzielle Ergebnisse, noch keinen allgemein anerkannten axiomatischen Rahmen für die Untersuchung der Kausalität gibt, der die Rolle der Wahrscheinlichkeitsräume bei der Untersuchung der Ungewissheit übernimmt. Da Eingriffe in die Welt im Allgemeinen nicht dazu führen, dass sich die Welt auf deterministische Weise verhält, sondern nach den meisten Eingriffen Ungewissheit darüber besteht, wie sich die Ereignisse in der Folge entfalten werden, spielt die Wahrscheinlichkeitstheorie eine grundlegende Rolle in jeder Theorie der Kausalität. Ausgehend von diesem Standpunkt schlagen wir einen axiomatischen Rahmen der Kausalität vor, der als Kausalräume bezeichnet wird und direkt auf Wahrscheinlichkeitsräumen aufbaut.

Der zweite Teil der Arbeit befasst sich mit verschiedenen Aspekten der Kernelregression. Regression ist ein Konzept, das in der Statistik und im maschinellen Lernen allgegenwärtig ist und eine endlose Liste von Anwendungen in einer Vielzahl von Bereichen hat, und Regressionstechniken, die auf Kerneln basieren, gehören zu den beliebtesten und einflussreichsten in der Statistik- und maschinellen Lernforschung. Es ist daher nur natürlich, dass sie auch von Theoretikern hinsichtlich ihrer Eigenschaften, Garantien und Grenzen viel Aufmerksamkeit erhalten haben. In dieser Arbeit leisten wir einen bescheidenen Beitrag zu mehreren Aspekten dieser Theorien. Zunächst erörtern wir die Kernel Conditional Mean Embeddings, die den Forschern schon seit über einem Jahrzehnt bekannt sind. Unser Beitrag besteht darin, dass wir sie als Bochner bedingte Erwartungen betrachten, im Gegensatz zu Operatoren zwischen reproduzierenden Kernel-Hilbert-Räumen (RKHS), wie es in der Literatur vorherrschend war, und daher ist ihre Schätzung genau ein Regressionsproblem, bei dem der Ausgangsraum ein RKHS ist. Der Hypothesenraum, in dem diese Regression durchgeführt wird, ist selbst ein (vektorwertiger) RKHS, und eine solche Technik ist allgemein als Kernel-Ridge-Regression bekannt.

Als Nächstes schlagen wir eine besondere Form der Kernel-Ridge-Regression vor, die so genannte U-Statistik-Regression, und wenden diese und die zuvor untersuchten bedingten Mittelwert-Einbettungen auf die Untersuchung bedingter Verteilungseffekte im Rahmen potenzieller Ergebnisse an, die in den Bereichen Medizin und Sozialwissenschaften weit verbreitet sind. Die Arbeit nimmt dann eine eher theoretische Wendung, um lerntheoretische und empirische prozesstheoretische Aspekte der Regression mit unendlich-dimensionalen Ausgaberäumen zu untersuchen, was natürlich vorkommen kann, wenn die Ausgaben selbst Funktionen sind, und von denen Kernel bedingte Mittelwert-Einbettungen ein besonderer Fall sind. Wir erweitern die bestehende Theorie der empirischen Prozesse, ein unverzichtbares Werkzeug in der statistischen Lerntheorie, das jedoch bisher nur für Klassen von reellwertigen Funktionen entwickelt wurde, um Klassen von (möglicherweise unendlich-dimensionalen) vektorwertigen Funktionen zu berücksichtigen; insbesondere schlagen wir Schranken für die metrische Entropie von Klassen glatter vektorwertiger Funktionen vor. Wir werfen auch einen Blick auf den Spezialfall der vektorwertigen Kernel-Ridge-Regression und beweisen ein Konsistenzergebnis, das nicht auf der empirischen Prozesstheorie, sondern auf den leistungsstarken Integraloperatortechniken basiert, die bei der Analyse der Kernel-Ridge-Regression beliebt sind.

Acknowledgements

My four years of doctoral studies in Tübingen have been some of the happiest and most memorable, both inside and outside research, in my already privileged life. I would like to express my heartfelt gratitude to those who have contributed to it.

None of this would have been possible without *Krikamol Muandet*, who, as a group leader at the time in the Empirical Inference Department, Max Planck Institute for Intelligent Systems, Tübingen, took a chance on an applicant without any prior publications. He was always by my side, sometimes as a sharp and insightful supervisor, sometimes as a considerate and kind-hearted mentor, and always, as a great friend. I cannot emphasise enough how much of a positive impact he had on my time in Tübingen, and I know for certain that his students and colleagues at his current institution, CISPA – Helmholtz Center for Information Security, Saarbrücken, are extremely fortunate to have him. I wish him the best of luck.

I would like to say a big thank you to my supervisor and director of the lab, *Bernhard Schölkopf*, for also giving me the chance to embark on the PhD journey and supporting me throughout my research. I remember clearly one zoom meeting during which he gave me the license to spend a few months on the axiomatisation of causality project, and assured me that it would be okay if nothing came out of it. I cannot imagine that many other places would give such freedom and trust in the research projects initiated by students, yet offer sincere encouragement and valuable advice, and for that I am extremely grateful. I would also like to say a special thank you to *Sabrina Rehbaum*, *Ann-Sophie Bähr*, *Lidia Pavel* and *Vincent Berenz* at the MPI for all the incredibly warm and supportive atmosphere that they foster through their work, and I would also like to say a big thank you to all cleaning, maintenance, kitchen, IT and other support staff members at the MPI, without whose hard work the lab could not function.

My Master's thesis supervisor, *Sara van de Geer* at ETH Zürich, gave me the first taste of research, and she was instrumental in my obtaining a PhD position, with her advice and reference letters. I would like to express my sincere gratitude to her for all the help she has given me in kickstarting my research career.

It is in Tübingen that I learnt the value of collaboration in research, and to that end, I would like to express my sincere gratitude to *Simon Buchholz* at the

MPI, who never fails to impress me with his sharp-mindedness and mathematical prowess. I would also like to say a big thank you to all my other collaborators, without whom my research could not have taken the shape it has: *Uri Shalit* at Technion, *Mihir Dhanakshirur* at Indian Institute of Science, *Felix Laumann* and *Mauricio Barahona* at Imperial College London, *Julius von Kügelgen* at the MPI, and *Shiva Kasiviswanathan* and *Patrick Blöbaum* at AWS Santa Clara.

Lastly, my most wholehearted gratitude is reserved for my family and friends, past, present and future, without whom nothing would have meaning.

Contents

Chapter 0

Introduction

This thesis is divided into two parts that treat rather distinct subfields of machine learning research. Part I is about causality, whereby we propose a novel, measure-theoretic framework of causality, called *causal spaces*. Part II is about various aspects of arguably the most ubiquitous technique in statistics and machine learning, namely *regression*, with particular focus on kernel methods and vector-valued output spaces. Even though these two topics are rather distinct in nature, we approach both of them with an emphasis on theoretical investigation. Accordingly, the Introduction is divided to reflect this structure of the paper: in Section 0.1, we will introduce the notion of causality, and why we felt we needed a new framework of causality in addition to the (excellent) existing frameworks, and in Section 0.2, we will introduce and summarise our contributions in regression and kernel methods. Section 0.3 lists all the papers that this thesis is based on, and highlights the correspondence between the chapters of this thesis and the papers.

0.1 The Mathematisation of Causality

Causal thinking is undoubtedly one of the hallmarks of human intelligence. I find it hard to disagree with the opening lines of the celebrated "The Book of Why" by Judea Pearl, the pioneer of modern research on causality (Pearl and Mackenzie, 2018), and since I find it even harder to put it in a pithier way, I will quote it directly:

"Some tens of thousands of years ago, humans began to realize that certain things cause other things and that tinkering with the former can change the latter. No other species grasps this, certainly not to the extent that we do. From this discovery came organized societies, then towns and cities, and eventually the science- and technology-based civilization we enjoy today. All because we asked a simple question: Why?"

Due to such importance of the notion in human cognition and behaviour, causality has duly received a huge amount of attention from researchers in a wide range of domains, including, but not limited to, philosophy (Lewis, 2013; Woodward, 2005; Collins et al., 2004), psychology (Waldmann, 2017), statistics (Pearl et al., 2016; Spirtes et al., 2000) including social, biological and medical sciences (Russo, 2010; Imbens and Rubin, 2015; Illari et al., 2011; Hernan and Robins, 2020), mechanics and law (Beebee et al., 2009). In the recent years, the machine learning community has also taken up a rapidly growing interest in the subject (Peters et al., 2017; Schölkopf, 2022; Schölkopf and von Kügelgen, 2022; Bareinboim et al., 2022), in particular in representation learning (Schölkopf et al., 2021; Mitrovic et al., 2020; Wang and Jordan, 2021; Von Kügelgen et al., 2021; Brehmer et al., 2022; Locatello et al., 2019) and natural language processing (Jin et al., 2022; Feder et al., 2022).

On the other hand, in order to describe, analyse and make predictions about real-world phenomena, *mathematical modelling* is an indispensable and universal tool, whereby mathematical models and frameworks are developed to abstractly represent a concept or a system. To name just a few, *dynamical systems* are widely used to describe how systems evolve over time (Arrowsmith and Place, 1990), for example, the swinging of a pendulum, the flow of water down a pipe, or the number of birds in a particular migratory bird sanctuary. These are typically a system of *differential equations* (Braun and Golubitsky, 1983), which can describe a wider range of phenomena, in which the variable with respect to which the derivative is calculated is not necessarily time. Further, game theoretic models are used to model strategic interactions between rational agents (Fudenberg and Tirole, 1991), and probability theory (Çınlar, 2011) and statistical models (Dobson, 2013) are used to model the concept of uncertainty, randomness or stochasticity that are innate to human intelligence.

In the same vein, in order to model and analyse the notion of causality, we need a mathematical framework that can encode causal information, and over the years, many have been proposed. Most prominently, there are the *structural causal models* (SCMs) (Pearl, 2009; Peters et al., 2017), based most often on directed acyclic graphs (DAGs). Here, the theory of causality is built around variables and structural equations, and probability only enters the picture through a distribution on the exogeneous variables (Janzing and Schölkopf, 2010). Efforts have been made to axiomatise causality based on this framework (Galles and Pearl, 1998; Halpern, 2000; Ibeling and Icard, 2020), but models based on structural equations or graphs inevitably rely on assumptions even for the definitions themselves, such as being confined to a finite number of variables, the issue of solvability in the case of non-recursive (or cyclic) cases, that all common causes (whether latent or observed) are modelled, or that the variables in the model do not causally affect anything outside the model. Hence, these cannot be said to be an "axiomatic definition" in the strictest sense.

The *potential outcomes framework* is a major competing model, most often used in economics, social sciences or medicine research, in which we have a designated *treatment* variable, whose causal effect we are interested in, and for each value of the treatment variable, we have a separate, *potential outcome* variable (Imbens and Rubin, 2015; Hernan and Robins, 2020). There are other, perhaps lesser-known approaches to model causality, such as that based on decision theory (Dawid, 2021; Schenone, 2018), on category theory (Jacobs et al., 2019; Fritz et al., 2022), on an agent explicitly performing actions that transform the state space (Cohen, 2022), or settable systems (White and Chalak, 2009).

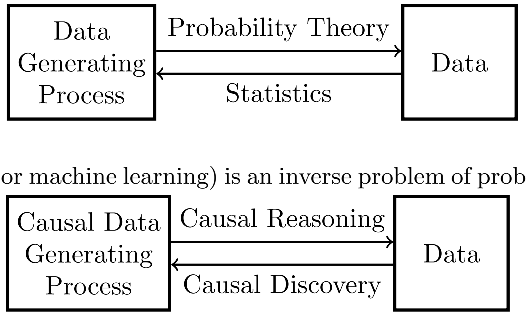
The starting point of this thesis is the that probability theory and statistics (Figure 1a) cannot encode the notion of causality, but will nevertheless play a central role in any theory of causality, since only in very special cases will interventions lead to deterministic changes to the world. Then we observe that the forwards direction of Figure 1a, i.e. probability theory, has a set of axioms based on measure theory that are widely accepted and used[1]. Hence, we argue that it is natural to take the primitive objects of this framework as the basic building blocks, and propose an axiomatic framework called *causal spaces* for the forwards direction of Figure 1b. Despite the fact that all of the existing mathematical frameworks of causality recognise the crucial role that probability plays / should play in any causal theory, it is surprising that few of them try to build directly upon the axioms of probability theory, and those that do fall short in different ways (see below).

As such, perhaps the works that are the most relevant to this thesis are those that have already recognised the need for an axiomatisation of causality based on measure-theoretic probability theory. Ortega (2015) uses a particular form of a *tree* to define a causal space, and in so doing, uses an alternative, Bayesian set-up of probability theory (Jaynes, 2003). It has an obvious drawback that it only considers *countable* sets of "realisations", clearly ruling out many interesting and commonly-occurring cases, and also does not seem to accommodate cycles. Heymann et al. (2021) define the *information dependency model* based on measurable spaces to encode causal information. We find this to be a highly interesting and relevant approach, but the issue of cycles and solvability arises, and again, only countable sets of outcomes are considered, with the authors admitting that the results are likely not to hold with uncountable sets. Moreover, probabilities and interventions require additional work to be taken care of. Lastly, Cabreros and Storey (2019) attempt to provide a measure-theoretic grounding to the potential outcomes framework, but thereby confine attention to the setting of a finite number of variables, and even restrict the random variables to be discrete.

Causal spaces will add to probability spaces in such a way that places the concept of *manipulations* at the heart; more precisely, causal spaces will encode information about what happens to a system when one makes changes to some parts of that system. This manipulative philosophy towards causality is shared by many philosophers (Woodward, 2005), and is the essence behind almost all causal frameworks proposed and adopted in the statistics/machine learning community that we are aware of.

We show that causal spaces strictly generalise (the interventional aspects of) existing frameworks, i.e. given any configuration of, for example, a struc-

[1] Kolmogorov's axiomatisation is without doubt the standard in probability theory. However, we are aware of other, less popular frameworks, for example, one that is more amenable to Bayesian probability (Jaynes, 2003), one based on game theory (Vovk and Shafer, 2014) and imprecise probabilities (Walley, 1991).



(a) Statistics (or machine learning) is an inverse problem of probability theory.

(b) Causal discovery is an inverse problem of causal reasoning.

Figure 1: Data generating processes and data.

tural causal model or potential outcomes framework, we can construct a causal space that can carry the same (interventional) information. Further, we show that causal spaces can seamlessly support situations where existing frameworks struggle, for example those with hidden confounders, cyclic causal relationships or continuous-time stochastic processes.

In the development of a mathematical theory, there are always ways to analyse multiple structures in a coherent manner. For example, in vector spaces, we have the notions of subspaces, product spaces and maps between vector spaces. After proposing causal spaces in Chapter 1, we then discuss operations on multiple causal spaces in Chapter 2. Chapter 1 only consider the development of *single* causal spaces, and omit the discussion of construction of new causal spaces from existing ones or maps between causal spaces. The latter is of particular interest to researchers in causality for the purpose of *abstraction*. When systems, humans or animals perceive the world, they consider different levels of detail depending on their ability to perceive and retain information and their level of interest. It is therefore crucial to connect the mathematical representations at varying levels of granularity in a coherent way.

In probability spaces, such notions are well-established. Product measures give rise to independent random variables, and measurable maps and probability kernels between probability spaces give rise to pushforward measures, which can be interpreted as abstractions or inclusions. Based on these concepts, and using the fact that causal spaces are a direct extension of probability spaces, we develop the notions of *product causal spaces* and *causal transformations*.

Seminal works on maps between causal frameworks lie in the field of SCMs (Rubenstein et al., 2017; Beckers and Halpern, 2019), where the notions of exact transformations, uniform transformation, abstraction, strong abstraction and constructive abstraction are proposed. Beckers et al. (2020) then relax these to an approximate notion. Massidda et al. (2023) extended the notions to soft interventions, and Zečević et al. (2023) to continually updated abstractions. Causal feature learning is a closely related approach, that also aims to learn higher level features (Chalupka et al., 2015, 2016, 2017) There are also approaches based on category theory (Rischel and Weichwald, 2021; Otsuka and Saigo, 2022, 2023) and probabilistic logic (Ibeling and Icard, 2023), all grounded in SCMs; see (Zennaro, 2022) for a review.

The notion of causal abstraction in the SCM framework has found applications in interpretations of neural networks (Geiger et al., 2021, 2023) as well as solving causal inference tasks (identification, estimation and sampling) at different levels of granularity with neural networks (Xia and Bareinboim, 2024). Moreover, Zennaro et al. (2023) proposed a way of *learning* an abstraction from partial information about the abstraction, and demonstrates an application of causal abstraction in the SCM framework in the context of electric vehicle battery manufacturing and Kekić et al. (2023) learn an abstraction that explains a specific target.

0.2 Kernel Regression

Regression is perhaps *the* most popular and widely-used technique in all of statistics and machine learning, and comes with a huge array of subfields according to what models are used (e.g. linear regression (Montgomery et al., 2021), nonparametric regression (Wasserman, 2006; Györfi et al., 2006) including kernel ridge regression (Vovk, 2013) or neural networks (Goodfellow et al., 2016)), or whether the researcher is interested in applications (Lewis-Beck and Lewis-Beck, 2015) or theory (a field called *statistical learning theory* (Vapnik, 1998)). In this thesis, we make a modest contribution to a selection of them, introduced in separate subsections below.

0.2.1 Kernel Conditional Mean Embeddings

The author's research career began with an interest in kernel methods, in particular, the embedding of distributions into *reproducing kernel Hilbert spaces* (RKHSs). The idea of embedding probability distributions into an RKHS, a space associated to a positive definite kernel, has received a lot of attention in the past decades (Berlinet and Thomas-Agnan, 2004; Smola et al., 2007), and has found a wealth of successful applications, such as independence testing (Gretton et al., 2008), two-sample testing (Gretton et al., 2012), learning on distributions (Muandet et al., 2012; Lopez-Paz et al., 2015; Szabó et al., 2016), goodness-of-fit testing (Chwialkowski et al., 2016; Liu et al., 2016) and probabilistic programming (Schölkopf et al., 2015; Simon-Gabriel et al., 2016), among others – see review (Muandet et al., 2017). It extends the idea of kernelising linear methods by embedding data points into high- (and often infinite-)dimensional RKHSs, which has been applied, for example, in ridge regression, spectral clustering, support vector machines and principal component analysis among others (Scholkopf and Smola, 2001; Hofmann et al., 2008; Steinwart and Christmann, 2008).

Conditional distributions can also be embedded into RKHSs in a similar manner (Song et al., 2013),(Muandet et al., 2017, Chapter 4). Compared to unconditional distributions, conditional distributions can represent more complicated relations between random variables, and so conditional mean embeddings (CMEs) have the potential to unlock the arsenal of kernel mean embeddings to a wider setting. Indeed, CMEs have been applied successfully to dynamical systems (Song et al., 2009), inference on graphical models (Song et al., 2010b), probabilistic inference via kernel sum and product rules (Song et al., 2013), reinforcement learning (Grünewälder et al., 2012b; Nishiyama et al., 2012), kernelising the Bayes rule and applying it to nonparametric state-space models (Fukumizu et al., 2013) and causal inference (Mitrovic et al., 2018) to name a few.

Despite such progress, the prevalent definition of the CME based on composing cross-covariance operators (Song et al., 2009) relied on some stringent assumptions, which are often violated and hinder its analysis. Klebanov et al. (2020) recently attempted to clarify and weaken some of these assumptions, but strong and hard-to-verify conditions still persist. Grünewälder et al. (2012a) provided a regression interpretation, but here, only the existence of the CME is shown, without an explicit expression. The main contribution in Chapter 3 is to remove these stringent assumptions using a novel measure-theoretic approach to the CME. This approach requires drastically weaker assumptions, and comes in an explicit expression. We believe this will enable a more principled analysis of its theoretical properties, and open doors to new application areas. We derive an empirical estimate based on vector-valued regression along with an in-depth theoretical analysis, including universal consistency. In particular, we relax the assumption of Grünewälder et al. (2012a) to allow for infinite-dimensional RKHSs.

As natural by-products, we obtain quantities that are extensions of the maximum mean discrepancy (MMD) and the Hilbert-Schmidt independence criterion (HSIC) to the conditional setting, which we call the *maximum conditional mean discrepancy* (MCMD) and the *Hilbert-Schmidt conditional independence criterion* (HSCIC). We demonstrate their properties through simulation experiments.

0.2.2 Conditional Distributional Treatment Effect

In Chapter 4, we discuss a particular form of treatment effect analysis in the potential outcomes framework, namely, the *conditional distributional treatment effect*, where we propose applying two types of kernel regression, the aforementioned kernel conditional mean embeddings and U-statistic regression that is newly proposed in Chapter 4. Analysing the effect of a treatment (medical drug, economic programme, etc.) has long been a problem of great importance, and has attracted researchers from diverse domains, including econometrics Imbens and Wooldridge (2009), political sciences Künzel et al. (2019), healthcare Foster et al. (2011) and social sciences Imbens and Rubin (2015). The field has naturally received much attention of statisticians over the years Rosenbaum

(2002); Rubin (2005); Imbens and Rubin (2015), and in the past few years, the machine learning community has started applying its own armoury to this problem.

Traditional methods for treatment effect evaluation focus on the analysis of the average treatment effect (ATE), such as an increase or decrease in average income, inequality or poverty, aggregated over the population. However, the ATE is not informative about the individual responses to the intervention and how the treatment impact varies across individuals (known as *treatment effect heterogeneity*). The study of conditional average treatment effect (CATE) has been proposed to analyse such heterogeneity in the mean treatment effect. Although sufficient in many cases, the CATE is still an average. As such, it fails to capture information about distributional aspects of the treatment beyond the mean. A significant amount of interest exists for developing methods that can analyse distributional treatment effects conditioned on the covariates Chang et al. (2015); Bitler et al. (2017); Shen (2019); Chernozhukov et al. (2020); Hohberg et al. (2020); Briseño Sanchez et al. (2020).

In the past few years the machine learning community has focused much effort on models for estimating the CATE function. Some approaches include Gaussian processes Alaa and van der Schaar (2017, 2018), Bayesian regression trees Hill (2011); Hahn et al. (2020), random forests Wager and Athey (2018), neural networks Johansson et al. (2016); Shalit et al. (2017); Louizos et al. (2017); Atan et al. (2018); Shi et al. (2019), GANs Yoon et al. (2018), boosting and adaptive regression splines Powers et al. (2018) and kernel mean embeddings Singh et al. (2020).

Distributional extensions of the ATE have been considered by many authors. Abadie (2002) tested the hypotheses of equality and stochastic dominance of the marginal outcome distributions P_{Y_0} and P_{Y_1}, whereas Kim et al. (2018); Muandet et al. (2018) focus on estimating P_{Y_0} and P_{Y_1}, or some distance between them. These works do not consider treatment effect heterogeneity. Singh et al. (2020, Appendix C) consider CATE as well as distributional treatment effect,and while it seems that the ideas can straightforwardly be extended to conditional distributional treatment effect, it is not explicitly considered in the paper.

Interest has also always existed for hypothesis tests in the context of treatment effect analysis, especially in econometrics (Imbens and Wooldridge, 2009, Sections 3.3 and 5.12). Abadie (2002) tested the equality between the marginal distributions of Y_0 and Y_1, while Crump et al. (2008) tested for the equality of $\mathbb{E}[Y_1|X]$ and $\mathbb{E}[Y_0|X]$. Lee and Whang (2009); Lee (2009); Chang et al. (2015); Shen (2019) were interested, among others, in the hypothesis of the equality of $P_{Y_1|X}$ and $P_{Y_0|X}$, which we consider in this thesis.

The *conditional distributional treatment effect* (CoDiTE) incorporates both distributional considerations of treatment effects *and* treatment effect heterogeneity. Interest has been growing, especially in the econometrics literature, for such analyses – indeed, Bitler et al. (2017) provided concrete evidence that in some settings, the CATE does not suffice. Existing works that analyse the CoDiTE can be split into three categories, depending on how distributions are characterised: (i) quantiles, (ii) cumulative distributional functions, and (iii) specific distributional parameters, such as the mean, variance, skewness, etc. In category (i), quantile regression is a powerful tool Koenker (2005); however, in order to get a distributional picture via quantiles, one needs to estimate a large number of quantiles, and issues of crossing quantiles arise, whereby estimated quantiles are non-monotone. In category (ii), Chernozhukov et al. (2013, 2020) propose splitting $\mathcal{Y}$ into a grid and regressing for the cumulative distribution function at each point in the grid, but this also brings issues of non-monotonicity of the cumulative distribution function, similar to crossing quantiles. Shen (2019) estimates the cumulative distribution functions $P(Y_0 < y^*)$ and $P(Y_1 < y^*)$ for each $y^* \in \mathcal{Y}$ given each value of $X = x$ by essentially applying the Nadaraya-Watson conditional U-statistic of Stute (1991) to the U-kernel $h(y) = \mathbf{1}(y \leq y^*)$. In category (iii), generalised additive models for location, scale and shape (GAMLSS) Stasinopoulos et al. (2017) have been applied for CoDiTE analysis Hohberg et al. (2020); Briseño Sanchez et al. (2020), but being a parametric model, despite its flexibility, the researcher has to choose a model beforehand to proceed, and issues of model misspecification are unavoidable.

The contributions of Chapter 4 are as follows. Firstly, we formally define the CoDiTE associated with a chosen distance function between distributions. Then we use kernel conditional mean embeddings to analyse the CoDiTE associated with the *maximum mean discrepancy* Gretton et al. (2012). Coupled with a statistical hypothesis test, this can determine *whether* there exists any effect of the treatment, conditioned on a set of covariates. Finally, we use *conditional witness functions* and *U-statistic regression* to investigate *what kind* of effect the treatment has. We characterise distributions in two ways – first as elements in a reproducing kernel Hilbert space via kernel conditional mean embeddings, which, to the best of our knowledge, is a novel attempt in the treatment effect literature, and secondly via specific distributional parameters, as in category (iii) above. The former characterisation gives us a novel way of testing for the equality of conditional distributions, as well as an exploratory tool for density comparison between the groups via conditional witness functions. For the latter characterisation, we provide, to the best of our knowledge, a novel U-statistic regression technique by generalising kernel ridge regression, which, in contrast to GAMLSS, is fully nonparametric. Neither characterisation requires the estimation of a large number of quantities, unlike characterisations via quantiles or cumulative distribution functions.

0.2.3 Empirical Process Theory

Thence, the thesis takes a more theoretical turn in Chapter 5, where we take a look at two main techniques in analysing kernel ridge regression, namely empirical process theory and integral operator technique.

Empirical process theory is an important branch of probability theory that deals with the empirical measure $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ based on random independent and identically distributed (i.i.d.) copies $X_1, ..., X_n$ of a random variable X on a domain $\mathcal{X}$, and stochastic processes of the form $\{P_n f - Pf : f \in \mathcal{F}\}$, where $\mathcal{F}$ is a class of functions $\mathcal{X} \to \mathbb{R}$. Due to its very nature, the theory has found a wealth of applications in statistics (van der Vaart and Wellner, 1996; van de Geer, 2000; Kosorok, 2008; Shorack and Wellner, 2009; Dudley, 2014). In particular, it has been the major tool in analysing properties of estimators in supervised learning, both in regression and classification (Györfi et al., 2006; Steinwart and Christmann, 2008; Shalev-Shwartz and Ben-David, 2014).

In the traditional (and still dominant) supervised learning setting, the output space is (a subset of) $\mathbb{R}$, but there is a rapidly growing literature in machine learning and statistics on learning vector-valued functions (Micchelli and Pontil, 2005; Álvarez et al., 2012), and efforts are already under way to explore ways to make them faster and more robust (Laforgue et al., 2020; Lambert et al., 2022; Ahmad et al., 2022). This occurs, for example, in multi-task or multi-output learning (Evgeniou et al., 2005; Yousefi et al., 2018; Xu et al., 2019; Reeve and Kaban, 2020), functional response models (Morris, 2015; Kadri et al., 2016; Brault, 2017; Saha and Palaniappan, 2020), kernel conditional mean embeddings (Grünewälder et al., 2012a; Park and Muandet, 2020a) or structured prediction (Ciliberto et al., 2020; Laforgue et al., 2020), among others. Very recently, there is even an interest in the more general setting of learning mappings between two metric spaces (Hanneke et al., 2020; Cohen and Kontorovich, 2022).

There are valuable works analysing the properties of vector-valued regressors with specific algorithms, notably integral operator techniques in vector-valued reproducing kernel Hilbert space regression (Caponnetto and De Vito, 2006; Kadri et al., 2016; Singh et al., 2019; Cabannes et al., 2021), and we highlight our own contributions in Chapter 5 too. Moreover, in the form of (local) Rademacher complexities, empirical process theoretic techniques have been applied to cases where the output space is finite dimensional (Yousefi et al., 2018; Li et al., 2019; Reeve and Kaban, 2020; Wu et al., 2021). However, as general empirical process theory is developed, to the best of our knowledge, exclusively for classes of real-valued functions, the powerful armoury of empirical process theory has not been utilised fully to analyse vector-valued learning problems. The aim of this Chapter is to provide some first steps towards developing a theory of empirical processes with vector-valued functions.

An indispensable object in empirical process theory is metric entropy of function classes[2], and one of the most frequently used function classes is that

[2] In the usual theory of empirical processes with real-valued functions, there are two major tools. The first is to consider the entropy with respect to the empirical measure P_n. One usually requires this entropy to be uniformly bounded over all realisations of the samples $X_1, ..., X_n$, and the most widely-used example of function classes that satisfy this property are the celebrated *Vapnik-Chervonenkis (VC) subgraph classes*. The second tool is what is known as *entropy with bracketing* with respect to the underlying measure P (see, for example, van de Geer (2000, p.122, Theorem 2.4.1 and p.129, Section 2.5.2), van de Geer (2000, Sections 3.1 and 5.5) and Dudley (2014, Chapter 7)). However, both VC subgraph classes and entropy with bracketing make explicit use of the fact that the output space $\mathbb{R}$ is *totally-ordered*, and makes use of objects such as $\{x \in \mathcal{X} : x \leq g(x_0)\}$ and $\{x \in \mathcal{X} : g_1(x_0) \leq x \leq g_2(x_0)\}$, where $g, g_1, g_2 \in \mathcal{G}$ and $x_0 \in \mathcal{X}$. A direct extension is clearly not possible when our output space $\mathcal{Y}$ has any dimension greater than 1, and an attempt at an extension is even more difficult when $\mathcal{Y}$ is infinite-dimensional. In this thesis, we do not investigate whether it is possible to obtain meaningful results by extending these ideas, and leave it for future work.

of smooth functions. In our main results, we investigate how we can bound the entropy of classes of smooth vector-valued functions. When the output space is infinite-dimensional, bounding the entropy becomes far less trivial, compared to the case of real-valued function classes. For example, seemingly benign function classes such as the classes of constant functions onto the unit ball clearly has infinite entropy with respect to any reasonable metric, since the unit ball in an infinite-dimensional Hilbert space is not totally bounded (Bollobás, 1999, p.62, Corollary 6).

0.3 Underlying Manuscripts

This thesis is based on six manuscripts that were written during the course of my PhD studies. The correspondence between chapters and these papers is outlined below.

Chapter 1 Junhyung Park, Simon Buchholz, Bernhard Schölkopf, Krikamol Muandet, "A Measure-Theoretic Axiomatisation of Causality", *NeurIPS 2023.* (Park et al., 2023).

This paper was one of 77 papers chosen for oral presentation out of 12343 submissions.

Chapter 2 Simon Buchholz*, Junhyung Park*, Bernhard Schölkopf, "Products, Abstractions and Inclusions of Causal Spaces", will appear in *UAI 2024.* * means equal contribution.

Chapter 3 Junhyung Park, Krikamol Muandet, "A Measure-Theoretic Approach to Kernel Conditional Mean Embeddings", *NeurIPS 2020.* (Park and Muandet, 2020a).

Chapter 4 Junhyung Park, Uri Shalit, Bernhard Schölkopf, Krikamol Muandet, "Conditional Distributional Treatment Effect with Kernel Conditional Mean Embeddings and U-Statistic Regression", *ICML 2021.* (Park et al., 2021).

Chapter 5 Junhyung Park, Krikamol Muandet, "Regularised Least Squares Regression with Infinite-Dimensional Output Space", *unsubmitted arXiv notes, 2020.* (Park and Muandet, 2020b).

Junhyung Park, Krikamol Muandet, "Towards Empirical Process Theory for Vector-Valued Functions: Metric Entropy of Smooth Function Classes", *ALT 2023.* (Park and Muandet, 2023).

I had the initial idea for all of these papers, and the concretisation of the concepts were done along with my supervisors, Krikamol Muandet (for all the papers on which he is an author), Bernhard Schölkopf (for all the papers on which he is an author) and Uri Shalit (for Park et al. (2021)). The supervisors also gave me advice throughout the projects. The majority of the work was carried out by myself, *except* the UAI paper on which Chapter 2 is based, for which Simon Buchholz is a shared first author. I carried out the work for products of causal spaces, causal independence and comparison with abstraction in the SCM framework, and Simon carried out the work for transformations of causal spaces. Simon Buchholz also checked all of the proofs for Park et al. (2023), and offered advice and insights throughout the project.

There are also a few other articles that I participated in, which are not included in this thesis:

- Mihir Dhanakshirur, Felix Laumann, Junhyung Park, Mauricio Barahona, "A Continuous Structural Intervention Distance to Compare Causal Graphs", *submitted.* (Dhanakshirur et al., 2023).

- Felix Laumann, Julius von Kügelgen, Junhyung Park, Bernhard Schölkopf, Mauricio Barahona, "Kernel-based Independence Tests for Causal Structure Learning on Functional Data", in *Entropy 2023.* (Laumann et al., 2023).

- Junhyung Park, Patrick Blöbaum, Shiva Kasiviswanathan, "Overfitting and Generalization for Regression with Trained Two-Layer ReLU Networks", *submitted.*

Part I

Causality

Chapter 1

Causal Spaces

1.1 Background: Measure Theory and Probability Theory

First, we compactly recall some basic facts about measure and probability theory that we need for the development in this Chapter. Please see Çınlar (2011) for more details.

1.1.1 Measure Theory

Suppose that E is a set. We first define the notion of a σ-algebra. A non-empty collection $\mathcal{E}$ of E is called a *σ-algebra* on E if it is closed under complements and countable unions, that is, if

(i) $A \in \mathcal{E} \implies E \backslash A \in \mathcal{E}$;

(ii) $A_1, A_2, ... \in \mathcal{E} \implies \cup_{n=1}^{\infty} A_n \in \mathcal{E}$

(Çınlar, 2011, p.2). We call $\{\emptyset, E\}$ the *trivial σ-algebra* of E. If $\mathcal{C}$ is an arbitrary collection of subsets of E, then the smallest σ-algebra that contains $\mathcal{C}$, or equivalently, the intersection of all σ-algebras that contain $\mathcal{C}$, is called the *σ-algebra generated by* $\mathcal{C}$, and is denoted $\sigma\mathcal{C}$.

A *measurable space* is a pair $(E, \mathcal{E})$, where E is a set and $\mathcal{E}$ is a σ-algebra on E (Çınlar, 2011, p.4).

Suppose $(E, \mathcal{E})$ and $(F, \mathcal{F})$ are measurable spaces. For $A \in \mathcal{E}$ and $B \in \mathcal{F}$, we define the *measurable rectangle* $A \times B$ as the set of all pairs (x, y) with $x \in A$ and $y \in B$. We define the *product σ-algebra* $\mathcal{E} \otimes \mathcal{F}$ on $E \times F$ as the σ-algebra generated by the collection of all measurable rectangles. The measurable space $(E \times F, \mathcal{E} \otimes \mathcal{F})$ is the *product* of $(E, \mathcal{E})$ and $(F, \mathcal{F})$ (Çınlar, 2011, p.4). More generally, if $(E_1, \mathcal{E}_1), ..., (E_n, \mathcal{E}_n)$ are measurable spaces, their product is

$$\bigotimes_{i=1}^{n}(E_i, \mathcal{E}_i) = (\bigtimes_{i=1}^{n} E_i, \bigotimes_{i=1}^{n} \mathcal{E}_i),$$

where $E_1 \times ... \times E_n$ is the set of all n-tuples $(x_1, ..., x_n)$ with x_i in E_i for $i = 1, ..., n$ and $\mathcal{E}_1 \otimes ... \otimes \mathcal{E}_n$ is the σ-algebra generated by the *measurable rectangles* $A_1 \times ... \times A_n$ with A_i in $\mathcal{E}_i$ for $i = 1, ..., n$ (Çınlar, 2011, p.44). If T is an arbitrary (countable or uncountable) index set and $(E_t, \mathcal{E}_t)$ is a measurable space for each $t \in T$, the *product space* of $\{E_t : t \in T\}$ is the set $\bigtimes_{t \in T} E_t$ of all collections $(x_t)_{t \in T}$ with $x_t \in E_t$ for each $t \in T$. A rectangle in $\bigtimes_{t \in T} E_t$ is a subset of the form

$$\bigtimes_{t \in T} A_t = \{x = (x_t)_{t \in T} \in \bigtimes_{t \in T} E_t : x_t \in A_t \text{ for each } t \text{ in } T\}$$

where A_t differs from E_t for only a finite number of t. It is said to be measurable if $A_t \in \mathcal{E}_t$ for every t (for which A_t differs from E_t). The σ-algebra on $\bigtimes_{t \in T} E_t$ generated by the collection of all measurable rectangles is called the *product σ-algebra* and is denoted by $\bigotimes_{t \in T} \mathcal{E}_t$ (Çınlar, 2011, p.45).

A collection $\mathcal{C}$ of subsets of E is called a p-system if it is closed under intersections (Çınlar, 2011, p.2). If two measures μ and ν on a measurable space $(E, \mathcal{E})$ with $\mu(E) = \nu(E) < \infty$ agree on a p-system generating $\mathcal{E}$, then μ and ν are identical (Çınlar, 2011, p.16, Proposition 3.7).

Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. A mapping $f : E \to F$ is *measurable* if $f^{-1}B \in \mathcal{E}$ for every $B \in \mathcal{F}$ (Çınlar, 2011, p.6).

Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. Let f be a bijection between E and F, and let $\hat{f}$ denote its functional inverse. Then, f is an *isomorphism* if f is measurable relative to $\mathcal{E}$ and $\mathcal{F}$, and $\hat{f}$ is measurable with respect to $\mathcal{F}$ and $\mathcal{E}$. The measurable spaces $(E, \mathcal{E})$ and $(F, \mathcal{F})$ are *isomorphic* if there exists an isomorphism between them (Çınlar, 2011, p.11).

A measurable space $(E, \mathcal{E})$ is a *standard measurable space* if it is isomorphic to $(F, \mathcal{B}_F)$ for some Borel subset F of $\mathbb{R}$. Polish spaces with their Borel σ-algebra are standard measurable spaces (Çınlar, 2011, p.11).

Let $A \subset E$. Its *indicator*, denoted by 1_A, is the function defined by

$$1_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

(Çınlar, 2011, p.8). Obviously, 1_A is $\mathcal{E}$-measurable if and only if $A \in \mathcal{E}$. A function $f : E \to \mathbb{R}$ is said to be *simple* if it is of the form

$$f = \sum_{i=1}^{n} a_i 1_{A_i}$$

for some $n \in \mathbb{N}$, $a_1, ..., a_n \in \mathbb{R}$ and $A_1, ..., A_n \in \mathcal{E}$ (Çınlar, 2011, p.8). The $A_1, ..., A_n \in \mathcal{E}$ can be chosen to be a measurable partition of E, and is then called the *canonical form* of the simple function f. A positive function on E is $\mathcal{E}$-measurable if and only if it is the limit of an increasing sequence of positive simple functions (Çınlar, 2011, p.10, Theorem 2.17).

A *measure* on a measurable space $(E, \mathcal{E})$ is a mapping $\mu : \mathcal{E} \to [0, \infty]$ such that

(i) $\mu(\emptyset) = 0$;

(ii) $\mu(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$ for every disjoint sequence (A_n) in $\mathcal{E}$

(Çınlar, 2011, p.14). A *measure space* is a triplet $(E, \mathcal{E}, \mu)$, where $(E, \mathcal{E})$ is a measurable space and μ is a measure on it.

A measurable set B is said to be *negligible* if $\mu(B) = 0$, and an arbitrary subset of E is said to be *negligible* if it is contained in a measurable negligible set. The measure space is said to be *complete* if every negligible set is measurable (Çınlar, 2011, p.17).

Next, we review the notion of integration of a real-valued function $f : E \to \mathbb{R}$ with respect to μ (Çınlar, 2011, p.20, Definition 4.3).

(a) Let $f : E \to [0, \infty]$ be simple. If its canonical form is $f = \sum_{i=1}^{n} a_i 1_{A_i}$ with $a_i \in \mathbb{R}$, then we define

$$\int f d\mu = \sum_{i=1}^{n} a_i \mu(A_i).$$

(b) Suppose $f : E \to [0, \infty]$ is measurable. Then by above, we have a sequence (f_n) of positive simple functions such that $f_n \to f$ pointwise. Then we define

$$\int f d\mu = \lim_{n\to\infty} \int f_n d\mu,$$

where $\int f_n d\mu$ is defined for each n by (a).

(c) Suppose $f : E \to [-\infty, \infty]$ is measurable. Then $f^+ = \max\{f, 0\}$ and $f^- = -\min\{f, 0\}$ are measurable and positive, so we can define $\int f^+ d\mu$ and $\int f^- d\mu$ as in (b). Then we define

$$\int f d\mu = \int f^+ d\mu - \int f^- d\mu$$

provided that at least one term on the right be positive. Otherwise, $\int f d\mu$ is undefined. If $\int f^+ d\mu < \infty$ and $\int f^- d\mu < \infty$, then we say that f is *integrable*.

Finally, we review the notion of *transition kernels*, which are crucial in the consideration of conditional distributions. Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. Let K be a mapping $E \times \mathcal{F} \to [0, \infty]$. Then, K is called a *transition kernel* from $(E, \mathcal{E})$ into $(F, \mathcal{F})$ if

(a) the mapping $x \mapsto K(x, B)$ is measurable for every set $B \in \mathcal{F}$; and

(b) the mapping $B \mapsto K(x, B)$ is a measure on $(F, \mathcal{F})$ for every $x \in E$.

A transition kernel from $(E, \mathscr{E})$ into $(F, \mathscr{F})$ is called a *probability transition kernel* if $K(x, F) = 1$ for all $x \in E$. A probability transition kernel K from $(E, \mathscr{E})$ into $(E, \mathscr{E})$ is called a *Markov kernel on* $(E, \mathscr{E})$ (Çınlar, 2011, p.37,39,40).

1.1.2 Probability Theory

Now we translate the above measure-theoretic notions into the language of probability theory, and introduce some additional concepts. A *probability space* is a measure space $(\Omega, \mathcal{H}, \mathbb{P})$ such that $\mathbb{P}(\Omega) = 1$ (Çınlar, 2011, p.49). We call Ω the *sample space*, and each element $\omega \in \Omega$ an *outcome*. We call $\mathcal{H}$ a collection of *events*, and for any $A \in \mathcal{H}$, we read $\mathbb{P}(A)$ as the *probability that the event A occurs* (Çınlar, 2011, p.50).

A *random variable* taking values in a measurable space $(E, \mathcal{E})$ is a function $X : \Omega \to E$, measurable with respect to $\mathcal{H}$ and $\mathcal{E}$. The *distribution* of X is the measure μ on $(E, \mathcal{E})$ defined by $\mu(A) = \mathbb{P}(X^{-1}A)$ (Çınlar, 2011, p.51). For an arbitrary set T, let X_t be a random variable taking values in $(E, \mathcal{E})$ for each $t \in T$. Then the collection $\{X_t : t \in T\}$ is called a *stochastic process* with *state space* $(E, \mathcal{E})$ and *parameter set* T (Çınlar, 2011, p.53).

Henceforth, random variables are defined on $(\Omega, \mathcal{H}, \mathbb{P})$ and take values in $[-\infty, \infty]$. We define the *expectation* of a random variable $X : \Omega \to [-\infty, \infty]$ as $\mathbb{E}[X] = \int_\Omega X d\mathbb{P}$ (Çınlar, 2011, p.57-58). We also define the *conditional expectation* (Çınlar, 2011, p.140, Definition 1.3). Suppose $\mathcal{F}$ is a sub-σ-algebra of $\mathcal{H}$.

(a) Suppose X is a positive random variable. Then the *conditional expectation of X given $\mathcal{F}$* is any positive random variable $\mathbb{E}_\mathcal{F} X$ satisfying

$$\mathbb{E}[VX] = \mathbb{E}\left[V \mathbb{E}_\mathcal{F} X\right]$$

for all $V : \Omega \to [0, \infty]$ measurable with respect to $\mathcal{F}$.

(b) Suppose $X : \Omega \to [-\infty, \infty]$ is a random variable. If $\mathbb{E}[X]$ exists, then we define

$$\mathbb{E}_\mathcal{F} X = \mathbb{E}_\mathcal{F} X^+ - \mathbb{E}_\mathcal{F} X^-,$$

where $\mathbb{E}_\mathcal{F} X^+$ and $\mathbb{E}_\mathcal{F} X^-$ are defined in (a).

Next, we define *conditional probabilities*, and regular versions thereof (Çınlar, 2011, pp.149-151). Suppose $H \in \mathcal{H}$, and let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{H}$. Then the *conditional probability* of H given $\mathcal{F}$ is defined as

$$\mathbb{P}_\mathcal{F} H = \mathbb{E}_\mathcal{F} 1_H.$$

Let $Q(H)$ be a version of $\mathbb{P}_\mathcal{F} H$ for every $H \in \mathcal{H}$. Then $Q : (\omega, H) \mapsto Q_\omega(H)$ is said to be a *regular version* of the conditional probability $\mathbb{P}_\mathcal{F}$ provided that Q be a probability transition kernel from $(\Omega, \mathcal{F})$ into $(\Omega, \mathcal{H})$. Regular versions exist if $(\Omega, \mathcal{H})$ is a standard measurable space (Çınlar, 2011, p.151, Theorem 2.7).

The *conditional distribution* of a random variable X given $\mathcal{F}$ is any transition probability kernel $L : (\omega, B) \mapsto L_\omega(B)$ from $(\Omega, \mathcal{F})$ into $(E, \mathcal{E})$ such that

$$P_\mathcal{F}\{Y \in B\} = L(B) \qquad \text{for all } B \in \mathcal{E}.$$

If $(E, \mathcal{E})$ is a standard measurable space, then a version of the conditional distribution of X given $\mathcal{F}$ exists (Çınlar, 2011, p.151).

Suppose that T is a totally ordered set, i.e. whenever $r, s, t \in T$ with $r < s$ and $s < t$, we have $r < t$ and for any $s, t \in T$, exactly one of $s < t, s = t$ and $t < s$ holds (Enderton, 1977, p.62). For each $t \in T$, let $\mathcal{F}_t$ be a sub-σ-algebra of $\mathcal{H}$. The family $\mathcal{F} = \{\mathcal{F}_t : t \in T\}$ is called a *filtration* provided that $\mathcal{F}_s \subset \mathcal{F}_t$ for $s < t$ (Çınlar, 2011, p.79). A *filtered probability space* $(\Omega, \mathcal{H}, \mathcal{F}, \mathbb{P})$ is a probability space $(\Omega, \mathcal{H}, \mathbb{P})$ endowed with a filtration $\mathcal{F}$.

Finally, we review the notion of *independence* and *conditional independence*. For a fixed integer $n \geq 2$, let $\mathcal{F}_1, ..., \mathcal{F}_n$ be sub-σ-algebras of $\mathcal{H}$. Then $\{\mathcal{F}_1, ..., \mathcal{F}_n\}$ is called an *independency* if

$$\mathbb{P}\left(H_1 \cap ... \cap H_n\right) = \mathbb{P}\left(H_1\right) ... \mathbb{P}\left(H_n\right)$$

for all $H_1 \in \mathcal{F}_1, ..., H_n \in \mathcal{F}_n$. Let T be an arbitrary index set. Let $\mathcal{F}_t$ be a sub-σ-algebra of $\mathcal{H}$ for each $t \in T$. The collection $\{\mathcal{F}_t : t \in T\}$ is called an *independency* if its every finite subset is an independency (Çınlar, 2011, p.82).

Moreover, $\mathcal{F}_1, ..., \mathcal{F}_n$ are said to be *conditional independent* given $\mathcal{F}$ if

$$\mathbb{P}_{\mathcal{F}}\left(H_1 \cap ... \cap H_n\right) = \mathbb{P}_{\mathcal{F}}\left(H_1\right) ... \mathbb{P}_{\mathcal{F}}\left(H_n\right)$$

for all $H_1 \in \mathcal{F}_1, ..., H_n \in \mathcal{F}_n$ (Çınlar, 2011, p.158).

1.2 Causal Spaces

In the development of probability theory, one starts by assuming the existence of a probability space $(\Omega, \mathcal{H}, \mathbb{P})$. However, the actual construction of probability spaces that can carry random variables corresponding to desired random experiments is done through (repeated applications of) two main results – those of Ionescu-Tulcea and Kolmogorov (Çınlar, 2011, p.160, Chapter IV, Section 4); the former constructs a probability space that can carry a finite or countably infinite chain of trials, and the latter shows the existence of a probability space that can carry a process with an arbitrary index set. In both cases, the measurable space $(\Omega, \mathcal{H})$ is constructed as a product space:

(i) for a finite set of trials, each taking place in some measurable space $(E_t, \mathcal{E}_t), t = 1, ..., n$, we have $(\Omega, \mathcal{H}) = \otimes_{t=1}^{n}(E_t, \mathcal{E}_t)$;

(ii) for a countably infinite set of trials, each taking place in some measurable space $(E_t, \mathcal{E}_t)$, $t \in \mathbb{N}$, we have $(\Omega, \mathcal{H}) = \otimes_{t \in \mathbb{N}}(E_t, \mathcal{E}_t)$;

(iii) for a process $\{X_t : t \in T\}$ with an arbitrary index set T, we assume that all the X_t live in the same standard measurable space $(E, \mathcal{E})$, and let $(\Omega, \mathcal{H}) = (E, \mathcal{E})^T = \otimes_{t \in T}(E, \mathcal{E})$.

In the construction of a *causal space*, we will take as our starting point a probability space $(\Omega, \mathcal{H}, \mathbb{P})$, where the measure $\mathbb{P}$ is defined on a product measurable space $(\Omega, \mathcal{H}) = \otimes_{t \in T}(E_t, \mathcal{E}_t)$ with the $(E_t, \mathcal{E}_t)$ being the same standard measurable space if T is uncountable. Denote by $\mathcal{P}(T)$ the power set of T, and for $S \in \mathcal{P}(T)$, we denote by $\mathcal{H}_S$ the sub-σ-algebra of $\mathcal{H} = \otimes_{t \in T}\mathcal{E}_t$ generated by measurable rectangles $\times_{t \in T} A_t$, where $A_t \in \mathcal{E}_t$ differs from E_t only for $t \in S$. In particular, $\mathcal{H}_\emptyset = \{\emptyset, \mathcal{H}\}$ is the trivial σ-algebra of $\Omega = \times_{t \in T} E_t$. Also, we denote by Ω_S the subspace $\times_{s \in S} E_s$ of $\Omega = \times_{t \in T} E_t$, and for $T \supseteq S \supseteq U$, we let π_{SU} denote the natural projection from Ω_S onto Ω_U.

Definition 1.2.1. A *causal space* is defined as the quadruple $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, where $(\Omega, \mathcal{H}, \mathbb{P}) = (\times_{t \in T} E_t, \otimes_{t \in T}\mathcal{E}_t, \mathbb{P})$ is a probability space and $\mathbb{K} = \{K_S : S \in \mathcal{P}(T)\}$, called the *causal mechanism*, is a collection of transition probability kernels K_S from $(\Omega, \mathcal{H}_S)$ into $(\Omega, \mathcal{H})$, called the *causal kernel on* $\mathcal{H}_S$, that satisfy the following axioms:

(i) for all $A \in \mathcal{H}$ and $\omega \in \Omega$,

$$K_\emptyset(\omega, A) = \mathbb{P}(A);$$

(ii) for all $\omega \in \Omega$, $A \in \mathcal{H}_S$ and $B \in \mathcal{H}$,

$$K_S(\omega, A \cap B) = 1_A(\omega) K_S(\omega, B) = \delta_\omega(A) K_S(\omega, B);$$

in particular, for $A \in \mathcal{H}_S$, $K_S(\omega, A) = 1_A(\omega) K_S(\omega, \Omega) = 1_A(\omega) = \delta_\omega(A)$.

Here, the probability measure $\mathbb{P}$ should be viewed as the "observational measure", and the causal mechanism $\mathbb{K}$, consisting of causal kernels K_S for $S \in \mathcal{P}(T)$, contains the "causal information" of the space, by directly specifying the interventional distributions. We write $1_A(\omega)$ when viewed as a function in ω for a fixed A, and $\delta_\omega(A)$ when viewed as a measure for a fixed $\omega \in \Omega$. Note that $\mathbb{K}$ cannot be determined "independently" of the probability measure $\mathbb{P}$, since, for example, $K_\emptyset$ is clearly dependent on $\mathbb{P}$ by (i).

Before we discuss the meaning of the two axioms, we immediately give the definition of an *intervention*. An intervention is carried out on a sub-σ-algebra of the form $\mathcal{H}_U$ for some $U \in \mathcal{P}(T)$. In the following, for $S \in \mathcal{P}(T)$, we denote $\omega_S = \pi_{TS}(\omega)$. Then note that $\Omega = \Omega_S \times \Omega_{T \setminus S}$ and for any $\omega \in \Omega$, we can decompose it into components as $\omega = (\omega_S, \omega_{T \setminus S})$. Then $K_S(\omega, A) = K_S((\omega_S, \omega_{T \setminus S}), A)$ for any $A \in \mathcal{H}$ only depends on the first ω_S component of $\omega = (\omega_S, \omega_{T \setminus S})$. As a slight abuse of notation, we will sometimes write $K_S(\omega_S, A)$ for conciseness.

Definition 1.2.2. Let $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t \in T} E_t, \otimes_{t \in T}\mathcal{E}_t, \mathbb{P}, \mathbb{K})$ be a causal space, and $U \in \mathcal{P}(T)$, $\mathbb{Q}$ a probability measure on $(\Omega, \mathcal{H}_U)$ and $\mathbb{L} = \{L_V : V \in \mathcal{P}(U)\}$ a causal mechanism on $(\Omega, \mathcal{H}_U, \mathbb{Q})$. An *intervention on* $\mathcal{H}_U$ *via* $(\mathbb{Q}, \mathbb{L})$ is a new causal space $(\Omega, \mathcal{H}, \mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}, \mathbb{K}^{\mathrm{do}(U,\mathbb{Q},\mathbb{L})})$, where the *intervention measure* $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}$ is a probability measure on $(\Omega, \mathcal{H})$ defined, for $A \in \mathcal{H}$, by

$$\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(A) = \int \mathbb{Q}(d\omega) K_U(\omega, A) \tag{1}$$

and $\mathbb{K}^{\mathrm{do}(U,\mathbb{Q},\mathbb{L})} = \{K_S^{\mathrm{do}(U,\mathbb{Q},\mathbb{L})} : S \in \mathcal{P}(T)\}$ is the *intervention causal mechanism* whose *intervention causal kernels* are

$$K_S^{\mathrm{do}(U,\mathbb{Q},\mathbb{L})}(\omega, A) = \int L_{S\cap U}(\omega_{S\cap U}, d\omega'_U) K_{S\cup U}((\omega_{S\setminus U}, \omega'_U), A). \tag{2}$$

The intuition behind these definitions is as follows. Starting from the probability space $(\Omega, \mathcal{H}, \mathbb{P})$, we choose a "subspace" on which to intervene, namely a sub-σ-algebra $\mathcal{H}_U$ of $\mathcal{H}$. The *intervention* is the process of placing any desired measure $\mathbb{Q}$ on this "subspace" $(\Omega, \mathcal{H}_U)$, along with an *internal* causal mechanism $\mathbb{L}$ on this "subspace"[1]. The causal kernel K_U corresponding to the "subspace" $\mathcal{H}_U$, which is encoded in the original causal space, determines what the *intervention measure* on the whole space $\mathcal{H}$ will be, via equation (1). For the causal kernels after intervention, the causal effect first takes place within $\mathcal{H}_U$ via the internal causal mechanism $\mathbb{L}$, then propagates to the rest of $\mathcal{H}$ via equation (2).

The definition of intervening on a σ-algebra of the form $\mathcal{H}_U$ given in Definition 1.2.2 sheds light on the two axioms of causal spaces given in Definition 1.2.1.

Remark 1.2.3. Trivial Intervention Axiom (i) in Definition 1.2.1 ensures that intervening on the trivial σ-algebra (i.e. not intervening at all) leaves the probability measure intact, i.e. writing $\mathbb{Q}$ for the trivial probability measure on $\{\emptyset, \Omega\}$, we have $\mathbb{P}^{\mathrm{do}(\emptyset,\mathbb{Q})} = \mathbb{P}$.

Interventional Determinism Axiom (ii) of Definition 1.2.1 ensures that for any $A \in \mathcal{H}_U$, we have $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(A) = \mathbb{Q}(A)$, which means that if we intervene on the causal space by giving $\mathcal{H}_U$ a particular probability measure $\mathbb{Q}$, then $\mathcal{H}_U$ indeed has that measure with respect to the intervention probability measure.

The following example should serve as further clarification of the concepts.

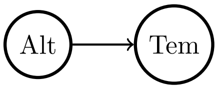


Figure 1.1: Altitude and Temperature.

Example 1.2.4. *Let $E_1 = E_2 = \mathbb{R}$, and $\mathcal{E}_1, \mathcal{E}_2$ be Lebesgue σ-algebras on E_1 and E_2. Each $e_1 \in E_1$ and $e_2 \in E_2$ respectively represent the altitude in metres and temperature in Celsius of a random location. For simplicity, we assume a jointly Gaussian measure $\mathbb{P}$ on $(\Omega, \mathcal{H}) = (E_1 \times E_2, \mathcal{E}_1 \otimes \mathcal{E}_2)$, say with mean vector $\begin{pmatrix} 1000 \\ 10 \end{pmatrix}$ and covariance matrix $\begin{pmatrix} 300 & -15 \\ -15 & 1 \end{pmatrix}$. For each $e_1 \in E_1$ and $A \in \mathcal{E}_2$, we let $K_1(e_1, A)$ be the conditional measure of $\mathbb{P}$ given e_1, i.e. Gaussian with mean $\frac{1200-e_1}{20}$ and variance $\frac{1}{4}$. This represents the fact that, if we intervene*

[1] Choosing $\mathbb{Q}$ to have measure 1 on a single element would correspond to what is known as a "hard intervention" in the SCM literature. Letting $\mathbb{Q}$ and $\mathbb{L}$ be arbitrary would allow us to obtain any "soft intervention".

by fixing the altitude of a location, then the temperature of that location will be causally affected. However, if we intervene by fixing a temperature of a location, say by manually heating up or cooling down a place, then we expect that this has no causal effect on the altitude of the place. This can be represented by the causal kernel $K_2(e_2, B) = \mathbb{P}(B)$ *for each* $B \in \mathcal{E}_1$*, i.e. Gaussian measure with mean* 1000 *and variance* 300*, regardless of the value of* e_2*. The corresponding "causal graph" would be Figure 1.1. If we intervene on* $\mathcal{E}_1$ *with measure* δ_{1000}*, i.e. we fix the altitude at* $1000m$*, then the intervention measure* $\mathbb{P}^{\text{do}(1,\delta_{1000})}$ *on* $(E_2, \mathcal{E}_2)$ *would be Gaussian with mean* 10 *and variance* $\frac{1}{4}$*. If we intervene on* $\mathcal{E}_2$ *with any measure* $\mathbb{Q}$*, the intervention measure* $\mathbb{P}^{\text{do}(2,\mathbb{Q})}$ *on* $(E_1, \mathcal{E}_1)$ *would still be Gaussian with mean* 1000 *and variance* 300*.*

The following theorem proves that the intervention measure and causal mechanism are indeed valid.

Theorem 1.2.5. *From Definition 1.2.2,* $\mathbb{P}^{\text{do}(U,\mathbb{Q})}$ *is indeed a measure on* $(\Omega, \mathcal{H})$*, and* $\mathbb{K}^{\text{do}(U,\mathbb{Q},\mathbb{L})}$ *is indeed a valid causal mechanism on* $(\Omega, \mathcal{H}, \mathbb{P}^{\text{do}(U,\mathbb{Q})})$*, i.e. they satisfy the axioms of Definition 1.2.1.*

To end this Section, we make a couple of further remarks on the definition of causal spaces.

Remark 1.2.6. (i) We require causal spaces to be built on top of product probability spaces, as opposed to general probability spaces, and causal kernels are defined on sub-σ-algebras of $\mathcal{H}$ of the form $\mathcal{H}_S$ for $S \in \mathcal{P}(T)$, as opposed to general sub-σ-algebras of $\mathcal{H}$. This is because, for two events that are not in separate components of a product space, one can always intervene on one of those events in such a way that the measure on the other event will have to change, meaning the causal kernel cannot be decoupled from the intervention itself. For example, in a dice-roll with outcomes $\{1, 2, 3, 4, 5, 6\}$ each with probability $\frac{1}{6}$, if we intervene to give measure 1 to roll 6, then the other outcomes are forced to have measure 0. Only if we consider separate components of product measurable spaces can we set meaningful causal relationships that are decoupled from the act of intervention itself.

(ii) We do not distinguish between interventions that are practically possible and those that are not. For example, the "causal effect of sunlight on the moon's temperature" cannot be measured realistically, as it would require covering up the sun, but the information encoded in the causal kernel would still correspond to what would happen when we cover up the sun.

1.3 Causal Effects

In this section, we define what it means for a sub-σ-algebra of the form $\mathcal{H}_S$ to have a *causal effect* on an event $A \in \mathcal{H}$.

Definition 1.3.1. Let $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t\in T} E_t, \otimes_{t\in T} \mathcal{E}_t, \mathbb{P}, \mathbb{K})$ be a causal space, $U \in \mathcal{P}(T)$, $A \in \mathcal{H}$ an event and $\mathcal{F}$ a sub-σ-algebra of $\mathcal{H}$ (not necessarily of the form $\mathcal{H}_S$ for some $S \in \mathcal{P}(T)$).

(i) If $K_S(\omega, A) = K_{S\setminus U}(\omega, A)$ for all $S \in \mathcal{P}(T)$ and all $\omega \in \Omega$, then we say that $\mathcal{H}_U$ has *no causal effect on* A, or that $\mathcal{H}_U$ is *non-causal to* A.

We say that $\mathcal{H}_U$ has *no causal effect on* $\mathcal{F}$, or that $\mathcal{H}_U$ is *non-causal to* $\mathcal{F}$, if, for all $A \in \mathcal{F}$, $\mathcal{H}_U$ has no causal effect on A.

(ii) If there exists $\omega \in \Omega$ such that $K_U(\omega, A) \neq \mathbb{P}(A)$, then we say that $\mathcal{H}_U$ has an *active causal effect on* A, or that $\mathcal{H}_U$ is *actively causal to* A.

We say that $\mathcal{H}_U$ has an *active causal effect on* $\mathcal{F}$, or that $\mathcal{H}_U$ is *actively causal to* $\mathcal{F}$, if $\mathcal{H}_U$ has an active causal effect on some $A \in \mathcal{F}$.

(iii) Otherwise, we say that $\mathcal{H}_U$ has a *dormant causal effect on* A, or that $\mathcal{H}_U$ is *dormantly causal to* A.

We say that $\mathcal{H}_U$ has a *dormant causal effect on* $\mathcal{F}$, or that $\mathcal{H}_U$ is *dormantly causal to* $\mathcal{F}$, if $\mathcal{H}_U$ does not have an active causal effect on any event in $\mathcal{F}$ and there exists $A \in \mathcal{F}$ on which $\mathcal{H}_U$ has a dormant causal effect.

Sometimes, we will say that $\mathcal{H}_U$ has a *causal effect* on A to mean that $\mathcal{H}_U$ has either an active or a dormant causal effect on A.

The intuition is as follows. For any $S \in \mathcal{P}(T)$ and any fixed event $A \in \mathcal{H}$, consider the function $\omega_S \mapsto K_S((\omega_{S\cap U}, \omega_{S\setminus U}), A)$. If $\mathcal{H}_U$ has no causal effect on A, then it means that the causal kernel does not depend on the $\omega_{S\cap U}$ component of ω_S. Since this has to hold for all $S \in \mathcal{P}(T)$, it means that it is possible to have, for example, $K_U(\omega, A) = \mathbb{P}(A)$ for all $\omega \in \Omega$ and yet for $\mathcal{H}_U$ to have a causal effect on A. This would be precisely the case where $\mathcal{H}_U$ has a dormant causal effect on A, and it means that, for some $S \in \mathcal{P}(T)$, $\omega_S \mapsto K_S((\omega_{S\cap U}, \omega_{S\setminus U}, A)$ does depend on the $\omega_{S\cap U}$ component.

Remark 1.3.2. We collect some straightforward but important special cases.

(a) If $\mathcal{H}_U$ has no causal effect on A, then letting $S = U$ in Definition 1.3.1(i) and applying Definition 1.2.1(i), we can see that, for all $\omega \in \Omega$,

$$K_U(\omega, A) = K_{U\setminus U}(\omega, A) = K_{\emptyset}(\omega, A) = \mathbb{P}(A).$$

In particular, this means that $\mathcal{H}_U$ cannot have both no causal effect and active causal effect on A.

(b) It is immediate that the trivial σ-algebra $\mathcal{H}_{\emptyset} = \{\emptyset, \Omega\}$ has no causal effect on any event $A \in \mathcal{H}$. Conversely, it is also clear that $\mathcal{H}_U$ for any $U \in \mathcal{P}(T)$ has no causal effect on the trivial σ-algebra.

(c) Let $U \in \mathcal{P}(T)$ and $\mathcal{F}$ a sub-σ-algebra of $\mathcal{H}$. If $\mathcal{H}_U \cap \mathcal{F} \neq \{\emptyset, \Omega\}$, then $\mathcal{H}_U$ has an active causal effect on $\mathcal{F}$, since, for $A \in \mathcal{H}_U \cap \mathcal{F}$ with $A \neq \emptyset$

and $A \neq \Omega$, Definition 1.2.1(ii) tells us that $K_U(\cdot, A) = 1_A(\cdot) \neq \mathbb{P}(A)$. In particular, $\mathcal{H}_U$ has an active causal effect on itself. Further, the full σ-algebra $\mathcal{H} = \mathcal{H}_T$ has an active causal effect on all of its sub-σ-algebras except the trivial σ-algebra, and every $\mathcal{H}_U, U \in \mathcal{P}(T)$ except the trivial σ-algebra has an active causal effect on the full σ-algebra $\mathcal{H}$.

(d) Let $U \in \mathcal{P}(T)$ and $\mathcal{F}_1, \mathcal{F}_2$ be sub-σ-algebras of $\mathcal{H}$. If $\mathcal{F}_1 \subseteq \mathcal{F}_2$ and $\mathcal{H}_U$ has no causal effect on $\mathcal{F}_2$, then it is clear that $\mathcal{H}_U$ has no causal effect on $\mathcal{F}_1$.

(e) If $\mathcal{H}_U$ has no causal effect on an event A, then for any $V \in \mathcal{P}(T)$ with $V \subseteq U$, $\mathcal{H}_V$ has no causal effect on A. Indeed, take any $S \in \mathcal{P}(T)$. Then using the fact that $\mathcal{H}_U$ has no causal effect on A, see that, for any $\omega \in \Omega$,

$$\begin{aligned} K_{S\setminus V}(\omega, A) &= K_{(S\setminus V)\setminus U}(\omega, A) && \text{applying Definition 1.3.1(i) with } S \setminus V \\ &= K_{S\setminus U}(\omega, A) && \text{since } V \subseteq U \\ &= K_S(\omega, A) && \text{applying Definition 1.3.1(i) with } S. \end{aligned}$$

Since $S \in \mathcal{P}(T)$ was arbitrary, we have that $\mathcal{H}_V$ has no causal effect on A.

(f) Contrapositively, if $U, V \in \mathcal{P}(T)$ with $V \subseteq U$ and $\mathcal{H}_V$ has a causal effect on A, then $\mathcal{H}_U$ has a causal effect on A.

(g) If $U \in \mathcal{P}(T)$ has no causal effect on A, then for any $V \in \mathcal{P}(T)$, we have

$$K_V(\omega, A) = K_{U\cup V}(\omega, A).$$

Indeed, since $U \setminus V$ has no causal effect on A by (e),

$$\begin{aligned} K_{U\cup V}(\omega, A) &= K_{(U\cup V)\setminus(U\setminus V)}(\omega, A) \\ &= K_V(\omega, A) && \text{since } (U \cup V) \setminus (U \setminus V) = V. \end{aligned}$$

(h) If $U, V \in \mathcal{P}(T)$ and neither $\mathcal{H}_U$ nor $\mathcal{H}_V$ has a causal effect on A, then $\mathcal{H}_{U\cup V}$ has no causal effect on A. Indeed, for any $S \in \mathcal{P}(T)$ and any $\omega \in \Omega$,

$$\begin{aligned} K_{S\setminus(U\cup V)}(\omega, A) &= K_{(S\setminus U)\setminus V}(\omega, A) \\ &= K_{S\setminus U}(\omega, A) && \text{as } V \text{ has no causal effect on } A \\ &= K_S(\omega, A) && \text{as } U \text{ has no causal effect on } A. \end{aligned}$$

Since $S \in \mathcal{P}(T)$ was arbitrary, $\mathcal{H}_{U\cup V}$ has no causal effect on A.

(i) Contrapositively, if $U, V \in \mathcal{P}(T)$ and $\mathcal{H}_{U\cup V}$ has a causal effect on A, then either $\mathcal{H}_U$ or $\mathcal{H}_V$ has a causal effect on A.

Following the definition of no causal effect, we define the notion of a *trivial causal kernel*.

Definition 1.3.3. Let $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t\in T} E_t, \otimes_{t\in T} \mathcal{E}_t, \mathbb{P}, \mathbb{K})$ be a causal space, and $U \in \mathcal{P}(T)$. We say that the causal kernel K_U is *trivial* if $\mathcal{H}_U$ has no causal effect on $\mathcal{H}_{T\setminus U}$.

Note that we can decompose $\mathcal{H}$ as $\mathcal{H} = \mathcal{H}_U \otimes \mathcal{H}_{T\setminus U}$, and so $\mathcal{H}$ is generated by events of the form $A \times B$ for $A \in \mathcal{H}_U$ and $B \in \mathcal{H}_{T\setminus U}$. But if K_U is trivial, then we have, by Axiom 1.2.1(ii), $K_U(\omega, A \times B) = 1_A(\omega)\mathbb{P}(B)$ for such a rectangle.

We also define a "conditional" version of causal effects.

Definition 1.3.4. Let $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t\in T} E_t, \otimes_{t\in T} \mathcal{E}_t, \mathbb{P}, \mathbb{K})$ be a causal space, $U, V \in \mathcal{P}(T)$, $A \in \mathcal{H}$ an event and $\mathcal{F}$ a sub-σ-algebra of $\mathcal{H}$ (not necessarily of the form $\mathcal{H}_S$ for some $S \in \mathcal{P}(T)$).

(i) If $K_{S\cup V}(\omega, A) = K_{(S\cup V)\setminus(U\setminus V)}(\omega, A)$ for all $S \in \mathcal{P}(T)$ and all $\omega \in \Omega$, then we say that $\mathcal{H}_U$ has *no causal effect on A given $\mathcal{H}_V$*, or that $\mathcal{H}_U$ is *non-causal to A given $\mathcal{H}_V$*.

We say that $\mathcal{H}_U$ has *no causal effect on $\mathcal{F}$ given $\mathcal{H}_V$*, or that $\mathcal{H}_U$ is *non-causal to $\mathcal{F}$ given $\mathcal{H}_V$*, if, for all $A \in \mathcal{F}$, $\mathcal{H}_U$ has no causal effect on A given $\mathcal{H}_V$.

(ii) If there exists $\omega \in \Omega$ such that $K_{U\cup V}(\omega, A) \neq K_V(\omega, A)$, then we say that $\mathcal{H}_U$ has an *active causal effect on A given $\mathcal{H}_V$*, or that $\mathcal{H}_U$ is *actively causal to A given $\mathcal{H}_V$*.

We say that $\mathcal{H}_U$ has an *active causal effect on $\mathcal{F}$ given $\mathcal{H}_V$*, or that $\mathcal{H}_U$ is *actively causal to $\mathcal{F}$ given $\mathcal{H}_V$*, if $\mathcal{H}_U$ has an active causal effect on some $A \in \mathcal{F}$.

(iii) Otherwise, we say that $\mathcal{H}_U$ has a *dormant causal effect on A given $\mathcal{H}_V$*, or that $\mathcal{H}_U$ is *dormantly causal to A given $\mathcal{H}_V$*.

We say that $\mathcal{H}_U$ has a *dormant causal effect on $\mathcal{F}$ given $\mathcal{H}_V$*, or that $\mathcal{H}_U$ is *dormantly causal to $\mathcal{F}$ given $\mathcal{H}_V$*, if $\mathcal{H}_U$ does not have an active causal effect on any event in $\mathcal{F}$ given $\mathcal{H}_V$ and there exists $A \in \mathcal{F}$ on which $\mathcal{H}_U$ has a dormant causal effect given $\mathcal{H}_V$.

Sometimes, we will say that $\mathcal{H}_U$ has a *causal effect on A given $\mathcal{H}_V$* to mean that $\mathcal{H}_U$ has either an active or a dormant causal effect on A given $\mathcal{H}_V$.

The intuition is as follows. For any fixed $S \in \mathcal{P}(T)$ and any fixed event $A \in \mathcal{H}$. consider the function $\omega_{S\cup V} \mapsto K_{S\cup V}((\omega_{(S\cup V)\setminus(U\setminus V)}, \omega_{S\cap(U\setminus V)}), A)$. If $\mathcal{H}_U$ has no causal effect on A given $\mathcal{H}_V$, then it means that the causal kernel does not depend on the $\omega_{S\cap(U\setminus V)}$ component of $\omega_{S\cup V}$; in other words, $\mathcal{H}_U$ only has an influence on A through its V component.

We collect some important special cases in the following remark.

Remark 1.3.5. (a) Letting $V = U$, we always have

$$K_{S\cup U}(\omega, A) = K_{(S\cup U)\setminus(U\setminus U)}(\omega, A) = K_{S\cup U}(\omega, A)$$

for all $\omega \in \Omega$ and $A \in \mathcal{H}$, which means that $\mathcal{H}_U$ has no causal effect on any event $A \in \mathcal{H}$ given itself.

(b) If $\mathcal{H}_U$ has no causal effect on A given $\mathcal{H}_V$, then letting $U = S$ in Definition 1.3.4(i), we see that, for all $\omega \in \Omega$,

$$K_{U\cup V}(\omega, A) = K_V(\omega, A).$$

In particular, this means that $\mathcal{H}_U$ cannot have both no causal effect and active causal effect on A given $\mathcal{H}_V$.

(c) The case $V = \emptyset$ reduces Definition 1.3.4 to Definition 1.3.1, i.e. $\mathcal{H}_U$ having no causal effect in the sense of Definition 1.3.1 is the same as $\mathcal{H}_U$ having no causal effect given $\{\emptyset, \Omega\}$ in the sense of Definition 1.3.4, etc.

(d) It is possible for $\mathcal{H}_U$ to be causal to an event A, and for there to exist $V \in \mathcal{P}(T)$ such that $\mathcal{H}_U$ has no causal effect on A given $\mathcal{H}_V$. However, if $\mathcal{H}_U$ has no causal effect on A, then for any $V \in \mathcal{P}(T)$, $\mathcal{H}_U$ has no causal effect on A given $\mathcal{H}_V$. To see this, note that Remark 1.3.2(e) tells us that $U \setminus V$ also does not have any causal effect on A. Then given any $S \in \mathcal{P}(T)$,

$$K_{S\cup V}(\omega, A) = K_{(S\cup V)\setminus(U\setminus V)}(\omega, A),$$

applying Definition 1.3.1(i) to $S \cup V$. Since $S \in \mathcal{P}(T)$ was arbitrary, $\mathcal{H}_U$ has no causal effect on A given $\mathcal{H}_V$.

1.4 Comparison with Existing Frameworks

In this section, we show how causal spaces can encode the interventional aspects of the two most widely-used frameworks of causality, i.e. structural causal models and the potential outcomes.

1.4.1 Structural Causal Models (SCMs)

Consider an SCM in its most basic form, given in the following definition.

Definition 1.4.1 ((Peters et al., 2017, p.83, Definition 6.2))**.** A structural causal model $\mathfrak{C} = (\mathbf{S}, \tilde{\mathbb{P}})$ consists of a collection $\mathbf{S}$ of d (structural) assignments $X_j := f_j(\mathbf{PA}_j, N_j), j = 1, ..., d$, where $\mathbf{PA}_j \subseteq \{X_1, ..., X_d\}\setminus\{X_j\}$ are called the *parents of* X_j and N_j are the *noise* variables; and a distribution $\tilde{\mathbb{P}}$ over the noise variables such that they are jointly independent.

The graph $\mathcal{G}$ of an SCM is obtained by creating one vertex for each X_j and drawing directed edges from each parent in $\mathbf{PA}_j$ to X_j. This graph is assumed to be acyclic.

Below, we show that a unique causal space that corresponds to such an SCM can be constructed.

First, we let the variables $X_j, j = 1, ..., d$ take values in measurable spaces $(E_j, \mathcal{E}_j)$ respectively, and let $(\Omega, \mathcal{H}) = \otimes_{j=1}^d (E_j, \mathcal{E}_j)$. An SCM $\mathfrak{C}$ entails a unique distribution $\mathbb{P}$ over the variables $\mathbf{X} = (X_1, ..., X_d)$ by the propagation of the noise distribution $\tilde{\mathbb{P}}$ through the structural equations f_j (Peters et al., 2017, p.84, Proposition 6.3), and we take this $\mathbb{P}$ as the observational measure of the causal space. More precisely, assuming $\{1, ..., d\}$ is a topological ordering, we have, for $A_j \in \mathcal{E}_j, j = 1, ..., d$,

$$\begin{aligned}
&\mathbb{P}(A_1 \times E_2 \times ... \times E_d) = \tilde{\mathbb{P}}(\{n_1 : f_1(n_1) \in A_1\})\\
&\mathbb{P}(A_1 \times A_2 \times E_3 \times ... \times E_d)\\
&\qquad = \tilde{\mathbb{P}}(\{(n_1, n_2) : (f_1(n_1), f_2(f_1(n_1), n_2)) \in A_1 \times A_2\})\\
&\qquad \vdots\\
&\mathbb{P}(A_1 \times ... \times A_d)\\
&\qquad = \tilde{\mathbb{P}}(\{(n_1, ..., n_d) : (f_1(n_1), ..., f_d(f_1(n_1), ..., n_d)) \in A_1 \times ... \times A_d\}).
\end{aligned}$$

Finally, for each $S \in \mathcal{P}(\{1, ..., d\})$ and for each $\omega \in \Omega$, define $f_j^{S,\omega} = f_j$ if $j \notin S$ and $f_j^{S,\omega} = \omega_j$ if $j \in S$. Then we have

$$\begin{aligned}
&K_S(\omega, A_1 \times ... \times A_d)\\
&\qquad = \tilde{\mathbb{P}}(\{(n_1, ..., n_d) : (f_1^{S,\omega}(n_1), ..., f_d^{S,\omega}(f_1^{S,\omega}(n_1), ..., n_d)) \in A_1 \times ... \times A_d\}).
\end{aligned}$$

This uniquely specifies the causal space $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$ that corresponds to the SCM $\mathfrak{C}$. While this shows that causal spaces strictly generalise (interventional aspects of) SCMs, there are fundamental philosophical differences between the two approaches, as highlighted in the following remark.

Remark 1.4.2. (i) The "system" in an SCM can be viewed as the collection of all variables $X_1, ..., X_d$, and the "subsystems" the individual variables or the groups of variables. Each *structural equation* f_j encodes how the whole system, when intervened on, affects a subsystem X_j, i.e. how the collection of all other variables affects the individual variables (even though, in the end, the equations only depend on the parents). This way of encoding causal effects seems somewhat inconsistent with the philosophy laid out in the Introduction, that we are interested in what happens to the "system" when we intervene on a "subsystem". It also seems inconsistent with the actual action taken, which is to intervene on subsystems, not the whole system, or the parents of a particular variable.

In contrast, the causal kernels encode exactly what happens to the whole system, i.e. what measure we get on the whole measurable space $(\Omega, \mathcal{H})$, when we intervene on a "subsystem", i.e. put a desired measure on a sub-σ-algebra of $\mathcal{H}$[2].

(ii) The primitive objects of SCMs are the variables X_j, the structural equations f_j and the distribution $P_{\mathbf{N}}$ over the noise variables. The observational distribution, as well as the interventional distributions, are derived

[2] In this sense, some philosophy is shared with *generalised structural equation models (GSEMs)* (Peters and Halpern, 2021).

from these objects. It turns out that unique existence of observational and interventional distributions are not guaranteed, and can only be shown under the acyclicity assumption or rather stringent and hard-to-verify conditions on the structural equations and the noise distributions (Bongers et al., 2021). Moreover, it means that the observational and interventional distributions are not decoupled, and rather are linked through the structural equations f_j, and as a result, it is not possible to encode the full range of observational and interventional distributions using just the variables of interest (see Example 1.5.1).

In contrast, in causal spaces, the observational distribution $\mathbb{P}$, as well as the interventional distributions (via the causal kernels), are the primitive objects. Not only does this automatically guarantee their unique existence, but it also allows the interventional distributions (i.e. the causal information) to be completely decoupled from the observational distribution.

(iii) Galles and Pearl (1998, Section 3) propose three axioms of counterfactuals based on SCMs (called *causal models* in that paper), namely, composition, effectiveness and reversibility. Even though these three concepts can be carried over to causal spaces, the mathematics through which they are represented needs to be adapted, since the tools that are used in causal spaces are different from those used in causal models of Galles and Pearl (1998). In particular, we work directly with measures as the primitive objects, whereas Galles and Pearl (1998) use the structural equations as the primitive objects, and the probabilities only enter through a measure on the exogenous variables. Thus, the three properties can be phrased in the causal space language as follows:

Composition For $S, R \subseteq T$, denote by $\mathbb{Q}'$ the measure on $\mathcal{H}_{S\cup R}$ obtained by restricting $\mathbb{P}^{\text{do}(S,\mathbb{Q})}$. Then $\mathbb{P}^{\text{do}(S,\mathbb{Q})} = \mathbb{P}^{\text{do}(S\cup R,\mathbb{Q}')}$. In words, intervening on $\mathcal{H}_S$ via the measure $\mathbb{Q}$ is the same as intervening on $\mathcal{H}_{S\cup R}$ via the measure that it would have if we intervened on $\mathcal{H}_S$ via $\mathbb{Q}$.

This is not in general true. A counterexample can be demonstrated with a simple SCM, where X_1, X_2 and X_3 causally affect Y, in a way that depends not only on the marginal distributions of X_1, X_2 and X_3 but their joint distribution, and X_1, X_2 and X_3 have no causal relationships among them. Then intervening on X_1 with some measure $\mathbb{Q}$ cannot be the same as intervening on X_1 and X_2 with $\mathbb{Q}\otimes\mathbb{P}$, since such an intervention would change the joint distribution of X_1, X_2 and X_3, even if we give them the same marginal distributions.

Effectiveness For $S \subseteq R \subseteq T$, if we intervene on $\mathcal{H}_R$ via a measure $\mathbb{Q}$, then $\mathcal{H}_S$ has measure $\mathbb{Q}$ restricted to $\mathcal{H}_S$.

This is indeed guaranteed by interventional determinism (Definition 1.2.1(ii)), so effectiveness continues to hold in causal spaces.

Reversibility For $S, R, U \subseteq T$, let $\mathbb{Q}$ be some measure on $\mathcal{H}_S$, and $\mathbb{Q}_1$ and $\mathbb{Q}_2$ be measures on $\mathcal{H}_{S\cup R}$ and $\mathcal{H}_{S\cup U}$ respectively such that they coincide with $\mathbb{Q}$ when restricted to $\mathcal{H}_S$. Then if $\mathbb{P}^{\text{do}(S\cup R,\mathbb{Q}_1)}(B) = \mathbb{Q}_2(B)$ for all $B \in \mathcal{H}_U$ and if $\mathbb{P}^{\text{do}(S\cup U,\mathbb{Q}_2)}(C) = \mathbb{Q}_1(C)$ for all $C \in \mathcal{H}_R$, then $\mathbb{P}^{\text{do}(S,\mathbb{Q})}(A) = \mathbb{Q}_1(A)$ for all $A \in \mathcal{H}_R$.

This does not hold in general in causal spaces; in fact, Example 1.5.2 is a counterexample of this, with $S = \emptyset$.

1.4.2 Potential Outcomes (PO) Framework

In the PO framework, the treatment and outcome variables of interest are fixed in advance. Although much of the literature begins with individual units, these units are in the end i.i.d. copies of random variables under the stable unit treatment value assumption (SUTVA), and that is how we begin.

Denote by $(\tilde{\Omega}, \tilde{\mathcal{H}}, \tilde{\mathbb{P}})$ the underlying probability space. Let $Z : \tilde{\Omega} \to \mathcal{Z}$ be the "treatment" variable, taking values in a measurable space $(\mathcal{Z}, \mathfrak{Z})$. Then for each value z of the treatment, there is a separate random variable $Y_z : \tilde{\Omega} \to \mathcal{Y}$, called the "potential outcome given $Z = z$" taking values in a measurable space $(\mathcal{Y}, \mathfrak{Y})$; we also have the "observed outcome", which is the potential outcome consistent with the treatment, i.e. $Y = Y_Z$. The researcher is interested in quantities such as the "average treatment effect", $\tilde{\mathbb{E}}[Y_{z_1} - Y_{z_2}]$, where $\tilde{\mathbb{E}}$ is the expectation with respect to $\tilde{\mathbb{P}}$, to measure the causal effect of the treatment. Often, there are other, "pre-treatment variables" or "covariates", which we denote by $X : \tilde{\Omega} \to \mathcal{X}$, taking values in a measurable space $(\mathcal{X}, \mathfrak{X})$. Given these, another object of interest is the "conditional average treatment effect", defined as $\tilde{\mathbb{E}}[Y_{z_1} - Y_{z_2} \mid X]$.

It is relatively straightforward to construct a causal space that can carry this framework. We define $\Omega = \mathcal{Z} \times \mathcal{Y} \times \mathcal{X}$ and $\mathcal{H} = \mathfrak{Z} \otimes \mathfrak{Y} \otimes \mathfrak{X}$. We also define $\mathbb{P}$, for each $A \in \mathfrak{Z}$, $B \in \mathfrak{Y}$ and $C \in \mathfrak{X}$, as $\mathbb{P}(A \times B \times C) = \tilde{\mathbb{P}}(Z \in A, Y \in B, X \in C)$. As for causal kernels, we are essentially only interested in $K_Z(z, B)$ for $B \in \mathfrak{Y}$, and we define these to be $K_Z(z, B) = \tilde{\mathbb{P}}(Y_z \in B)$.

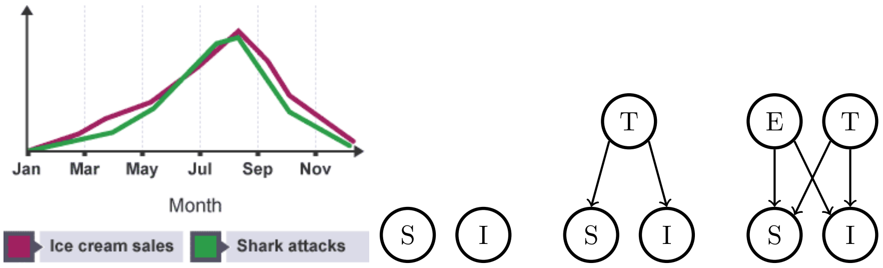
1.5 Examples

In this section, we give a few more concrete constructions of causal spaces. In particular they are designed to highlight cases which are hard to represent with existing frameworks, but which have natural representations in terms of causal spaces. Comparisons are made particularly with SCMs.

1.5.1 Confounders

The following example highlights the fact that, with graphical models, there is no way to encode correlation but no causation between two variables, using just the variables of interest.

Example 1.5.1. *Consider the popular example of monthly ice cream sales and shark attacks in the US (Figure 1.2a), that shows that correlation does not imply causation. This cannot be encoded by an SCM with just two variables as in Figure 1.2b, since no causation means no arrows between the variables,*



(a) (b) (c) (d)

Figure 1.2: Correlation but no causation between ice-cream sales and shark attacks. S stands for the number of shark attacks, I for ice cream sales, T for temperature and E for economy.

which in turn also means no dependence. One needs to add the common causes into the model (whether observed or latent), the most obvious one being the temperature (high temperatures make people desire ice cream more, as well as to go to the beach more), as seen in Figure 1.2c. Now we have a model in which both dependence and no causation are captured. But can we stop here? There are probably other factors that affect both variables, such as the economy (the better the economic situation, the more likely people are to buy ice cream, and to take beach holidays) – see Figure 1.2d. Not only is the model starting to lose parsimony, but as soon as we stop adding variables to the model, we are making an assumption that there are no further confounding variables out there in the world[3].

In contrast, causal spaces allow us to model any observational and causal relationships with just the variables that we were interested in, without any restrictions or the need to add more variables. In this particular example, we would take as our causal space $(E_1 \times E_2, \mathcal{E}_1 \otimes \mathcal{E}_2, \mathbb{P}, \mathbb{K})$*, where* $E_1 = E_2 = \mathbb{R}$ *with values in* E_1 *and* E_2 *corresponding to ice cream sales and shark attacks respectively, and* $\mathcal{E}_1 = \mathcal{E}_2$ *being Lebesgue* σ*-algebras. Then we can let* $\mathbb{P}$ *be a measure that has a strong dependence between any* $A \in \mathcal{E}_1$ *and* $B \in \mathcal{E}_2$*, but let the causal kernels be* $K_1(x, B) = \mathbb{P}(B)$ *for any* $x \in E_1$ *and* $B \in \mathcal{E}_2$*, and likewise* $K_2(x, A) = \mathbb{P}(A)$ *for any* $x \in E_2$ *and* $A \in \mathcal{E}_1$*.*

Nancy Cartwright argued against the completeness of causal Markov condition, using an example of two factories (Cartwright, 1999, p.108), in which there may not even be any confounders between dependent variables, not even an unobserved one. If we accept her position, then there are situations which SCMs would not be able to represent, whereas causal spaces would have no problems at all.

[3] One solution could be to add a single "variable" that collects all of the confounders into one, but then the numerical value of this "variable", as well as its distribution and the structural equations from this "variable" into S and I, would be completely meaningless.

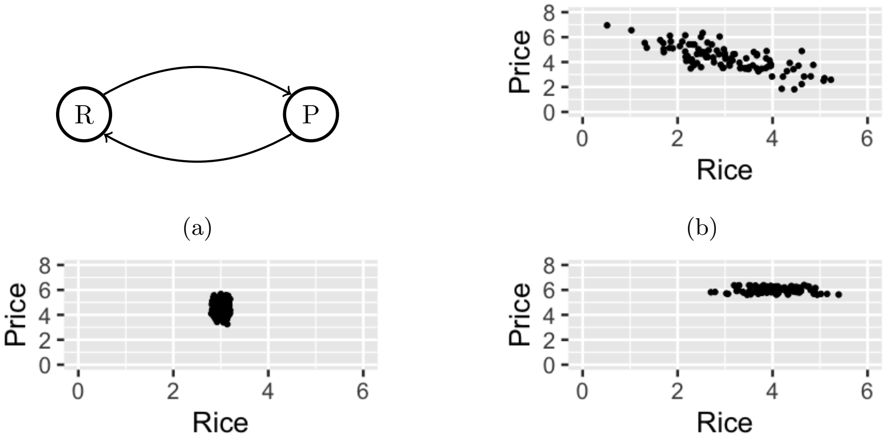
1.5.2 Cycles

As mentioned before, cycles in SCMs cause serious problems, namely that observational and interventional distributions that are consistent with the given structural equations and noise distribution may not exist, and when they do, they may not exist uniquely. This is an artefact of the fact that these distributions are derived from the structural equations rather than taken as the primitive objects. In the vast majority of the cases, cycles are excluded from consideration from the beginning and only directed acyclic graphs (DAGs) are considered. Some works study the *solvability* of cyclic SCMs (Halpern, 2000; Bongers et al., 2021), where the authors investigate under what conditions on the structural equations and the noise variables there exist random variables and distributions that *solve* the given structural equations, and if so, when that happens *uniquely*. Other works have allowed cycles to exist, but restricted the definition of an SCM only to those that have a unique solution (Halpern, 2000; Pearl, 2009; Rubenstein et al., 2017).

Of course, cyclic causal relationships abound in the real world. In our proposed causal space, given two sub-σ-algebras $\mathscr{H}_S$ and $\mathscr{H}_U$ of $\mathscr{H}$, nothing stops both of them from having a causal effect on the other (see Definition 1.3.1 for a precise definition of causal effects), but we are still guaranteed to have a unique causal space, both before intervention and after intervention on either $\mathscr{H}_S$ or $\mathscr{H}_U$. The following is an example of a situation with "cyclic" causal relationship.

Example 1.5.2. *We want to model the relationship between the amount of rice in the market and its price per kg. Take as the probability space* $(E_1 \times E_2, \mathcal{E}_1 \otimes \mathcal{E}_2, \mathbb{P})$*, where* $E_1 = E_2 = \mathbb{R}$ *with values in* E_1 *and* E_2 *representing the amount of rice in the market in million tonnes and the price of rice per kg in KRW respectively,* $\mathcal{E}_1, \mathcal{E}_2$ *are Lebesgue* σ*-algebras and* $\mathbb{P}$ *is for simplicity taken to be jointly Gaussian. Without any intervention, the higher the yield, the more rice there is in the market and lower the price, as in Figure 1.3b. If the government intervenes on the market by buying up extra rice or releasing rice into the market from its stock, with the goal of stabilising supply at 3 million tonnes, then the price will stabilise accordingly, say with Gaussian distribution with mean 4.5 and standard deviation 0.5, as in Figure 1.3c. The corresponding causal kernel will be* $K_1(3, x) = \frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-4.5}{0.5})^2}$*. On the other hand, if the government fixes the price of rice at a price, say at 6,000 KRW per kg, then the farmers will be incentivised to produce more, say with Gaussian distribution with mean 4 and standard deviation 0.5, as in Figure 1.3d. The corresponding causal kernel will be* $K_2(6, y) = \frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y-4}{0.5})^2}$*.*

Causal spaces treat causal effects really as what happens after an intervention takes place, and with this approach, cycles can be rather naturally encoded, as shown above. We do not view cyclic causal relationships as an equilibrium of a dynamical system, or require it to be viewed as an acyclic stochastic process, as done by some authors (Peters et al., 2017, p.85, Remark 6.5).



(a)

(b)

(c)

(d)

Figure 1.3: Rice in the market in million tonnes and price per kg in KRW.

1.5.3 Continuous-time Stochastic Processes, and Parents

A very well established sub-field of probability theory is the field of stochastic processes, in which the index set representing (most often) time can be either discrete or continuous, and in both cases, infinite. However, most causal models start by assuming a finite number of variables, which immediately rules out considering stochastic processes, and efforts to extend to infinite number of variables usually consider only discrete time steps (Peters et al., 2017, Chapter 10) or dynamical systems (Bongers et al., 2018; Peters et al., 2022; Blom et al., 2020; Rubenstein et al., 2018). Since probability spaces have proven to accommodate continuous time stochastic processes in a natural way, it is natural to believe that causal spaces, being built up from probability spaces, should be able to enable the incorporation of the concept of causality in the theory of stochastic processes.

Let W be a totally-ordered set, in particular $W = \mathbb{N} = \{0, 1, ...\}$, $W = \mathbb{Z} = \{..., -2, -1, 0, 1, 2, ...\}$, $W = \mathbb{R}_+ = [0, \infty)$ or $W = \mathbb{R} = (-\infty, \infty)$ considered as the time set. Then, we consider causal spaces of the form $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t\in T} E_t, \otimes_{t\in T} \mathcal{E}_t, \mathbb{P}, \mathbb{K})$, where the index set T can be written as $T = W \times \tilde{T}$ for some other index set $\tilde{T}$. The following notion captures the intuition that causation can only go forwards in time.

Definition 1.5.3. Let $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t\in T} E_t, \otimes_{t\in T} \mathcal{E}_t, \mathbb{P}, \mathbb{K})$ be a causal space, where the index set T can be written as $T = W \times \tilde{T}$, with W representing time. Then we say that the causal mechanism $\mathbb{K}$ *respects time*, or that $\mathbb{K}$ is a *time-respecting causal mechanism*, if, for all $w_1, w_2 \in W$ with $w_1 < w_2$, we have that $\mathcal{H}_{w_2 \times \tilde{T}}$ has no causal effect (in the sense of Definition 1.3.1) on $\mathcal{H}_{w_1 \times \tilde{T}}$.

In a causal space where the index set T has a time component, the fact that

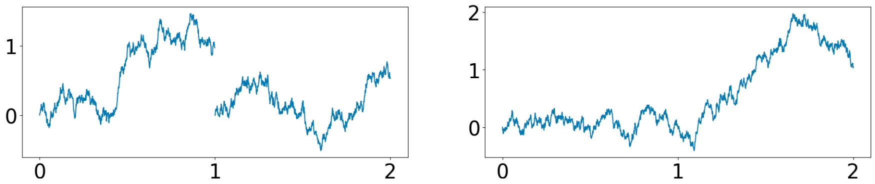


Figure 1.4: 1-dimensional Brownian motion, intervened and conditioned to have value 0 at time 1.

causal mechanism $\mathbb{K}$ respects time means that the past can affect the future, but the future cannot affect the past. This already distinguishes itself from the concept of conditioning – conditioning on the future does have implications for past events. We illustrate this point in the example of a Brownian motion.

Example 1.5.4. *Take $(\times_{t\in\mathbb{R}_+}E_t, \otimes_{t\in\mathbb{R}_+}\mathcal{E}_t, \mathbb{P}, \mathbb{K})$, where, for each $t \in \mathbb{R}_+$, $E_t = \mathbb{R}$ and $\mathcal{E}_t$ is the Lebesgue σ-algebra, and $\mathbb{P}$ is the Wiener measure. For $s < t$, we have causal kernels $K_s(x, y) = \frac{1}{\sqrt{2\pi(t-s)}}e^{-\frac{1}{2(t-s)}(y-x)^2}$ and $K_t(x, y) = \frac{1}{\sqrt{2\pi s}}e^{-\frac{1}{2s}y^2}$. The former says that, if we intervene by setting the value of the process to x at time s, then the process starts again from x, whereas the latter says that if we intervene at time t, the past values at time s are not affected. On the left-hand plot of Figure 1.4, we set the value of the process at time 1 to 0. The past values of the process are not affected, and there is a discontinuity at time 1 where the process starts again from 0. Contrast this to the right-hand plot, where we condition on the process having value 0 at time 1. This does affect past values, and creates a Brownian bridge from time 0 to time 1.*

Note, Brownian motion is not differentiable, so no approach based on dynamical systems is applicable.

Remark 1.5.5. The concept of *parents* is central in SCMs – the structural equations are defined on the parents of each variable. However, continuous time is dense, so given two distinct points in time, there is always a time point in between. Suppose we have a one-dimensional continuous time Markov process $(X_t)_{t\in\mathbb{R}}$ (Çınlar, 2011, p.169), and a point t_0 in time. Then for any $t < t_0$, X_t has a causal effect on X_{t_0}, but there always exists some t' with $t < t' < t_0$ such that conditioned on $X_{t'}$, X_t does not have a causal effect on X_{t_0}, meaning X_t cannot be a parent of X_{t_0}. In such a case, X_{t_0} cannot be said to have any parents, and hence no corresponding SCM can be defined.

1.6 Interventions

In this section, we provide a few more definitions and results related to the notion of interventions, introduced in Definition 1.2.2.

Remark 1.6.1. First, we make a few remarks on how the intervention causal kernels $K_S^{\text{do}(U,\mathbb{Q},\mathbb{L})}$ behave in some special cases, depending on the relationship between U and S.

(a) For $S \in \mathcal{P}(T)$ with $U \subseteq S$, we have, for all $\omega \in \Omega$ and all $A \in \mathcal{H}$,

$$\begin{aligned} K_S^{\text{do}(U,\mathbb{Q},\mathbb{L})}(\omega, A) &= \int L_U(\omega_U, d\omega'_U) K_S((\omega_{S\setminus U}, \omega'_U), A) \\ &= \int \delta_{\omega_U}(d\omega'_U) K_S((\omega_{S\setminus U}, \omega'_U), A) \\ &= K_S((\omega_{S\setminus U}, \omega_U), A) \\ &= K_S(\omega, A). \end{aligned}$$

This means that, after an intervention on $\mathcal{H}_U$, subsequent interventions on $\mathcal{H}_S$ with $\mathcal{H}_U \subseteq \mathcal{H}_S$ simply overwrite the original intervention. Note that this is reminiscent of the “partial ordering on the set of interventions” in (Rubenstein et al., 2017), but in our setting, this is given by the partial ordering induced by the inclusion structure of sub-σ-algebras of $\mathcal{H}$.

(b) For $S \in \mathcal{P}(T)$ with $S \subseteq U$,

$$K_S^{\text{do}(U,\mathbb{Q},\mathbb{L})}(\omega, A) = \int L_S(\omega_S, d\omega'_U) K_U(\omega'_U, A)$$

for all $\omega \in \Omega$ and $A \in \mathcal{H}$, i.e. $K_S^{\text{do}(U,\mathbb{Q},\mathbb{L})}$ is a product of the two kernels K_U and L_S (Çınlar, 2011, p.39); in particular, $K_S^{\text{do}(U,\mathbb{Q},\mathbb{L})}(\omega, A) = L_S(\omega, A)$ for all $A \in \mathcal{H}_U$.

(c) For $S \in \mathcal{P}(T)$ with $S \cap U = \emptyset$,

$$\begin{aligned} K_S^{\text{do}(U,\mathbb{Q},\mathbb{L})}(\omega, A) &= \int L_\emptyset(\omega_\emptyset, d\omega'_U) K_{S\cup U}((\omega_S, \omega'_U), A) \\ &= \int \mathbb{Q}(d\omega'_U) K_{S\cup U}((\omega_S, \omega'_U), A) \end{aligned}$$

for all $\omega \in \Omega$ and $A \in \mathcal{H}$, i.e. the effect of intervening on $\mathcal{H}_U$ with $\mathbb{Q}$ then $\mathcal{H}_S$ is the same as intervening on $\mathcal{H}_{U\cup S}$ with a product measure of $\mathbb{Q}$ on $\mathcal{H}_U$ and whatever measure we place on $\mathcal{H}_S$.

We give it a name for the special case in which the internal causal kernels are all trivial (see Definition 1.3.3).

Definition 1.6.2. Let $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t\in T} E_t, \otimes_{t\in T} \mathcal{E}_t, \mathbb{P}, \mathbb{K})$ be a causal space, and $U \in \mathcal{P}(T)$ and $\mathbb{Q}$ a probability measure on $(\Omega, \mathcal{H}_U)$. A *hard intervention on $\mathcal{H}_U$ via $\mathbb{Q}$* is a new causal space $(\Omega, \mathcal{H}, \mathbb{P}^{\text{do}(U,\mathbb{Q})}, \mathbb{K}^{\text{do}(U,\mathbb{Q},\text{hard})})$, where the intervention measure $\mathbb{P}^{\text{do}(U,\mathbb{Q})}$ is a probability measure $(\Omega, \mathcal{H})$ defined in the same way as in Definition 1.2.2, and the intervention causal mechanism $\mathbb{K}^{\text{do}(U,\mathbb{Q},\text{hard})} = \{K_S^{\text{do}(U,\mathbb{Q},\text{hard})} : S \in \mathcal{P}(T)\}$ consists of causal kernels that are obtained from the intervention causal kernels in Definition 1.2.2 in which $L_{S\cap U}$ is a trivial causal kernel, i.e. one that has no causal effect on $\mathcal{H}_{U\setminus S}$.

From the discussion following Definition 1.3.3, we have that, for $A \in \mathcal{H}_{S\cap U}$ and $B \in \mathcal{H}_{U\setminus S}$, $L_{S\cap U}(\omega, A \times B) = 1_A(\omega_{S\cap U})\mathbb{Q}(B)$.

The next result gives an explicit expression for the causal kernels obtained after a hard intervention.

Theorem 1.6.3. *Let $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t\in T}E_t, \otimes_{t\in T}\mathcal{E}_t, \mathbb{P}, \mathbb{K})$ be a causal space, and $U \in \mathcal{P}(T)$ and $\mathbb{Q}$ a probability measure on $(\Omega, \mathcal{H}_U)$. Then after a hard intervention on $\mathcal{H}_U$ via $\mathbb{Q}$, the intervention causal kernels $K_S^{\mathrm{do}(U,\mathbb{Q},\mathrm{hard})}$ are given by*

$$\begin{aligned} K_S^{\mathrm{do}(U,\mathbb{Q},\mathrm{hard})}(\omega, A) &= K_S^{\mathrm{do}(U,\mathbb{Q},\mathrm{hard})}(\omega_S, A) \\ &= \int \mathbb{Q}(d\omega'_{U\setminus S})K_{S\cup U}((\omega_S, \omega'_{U\setminus S}), A). \end{aligned}$$

Intuitively, hard interventions do not encode any internal causal relationships within $\mathcal{H}_U$, so after we subsequently intervene on $\mathcal{H}_S$, the measure $\mathbb{Q}$ that we originally imposed on $\mathcal{H}_U$ remains on $\mathcal{H}_{U\setminus S}$.

The following lemma contains a couple of results about particular sub-σ-algebras having no causal effects on particular events in the intervention causal space, regardless of the measure and causal mechanism that was used for the intervention.

Lemma 1.6.4. *Let $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t\in T}E_t, \otimes_{t\in T}\mathcal{E}_t, \mathbb{P}, \mathbb{K})$ be a causal space, and $U \in \mathcal{P}(T)$, $\mathbb{Q}$ a probability measure on $(\Omega, \mathcal{H}_U)$ and $\mathbb{L} = \{L_V : V \in \mathcal{P}(U)\}$ a causal mechanism on $(\Omega, \mathcal{H}_U, \mathbb{Q})$. Suppose we intervene on $\mathcal{H}_U$ via $(\mathbb{Q}, \mathbb{L})$.*

(i) For $A \in \mathcal{H}_U$ and $V \in \mathcal{P}(T)$ with $V \cap U = \emptyset$, $\mathcal{H}_V$ has no causal effect on A in the intervention causal space $(\Omega, \mathcal{H}, \mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}, \mathbb{K}^{\mathrm{do}(U,\mathbb{Q},\mathbb{L})})$, i.e. events in the σ-algebra $\mathcal{H}_U$ on which intervention took place are not causally affected by σ-algebras outside $\mathcal{H}_U$.

(ii) Again, let $V \in \mathcal{P}(T)$ with $V \cap U = \emptyset$, and also let $A \in \mathcal{H}$ be any event. If, in the original causal space, $\mathcal{H}_V$ had no causal effect on A, then in the intervention causal space, $\mathcal{H}_V$ has no causal effect on A either.

(iii) Now let $V \in \mathcal{P}(T)$, $A \in \mathcal{H}$ any event and suppose that the intervention on $\mathcal{H}_U$ via $\mathbb{Q}$ is hard. Then if $\mathcal{H}_V$ had no causal effect on A in the original causal space, then $\mathcal{H}_V$ has no causal effect on A in the intervention causal space either.

Lemma 1.6.4(ii) and (iii) tell us that, if $\mathcal{H}_V$ had no causal effect on A in the original causal space, then by intervening on $\mathcal{H}_U$ with $V \cap U = \emptyset$ or by any hard intervention, we cannot create a causal effect from $\mathcal{H}_v$ on A. However, by intervening on a sub-σ-algebra that contains both $\mathcal{H}_V$ and (a part of) A, and manipulating the internal causal mechanism $\mathbb{L}$ appropriately, it is clear that we can create a causal effect from $\mathcal{H}_V$.

The next result tells us that if a sub-σ-algebra $\mathcal{H}_U$ has a dormant causal effect on an event A, then there is a sub-σ-algebra of $\mathcal{H}_U$ and a hard intervention after which that sub-σ-algebra has an active causal effect on A.

Lemma 1.6.5. *Let $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t \in T} E_t, \otimes_{t \in T} \mathcal{E}_t, \mathbb{P}, \mathbb{K})$ be a causal space, and $U \in \mathcal{P}(T)$. For an event $A \in \mathcal{H}$, if $\mathcal{H}_U$ has a dormant causal effect on A in the original causal space, then there exists a hard intervention and a subset $V \subseteq U$ such that in the intervention causal space, $\mathcal{H}_V$ has an active causal effect on A.*

The next result is about what happens to a causal effect of a sub-σ-algebra that has no causal effect on an event conditioned on another sub-σ-algebra, after intervening on that sub-σ-algebra.

Lemma 1.6.6. *Let $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t \in T} E_t, \otimes_{t \in T} \mathcal{E}_t, \mathbb{P}, \mathbb{K})$ be a causal space, and $U, V \in \mathcal{P}(T)$. For an event $A \in \mathcal{H}$, suppose that $\mathcal{H}_U$ has no causal effect on A given $\mathcal{H}_V$ (see Definition 1.3.4). Then after an intervention on $\mathcal{H}_V$ via any $(\mathbb{Q}, \mathbb{L})$, $\mathcal{H}_{U \setminus V}$ has no causal effect on A.*

The next result shows that, under a hard intervention, a time-respecting causal mechanism stays time-respecting.

Theorem 1.6.7. *Let $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t \in T} E_t, \otimes_{t \in T} \mathcal{E}_t, \mathbb{P}, \mathbb{K})$ be a causal space, where the index set T can be written as $T = W \times \tilde{T}$, with W representing time and $\mathbb{K}$ respecting time. Take any $U \in \mathcal{P}(T)$ and any probability measure $\mathbb{Q}$ on $\mathcal{H}_U$. Then the intervention causal mechanism $\mathbb{K}^{\text{do}(U, \mathbb{Q}, \text{hard})}$ also respects time.*

1.7 Sources

In causal spaces, the observational distribution $\mathbb{P}$ and the causal mechanism $\mathbb{K}$ are completely decoupled. In Section 1.4.1, we give a detailed argument as to why this is desirable, but of course, there is no doubt that the special case in which the causal kernels coincide with conditional measures with respect to $\mathbb{P}$ is worth studying. To that end, we introduce the notion of *sources*.

Definition 1.7.1. Let $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t \in T} E_t, \otimes_{t \in T} \mathcal{E}_t, \mathbb{P}, \mathbb{K})$ be a causal space, $U \in \mathcal{P}(T)$, $A \in \mathcal{H}$ an event and $\mathcal{F}$ a sub-σ-algebra of $\mathcal{H}$. We say that $\mathcal{H}_U$ is a *(local) source* of A if $K_U(\cdot, A)$ is a version of the conditional probability $\mathbb{P}_{\mathcal{H}_U}(A)$. We say that $\mathcal{H}_U$ is a *(local) source* of $\mathcal{F}$ if $\mathcal{H}_U$ is a source of all $A \in \mathcal{F}$. We say that $\mathcal{H}_U$ is a *global source* of the causal space if $\mathcal{H}_U$ is a source of all $A \in \mathcal{H}$.

Clearly, source σ-algebras are not unique (whether local or global). It is easy to see that $\mathcal{H}_\emptyset = \{\emptyset, \Omega\}$ and $\mathcal{H} = \mathcal{H}_T = \otimes_{t \in T} \mathcal{E}_t$ are global sources, and axiom (ii) of Definition 1.2.1 implies that any $\mathcal{H}_S$ is a local source of any of its sub-σ-algebras, including itself, since, for any $A \in \mathcal{H}_U$, $\mathbb{P}_{\mathcal{H}_U}(A) = 1_A$. Also, a sub-$\sigma$-algebra of a source is not necessarily a source, nor is a σ-algebra that contains a source necessarily a source (whether local or global). In Example 1.2.4 above, altitude is a source of temperature (and hence a global source), since the causal kernel corresponding to temperature coincides with the conditional measure given altitude, but temperature is not a source of altitude.

When we intervene on $\mathcal{H}_U$ (via any $(\mathbb{Q}, \mathbb{L})$), $\mathcal{H}_U$ becomes a global source. This precisely coincides with the "gold standard" that is randomised control trials in causal inference, i.e. the idea that, if we are able to intervene on $\mathcal{H}_U$, then the causal effect of $\mathcal{H}_U$ on any event can be obtained by first intervening on $\mathcal{H}_U$, then considering the conditional distribution on $\mathcal{H}_U$. Next is a theorem showing that when one intervenes on $\mathcal{H}_U$, then $\mathcal{H}_U$ becomes a source.

Theorem 1.7.2. *Let $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t\in T} E_t, \otimes_{t\in T} \mathcal{E}_t, \mathbb{P}, \mathbb{K})$ be a causal space, and let $U \in \mathcal{P}(T)$.*

(i) For any measure $\mathbb{Q}$ on $\mathcal{H}_U$ and any causal mechanism $\mathbb{L}$ on $(\Omega, \mathcal{H}_U, \mathbb{Q})$, the causal kernel $K_U^{\mathrm{do}(U,\mathbb{Q},\mathbb{L})} = K_U$ is a version of $\mathbb{P}_{\mathcal{H}_U}^{\mathrm{do}(U,\mathbb{Q})}$, which means that $\mathcal{H}_U$ is a global source σ-algebra of $(\Omega, \mathcal{H}, \mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}, \mathbb{K}^{\mathrm{do}(U,\mathbb{Q},\mathbb{L})})$.

(ii) Suppose $V \in \mathcal{P}(T)$ with $V \subseteq U$. Suppose that the measure $\mathbb{Q}$ on $(\Omega, \mathcal{H}_U)$ factorises over $\mathcal{H}_V$ and $\mathcal{H}_{U\setminus V}$, i.e. for any $A \in \mathcal{H}_V$ and $B \in \mathcal{H}_{U\setminus V}$, $\mathbb{Q}(A\cap B) = \mathbb{Q}(A)\mathbb{Q}(B)$. Then after a hard intervention on $\mathcal{H}_U$ via $\mathbb{Q}$, the causal kernel $K_V^{\mathrm{do}(U,\mathbb{Q},\mathrm{hard})}$ is a version of $\mathbb{P}_V^{\mathrm{do}(U,\mathbb{Q})}$, which means that $\mathcal{H}_V$ is a global source σ-algebra of $(\Omega, \mathcal{H}, \mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}, \mathbb{K}^{\mathrm{do}(U,\mathbb{Q},\mathrm{hard})})$.

Let $A \in \mathcal{H}$ be an event, and $U \in \mathcal{P}(T)$. By the definition of the intervention measure (Definition 1.2.2), we always have

$$\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(A) = \int \mathbb{Q}(d\omega) K_U(\omega, A),$$

hence $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(A)$ can be written in terms of $\mathbb{P}$ and $\mathbb{Q}$ if $K_U(\omega, A)$ can be written in terms of $\mathbb{P}$. This can be seen to occur in three trivial cases: first, if $\mathcal{H}_U$ is a local source of A (see Definition 1.7.1), in which case $K_U(\omega, A) = \mathbb{P}_{\mathcal{H}_U}(\omega, A)$; secondly, if $\mathcal{H}_U$ has no causal effect on A (see Definition 1.3.1), in which case $K_U(\omega, A) = \mathbb{P}(A)$; and finally, if $A \in \mathcal{H}_U$, in which case, by intervention determinism (Definition 1.2.1(ii), we have $K_U(\omega, A) = 1_A(\omega)$. In the latter case, we do not even have dependence on $\mathbb{P}$. Can we generalise these results?

Lemma 1.7.3. *Let $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t\in T} E_t, \otimes_{t\in T} \mathcal{E}_t, \mathbb{P}, \mathbb{K})$ be a causal space. Let $A \in \mathcal{H}$ be an event, and $U \in \mathcal{P}(T)$. If there exists a sub-σ-algebra $\mathcal{G}$ of $\mathcal{H}$ (not necessarily of the form $\mathcal{H}_V$ for some $V \in \mathcal{P}(T)$) such that*

(i) the conditional probability $\mathbb{P}_{\mathcal{H}_U \vee \mathcal{G}}^{\mathrm{do}(U,\mathbb{Q})}(\cdot, A)$ can be written in terms of $\mathbb{P}$ and $\mathbb{Q}$;

(ii) the causal kernel $K_U(\cdot, B)$ can be written in terms of $\mathbb{P}$ for all $B \in \mathcal{G}$;

then $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(A)$ can be written in terms of $\mathbb{P}$ and $\mathbb{Q}$.

Remark 1.7.4. The three cases discussed in the paragraph above Lemma 1.7.3 are special cases of the Lemma with $\mathcal{G}$ being any sub-σ-algebra of $\mathcal{H}$ with $\{\emptyset, \Omega\} \subseteq \mathcal{G} \subseteq \mathcal{H}_U$. In this case, condition (ii) is trivially satisfied since we have $K_U(\cdot, B) = 1_B(\cdot)$ by intervention determinism (Definition 1.2.1(ii)), and for condition (i), by Theorem 1.7.2(i), we have $\mathbb{P}_{\mathcal{H}_U}^{\mathrm{do}(U,\mathbb{Q})}(\cdot, A) = K_U(\cdot, A)$, which means that the problem reduces to checking if $K_U(\cdot, A)$ can be written in terms of $\mathbb{P}$.

Corollary 1.7.5. *Let* $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t\in T} E_t, \otimes_{t\in T} \mathcal{E}_t, \mathbb{P}, \mathbb{K})$ *be a causal space. Let* $A \in \mathcal{H}$ *be an event, and* $U \in \mathcal{P}(T)$*. If there exists a* $V \in \mathcal{P}(T)$ *such that condition (i) of Lemma 1.7.3 is satisfied with* $\mathcal{G} = \mathcal{H}_V$ *and one of the following conditions is satisfied:*

(a) $\mathcal{H}_U$ *is a local source of* $\mathcal{H}_V$*; or*

(b) $\mathcal{H}_U$ *has no causal effect on* $\mathcal{H}_V$*; or*

(c) $V \subseteq U$*,*

then $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(A)$ *can be written in terms of* $\mathbb{P}$ *and* $\mathbb{Q}$*.*

The above is reminiscent of “valid adjustments” in the context of structural causal models (Peters et al., 2017, p.115, Proposition 6.41), and in fact contains the valid adjustments.

Chapter 2

Operations on Multiple Causal Spaces

The notations and definitions in this Chapter are carried straight over from Chapter 1.

2.1 Product Causal Spaces and Causal Independence

We first give the definition of the product of causal kernels, and the product of causal spaces. This constitutes the simplest way of constructing new causal spaces from existing ones.

Definition 2.1.1. Suppose $\mathcal{C}^1 = (\Omega^1, \mathcal{H}^1, \mathbb{P}^1, \mathbb{K}^1)$ and $\mathcal{C}^2 = (\Omega^2, \mathcal{H}^2, \mathbb{P}^2, \mathbb{K}^2)$ with $\Omega^1 = \times_{t \in T^1} E_t$ and $\Omega^2 = \times_{t \in T^2} E_t$ are two causal spaces. For all $S^1 \subseteq T^1$ and $S^2 \subseteq T^2$, and for a pair of causal kernels $K^1_{S^1} \in \mathbb{K}^1$ and $K^2_{S^2} \in \mathbb{K}^2$, we define the *product causal kernel* $K^1_{S^1} \otimes K^2_{S^2}$, for $\omega = (\omega_1, \omega_2) \in \Omega^1_{S^1} \times \Omega^2_{S^2}$ and events $A_1 \in \mathcal{H}^1$ and $A_2 \in \mathcal{H}^2$, by

$$K^1_{S^1} \otimes K^2_{S^2}(\omega, A_1 \times A_2) = K^1_{S^1}(\omega_1, A_1) K^2_{S^2}(\omega_2, A_2).$$

This can then be extended to all of $\mathcal{H}^1 \otimes \mathcal{H}^2$ since the rectangles $A_1 \times A_2$ with $A_1 \in \mathcal{H}^1$ and $A_2 \in \mathcal{H}^2$ generate $\mathcal{H}^1 \otimes \mathcal{H}^2$. Then we define the *product causal space*

$$\mathcal{C}^1 \otimes \mathcal{C}^2 = (\Omega^1 \times \Omega^2, \mathcal{H}^1 \otimes \mathcal{H}^2, \mathbb{P}^1 \otimes \mathbb{P}^2, \mathbb{K}^1 \otimes \mathbb{K}^2)$$

where the product causal mechanism $\mathbb{K}^1 \otimes \mathbb{K}^2$ is the unique family of kernels of the form $(K^1 \otimes K^2)_{S^1 \cup S^2} = K^1_{S^1} \otimes K^2_{S^2}$ for $S^1 \subseteq T^1$ and $S^2 \subseteq T^2$.

We first check that this procedure indeed produces a valid causal space.

Lemma 2.1.2. *The product causal space $\mathcal{C}^1 \otimes \mathcal{C}^2$ as defined in Definition 2.1.1 is a causal space.*

Note that it is only for the sake of simplicity of presentation that we presented the notion of products only for two probability spaces. Indeed, we can easily extend the definition to arbitrary products of causal kernels and causal spaces, just like it is possible for products of probability spaces.

When we take a product of causal spaces, the corresponding components in the resulting causal space do not have a causal effect on each other, as the following result shows.

Lemma 2.1.3. *Let* $\mathcal{C}^1 = (\Omega^1, \mathcal{H}^1, \mathbb{P}^1, \mathbb{K}^1)$ *and* $\mathcal{C}^2 = (\Omega^2, \mathcal{H}^2, \mathbb{P}^2, \mathbb{K}^2)$ *with* $\Omega^1 = \times_{t \in T^1} E_t$ *and* $\Omega^2 = \times_{t \in T^2} E_t$ *be two causal spaces. Then in* $\mathcal{C}^1 \otimes \mathcal{C}^2$*,*

(i) $\mathcal{H}_{T^1}$ *has no causal effect on* $\mathcal{H}_{T^2}$*, and* $\mathcal{H}_{T^2}$ *has no causal effect on* $\mathcal{H}_{T^1}$*;*

(ii) $\mathcal{H}_{T^1}$ *and* $\mathcal{H}_{T^2}$ *are (local) sources of each other.*

Product causal spaces are analogous to *connected components* in graphical models – see, for example, (Sadeghi and Soo, 2023). When forming a product of causal spaces that each arise from an SCM, then each component in the product would be a connected component, and the components would not have any causal effect on each other.

2.1.1 Causal Independence

Recall that, in probability spaces, two events A and B are *independent* with respect to the measure $\mathbb{P}$ if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, i.e. the probability measure is the product measure. Moreover, two σ-algebras are independent if each pair of events from the two σ-algebras are independent[1]. Similarly, for a sub-σ-algebra $\mathcal{F}$ of $\mathcal{H}$, two events A and B are *conditionally independent given* $\mathcal{F}$ if $\mathbb{P}_{\mathcal{F}}(A \cap B) = \mathbb{P}_{\mathcal{F}}(A)\mathbb{P}_{\mathcal{F}}(B)$ almost surely, and two σ-algebras are conditionally independent given $\mathcal{F}$ if each pair of events from the two σ-algebras are conditionally independent given $\mathcal{F}$.

The concept of (conditional) independence is arguably one of the most important in probability theory. Now we give an analogous definition of *causal independence*.

Definition 2.1.4. Let $\mathcal{C} = (\Omega = \times_{t \in T} E_t, \mathcal{H} = \otimes_{t \in T} \mathcal{E}_t, \mathbb{P}, \mathbb{K})$ be a causal space. Then for $U \subseteq T$, two events $A, B \in \mathcal{H}$ are *causally independent on* $\mathcal{H}_U$ if, for all $\omega \in \Omega$,

$$K_U(\omega, A \cap B) = K_U(\omega, A) K_U(\omega, B).$$

We say that two sub-σ-algebras $\mathcal{F}_1$ and $\mathcal{F}_2$ are *causally independent on* $\mathcal{H}_U$ if each pair of events from $\mathcal{F}_1$ and $\mathcal{F}_2$ are causally independent on $\mathcal{H}_U$.

More generally, we say that a finite collection of σ-algebras $\mathcal{F}_1, ..., \mathcal{F}_n$ are a *causal independency on* $\mathcal{H}_U$ if, for all $A_1 \in \mathcal{F}_1, ..., A_n \in \mathcal{F}_n$, we have

$$K_U(\omega, A_1 \cap ... \cap A_n) = K_U(\omega, A_1) ... K_U(\omega, A_n).$$

[1] Many authors take the view that the notion of independence is truly where probability theory starts, as a distinct theory from measure theory (Çınlar, 2011, p.82, Section II.5).

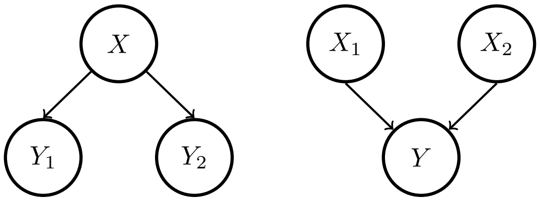


Figure 2.1: Graphs of SCMs in Example 2.1.5.

Moreover, if I is an arbitrary index set, and $\mathcal{F}_i$ is a sub-σ-algebra of $\mathcal{H}$ for each $i \in I$, then the collection $\{\mathcal{F}_i : i \in I\}$ is called a *causal independency given* $\mathcal{H}_U$ if its every finite subset is a causal independency on $\mathcal{H}_U$.

Semantically, causal independence should be interpreted as follows: if A and B are causally independent on $\mathcal{H}_U$, then they are independent once an intervention has been carried out on $\mathcal{H}_U$. Note also that causal independence is really about the causal kernels, and has nothing to do with the probability measure $\mathbb{P}$ of the causal space. Indeed, it is possible for A and B to be causally independent but not probabilistically independent, or causally independent but not conditionally independent, or vice versa. Let us illustrate with the following simple examples.

Example 2.1.5. *We use the language of SCMs, because they are convenient framework that fits into the framework of causal spaces.*

(i) Consider three variables X, Y_1 and Y_2 related through the equations

$$X = N, \quad Y_1 = X + U_1, \quad Y_2 = X + U_2,$$

where N, U_1 and U_2 are standard normal variables (see Figure 2.1 left). We denote by $\mathbb{P}$ their joint distribution on $\mathbb{R}^3$, and we identify this SCM with the causal space $(\mathbb{R}^3, \mathcal{B}(\mathbb{R}^3), \mathbb{P}, \mathbb{K})$[2], where $\mathbb{K}$ is obtained via the above structural equations. Then it is clear to see that Y_1 and Y_2 are causally independent on $\mathcal{H}_X$, since, for every x, and $A, B \in \mathcal{B}(\mathbb{R})$, $K_X(x, \{Y_1 \in A, Y_2 \in B\})$ is bivariate-normally distributed with mean (x, x) and identity covariance matrix, and so

$$K_X(x, \{Y_1 \in A, Y_2 \in B\}) = K_X(x, \{Y_1 \in A\}) K_X(x, \{Y_2 \in B\}).$$

By the same reasoning, Y_1 and Y_2 are conditionally independent given $\mathcal{H}_X$. However, it is clear that they are unconditionally dependent, because they both depend on the value of X.

(ii) Now consider three variables X_1, X_2 and Y related through the equations

$$X_1 = N_1, \quad X_2 = N_2, \quad Y = X_1 + X_2 + U$$

[2] Here, $\mathcal{B}$ represents the Borel σ-algebra.

where N_1, N_2 and U are standard normal variables (see Figure 2.1 right). We denote by $\mathbb{P}$ their joint distribution on $\mathbb{R}^3$, and we identify this SCM with the causal space $(\mathbb{R}^3, \mathcal{B}(\mathbb{R}^3), \mathbb{P}, \mathbb{K})$, where $\mathbb{K}$ is obtained via the above structural equations. Then it is clear that X_1 and X_2 are probabilistically independent. They are also causally independent on $\mathcal{H}_Y$, since, for any $A, B \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned} K_Y(y, \{X_1 \in A, X_2 \in B\}) &= \mathbb{P}(X_1 \in A, X_2 \in B) \\ &= \mathbb{P}(X_1 \in A)\mathbb{P}(X_2 \in B). \end{aligned}$$

However, it is clear that they are conditionally dependent given $\mathcal{H}_Y$.

2.2 Transformations of Causal Spaces

Consider causal spaces $\mathcal{C}^1 = (\Omega^1, \mathcal{H}^1, \mathbb{P}^1, \mathbb{K}^1)$ and $\mathcal{C}^2 = (\Omega^2, \mathcal{H}^2, \mathbb{P}^2, \mathbb{K}^2)$ with $\Omega^1 = \times_{t \in T^1} E_t$ and $\Omega^2 = \times_{t \in T^2} E_t$. We want to define transformations between causal spaces $\mathcal{C}^1$ and $\mathcal{C}^2$. These transformations shall, on the one hand, preserve aspects of the causal structure, i.e., the spaces $\mathcal{C}^1$ and $\mathcal{C}^2$ shall still describe essentially the same system. On the other hand, they shall be flexible so that different types of mappings between causal spaces can be captured.

We focus on transformations that preserve individual variables or combine them in a meaningful way. This relation will be encoded by a map $\rho : T^1 \to T^2$, which can be interpreted as encoding the fact that $S \subseteq T^2$ depends only on the variables indexed by $\rho^{-1}(S)$. Deterministic maps are not sufficiently expressive for our purposes and we therefore focus on stochastic maps, i.e., on probability kernels from measurable spaces $(\Omega^1, \mathcal{H}^1)$ to $(\Omega^2, \mathcal{H}^2)$.

Definition 2.2.1. Suppose that $\kappa : \Omega^1 \times \mathcal{H}^2 \to [0, 1]$ is a probability kernel and $\rho : T^1 \to T^2$ is a map. Then we call the pair (κ, ρ) *admissible* if $\kappa(\cdot, A)$ is $\mathcal{H}^1_{\rho^{-1}(S)}$ measurable for all $S \subset \rho(T^1)$ and $A \in \mathcal{H}^2_S$.

One difference between probability theory and causality seems to be that the latter requires the notion of variables (equivalently a product structure of the underlying space) that define entities that can be intervened upon. For a meaningful relation between two causal spaces, their interventions should be related, which requires some preservation of variables. The definition of admissible maps captures the fact that variables from $\rho^{-1}(S)$ are combined to form a new summary collection of variables indexed by S.

We now require maps between causal spaces to respect the distributional and interventional structure in the following sense.

Definition 2.2.2. A *transformation* of causal spaces, or a *causal transformation*, $\varphi : \mathcal{C}^1 \to \mathcal{C}^2$ is an admissible pair $\varphi = (\kappa, \rho)$ satisfying the following two properties.

(i) The map satisfies *distributional consistency*, i.e., for $A \in \mathcal{H}^2$

$$\int \mathbb{P}^1(d\omega)\kappa(\omega, A) = \mathbb{P}^2(A). \tag{2.1}$$

$$\begin{tikzcd}
\mathcal{C}^1 \arrow[r, "\kappa"] \arrow[d, "K^1_{\rho^{-1}(S)}"] & \mathcal{C}^2 \arrow[d, "K^2_S"'] \\
\mathcal{C}^{1,\mathrm{do}(\rho^{-1}(S))} \arrow[r, "\kappa"] & \mathcal{C}^{2,\mathrm{do}(S)}
\end{tikzcd}$$

Figure 2.2: Interventional Consistency Definition 2.2.2 Equation (2.2) – intervention and transformation commute.

(ii) The map satisfies *interventional consistency*, i.e., for all $A \in \mathcal{H}^2_{\rho(T^1)}$, $S \subset \rho(T^1)$, and $\omega \in \Omega^1$ the following holds

$$\int K^1_{\rho^{-1}(S)}(\omega, d\omega')\kappa(\omega', A) = \int \kappa(\omega, d\omega')K^2_S(\omega', A). \tag{2.2}$$

Interventional consistency requires that interventions and causal transformations commute, i.e., the result of first intervening and then applying the transformation is the same as intervening on the target after the transformation – see Figure 2.2.

We emphasise that in Definition 2.2.1 and 2.2.2 we do not prescribe conditions for added components indexed by $T^2 \setminus \rho(T^1)$.

Further, we remark that as a special case, we can accommodate deterministic maps $f : \Omega_1 \to \Omega_2$ by considering the associated probability kernel $\kappa_f(\omega, A) = \mathbf{1}_A(f(\omega))$. In this case, the admissibility condition reduces to the statement that $\pi_S \circ f$ is measurable with respect to $\mathcal{H}^1_{\rho^{-1}(S)}$ for all $S \subset \rho(T^1)$ and distributional consistency becomes, for $A \in \mathcal{H}^2$,

$$\begin{aligned}
\mathbb{P}^2(A) &= \int \mathbb{P}^1(d\omega)\kappa(\omega, A) \\
&= \int \mathbb{P}^1(d\omega)\mathbf{1}_A(f(\omega)) \\
&= \mathbb{P}^1(f^{-1}(A))
\end{aligned}$$

so $f_*\mathbb{P}^1 = \mathbb{P}^2$ is the pushforward measure of $\mathbb{P}^1$ along f. Interventional consistency then reads

$$K^1_{\rho^{-1}(S)}(\omega, f^{-1}(A)) = K^2_S(f(\omega), A) \tag{2.3}$$

for all $A \in \mathcal{H}^2_{\rho(T^1)}$, $S \subset \rho(T^1)$, and $\omega \in \Omega^1$. Alternatively this can be expressed as

$$f_*K^1_{\rho^{-1}(S)}(\omega, A) = K^2_S(f(\omega), A).$$

where the push-forward acts on the measure defined by the probability kernel for some fixed ω. Henceforth, with a slight abuse of notation, we denote deterministic maps by (f, ρ) without resorting to the associated probability kernel.

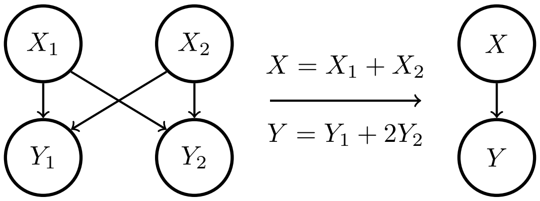


Figure 2.3: Abstraction of SCMs in Example 2.2.3.

2.2.1 Examples

Let us provide four prototypical examples of maps between causal spaces that are covered by this definition. Again, we make use of the language of SCMs.

Example 2.2.3. *We consider four variables X_1, X_2, Y_1, and Y_2 which are related through the equations*

$$\begin{aligned} X_1 &= N_1, & X_2 &= N_2, \\ Y_1 &= 3X_1 + X_2 + U_1, & Y_2 &= X_2 + U_2 \end{aligned}$$

where U_1, U_2, N_1, N_2 are independent standard normal variables. We denote by $\mathbb{P}$ their joint distribution on $\mathbb{R}^4$. Consider

$$X = N, \qquad Y = 3X + U$$

where $N \sim N(0,2)$ and $U \sim N(0,5)$. Denote their joint distribution on $\mathbb{R}^2$ by $\mathbb{Q}$. We identify the two SCMs with causal spaces $(\mathbb{R}^4, \mathcal{B}(\mathbb{R}^4), \mathbb{P}, \mathbb{K})$ and $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), \mathbb{Q}, \mathbb{L})$ as explained in Section 1.4.

Consider the deterministic map $f : \mathbb{R}^4 \to \mathbb{R}^2$ given by $f(x_1, x_2, y_1, y_2) = (x_1 + x_2, y_1 + 2y_2)$ and the map $\rho : [4] \to [2]$ given by $\rho(1) = \rho(2) = 1$, $\rho(3) = \rho(4) = 2$. Clearly, the pair (f, ρ) is admissible as defined in Definition 2.2.1. It can be checked that

$$\begin{aligned} \mathbb{E}[(X_1 + X_2)^2] &= 2 = \mathbb{E}[X^2] \\ \mathbb{E}[(Y_1 + 2Y_2)^2] &= 23 = \mathbb{E}[Y^2] \\ \mathbb{E}[(X_1 + X_2)(Y_1 + 2Y_2)] &= 6 = \mathbb{E}[XY] \end{aligned}$$

which implies that $f_*\mathbb{P} = \mathbb{Q}$ because both distributions are centred Gaussian and their covariance matrices agree.

The non-trivial causal consistency relation (2.2) concerns interventions on $\{X_1, X_2\}$ and X and on $\{Y_1, Y_2\}$ and Y. Note that

$$K_{\{X_1,X_2\}}((x_1, x_2, y_1, y_2), \cdot) = \delta_{(x_1,x_2)} \otimes N\left(\begin{pmatrix} 3x_1 + x_2 \\ x_2 \end{pmatrix}, \mathrm{Id}_2\right).$$

Then we obtain

$$f_*K_{\{X_1,X_2\}}((x_1, x_2, y_1, y_2), \cdot) = \delta_{x_1+x_2} \otimes N(3x_1 + 3x_2, 5).$$

$$\begin{tikzcd}\mathcal{C}^1 \arrow[r, "\rho(t)=t", "\kappa(\omega,\cdot)=\delta_\omega\otimes\mathbb{P}^2"'] & \mathcal{C}^1\otimes\mathcal{C}^2\end{tikzcd}$$

Figure 2.4: Inclusions of component causal spaces into the product (Example 2.2.4).

On the other hand, we find

$$\begin{aligned} L_X((x,y),\cdot) &= \delta_x \otimes N(3x,5) \\ \Rightarrow L_X(f(x_1,x_2,y_1,y_2),\cdot) &= \delta_{x_1+x_2}\otimes N(3x_1+3x_2,5) \end{aligned}$$

so that we see that (2.3) *holds in this case. Similarly, we obtain*

$$\begin{aligned} K_{\{Y_1,Y_2\}}((x_1,x_2,y_1,y_2),\cdot) &= N(0,\mathrm{Id}_1)\otimes\delta_{(y_1,y_2)}, \\ L_Y((x,y),\cdot) &= N(0,2)\otimes\delta_y. \end{aligned}$$

We again find

$$f_* K_{\{Y_1,Y_2\}}((x_1,x_2,y_1,y_2),\cdot) = L_Y((x_1+x_2,y_1+2y_2),\cdot).$$

This example shows abstraction, i.e., we obtain a transformation to a more coarse-grained view of the system. Note that interventional consistency is quite restrictive to satisfy, e.g., here it is crucial that all distributions are Gaussian so that all conditional distributions are also Gaussian.

Next, we consider an example that allows us to embed a causal space in a larger space that adds an independent disjoint system. For this, we make use of the definition of product causal spaces (Definition 2.1.1). In this case, the transformation is stochastic.

Example 2.2.4. *Let* $\mathcal{C}^1=(\Omega^1,\mathcal{H}^1,\mathbb{P}^1,\mathbb{K}^1)$ *and* $\mathcal{C}^2=(\Omega^2,\mathcal{H}^2,\mathbb{P}^2,\mathbb{K}^2)$ *be two causal spaces, with* $\Omega^1=\times_{t\in T^1}E_t$ *and* $\Omega^2=\times_{t\in T^2}E_t$*. We define an inclusion map* $(\kappa,\rho):\mathcal{C}^1\to\mathcal{C}^1\otimes\mathcal{C}^2$ *by considering* $\rho(t)=t$ *for* $t\in T^1$ *and* $\kappa(\omega,\cdot)=\delta_\omega\otimes\mathbb{P}^2$ *(see Figure 2.4). This pair is clearly admissible and satisfies distributional consistency:*

$$\begin{aligned} \int \mathbb{P}^1(d\omega)\kappa(\omega,A_1\times A_2) &= \int \mathbb{P}^1(d\omega)\mathbf{1}_{A_1}(\omega)\mathbb{P}^2(A_2) \\ &= \mathbb{P}^1(A_1)\mathbb{P}^2(A_2). \end{aligned}$$

Moreover, for any $S\subset T^1$, $\omega\in\Omega^1$, $A_1\in\mathcal{H}^1$ *and* $A_2\in\mathcal{H}^2$*, we have*

$$\int K^1_S(\omega,d\omega')\kappa(\omega',A_1\times A_2)=\mathbb{P}^2(A_2)K^1_S(\omega,A_1)$$

and also,

$$\int \kappa(\omega,d\omega_1'd\omega_2')K^1_S\otimes K^2_\emptyset((\omega_1',\omega_2'),A_1\times A_2)$$

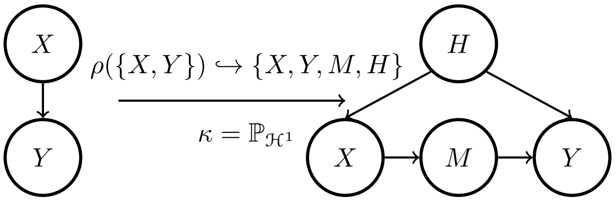

X
Y
$\rho(\{X,Y\}) \hookrightarrow \{X,Y,M,H\}$
$\kappa = \mathbb{P}_{\mathcal{H}^1}$
H
X
M
Y

Figure 2.5: Inclusions of SCMs (Example 2.2.5).

$$\begin{aligned}
&= \int \kappa(\omega, d\omega_1' d\omega_2') K_S^1(\omega_1', A_1) K_\emptyset^2(\omega_2', A_2)\\
&= K_S^1(\omega, A_1) \int \mathbb{P}^2(d\omega_2')\mathbb{P}^2(A_2)\\
&= \mathbb{P}^2(A_2) K_S^1(\omega, A_1).
\end{aligned}$$

where we used the condition on $K_\emptyset$ in Definition 1.2.1. By the usual monotone convergence theorem arguments, we have that, for any $A \in \mathcal{H}^1 \otimes \mathcal{H}^2$,

$$\int K_S^1(\omega, d\omega')\kappa(\omega', A) = \int \kappa(\omega, d\omega_1' d\omega_2') K_S^1 \otimes K_\emptyset^2((\omega_1', \omega_2'), A_1 \times A_2).$$

Thus, interventional consistency holds, in this case even for all sets A, not just for those measurable with respect to $\mathcal{H}^2_{\rho(T^1)}$.

This shows that we can consider causal maps including our system into a larger system containing additional independent components.

Finally, we consider a more involved embedding example.

Example 2.2.5. *Consider the following SCM*

$$\begin{aligned}
H &= N_H, & X &= H + N_X,\\
M &= X + N_M, & Y &= M + H + N_Y.
\end{aligned}$$

We denote the joint distribution of (X, Y, M, H) by $\mathbb{P}$, and the marginal distribution on (X, Y) by $\mathbb{P}^{XY}$.

We consider a causal space $\mathcal{C}^1 = (\Omega^1, \mathcal{H}^1, \mathbb{P}^{XY}, \mathbb{K})$ that represents the pair (X, Y), where $\Omega^1 = \mathbb{R}^2$ and $\mathcal{H}^1 = \mathcal{B}(\mathbb{R}^2)$, and a causal space $\mathcal{C}^2 = (\Omega^2, \mathcal{H}^2, \mathbb{P}, \mathbb{L})$ representing the full SCM, where $\Omega^2 = \mathbb{R}^4$ and $\mathcal{H}^2 = \mathcal{B}(\mathbb{R}^4)$, i.e., it contains in addition a mediator and a confounder. The causal mechanisms $\mathbb{K}$ and $\mathbb{L}$ are derived from the SCM. Then we consider the obvious ρ that embeds $\{X, Y\}$ into $\{X, Y, M, H\}$ and, for $A \in \mathcal{H}^2$,

$$\kappa(\cdot, A) = \mathbb{P}_{\mathcal{H}^1}(A).$$

Clearly, this pair is admissible because on the variables X and Y we use the identity transformation. Distributional consistency follows by

$$\int \kappa((x, y), A)\mathbb{P}^{XY}(d(x, y)) = \int \mathbb{P}_{\mathcal{H}^1}(A) d\mathbb{P}^{XY}$$

$$= \mathbb{P}(A).$$

Interventional consistency also holds so that (κ, ρ) is indeed a causal transformation. For a proof of this fact we refer to the more general result in Lemma 2.4.3.

This example therefore shows that we can embed a system in a larger system that captures a more accurate description.

2.2.2 Abstractions

Note that Example 2.2.3 is different from Examples 2.2.4 and 2.2.5 in that it compresses the representation while the other two consider an extension of the system. As these are different objectives, we consider the following definition.

Definition 2.2.6. The maps (κ, ρ) between measurable spaces $(\Omega^1, \mathcal{H}^1) = \otimes_{t \in T^1}(E_t, \mathcal{E}_t)$ and $(\Omega^2, \mathcal{H}^2) = \otimes_{t \in T^2}(E_t, \mathcal{E}_t)$ is called an *abstraction* if $\rho : T^1 \to T^2$ is surjective.

In the case of abstractions it is often sufficient to consider deterministic maps, motivating the following definition.

Definition 2.2.7. An abstraction (κ, ρ) is called a *perfect abstraction* if κ is deterministic, i.e., $\kappa = \kappa_f$ for some measurable $f : \Omega^1 \to \Omega^2$, and moreover f is surjective.

We finally remark that one further setting of potential interest would be to consider the inverse of an abstraction, i.e., a setting where a summary variable X is mapped to a more detailed description (X_1, X_2). However, to accommodate such transformations we need a slightly different framework than the one presented here. Roughly, we need to consider $\rho : T^1 \to \mathcal{P}(T^2)$ with $\rho(t_1) \cap \rho(t_1') = \emptyset$ for $t_1, t_1' \in T_1$, and interventions on all sets $S \subset T^1$ can be expressed as interventions on the target $\mathcal{C}^2$ (i.e., the more fine-grained representations), while this is reversed in our case so that those two settings are dual to each other.

We do not pursue this here any further, as those transformations are of more limited interest and applicability. Let us emphasise nevertheless that it seems ambitious to handle all cases in one framework. Indeed, combining variables in a summary variable or splitting variables in a more fine-grained description are meaningful operations, but it is less clear to interpret in a causal manner a definition of a transformation $(X_1, X_2) \to (Y_1, Y_2)$ that allows both at the same time. For example, intervening on X_1, in general, then does not correspond to a meaningful causal operation on the variables (Y_1, Y_2). We also remark that this attempt has not been made in the SCM literature, where the focus is almost exclusively on abstractions.

2.3 Comparison with Abstraction in the SCM framework

Rubenstein et al. (2017) gives the definition of *exact transformations* between SCMs. While being the seminal work on the theory of causal abstractions, it is probably also the most relevant to compare to our proposals. We first recall some essential aspects of their definition of SCMs (or SEMs, for structural equation models, by their nomenclature)[3].

Definition 2.3.1 ((Rubenstein et al., 2017, Definition 1))**.** Let $\mathbb{I}_X$ be an index set. An SEM $\mathcal{M}_X$ over variables $X = (X_i : i \in \mathbb{I}_X$ taking values in $\mathcal{X}$ is a tuple $(\mathcal{S}_X, \mathbb{P}_E)$, where

- $\mathcal{S}_X$ is a set of structural equations, i.e. the set of equations $X_i = f_i(X, E_i)$ for $i \in \mathbb{I}_X$;
- $\mathbb{P}_E$ is a distribution over the exogenous variables $E = (E_i : i \in \mathbb{I}_X)$.

Note that their definition of SCMs is a bit more general than standard ones in the literature (e.g. (Peters et al., 2017, p.83, Definition 6.2)), in that they allow, for example, cycles and latent confounders, but they simply insist that there must be a unique solution to any interventions. They also consider a specific set of "allowed interventions", rather than considering all possible interventions. We also recall some essential aspects of the notion of exact transformations.

Definition 2.3.2 ((Rubenstein et al., 2017, Definition 3))**.** Let $\mathcal{M}_X$ and $\mathcal{M}_Y$ be SCMs, and $\tau : \mathcal{X} \to \mathcal{Y}$ a function. We say that $\mathcal{M}_Y$ is an *exact τ-transformation* of $\mathcal{M}_X$ if, there exists a surjective mapping ω of the interventions such that for any intervention i, $\mathbb{P}^i_{\tau(X)} = \mathbb{P}^{\omega(i)}_Y$.

Note that this definition is trying to capture the same concept as our notion of interventional consistency given in (2.2): that interventions and transformations commute. However, there are several aspects in which our proposal is more appealing.

- They only consider deterministic maps $\tau : \mathcal{X} \to \mathcal{Y}$, whereas we allow the map ρ to be stochastic.
- They have to find a separate map ω *between the interventions themselves*, whereas our map ρ also determines the transformation of the causal kernels.
- By insisting on surjectivity of ω, they only allow the consideration of abstraction, whereas we can consider more general transformations of causal spaces, such as inclusions considered in Example 2.2.4.

Nevertheless, restricted to considerations amenable to both approaches, the notions coincide. For example, we return to Example 2.2.3, where we already showed that $f_*\mathbb{P} = \mathbb{Q}$, $f_*K_{\{X_1,X_2\}} = L_X$ and $f_*K_{\{Y_1,Y_2\}} = L_Y$, which implies that two-variable SCM is an exact transformation of the four-variable SCM according to Definition 2.3.2.

[3] In this section, some imported notations might clash with ours; the clashes are restricted to this section and should not cause any confusion.

Finally, we mention that Beckers and Halpern (2019) criticise exact transformations of Rubenstein et al. (2017) on the basis that probabilities and allowed interventions can mask significant differences between SCMs, and then proceed to propose definitions of abstractions that depend only on the structural equations, independently of probabilities. We remark that this criticism is not valid in our framework, in that the interventional consistency of our transformations is imposed independently of probabilities, making it impossible to mask them with the choice of probability measures. That this is possible with SCMs is an artefact of the fact that in SCMs, the observational and interventional measures are coupled through the exogenous distribution, whereas in causal spaces they are completely decoupled.

Moreover, we consider all possible interventions rather than a reduced set of allowed interventions. We also remark that, since probabilities and causal kernels are the primitive objects in our framework, rather than being derived by other primitive objects (namely the structural equations), it does not make sense for the transformation to be defined independently of probabilities, as done by Beckers and Halpern (2019).

2.4 Further Properties of Causal Transformations

In this section we investigate various properties of causal transformations and connect them to the notions introduced in Chapter 1.

First, we have the following lemma on the composition of causal transformations. Recall that for two probability kernels $\kappa_1 : \Omega^1 \times \mathcal{H}^2 \to [0,1]$ mapping $(\Omega^1, \mathcal{H}^1)$ to $(\Omega^2, \mathcal{H}^2)$ and $\kappa_2 : \Omega^2 \times \mathcal{H}^3 \to [0,1]$ mapping $(\Omega^2, \mathcal{H}^2)$ to $(\Omega^3, \mathcal{H}^3)$ the concatenation defined by (Çınlar, 2011, p.39)

$$\kappa_1 \circ \kappa_2(\omega_1, A) = \int \kappa_1(\omega_1, d\omega_2)\kappa_2(\omega_2, A)$$

defines a probability kernel from $(\Omega^1, \mathcal{H}^1)$ to $(\Omega^3, \mathcal{H}^3)$.

Lemma 2.4.1. *Let $(\kappa_1, \rho_1) : \mathcal{C}^1 \to \mathcal{C}^2$ and $(\kappa_2, \rho_2) : \mathcal{C}^2 \to \mathcal{C}^3$ be causal transformations. If (κ_1, ρ_1) is an abstraction then $(\kappa_3, \rho_3) = (\kappa_1 \circ \kappa_2, \rho_1 \circ \rho_2) : \mathcal{C}^1 \to \mathcal{C}^3$ is a causal transformation.*

We remark that, unfortunately, we cannot remove the assumption that the first transformation is an abstraction. Let us clarify this through an example.

Example 2.4.2. *Consider an SCM with equations*

$$\begin{aligned} X_1 &= N_1, \\ X_2 &= N_2, \\ Y &= X_1 + X_2 + N_Y \end{aligned}$$

where N_1, N_2, and N_Y follow independent standard normal distributions. Then we can consider the causal space $\mathcal{C}^1$ containing (X_1, Y), the causal space $\mathcal{C}^2$ containing (X_1, X_2, Y) and an abstraction $\mathcal{C}^3$ containing $(X_1 + X_2, Y)$. Then we can embed $\mathcal{C}^1 \to \mathcal{C}^2$ and there is an abstraction $\mathcal{C}^2 \to \mathcal{C}^3$ which are both transformations of causal spaces.

However, their concatenation is not a causal transformation because it is not even admissible (and also interventional consistency does not hold for the intervention K^3_X as this cannot be expressed by $K^1_{X_1}$). Note that

$$P_{X_2|X_1=x_1,Y=y} = N((y - x_1)/2, 1/2)$$

and therefore we have

$$\kappa_1((x_1, y), \cdot) = \delta_{x_1} \otimes N((y - x_1)/2, 1/2) \otimes \delta_y.$$

We also have $\kappa_2((x_1, x_2, y), \cdot) = \delta_{(x_1+x_2,y)}$. Thus, their concatenation is given by

$$\kappa_3((x_1, y), \cdot) = N((y + x_1)/2, 1/2) \otimes \delta_y.$$

So the first coordinate is not measurable with respect to $\mathcal{H}^1_{X_1}$.

This shows that we lose measurability along the concatenation because the variables added in the more complete description $\mathcal{C}^2$ may depend on all other variables.

Let us now generalise Example 2.2.5 to general SCMs.

Lemma 2.4.3. *Consider an acyclic SCM on variables $(X_1, \ldots, X_d) \in \mathbb{R}^d$ with observational distribution $\mathbb{P}$. Let $S \subset [d]$, $R = S^c = [d] \setminus S$ and consider causal spaces $\mathcal{C}^1 = (\Omega^1, \mathcal{H}^1, \mathbb{P}^S, \mathbb{K})$ and $\mathcal{C}^2 = (\Omega^2, \mathcal{H}^2, \mathbb{P}, \mathbb{L})$, where we have $(\Omega^1, \mathcal{H}^1) = (\mathbb{R}^{|S|}, \mathcal{B}(\mathbb{R}^{|S|}))$ and $(\Omega^2, \mathcal{H}^2) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Moreover, $\mathbb{P}^S$ is the marginal distribution on the variables in S, and the causal mechanisms $\mathbb{K}$ and $\mathbb{L}$ are derived from the SCM. In particular, $\mathbb{K}$ is a marginalisation of $\mathbb{L}$, namely, for any $\omega \in \Omega^2$, any event $A \in \mathcal{H}^1$ and any $S' \subseteq S$, we have that $K_{S'}(\omega, A) = L_{S'}(\omega, A)$.*

Consider the map $\rho : S \hookrightarrow [d]$ and $\kappa(\cdot, A) = \mathbb{P}_{\mathcal{H}^1}(A)$. Then (ρ, κ) is a causal transformation from $\mathcal{C}^1$ to $\mathcal{C}^2$.

We now investigate to what degree distributional and interventional consistency determines the causal structure on the target space. We show that generally the causal structure on $\mathcal{H}^2_{\rho(T)}$ is quite rigid.

Lemma 2.4.4. *Let $\mathcal{C}^2 = (\Omega^2, \mathcal{H}^2, \mathbb{P}^2, \mathbb{K}^2)$ and $\tilde{\mathcal{C}}^2 = (\Omega^2, \mathcal{H}^2, \tilde{\mathbb{P}}^2, \tilde{\mathbb{K}}^2)$ be two causal spaces with the same underlying measurable space.*

Let (κ, ρ) be an admissible pair for the measurable spaces $(\Omega^1, \mathcal{H}^1)$ and $(\Omega^2, \mathcal{H}^2)$. Assume that the pair (κ, ρ) defines causal transformations $\varphi : \mathcal{C}^1 \to \mathcal{C}^2$ and $\tilde{\varphi} : \mathcal{C}^1 \to \tilde{\mathcal{C}}^2$ be a causal transformations.

Then $\mathbb{P}^2 = \tilde{\mathbb{P}}^2$, and for all $A \in \mathcal{H}^2_{\rho(T^1)}$ and any $S \subseteq T^2$

$$K^2_S(\omega, A) = \tilde{K}^2_S(\omega, A) \qquad \textit{for } \mathbb{P}^2 = \tilde{\mathbb{P}}^2\textit{-a. e. } \omega \in \Omega^2.$$

We cannot expect to derive much stronger results for general causal transformation because interventional consistency does not restrict $K^2(\omega, A)$ for ω not in the support of $\mathbb{P}^2$ or $A \notin \mathcal{H}^2_{\rho(T)}$. For example, in the setting of Example 2.2.4, the causal structure on the second factor is arbitrary.

However, when we consider deterministic transformations (f, ρ) such that $f : (\Omega^1, \mathcal{H}^1) \to (\Omega^2, \mathcal{H}^2)$ and ρ are surjective, then there is at most one causal structure on the target space $(\Omega^2, \mathcal{H}^2)$ such that the pair (f, ρ) is a causal transformation (and thus a perfect abstraction).

Lemma 2.4.5. *Suppose (f, ρ) is an admissible pair for the causal space $\mathcal{C}^1 = (\Omega^1, \mathcal{H}^1, \mathbb{P}^1, \mathbb{K}^1)$ to the measurable space $X^2 = (\Omega^2, \mathcal{H}^2)$ and assume that ρ is surjective and $f : \Omega_1 \to \Omega_2$ measurable. If f is surjective, there exists at most one causal space $\mathcal{C}^2 = (\Omega^2, \mathcal{H}^2, \mathbb{P}^2, \mathbb{K}^2)$ such that $(f, \rho) : \mathcal{C}^1 \to \mathcal{C}^2$ is a causal transformation.*

If, in addition, $K^1_{\rho^{-1}(S^2)}(\cdot, A)$ is measurable with respect to $f^{-1}(\mathcal{H}^2_{S^2})$ for all $A \in f^{-1}(\mathcal{H}^2)$ and all $S^2 \subset T^2$ then a unique causal space $\mathcal{C}^2$ exists such that $(f, \rho) : \mathcal{C}^1 \to \mathcal{C}^2$ is a causal transformation.

To motivate the measurability condition for $K^1(\cdot, A)$, we remark that interventional consistency requires $K^1(\omega, A) = K^1(\omega', A)$ for ω and ω' with $f(\omega) = f(\omega')$, and the measurability condition in the result is a slightly stronger condition than this.

Next, we show that interventions on a space can be pushed forward along a perfect abstraction.

Lemma 2.4.6. *Let $\mathcal{C}^1 = (\Omega^1, \mathcal{H}^1, \mathbb{P}^1, \mathbb{K}^1)$ with $(\Omega^1, \mathcal{H}^1)$ a product with index set T^1 and $\mathcal{C}^2 = (\Omega^2, \mathcal{H}^2, \mathbb{P}^2, \mathbb{K}^2)$ with $(\Omega^2, \mathcal{H}^2)$ a product with index set T^2 be causal spaces, and let $(f, \rho) : \mathcal{C}^1 \to \mathcal{C}^2$ be a perfect abstraction.*

Let $U^1 = \rho^{-1}(U^2) \subseteq T^1$ for some $U^2 \subseteq T^2$. Let $\mathbb{Q}^1$ be a probability measure on $(\Omega^1, \mathcal{H}^1_{U^1})$ and $\mathbb{L}^1$ a causal mechanism on $(\Omega^1, \mathcal{H}^1_{U^1}, \mathbb{Q}^1)$. Suppose that, for all $S \subseteq U^2$ and $A \in \mathcal{H}^1$, the map $L^1_{\rho^{-1}(S)}(\cdot, A)$ is measurable with respect to $f^{-1}(\mathcal{H}^2_S)$, and consider the intervened causal spaces

$$\mathcal{C}^1_I = (\Omega^1, \mathcal{H}^1, (\mathbb{P}^1)^{\mathrm{do}(U^1, \mathbb{Q}^1)}, (\mathbb{K}^1)^{\mathrm{do}(U^1, \mathbb{Q}^1, \mathbb{L}^1)}),$$
$$\mathcal{C}^2_I = (\Omega^2, \mathcal{H}^2, (\mathbb{P}^2)^{\mathrm{do}(U^2, \mathbb{Q}^2)}, (\mathbb{K}^2)^{\mathrm{do}(U^2, \mathbb{Q}^2, \mathbb{L}^2)}),$$

where $\mathbb{Q}^2 = f_* \mathbb{Q}^1$ and $\mathbb{L}^2$ is the unique family of kernels satisfying

$$L^2_S(f(\omega), A) = L^1_{\rho^{-1}(S)}(\omega, f^{-1}(A))$$

for all $\omega \in \Omega^1$, $A \in \mathcal{H}^2$, and $S \subseteq U^2$. Then $(f, \rho) : \mathcal{C}^1_I \to \mathcal{C}^2_I$ is a perfect abstraction.

We now study whether causal effects in target and domain of a causal transformation can be related. Our first results shows that, for perfect abstractions having no causal effect in the domain implies, there is also no causal effect in the target.

Lemma 2.4.7. *Let $(f,\rho) : \mathcal{C}^1 \to \mathcal{C}^2$ be a perfect abstraction. Consider two sets $U^2, V^2 \subset T^2$ and denote $U^1 = \rho^{-1}(U^2)$ and $V^1 = \rho^{-1}(V^2)$. If $\mathcal{H}^1_{U^1}$ has no causal effect on $\mathcal{H}^1_{V^1}$ in $\mathcal{C}^1$, then $\mathcal{H}^2_{U^2}$ has no causal effect on $\mathcal{H}^2_{V^2}$ in $\mathcal{C}^2$.*

On the other hand, we can show that when there is an active causal effect in the target space, there is also an active causal effect in the domain.

Lemma 2.4.8. *Let $(f,\rho) : \mathcal{C}^1 \to \mathcal{C}^2$ be a perfect abstraction. Consider two sets $U^2, V^2 \subset T^2$ and denote $U^1 = \rho^{-1}(U^2)$ and $V^1 = \rho^{-1}(V^2)$. Assume that $\mathcal{H}^2_{U^2}$ has an active causal effect on $\mathcal{H}^2_{V^2}$ in $\mathcal{C}^2$. Then $\mathcal{H}^1_{U^1}$ has an active causal effect on $\mathcal{H}^1_{V^1}$ in $\mathcal{C}^1$.*

The reverse statements are not true, i.e., if there is no causal effect in the target there might be a causal effect in the domain, and if there is an active causal effect in the domain this does not imply that there is a causal effect in the target, which can be seen by considering a target space with only a single point.

We can also study causal effects in the context of embedding transformations, as in Lemma 2.4.3. Then we see directly that active causal effects are preserved. On the other hand, it is straightforward to construct examples where there is no causal effect in a subsystem, but there is a causal effect in a larger system. This can be achieved by a violation of faithfulness.

Example 2.4.9. *Consider the SCM*

$$\begin{aligned} X &= N_X, \\ M &= N_X + N_M, \\ Y &= M - X + N_Y. \end{aligned}$$

Then there is no causal effect from $\sigma(X)$ to $\sigma(Y)$ in the system (X,Y) but there is a causal effect in the complete system.

Finally, we show that similar results can be established for sources. Indeed, perfect abstraction preserve sources in the following sense.

Lemma 2.4.10. *Let $(f,\rho) : \mathcal{C}^1 \to \mathcal{C}^2$ be a perfect abstraction. Consider two sets $U^2, V^2 \subset T^2$ and denote $U^1 = \rho^{-1}(U^2)$ and $V^1 = \rho^{-1}(V^2)$. Assume that $\mathcal{H}^1_{U^1}$ is a local source of $\mathcal{H}^1_{V^1}$ in $\mathcal{C}^1$. Then $\mathcal{H}^2_{U^2}$ is a local source of $\mathcal{H}^2_{V^2}$ in $\mathcal{C}^2$.*

In particular, this implies that if $\mathcal{H}^1_{U^1}$ is a global source then $\mathcal{H}^2_{U^2}$ also is a global source.

Similar to our results for causal effects, the existence of sources in the abstracted space does not guarantee the existence of sources in the domain space. Note that local sources are preserved in the setting of Lemma 2.4.3. On the other hand, global sources are clearly not preserved, as we can add a global source to the system.

Part II

Kernel Regression

Chapter 3

Kernel Conditional Mean Embeddings

3.1 Preliminaries

We take $(\Omega, \mathcal{F}, P)$ as the underlying probability space. Let $(\mathcal{X}, \mathfrak{X})$, $(\mathcal{Y}, \mathfrak{Y})$ and $(\mathcal{Z}, \mathfrak{Z})$ be separable measurable spaces, and let $X : \Omega \to \mathcal{X}$, $Y : \Omega \to \mathcal{Y}$ and $Z : \Omega \to \mathcal{Z}$ be random variables with distributions P_X, P_Y and P_Z. We will use Z as the conditioning variable throughout.

3.1.1 Positive definite kernels and RKHS embeddings

Let $\mathcal{H}_\mathcal{X}$ be a vector space of $\mathcal{X} \to \mathbb{R}$ functions, endowed with a Hilbert space structure via an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_\mathcal{X}}$. A symmetric function $k_\mathcal{X} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *reproducing kernel* of $\mathcal{H}_\mathcal{X}$ if and only if: 1. $\forall x \in \mathcal{X}$, $k_\mathcal{X}(x, \cdot) \in \mathcal{H}_\mathcal{X}$; 2. $\forall x \in \mathcal{X}$ and $\forall f \in \mathcal{H}_\mathcal{X}$, $f(x) = \langle f, k_\mathcal{X}(x, \cdot) \rangle_{\mathcal{H}_\mathcal{X}}$. A space $\mathcal{H}_\mathcal{X}$ which possesses a reproducing kernel is called a *reproducing kernel Hilbert space* (RKHS) (Berlinet and Thomas-Agnan, 2004). Throughout the rest of this thesis, we assume that all RKHSs are *separable*. This is not a restrictive assumption, since it is satisfied if, for example, $k_\mathcal{X}$ is a continuous kernel (Steinwart and Christmann, 2008, p.130, Lemma 4.33) (for further details, please see Owhadi and Scovel (2017)). Given a distribution P_X on $\mathcal{X}$, assuming the integrability condition

$$\int_\mathcal{X} \sqrt{k_\mathcal{X}(x, x)} dP_X(x) < \infty, \tag{3.1}$$

we define the *kernel mean embedding* $\mu_{P_X}(\cdot) = \int_\mathcal{X} k_\mathcal{X}(x, \cdot) dP_X(x)$ of P_X, where the integral is a *Bochner integral* (Dinculeanu, 2000, p.15, Def. 35). We will later show a conditional analogue of the following lemma.

Lemma 3.1.1 (Smola et al. (2007)). *For each $f \in \mathcal{H}_\mathcal{X}$, $\int_\mathcal{X} f(x) dP_X(x) = \langle f, \mu_{P_X} \rangle_{\mathcal{H}_\mathcal{X}}$.*

Next, suppose $\mathcal{H}_{\mathcal{Y}}$ is an RKHS of functions on $\mathcal{Y}$ with kernel $k_{\mathcal{Y}}$, and consider the *tensor product RKHS* $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ (see (Weidmann, 1980, pp.47-48) for a definition of tensor product Hilbert spaces).

Theorem 3.1.2 ((Berlinet and Thomas-Agnan, 2004, p.31, Theorem 13))**.** *The tensor product $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ is generated by the functions $f \otimes g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, with $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$ defined by $(f \otimes g)(x, y) = f(x)g(y)$. Moreover, $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ is an RKHS of functions on $\mathcal{X} \times \mathcal{Y}$ with kernel $(k_{\mathcal{X}} \otimes k_{\mathcal{Y}})((x_1, y_1), (x_2, y_2)) = k_{\mathcal{X}}(x_1, x_2)k_{\mathcal{Y}}(y_1, y_2)$.*

Now let us impose a slightly stronger integrability condition:

$$\mathbb{E}_X[k_{\mathcal{X}}(X, X)] < \infty, \quad \mathbb{E}_Y[k_{\mathcal{Y}}(Y, Y)] < \infty. \tag{3.2}$$

This ensures that $k_{\mathcal{X}}(X, \cdot) \otimes k_{\mathcal{Y}}(Y, \cdot)$ is Bochner P_{XY}-integrable, and so $\mu_{P_{XY}} := \mathbb{E}_{XY}[k_{\mathcal{X}}(X, \cdot) \otimes k_{\mathcal{Y}}(Y, \cdot)] \in \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$. The next lemma is analogous to Lemma 3.1.1:

Lemma 3.1.3 ((Fukumizu et al., 2004, Theorem 1))**.** *For $f \in \mathcal{H}_{\mathcal{X}}$, $g \in \mathcal{H}_{\mathcal{Y}}$, $\langle f \otimes g, \mu_{P_{XY}} \rangle_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}} = \mathbb{E}_{XY}[f(X)g(Y)]$.*

As a consequence, for any pair $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$, we have $\langle f \otimes g, \mu_{P_{XY}} - \mu_{P_X} \otimes \mu_{P_Y} \rangle_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}} = \mathrm{Cov}_{XY}[f(X), g(Y)]$. There exists an isometric isomorphism $T : \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}} \to \mathrm{HS}(\mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{Y}})$, where $\mathrm{HS}(\mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{Y}})$ is the space of Hilbert-Schmidt operators from $\mathcal{H}_{\mathcal{X}}$ to $\mathcal{H}_{\mathcal{Y}}$. The (centred) *cross-covariance operator* is defined as $\mathcal{C}_{YX} := T(\mu_{P_{XY}} - \mu_{P_X} \otimes \mu_{P_Y})$ (Fukumizu et al., 2004, Theorem 1). The object $T(\mu_{P_{XY}})$ is referred to as the *uncentred cross-covariance operator* in the literature (Song et al., 2010a, Section 3.2).

The notion of *characteristic kernels* is essential, since it tells us that the associated RKHSs are rich enough to enable us to distinguish different distributions from their embeddings.

Definition 3.1.4 (Fukumizu et al. (2008))**.** A positive definite kernel $k_{\mathcal{X}}$ is *characteristic* to a set $\mathcal{P}$ of probability measures defined on $\mathcal{X}$ if the map $\mathcal{P} \to \mathcal{H}_{\mathcal{X}} : P_X \mapsto \mu_{P_X}$ is injective.

Sriperumbudur et al. (2010) discusses various characterisations of characteristic kernels and show that the well-known Gaussian and Laplacian kernels are characteristic. We then have a metric on $\mathcal{P}$ via $\|\mu_{P_X} - \mu_{P_{X'}}\|_{\mathcal{H}_{\mathcal{X}}}$ for $P_X, P_{X'} \in \mathcal{P}$, which is the definition of the MMD (Gretton et al., 2007). Furthermore, the HSIC is defined as the Hilbert-Schmidt norm of $\mathcal{C}_{YX}$, or equivalently, $\|\mu_{P_{XY}} - \mu_{P_X} \otimes \mu_{P_Y}\|_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}}$ (Gretton et al., 2005). If $k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$ is characteristic, then HSIC $= 0$ if and only if $X \perp\!\!\!\perp Y$.

3.1.2 Conditioning

Recall that we reviewed some essentials of probability theory with real-valued random variables in Section 1.1. We briefly review the concept of conditioning in measure-theoretic probability theory with Banach space-valued random variables. We consider a sub-σ-algebra $\mathcal{E}$ of $\mathcal{F}$ and a Banach space $\mathcal{H}$.

Definition 3.1.5 (Conditional Expectation, (Dinculeanu, 2000, p.45, Definition 38))**.** Suppose H is a Bochner P-integrable, $\mathcal{H}$-valued random variable. Then the *conditional expectation* of H given $\mathcal{E}$ is any $\mathcal{E}$-measurable, Bochner P-integrable, $\mathcal{H}$-valued random variable H' such that $\int_A H dP = \int_A H' dP \; \forall A \in \mathcal{E}$. Any H' satisfying this condition is a *version* of $\mathbb{E}[H \mid \mathcal{E}]$. We write $\mathbb{E}[H \mid Z]$ to mean $\mathbb{E}[H \mid \sigma(Z)]$, where $\sigma(Z)$ is the sub-σ-algebra of $\mathcal{F}$ generated by the random variable Z.

The (almost sure) uniqueness of the conditional expectation is shown in (Dinculeanu, 2000, p.44, Proposition 37), and the existence in (Dinculeanu, 2000, pp.45-46, Theorems 39 and 50). The following theorem, proved in Appendix A.3, is the reason why a regular version is important. It means that, roughly speaking, the conditional expectation is indeed obtained by integration with respect to the conditional measure.

Theorem 3.1.6 (Adapted from (Çınlar, 2011, p.150, Proposition 2.5))**.** *Suppose that $P(\cdot \mid \mathcal{E})$ admits a regular version Q. Then $QH : \Omega \to \mathcal{H}$ with $\omega \mapsto Q_\omega H = \int_\Omega H(\omega') Q_\omega(d\omega')$ is a version of $\mathbb{E}[H \mid \mathcal{E}]$ for every Bochner P-integrable H.*

3.1.3 Vector-valued RKHS regression

In this subsection, we introduce the theory of vector-valued RKHS regression, based on operator-valued kernels. For a more comprehensive treatment, see Chapter 5. Let $\mathcal{H}$ be a Hilbert space, which will be the output space of regression.

Definition 3.1.7 ((Carmeli et al., 2006, Definition 1))**.** An *$\mathcal{H}$-valued RKHS* on $\mathcal{Z}$ is a Hilbert space $\mathcal{G}$ such that 1. the elements of $\mathcal{G}$ are functions $\mathcal{Z} \to \mathcal{H}$; 2. $\forall z \in \mathcal{Z}$, $\exists C_z > 0$ such that $\|F(z)\|_\mathcal{H} \leq C_z \|F\|_\mathcal{G} \; \forall F \in \mathcal{G}$.

Next, we let $\mathcal{L}(\mathcal{H})$ denote the Banach space of bounded linear operators from $\mathcal{H}$ into itself.

Definition 3.1.8 ((Carmeli et al., 2006, Definition 2))**.** A *$\mathcal{H}$-kernel of positive type* on $\mathcal{Z} \times \mathcal{Z}$ is a map $\Gamma : \mathcal{Z} \times \mathcal{Z} \to \mathcal{L}(\mathcal{H})$ such that $\forall N \in \mathbb{N}$, $\forall z_1, ..., z_N \in \mathcal{Z}$ and $\forall c_1, ..., c_N \in \mathbb{R}$, $\sum_{i,j=1}^N c_i c_j \langle \Gamma(z_j, z_i) h, h \rangle_\mathcal{H} \geq 0 \; \forall h \in \mathcal{H}$.

Analogously to the scalar case, it can be shown that any $\mathcal{H}$-valued RKHS $\mathcal{G}$ possesses a *reproducing kernel*, which is an $\mathcal{H}$-kernel of positive type Γ satisfying, for any $z, z' \in \mathcal{Z}$, $h, h' \in \mathcal{H}$ and $F \in \mathcal{G}$, $\langle F(z), h \rangle_\mathcal{H} = \langle F, \Gamma(\cdot, z) h \rangle_\mathcal{G}$ and $\langle h, \Gamma(z, z')(h') \rangle_\mathcal{H} = \langle \Gamma(\cdot, z)(h), \Gamma(\cdot, z')(h') \rangle_\mathcal{G}$.

Now suppose we want to perform regression with input space $\mathcal{Z}$ and output space $\mathcal{H}$, by minimising

$$\frac{1}{n} \sum_{j=1}^n \|h_j - F(z_j)\|_\mathcal{H}^2 + \lambda \|F\|_\mathcal{G}^2, \tag{3.3}$$

where $\lambda > 0$ is a regularisation parameter and $\{(z_j, h_j) : j = 1, ..., n\} \subseteq \mathcal{Z} \times \mathcal{H}$. There is a corresponding representer theorem (here, δ_{jl} is the Kronecker delta):

Theorem 3.1.9 ((Micchelli and Pontil, 2005, Theorem 4.1))**.** *If $\hat{F}$ minimises (3.3) in $\mathcal{G}$, it is unique and has the form $\hat{F} = \sum_{j=1}^{n} \Gamma(\cdot, z_j)(u_j)$ where the coefficients $\{u_j : j = 1, ..., n\} \subseteq \mathcal{H}$ are the unique solution of the linear equations $\sum_{l=1}^{n}(\Gamma(z_j, z_l) + n\lambda\delta_{jl})(u_l) = h_j, j = 1, ..., n$.*

3.1.4 Generalised Jensen's Inequality

In Section 3.3, we require a version of Jensen's inequality generalised to (possibly) infinite-dimensional vector spaces, because our random variable takes values in $\mathcal{H}_\mathcal{X}$, and our convex function is $\|\cdot\|^2_{\mathcal{H}_\mathcal{X}} : \mathcal{H}_\mathcal{X} \to \mathbb{R}$. Note that this square norm function is indeed convex, since, for any $t \in [0,1]$ and any pair $f, g \in \mathcal{H}_\mathcal{X}$,

$$\begin{aligned} \|tf + (1-t)g\|^2_{\mathcal{H}_\mathcal{X}} &\leq (t\|f\|_{\mathcal{H}_\mathcal{X}} + (1-t)\|g\|_{\mathcal{H}_\mathcal{X}})^2 && \text{by the triangle inequality} \\ &\leq t\|f\|^2_{\mathcal{H}_\mathcal{X}} + (1-t)\|g\|^2_{\mathcal{H}_\mathcal{X}}, && \text{by the convexity of } x \mapsto x^2. \end{aligned}$$

The following theorem generalises Jensen's inequality to infinite-dimensional vector spaces.

Theorem 3.1.10 ((Perlman, 1974), Theorem 3.10)**.** *Suppose $\mathcal{T}$ is a real Hausdorff locally convex (possibly infinite-dimensional) linear topological space, and let C be a closed convex subset of $\mathcal{T}$. Suppose $(\Omega, \mathcal{F}, P)$ is a probability space, and $V : \Omega \to \mathfrak{T}$ a Pettis-integrable random variable such that $V(\Omega) \subseteq C$. Let $f : C \to [-\infty, \infty)$ be a convex, lower semi-continuous extended-real-valued function such that $\mathbb{E}[f(V)]$ exists. Then*

$$f(\mathbb{E}[V]) \leq \mathbb{E}[f(V)].$$

We will actually apply generalised Jensen's inequality with conditional expectations, so we need the following theorem. The proof is in Appendix A.3.

Theorem 3.1.11 (Generalised Conditional Jensen's Inequality)**.** *Suppose $\mathcal{T}$ is a real Hausdorff locally convex (possibly infinite-dimensional) linear topological space, and let C be a closed convex subset of $\mathcal{T}$. Suppose $(\Omega, \mathcal{F}, P)$ is a probability space, and $V : \Omega \to \mathcal{T}$ a Pettis-integrable random variable such that $V(\Omega) \subseteq C$. Let $f : C \to [-\infty, \infty)$ be a convex, lower semi-continuous extended-real-valued function such that $\mathbb{E}[f(V)]$ exists. Suppose $\mathcal{E}$ is a sub-σ-algebra of $\mathcal{F}$. Then*

$$f(\mathbb{E}[V \mid \mathcal{E}]) \leq \mathbb{E}[f(V) \mid \mathcal{E}].$$

In the context of Section 3.3, $\mathcal{H}_\mathcal{X}$ is real and Hausdorff, and locally convex (because it is a normed space). We take the closed convex subset to be the whole space $\mathcal{H}_\mathcal{X}$ itself. The function $\|\cdot\|^2_{\mathcal{H}_\mathcal{X}} : \mathcal{H}_\mathcal{X} \to \mathbb{R}$ is convex (as shown above) and continuous, and finally, since Bochner-integrability implies Pettis integrability, all the conditions of Theorem 3.1.11 are satisfied.

3.2 Conditional mean embedding

We are now ready to introduce a formal definition of the conditional mean embedding of X given Z.

Definition 3.2.1. Assuming X satisfies the integrability condition (3.1), we define the *conditional mean embedding* of X given Z as $\mu_{P_{X|Z}} := \mathbb{E}[k_{\mathcal{X}}(X,\cdot) \mid Z]$.

This is a direct extension of the unconditional kernel mean embedding, $\mu_{P_X} = \mathbb{E}[k_{\mathcal{X}}(X,\cdot)]$, but instead of being a fixed element in $\mathcal{H}_{\mathcal{X}}$, $\mu_{P_{X|Z}}$ is a Z-measurable random variable taking values in $\mathcal{H}_{\mathcal{X}}$ (see Definition 3.1.5). Also, for any function $f : \mathcal{X} \to \mathbb{R}$, $\mathbb{E}[f(X) \mid Z]$ is a real-valued Z-measurable random variable. The following lemma is analogous to Lemma 3.1.1.

Lemma 3.2.2. *For any $f \in \mathcal{H}_{\mathcal{X}}$, $\mathbb{E}[f(X) \mid Z] = \langle f, \mu_{P_{X|Z}} \rangle_{\mathcal{H}_{\mathcal{X}}}$ almost surely.*

Next, assuming X and Y satisfy (3.2), we define $\mu_{P_{XY|Z}} := \mathbb{E}[k_{\mathcal{X}}(X,\cdot) \otimes k_{\mathcal{Y}}(Y,\cdot) \mid Z]$, a Z-measurable, $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$-valued random variable. We have the following analogy of Lemma 3.1.3:

Lemma 3.2.3. *For any pair $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$, $\mathbb{E}[f(X)g(Y) \mid Z] = \langle f \otimes g, \mu_{P_{XY|Z}} \rangle_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}}$ almost surely.*

By Lemmas 3.2.2 and 3.2.3, for any pair $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$,

$$\begin{aligned}\langle f \otimes g, \mu_{P_{XY|Z}} - \mu_{P_{X|Z}} \otimes \mu_{P_{Y|Z}} \rangle_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}} &= \operatorname{Cov}(f(X), g(Y) \mid Z) \\ &= \mathbb{E}[f(X)g(Y) \mid Z] - \mathbb{E}[f(X) \mid Z]\mathbb{E}[g(Y) \mid Z]\end{aligned}$$

almost surely. Hence, we define the *conditional cross-covariance operator* as $\mathcal{C}_{YX|Z} := T(\mu_{P_{XY|Z}} - \mu_{P_{X|Z}} \otimes \mu_{P_{Y|Z}})$ (see Section 3.1.1 for the definition of T).

3.2.1 Comparison with existing definitions

As previously mentioned, the idea of CMEs and conditional cross-covariance operators is not a novel one, yet our development of the theory above differs significantly from the existing works. In this subsection, we review the previous approaches and compare them to ours.

The prevalent definition of CMEs in the literature is the following. We first need to endow the conditioning space $\mathcal{Z}$ with a scalar kernel, say $k_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, with corresponding RKHS $\mathcal{H}_{\mathcal{Z}}$.

Definition 3.2.4 ((Song et al., 2009, Definition 3))**.** The conditional mean embedding of the conditional distribution $P(X \mid Z)$ is the operator $\mathcal{U}_{X|Z} : \mathcal{H}_{\mathcal{Z}} \to \mathcal{H}_{\mathcal{X}}$ defined by $\mathcal{U}_{X|Z} = \mathcal{C}_{XZ}\mathcal{C}_{ZZ}^{-1}$, where $\mathcal{C}_{XZ}$ and $\mathcal{C}_{ZZ}$ are unconditional (cross-)covariance operators as defined in Section 3.1.1.

As noted by (Song et al., 2009), the motivation for this comes from (Fukumizu et al., 2004, Theorem 2), which states that for any $f \in \mathcal{H}_{\mathcal{X}}$, if $\mathbb{E}[f(X) \mid Z = \cdot] \in \mathcal{H}_{\mathcal{Z}}$, then $\mathcal{C}_{ZZ}\mathbb{E}[f(X) \mid Z = \cdot] = \mathcal{C}_{ZX} f$. This relation can be used to prove the following theorem, which is analogous to Lemma 3.2.2.

Theorem 3.2.5 ((Song et al., 2009, Theorem 4))**.** *For $f \in \mathcal{H}_\mathcal{X}$, assuming $\mathbb{E}[f(X) \mid Z = \cdot] \in \mathcal{H}_\mathcal{Z}$, $\mathcal{U}_{X|Z}$ satisfies: 1. $\mu_{X|z} := \mathbb{E}[k_\mathcal{X}(X, \cdot) \mid Z = z] = \mathcal{U}_{X|Z} k_\mathcal{Z}(z, \cdot)$; 2. $\mathbb{E}[f(X) \mid Z = z] = \langle f, \mu_{X|z} \rangle_{\mathcal{H}_\mathcal{X}}$.*

Now we highlight the key differences between this approach and ours. Firstly, this approach requires the endowment of a kernel $k_\mathcal{Z}$ on the conditioning space $\mathcal{Z}$, and defines the CME as an *operator* from $\mathcal{H}_\mathcal{Z}$ to $\mathcal{H}_\mathcal{X}$. By contrast, Definition 3.2.1 did not consider any kernel or function on $\mathcal{Z}$, and defined the CME as a *Bochner conditional expectation* given $\sigma(Z)$. We argue that it is more natural not to endow the *conditioning space* with a kernel before the estimation stage. Secondly, the operator-based approach assumes that $\mathbb{E}[f(X)|Z = \cdot]$, as a function in z, lives in $\mathcal{H}_\mathcal{Z}$. This is a severe restriction; it is stated in (Song et al., 2009) that this assumption, while true for finite domains with characteristic kernels, is not necessarily true for continuous domains, and (Fukumizu et al., 2013) gives a simple counterexample using the Gaussian kernel. Lastly, it also assumes that $\mathcal{C}_{ZZ}^{-1}$ exists, which is another unrealistic assumption. (Fukumizu et al., 2013) mentions that this assumption is too strong in many situations, and gives a counterexample using the Gaussian kernel. The most common remedy is to resort to the regularised version $\mathcal{C}_{XZ}(\mathcal{C}_{ZZ} + \lambda I)^{-1}$ and treat it as an approximation of $\mathcal{U}_{X|Z}$. These assumptions have been clarified and slightly weakened in (Klebanov et al., 2020), but strong and hard-to-verify conditions persist. In contrast, Definition 3.2.1 extend the notions of kernel mean embedding, expectation operator and cross-covariance operator to the conditional setting simply by using the formal definition of conditional expectations (Definition 3.1.5), and the subsequent result in Lemma 3.2.2, analogous to (Song et al., 2009, Theorem 4), does not rely on any assumptions.

A regression interpretation is given in Grünewälder et al. (2012a), by showing the *existence*, for each $z \in \mathcal{Z}$, of $\mu(z) \in \mathcal{H}_\mathcal{X}$ that satisfies $\mathbb{E}[h(X) \mid Z = z] = \langle h, \mu(z) \rangle_{\mathcal{H}_\mathcal{X}}$. However, no explicit expression for $\mu(z)$ is provided. In contrast, our definition provides an explicit expression $\mu_{P_{X|Z}} = \mathbb{E}[k_\mathcal{X}(X, \cdot) \mid Z]$.

In (Fukumizu et al., 2004, Section A.2), the conditional cross-covariance operator is defined, but in a significantly different way. It is defined as $\Sigma_{YX|Z} := \mathcal{C}_{YX} - \mathcal{C}_{YZ}\tilde{\mathcal{C}}_{ZZ}^{-1}\mathcal{C}_{ZX}$, where $\tilde{\mathcal{C}}_{ZZ}^{-1}$ is the right inverse of $\mathcal{C}_{ZZ}$ on $(\text{Ker}\mathcal{C}_{ZZ})^\perp$. This has the property that, for all $f \in \mathcal{H}_\mathcal{X}$ and $g \in \mathcal{H}_\mathcal{Y}$,

$$\langle g, \Sigma_{YX|Z} f \rangle_{\mathcal{H}_\mathcal{Y}} = \mathbb{E}[\text{Cov}(f(X), g(Y) \mid Z)].$$

Note that this is different to our relation stated after Lemma 3.2.3; the conditional covariance is integrated out over $\mathcal{Z}$. In fact, this difference is explicitly noted by Song et al. (2009).

3.3 Empirical estimates

In this section, we discuss how we can obtain empirical estimates of $\mu_{P_{X|Z}} = \mathbb{E}[k_\mathcal{X}(X, \cdot) \mid Z]$.

Theorem 3.3.1. *Denote the Borel σ-algebra of $\mathcal{H}_\mathcal{X}$ by $\mathcal{B}(\mathcal{H}_\mathcal{X})$. Then we can write $\mu_{P_{X|Z}} = F_{P_{X|Z}} \circ Z$, where $F_{P_{X|Z}} : \mathcal{Z} \to \mathcal{H}_\mathcal{X}$ is some deterministic function, measurable with respect to $\mathfrak{Z}$ and $\mathcal{B}(\mathcal{H}_\mathcal{X})$.*

Hence, estimating $\mu_{P_{X|Z}}$ boils down to estimating the function $F_{P_{X|Z}}$, which is exactly the setting for vector-valued regression (Section 3.1.3) with input space $\mathcal{Z}$ and output space $\mathcal{H}_\mathcal{X}$. In contrast to Grünewälder et al. (2012a), where regression is motivated by applying the Riesz representation theorem conditioned on each value of $z \in \mathcal{Z}$, we derive the CME as an explicit function of Z, which we argue is a more principled way to motivate regression. Moreover, for continuous Z, the event $Z = z$ has measure 0, so it is not measure-theoretically rigorous to apply the Riesz representation theorem conditioned on $Z = z$.

The natural optimisation problem is to minimise the loss

$$\mathcal{E}_{X|Z}(F) := \mathbb{E}_Z[\|F_{P_{X|Z}}(Z) - F(Z)\|^2_{\mathcal{H}_\mathcal{X}}]$$

among all $F \in \mathcal{G}_{\mathcal{X}\mathcal{Z}}$, where $\mathcal{G}_{\mathcal{X}\mathcal{Z}}$ is a vector-valued RKHS of functions $\mathcal{Z} \to \mathcal{H}_\mathcal{X}$. For simplicity, we endow $\mathcal{G}_{\mathcal{X}\mathcal{Z}}$ with a kernel $l_{\mathcal{X}\mathcal{Z}}(z, z') = k_\mathcal{Z}(z, z')\mathrm{Id}$, where $k_\mathcal{Z}(\cdot, \cdot)$ is a scalar kernel on $\mathcal{Z}$.[1]

We cannot minimise $\mathcal{E}_{X|Z}$ directly, since we do not observe samples from $\mu_{P_{X|Z}}$, but only the pairs (x_i, z_i) from (X, Z). We bound this with a surrogate loss $\tilde{\mathcal{E}}_{X|Z}$ that has a sample-based version:

$$\begin{aligned}
\mathcal{E}_{X|Z}(F) &= \mathbb{E}_Z[\|\mathbb{E}_{X|Z}[k_\mathcal{X}(X, \cdot) - F(Z) \mid Z]\|^2_{\mathcal{H}_\mathcal{X}}] \\
&\leq \mathbb{E}_Z\mathbb{E}_{X|Z}[\|k_\mathcal{X}(X, \cdot) - F(Z)\|^2_{\mathcal{H}_\mathcal{X}} \mid Z] \\
&= \mathbb{E}_{X,Z}[\|k_\mathcal{X}(X, \cdot) - F(Z)\|^2_{\mathcal{H}_\mathcal{X}}] \\
&=: \tilde{\mathcal{E}}_{X|Z}(F),
\end{aligned}$$

where we used generalised conditional Jensen's inequality (see Section 3.1.4, or Perlman (1974)). Section 3.3.1 discusses the meaning of this surrogate loss. We replace the surrogate population loss with a regularised empirical loss based on samples $\{(x_i, z_i)\}_{i=1}^n$ from the joint distribution P_{XZ}:

$$\hat{\mathcal{E}}_{X|Z,n,\lambda}(F) := \frac{1}{n}\sum_{i=1}^n \|k_\mathcal{X}(x_i, \cdot) - F(z_i)\|^2_{\mathcal{H}_\mathcal{X}} + \lambda\|F\|^2_{\mathcal{G}_{\mathcal{X}\mathcal{Z}}},$$

where $\lambda > 0$ is a regularisation parameter. We see that this loss functional is exactly in the form of (3.3). Therefore, by Theorem 3.1.9, the minimiser $\hat{F}_{P_{X|Z},n,\lambda}$ of $\hat{\mathcal{E}}_{X|Z,n,\lambda}$ is $\hat{F}_{P_{X|Z},n,\lambda}(\cdot) = \mathbf{k}_Z^T(\cdot)\mathbf{f}$, where $\mathbf{k}_Z(\cdot) := (k_Z(z_1, \cdot), ..., k_Z(z_n, \cdot))^T$, $\mathbf{f} := (f_1, ..., f_n)^T$ and the coefficients $f_i \in \mathcal{H}_\mathcal{X}$ are the unique solutions of

[1] $\mathcal{E}_{X|Z}$ is not the only loss function, nor is $l_{\mathcal{X}\mathcal{Z}}$ the only kernel, that we can use for this problem. Kadri et al. (2016) discuss various operator-valued kernels that can be used (albeit without closed-form solutions) and Laforgue et al. (2020) discuss other loss functions that can be used for more robust estimates. We view this flexibility to facilitate other loss and kernel functions in the regression set-up, although not explored in depth in this work, as a significant advantage over the previous approaches.

the linear equations $(\mathbf{K}_Z + n\lambda\mathbf{I})\mathbf{f} = \mathbf{k}_X$, where $[\mathbf{K}_Z]_{ij} := k_{\mathcal{Z}}(z_i, z_j)$, $\mathbf{k}_X := (k_{\mathcal{X}}(x_1, \cdot), ..., k_{\mathcal{X}}(x_n, \cdot))^T$ and $\mathbf{I}$ is the $n \times n$ identity matrix. Hence, the coefficients are $\mathbf{f} = \mathbf{W}\mathbf{k}_X$, where $\mathbf{W} = (\mathbf{K}_Z + n\lambda\mathbf{I})^{-1}$. Finally, substituting this into the expression for $\hat{F}_{P_{X|Z},n,\lambda}(\cdot)$, we have

$$\hat{F}_{P_{X|Z},n,\lambda}(\cdot) = \mathbf{k}_Z^T(\cdot)\mathbf{W}\mathbf{k}_X \in \mathcal{G}_{\mathcal{X}\mathcal{Z}}. \tag{3.4}$$

3.3.1 Surrogate loss, universality and consistency

In this subsection, we investigate the meaning and consequences of using the surrogate loss $\tilde{\mathcal{E}}_{X|Z}$ instead of the original $\mathcal{E}_{X|Z}$, as well as the universal consistency property of our learning algorithm.

Denote by $L^2(\mathcal{Z}, P_Z; \mathcal{H}_{\mathcal{X}})$ the Banach space of (equivalence classes of) measurable functions $F : \mathcal{Z} \to \mathcal{H}_{\mathcal{X}}$ such that $\|F(\cdot)\|^2_{\mathcal{H}_{\mathcal{X}}}$ is P_Z-integrable, with norm $\|F\|_2 = (\int_{\mathcal{Z}} \|F(z)\|^2_{\mathcal{H}_{\mathcal{X}}} dP_Z(z))^{\frac{1}{2}}$. We can note that the true function $F_{P_{X|Z}}$ belongs to $L^2(\mathcal{Z}, P_Z; \mathcal{H}_{\mathcal{X}})$, because Theorem 3.3.1 tells us that $F_{P_{X|Z}}$ is indeed measurable, and by Theorem 3.1.11 and (3.2),

$$\begin{aligned}\int_{\mathcal{Z}} \|F_{P_{X|Z}}(z)\|^2_{\mathcal{H}_{\mathcal{X}}} dP_Z(z) &= \mathbb{E}_Z[\|\mathbb{E}_{X|Z}[k_{\mathcal{X}}(X, \cdot) \mid Z]\|^2_{\mathcal{H}_{\mathcal{X}}}] \\ &\leq \mathbb{E}_Z[\mathbb{E}_{X|Z}[\|k_{\mathcal{X}}(X, \cdot)\|^2_{\mathcal{H}_{\mathcal{X}}} \mid Z]] \\ &= \mathbb{E}_X[\|k_{\mathcal{X}}(X, \cdot)\|^2_{\mathcal{H}_{\mathcal{X}}}] \\ &< \infty.\end{aligned}$$

The true function $F_{P_{X|Z}}$ is the unique minimiser in $L^2(\mathcal{Z}, P_Z; \mathcal{H}_{\mathcal{X}})$ of both $\mathcal{E}_{X|Z}$ and $\tilde{\mathcal{E}}_{X|Z}$:

Theorem 3.3.2. *$F_{P_{X|Z}}$ minimises both $\tilde{\mathcal{E}}_{X|Z}$ and $\mathcal{E}_{X|Z}$ over $L^2(\mathcal{Z}, P_Z; \mathcal{H}_{\mathcal{X}})$. Moreover, it is almost surely equal to any other minimiser of the loss functionals.*

Note the difference in the statement of Theorem 3.3.2 from (Grünewälder et al., 2012a, Theorem 3.1), which only considers the minimisation of the loss functionals in $\mathcal{G}_{\mathcal{X}\mathcal{Z}}$, whereas we consider the larger space $L^2(\mathcal{Z}, P_Z; \mathcal{H}_{\mathcal{X}})$. Next, we discuss the concepts of *universal kernels* and *universal consistency.*

Definition 3.3.3 ((Carmeli et al., 2010, Definition 2))**.** A kernel $l_{\mathcal{X}\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathcal{L}(\mathcal{H}_{\mathcal{X}})$ with RKHS $\mathcal{G}_{\mathcal{X}\mathcal{Z}}$ is $\mathcal{C}_0$ if $\mathcal{G}_{\mathcal{X}\mathcal{Z}}$ is a subspace of $\mathcal{C}_0(\mathcal{Z}, \mathcal{H}_{\mathcal{X}})$, the space of continuous functions $\mathcal{Z} \to \mathcal{H}_{\mathcal{X}}$ vanishing at infinity. The kernel $l_{\mathcal{X}\mathcal{Z}}$ is $\mathcal{C}_0$-*universal* if is is $\mathcal{C}_0$ and $\mathcal{G}_{\mathcal{X}\mathcal{Z}}$ is dense in $L^2(\mathcal{Z}, P_Z; \mathcal{H}_{\mathcal{X}})$ for any measure P_Z on $\mathcal{Z}$.

Carmeli et al. (2010, Example 14) shows that $l_{\mathcal{X}\mathcal{Z}} = k_{\mathcal{Z}}(\cdot, \cdot)\text{Id}$ is $\mathcal{C}_0$-universal if $k_{\mathcal{Z}}$ is a universal scalar kernel, which in turn is guaranteed if $k_{\mathcal{Z}}$ is Gaussian or Laplacian, for example (Steinwart, 2001).

The consistency result with optimal rate $\mathcal{O}_p(\frac{\log n}{n})$ in (Grünewälder et al., 2012a, Corollaries 4.1, 4.2) is based on Caponnetto and De Vito (2006), and assumes, as well as some distributional assumptions, that $\mathcal{H}_{\mathcal{X}}$ is finite-dimensional, which is not true for many common choices of $k_{\mathcal{X}}$. In (Song et al., 2009, Theorem 6), (Song et al., 2010b, Theorem 1) and (Fukumizu, 2015, Theorem 1.3.2), consistency is also shown under various assumptions, with rates at best $\mathcal{O}_p(n^{-\frac{1}{4}})$. In Theorem 3.3.4, we prove universal consistency without any distributional assumptions, and in Theorem 3.3.5, we show that a convergence rate of $\mathcal{O}_p(n^{-1/4})$ can be achieved with a simple smoothness assumption that $F_{P_{X|Z}} \in \mathcal{G}_{\mathcal{X}\mathcal{Z}}$ (sometimes referred to as the *well-specified case*; see Szabó et al. (2016)). In particular, both results relax the finite-dimensionality assumption on $\mathcal{H}_{\mathcal{X}}$ of Grünewälder et al. (2012a).

Theorem 3.3.4. *Suppose that $k_{\mathcal{X}}$ and $k_{\mathcal{Z}}$ are bounded kernels, i.e. there are $B_{\mathcal{Z}}, B_{\mathcal{X}} > 0$ with $\sup_{z\in\mathcal{Z}} k_{\mathcal{Z}}(z,z) \leq B^2_{\mathcal{Z}}$, $\sup_{x\in\mathcal{X}} k_{\mathcal{X}}(x,x) \leq B^2_{\mathcal{X}}$, and that the operator-valued kernel $l_{\mathcal{X}\mathcal{Z}}$ is $\mathcal{C}_0$-universal. Let the regularisation parameter λ_n decay to 0 at a slower rate than $\mathcal{O}(n^{-1/2})$. Then the learning algorithm that yields $\hat{F}_{P_{X|Z},n,\lambda_n}$ is universally consistent, i.e. for any joint distribution P_{XZ}, $\epsilon > 0$ and $\delta > 0$, $P_{XZ}(\tilde{\mathcal{E}}_{X|Z}(\hat{F}_{P_{X|Z},n,\lambda_n}) - \tilde{\mathcal{E}}_{X|Z}(F_{P_{X|Z}}) > \epsilon) < \delta$ for sufficiently large n.*

Theorem 3.3.5. *Assume further that $F_{P_{X|Z}} \in \mathcal{G}_{\mathcal{X}\mathcal{Z}}$. Then with probability at least $1-\delta$,*

$$\begin{aligned}\tilde{\mathcal{E}}_{X|Z}(\hat{F}_{P_{X|Z},n,\lambda_n}) - \tilde{\mathcal{E}}_{X|Z}(F_{P_{X|Z}}) &\leq \lambda_n \left\|F_{P_{X|Z}}\right\|^2_{\mathcal{G}_{\mathcal{X}\mathcal{Z}}} \\ &+ \frac{2\ln\left(\frac{4}{\delta}\right)}{3n\lambda_n}\left(1+\sqrt{1+\frac{18n}{\ln\left(\frac{4}{\delta}\right)}}\right) \\ &\quad \left(\left(B_{\mathcal{Z}}\left\|F_{P_{X|Z}}\right\|_{\mathcal{G}_{\mathcal{X}\mathcal{Z}}} + B_{\mathcal{X}}\right)^2 \lambda_n + B^2_{\mathcal{X}}\left(B_{\mathcal{Z}} + \sqrt{\lambda_n}\right)^2\right)\end{aligned}$$

In particular, if $\lambda_n = \mathcal{O}(n^{-1/4})$, then $\tilde{\mathcal{E}}_{X|Z}(\hat{F}_{P_{X|Z},n,\lambda_n}) - \tilde{\mathcal{E}}_{X|Z}(F_{P_{X|Z}}) = \mathcal{O}_p(n^{-1/4})$. The boundedness assumption is satisfied with many commonly used kernels, such as the Gaussian and Laplacian, and hence is not a restrictive condition. Note that some smoothness assumption on $F_{P_{X|Z}}$ or other distributional assumptions are necessary to achieve universal convergence rates, otherwise the rates can be arbitrarily slow – for more discussion, see e.g. (Vapnik, 1998, p.56), (Devroye et al., 1996, p.114, Theorem 7.2) or (Györfi et al., 2006, p.32, Theorem 3.1). It is likely that better (and even optimal) rates can be achieved with further assumptions (see e.g. Caponnetto and De Vito (2006); Steinwart et al. (2009); Blanchard and Mücke (2018) for results with real or finite-dimensional output spaces), but we leave further investigation of learning rates with infinite-dimensional output spaces as future work.

Theorem 3.3.4 is stated with respect to the surrogate loss $\tilde{\mathcal{E}}_{X|Z}$, not the original loss $\mathcal{E}_{X|Z}$. Let us now investigate its implications with respect to the original loss. Write $\eta = \tilde{\mathcal{E}}_{X|Z}(F_{P_{X|Z}})$. Since $\tilde{\mathcal{E}}_{X|Z}(\hat{F}_{P_{X|Z},n,\lambda_n}) \geq \mathcal{E}_{X|Z}(\hat{F}_{P_{X|Z},n,\lambda_n})$, a consequence of Theorem 3.3.4 is that

$$\lim_{n\to\infty} P_{XZ}(\mathcal{E}_{X|Z}(\hat{F}_{P_{X|Z},n,\lambda_n}) > \epsilon + \eta) \leq \lim_{n\to\infty} P_{XZ}(\tilde{\mathcal{E}}_{X|Z}(\hat{F}_{P_{X|Z},n,\lambda_n}) - \eta > \epsilon)$$

$$= 0$$

for any $\epsilon > 0$. This shows that, in the limit as $n \to \infty$, the loss $\mathcal{E}_{X|Z}(\hat{F}_{P_{X|Z},n,\lambda_n})$ is at most an arbitrarily small amount larger than η with high probability.

It remains to investigate what η represents, and how large it is. The law of total expectation gives

$$\eta = \mathbb{E}[\|k_{\mathcal{X}}(X,\cdot) - F_{P_{X|Z}}(Z)\|^2_{\mathcal{H}_{\mathcal{X}}}] = \mathbb{E}[\mathbb{E}[\|k_{\mathcal{X}}(X,\cdot) - \mathbb{E}[k_{\mathcal{X}}(X,\cdot) \mid Z]\|^2_{\mathcal{H}_{\mathcal{X}}} \mid Z]].$$

Here, the integrand $\mathbb{E}[\|k_{\mathcal{X}}(X,\cdot) - \mathbb{E}[k_{\mathcal{X}}(X,\cdot)| \mid Z]\|^2_{\mathcal{H}_{\mathcal{X}}} \mid Z]$ is the *variance* of $k_{\mathcal{X}}(X,\cdot)$ given Z (see (Bharucha-Reid, 1972, p.24) for the definition of the variance of Banach-space valued random variables), and by integrating over $\mathcal{Z}$ in the outer integral, η represents the "expected variance" of $k_{\mathcal{X}}(X,\cdot)$.

Suppose X is measurable with respect to Z, i.e. $F_{P_{X|Z}}$ has no noise. Then $\mathbb{E}[k_{\mathcal{X}}(X,\cdot) \mid Z] = k_{\mathcal{X}}(X,\cdot)$, and consequently, $\eta = 0$. In this case, we have universal consistency in both the surrogate loss $\tilde{\mathcal{E}}_{X|Z}$ and the original loss $\mathcal{E}_{X|Z}$. On the other hand, η will be large if information about Z tells us little about X, and subsequently $k_{\mathcal{X}}(X,\cdot) \in \mathcal{H}_{\mathcal{X}}$. In the extreme case where X and Z are independent, we have $\mathbb{E}[k_{\mathcal{X}}(X,\cdot) \mid Z] = \mathbb{E}[k_{\mathcal{X}}(X,\cdot)]$, and $\eta = \mathbb{E}[\|k_{\mathcal{X}}(X,\cdot) - \mathbb{E}[k_{\mathcal{X}}(X,\cdot)]\|^2_{\mathcal{H}_{\mathcal{X}}}]$, which is precisely the variance of $k_{\mathcal{X}}(X,\cdot)$ in $\mathcal{H}_{\mathcal{X}}$. Hence, η represents the irreducible loss of the true function due to noise in X, and the surrogate loss represents the loss functional taking noise into account, while the original loss measures the deviance from the true conditional expectation.

3.4 Discrepancy between conditional distributions and conditional independence

In this section, we propose conditional analogues of the maximum mean discrepancy (MMD) and the Hilbert-Schmidt independence criterion (HSIC), to measure, respectively, the discrepancy between conditional distributions and conditional independence.

3.4.1 Maximum conditional mean discrepancy

Let $X' : \Omega \to \mathcal{X}$, $Z' : \Omega \to \mathcal{Z}$ be additional random variables, with

$$\int_{\mathcal{X}} \sqrt{k_{\mathcal{X}}(x',x')} dP_{X'}(x') < \infty.$$

Following Theorem 3.3.1, we write $\mu_{P_{X|Z}} = F_{P_{X|Z}} \circ Z$ and $\mu_{P_{X'|Z'}} = F_{P_{X'|Z'}} \circ Z'$.

Definition 3.4.1. The *maximum conditional mean discrepancy* (MCMD) between $P_{X|Z}$ and $P_{X'|Z'}$ is defined as the function $\mathcal{Z} \to \mathbb{R}$ given by

$$M_{P_{X|Z},P_{X'|Z'}}(z) = \|F_{P_{X|Z}}(z) - F_{P_{X'|Z'}}(z)\|_{\mathcal{H}_{\mathcal{X}}}.$$

Using $\{(x_i, z_i)\}_{i=1}^n, \{(x'_j, z'_j)\}_{j=1}^m$ from joint distributions $P_{XZ}, P_{X'Z'}$, we obtain a closed-form, plug-in estimate from (3.4) for the square of the MCMD function as

$$\begin{aligned}\hat{M}^2_{P_{X|Z},P_{X'|Z'}}(\cdot) &= \|\hat{F}_{P_{X|Z},n,\lambda}(\cdot) - \hat{F}_{P_{X'|Z'},m,\lambda'}(\cdot)\|^2_{\mathcal{H}_\mathcal{X}} \\ &= \mathbf{k}_Z^T(\cdot)\mathbf{W}_Z\mathbf{K}_X\mathbf{W}_Z^T\mathbf{k}_Z(\cdot) \\ &\quad - 2\mathbf{k}_Z^T(\cdot)\mathbf{W}_Z\mathbf{K}_{XX'}\mathbf{W}_{Z'}^T\mathbf{k}_{Z'}(\cdot) \\ &\quad\quad + \mathbf{k}_{Z'}^T(\cdot)\mathbf{W}_{Z'}\mathbf{K}_{X'}\mathbf{W}_{Z'}^T\mathbf{k}_{Z'}(\cdot),\end{aligned}$$

where

$$\begin{aligned}[\mathbf{K}_X]_{ij} &= k_\mathcal{X}(x_i, x_j), \\ [\mathbf{K}_{X'}]_{ij} &= k_\mathcal{X}(x'_i, x'_j), \\ [\mathbf{K}_{XX'}]_{ij} &= k_\mathcal{X}(x_i, x'_j), \\ [\mathbf{K}_{Z'}]_{ij} &= k_\mathcal{X}(z'_i, z'_j), \\ \mathbf{k}_{Z'}(\cdot) &= (k_Z(z'_1, \cdot), ..., k_Z(z'_m, \cdot))^T, \\ \mathbf{W}_Z &= (\mathbf{K}_Z + n\lambda\mathbf{I}_n)^{-1}, \\ \mathbf{W}_{Z'} &= (\mathbf{K}_{Z'} + m\lambda'\mathbf{I}_m)^{-1}.\end{aligned}$$

The term "maximum mean discrepancy" stems from the equality $\|\mu_{P_X} - \mu_{P_{X'}}\|_{\mathcal{H}_\mathcal{X}} = \sup_{f\in\mathcal{B}_\mathcal{X}}|\mathbb{E}_X[f(X)] - \mathbb{E}_{X'}[f(X')]|$ (Gretton et al., 2007; Sriperumbudur et al., 2010), where $\mathcal{B}_\mathcal{X} := \{f \in \mathcal{H}_\mathcal{X} \mid \|f\|_{\mathcal{H}_\mathcal{X}} \leq 1\}$. The supremum is attained by the *witness function*, $\frac{\mu_{P_X}-\mu_{P_{X'}}}{\|\mu_{P_X}-\mu_{P_{X'}}\|_{\mathcal{H}_\mathcal{X}}}$ (Gretton et al., 2012). Using Lemma 3.2.2, the analogous (almost sure) equality for the MCMD is $\sup_{f\in\mathcal{B}_\mathcal{X}}|\mathbb{E}_{X|Z}[f(X) \mid Z] - \mathbb{E}_{X'|Z'}[f(X') \mid Z']| = \|\mu_{P_{X|Z}} - \mu_{P_{X'|Z'}}\|_{\mathcal{H}_\mathcal{X}}$. We define the *conditional witness function* as the $\mathcal{H}_\mathcal{X}$-valued random variable

$$\frac{\mu_{P_{X|Z}} - \mu_{P_{X'|Z'}}}{\|\mu_{P_{X|Z}} - \mu_{P_{X'|Z'}}\|_{\mathcal{H}_\mathcal{X}}}.$$

We can informally think of $\text{MCMD}_{P_{X|Z},P_{X'|Z'}}(z)$ as "MMD between $P_{X|Z=z}$ and $P_{X'|Z'=z}$". However, we do not have i.i.d. samples from $P_{X|Z=z}$ and $P_{X'|Z'=z}$, and hence the estimation cannot be done by U- or V-statistic procedures as done for the MMD. The following theorem says that, with characteristic kernels, the MCMD can indeed act as a discrepancy measure between conditional distributions.

Theorem 3.4.2. *Suppose that $k_\mathcal{X}$ is characteristic, that P_Z and $P_{Z'}$ are absolutely continuous with respect to each other, and that $P(\cdot \mid Z)$ and $P(\cdot \mid Z')$ admit regular versions. Then $M_{P_{X|Z},P_{X'|Z'}} = 0$ almost everywhere if and only if, for almost all $z \in \mathcal{Z}$, $P_{X|Z=z}(B) = P_{X'|Z'=z}(B)$ for all $B \in \mathfrak{X}$.*

By (Çınlar, 2011, p.11 & p.151, Theorem 2.10), we know that the space $(\Omega, \mathcal{F})$ being a Polish space with its Borel σ-algebra is a sufficient condition for

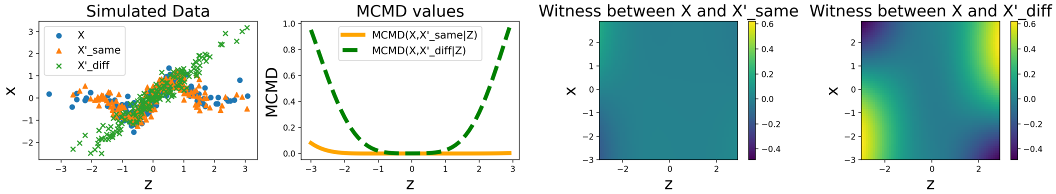


Figure 3.1: We see that $\mathrm{MCMD}(X, X'_{\text{same}}|Z) \approx 0\ \forall Z$. Near $Z = 0$, where the dependence on Z of X and X'_{diff} are similar, $\mathrm{MCMD}(X, X'_{\text{diff}}|Z) \approx 0$, whereas away from 0, the dependence on Z of X and X'_{diff} are different, and so $\mathrm{MCMD}(X, X'_{\text{diff}}|Z) > 0$. We also see that the conditional witness function between X and X'_{same} gives 0 at all values of X given any value of Z, whereas we have a saddle-like function between X and X'_{diff}, with non-zero functions in X in the regions of Z away from 0.

$P(\cdot \mid \mathcal{E})$ to have a regular version for any sub-σ-algebra $\mathcal{E}$ of $\mathcal{F}$. Hence, the assumption that $P(\cdot \mid Z)$ admits a regular version is not a restrictive one.

The MCMD is reminiscent of the *conditional maximum mean discrepancy* of (Ren et al., 2016), defined as the Hilbert-Schmidt norm of the operator $\mathcal{U}_{X|Z} - \mathcal{U}_{X'|Z}$ (see Definition 3.2.4). However, due to previously discussed assumptions, $\mathcal{U}_{X|Z}$ and $\mathcal{U}_{X'|Z}$ often do not even exist, and/or do not have the desired properties of Theorem 3.2.5, so even at population level, $\mathcal{U}_{X|Z} - \mathcal{U}_{X'|Z}$ is often not an exact measure of discrepancy between conditional distributions, unlike the MCMD. Moreover, Ren et al. (2016) only considers the case when the conditioning variable is the same.

3.4.2 Hilbert-Schmidt conditional independence criterion

In this subsection, we introduce a novel criterion of conditional independence.

Definition 3.4.3. We define the *Hilbert-Schmidt Conditional Independence Criterion* between X and Y given Z to be $\mathrm{HSCIC}(X, Y \mid Z) = \|\mu_{P_{XY|Z}} - \mu_{P_{X|Z}} \otimes \mu_{P_{Y|Z}}\|_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}}$.

We can write $\mathrm{HSCIC}(X, Y \mid Z) = H_{X,Y|Z} \circ Z$ for some $H_{X,Y|Z} : \mathcal{Z} \to \mathbb{R}$. Given a sample $\{(x_i, y_i, z_i)\}_{i=1}^n$ from P_{XYZ}, we obtain a plug-in, closed-form estimate of $H^2_{X,Y|Z}(\cdot)$ as follows:

$$\begin{aligned}\hat{H}^2_{X,Y|Z}(\cdot) = \mathbf{k}_Z^T(\cdot)\mathbf{W}(\mathbf{K}_X \odot \mathbf{K}_Y)\mathbf{W}^T\mathbf{k}_Z(\cdot) \\ - 2\mathbf{k}_Z^T(\cdot)\mathbf{W}((\mathbf{K}_X\mathbf{W}^T\mathbf{k}_Z(\cdot)) \odot (\mathbf{K}_Y\mathbf{W}^T\mathbf{k}_Z(\cdot))) \\ + (\mathbf{k}_Z^T(\cdot)\mathbf{W}\mathbf{K}_X\mathbf{W}^T\mathbf{k}_Z(\cdot))(\mathbf{k}_Z^T(\cdot)\mathbf{W}\mathbf{K}_Y\mathbf{W}^T\mathbf{k}_Z(\cdot))\end{aligned}$$

where $[\mathbf{K}_Y]_{ij} := k_\mathcal{Y}(y_i, y_j)$ and $\odot$ denotes elementwise multiplication of matrices.

Casting aside measure-theoretic issues arising from conditioning on an event of probability 0, we can conceptually think of the realisation of the HSCIC at each $z = Z(\omega)$ as "the HSIC between $P_{X|Z=z}$ and $P_{Y|Z=z}$". Again, we do not have multiple samples from each distribution $P_{X|Z=z}$ and $P_{Y|Z=z}$, so the estimation cannot be done by U- or V-statistic procedures as done for HSIC. The following theorem shows that HSCIC is a measure of conditional independence.

Theorem 3.4.4. *Suppose $k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$ is a characteristic kernel*[2] *on $\mathcal{X} \times \mathcal{Y}$, and that $P(\cdot \mid Z)$ admits a regular version. Then* $\mathrm{HSCIC}(X, Y \mid Z) = 0$ *almost surely if and only if $X \perp\!\!\!\perp Y \mid Z$.*

Sheng and Sriperumbudur (2019) also proposed a similar criterion with the same nomenclature (HSCIC). However, they omit the discussion of CMEs entirely, and define the HSCIC as the HSIC between $P_{XY|Z=z}$ and $P_{X|Z=z}P_{Y|Z=z}$, without considerations for conditioning on an event of measure 0. Their focus is more on investigating connections to distance-based measures (Wang et al., 2015; Sejdinovic et al., 2013). Fukumizu et al. (2008) propose I^{COND}, defined as the squared Hilbert-Schmidt norm of the normalised conditional cross-covariance operator $V_{\ddot{Y}\ddot{X}|Z} := \mathcal{C}_{\ddot{Y}\ddot{Y}}^{-1/2}\Sigma_{\ddot{Y}\ddot{X}|Z}\mathcal{C}_{\ddot{X}\ddot{X}}^{-1/2}$, where $\ddot{X} := (X, Z)$ and $\ddot{Y} := (Y, Z)$. As discussed, these operator-based definitions rely on a number of strong assumptions that will often mean that $V_{\ddot{Y}\ddot{X}|Z}$ does not exist, or it does not satisfy the conditions for it to be used as an exact criterion even at population level. On the other hand, the HSCIC defined as in Definition 3.4.3 is an exact mathematical criterion of conditional independence at population level. Note that I^{COND} is a single-value criterion, whereas the HSCIC is a random criterion.

3.4.3 Experiments

We carry out simulations to demonstrate the behaviour of the MCMD and HSCIC. In all simulations, we use the Gaussian kernel $k_{\mathcal{X}}(x, x') = k_{\mathcal{Y}}(x, x') = k_{\mathcal{Z}}(x, x') = e^{-\frac{1}{2}\sigma_X \|x-x'\|_2^2}$ with hyperparameter $\sigma_X = 0.1$, and regularisation parameter $\lambda = 0.01$.

In Figure 3.1, we simulate 500 samples from

$$\begin{aligned} Z, Z' &\sim \mathcal{N}(0,1) \\ X &= e^{-0.5Z^2}\sin(2Z) + N_X \\ X'_{\text{same}} &= e^{-0.5Z'^2}\sin(2Z') + N_X \\ X'_{\text{diff}} &= Z' + N_X, \end{aligned}$$

where $N_X \sim 0.3\mathcal{N}(0, 1)$ is the (additive) noise variable. The first plot shows simulated data, the second MCMD values against Z, and the heatmaps show the (unnormalised) conditional witness function, whose norm gives the MCMD.

[2] See (Szabó and Sriperumbudur, 2017) for a detailed discussion on characteristic tensor product kernels.

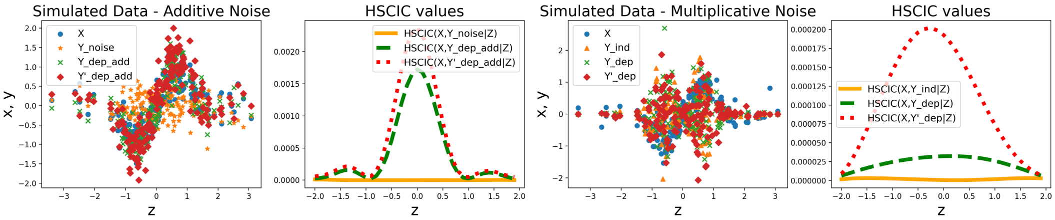


Figure 3.2: We see that $\mathrm{HSCIC}(X, Y_{\mathrm{noise}}|Z) \approx 0$ (left) and $\mathrm{HSCIC}(X, Y_{\mathrm{ind}}|Z) \approx 0$ (right) for all Z, whereas $\mathrm{HSCIC}(X, Y_{\mathrm{dep\_add}}|Z) > 0$, $\mathrm{HSCIC}(X, Y'_{\mathrm{dep\_add}}|Z) > 0$, $\mathrm{HSCIC}(X, Y_{\mathrm{dep}}|Z) > 0$, $\mathrm{HSCIC}(X, Y'_{\mathrm{dep}}|Z) > 0$. In particular, the dependence of $Y'_{\mathrm{dep\_add}}$ and Y'_{dep} on X is greater than that of $Y_{\mathrm{dep\_add}}$ and Y_{dep}, and is represented by larger values of $\mathrm{HSCIC}(X, Y'_{\mathrm{dep\_add}}|Z)$ and $\mathrm{HSCIC}(X, Y'_{\mathrm{dep}}|Z)$ compared to $\mathrm{HSCIC}(X, Y_{\mathrm{dep}}|Z)$ and $\mathrm{HSCIC}(X, Y_{\mathrm{dep\_add}}|Z)$.

In Figure 3.2, on the left, we simulate 500 samples from the additive noise model,

$$\begin{aligned}
Z &\sim \mathcal{N}(0,1)\\
X &= e^{-0.5Z^2}\sin(2Z) + N_X\\
Y_{\mathrm{noise}} &= N_Y\\
Y_{\mathrm{dep\_add}} &= e^{-0.5Z^2}\sin(2Z) + N_X + 0.2X\\
Y'_{\mathrm{dep\_add}} &= e^{-0.5Z^2}\sin(2Z) + N_X + 0.4X,
\end{aligned}$$

where $N_X \sim 0.3\mathcal{N}(0,1)$ is the (additive) noise variable. On the right, we simulate 500 samples from the multiplicative noise model,

$$\begin{aligned}
Z &\sim \mathcal{N}(0,1)\\
X = Y_{\mathrm{ind}} &= e^{-0.5Z^2}\sin(2Z)N_X\\
Y_{\mathrm{dep}} &= e^{-0.5Z^2}\sin(2Z)N_Y + 0.2X\\
Y'_{\mathrm{dep}} &= e^{-0.5Z^2}\sin(2Z)N_Y + 0.4X,
\end{aligned}$$

where $N_X, N_Y \sim 0.3\mathcal{N}(0,1)$ are the (multiplicative) noise variables.

Chapter 4

Kernel Regression for Treatment Effect

4.1 Problem Set-Up

As in Chapter 3, we take $(\Omega, \mathcal{F}, P)$ as the underlying probability space, $\mathcal{X}$ as the input space and $\mathcal{Y} \subseteq \mathbb{R}$ as the output space. Let $Z : \Omega \to \{0, 1\}$, $X : \Omega \to \mathcal{X}$ and $Y_0, Y_1, Y : \Omega \to \mathcal{Y}$ be random variables representing, respectively, the treatment assignment, covariates, the potential outcomes under control and treatment, and the observed outcome, i.e. $Y = Y_0(1 - Z) + Y_1 Z$. For example, Z may indicate whether a subject is administered a medical treatment ($Z = 1$) or not ($Z = 0$). The potential outcomes Y_1, Y_0 respectively correspond to subject's responses had they received treatment or not. The covariates X correspond to subject's characteristics such as age, gender, race that could influence both the potential outcomes and the choice of treatment. We denote the distributions of random variables by subscripting P, e.g. P_X for the distribution of X. We also impose, as we did in Chapter 3, the mild condition that conditional distribution $P(\cdot \mid X)$ admits a *regular version* (Çınlar, 2011, p.150, Definition 2.4, Proposition 2.5).

Each unit $i = 1, ..., n$ is associated with an independent copy $(X_i, Z_i, Y_{0i}, Y_{1i})$ of (X, Z, Y_0, Y_1). However, for each $i = 1, ..., n$, we observe either Y_{0i} or Y_{1i}; this missing value problem is known as the *fundamental problem of causal inference* (Holland, 1986), preventing us from directly computing the difference in the outcomes under treatment and control for each unit. As a result, we only have access to samples $\{(x_i, z_i, y_i)\}_{i=1}^n$ of (X, Z, Y). We write $n_0 = \sum_{i=1}^n \mathbf{1}_{z_i=0}$ and $n_1 = \sum_{i=1}^n \mathbf{1}_{z_i=1}$ for the control and treatment sample sizes, and denote the control and treatment samples by $\{(x_i^0, y_i^0)\}_{i=1}^{n_0}$ and $\{(x_i^1, y_i^1)\}_{i=1}^{n_1}$.

We assume *strong ignorability* Rosenbaum and Rubin (1983):

unconfoundedness $Z \perp\!\!\!\perp (Y_0, Y_1) \mid X$; and

overlap $0 < e(X) = P(Z = 1 \mid X) = \mathbb{E}[Z \mid X] < 1$.

Causal treatment effects are then identifiable from observational data, since

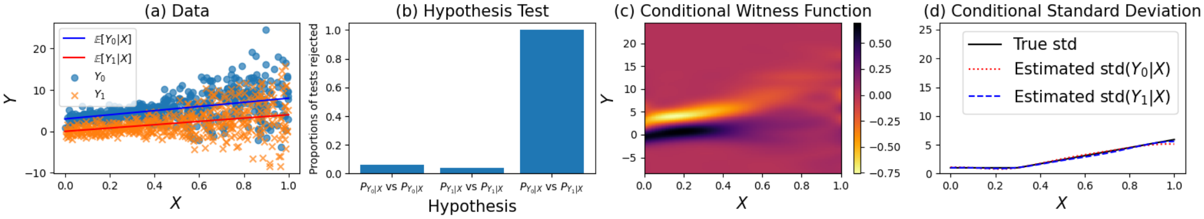


Figure 4.1: **Toy illustration of higher-order heterogeneity that cannot be captured by CATE. (a) Data.** $X \sim \text{Uniform}[0,1]$, $Y_0 = 3 + 5X + \mathbf{1}_{X<0.3}N + 7\mathbf{1}_{X\geq 0.3}(1 + (X - 0.3))N$ and $Y_1 = 4X + \mathbf{1}_{X<0.3}N + 7\mathbf{1}_{X\geq 0.3}(1 + (X - 0.3))N$, where $N \sim \mathcal{N}(0,1)$; in particular, the CATE is increasing with X. **(b) Hypothesis test** (Section 4.3.2) Each of the hypotheses $P_{Y_0|X} \equiv P_{Y_0|X}$, $P_{Y_1|X} \equiv P_{Y_1|X}$ and $P_{Y_0|X} \equiv P_{Y_1|X}$ are tested 100 times. The last (false) hypothesis is rejected in most tests, while the first two (true) hypotheses are not rejected in most tests, meaning that both type I and type II errors are low. **(c) Conditional witness function** (Section 4.4.1). The conditional witness function is close to zero for all Y at $X \geq 0.5$, demonstrating that $P_{Y_0|X}$ and $P_{Y_1|X}$ are similar in this region of $\mathcal{X}$. For $X < 0.4$, the witness function is positive in regions where the density of Y_1 is higher than that of Y_0, and negative in regions where the density of Y_0 is higher than that of Y_1. **(d) U-statistic regression** (Section 4.4.2). True conditional standard deviation (in black) is estimated (in red and blue for control and treatment groups respectively) as a function of X via U-statistic regression (since variance is a U-statistic) and the square-root operation. We see that the standard deviation increases linearly for $X \geq 0.3$.

$P_{Y_0|X} = P_{Y_0|X,Z=0} = P_{Y|X,Z=0}$, and similarly for $P_{Y_1|X}$. The quantity $e(X)$ is the *propensity score*. In a *randomised experiment*, $e(X)$ is known and controlled (Imbens and Rubin, 2015, p.40, Definition 3.10).

The usual objects of interest in the treatment effect literature are the *average treatment effect* (ATE), $\mathbb{E}[Y_1 - Y_0]$, and the *conditional average treatment effect* (CATE), $T(x) = \mathbb{E}[Y_1 - Y_0 \mid X = x]$. We propose to extend the analysis to compare other aspects of the conditional distributions, $P_{Y_0|X}$ and $P_{Y_1|X}$. One compelling reason to do this is that estimating CATE is inherently a problem of *comparing two means*, and as such, is only meaningful if the corresponding variances are given. Consider the toy example in Figure 4.1. The CATE is constructed to be increasing with X, but taking into account the variance, the treatment effect is clearly more pronounced for small values of X. For example, the probability of Y_1 being greater than Y_0 is much higher for smaller values of X.

Beyond the mean and variance, researchers may also be interested in other higher-moment treatment effect heterogeneity, such as Gini's mean difference or skewness, or indeed how the entire conditional densities of the control and treatment groups differ given the covariates, in an exploratory fashion. Panels (b), (c) and (d) in Figure 4.1 demonstrate each of the steps we propose in this paper applied to this toy dataset: hypothesis testing of equality of conditional distributions, the conditional witness function and U-statistic regression (variance, in this instance), respectively.

4.1.1 U-Statistics

Suppose $Y_1, Y_2, ..., Y_r$ are independent copies of the random variable Y, i.e. they are independent and all have distribution P_Y. Let $h : \mathcal{Y}^r \to \mathbb{R}$ be a symmetric function (called a *kernel* in the U-statistics literature; confusion must be avoided with the reproducing kernel used throughout this paper), i.e. for any permutation π of $\{1, ..., r\}$, we have $h(y_1, ..., y_r) = h(y_{\pi(1)}, ..., y_{\pi(r)})$. Suppose we would like to estimate a function of the form

$$\theta(P_Y) = \mathbb{E}\left[h\left(Y_1, ..., Y_r\right)\right] = \int_{\mathcal{Y}} ... \int_{\mathcal{Y}} h\left(y_1, ..., y_r\right) dP_Y(y_1)...dP_Y(y_r).$$

The corresponding *U-statistic* for an unbiased estimation of $\theta(P_Y)$ based on a sample $Y_1, ..., Y_n$ of size $n \geq r$ is given by

$$\hat{\theta}(P_Y) = \frac{1}{\binom{n}{r}} \sum h\left(Y_{i_1}, ..., Y_{i_r}\right),$$

where $\binom{n}{r}$ is the binomial coefficient and the summation is over the $\binom{n}{r}$ combinations of r distinct elements $\{i_1, ..., i_r\}$ from $\{1, ..., n\}$. Clearly, since the expectation of each summand yields $\theta(P_Y)$, we have $\mathbb{E}[\hat{\theta}(P_Y)] = \theta(P_Y)$, so U-statistics are unbiased estimators.

Some examples of h and the corresponding estimator include the sample mean $h(y) = y$, the sample variance $h(y_1, y_2) = \frac{1}{2}(y_1 - y_2)^2$, the sample cumulative distribution up to y^* $h(y) = \mathbf{1}(y \leq y^*)$, the k^{th} sample raw moment $h(y) = y^k$ and Gini's mean difference $h(y_1, y_2) = |y_1 - y_2|$.

To the best of our knowledge, Stute (1991) was the first to consider a conditional counterpart of U-statistics. Let $X_1, ..., X_r$ be independent copies of the random variable X. We are now interested in the estimation of the following quantity:

$$\theta\left(P_{Y|X}\right) = \mathbb{E}\left[h\left(Y_1, ..., Y_r\right) \mid X_1, ..., X_r\right].$$

By Çınlar (2011, p.146, Theorem 1.17), $\theta(P_{Y|X})$ can be considered as a function $\mathcal{X}^r \to \mathbb{R}$, such that for each r-tuple $\{x_1, ..., x_r\}$, we have

$$\theta\left(P_{Y|X}\right)(x_1, ..., x_r) = \mathbb{E}\left[h\left(Y_1, ..., Y_r\right) \mid X_1 = x_1, ..., X_r = x_r\right].$$

The simplest case is when $r = 1$ and $h(y) = y$. In this case, the estimand reduces to $f(X) = \mathbb{E}[Y|X]$, which is the usual regression problem for which a plethora of methods exist. Suppose we have a sample $\{(X_i, Y_i)\}_{i=1}^n$. One such regression method is the Nadaraya-Watson kernel smoother:

$$\hat{f}(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x - X_i}{a}\right)}{\sum_{i=1}^n K\left(\frac{x - X_i}{a}\right)},$$

where K is the so-called “smoothing kernel” and a is the bandwidth. This was extended by Stute (1991) to $r \geq 1$ and more general h:

$$\hat{\theta}\left(P_{Y|X}\right)\left(x_1, ..., x_r\right) = \frac{\sum h\left(Y_{i_1}, ..., Y_{i_r}\right) \prod_{j=1}^{r} K\left(\frac{x_j - X_{i_j}}{a}\right)}{\sum \prod_{j=1}^{r} K\left(\frac{x_j - X_{i_j}}{a}\right)},$$

where the sums are over the $\binom{n}{r}$ combinations of r distinct elements $\{i_1, ..., i_r\}$ from $\{1, ..., n\}$ as before. Derumigny (2019) considers a parametric model of the form

$$\Lambda\left(\theta\left(P_{Y|X}\right)\left(x_1, ..., x_r\right)\right) = \boldsymbol{\psi}\left(x_1, ..., x_r\right)^T \beta^*,$$

where Λ is a strictly increasing and continuously differentiable “link function” such that the range of $\Lambda \circ \theta$ is exactly $\mathbb{R}$, $\beta^* \in \mathbb{R}^s$ is the true parameter and $\boldsymbol{\psi}(\cdot) = \left(\psi_1(\cdot), ..., \psi_s(\cdot)\right)^T \in \mathbb{R}^s$ is some basis, such as polynomials, exponentials, indicator functions etc. However, the estimation of β^* still makes use of the Nadaraya-Watson kernel smoothers considered above.

Of course, Nadaraya-Watson kernel smoothers are far from being the only method of regression that can be extended to estimate conditional U-statistics, and in the main body of the paper (Section 4.4.2), we consider extending kernel ridge regression for this purpose.

4.2 Conditional Distributional Treatment Effect

In this section, we generalise the notion of CATE to account for distributional differences between treatment and control groups, rather than just the mean difference.

Definition 4.2.1. Let D be some distance function between probability measures. We define the *conditional distributional treatment effect* (CoDiTE) associated with D as

$$U_D(x) = D(P_{Y_0|X=x}, P_{Y_1|X=x}).$$

Here, the choice of D depends on what characterisation of distributions is used. For example, if $D(P_{Y_0|X=x}, P_{Y_1|X=x}) = \mathbb{E}[Y_1 \mid X = x] - \mathbb{E}[Y_0 \mid X = x]$, we recover the CATE, i.e. $U_D(x) = T(x)$, thereby showing that the CoDiTE is a strict generalisation of the CATE. Different choices of D will require different estimators.

The usual performance metric of a CATE estimator $\hat{T}$ is the *precision of estimating heterogeneous effects* (PEHE) (first proposed in sample form by Hill (2011, Section 4.3); we report the population-level definition, found in, for example, Alaa and Van Der Schaar (2019, Eqn. (5)):

$$\|\hat{T} - T\|_2^2 = \mathbb{E}[|\hat{T}(X) - T(X)|^2].$$

We propose a performance metric of an estimator of the CoDiTE in an exactly analogous manner.

Definition 4.2.2. Given a distance function D, for an estimator $\hat{U}_D$ of U_D, we define the *precision of estimating heterogeneous distributional effects* (PEHDE) as

$$\psi_D(\hat{U}_D) = \|\hat{U}_D - U_D\|_2^2 = \mathbb{E}[|\hat{U}_D(X) - U_D(X)|^2].$$

Again, if D measures the difference in expectations, then the associated PEHDE ψ_D reduces to the usual PEHE.

Henceforth, we explore different choices of the distance function D, as well as methods of estimating the corresponding CoDiTE U_D, to answer the following questions:

Q1 Are $P_{Y_0|X}$ and $P_{Y_1|X}$ different? In other words, is there any distributional effect of the treatment? (Section 4.3)

Q2 If so, how does the distribution of the treatment group differ from that of the control group? (Section 4.4)

4.3 CoDiTE associated with MMD via CMEs

In this section, we answer Q1, i.e. we investigate whether the treatment has any effect at all. To this end we choose D to be the MMD with the associated kernel l being characteristic. Then writing $\mu_{Y_0|X}$ and $\mu_{Y_1|X}$ for the CMEs of Y_0 and Y_1 given X respectively (c.f. Section 3.1.1), we have

$$\begin{aligned} U_{\text{MMD}}(x) &= \text{MMD}(P_{Y_0|X=x}, P_{Y_1|X=x}) \\ &= \|\mu_{Y_1|X=x} - \mu_{Y_0|X=x}\|_{\mathcal{H}}. \end{aligned} \tag{4.1}$$

Since l is characteristic, $P_{Y_0|X=x}$ and $P_{Y_1|X=x}$ are the same distribution if and only if $\text{MMD}(P_{Y_0|X=x}, P_{Y_1|X=x}) = 0$. What makes the MMD a particularly convenient choice is that for each $x \in \mathcal{X}$, $P_{Y_0|X=x}$ and $P_{Y_1|X=x}$ are represented by individual elements $\mu_{Y_0|X=x}$ and $\mu_{Y_1|X=x}$ in the RKHS $\mathcal{H}$, which means that we can estimate the associated CoDiTE simply by performing regression with $\mathcal{X}$ as the input space and $\mathcal{H}$ as the output space, as will be shown in the next section.

4.3.1 Estimation and Consistency

We now discuss how to obtain empirical estimates of $U_{\text{MMD}}(x)$. Recall that, by the unconfoundedness assumption, we can estimate $\mu_{Y_0|X}$ and $\mu_{Y_1|X}$ separately from control and treatment samples respectively. We perform operator-valued kernel regression in separate vector-valued RKHSs $\mathcal{G}_0$ and $\mathcal{G}_1$, endowed with kernels $\Gamma_0(\cdot,\cdot) = k_0(\cdot,\cdot)\text{Id}$ and $\Gamma_1(\cdot,\cdot) = k_1(\cdot,\cdot)\text{Id}$, where $k_0, k_1 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are scalar-valued kernel and $\text{Id} : \mathcal{H} \to \mathcal{H}$ is the identity operator. Following Section 3.3, the empirical estimates $\hat{\mu}_{Y_0|X}$ and $\hat{\mu}_{Y_1|X}$ of $\mu_{Y_0|X}$ and $\mu_{Y_1|X}$ are constructed, for each $x \in \mathcal{X}$, as

$$\begin{aligned} &\hat{\mu}_{Y_0|X=x} = \boldsymbol{k}_0^T(x)\mathbf{W}_0\boldsymbol{l}_0 \in \mathcal{G}_0 \\ \text{and} \quad &\hat{\mu}_{Y_1|X=x} = \boldsymbol{k}_1^T(x)\mathbf{W}_1\boldsymbol{l}_1 \in \mathcal{G}_1, \end{aligned} \tag{4.2}$$

where

$$\begin{aligned}
\mathbf{W}_0 &= (\mathbf{K}_0 + n_0\lambda^0_{n_0}\mathbf{I}_{n_0})^{-1},\\
\mathbf{W}_1 &= (\mathbf{K}_1 + n_1\lambda^1_{n_1}\mathbf{I}_{n_1})^{-1},\\
[\mathbf{K}_0]_{1\le i,j\le n_0} &= k_0(x^0_i, x^0_j),\\
[\mathbf{K}_1]_{1\le i,j\le n_1} &= k_1(x^1_i, x^1_j),\\
\lambda^0_{n_0}, \lambda^1_{n_1} &> 0 \text{ are regularisation parameters},\\
\mathbf{I}_{n_0}, \mathbf{I}_{n_1} &\text{ are identity matrices},\\
\boldsymbol{k}_0(x) &= (k_0(x^0_1, x), ..., k_0(x^0_{n_0}, x))^T,\\
\boldsymbol{k}_1(x) &= (k_1(x^1_1, x), ..., k_1(x^1_{n_1}, x))^T,\\
\boldsymbol{l}_0 &= (l(y^0_1, \cdot), ..., l(y^0_{n_0}, \cdot))^T,\\
\boldsymbol{l}_1 &= (l(y^1_1, \cdot), ..., l(y^1_{n_1}, \cdot))^T.
\end{aligned}$$

By plugging in the estimates (4.2) in the expression (4.1) for U_{MMD}, we can construct $\hat{U}_{\text{MMD}}$ as

$$\hat{U}_{\text{MMD}}(x) = \|\hat{\mu}_{Y_1|X=x} - \hat{\mu}_{Y_0|X=x}\|_{\mathcal{H}}.$$

The next lemma establishes a closed-form expression for $\hat{U}_{\text{MMD}}$ based on the control and treatment samples.

Lemma 4.3.1. *For each $x \in \mathcal{X}$, we have*

$$\begin{aligned}
\hat{U}^2_{\text{MMD}}(x) = &\boldsymbol{k}^T_0(x)\mathbf{W}_0\mathbf{L}_0\mathbf{W}^T_0\boldsymbol{k}_0(x)\\
&- 2\boldsymbol{k}^T_0(x)\mathbf{W}_0\mathbf{L}\mathbf{W}^T_1\boldsymbol{k}_1(x)\\
&+ \boldsymbol{k}^T_1(x)\mathbf{W}_1\mathbf{L}_1\mathbf{W}^T_1\boldsymbol{k}_1(x),
\end{aligned}$$

where $[\mathbf{L}_0]_{1\le i,j\le n_0} = l(y^0_i, y^0_j)$, $[\mathbf{L}]_{1\le i\le n_0, 1\le j\le n_1} = l(y^0_i, y^1_j)$ and $[\mathbf{L}_1]_{1\le i,j\le n_1} = l(y^1_i, y^1_j)$.

The next theorem shows that, using *universal kernels* Γ_0, Γ_1 (Carmeli et al., 2010, Definition 4.1), $\hat{U}_{\text{MMD}}$ is universally consistent with respect to the PEHDE.

Theorem 4.3.2 (Universal consistency)**.** *Suppose that k_0, k_1 and l are bounded, that Γ_0 and Γ_1 are universal, and that $\lambda^0_{n_0}$ and $\lambda^1_{n_1}$ decay at slower rates than $\mathcal{O}(n_0^{-1/2})$ and $\mathcal{O}(n_1^{-1/2})$ respectively. Then as $n_0, n_1 \to \infty$,*

$$\psi_{\text{MMD}}(\hat{U}_{\text{MMD}}) = \mathbb{E}[(\hat{U}_{\text{MMD}}(X) - U_{\text{MMD}}(X))^2] \xrightarrow{p} 0.$$

4.3.2 Statistical Hypothesis Testing

We are interested in whether or not the two conditional distributions $P_{Y_0|X}$ and $P_{Y_1|X}$, corresponding to control and treatment, are equal. The hypotheses are then

Algorithm 1 Kernel conditional discrepancy (KCD) test of conditional distributional treatment effect

```
Input: data $\{(x_i, z_i, y_i)\}_{i=1}^n$, significant level $\alpha$, kernels $k_0, k_1, l$, regularisation parameters $\lambda^0_{n_0}, \lambda^1_{n_1}$, no. of permutations $m$.
Calculate $\hat{t}$ using Lemma 4.3.4 based on the input data.
KLR of $\{z_i\}_{i=1}^n$ against $\{x_i\}_{i=1}^n$ to obtain $\hat{e}(x_i)$.
for $k = 1$ to $m$ do
  For each $i = 1, ..., n$, sample $\tilde{z}_i \sim \text{Bernoulli}(\hat{e}(x_i))$.
  Calculate $\hat{t}_k$ from the new dataset $\{x_i, \tilde{z}_i, y_i\}_{i=1}^n$.
end for
Calculate the $p$-value as $p = \frac{1+\sum_{l=1}^m \mathbf{1}\{\hat{t}_l > \hat{t}\}}{1+m}$.
if $p < \alpha$ then
  Reject $H_0$.
end if
```

H_0**:** $P_{Y_0|X=x}(\cdot) = P_{Y_1|X=x}(\cdot)$ P_X-almost everywhere.

H_1**:** There exists $A \subseteq \mathcal{X}$ of positive measure such that $P_{Y_0|X=x}(\cdot) \neq P_{Y_1|X=x}(\cdot)$ for all $x \in A$.

The null hypothesis H_0 means that the treatment has no effect for any of the covariates, whereas the alternative hypothesis H_1 means that the treatment has an effect on *some* of the covariates, where the effect is distributional. For notational simplicity, we write $P_{Y_0|X} \equiv P_{Y_1|X}$ if H_0 holds.

We use the following criterion for $P_{Y_0|X} \equiv P_{Y_1|X}$, which we call the *kernel conditional discrepancy* (KCD):

$$t = \mathbb{E}[\|\mu_{Y_1|X} - \mu_{Y_0|X}\|^2_{\mathcal{H}}].$$

The following lemma tells us that t can indeed be used as a criterion of $P_{Y_0|X} \equiv P_{Y_1|X}$.

Lemma 4.3.3. *If l is a characteristic kernel, $P_{Y_0|X} \equiv P_{Y_1|X}$ if and only if $t = 0$.*

Next, we define a plug-in estimate $\hat{t}$ of t, which we will use as the test statistic of our hypothesis test:

$$\hat{t} = \frac{1}{n}\sum_{i=1}^n \left\|\hat{\mu}_{Y_1|X=x_i} - \hat{\mu}_{Y_0|X=x_i}\right\|^2_{\mathcal{H}}.$$

Then we have a closed-form expression for $\hat{t}$ as follows.

Lemma 4.3.4. *We have*

$$\hat{t} = \frac{1}{n}\text{Tr}\left(\tilde{\mathbf{K}}_0\mathbf{W}_0\mathbf{L}_0\mathbf{W}_0^T\tilde{\mathbf{K}}_0^T\right)$$

$$
\begin{aligned}
&- \frac{2}{n}\mathrm{Tr}\left(\tilde{\mathbf{K}}_0\mathbf{W}_0\mathbf{L}\mathbf{W}_1^T\tilde{\mathbf{K}}_1^T\right)\\
&+ \frac{1}{n}\mathrm{Tr}\left(\tilde{\mathbf{K}}_1\mathbf{W}_1\mathbf{L}_1\mathbf{W}_1^T\tilde{\mathbf{K}}_1^T\right),
\end{aligned}
$$

where $\mathbf{L}_0, \mathbf{L}_1$ *and* $\mathbf{L}$ *are as defined in Lemma 4.3.1 and* $[\tilde{\mathbf{K}}_0]_{1\leq i\leq n, 1\leq j\leq n_0} = k_0(x_i, x_j^0)$ *and* $[\tilde{\mathbf{K}}_1]_{1\leq i\leq n, 1\leq j\leq n_1} = k_1(x_i, x_j^1)$.

The consistency of $\hat{t}$ in the limit of infinite data is shown in the following theorem.

Theorem 4.3.5. *Under the same assumptions as in Theorem 4.3.2, we have* $\hat{t} \xrightarrow{p} t$ *as* $n_0, n_1 \to \infty$.

Unfortunately, it is extremely difficult to compute the (asymptotic) null distribution of $\hat{t}$ analytically, and so we resort to resampling the treatment labels to simulate the null distribution. To ensure that our resampling scheme respects the control and treatment covariate distributions $P_{X|Z=0}$ and $P_{X|Z=1}$, we follow the conditional resampling scheme of Rosenbaum (1984). We first estimate the propensity score $e(x_i)$ for each datapoint x_i (e.g. using kernel logistic regression (KLR) Zhu and Hastie (2005); Marteau-Ferey et al. (2019)), and then resample each data label from this estimated propensity score. By repeating this resampling procedure and computing the test statistic on each resampled dataset, we can simulate from the null distribution of the test statistic. Finally, the test statistic computed from the original dataset is compared to this simulated null distribution, and the null hypothesis is rejected or not rejected accordingly. The exact procedure is summarised in Algorithm 1.

4.4 Understanding the CoDiTE

After determining *whether* $P_{Y_0|X}$ and $P_{Y_1|X}$ are different via MMD-associated CoDiTE and hypothesis testing, we now turn to Q2, i.e. we investigate *how* they are different.

4.4.1 Conditional Witness Functions

For two real-valued random variables, the witness function between them is a useful tool for visualising where their densities differ, without explicitly estimating the densities (Gretton et al., 2012, Figure 1; Lloyd and Ghahramani, 2015, Figure 1). We extend this to the conditional case with the (unnormalised) *conditional witness function* $\mu_{Y_1|X} - \mu_{Y_0|X}$.

Let us fix $x \in \mathcal{X}$. The witness function between $P_{Y_1|X=x}$ and $P_{Y_0|X=x}$ is $\mu_{Y_1|X=x} - \mu_{Y_0|X=x} : \mathcal{Y} \to \mathbb{R}$. For $y \in \mathcal{Y}$ in regions where the density of $P_{Y_1|X=x}$ is greater than that of $P_{Y_0|X=x}$, we have $\mu_{Y_1|X=x}(y) - \mu_{Y_0|X=x}(y) > 0$. For y in regions where the converse is true, we similarly have $\mu_{Y_1|X=x}(y) - \mu_{Y_0|X=x}(y) <$

0. The greater the difference in density, the greater the magnitude of the witness function. For each $y \in \mathcal{Y}$, the associated CoDiTE is

$$U_{\text{witness},y}(x) = \mu_{Y_1|X=x}(y) - \mu_{Y_0|X=x}(y).$$

The estimates in (4.2) can be plugged in to obtain the estimate $\hat{U}_{\text{witness},y} = \hat{\mu}_{Y_1|X=x}(y) - \hat{\mu}_{Y_0|X=x}(y)$. Since convergence in the RKHS norm implies pointwise convergence (Berlinet and Thomas-Agnan, 2004, p.10, Corollary 1), Theorem 4.3.2 implies the consistency of $\hat{U}_{\text{witness},y}$ with respect to the corresponding PEHDE. Clearly, if X is more than 1-dimensional, heat maps as in Figure 4.1(c) cannot be plotted; however, fixing a particular $x \in \mathcal{X}$, $\hat{\mu}_{Y_1|X=x} - \hat{\mu}_{Y_0|X=x}$ can be plotted against y, since $Y \subseteq \mathbb{R}$. Such plots will be informative of where the density of $P_{Y_1|X=x}$ is greater than that of $P_{Y_0|X=x}$ and vice versa.

4.4.2 CoDiTE via U-statistic Regression

Next, we consider CoDiTE on specific distributional quantities, such as the mean, variance or skewness, or some function thereof. For example, in addition to the CATE, Briseño Sanchez et al. (2020, Eqn. (2)) were interested in the treatment effect on the standard deviation $U_D(x) = \text{std}(Y_1|X = x) - \text{std}(Y_0|X = x)$. Our motivating example in Figure 4.1 could inspire a "standardised" version of the CATE[1]:

$$U_D(x) = \frac{\mathbb{E}[Y_1|X = x] - \mathbb{E}[Y_0|X = x]}{\sqrt{\text{Var}(Y_1|X = x) + \text{Var}(Y_0|X = x)}}. \tag{4.3}$$

Many of these quantities can be represented as the expectation of a U-kernel, i.e. $\mathbb{E}[h(Y_1, ..., Y_r)]$ (c.f. Section 4.1.1). For example, $h(y) = y$ gives the mean, $h(y_1, y_2) = \frac{1}{2}(y_1 - y_2)^2$ gives the variance and $h(y_1, y_2) = |y_1 - y_2|$ gives Gini's mean difference. We consider their conditional counterparts, i.e.

$$\theta(P_{Y_0|X}) = \mathbb{E}[h(Y_{01}, ..., Y_{0r})|X_1, ..., X_r],$$
$$\theta(P_{Y_1|X}) = \mathbb{E}[h(Y_{11}, ..., Y_{1r})|X_1, ..., X_r]$$

(c.f. Section 4.1.1). By Çınlar (2011, p.146, Theorem 1.17), there exist functions $F_0, F_1 : \mathcal{X}^r \to \mathbb{R}$ such that $F_0(X_1, ..., X_r) = \theta(P_{Y_0|X})$ and $F_1(X_1, ..., X_r) = \theta(P_{Y_1|X})$.

Estimation of F_0 and F_1 can be done via U-statistic regression, by generalising kernel ridge regression as follows. As in Section 4.3.1, let $k_0 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel on $\mathcal{X}$ with RKHS $\mathcal{H}_0$. Then if we define $k_0^r : \mathcal{X}^r \times \mathcal{X}^r \to \mathbb{R}$ as

$$k_0^r((x_1, ..., x_r), (x_1', ..., x_r')) = k_0(x_1, x_1')...k_0(x_r, x_r'),$$

[1] In practice, if the CoDiTE involves ratios of estimated quantities, we do not recommend plugging in the estimates directly into the ratio, since, if the denominator is small, then a small error in the estimation of the denominator will result in a large error in the overall CoDiTE estimation. Instead, we recommend that the practitioner estimate the numerator and the denominator separately and interpret the results directly from the raw estimates.

Method	**Setting SN**	
	Control	Treatment
GAMLSS	0.17 ± 0.031	0.767 ± 0.414
U-regression KRR	$\mathbf{0.13 \pm 0.059}$	$\mathbf{0.16 \pm 0.059}$
Method	**Setting LN**	
	Control	Treatment
GAMLSS	3.3 ± 0.55	15.44 ± 8.128
U-regression KRR	$\mathbf{1.1 \pm 0.31}$	$\mathbf{2.16 \pm 0.61}$
Method	**Setting HN**	
	Control	Treatment
GAMLSS	2.27 ± 0.44	10.91 ± 5.42
U-regression KRR	$\mathbf{0.7 \pm 0.25}$	$\mathbf{1.39 \pm 0.47}$

Table 4.1: Root mean square error in estimating the conditional standard deviation, with standard error from 100 simulations, for GAMLSS (implemented via the R package `gamlss` Rigby and Stasinopoulos (2005)) and our U-statistic regression via generalised kernel ridge regression (U-regression KRR; implemented via the Falkon library on Python Rudi et al. (2017); Meanti et al. (2020)). Lower is better.

Berlinet and Thomas-Agnan (2004, p.31, Theorem 13) tells us that k_0^r is a reproducing kernel on $\mathcal{X}^r$ with RKHS $\mathcal{H}_0^r = \mathcal{H}_0 \otimes ... \otimes \mathcal{H}_0$, the r-times tensor product of $\mathcal{H}_0$, whose elements are functions $\mathcal{X}^r \to \mathbb{R}$. We estimate F_0 in $\mathcal{H}_0^r$. Given any $F \in \mathcal{H}_0^r$, the natural least-squares risk is

$$\mathcal{E}(F) = \mathbb{E}[(F(X_1, ..., X_r) - h(Y_{01}, ..., Y_{0r}))^2].$$

Recalling the control sample $\{(x_i^0, y_i^0)\}_{i=1}^{n_0}$, we solve the following regularised least-squares problem:

$$\hat{F}_0 = \operatorname*{arg\,min}_{F \in \mathcal{H}_0^r} \left\{ \hat{\mathcal{E}}(F) + \lambda_{n_0}^0 \|F\|_{\mathcal{H}_0^r}^2 \right\} \tag{4.4}$$

where the empirical least-squares risk $\hat{\mathcal{E}}$ is defined as

$$\hat{\mathcal{E}}(F) = \frac{1}{\binom{n_0}{r}} \sum \left(F(x_{i_1}^0, ..., x_{i_r}^0) - h(y_{i_1}^0, ..., y_{i_r}^0) \right)^2,$$

with the summation over the $\binom{n_0}{r}$ combinations of r distinct elements $\{i_1, ..., i_r\}$ from $\{1, ..., n_0\}$. Note that $\hat{\mathcal{E}}(F)$ is itself a U-statistic for the estimation of $\mathcal{E}(F)$. The following is a representer theorem for the problem in (4.4).

Theorem 4.4.1. *The solution $\hat{F}_0$ to the problem in (4.4) is*

$$\hat{F}_0(x_1, ..., x_r) = \sum_{i_1, ..., i_r}^{n_0} k_0(x_{i_1}^0, x_1) ... k_0(x_{i_r}^0, x_r) c_{i_1, ..., i_r}^0$$

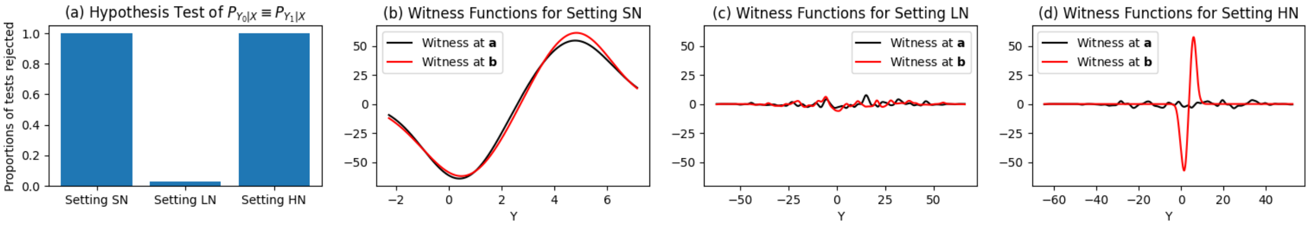


Figure 4.2: **Hypothesis testing and witness functions on the IHDP dataset.** (a) Hypothesis test is conducted on 100 simulations for each setting, with the bar chart showing proportion of tests rejected for each setting. In setting "LN", where the variance overwhelms the CATE, the test does not reject the hypothesis $P_{Y_0|X} \equiv P_{Y_1|X}$, whereas in the other two settings, the hypothesis is rejected. (b) At both $X = \mathbf{a}$ and $X = \mathbf{b}$, the density of the control group is larger than that of the treatment group around $Y = 0$, and the reverse is true around $Y = 4$, showing the marked effect of the treatment. (c) At both $X = \mathbf{a}$ and $X = \mathbf{b}$, the density of the control and treatment groups are roughly equal for all Y. (d) At $X = \mathbf{a}$, where the variance engulfs the CATE, the density of the control and treatment groups are roughly equal for all Y, whereas at $X = \mathbf{b}$, the witness function clearly shows where the density of one group dominates the other. The juxtaposition of witness functions at different points in the covariate space is an exploratory tool to compare the relative strength of the treatment effect.

where the coefficients $c^0_{i_1,\ldots,i_r} \in \mathbb{R}$ are the unique solution of the n^r linear equations,

$$\sum_{j_1,\ldots,j_r=1}^{n_0} \left(k_0\left(x^0_{i_1}, x^0_{j_1}\right) \ldots k_0\left(x^0_{i_r}, x^0_{j_r}\right) + \binom{n_0}{r} \lambda^0_{n_0} \delta_{i_1 j_1} \ldots \delta_{i_r j_r} \right) c^0_{j_1,\ldots,j_r}$$
$$= h\left(y^0_{i_1}, \ldots, y^0_{i_r}\right).$$

Note that if $r = 1$ and $h(y) = y$, we recover the usual kernel ridge regression. The following result shows that this estimation procedure is universally consistent.

Theorem 4.4.2. *Suppose k^r_0 is a bounded and universal kernel and that $\lambda^0_{n_0}$ decays at a slower rate than $\mathcal{O}(n_0^{-1/2})$. Then as $n_0 \to \infty$,*

$$\mathbb{E}\left[\left(\hat{F}_0\left(X_1, \ldots, X_r\right) - F_0\left(X_1, \ldots, X_r\right)\right)^2\right] \xrightarrow{p} 0.$$

A consistent estimate $\hat{F}_1$ of F_1 is obtained by exactly the same procedure, using the treatment sample $\{(x^1_i, y^1_i)\}_{i=1}^{n_1}$.

4.5 Experiments

4.5.1 Semi-synthetic IHDP Data

We demonstrate the use of our methods on the Infant Health and Development Program (IHDP) dataset (Hill, 2011, Section 4). The covariates are taken from a randomised control trial, from which a non-random portion is removed to imitate an observational study. The reason for its popularity in the CATE literature is that, for each datapoint, the outcome is simulated for both treatment and control, enabling cross-validation and evaluation, which is usually not possible in observational studies due to the missing counterfactuals. Existing works first define the noiseless response surfaces for the control and treatment groups, and generate realisations of the potential outcomes by applying Gaussian noise with constant variance across the whole dataset.

This last assumption of constant variance is somewhat unrealistic, but of little importance in evaluating CATE estimators. In our experiments, we modify the data generating process in three different ways, all of which have the same parallel linear mean response surfaces, with the CATE of 4 ("response surface A" in Hill (2011)). In setting "SN" ("small noise"), the standard deviation of the noise is constant at 1, so that the CATE of 4 translates to a meaningful treatment effect. In setting "LN" ("large noise"), the standard deviation of the noise is constant at 20, meaning that the mean difference in the response surfaces is negligible in comparison. In this case, our test does not reject the hypothesis that the two conditional distributions are the same, and there is no case for further investigation (see middle bar in Figure 4.2(a)). In setting "HN" ("heterogeneous noise"), the standard deviation is heterogeneous across the dataset, so that the standard deviation is 1 for some data points while others have standard deviation of 20.

The data consists of 25 covariates: birth weight, head circumference, weeks born preterm, birth order, first born, neonatal health index, sex, twin status, whether or not the mother smoked during pregnancy, whether or not the mother drank alcohol during pregnancy, whether or not the mother took drugs during pregnancy, the mother's age, marital status, education attainment, whether or not the mother worked during pregnancy, whether she received prenatal care, and 7 dummy variables for the 8 sites in which the family resided at the start of the intervention.

These covariates are originally taken from a randomised experiment, and included information about the ethnicity of the mothers. Hill (2011) removed all children with nonwhite mothers from the treatment group, which is clearly a non-random (biased) portion of the data, thereby imitating an observational study. This leaves 608 children in the control group and 139 in the treatment group. The overlap condition is now only satisfied for the treatment group.

In creating the parallel linear response surfaces, which are used in all three of the settings "SN", "LN" and "HN", we let $\mathbb{E}[Y_0|X] = \beta X$ and $\mathbb{E}[Y_1|X] = \beta X + 4$, where the 25-dimensional coefficient vector β is generated in the same way as in Alaa and Schaar (2018): for the 6 continuous variables (birth weight, head circumference, weeks born preterm, birth order, neonatal health index, mother's age), the corresponding coefficients is sampled from $\{0, 0.1, 0.2, 0.3, 0.4\}$ with probabilities $\{0.5, 0.125, 0.125, 0.125, 0.125\}$ respectively, whereas for the other 19 binary variables, the coefficients are sampled from $\{0, 0.1, 0.2, 0.3, 0.4\}$ with probabilities $\{0.6, 0.1, 0.1, 0.1, 0.1\}$ respectively.

Finally, we generate realisations of the potential outcomes by adding noise to the mean response surfaces. We let $Y_0 = \beta X + \epsilon(X)$ and $Y_1 = \beta X + 4 + \epsilon(X)$, where $\epsilon(X) = \epsilon_{\text{SN}}$ in setting "SN", $\epsilon(X) = \epsilon_{\text{LN}}$ in setting "LN" and $\epsilon(X) = X_6\epsilon_{\text{SN}} + (1 - X_6)\epsilon_{\text{LN}}$ in setting "HN", with $\epsilon_{\text{SN}} \sim \mathcal{N}(0, 1^2)$ and $\epsilon_{\text{LN}} \sim \mathcal{N}(0, 20^2)$. The covariate X_6 corresponds to the sex of the child, and was chosen because there are roughly the same number of each sex in both the control and the treatment groups.

In setting "HN", let us consider points $\mathbf{a}, \mathbf{b} \in \mathcal{X}$ with $\text{sd}(Y|X = \mathbf{a}) = 20$ and $\text{sd}(Y|X = \mathbf{b}) = 1$. Then even though the CATE at $\mathbf{a}$ and $\mathbf{b}$ are equal at 4, we have $\text{std}(Y_1 - Y_0|X = \mathbf{a}) \gg \text{std}(Y_1 - Y_0|X = \mathbf{b})$, such that there is a pronounced treatment effect at $\mathbf{b}$, while the variance engulfs the treatment effect at $\mathbf{a}$. The comparative magnitudes of the witness functions conditioned on $\mathbf{a}$ and $\mathbf{b}$ confirm this heterogeneity (see Figure 4.2(d)). In Table 4.1, the quality of estimation of the standard deviation via our U-statistic regression is compared with GAMLSS (Stasinopoulos et al., 2017) estimation for each setting.

An immediate benefit is a better understanding of the treatment. Even a perfect CATE estimator cannot capture such heterogeneity in distributional treatment effect (variance, in this case). As argued in Section 4.1, any method that involves comparing mean values (of which CATE is one) should also take into account the variance for it to be meaningful. This will give a clearer picture of the subpopulations on which there is a marked treatment effect, and those on which it is weaker, than relying on the CATE alone. Such knowledge should in turn influence policy decisions, in terms of which subpopulations should be targeted. We note that recently Jesson et al. (2020) considered CATE uncertainty in IHDP in the context of a different task: making or deferring treatment recommendations while using Bayesian neural networks, focusing on cases where overlap fails or under covariate shift; however, distributional considerations can be important even when overlap is satisfied and no covariate shift takes place.

4.5.2 Real Outcomes: LaLonde Data

In this section, we apply the proposed methods to LaLonde's well-known National Supported Work (NSW) dataset (LaLonde, 1986; Dehejia and Wahba, 1999) which has been used widely to evaluate estimators of treatment effects. The outcome of interest Y is the real earnings in 1978, with treatment Z being the job training. We refer the interested readers to Dehejia and Wahba (1999, Sec. 2.1) for a detailed description of the dataset. As income distributions are known to be skewed to the right, it may be interesting to investigate not only the CATE, but the entire distributions.

The test rejects the hypothesis $P_{Y_0|X} \equiv P_{Y_1|X}$ with p-value of 0.013. As a demonstration of the kind of exploratory analysis that can be conducted using

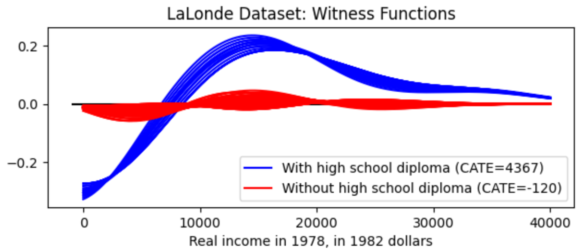


Figure 4.3: **Witness functions for Black, unmarried participant up to the age of 25, unemployed in both 1974 and 1975.** Each curve (witness function) corresponds to an individual in this subset.

the conditional witness functions, we focus our attention on a subset of the data on which the overlap condition is satisfied – Black, unmarried participants up to the age of 25, who were unemployed in both 1974 and 1975. Figure 4.3 shows the witness function for each individual in this subset, with the colour of the curve delineating whether the corresponding individual has a high school diploma.

We can see clearly that for those without a high school diploma, the treatment effect is not so pronounced, whereas there is a marked treatment effect for those with it. Negative values of the witness function for small income values mean that we are more likely to get small income values from the control group than the treatment group, whereas larger income values are more likely to come from the treatment group, as indicated by the positive values of the witness functions. In particular, the tail of the blue curves to the right implies a skewness of the density of the treated group relative to the control group, and the treatment group continues to have larger density than the control group for high income values (> 25000), albeit to a lesser extent. Such comparison of densities in different regions of $\mathcal{Y}$ is not possible with the CATE, which is a simple difference of the means between the control and treated groups.

Chapter 5

Vector-Valued Regression

In this Chapter, we treat the two main branches of learning theory analysis of kernel regression with vector-valued output space, namely, the integral operator technique and empirical processes.

5.1 Integral Operators

As in the previous chapters, let us take $(\Omega, \mathcal{F}, P)$ as the underlying probability space. Suppose $(\mathcal{X}, \mathfrak{X})$ is a separable measurable space, and that $\mathcal{Y}$ is a (potentially infinite-dimensional) separable Hilbert space with associated inner product and norm denoted by $\langle \cdot, \cdot \rangle_{\mathcal{Y}}$ and $\|\cdot\|_{\mathcal{Y}}$. Denote the Borel σ-algebra of $\mathcal{Y}$ as $\mathfrak{Y}$. Suppose $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ are random variables, with distributions $P_X(A) = P(X^{-1}(A))$ for $A \in \mathfrak{X}$ and $P_Y(B) = P(Y^{-1}(B))$ for $B \in \mathfrak{Y}$. Further, we denote by P_{XY} the joint distribution of X and Y. In order for regression of Y on X to be possible, the following assumption that Y has finite variance is a minimal requirement:

Assumption 5.1.1. We have $\mathbb{E}\left[\|Y\|_{\mathcal{Y}}^2\right] < \infty$.

Assumption 5.1.1 also implies that $\mathbb{E}[\|Y\|_{\mathcal{Y}}] < \infty$, which means that Y is Bochner-integrable (Dinculeanu, 2000, p.15, Definition 35). Hence, we can define its conditional expectation $\mathbb{E}[Y \mid X]$ as an X-measurable, Bochner-P_X-integrable random variable taking values in $\mathcal{Y}$, according to Dinculeanu (2000, p.45, Definition 38). In the rest of this chapter, we let $\mathbb{E}[Y \mid X]$ be any particular version thereof, and talk about *the* conditional expectation of Y given X. Since $\mathbb{E}[Y \mid X]$ is an X-measurable random variable, we can write

$$\mathbb{E}[Y \mid X] = f^*(X). \tag{5.1}$$

for some deterministic measurable function $f^* : \mathcal{X} \to \mathcal{Y}$. It is this function f^* that we aim to estimate via regression.

Denote by $L^2(\mathcal{X}, P_X; \mathcal{Y})$ the Bochner space with output in $\mathcal{Y}$, i.e. the Hilbert space of (equivalence classes of) measurable functions $f : \mathcal{X} \to \mathcal{Y}$ such that

$\|f(\cdot)\|_{\mathcal{Y}}^2$ is P_X-integrable, with inner product $\langle f_1, f_2\rangle_2 = \mathbb{E}[\langle f_1(X), f_2(X)\rangle_{\mathcal{Y}}]$. Denote its corresponding norm by $\|\cdot\|_2$. Then by Jensen's inequality and Assumption 5.1.1, we have $f^* \in L^2(\mathcal{X}, P_X; \mathcal{Y})$:

$$\mathbb{E}\left[\|f^*(X)\|_{\mathcal{Y}}^2\right] = \mathbb{E}\left[\|\mathbb{E}\left[Y \mid X\right]\|_{\mathcal{Y}}^2\right] \leq \mathbb{E}\left[\mathbb{E}\left[\|Y\|_{\mathcal{Y}}^2 \mid X\right]\right] = \mathbb{E}\left[\|Y\|_{\mathcal{Y}}^2\right] < \infty.$$

5.1.1 Vector-Valued Reproducing Kernel Hilbert Spaces

In this report, regression for $f^* \in L^2(\mathcal{X}, P_X; \mathcal{Y})$ will be carried out in a *fixed* vector-valued reproducing kernel Hilbert space, the well-known theory of which we briefly review here.

Suppose that $\mathcal{H}$ is a Hilbert space of functions $\mathcal{X} \to \mathcal{Y}$, with inner product and norm denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}}$ respectively. For any $n \in \mathbb{N}$, we denote by $\mathcal{X}^n$ and $\mathcal{Y}^n$ the n-fold direct sums of $\mathcal{X}$ and $\mathcal{Y}$ respectively; in particular, $\mathcal{Y}^n$ is a Hilbert space with inner product $\langle (y_1, ..., y_n)^T, (y_1', ..., y_n')^T\rangle_{\mathcal{Y}^n} = \sum_{i=1}^n \langle y_i, y_i'\rangle_{\mathcal{Y}}$. For any $\mathbf{x} = (x_1, ..., x_n)^T \in \mathcal{X}^n$, we define the *evaluation operator* (or *sampling operator*) by

$$\begin{aligned} S_{\mathbf{x}} :& \mathcal{H} \to \mathcal{Y}^n \\ & f \mapsto \frac{1}{n}\left(f\left(x_1\right), ..., f\left(x_n\right)\right)^T. \end{aligned}$$

Then $\mathcal{H}$ is a *vector-valued reproducing kernel Hilbert space* (vvRKHS) if the evaluation map $S_x : \mathcal{H} \to \mathcal{Y}$ is continuous for all $x \in \mathcal{X}$ (Carmeli et al., 2006, Definition 2.1). This immediately implies that $S_{\mathbf{x}} : \mathcal{H} \to \mathcal{Y}^n$ is continuous for all $n \in \mathbb{N}$ and $\mathbf{x} \in \mathcal{X}^n$. We define the *operator-valued kernel* $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$, where $\mathcal{L}(\mathcal{Y})$ is the Banach space of continuous linear operators from $\mathcal{Y}$ to itself, by

$$K\left(x, x'\right)(y) = S_x S_{x'}^* y, \qquad \text{i.e.} \qquad K\left(\cdot, x'\right)(y) = S_{x'}^*(y).$$

Then we can easily deduce the *reproducing property*. For any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$\langle y, f(x)\rangle_{\mathcal{Y}} = \langle y, S_x(f)\rangle_{\mathcal{Y}} = \langle S_x^*(y), f\rangle_{\mathcal{H}} = \langle K(\cdot, x)(y), f\rangle_{\mathcal{H}}.$$

For arbitrary $n \in \mathbb{N}$ and $\mathbf{x} = (x_1, ..., x_n)^T \in \mathcal{X}^n$, the adjoint of the sampling operator, $S_{\mathbf{x}}^* : \mathcal{Y}^n \to \mathcal{H}$, is given by

$$S_{\mathbf{x}}^* \mathbf{y} = \frac{1}{n}\sum_{i=1}^n K(x_i, \cdot) y_i, \qquad \text{for } \mathbf{y} = (y_1, ..., y_n)^T, y_i \in \mathcal{Y},$$

since, by the reproducing property, for any $f \in \mathcal{H}$ and $\mathbf{y} \in \mathcal{Y}^n$,

$$\begin{aligned} \langle S_{\mathbf{x}} f, \mathbf{y}\rangle_{\mathcal{Y}^n} &= \frac{1}{n}\sum_{i=1}^n \langle f\left(x_i\right), y_i\rangle_{\mathcal{Y}} \\ &= \frac{1}{n}\sum_{i=1}^n \langle f, K\left(x_i, \cdot\right) y_i\rangle_{\mathcal{H}} \end{aligned}$$

$$= \left\langle f, \frac{1}{n} \sum_{i=1}^{n} K\left(x_i, \cdot\right) y_i \right\rangle_{\mathcal{H}}.$$

Assumption 5.1.2. We henceforth assume that $\mathcal{H}$ is separable, and that the kernel K is bounded:

$$\sup_{x \in \mathcal{X}} \|K(x,x)\|_{\text{op}} = \sup_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}, \|y\|_{\mathcal{Y}} \leq 1} \|K(x,x)(y)\|_{\mathcal{Y}} < B, \qquad \text{for some } B > 0.$$

For a fixed $f \in \mathcal{H}$, Assumption 5.1.2 allows us to bound $\|f(\cdot)\|_{\mathcal{Y}}$ uniformly over $\mathcal{X}$, and hence the operator norm of $S_{\mathbf{x}}$ uniformly over $\mathcal{X}^n$.

Lemma 5.1.3. *Suppose Assumption 5.1.2 holds. Then*

(i) For all $f \in \mathcal{H}$,

$$\sup_{x \in \mathcal{X}} \|f(x)\|_{\mathcal{Y}} \leq \sqrt{B} \|f\|_{\mathcal{H}}.$$

(ii) For all $n \in \mathbb{N}$,

$$\sup_{\mathbf{x} \in \mathcal{X}^n} \|S_{\mathbf{x}}\|_{\text{op}}^2 \leq \frac{B}{n}.$$

Lemma 5.1.3(i) immediately implies that $\mathcal{H} \subseteq L^2(\mathcal{X}, P_X; \mathcal{Y})$, since, for any $f \in \mathcal{H}$, $\mathbb{E}\left[\|f(X)\|_{\mathcal{Y}}^2\right] \leq B \|f\|_{\mathcal{H}}^2 < \infty$, and the inclusion $\iota : \mathcal{H} \to L^2(\mathcal{X}, P_X; \mathcal{Y})$ is a bounded linear operator with $\|\iota\|_{\text{op}} \leq \sqrt{B}$:

$$\|\iota(f)\|_2 = \sqrt{\mathbb{E}\left[\|f(X)\|_{\mathcal{Y}}^2\right]} \leq \sqrt{B} \|f\|_{\mathcal{H}}, \qquad \text{for all } f \in \mathcal{H}.$$

Denote the adjoint of the inclusion by $\iota^* : L^2(\mathcal{X}, P_X; \mathcal{Y}) \to \mathcal{H}$. Then $\iota^* \circ \iota : \mathcal{H} \to \mathcal{H}$ and $\iota \circ \iota^* : L^2(\mathcal{X}, P_X; \mathcal{Y}) \to L^2(\mathcal{X}, P_X; \mathcal{Y})$ are self-adjoint operators.

Let $\{(X_i, Y_i)\}_{i=1}^n$ be i.i.d. copies of (X, Y), and denote by $\mathbf{X}$ and $\mathbf{Y}$ the random vectors $(X_1, ..., X_n)^T \in \mathcal{X}^n$ and $(Y_1, ..., Y_n)^T \in \mathcal{Y}^n$. Then the operators $S_{\mathbf{X}} : \mathcal{H} \to \mathcal{Y}^n$ and $S_{\mathbf{X}}^* : \mathcal{Y}^n \to \mathcal{H}$, given by $S_{\mathbf{X}}(f) = \frac{1}{n}(f(X_1), ..., f(X_n))^T$ and $S_{\mathbf{X}}^*((y_1, ..., y_n)^T) = \frac{1}{n}\sum_{i=1}^n K(X_i, \cdot) y_i$ respectively, are random.

Lemma 5.1.4. *We state and prove some results about the inclusion operator and its adjoint.*

(i) An explicit integral expression for $\iota^* : L^2(\mathcal{X}, P_X; \mathcal{Y}) \to \mathcal{H}$ can be given as

$$\iota^*(f)(\cdot) = \mathbb{E}\left[K(\cdot, X) f(X)\right] \qquad \text{for } f \in L^2(\mathcal{X}, P_X; \mathcal{Y}).$$

(ii) For any $f \in L^2(\mathcal{X}, P_X; \mathcal{Y})$ and any $n \in \mathbb{N}$,

$$\iota^*(f) = \mathbb{E}\left[S_{\mathbf{X}}^*\left((f(X_1), ..., f(X_n))^T\right)\right].$$

(iii) For any $f \in \mathcal{H}$ and any $n \in \mathbb{N}$,

$$\iota^* \circ \iota\left(f\right) = \mathbb{E}\left[nS_{\mathbf{X}}^* \circ S_{\mathbf{X}}\left(f\right)\right].$$

Although the inclusion operator ι is a compact (in fact, Hilbert-Schmidt) operator if $\mathcal{Y}$ is $\mathbb{R}$ (Steinwart and Christmann, 2008, p. 127, Theorem 4.27), this is not true in the general case we consider in this thesis. Indeed, consider the following counterexample, in which $K(x, x') = k(x, x')\text{Id}$, where $k(\cdot, \cdot)$ is a bounded scalar kernel with $k(x_0, x_0) = 1$ for some $x_0 \in \mathcal{X}$ and $\text{Id} : \mathcal{Y} \to \mathcal{Y}$ is the identity operator. Let $\{y_i\}_{i=1}^{\infty}$ be a (countable, by separability assumption) orthonormal basis of $\mathcal{Y}$. Then $\{K(x_0, \cdot)y_i\}_{i=1}^{\infty}$ form a bounded sequence in $\mathcal{H}$, since, by the reproducing property,

$$\|K(x_0, \cdot)y_i\|_{\mathcal{H}}^2 = \langle y_i, K(x_0, x_0)y_i\rangle_{\mathcal{Y}} = \|y_i\|_{\mathcal{Y}}^2 = 1.$$

However, the sequence $\{\iota(K(x_0, \cdot)y_i)\}_{i=1}^{\infty}$ in $L^2(\mathcal{X}, P_X; \mathcal{Y})$ cannot have a convergent subsequence, since, for any $i \neq j$,

$$\begin{aligned}\left\|\iota\left(K\left(x_0, \cdot\right)y_i\right) - \iota\left(K\left(x_0, \cdot\right)y_j\right)\right\|_2^2 &= \mathbb{E}\left[\|k(x_0, X)y_i - k(x_0, X)y_j\|_{\mathcal{Y}}^2\right] \\ &= 2\mathbb{E}\left[k(x_0, X)^2\right] \\ &> 0.\end{aligned}$$

Hence ι is not a compact operator[1].

The self-adjoint operator $\iota \circ \iota^*$ is also not compact. Indeed, let $\{y_i\}_{i=1}^{\infty}$ be an orthonormal basis of $\mathcal{Y}$ again, and consider the sequence of functions $f_i \in L^2(\mathcal{X}, P_X; \mathcal{Y})$ given by $f_i(x) = y_i$ for all $x \in \mathcal{X}$. Also, consider again the kernel $K(x, x') = k(x, x')\text{Id}$, where $k(\cdot, \cdot)$ is a scalar kernel and $\text{Id} : \mathcal{Y} \to \mathcal{Y}$ is the identity operator. Then $\|f_i\|_2^2 = \mathbb{E}\left[\|f_i(X)\|_{\mathcal{Y}}^2\right] = \|y_i\|_{\mathcal{Y}}^2 = 1$, so the sequence is bounded, but for any $i \neq j$,

$$\begin{aligned}&\|\iota \circ \iota^*(f_i) - \iota \circ \iota^*(f_j)\|_2^2 \\ &= \mathbb{E}_{X_1}\left[\|\mathbb{E}_{X_2}\left[K(X_1, X_2)f_i(X_2)\right] - \mathbb{E}_{X_2}\left[K(X_1, X_2)f_j(X_2)\right]\|_{\mathcal{Y}}^2\right] \\ &= \mathbb{E}_{X_1}\left[\|\mathbb{E}_{X_2}\left[k(X_1, X_2)\right]y_i - \mathbb{E}_{X_2}\left[k(X_1, X_2)\right]y_j\|_{\mathcal{Y}}^2\right] \\ &= \mathbb{E}_{X_1}\left[\mathbb{E}_{X_2}\left[k(X_1, X_2)\right]^2\|y_i - y_j\|_{\mathcal{Y}}^2\right] \\ &= 2\mathbb{E}_{X_1}\left[\mathbb{E}_{X_2}\left[k(X_1, X_2)\right]^2\right] \\ &> 0,\end{aligned}$$

using the expression for ι^* given in Lemma 5.1.4(i). So the sequence $\{\iota \circ \iota^*(f_i)\}_{i=1}^{\infty}$ in $L^2(\mathcal{X}, P_X; \mathcal{Y})$ cannot have a convergent subsequence, which in turn implies that $\iota \circ \iota^*$ is not compact.

[1] See Bollobás (1999, p.186) for the definition and equivalent formulations of compact operators. This counterexample does not contradict Carmeli et al. (2006, Proposition 4.8), which says that ι is compact if $K(x, x) : \mathcal{Y} \to \mathcal{Y}$ is compact for all $x \in \mathcal{X}$ and $\mathbb{E}\left[\|K(X, X)\|_{\text{op}}\right] < \infty$, since $K(x, x) = k(x, x)\text{Id}$ is clearly not a compact operator.

5.1.2 Regularised Least-Squares Regression

As above, take i.i.d. copies $\{(X_i, Y_i)\}_{i=1}^n$ of (X, Y). We define the unregularised population, regularised population, unregularised empirical and regularised empirical risk functions with respect to the squared-loss as follows:

$$\begin{aligned} R(f) &= \mathbb{E}\left[\|f(X) - Y\|_{\mathcal{Y}}^2\right]; \\ R_\lambda(f) &= \mathbb{E}\left[\|f(X) - Y\|_{\mathcal{Y}}^2\right] + \lambda \|f\|_{\mathcal{H}}^2; \\ R_n(f) &= \frac{1}{n}\sum_{i=1}^n \|f(X_i) - Y_i\|_{\mathcal{Y}}^2; \text{ and} \\ R_{n,\lambda}(f) &= \frac{1}{n}\sum_{i=1}^n \|f(X_i) - Y_i\|_{\mathcal{Y}}^2 + \lambda \|f\|_{\mathcal{H}}^2, \end{aligned} \tag{5.2}$$

where $\lambda > 0$ is a regularisation parameter. Here, R and R_n is defined for any $f \in L^2(\mathcal{X}, P_X; \mathcal{Y})$, but R_λ and $R_{n,\lambda}$ are only defined for $f \in \mathcal{H}$. Also, the population risks R and R_λ are deterministic functions of F, whereas the empirical risks R_n and $R_{n,\lambda}$ are random, varying with the random sample $\{(X_i, Y_i)\}_{i=1}^n$.

The following decomposition of the population risk is well-known; see, for example, Cucker and Smale (2002, Propoisition 1).

Lemma 5.1.5. *We have the following decomposition of the risk R:*

$$R(f) = \mathbb{E}\left[\|f(X) - f^*(X)\|_{\mathcal{Y}}^2\right] + R(f^*).$$

It is immediate that f^* is the minimiser of R in $L^2(\mathcal{X}, P_X; \mathcal{Y})$.

Lemma 5.1.6. *We formulate the minimisers in $\mathcal{H}$ of the regularised risks R_λ and $R_{n,\lambda}$ in terms of the inclusion and evaluation operators. Similar results can be found in many places in the literature, for example Micchelli and Pontil (2005, Section 4) or Engl et al. (1996, p.117, Theorem 5.1).*

(i) The minimiser f_λ of the risk R_λ in $\mathcal{H}$ is unique and is given by

$$f_\lambda := \operatorname*{arg\,min}_{f \in \mathcal{H}} R_\lambda(f) = \left(\iota^* \circ \iota + \lambda \mathrm{Id}_{\mathcal{H}}\right)^{-1} \iota^* f^* = \iota^* \left(\iota \circ \iota^* + \lambda \mathrm{Id}_2\right)^{-1} f^*,$$

where $\mathrm{Id}_{\mathcal{H}} : \mathcal{H} \to \mathcal{H}$ and $\mathrm{Id}_2 : L^2(\mathcal{X}, P_X; \mathcal{Y}) \to L^2(\mathcal{X}, P_X; \mathcal{Y})$ are the identity operators.

(ii) The minimiser $\hat{f}_{n,\lambda}$ of the risk $R_{n,\lambda}$ in $\mathcal{H}$ is unique and is given by

$$\begin{aligned} \hat{f}_{n,\lambda} &:= \operatorname*{arg\,min}_{f \in \mathcal{H}} R_{n,\lambda}(f) \\ &= \left(n S_{\mathbf{X}}^* \circ S_{\mathbf{X}} + \lambda \mathrm{Id}_{\mathcal{H}}\right)^{-1} S_{\mathbf{X}}^* \mathbf{Y} \\ &= S_{\mathbf{X}}^* \left(n S_{\mathbf{X}} \circ S_{\mathbf{X}}^* + \lambda \mathrm{Id}_{\mathcal{Y}^n}\right)^{-1} \mathbf{Y}, \end{aligned}$$

where $\mathrm{Id}_{\mathcal{Y}^n} : \mathcal{Y}^n \to \mathcal{Y}^n$ is the identity operator.

5.1.3 Universal Consistency

Our goal in this subsection is to investigate the convergence to 0 in probability of

$$R\left(\hat{f}_{n,\lambda}\right) - R\left(f^*\right) = \mathbb{E}\left[\left\|\hat{f}_{n,\lambda}(X) - f^*(X)\right\|_{\mathcal{Y}}^2\right] = \left\|\iota\hat{f}_{n,\lambda} - f^*\right\|_2^2,$$

where the equality comes from Lemma 5.1.5. We first consider the case where the measure is fixed, i.e. the distributions P_{XY}, P_X and P_Y, the regression function f^*, the function space $L^2(\mathcal{X}, P_X; \mathcal{Y})$ as well as the operator ι, are fixed. In Section 5.1.4, we will consider a uniform rate of convergence over a class of distributions.

We split the above using the triangle inequality into estimation and approximation errors:

$$\left\|\iota\hat{f}_{n,\lambda} - f^*\right\|_2 \leq \left\|\iota\hat{f}_{n,\lambda} - \iota f_\lambda\right\|_2 + \left\|\iota f_\lambda - f^*\right\|_2.$$

Proposition 5.1.7 shows, under the assumption that $\mathcal{H}$ is dense in $L^2(\mathcal{X}, P_X; \mathcal{Y})$, the convergence of the second term to 0 as $\lambda \to 0$, and Proposition 5.1.8 shows the convergence of the first term in probability to 0 as $n \to \infty$ and $\lambda \to 0$. Theorem 5.1.10 then brings them together to show the consistency of $\hat{f}_{n,\lambda}$.

Proposition 5.1.7 (Approximation Error). *If $\iota\mathcal{H}$ is dense in $L^2(\mathcal{X}, P_X; \mathcal{Y})$, then $\|f^* - \iota f_\lambda\|_2^2 \to 0$ as $\lambda \to 0$.*

Proposition 5.1.8 (Estimation Error). *Take any $\delta > 0$. Then*

$$P\left(\left\|\hat{f}_{n,\lambda} - f_\lambda\right\|_{\mathcal{H}}^2 \geq \frac{B\mathbb{E}\left[\|Y\|_{\mathcal{Y}}^2\right]}{n\lambda^2\delta}\right) \leq \delta.$$

In particular, if $\lambda = \lambda_n$ depends on n and converges to 0 at a slower rate than $\mathcal{O}(n^{-1/2})$, then

$$\left\|\hat{f}_{n,\lambda_n} - f_{\lambda_n}\right\|_{\mathcal{H}}^2 \xrightarrow{P} 0.$$

Remark 5.1.9. Under additional assumptions on the underlying distribution, we can obtain tighter bounds in Proposition 5.1.8, by using exponential probabilistic inequalities like Bernstein's inequality, instead of Chebyshev's inequality like we did above. This is indeed done, for example, in Smale and Zhou (2007, Theorem 1) for real output spaces and Singh et al. (2019, Theorem 2) for RKHS output spaces in the context of conditional mean embeddings, by assuming that Y is almost surely bounded, not just square integrable as we assumed in Assumption 5.1.1.

Theorem 5.1.10. *Suppose $\iota\mathcal{H}$ is dense in $L^2(\mathcal{X}, P_X; \mathcal{Y})$. Suppose that $\lambda = \lambda_n$ depends on the sample size n, and converges to 0 at a slower rate than $\mathcal{O}(n^{-1/2})$. Then we have*

$$R\left(\hat{f}_{n,\lambda_n}\right) - R\left(f^*\right) = \mathbb{E}\left[\left\|\hat{f}_{n,\lambda_n}(X) - f^*(X)\right\|_{\mathcal{Y}}^2\right] = \left\|\iota\hat{f}_{n,\lambda_n} - f^*\right\|_2^2 \xrightarrow{P} 0.$$

5.1.4 Uniform Rates in the Well-Specified Case

In our work above, possible bottlenecks are $\mathbb{E}[\|Y\|_{\mathcal{Y}}^2]$ in Proposition 5.1.8 being arbitrarily large, or f_ϵ in the proof of Proposition 5.1.7 having arbitrarily large norm in $\mathcal{H}$. In the next result, we consider a class of measures over which the rate of convergence is uniform. In particular, any measure in this class of measures is conditioned to have the conditional expectation f^* of Y given X in $\mathcal{H}$, i.e. there exists some $f_{\mathcal{H}}^* \in \mathcal{H}$ such that $\iota f_{\mathcal{H}}^* = f^*$. This is known as the *well-specified* case (Szabó et al., 2016, p.2).

Theorem 5.1.11. *For constants $M, C > 0$, define $\mathcal{P}(M,C)$ as the class of measures such that*

(i) $\mathbb{E}\left[\|Y\|_{\mathcal{Y}}^2\right] \leq M$, *and*

(ii) $f^* = \iota f_{\mathcal{H}}^*$ *for some* $f_{\mathcal{H}}^* \in \mathcal{H}$ *with* $\|f_{\mathcal{H}}^*\|_{\mathcal{H}}^2 \leq C$.

Let $\mathcal{H}$ be dense in $L^2(\mathcal{X}, P_X; \mathcal{Y})$ for all $P \in \mathcal{P}(M,C)$. Then

$$\sup_{P \in \mathcal{P}(M,C)} P\left(\left\|\iota \hat{f}_{n,\lambda} - f^*\right\|_2^2 \geq \frac{2B^2 M}{n\lambda^2\delta} + 2\lambda C\right) \leq \delta.$$

In particular, if $\lambda = \lambda_n$ depends on the sample size n and converges to 0 at the rate of $\mathcal{O}(n^{-1/4})$, then $R(\hat{f}_{n,\lambda_n}) - R(f^*) = \mathcal{O}_P(n^{-1/4})$ uniformly over the class $\mathcal{P}(M,C)$ of measures.

5.2 Empirical Process Theory for Vector-Valued Functions

Recall that $(\Omega, \mathcal{F}, \mathbb{P})$ is the underlying probability space, and $\mathcal{Y}$ is a separable Hilbert space over $\mathbb{R}$, with its inner product and norm denoted by $\langle\cdot,\cdot\rangle_{\mathcal{Y}}$ and $\|\cdot\|_{\mathcal{Y}}$ respectively. We denote by $\mathfrak{Y}$ the Borel σ-algebra of $\mathcal{Y}$, i.e. the σ-algebra generated by the open subsets of $\mathcal{Y}$. Let $(\mathcal{X}, \mathfrak{X})$ be a measurable set, and Q a probability measure on it.

Let $X : \Omega \to \mathcal{X}$ be a random variable, and let $X_1, X_2, ...$ be i.i.d. copies of X. Denote by P its distribution, i.e. for $A \in \mathfrak{X}$, $P(A) = \mathbb{P}(X^{-1}(A))$, and by P_n the empirical measure on $\mathcal{X}$ based on $X_1, ..., X_n$, i.e.

$$P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}, \qquad \text{where, for } A \in \mathfrak{X}, \delta_{X_i}(A) = \begin{cases} 0 & \text{if } X_i \notin A \\ 1 & \text{if } X_i \in A \end{cases}.$$

For a function $g \in L^1(\mathcal{X}, Q; \mathcal{Y})$, we adopt the notation $Qg = \int g dQ$. Hence,

$$Pg = \int g dP \qquad \text{and} \qquad P_n g = \frac{1}{n}\sum_{i=1}^n g(X_i).$$

Note that the integral Pg is a Bochner integral, and that we have $Pg, P_n g \in \mathcal{Y}$. Now, for fixed g, the law of large numbers in Hilbert (more generally, Banach) spaces (Mourier, 1953) tells us that $P_n g$ converges to Pg. One of the pillars of empirical process theory is to consider the convergence of $P_n g$ to Pg not for a fixed g, but uniformly over a class of functions. Let $\mathcal{G} \subset L^1(\mathcal{X}, P; \mathcal{Y})$. For a measure Q on $\mathcal{X}$, we denote $\|Q\|_{\mathcal{G}} := \sup_{g\in\mathcal{G}} \|Qg\|_{\mathcal{Y}}$.

Definition 5.2.1. We say that the class $\mathcal{G}$ is a *Glivenko Cantelli (GC)* class, or that it satisfies the *uniform law of large numbers* (with respect to the measure P) if $\|P_n - P\|_{\mathcal{G}} = \sup_{g\in\mathcal{G}} \|P_n g - Pg\|_{\mathcal{Y}} \xrightarrow{P} 0$.

Definition 5.2.1 could have been defined in terms of the weak convergence in Hilbert spaces, i.e. $y_n \to y_0$ if $\langle y, y_n\rangle_{\mathcal{Y}} \to \langle y, y_0\rangle_{\mathcal{Y}}$ for every $y \in \mathcal{Y}$. In this thesis, we only consider strong (norm) convergence. Next, we define the empirical process and the asymptotic equicontinuity.

Definition 5.2.2. We regard $\{\nu_n(g) = \sqrt{n}\,(P_n - P)\,g : g \in \mathcal{G}\}$ as a stochastic process with values in $\mathcal{Y}$ indexed by $\mathcal{G}$, and call it the *empirical process*.

We say that the empirical process $\{\nu_n(g) : g \in \mathcal{G}\}$ is *asymptotically equicontinuous* at $g_0 \in \mathcal{G}$ if, for every sequence $\{\hat{g}_n\} \subset \mathcal{G}$ with $\|\hat{g}_n - g_0\|_{2,P} \xrightarrow{P} 0$, we have $\|\nu_n(\hat{g}_n) - \nu_n(g_0)\|_{\mathcal{Y}} \xrightarrow{P} 0$.

In the next few subsections, we state and prove some basic empirical process-theoretic results, adapted to our setting of vector-valued functions.

5.2.1 Symmetrisation

Symmetrisation is an indispensable technique in empirical process theory. Let $X_1', ..., X_n'$ be another set of independent copies of X, independent of $X_1, ..., X_n$. Denote by P_n' the empirical measure on $X_1', ..., X_n'$, i.e. $P_n' = \frac{1}{n}\sum_{i=1}^n \delta_{X_i'}$.

Lemma 5.2.3. *We have*

$$\mathbb{E}\left[\|P_n - P\|_{\mathcal{G}}\right] \leq \mathbb{E}\left[\|P_n - P_n'\|_{\mathcal{G}}\right].$$

We let $\{\sigma_i\}_{i=1}^n$ be a *Rademacher sequence*, i.e. a sequence of independent random variables σ_i with

$$\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = \frac{1}{2}, \qquad \text{for all } i = 1, ..., n.$$

We define the symmetrised empirical measures $P_n^\sigma = \frac{1}{n}\sum_{i=1}^n \sigma_i \delta_{X_i}$ and $P_n'^\sigma = \frac{1}{n}\sum_{i=1}^n \sigma_i \delta_{X_i'}$, and denote

$$P_n^\sigma g = \frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i) \qquad \text{and} \qquad P_n'^\sigma g = \frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i').$$

Lemma 5.2.4 (Symmetrisation with means)**.** *We have*

$$\mathbb{E}\left[\|P_n - P\|_{\mathcal{G}}\right] \leq 2\mathbb{E}\left[\|P_n^\sigma\|_{\mathcal{G}}\right]$$

Lemma 5.2.5 (Symmetrisation with probabilities)**.** *Let $a > 0$. Suppose that for all $g \in \mathcal{G}$,*

$$\mathbb{P}\left(\|(P_n - P)\,g\|_{\mathcal{Y}} > \frac{a}{2}\right) \leq \frac{1}{2}.$$

Then

$$\mathbb{P}\left(\|P_n - P\|_{\mathcal{G}} > a\right) \leq 4\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{G}} > \frac{a}{4}\right).$$

A simple application of the above symmetrisation argument and Hoeffding's inequality in Hilbert spaces shows that finite function classes are Glivenko-Cantelli.

Lemma 5.2.6. *Let $\mathcal{G} = \{g_1, ..., g_N\} \in L^1(\mathcal{X}, P; \mathcal{Y})$ be a finite class of functions with cardinality $N > 1$. Then we have*

$$\|P_n - P\|_{\mathcal{G}} \to 0.$$

5.2.2 Uniform law of large numbers

We start with a definition.

Definition 5.2.7 (Adapted from van de Geer (2000, p.26, Definition 3.1))**.** The function $G : \mathcal{X} \to \mathbb{R}$ defined by $G(\cdot) = \sup_{g \in \mathcal{G}} \|g(\cdot)\|_{\mathcal{Y}}$ is called the *envelope* of $\mathcal{G}$.

The following is a uniform law of large numbers based on conditions on the entropy $H(\delta, \mathcal{G}, \|\cdot\|_{1,P_n})$ and the envelope G.

Theorem 5.2.8. *Suppose that*

$$G \in L^1(\mathcal{X}, P; \mathbb{R}) \qquad \text{and} \qquad \frac{1}{n} H(\delta, \mathcal{G}, \|\cdot\|_{1,P_n}) \xrightarrow{P} 0 \ \text{ for each } \delta > 0.$$

Then $\mathcal{G}$ is a Glivenko Cantelli class, i.e. $\|P_n - P\|_{\mathcal{G}} \xrightarrow{P} 0$.

5.2.3 Chaining and asymptotic equicontinuity with empirical entropy

In this subsection we show that, with additional conditions on the entropy of $\mathcal{G}$ (which we assume to be totally bounded with respect to the appropriate metric) and a technique called "chaining", we can derive explicit finite-sample bounds, and show the asymptotic continuity of the empirical process indexed by $\mathcal{G}$ (see Definition 5.2.2). As before, we work conditionally on the samples, and denote the σ-algebra generated by $X_1, ..., X_n$ as $\mathcal{F}_n$.

Suppose that $\mathcal{G}$ has an envelope $G \in L^2(\mathcal{X}, P; \mathbb{R})$ (see Definition 5.2.7). Then the quantity $R = \sup_{g \in \mathcal{G}} \|g\|_{2,P}$ is finite, since

$$R^2 = \sup_{g \in \mathcal{G}} \mathbb{E}\left[\|g(X)\|_{\mathcal{Y}}^2\right] \leq \mathbb{E}\left[\sup_{g \in \mathcal{G}} \|g(X)\|_{\mathcal{Y}}^2\right] = \mathbb{E}\left[G^2\right] < \infty.$$

Similarly, the quantity $R_n = \sup_{g \in \mathcal{G}} \|g\|_{2,P_n}$ is almost surely finite. We call R and R_n the *theoretical radius* and *empirical radius* of $\mathcal{G}$, respectively. Note that R_n is a random quantity, measurable with respect to $\mathcal{F}_n$.

Let us fix $S \in \mathbb{N}$. To ease the notation, for $s = 0, 1, ..., S$, write $N_s = N(2^{-s}R_n, \mathcal{G}, \|\cdot\|_{2,P_n})$ for the $2^{-s}R_n$-covering number of $\mathcal{G}$ with respect to the $\|\cdot\|_{2,P_n}$-metric, which we assume to be finite. Let $\{g_j^s\}_{j=1}^{N_s} \subset \mathcal{G}$ be a $2^{-s}R_n$-covering set of $\mathcal{G}$ with respect to the $\|\cdot\|_{2,P_n}$-metric. Note that $\{g^0\} = \{0\}$ is an R_n-covering set of $\mathcal{G}$, since, for any $g \in \mathcal{G}$, $\|g\|_{2,P_n} \leq R_n$. Similarly, write $H_s = \log N_s$ for each $s = 0, 1, ..., S$, for the corresponding entropy. Note that the quantities N_s and H_s, as well as the covering set $\{g_j^s\}_{j=1}^{N_s}$, are random quantities that are measurable with respect to $\mathcal{F}_n$.

Now fix $g \in \mathcal{G}$. Then define

$$\begin{aligned}
g^{S+1} &:= \operatorname*{arg\,min}_{\{g_j^{S+1}\}_{j=1}^{N_{S+1}}} \left\{\|g - g_j^{S+1}\|_{2,P_n}\right\} \\
g^{S} &:= \operatorname*{arg\,min}_{\{g_j^{S}\}_{j=1}^{N_{S}}} \left\{\|g^{S+1} - g_j^{S}\|_{2,P_n}\right\} \\
&\vdots \quad \vdots \\
g^{s} &:= \operatorname*{arg\,min}_{\{g_j^{s}\}_{j=1}^{N_{s}}} \left\{\|g^{s+1} - g_j^{s}\|_{2,P_n}\right\} \\
&\vdots \quad \vdots \\
g^0 &:= 0.
\end{aligned}$$

Proposition 5.2.9 (Chaining). *We fix $S \in \mathbb{N}$. Define*

$$J_n := \sum_{s=0}^{S} 2^{-s} R_n \sqrt{2H_{s+1}}.$$

(i) For all $t > 0$,

$$\mathbb{P}\left(\sup_{g \in \mathcal{G}} \left\|\sum_{s=0}^{S} P_n^\sigma \left(g^{s+1} - g^s\right)\right\|_{\mathcal{Y}} \geq \frac{\sqrt{2}J_n}{\sqrt{n}} + 6R_n\sqrt{\frac{1+t}{n}} \mid \mathcal{F}_n\right) \leq 2e^{-t}.$$

(ii) Suppose that $\varepsilon_1, ..., \varepsilon_n$ are i.i.d. Gaussian random variables in $\mathcal{Y}$ with mean 0 and covariance operator Q. Without loss of generality (by rescaling

if necessary), assume $\mathrm{Tr}Q = 1$. *For each* $g \in \mathcal{G}$, *we can consider the following inner product:*

$$\langle \varepsilon, g \rangle_{2,P_n} = \frac{1}{n} \sum_{i=1}^{n} \langle \varepsilon_i, g(X_i) \rangle_{\mathcal{Y}} .$$

Then for all $t > 0$,

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} \sum_{s=0}^{S} \left\langle \varepsilon, g^{s+1} - g^{s} \right\rangle_{2,P_n} \geq \frac{J_n}{\sqrt{n}} + 4R_n \sqrt{\frac{1+t}{n}} \mid \mathcal{F}_n \right) \leq e^{-t}.$$

Under additional conditions, we can use the previous lemma to show the asymptotic equicontinuity of the empirical process, $\{\sqrt{n}\,(P_n - P)\, g : g \in \mathcal{G}\}$. We continue to assume that the envelope $G = \sup_{g\in\mathcal{G}} \|g\|_{\mathcal{Y}}$ satisfies $G \in L^2(\mathcal{X}, P; \mathbb{R})$.

Theorem 5.2.10. *Suppose that* $\mathcal{G}$ *satisfies the "uniform entropy condition", i.e. there exists a decreasing function* $H : \mathbb{R} \to \mathbb{R}$ *satisfying*

$$\int_0^1 \sqrt{H(u)} du < \infty$$

such that, for all $u > 0$ *and any probability distribution* Q *with finite support,*

$$H(u \|G\|_{2,Q}, \mathcal{G}, \|\cdot\|_{2,Q}) \leq H(u).$$

Then the empirical process ν_n *is asymptotically equicontinuous.*

5.2.4 Peeling and Least-Squares Regression with Fixed Design and Gaussian Noise

Theorem 5.2.11. *Suppose that* $\varepsilon_1, ..., \varepsilon_n$ *are i.i.d. with Gaussian distribution with mean 0 and covariance operator* Q, *and that* $\mathrm{Tr}\, Q = 1$. *Further, suppose that*

$$J(\delta) := 4 \int_0^{\delta} \sqrt{2H(u, \mathcal{B}_{2,P_n}(\delta), \|\cdot\|_{2,P_n})} du < \infty,$$

for each $\delta > 0$ *and* $\frac{J(\delta)}{\delta^2}$ *is decreasing in* δ *where* $\mathcal{B}_{2,P_n}(\delta) := \{g \in \mathcal{G} : \|g\|_{2,P_n} \leq \delta\}$. *Then for all* $t \geq \frac{3}{8}$ *and all* δ_n *satisfying*

$$\sqrt{n}\delta_n^2 \geq 8 \left(J(\delta_n) + 4\delta_n \sqrt{1+t} + \delta_n \sqrt{\frac{8}{3} t} \right),$$

we have

$$\mathbb{P}\left(\|\hat{g}_n - g_0\|_{2,P_n} > \delta_n \right) \leq \left(1 + \frac{2}{e-1} \right) e^{-t}.$$

5.2.5 Empirical Risk Minimisation with Random Design

In this Section, we discuss the setting where we have an L-bounded, c-Lipschitz loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Suppose we have a given class $\mathcal{G}$ of functions $\mathcal{X} \to \mathcal{Y}$. Then given samples $(X_1, Y_1), ..., (X_n, Y_n)$, the empirical risk minimiser, which we assume exists, is given by

$$\hat{g}_n = \arg\min_{g \in \mathcal{G}} \hat{\mathcal{R}}_n(g), \qquad \hat{\mathcal{R}}_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(Y_i, g(X_i)).$$

We are interested in the convergence of $\hat{g}_n$ to the population risk minimiser,

$$g^* = \arg\min_{g \in \mathcal{G}} \mathcal{R}(g), \qquad \mathcal{R}(g) = \mathbb{E}[\mathcal{L}(Y, g(X))],$$

in terms of the population risk $\mathcal{R}$. First, see that

$$\begin{aligned} \mathcal{R}(\hat{g}_n) - \mathcal{R}(g^*) &= \mathcal{R}(\hat{g}_n) - \hat{\mathcal{R}}(\hat{g}_n) + \hat{\mathcal{R}}(\hat{g}_n) - \hat{\mathcal{R}}(g^*) + \hat{\mathcal{R}}(g^*) - \mathcal{R}(g^*) \\ &\leq \sup_{g \in \mathcal{G}} \left| \mathcal{R}(g) - \hat{\mathcal{R}}(g) \right| + \hat{\mathcal{R}}(g^*) - \mathcal{R}(g^*), \end{aligned}$$

where, going from the first line to the second, the first two terms on the right-hand side were bounded by the supremum over the whole function class $\mathcal{G}$ (since, although $\hat{g}_n$ varies as the samples and the size n of the dataset vary, it always lives in $\mathcal{G}$), the middle two terms were bounded above by 0 since the empirical risk minimiser $\hat{g}_n$ minimises $\hat{\mathcal{R}}$, and the last two terms remain unchanged.

Theorem 5.2.12. *Suppose the following uniform entropy condition is satisfied: there exists some function $H : \mathbb{R} \to \mathbb{R}$ satisfying*

$$\mathcal{J}(1) := 4 \int_0^1 \sqrt{2H(u)} du < \infty,$$

such that, for all $u > 0$ and any probability distribution Q with finite support,

$$H(uL, \mathcal{L} \circ \mathcal{G}, \|\cdot\|_{2,Q}) \leq H(u).$$

Then

$$\mathbb{P}\left(\sup_{g \in \mathcal{G}} \left| \mathcal{R}(g) - \hat{\mathcal{R}}(g) \right| > \frac{4\sqrt{2} L \mathcal{J}(1)}{\sqrt{n}} + 24L\sqrt{\frac{1+t}{n}} + \frac{L}{\sqrt{n}} \right) \leq 2e^{-t}.$$

5.3 Interlude: Vector-Valued Differential Calculus and Metric Spaces

Recall that $\mathcal{Y}$ is a Hilbert space. Suppose that U is an open subset of $\mathbb{R}^d$, and denote the Euclidean norm in $\mathbb{R}^d$ by $\|\cdot\|$. We say that $f_1, f_2 : U \to \mathcal{Y}$ are *tangent* at a point $a \in U$ (Cartan, 1967, p.28) if the quantity

$$m(r) = \sup_{\|x-a\| \leq r} \|f_1(x) - f_2(x)\|_{\mathcal{Y}},$$

which is defined for $r > 0$ small enough (since U is open), satisfies the condition

$$\lim_{r \to 0} \frac{m(r)}{r} = 0, \qquad \text{which we also write as} \qquad m(r) = o(r).$$

We say that the map $g : U \to \mathcal{Y}$ is *differentiable* at $a \in U$ if g is continuous at a and there exists a linear map $g'(a) : \mathbb{R}^d \to \mathcal{Y}$ such that the maps $x \mapsto g(x) - g(a)$ and $x \mapsto g'(a)(x - a)$ are tangent at a (Cartan, 1967, p.29). This condition is also written as

$$\|g(x) - g(a) - g'(a)(x-a)\|_{\mathcal{Y}} = o(\|x - a\|).$$

This immediately implies that $g'(a)$ is continuous, so $g'(a)$ belongs to $\mathcal{L}(\mathbb{R}^d, \mathcal{Y})$, the space of continuous linear operators from $\mathbb{R}^d$ into $\mathcal{Y}$. We call $g'(a) \in \mathcal{L}(\mathbb{R}^d, \mathcal{Y})$ the *derivative* of g at a. We say that g is differentiable on U if g is differentiable at every point in U, and the map $g' : U \to \mathcal{L}(\mathbb{R}^d, \mathcal{Y})$ is called the *derivative map* of g. We say that g is *continuously differentiable*, or *of class* C^1, if g is differentiable at every point of U and the map $g' : U \to \mathcal{L}(\mathbb{R}^d, \mathcal{Y})$ is continuous (Cartan, 1967, p.30).

Let $g : U \to \mathcal{Y}$ be a continuous map. For each $a = (a_1, ..., a_d) \in U$ and each $l = 1, ..., d$, consider the inclusion $\lambda_l : \mathbb{R} \to \mathbb{R}^d$ defined by

$$\lambda_l(x_l) = (a_1, ..., a_{l-1}, x_l, a_{l+1}, ..., a_d).$$

The composition $g \circ \lambda_l$ is defined on an open subset $\lambda_l^{-1}(\mathcal{X}) \subset \mathbb{R}$, which contains a_l. If g is differentiable at a, then for each $l = 1, ..., d$, the map $g \circ \lambda_l$ is differentiable at a_l (Cartan, 1967, p.38, Proposition 2.6.1). The derivative of $g \circ \lambda_l$ at a is called the *partial derivative* of g, denoted by $\partial_l g(a)$, and lives in $\mathcal{L}(\mathbb{R}, \mathcal{Y})$. But $\mathcal{L}(\mathbb{R}, \mathcal{Y})$ is isometrically isomorphic to $\mathcal{Y}$ (Cartan, 1967, p.20, Exemple 1), so we can view $\partial_l g(a)$ as an element of $\mathcal{Y}$. Moreover,

$$g'(a)(h) = g'(a)(h_1, ..., h_d) = \sum_{l=1}^{d} h_l \partial_l g(a), \qquad \text{for } h = (h_1, ..., h_d) \in \mathbb{R}^d.$$

Cartan (1967, p.40, Proposition 2.6.2) tells us that g is of class C^1 if and only if $\partial_l g : U \to \mathcal{Y}$ is continuous for each $l = 1, ..., d$.

Next, we consider higher-order derivatives. For $m \in \mathbb{N}$, a map $F : (\mathbb{R}^d)^m \to \mathcal{Y}$ is *m-linear* if, for each $k = 1, ..., m$ and any $a^{(1)}, ..., a^{(k-1)}, a^{(k+1)}, ..., a^{(m)} \in \mathbb{R}^d$, the map $x \mapsto F(a^{(1)}, ..., a^{(k-1)}, x, a^{(k+1)}, ..., a^{(m)})$ is linear from $\mathbb{R}^d$ into $\mathcal{Y}$ (Cartan, 1967, p.24). We say that F is an m-linear map from $\mathbb{R}^d$ into $\mathcal{Y}$, and denote by $\mathcal{L}_m(\mathbb{R}^d; \mathcal{Y})$ the space of all continuous m-linear maps from $\mathbb{R}^d$ into $\mathcal{Y}$[2]. The space $\mathcal{L}_m(\mathbb{R}^d; \mathcal{Y})$ can then be equipped with a natural operator norm defined by

$$\|F\|_{\text{op}} = \sup_{\|x^{(1)}\| \le 1, ..., \|x^{(m)}\| \le 1} \left\| F(x^{(1)}, ..., x^{(m)}) \right\|_{\mathcal{Y}}.$$

[2] Beware that $\mathcal{L}_m(\mathbb{R}^d; \mathcal{Y})$, the space of continuous m-linear maps from $\mathbb{R}^d$ into $\mathcal{Y}$, is different to $\mathcal{L}((\mathbb{R}^d)^m, \mathcal{Y})$, the space of continuous linear maps from $(\mathbb{R}^d)^m$ into $\mathcal{Y}$.

For any integer $m \in \mathbb{N}$, Coleman (2012, p.88, Theorem 4.4) tells us that

$$\Psi_m : \mathcal{L}(\mathbb{R}^d, \mathcal{L}_{m-1}(\mathbb{R}^d; \mathcal{Y})) \to \mathcal{L}_m(\mathbb{R}^d; \mathcal{Y}) \qquad \text{defined by}$$
$$\Psi_m(F)(x^{(1)}, x^{(2)}, ..., x^{(m)}) = F(x^{(1)})(x^{(2)}, ..., x^{(m)})$$

is an isometric isomorphism.

We say that $g : U \to \mathcal{Y}$ is *twice differentiable at* $a \in U$ if the derivative map $g' : U \to \mathcal{L}(\mathbb{R}^d, \mathcal{Y})$ is differentiable at a. We denote by $g''(a) = g^{(2)}(a) \in \mathcal{L}(\mathbb{R}^d, \mathcal{L}(\mathbb{R}^d, \mathcal{Y})) \simeq \mathcal{L}_2(\mathbb{R}^d; Y)$ the *second derivative of* g *at* a. We say that g is *twice differentiable on* U if it is twice differentiable at all points in U. Then we have a map $g^{(2)} : U \to \mathcal{L}_2(\mathbb{R}^d, \mathcal{Y})$. We say that g is *twice continuously differentiable on* U, or *of class* C^2 *on* U, if g is twice differentiable and if the map $g^{(2)}$ is continuous (Cartan, 1967, p.64). By continuing in this way, we say that g is m*-times differentiable at* $a \in U$ if $g^{(m-1)} : U \to \mathcal{L}_{m-1}(\mathbb{R}^d; \mathcal{Y})$ is differentiable at a, define the m^{th} *derivative* $g^{(m)}(a) \in \mathcal{L}_m(\mathbb{R}^d; \mathcal{Y})$ of g at a as the derivative of $g^{(m-1)}$ at a, and say that g is m-times differentiable on U if it is m-times differentiable at all points in U. We say that g is *of class* C^m *on* U if g is m-times differentiable at all points in U and the map $g^{(m)} : U \to \mathcal{L}_m(\mathbb{R}^d; \mathcal{Y})$ is continuous; we say that g is *of class* C^∞ if it is of class C^m for all $m \in \mathbb{N}$ (Cartan, 1967, pp.69–70).

Similarly, for $l_1 \in \{1, ..., d\}$, if the partial derivative $\partial_{l_1} g : U \to \mathcal{Y}$ is defined in some neighbourhood of $x \in U$ and is differentiable, then for $l_2 \in \{1, ..., d\}$ (which may or may not be distinct from l_1), we may define the second partial derivative $\partial_{l_1}\partial_{l_2} g(a) \in \mathcal{Y}$. If $l_1 = l_2 = l$, then we write $\partial_l \partial_l g = \partial_l^2 g$. Analogously to the first partial derivative, we have a formula that expresses the second derivative as a sum of second partial derivatives:

$$g''(a)((h_1^{(1)}, ..., h_d^{(1)}), (h_1^{(2)}, ..., h_d^{(2)})) = \sum_{l_1, l_2 = 1}^{d} h_{l_1}^{(1)} h_{l_2}^{(2)} \partial_{l_1} \partial_{l_2} g(a),$$

where $h^{(1)} = (h_1^{(1)}, ..., h_d^{(1)}), h^{(2)} = (h_1^{(2)}, ..., h_d^{(2)}) \in \mathbb{R}^d$ (Cartan, 1967, p.68, (5.2.5)). Continuing in the same way, we can define the m^{th} partial derivative $\partial_{l_1} ... \partial_{l_m} g(a) \in \mathcal{Y}$. Then writing $\mathbf{h} = (h^{(1)}, ..., h^{(m)}) \in (\mathbb{R}^d)^m$, we have

$$g^{(m)}(a)(\mathbf{h}) = \sum_{l_1, ..., l_m = 1}^{d} h_{l_1}^{(1)} ... h_{l_m}^{(m)} \partial_{l_1} ... \partial_{l_m} g(a).$$

Finally, we state the extension of Taylor's theorem to functions with values in $\mathcal{Y}$, with Lagrange's form of the remainder. To this end, for $a, b \in \mathbb{R}^d$, define the *segment* joining a and b as the set (Coleman, 2012, p.51).

$$[a, b] = \{x \in \mathbb{R}^d : x = va + (1 - v)b, v \in [0, 1]\}.$$

Theorem 5.3.1 (Cartan (1967, p.77, Théorème 5.6.2))**.** *Suppose that $g : U \to \mathcal{Y}$ is $(m+1)$-times differentiable, that the segment $[a, a+h]$ is contained in U and that, for some $K > 0$, we have*

$$\left\| g^{(m+1)}(x) \right\|_{\text{op}} \leq K \qquad \textit{for all } x \in U.$$

Then

$$\left\| g(a+h) - \sum_{k=0}^{m} \frac{1}{k!} g^{(k)}(a)((h)^k) \right\|_{\mathcal{Y}} \leq K \frac{\|h\|^{m+1}}{(m+1)!},$$

where we wrote $(h)^k = (h, ..., h) \in (\mathbb{R}^d)^k$ for $k = 1, ..., m$.

Write $\mathbb{N}_0 = \{0, 1, 2, ...\}$, and for $p = (p_1, ..., p_d) \in \mathbb{N}_0^d$, write $[p] := p_1 + ... + p_d$. Then we denote the p^{th} partial derivative $\partial_1^{p_1} ... \partial_d^{p_d} g(a)$ of g at $a \in U$ as $D^p g(a) \in \mathcal{Y}$. This is possible since the order of partial differentiation is immaterial by repeated application of Cartan (1967, p.69, Proposition 5.2.2). Hence, for each $k = 1, ..., m+1$, we have

$$g^{(k)}(a)((h)^k) = \sum_{l_1, \ldots, l_k = 1}^{d} h_{l_1} ... h_{l_k} \partial_{l_1} ... \partial_{l_k} g(a) = \sum_{[p]=k} \frac{k! h^p}{p!} D^p g(a),$$

where we wrote h^p as a shorthand for $h_1^{p_1} ... h_d^{p_d}$ and $p!$ for $p_1! ... p_d!$. Hence, using partial derivatives, we can express Taylor's theorem above as

$$\left\| g(a+h) - \sum_{[p] \leq m} \frac{h^p}{p!} D^p g(a) \right\|_{\mathcal{Y}} \leq K \frac{\|h\|^{m+1}}{(m+1)!}.$$

Finally, we introduce some notions from the theory of metric spaces. In particular, covering numbers play a central role in entropy discussions, and different notions of dimensions based on covering numbers will be used to restrict the range of partial derivatives of functions, leading up to entropy bounds in our main results (Section 5.4).

Suppose $(\mathcal{Z}, \rho)$ is a metric space. For $r > 0$ and $z_0 \in \mathcal{Z}$, the *ball of radius r centred at z_0* is $\mathcal{B}(z_0, r) = \{z \in \mathcal{Z} : \rho(z, z_0) \leq r\}$. For any $\delta > 0$, the *δ-covering number* of $(\mathcal{Z}, \rho)$, denoted by $N(\delta, \mathcal{Z}, \rho)$, is the minimum number of balls of radius δ with centres in $\mathcal{Z}$ required to cover $\mathcal{Z}$, i.e. the minimal N such that there exists a set $\{z_1, ..., z_N\} \subset \mathcal{Z}$ such that for all $z \in \mathcal{Z}$, there exists a $j = j(z) \in \{1, ..., N\}$ with $\rho(z, z_j) \leq \delta$ (we take $N(\delta, \mathcal{Z}, \rho) = \infty$ if no finite covering by closed balls with radius δ exists). We say that $\mathcal{Z}$ is *totally bounded* if $N(\delta, \mathcal{Z}, \rho) < \infty$ for all $\delta > 0$. We define the *δ-entropy* as $H(\delta, \mathcal{Z}, \rho) = \log N(\delta, \mathcal{Z}, \rho)$.

Let E be a subset of $(\mathcal{Z}, \rho)$. The *upper box-counting dimension* of E is

$$\tau_{\text{box}}(E) := \limsup_{\delta \to 0} \frac{H(\delta, E, \rho)}{-\log \delta}$$

(Robinson, 2010, p.32, Definition 3.1). It is immediate from the definition (Robinson, 2010, p.32, (3.3)) that if $\tau > \tau_{\text{box}}(E)$, then there exists $\delta_0 > 0$ such that for all $\delta < \delta_0$,

$$N(\delta, E, \rho) < \delta^{-\tau}. \tag{box}$$

A subset E of $(\mathcal{Z}, \rho)$ is said to be (M, τ)-*homogeneous* (or simply *homogeneous*) if the intersection of E with any closed ball of radius R can be covered by at most $M\left(\frac{R}{r}\right)^\tau$ closed balls of smaller radius r, i.e. $N(r, \mathcal{B}(z, R) \cap E, \rho) \leq M\left(\frac{R}{r}\right)^\tau$ for all $z \in E$ and $R > r$ (Robinson, 2010, p.83, Definition 9.1). The *Assouad dimension* (Robinson, 2010, p.85, Definition 9.5), sometimes also known as the *doubling dimension*, of E is

$$\tau_{\text{asd}}(E) := \inf\{\tau : E \text{ is } (M, \tau)\text{-homogeneous for some } M \geq 1\}.$$

5.4 Entropy of Classes of Smooth Vector-Valued Functions

In the usual empirical process theory with real-valued functions, classes of smooth functions on compact domains are some of the most frequently used examples that satisfy good entropy conditions (van de Geer, 2000, p.154, Example 9.3.2), (van der Vaart and Wellner, 1996, Section 2.7.1), (Dudley, 2014, Section 8.2). In this section, we give analogues of these results when the output space is the (not necessarily finite-dimensional) Hilbert space $\mathcal{Y}$.

Let $m \in \mathbb{N}$; this will determine the smoothness of our function class. Let $d \geq 1$, and take as our input space the unit cube in $\mathbb{R}^d$, $\mathcal{X} = \{x \in \mathbb{R}^d : 0 \leq x_j \leq 1, j = 1, ..., d\}$; this is only to simplify the exposition, and the subsequent results will clearly hold for any bounded convex subsets of $\mathbb{R}^d$.

In order to bound the entropy of classes of smooth real-valued functions, one bounds the absolute values of the range of the functions and their partial derivatives. When the output space is $\mathcal{Y}$, in particular, if $\mathcal{Y}$ has infinite dimensions, bounding the norm of the range is useless, because balls in infinite-dimensional spaces are not totally bounded. Therefore, to have any hope, the very least we need to do is to find a totally bounded subset $B \subset \mathcal{Y}$, and restrict our range and partial derivatives therein. As B is totally bounded, for some $K_B > 0$, $\|y\|_{\mathcal{Y}} \leq K_B$ for all $y \in B$.

Denote by $\mathcal{G}_B^m$ the set of m-times differentiable functions $g : \mathcal{X} \to \mathcal{Y}$ whose partial derivatives $D^p g : \mathcal{X} \to \mathcal{Y}$ of orders $[p] \leq m$ exist everywhere on the interior of $\mathcal{X}$, and such that $D^p g(x) \in B$ for all $x \in \mathcal{X}$ and $[p] \leq m$, where $D^0 g = g$. We present three results bounding $H(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty)$ for $\delta > 0$ sufficiently small, each with different assumptions on B. Theorem 5.4.1 assumes that B is homogeneous, i.e. we impose local entropy conditions. In Theorems 5.4.2 and 5.4.3, we impose global entropy conditions on B, the former with finite upper box-counting dimension, and the latter with $N(\delta, B, \|\cdot\|_{\mathcal{Y}})$ allowed to grow exponentially as δ decreases.

Theorem 5.4.1. *Let $B \subset \mathcal{Y}$ be totally bounded and (M, τ_{asd})-homogeneous. Then for sufficiently small $\delta > 0$, there exists some constant K depending on K_B, m, d, M and τ_{asd} such that*

$$H\left(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty\right) \leq K\delta^{-\frac{d}{m}}.$$

Theorem 5.4.1 gives the same rate for $\mathcal{G}_B^m$ as for smooth real-valued function classes (Dudley, 2014, p.288, Theorem 8.4(a)), which is a special case of the set-up in Theorem 5.4.1, since any bounded subset of $\mathbb{R}$ is a homogeneous subset (with Assouad dimension at most 1). In fact, Dudley (2014, Theorem 8.4(a)) shows that this rate of $\delta^{-\frac{d}{m}}$ cannot be improved, so the rate given in Theorem 5.4.1 is also optimal. We will later see from the proof that the dependence on τ_{asd} is linear.

Theorem 5.4.2. *Let B be a subset of $\mathcal{Y}$ with finite upper box-counting dimension τ_{box}. Then for sufficiently small $\delta > 0$, there exists some constant K depending on K_B, m, d and τ_{box} such that*

$$H\left(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty\right) \leq K\delta^{-\frac{d}{m}} \log\left(\frac{1}{\delta}\right).$$

Theorem 5.4.3. *Let B be a subset of $\mathcal{Y}$ with $N(\epsilon, B, \|\cdot\|_{\mathcal{Y}}) \leq \exp\{M\epsilon^{-\tau_{\text{exp}}}\}$ for some $M, \tau_{\text{exp}} > 0$. Then for sufficiently small $\delta > 0$, there is some constant K depending on K_B, m, d, M and τ_{exp} such that*

$$H\left(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty\right) \leq K\delta^{-\left(\frac{d}{m} + \tau_{\text{exp}}\right)}.$$

We can use results in Section 5.2 to show that we have uniform law of large numbers over $\mathcal{G}_B^m$, where B satisfies the conditions in any one of Theorems 5.4.1, 5.4.2 or 5.4.3.

Corollary 5.4.4. *The function class $\mathcal{G}_B^m$, where B is either homogeneous, has finite upper box-counting dimension or satisfies $N(\epsilon, B, \|\cdot\|_{\mathcal{Y}}) \leq \exp\{M\epsilon^{-\tau_{\text{exp}}}\}$ for some $\tau_{\text{exp}} > 0$, is Glivenko-Cantelli.*

Further, the empirical process defined by $\mathcal{G}_B^m$ (c.f. Definition 5.2.2) is asymptotically equicontinuous.

Corollary 5.4.5. *Suppose that B is either homogeneous, has finite upper box-counting dimension or satisfies $N(\epsilon, B, \|\cdot\|_{\mathcal{Y}}) \leq \exp\{M\epsilon^{-\tau_{\text{exp}}}\}$ for some $\tau_{\text{exp}} > 0$. Then the empirical process $\{\nu_n(g) = \sqrt{n}(P_n - P)g : g \in \mathcal{G}_B^m\}$ defined by $\mathcal{G}_B^m$ is asymptotically equicontinuous.*

5.4.1 Examples

With these results in hand, it is now of interest to investigate which interesting examples of output space $\mathcal{Y}$ and subsets B satisfy the conditions of Theorems 5.4.1, 5.4.2 and 5.4.3.

Example 5.4.6. *Suppose that $\mathcal{Y}$ is a finite-dimensional Hilbert space, say with dimension $d_{\mathcal{Y}}$. Then balls are totally bounded, so we can let B be of the form $B = \{y \in \mathcal{Y} : \|y\|_{\mathcal{Y}} \leq K\}$ for any $K > 0$. Moreover, subsets of finite-dimensional spaces are homogeneous with Assouad dimension at most $d_{\mathcal{Y}}$ (Robinson, 2010, p.85, Lemma 9.6(iii)), and so we can apply Theorem 5.4.1. The case $\mathcal{Y} = \mathbb{R}$ corresponds to the usual regression with real-valued output. If $\mathcal{Y} = \mathbb{R}^{d_{\mathcal{Y}}}$, it corresponds to the multi-task learning setting (Evgeniou et al., 2005; Yousefi et al., 2018; Xu et al., 2019).*

A prominent application of vector-valued output spaces will be when we have functional responses; example data sets include speech, diffusion tensor imaging, mass spectrometry and glaucoma (see Morris (2015); Kadri et al. (2016) and references therein). Let $\mathcal{X}'$ be a domain, and $\mathcal{Y} = L^2(\mathcal{X}', P'; \mathbb{R})$ the space of real-valued functions that are square-integrable with respect to some distribution P' on $\mathcal{X}'$. By considering interesting subsets of $\mathcal{Y}$, we can derive bounds on the entropy $H(\delta, \mathcal{G}^m_B, \|\cdot\|_\infty)$ using Theorems 5.4.1, 5.4.2 and 5.4.3. The next 4 examples are considered in this set-up.

Example 5.4.7. *Suppose that $\psi_1, ..., \psi_r \in \mathcal{Y}$, and let $B = \{f = \theta_1\psi_1 + ... + \theta_r\psi_r : \theta = (\theta_1, ..., \theta_r)^T \in \mathbb{R}^r, \|f\|_{2,P'} \leq R\}$ Then van de Geer (2000, p.20, Lemma 2.5) tells us that B is homogeneous, and so Theorem 5.4.1 applies. This corresponds to the case where the responses are finite-dimensional functions, or adopting the nomenclature of van de Geer (2000, p.152, Example 9.3.1), "linear regressors".*

Example 5.4.8. *More generally, function classes with finite Assouad dimensions have been considered in classification problems, and their generalisation properties analysed (Li and Long, 2007; Bshouty et al., 2009). If these functions form the responses of a regression problem, then Theorem 5.4.1 can again be applied. Examples of such function classes include halfspaces with respect to the uniform distribution (i.e. where P' is the uniform distribution) (Bshouty et al., 2009, Proposition 6).*

Example 5.4.9. *Let $\mathcal{X}'$ be compact in $\mathbb{R}^{d'}$ (in general, $d \neq d'$), and suppose that $B \subset \mathcal{Y}$ consists of smooth functions. More specifically, for some $m' \in \mathbb{N}$ and $M > 0$, let B be the set of all m'-times differentiable functions $f : \mathcal{X}' \to \mathbb{R}$ whose partial derivatives $D^q f : \mathcal{X}' \to \mathbb{R}$ of orders $[q] \leq m'$ exist everywhere on the interior of $\mathcal{X}'$, and such that $|D^q f(x')| \leq M$ for all $x' \in \mathcal{X}'$ and $[q] \leq m'$. Then applying the result for real-valued function classes (Dudley, 2014, p.288, Theorem 8.4) (or Theorem 5.4.1 with $\mathcal{Y} = \mathbb{R}$ and B being the ball of radius M), we have $N(\delta, B, \|\cdot\|_\infty) \leq \exp\{K'\delta^{-\frac{d'}{m'}}\}$ for some constant $K' > 0$. This in turn allows us to apply Theorem 5.4.3 to bound the entropy of $\mathcal{G}^m_B$ as*

$$H(\delta, \mathcal{G}^m_B, \|\cdot\|_\infty) \leq K\delta^{-\left(\frac{d}{m}+\frac{d'}{m'}\right)}$$

for some constant $K > 0$. So when the output space is itself a class of smooth (real-valued) functions, the smoothness of the two function classes simply add in the negative exponent of δ in the entropy.

Example 5.4.10. *Let B be a ball in a reproducing kernel Hilbert space (RKHS) with a $\mathcal{C}^\infty$ Mercer kernel (see Cucker and Smale (2002) for details), then Cucker and Smale (2002, Theorem D) tells us that for some constant K', we have $N(\delta, B, \|\cdot\|_\infty) \leq \exp\{K'\delta^{-\frac{2d}{h}}\}$ for any $h > d$. Then we can again apply Theorem 5.4.3 to bound $H(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty)$ by $K\delta^{-(\frac{d}{m}+\frac{2d}{h})}$ for some constant K and any $h > d$.*

5.5 Discussion on Rademacher Complexities

In this Section, we discuss the extension of the concept of Rademacher complexities to classes of vector-valued functions. We first give the definition of Rademacher complexities of classes of real-valued functions.

Definition 5.5.1 (Bartlett and Mendelson (2002, Definition 2))**.** Suppose $\mathcal{G}$ is a class of real-valued functions $\mathcal{X} \to \mathbb{R}$. Then the *empirical (or conditional) Rademacher complexity* of $\mathcal{G}$ is defined as

$$\hat{\mathfrak{R}}_n(\mathcal{G}) = \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i)\right| \mid X_1, ..., X_n\right],$$

where the expectation is taken with respect to $\{\sigma_i\}_{i=1}^n$. The Rademacher complexity of $\mathcal{G}$ is defined as

$$\mathfrak{R}_n(G) = \mathbb{E}\left[\hat{\mathfrak{R}}_n(G)\right].$$

Since this seminal definition, it was realised that the absolute value around $\frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i)$ was unnecessary (see, for example, Meir and Zhang (2003, paragraph between Corollary 4 and Lemma 5) or Maurer (2016, last paragraph of Section 1)). However, in order to facilitate the following direct extension to classes of vector-valued functions, we retain the absolute value sign.

Definition 5.5.2. Suppose $\mathcal{G}$ is a class of $\mathcal{X} \to \mathcal{Y}$ functions. Then the empirical (or conditional) Rademacher complexity of $\mathcal{G}$ is defined as

$$\hat{\mathfrak{R}}_n(\mathcal{G}) = \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\|\frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i)\right\|_{\mathcal{Y}} \mid X_1, ..., X_n\right] = \mathbb{E}\left[\|P_n^\sigma g\|_{\mathcal{G}} \mid X_1, ..., X_n\right],$$

using the notation from Section 5.2. The Rademacher complexity of $\mathcal{G}$ is defined as

$$\mathfrak{R}_n(G) = \mathbb{E}\left[\hat{\mathfrak{R}}_n(G)\right].$$

Note that our definition is different to the "vector-valued Rademacher complexity" already in use in the literature, mostly for $\mathcal{Y}$ being a finite-dimensional Euclidean space (Yousefi et al., 2018, Definition 1; Li et al., 2019, Definition 3), but also for $\mathcal{Y} = l_2$, the space of square-summable sequences (Maurer, 2016). These papers define the "Rademacher complexity" of vector-valued function classes not as in Definition 5.5.2, where we have one Rademacher variable σ_i per sample X_i, but introduce a Rademacher variable for every coordinate of $\mathcal{Y}$. The resulting quantity looks something like

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sum_{k}\sigma_i^k g_k(X_i)\mid X_1,...,X_n\right],$$

where g_k is the k^{th} coordinate of g with respect to a basis, and $\{\sigma_i^k\}_{i,k}$ are Rademacher random variables. For convenience, in what follows, we call this the "coordinate-wise Rademacher complexity", and denote it by $\hat{\mathfrak{R}}_n^{\text{coord}}(\mathcal{G})$.

While we recognise the usefulness of this definition, especially thanks to the contraction result shown in Maurer (2016), Cortes et al. (2016), Zatarain-Vera (2019) and Foster and Rakhlin (2019), for several reasons, we insist on using Definition 5.5.2. Firstly, as it is clear from the definition, and as admitted by Maurer (2016, paragraph just above Conjecture 2), Definition 5.5.2 is a more natural definition in view of the real-valued Rademacher complexity. Moreover, our work in Section 5.2.1 uses the empirical symmetrised measure $\frac{1}{n}\sum_{i=1}^{n}\sigma_i\delta_{X_i}$ to good effect and in a way that directly generalises from the real-valued case, which suggests that Definition 5.5.2 is natural. Finally, and perhaps most critically, the coordinate-wise Rademacher complexity is not independent of the choice of the basis of $\mathcal{Y}$. For a simple counterexample, let $\mathcal{X}=\mathcal{Y}=\mathbb{R}^2$, and $\mathcal{G}=\{g_1,g_2\}$, where g_1 is the orthogonal projection onto the line $y=x$, and g_2 is the orthogonal projection onto the line $y=-x$. This means that, letting $X_1=\begin{pmatrix}1\\0\end{pmatrix}$ and $X_2=\begin{pmatrix}0\\1\end{pmatrix}$, we have

$$g_1(X_1)=\begin{pmatrix}\frac{1}{2}\\\frac{1}{2}\end{pmatrix},\quad g_1(X_2)=\begin{pmatrix}\frac{1}{2}\\\frac{1}{2}\end{pmatrix},\quad g_2(X_1)=\begin{pmatrix}\frac{1}{2}\\-\frac{1}{2}\end{pmatrix},\quad g_2(X_2)=\begin{pmatrix}-\frac{1}{2}\\\frac{1}{2}\end{pmatrix}.$$

Then the coordinate-wise Rademacher complexity of $\mathcal{G}$ with respect to the standard basis $\{X_1,X_2\}$ is

$$\begin{aligned}\hat{\mathfrak{R}}_n^{\text{coord}}(\mathcal{G})&=\mathbb{E}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{2}\sum_{k=1}^{2}\sigma_i^k g_k(X_i)\right]\\&=\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\sigma_1^1\left(g(X_1)\right)_1+\sigma_1^2\left(g(X_1)\right)_2+\sigma_2^1\left(g(X_2)\right)_1+\sigma_2^2\left(g(X_2)\right)_2\right\}\right]\\&=\mathbb{E}\left[\frac{\sigma_1^1}{2}+\frac{\sigma_2^2}{2}+\sup_{g\in\mathcal{G}}\left\{\sigma_1^2\left(g(X_1)\right)_2+\sigma_2^1\left(g(X_2)\right)_1\right\}\right]\\&=\sup_{g\in\mathcal{G}}\left\{\left(g(X_1)\right)_2+\left(g(X_2)\right)_1\right\}+\sup_{g\in\mathcal{G}}\left\{-\left(g(X_1)\right)_2+\left(g(X_2)\right)_1\right\}\\&\qquad+\sup_{g\in\mathcal{G}}\left\{\left(g(X_1)\right)_2-\left(g(X_2)\right)_1\right\}+\sup_{g\in\mathcal{G}}\left\{-\left(g(X_1)\right)_2-\left(g(X_2)\right)_1\right\}\\&=1+0+0+1\\&=2.\end{aligned}$$

But if we use the orthonormal basis $\left\{ \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \right\}$, then we have

$$
\begin{aligned}
&(g_1(X_1))_1 = \frac{1}{\sqrt{2}}, \quad (g_1(X_1))_2 = 0, \quad (g_1(X_2))_1 = \frac{1}{\sqrt{2}} \quad (g_1(X_2))_2 = 0 \\
&(g_2(X_1))_1 = 0, \qquad (g_2(X_1))_2 = -\frac{1}{\sqrt{2}}, \qquad (g_2(X_2))_1 = 0, \quad (g_2(X_2))_2 = \frac{1}{\sqrt{2}}.
\end{aligned}
$$

So the complexity with respect to the standard basis $\{X_1, X_2\}$ is

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{g\in\mathcal{G}} \sum_{i=1}^{2}\sum_{k=1}^{2} \sigma_i^k g_k(X_i)\right] \\
&= \mathbb{E}\left[\sup_{g\in\mathcal{G}} \left\{\sigma_1^1 \left(g(X_1)\right)_1 + \sigma_1^2 \left(g(X_1)\right)_2 + \sigma_2^1 \left(g(X_2)\right)_1 + \sigma_2^2 \left(g(X_2)\right)_2\right\}\right] \\
&= \sup_{g\in\mathcal{G}} \left\{\left(g(X_1)\right)_1 + \left(g(X_1)\right)_2 + \left(g(X_2)\right)_1 + \left(g(X_2)\right)_2\right\} \\
&\qquad + \sup_{g\in\mathcal{G}} \left\{\left(g(X_1)\right)_1 + \left(g(X_1)\right)_2 + \left(g(X_2)\right)_1 - \left(g(X_2)\right)_2\right\} \\
&\qquad + \sup_{g\in\mathcal{G}} \left\{\left(g(X_1)\right)_1 + \left(g(X_1)\right)_2 - \left(g(X_2)\right)_1 + \left(g(X_2)\right)_2\right\} \\
&\qquad + \sup_{g\in\mathcal{G}} \left\{\left(g(X_1)\right)_1 - \left(g(X_1)\right)_2 + \left(g(X_2)\right)_1 + \left(g(X_2)\right)_2\right\} \\
&\qquad + \sup_{g\in\mathcal{G}} \left\{-\left(g(X_1)\right)_1 + \left(g(X_1)\right)_2 + \left(g(X_2)\right)_1 + \left(g(X_2)\right)_2\right\} \\
&\qquad + \sup_{g\in\mathcal{G}} \left\{\left(g(X_1)\right)_1 + \left(g(X_1)\right)_2 - \left(g(X_2)\right)_1 - \left(g(X_2)\right)_2\right\} \\
&\qquad + \sup_{g\in\mathcal{G}} \left\{\left(g(X_1)\right)_1 - \left(g(X_1)\right)_2 + \left(g(X_2)\right)_1 - \left(g(X_2)\right)_2\right\} \\
&\qquad + \sup_{g\in\mathcal{G}} \left\{-\left(g(X_1)\right)_1 + \left(g(X_1)\right)_2 + \left(g(X_2)\right)_1 - \left(g(X_2)\right)_2\right\} \\
&\qquad + \sup_{g\in\mathcal{G}} \left\{\left(g(X_1)\right)_1 - \left(g(X_1)\right)_2 - \left(g(X_2)\right)_1 + \left(g(X_2)\right)_2\right\} \\
&\qquad + \sup_{g\in\mathcal{G}} \left\{-\left(g(X_1)\right)_1 + \left(g(X_1)\right)_2 - \left(g(X_2)\right)_1 + \left(g(X_2)\right)_2\right\} \\
&\qquad + \sup_{g\in\mathcal{G}} \left\{-\left(g(X_1)\right)_1 - \left(g(X_1)\right)_2 + \left(g(X_2)\right)_1 + \left(g(X_2)\right)_2\right\} \\
&\qquad + \sup_{g\in\mathcal{G}} \left\{\left(g(X_1)\right)_1 - \left(g(X_1)\right)_2 - \left(g(X_2)\right)_1 - \left(g(X_2)\right)_2\right\} \\
&\qquad + \sup_{g\in\mathcal{G}} \left\{-\left(g(X_1)\right)_1 + \left(g(X_1)\right)_2 - \left(g(X_2)\right)_1 - \left(g(X_2)\right)_2\right\} \\
&\qquad + \sup_{g\in\mathcal{G}} \left\{-\left(g(X_1)\right)_1 - \left(g(X_1)\right)_2 + \left(g(X_2)\right)_1 - \left(g(X_2)\right)_2\right\} \\
&\qquad + \sup_{g\in\mathcal{G}} \left\{-\left(g(X_1)\right)_1 - \left(g(X_1)\right)_2 - \left(g(X_2)\right)_1 + \left(g(X_2)\right)_2\right\} \\
&\qquad + \sup_{g\in\mathcal{G}} \left\{-\left(g(X_1)\right)_1 - \left(g(X_1)\right)_2 - \left(g(X_2)\right)_1 - \left(g(X_2)\right)_2\right\}
\end{aligned}
$$

$$= \sqrt{2} + \sqrt{2} + 0 + \sqrt{2} + 0 + 0 + \sqrt{2} + 0 + \sqrt{2} + 0 + \sqrt{2} + 0 - \sqrt{2} + 0 + 0 + 0$$
$$= 5\sqrt{2}.$$

Hence, we see that the coordinate-wise Rademacher complexity is not independent of the chosen orthonormal basis. We deem this to be a critical issue with the coordinate-wise Rademacher complexity, because it is intuitively clear that the "complexity" of a function class should not depend on the choice of the basis of the output space. This is especially pertinent in our context, considering that our interest is primarily in the case when the output space $\mathcal{Y}$ is infinite-dimensional in which there may be no "standard basis".

One of the main ways of bounding the Rademacher complexity of real-valued function classes is to use the entropy. We show that the Rademacher complexity of vector-valued function classes $\mathcal{G}$ can be bounded using the entropy, a vector-valued analogue of Shalev-Shwartz and Ben-David (2014, p.338, Lemma 27.4). We use the chaining notation in Section 5.2.3, and also use Hoeffding's inequality in Hilbert spaces (Pinelis, 1992).

Theorem 5.5.3. *Let $S \in \mathbb{N}$ be any (large) integer. The empirical Rademacher complexity is bounded as*

$$\hat{\mathfrak{R}}_n(\mathcal{G}) \leq 2^{-(S+1)} R_n + \frac{2}{\sqrt{n}} J_n,$$

where we recall that $R_n = \sup_{g\in\mathcal{G}} \|g\|_{2,P_n}$ is the empirical radius and $J_n = \sum_{s=0}^{S} 2^{-s} R_n \sqrt{2H_{s+1}}$ is the uniform entropy bound.

When the Rademacher complexity is used in empirical risk minimisation for real-valued function classes $\mathcal{F}$, what we end up using is not the Rademacher complexity $\mathfrak{R}_n(\mathcal{F})$ of the function class itself, but that of the composition of the loss with the function class. The same is true for vector-valued empirical risk minimisation problems. More precisely, suppose we have a loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, and we denote by $\hat{g}_n$ the solution of the following empirical risk minimisation problem:

$$\hat{g}_n = \arg\min_{g\in\mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(Y_i, g(X_i)) = \arg\min_{g\in\mathcal{G}} \hat{\mathcal{R}}_n(g).$$

Denote by g^* the minimiser of the population risk:

$$g^* := \arg\min_{g\in\mathcal{G}} \mathbb{E}\left[\mathcal{L}(Y, g(X))\right] = \arg\min_{g\in\mathcal{G}} \mathcal{R}(g).$$

We want to know how fast $\mathcal{R}(\hat{g}_n)$ converges to the minimal risk $\mathcal{R}(g^*)$ as the sample size n increases. Here, actually, the standard result concerning Rademacher complexities applies directly – we will quote the following result.

Theorem 5.5.4 (Shalev-Shwartz and Ben-David (2014, p.328, Theorem 26.5))**.** *Assume that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $g \in \mathcal{G}$, we have $|\mathcal{L}(y, g(x))| \leq c$ for some* *constant $c > 0$. Then with probability at least $1 - \delta$, we have*

$$\mathcal{R}(\hat{g}_n) - \mathcal{R}(g^*) \leq 2\mathfrak{R}_n(\mathcal{L} \circ \mathcal{G}) + 5c\sqrt{\frac{2\log\left(\frac{8}{\delta}\right)}{n}}$$

where we used the notation $\mathcal{L} \circ \mathcal{G}$ for the class of functions $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ defined as

$$\mathcal{L} \circ \mathcal{G} := \{(x, y) \mapsto \mathcal{L}(y, g(x)) : g \in \mathcal{G}\}.$$

Now, the question is how to obtain a meaningful bound on the Rademacher complexity $\mathfrak{R}_n(\mathcal{L} \circ \mathcal{G})$ as $n \to \infty$. When $\mathcal{G}$ is a class of real-valued functions, the Contraction Lemma (Shalev-Shwartz and Ben-David, 2014, p.331, Lemma 26.9) tells us that if, for each $Y_i \in \mathbb{R}$, the map $y \mapsto \mathcal{L}(Y_i, y)$ is c-Lipschitz, then $\mathfrak{R}_n(\mathcal{L} \circ \mathcal{G})$ is bounded by $c\mathfrak{R}_n(\mathcal{G})$, so it is meaningful to work with $\mathfrak{R}_n(\mathcal{G})$. However, an analogue of this result when $\mathcal{G}$ is a class of $\mathcal{Y}$-valued functions is shown to be impossible via a counterexample, in Maurer (2016, Section 6).

As mentioned above, one of the main ways of bounding the Rademacher complexity is to use entropy. As our end goal is to bound the Rademacher complexity of $\mathcal{L} \circ \mathcal{G}$, there are two ways of going about this task with entropy. For real-valued function classes $\mathcal{F}$, what is commonly done is to bound the Rademacher complexity of $\mathcal{L} \circ \mathcal{F}$ with the Rademacher complexity of $\mathcal{F}$ using contraction, then to bound the Rademacher complexity of $\mathcal{F}$ by an expression involving the entropy, using chaining. As discussed before, contraction becomes difficult with vector-valued function classes. But we propose a different way that avoids contraction of Rademacher complexities. We can first bound the Rademacher complexity of $\mathcal{L} \circ \mathcal{G}$ with an expression involving the entropy of $\mathcal{L} \circ \mathcal{G}$, and use the following contraction result of entropies.

Lemma 5.5.5. *Suppose that for each $Y \in \mathcal{Y}$, the $\mathcal{Y} \to \mathbb{R}$ map $y \mapsto \mathcal{L}(Y, y)$ is c-Lipschitz for some constant $c > 0$, i.e. for $y_1, y_2 \in \mathcal{Y}$, $|\mathcal{L}(Y, y_1) - \mathcal{L}(Y, y_2)| \leq c\|y_1 - y_2\|_{\mathcal{Y}}$. Then for any $\delta > 0$, we have*

$$H(c\delta, \mathcal{L} \circ \mathcal{G}, \|\cdot\|_{2,P_n}) \leq H(\delta, \mathcal{G}, \|\cdot\|_{2,P_n}).$$

So for empirical risk minimisation problems with appropriate loss functions, it does make sense to consider the entropy of vector-valued function classes $\mathcal{G}$, while it remains as future work to investigate the use of the Rademacher complexity of $\mathcal{G}$.

Chapter 6

Conclusion and Future Directions

In this thesis, we detailed our contributions in two broad research domains – causality and kernel regression.

In Part I, we proposed a new, axiomatic framework thereof, based on measure theory, called *causal spaces*, by enriching probability spaces with causal kernels that encode information about what happens after an intervention. After the axiomatisation of the concept, as is always done in the development of any mathematical theory, we investigated several definitions and proved a number of results that stem from the axioms, including causal effects, interventions, sources, products, causal independence and transformations. We showed how the interventional aspects of existing frameworks can be captured by causal spaces, and finally we gave some explicit constructions, highlighting cases in which existing frameworks fall short.

Even if causal spaces prove with time to be the correct approach to axiomatise causality, there is much work to be done – in fact, all the more so in that case. Despite the start made in this thesis, we foresee that there are countless more objects to be defined and theorems to be proved. Perhaps most conspicuously, we only discussed the *interventional* aspects of the theory of causality, but the notion of *counterfactuals* is also seen as a key part of the theory, both *interventional counterfactuals* as advocated by Pearl's ladder of causation (Pearl and Mackenzie, 2018, Figure 1.2) and *backtracking counterfactuals* (Von Kügelgen et al., 2023). We leave this as essential future work. Only then will we be able to provide a full comparison with the counterfactual aspects of SCMs and the potential outcomes.

We also mention the distinction between *type* causality and *actual* causality. The former is a theory about general causality, involving statements such as "in general, smoking makes lung cancer more likely". Type causality is what we will be concerned with in this paper. Actual causality, on the other hand, is interested in whether a *particular* event was caused by a *particular* action, dealing with statements such as "Bob got lung cancer because he smoked for 30 years". It is an extremely interesting area of research that has far-reaching implications for concepts such as responsibility, blame, law, harm (Beckers et al., 2022, 2023), model explanation (Biradar et al., 2021) and algorithmic recourse (Karimi et al., 2022). Many definitions of actual causality have been proposed (Halpern and Pearl, 2005; Halpern, 2015, 2016), but the question of how to define actual causality is still not settled (Beckers, 2021). The current definitions of actual causality are all grounded on (variants) of SCMs, and though it was out of the scope of this thesis, it will be an interesting future research direction to consider how actual causality can be incorporated into our proposed framework.

Regarding operations on multiple causal spaces, one interesting direction could be to consider a categorical treatment. Although probability theory does not seem to be so amenable to a category-theoretic treatment as other mathematical objects, there have been some efforts to do so (Lynn, 2010; Adachi and Ryu, 2016; Cho and Jacobs, 2019; Fritz, 2020). As future work, it would be interesting to explore extensions of the transformations proposed here to formal category-theoretic morphisms between causal spaces.

As a final note on our contributions to causality, it must be stressed that our goal should *not* be understood as replacing existing frameworks. Indeed, causal spaces cannot compete in terms of interpretability, and in the vast majority of situations in which SCMs, potential outcomes or any of the other existing frameworks are suitable, we expect them to be much more useful. In particular, assumptions are unavoidable for identifiability from observational data, and those assumptions are much better captured by existing frameworks[1], However, just as measure-theoretic probability theory has its value despite not being useful for practitioners in applied statistics, we believe that it is a worthy endeavour to formally axiomatise causality.

Regarding kernel regression, our contributions were much more contained within existing fields. Kernel conditional mean embeddings have been around for over a decade, and our contributions in Chapter 3 were to provide a new interpretation of them as Bochner conditional expectations, which, compared to the previous operator-based approaches, it does not rely on stringent assumptions that are often violated in common situations. Using this new approach, we discussed how to obtain empirical estimates via natural vector-valued regression, and established some theoretical results based on this regression interpretation. Finally, we extended the notions of the MMD, witness function and HSIC to the conditional case.

In Chapter 4, we discussed the analysis of the conditional distributional treatment effect (CoDiTE). We first propose a new kernel-based hypothesis test via kernel conditional mean embeddings to see whether there exists any CoDiTE. Then we proceeded to investigate the nature of the treatment effect via conditional witness functions, revealing where and how much the conditional densities differ, and U-statistic regression, which is informative about the differences in

[1] Researchers from the potential outcomes community and the graphical model community are arguing as to which framework is better for which situations (Imbens, 2019; Pearl, 2009). We do not take part in this debate.

specific conditional distributional quantities.

We foresee that much of the work that has been done by the machine learning community on treatment effect analysis, although cast mostly in the context of CATE, applies for the CoDiTE. Examples include *meta learners* (Künzel et al., 2019), *model validation* (Alaa and Van Der Schaar, 2019), *subgroup analysis* (Su et al., 2009; Lee et al., 2020) and *covariate balancing* (Gretton et al., 2009; Kallus, 2018). A major obstacle in any covariate-conditional analysis of treatment effect is this: when the covariate space is high-dimensional, the accuracy and reliability of the estimates deteriorate significantly due to the curse of dimensionality, and we heavily rely on changes to be smooth across the covariate space. This limitation is present not only in methods presented in this paper, but any CATE or CoDiTE analysis. While out of scope for the present thesis, it is of interest to investigate how to mitigate this problem.

Perhaps the most innovative contribution of Part II was made in Chapter 5, where the theory of empirical processes was extended to the vector-valued case. In particular, we investigated the metric entropy of smooth functions, by restricting the partial derivatives to take values in totally bounded subsets with specific properties, leveraging theory from fractal geometry, and demonstrated its application in empirical risk minimisation.

There is a plethora of possible future research directions. Considering other classes of functions than those of smooth functions is a natural next step. Also, we let $\mathcal{Y}$ be a Hilbert space, primarily because some simplifications occur for Hoeffding's inequality and Gaussian measures, but extensions to Banach spaces should be possible. Moreover, we used compact subsets of $\mathbb{R}^d$ as our input space due to the ease in considering partial derivatives, but interesting applications exist for which the input space $\mathcal{X}$ is a subset of an infinite-dimensional space (Li et al., 2020; Nelsen and Stuart, 2021; Lu et al., 2021). On the more theoretical side, measurability questions for empirical processes and uniform central limit theorems involving Gaussian elements in vector spaces are interesting questions. Also, obtaining complementary lower bounds, so that our upper bounds are minimax optimal, is an interesting problem. With empirical risk minimisation, extensions to more general noise with vector-valued Bernstein's inequality or misspecified models are important.

Bibliography

A. Abadie. Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models. *Journal of the American statistical Association*, 97(457): 284–292, 2002.

T. Adachi and Y. Ryu. A Category of Probability Spaces. *arXiv preprint arXiv:1611.03630*, 2016.

T. E. Ahmad, P. Laforgue, and F. d'Alché Buc. p-Sparsified Sketches for Fast Multiple Output Kernel Methods. *arXiv preprint arXiv:2206.03827*, 2022.

A. Alaa and M. Schaar. Limits of Estimating Heterogeneous Treatment Effects: Guidelines for Practical Algorithm Design. In *International Conference on Machine Learning*, pages 129–138, 2018.

A. Alaa and M. Van Der Schaar. Validating Causal Inference Models via Influence Functions. In *International Conference on Machine Learning*, pages 191–201, 2019.

A. M. Alaa and M. van der Schaar. Bayesian Inference of Individualized Treatment Effects using Multi-Task Gaussian Processes. In *Advances in Neural Information Processing Systems*, pages 3424–3432, 2017.

A. M. Alaa and M. van der Schaar. Bayesian Nonparametric Causal Inference: Information Rates and Learning Algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):1031–1046, 2018.

M. A. Álvarez, L. Rosasco, N. D. Lawrence, et al. Kernels for Vector-Valued Functions: A Review. *Foundations and Trends® in Machine Learning*, 4(3): 195–266, 2012.

D. K. Arrowsmith and C. M. Place. *An Introduction to Dynamical Systems.* Cambridge university press, 1990.

O. Atan, J. Jordon, and M. van der Schaar. Deep-Treat: Learning Optimal Personalized Treatments from Observational Data using Neural Networks. In *AAAI*, pages 2071–2078, 2018.

E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. *On Pearl's Hierarchy and the Foundations of Causal Inference*, page 507–556. Association for Computing Machinery, New York, NY, USA, 1st edition, 2022.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3 (Nov):463–482, 2002.

S. Beckers. Causal Sufficiency and Actual Causation. *Journal of Philosophical Logic*, 50(6):1341–1374, 2021.

S. Beckers and J. Y. Halpern. Abstracting Causal Models. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 2678–2685, 2019.

S. Beckers, F. Eberhardt, and J. Y. Halpern. Approximate Causal Abstractions. In *Uncertainty in Artificial Intelligence*, pages 606–615. PMLR, 2020.

S. Beckers, H. Chockler, and J. Halpern. A Causal Analysis of Harm. *Advances in Neural Information Processing Systems*, 35:2365–2376, 2022.

S. Beckers, H. Chockler, and J. Y. Halpern. Quantifying Harm. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023)*, 2023.

H. Beebee, C. Hitchcock, and P. Menzies. *The Oxford Handbook of Causation*. Oxford University Press, 2009.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, 2004.

A. T. Bharucha-Reid. *Random Integral Equations*. Academic Press, 1972.

G. Biradar, V. Viswanathan, and Y. Zick. Model Explanations via the Axiomatic Causal Lens. *arXiv preprint arXiv:2109.03890*, 2021.

M. P. Bitler, J. B. Gelbach, and H. W. Hoynes. Can Variation in Subgroups' Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment. *Review of Economics and Statistics*, 99(4): 683–697, 2017.

G. Blanchard and N. Mücke. Optimal Rates for Regularization of Statistical Inverse Learning Problems. *Foundations of Computational Mathematics*, 18 (4):971–1013, 2018.

T. Blom, S. Bongers, and J. M. Mooij. Beyond Structural Causal Models: Causal Constraints Models. In *Uncertainty in Artificial Intelligence*, pages 585–594. PMLR, 2020.

B. Bollobás. *Linear Analysis: An Introductory Course*. Cambridge University Press, 1999.

S. Bongers, T. Blom, and J. M. Mooij. Causal Modeling of Dynamical Systems. *arXiv preprint arXiv:1803.08784*, 2018.

S. Bongers, P. Forré, J. Peters, and J. M. Mooij. Foundations of Structural Causal Models with Cycles and Latent Variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

R. Brault. *Large-Scale Operator-Valued Kernel Regression.* PhD thesis, Université Paris Saclay, 2017.

M. Braun and M. Golubitsky. *Differential Equations and their Applications*, volume 2. Springer, 1983.

J. Brehmer, P. De Haan, P. Lippe, and T. S. Cohen. Weakly Supervised Causal Representation Learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022.

G. Briseño Sanchez, M. Hohberg, A. Groll, and T. Kneib. Flexible Instrumental Variable Distributional Regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(4):1553–1574, 2020.

N. H. Bshouty, Y. Li, and P. M. Long. Using the Doubling Dimension to Analyze the Generalization of Learning Algorithms. *Journal of Computer and System Sciences*, 75(6):323–335, 2009.

V. A. Cabannes, F. R. Bach, and A. Rudi. Fast Rates for Structured Prediction. In *Conference on Learning Theory, COLT 2021*, pages 823–865, 2021.

I. Cabreros and J. D. Storey. Causal Models on Probability Spaces. *arXiv preprint arXiv:1907.01672*, 2019.

A. Caponnetto and E. De Vito. Optimal Rates for the Regularized Least-Squares Algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2006.

C. Carmeli, E. De Vito, and A. Toigo. Vector Valued Reproducing Kernel Hilbert Spaces of Integrable Functions and Mercer Theorem. *Analysis and Applications*, 4(04):377–408, 2006.

C. Carmeli, E. De Vito, A. Toigo, and V. Umanitá. Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

H. Cartan. *Calcul Différentiel.* Hermann, 1967.

N. Cartwright. *The Dappled World: A Study of the Boundaries of Science.* Cambridge University Press, 1999.

K. Chalupka, P. Perona, and F. Eberhardt. Visual causal feature learning. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, page 181–190, Arlington, Virginia, USA, 2015. AUAI Press. ISBN 9780996643108.

K. Chalupka, T. Bischoff, P. Perona, and F. Eberhardt. Unsupervised discovery of el nino using causal feature learning on microlevel climate data. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, page 72–81, Arlington, Virginia, USA, 2016. AUAI Press. ISBN 9780996643115.

K. Chalupka, F. Eberhardt, and P. Perona. Causal feature learning: an overview. *Behaviormetrika*, 44(1):137–164, 2017. doi: 10.1007/s41237-016-0008-2. URL https://doi.org/10.1007/s41237-016-0008-2.

M. Chang, S. Lee, and Y.-J. Whang. Nonparametric Tests of Conditional Treatment Effects with an Application to Single-Sex Schooling on Academic Achievements. *The Econometrics Journal*, 18(3):307–346, 2015.

V. Chernozhukov, I. Fernández-Val, and B. Melly. Inference on Counterfactual Distributions. *Econometrica*, 81(6):2205–2268, 2013.

V. Chernozhukov, I. Fernandez-Val, and M. Weidner. Network and Panel Quantile Effects via Distribution Regression. *Journal of Econometrics*, 2020.

K. Cho and B. Jacobs. Disintegration and Bayesian Inversion via String Diagrams. *Mathematical Structures in Computer Science*, 29(7):938–971, 2019.

K. Chwialkowski, H. Strathmann, and A. Gretton. A Kernel Test of Goodness of Fit. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 2606–2615, 2016.

C. Ciliberto, L. Rosasco, and A. Rudi. A General Framework for Consistent Structured Prediction with Implicit Loss Embeddings. *J. Mach. Learn. Res.*, 21(98):1–67, 2020.

E. Çınlar. *Probability and Stochastics*, volume 261. Springer Science & Business Media, 2011.

D. T. Cohen and A. Kontorovich. Metric-Valued Regression. *arXiv preprint arXiv:2202.03045*, 2022.

T. Cohen. Towards a Grounded Theory of Causation for Embodied AI. In *UAI 2022 Workshop on Causal Representation Learning*, 2022.

D. L. Cohn. *Measure Theory*. Birkhäuser, 2013.

R. Coleman. *Calculus on Normed Vector Spaces*. Springer Science & Business Media, 2012.

J. Collins, N. Hall, and L. A. Paul. *Causation and Counterfactuals*. The MIT Press, 2004.

J. B. Conway. *A Course in Functional Analysis*, volume 96. Springer, 1990.

C. Cortes, V. Kuznetsov, M. Mohri, and S. Yang. Structured Prediction Theory based on Factor Graph Complexity. *Advances in Neural Information Processing Systems*, 29:2514–2522, 2016.

R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Nonparametric Tests for Treatment Effect Heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405, 2008.

F. Cucker and S. Smale. On the Mathematical Foundations of Learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.

P. Dawid. Decision-Theoretic Foundations for Statistical Causality. *Journal of Causal Inference*, 9(1):39–77, 2021.

R. H. Dehejia and S. Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062, 1999.

A. Derumigny. Estimation of a Regular Conditional Functional by Conditional U-Statistics Regression. *arXiv preprint arXiv:1903.10914*, 2019.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer Science & Business Media, 1996.

M. Dhanakshirur, F. Laumann, J. Park, and M. Barahona. A Continuous Structural Intervention Distance to Compare Causal Graphs. *arXiv preprint arXiv:2307.16452*, 2023.

N. Dinculeanu. *Vector Integration and Stochastic Integration in Banach Spaces*, volume 48. John Wiley & Sons, 2000.

A. J. Dobson. *Introduction to Statistical Modelling.* Springer, 2013.

R. M. Dudley. *Uniform Central Limit Theorems*, volume 142. Cambridge university press, 2014.

R. M. Dudley. *Real Analysis and Probability.* CRC Press, 2018.

H. B. Enderton. *Elements of Set Theory.* Academic press, 1977.

H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*, volume 375. Springer Science & Business Media, 1996.

T. Evgeniou, C. A. Micchelli, M. Pontil, and J. Shawe-Taylor. Learning Multiple Tasks with Kernel Methods. *Journal of machine learning research*, 6(4), 2005.

A. Feder, K. A. Keith, E. Manzoor, R. Pryzant, D. Sridhar, Z. Wood-Doughty, J. Eisenstein, J. Grimmer, R. Reichart, M. E. Roberts, et al. Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. *Transactions of the Association for Computational Linguistics*, 10: 1138–1158, 2022.

D. J. Foster and A. Rakhlin. l_∞ Vector Contraction for Rademacher Complexity. *arXiv preprint arXiv:1911.06468*, 6, 2019.

J. C. Foster, J. M. Taylor, and S. J. Ruberg. Subgroup Identification from Randomized Clinical Trial Data. *Statistics in medicine*, 30(24):2867–2880, 2011.

T. Fritz. A Synthetic Approach to Markov Kernels, Conditional Independence and Theorems on Sufficient Statistics. *Advances in Mathematics*, 370:107239, 2020.

T. Fritz, T. Gonda, N. G. Houghton-Larsen, P. Perrone, and D. Stein. Dilations and Information Flow Axioms in Categorical Probability. *arXiv preprint arXiv:2211.02507*, 2022.

D. Fudenberg and J. Tirole. *Game Theory.* MIT press, 1991.

K. Fukumizu. Nonparametric Bayesian Inference with Kernel Mean Embedding. In *Modern Methodology and Applications in Spatial-Temporal Modeling*, pages 1–24. Springer, 2015.

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

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel Measures of Conditional Dependence. In *Advances in neural information processing systems*, pages 489–496, 2008.

K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes' Rule: Bayesian Inference with Positive Definite Kernels. *The Journal of Machine Learning Research*, 14(1):3753–3783, 2013.

D. Galles and J. Pearl. An Axiomatic Characterization of Causal Counterfactuals. *Foundations of Science*, 3(1):151–182, 1998.

A. Geiger, H. Lu, T. Icard, and C. Potts. Causal Abstractions of Neural Networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021.

A. Geiger, C. Potts, and T. Icard. Causal Abstraction for Faithful Model Interpretation. *arXiv preprint arXiv:2301.04709*, 2023.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning.* MIT press, 2016.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring Statistical Dependence with Hilbert-Schmidt Norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A Kernel Method for the Two-Sample-Problem. In *Advances in neural information processing systems*, pages 513–520, 2007.

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A Kernel Statistical Test of Independence. In *Advances in neural information processing systems*, pages 585–592, 2008.

A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate Shift by Kernel Mean Matching. *Dataset shift in machine learning*, 3(4):5, 2009.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(Mar): 723–773, 2012.

S. Grünewälder, G. Lever, L. Baldassarre, S. Patterson, A. Gretton, and M. Pontil. Conditional Mean Embeddings as Regressors. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1803–1810, 2012a.

S. Grünewälder, G. Lever, L. Baldassarre, M. Pontil, and A. Gretton. Modelling Transition Dynamics in MDPs with RKHS Embeddings. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1603–1610. Omnipress, 2012b.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer Science & Business Media, 2006.

P. R. Hahn, J. S. Murray, and C. M. Carvalho. Bayeisan Regression Tree Models for Causal Inference: Regularisation, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis*, 15(3):965–1056, 09 2020.

B. C. Hall. *Quantum Theory for Mathematicians.* Springer, 2013.

J. Halpern. A Modification of the Halpern-Pearl Definition of Causality. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

J. Y. Halpern. Axiomatizing Causal Reasoning. *Journal of Artificial Intelligence Research*, 12:317–337, 2000.

J. Y. Halpern. *Actual Causality.* MIT Press, 2016.

J. Y. Halpern and J. Pearl. Causes and Explanations: A Structural-Model Approach. Part I: Causes. *British Journal for the Philosophy of Science*, 56 (4):843–843, 2005.

S. Hanneke, A. Kontorovich, S. Sabato, and R. Weiss. Universal Bayes Consistency in Metric Spaces. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–33. IEEE, 2020.

M. Hernan and J. Robins. *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC, 2020.

B. Heymann, M. De Lara, and J.-P. Chancelier. Causal inference theory with information dependency models. *arXiv preprint arXiv:2108.03099*, 2021.

J. L. Hill. Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel Methods in Machine Learning. *The annals of statistics*, pages 1171–1220, 2008.

M. Hohberg, P. Pütz, and T. Kneib. Treatment Effects Beyond the Mean Using Distributional Regression: Methods and Guidance. *Plos one*, 15(2):e0226514, 2020.

P. W. Holland. Statistics and Causal Inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

D. Ibeling and T. Icard. Probabilistic Reasoning Across the Causal Hierarchy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10170–10177, 2020.

D. Ibeling and T. Icard. Comparing Causal Frameworks: Potential Outcomes, Structural Models, Graphs, and Abstractions. *arXiv preprint arXiv:2306.14351*, 2023.

P. M. Illari, F. Russo, and J. Williamson. *Causality in the Sciences.* Oxford University Press, 2011.

G. Imbens. Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. Technical report, National Bureau of Economic Research, 2019.

G. W. Imbens and D. B. Rubin. *Causal Inference in Statistics, Social, and Biomedical sciences.* Cambridge University Press, 2015.

G. W. Imbens and J. M. Wooldridge. Recent Developments in the Econometrics of Program Evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

B. Jacobs, A. Kissinger, and F. Zanasi. Causal Inference by String Diagram Surgery. In *Foundations of Software Science and Computation Structures: 22nd International Conference, FOSSACS 2019, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2019, Prague, Czech Republic, April 6–11, 2019, Proceedings 22*, pages 313–329. Springer, 2019.

D. Janzing and B. Schölkopf. Causal Inference Using the Algorithmic Markov Condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

E. T. Jaynes. *Probability Theory: The Logic of Science.* Cambridge university press, 2003.

A. Jesson, S. Mindermann, U. Shalit, and Y. Gal. Identifying Causal-Effect Inference Failure with Uncertainty-Aware Models. *Advances in Neural Information Processing Systems*, 33, 2020.

Z. Jin, A. Feder, and K. Zhang. CausalNLP Tutorial: An Introduction to Causality for Natural Language Processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 17–22, 2022.

F. Johansson, U. Shalit, and D. Sontag. Learning Representations for Counterfactual Inference. In *International conference on machine learning*, pages 3020–3029, 2016.

H. Kadri, E. Duflos, P. Preux, S. Canu, A. Rakotomamonjy, and J. Audiffren. Operator-Valued Kernels for Learning from Functional Response Data. *The Journal of Machine Learning Research*, 17(1):613–666, 2016.

N. Kallus. Optimal A Priori Balance in the Design of Controlled Experiments. *Journal of the Royal Statistical Society Series B*, 80(1):85–112, 2018.

A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Computing Surveys*, 55(5):1–29, 2022.

A. Kekić, B. Schölkopf, and M. Besserve. Targeted reduction of causal models, 2023.

K. Kim, J. Kim, and E. H. Kennedy. Causal Effects Based on Distributional Distances. *arXiv preprint arXiv:1806.02935*, 2018.

I. Klebanov, I. Schuster, and T. Sullivan. A Rigorous Theory of Conditional Mean Embeddings. *SIAM Journal on Mathematics of Data Science*, 2(3): 583–606, 2020.

R. Koenker. *Quantile Regression.* Cambridge University Press, 2005.

A. Kolmogorov. Bounds for the Minimal Number of Elements of an ε-net in Various Classes of Functions and Their Applications to the Question of Representability of Functions of Several Variables by Superpositions of Functions of Fewer Variables. *Uspekhi Mat. Nauk (NS)*, 10:192–194, 1955.

M. R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference.* Springer, 2008.

S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for Estimating Heterogeneous Treatment Effects using Machine Learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

P. Laforgue, A. Lambert, L. Brogat-Motte, and F. d'Alché Buc. Duality in RKHSs with Infinite Dimensional Outputs: Application to Robust Losses. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620, 1986.

A. Lambert, D. Bouche, Z. Szabo, and F. d'Alché Buc. Functional Output Regression with Infimal Convolution: Exploring the Huber and ε-Insensitive Losses. In *International Conference on Machine Learning*, pages 11844–11867. PMLR, 2022.

F. Laumann, J. Von Kügelgen, J. Park, B. Schölkopf, and M. Barahona. Kernel-based Independence Tests for Causal Structure Learning on Functional Data. *Entropy*, 25(12):1597, 2023.

H.-S. Lee, Y. Zhang, W. Zame, C. Shen, J.-W. Lee, and M. van der Schaar. Robust Recursive Partitioning for Heterogeneous Treatment Effects with Uncertainty Quantification. *Advances in Neural Information Processing Systems*, 33, 2020.

M.-J. Lee. Non-parametric Tests for Distributional Treatment Effect for Randomly Censored Responses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):243–264, 2009.

S. S. Lee and Y.-J. Whang. Nonparametric Tests of Conditional Treatment Effects. Technical report, Cowles Foundation for Research in Economics, Yale University, 2009.

D. Lewis. *Counterfactuals*. John Wiley & Sons, 2013.

C. Lewis-Beck and M. Lewis-Beck. *Applied Regression: An Introduction*, volume 22. Sage publications, 2015.

J. Li, Y. Liu, and W. Wang. Learning Vector-Valued Functions with Local Rademacher Complexity. *arXiv preprint arXiv:1909.04883*, 2019.

Y. Li and P. M. Long. Learnability and the Doubling Dimension. *Advances in neural information processing systems*, 19:889, 2007.

Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Neural Operator: Graph Kernel Network for Partial Differential Equations. *arXiv preprint arXiv:2003.03485*, 2020.

Q. Liu, J. Lee, and M. Jordan. A Kernelized Stein Discrepancy for Goodness-of-Fit Tests. In *International conference on machine learning*, pages 276–284, 2016.

J. R. Lloyd and Z. Ghahramani. Statistical Model Criticism using Kernel Two Sample Tests. *Advances in Neural Information Processing Systems*, 28:829–837, 2015.

F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin. Towards a Learning Theory of Cause-Effect Inference. In *International Conference on Machine Learning*, pages 1452–1461, 2015.

C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal Effect Inference with Deep Latent-Variable Models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis. Learning Nonlinear Operators via DeepONet based on the Universal Approximation Theorem of Operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

M. Lynn. Categories of Probability Spaces, 2010.

U. Marteau-Ferey, F. Bach, and A. Rudi. Globally Convergent Newton Methods for Ill-Conditioned Generalized Self-Concordant Losses. In *Advances in Neural Information Processing Systems*, 2019.

R. Massidda, A. Geiger, T. Icard, and D. Bacciu. Causal Abstraction with Soft Interventions. In *2nd Conference on Causal Learning and Reasoning*, 2023.

A. Maurer. A Vector-Contraction Inequality for Rademacher Complexities. In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer, 2016.

G. Meanti, L. Carratino, L. Rosasco, and A. Rudi. Kernel Methods Through the Roof: Handling Billions of Points Efficiently. *Advances in Neural Information Processing Systems*, 33, 2020.

R. Meir and T. Zhang. Generalization Error Bounds for Bayesian Mixture Algorithms. *Journal of Machine Learning Research*, 4(Oct):839–860, 2003.

C. A. Micchelli and M. Pontil. On Learning Vector-Valued Functions. *Neural computation*, 17(1):177–204, 2005.

J. Mitrovic, D. Sejdinovic, and Y. W. Teh. Causal Inference via Kernel Deviance Measures. In *Advances in Neural Information Processing Systems*, pages 6986–6994, 2018.

J. Mitrovic, B. McWilliams, J. C. Walker, L. H. Buesing, and C. Blundell. Representation learning via invariant causal mechanisms. In *International Conference on Learning Representations*, 2020.

D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis*. John Wiley & Sons, 2021.

J. S. Morris. Functional Regression. *Annual Review of Statistics and Its Application*, 2:321–359, 2015.

E. Mourier. Eléments Aléatoires dans un Espace de Banach. In *Annales de l'institut Henri Poincaré*, volume 13, pages 161–244, 1953.

K. Muandet, K. Fukumizu, F. Dinuzzo, and B. Schölkopf. Learning from Distributions via Support Measure Machines. In *Advances in neural information processing systems*, pages 10–18, 2012.

K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, et al. Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

K. Muandet, M. Kanagawa, S. Saengkyongam, and S. Marukatat. Counterfactual Mean Embedding. *arXiv preprint arXiv:1805.08845*, 2018.

N. H. Nelsen and A. M. Stuart. The Random Feature Model for Input-Output Maps between Banach Spaces. *SIAM Journal on Scientific Computing*, 43 (5):A3212–A3243, 2021.

Y. Nishiyama, A. Boularias, A. Gretton, and K. Fukumizu. Hilbert Space Embeddings of POMDPs. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 644–653. AUAI Press, 2012.

P. A. Ortega. Subjectivity, bayesianism, and causality. *Pattern Recognition Letters*, 64:63–70, 2015.

J. Otsuka and H. Saigo. On the Equivalence of Causal Models: A Category-Theoretic Approach. In *Conference on Causal Learning and Reasoning*, pages 634–646. PMLR, 2022.

J. Otsuka and H. Saigo. Process Theory of Causality: a Category-Theoretic Perspective. *Behaviormetrika*, pages 1–16, 2023.

H. Owhadi and C. Scovel. Separability of Reproducing Kernel Spaces. *Proceedings of the American Mathematical Society*, 145(5):2131–2138, 2017.

J. Park and K. Muandet. A Measure-Theoretic Approach to Kernel Conditional Mean Embeddings. In *Advances in Neural Information Processing Systems*, volume 33, pages 21247–21259, 2020a.

J. Park and K. Muandet. Regularised Least-Squares Regression with Infinite-Dimensional Output Space. *arXiv preprint arXiv:2010.10973*, 2020b.

J. Park and K. Muandet. Towards Empirical Process Theory for Vector-Valued Functions: Metric Entropy of Smooth Function Classes. In *International Conference on Algorithmic Learning Theory*, pages 1216–1260. PMLR, 2023.

J. Park, U. Shalit, B. Schölkopf, and K. Muandet. Conditional Distributional Treatment Effect with Kernel Conditional Mean Embeddings and U-Statistic Regression. In *International Conference on Machine Learning*, pages 8401–8412. PMLR, 2021.

J. Park, S. Buchholz, B. Schölkopf, and K. Muandet. A measure-theoretic axiomatisation of causality. In *Advances in Neural Information Processing Systems*, volume 36, pages 28510–28540, 2023.

J. Pearl. *Causality*. Cambridge university press, 2009.

J. Pearl and D. Mackenzie. *The Book of Why*. Basic Books, New York, 2018.

J. Pearl, M. Glymour, and N. P. Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, 2016.

M. D. Perlman. Jensen's Inequality for a Convex Vector-Valued Function on an Infinite-Dimensional Space. *Journal of Multivariate Analysis*, 4(1):52–65, 1974.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference*. The MIT Press, 2017.

J. Peters, S. Bauer, and N. Pfister. Causal Models for Dynamical Systems. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 671–690. ACM, 2022.

S. Peters and J. Y. Halpern. Causal Modeling with Infinitely Many Variables. *arXiv preprint arXiv:2112.09171*, 2021.

I. Pinelis. An Approach to Inequalities for the Distributions of Infinite-Dimensional Martingales. In *Probability in Banach Spaces, 8: Proceedings of the Eighth International Conference*, pages 128–134. Springer, 1992.

S. Powers, J. Qian, K. Jung, A. Schuler, N. H. Shah, T. Hastie, and R. Tibshirani. Some Methods for Heterogeneous Treatment Effect Estimation in High Dimensions. *Statistics in medicine*, 37(11):1767–1787, 2018.

R. Precup. *Methods in Nonlinear Integral Equations*. Springer Science & Business Media, 2002.

H. Reeve and A. Kaban. Optimistic Bounds for Multi-Output Learning. In *International Conference on Machine Learning*, pages 8030–8040. PMLR, 2020.

Y. Ren, J. Zhu, J. Li, and Y. Luo. Conditional Generative Moment-Matching Networks. In *Advances in Neural Information Processing Systems*, pages 2928–2936, 2016.

R. A. Rigby and D. M. Stasinopoulos. Generalized Additive Models for Location, Scale and Shape,(with discussion). *Applied Statistics*, 54:507–554, 2005.

E. F. Rischel and S. Weichwald. Compositional Abstraction Error and a Category of Causal Models. In *Uncertainty in Artificial Intelligence*, pages 1013–1023. PMLR, 2021.

J. C. Robinson. *Dimensions, Embeddings, and Attractors*, volume 186. Cambridge University Press, 2010.

P. R. Rosenbaum. Conditional Permutation Tests and the Propensity Score in Observational Studies. *Journal of the American Statistical Association*, 79 (387):565–574, 1984.

P. R. Rosenbaum. *Observational Studies*. Springer Science & Business Media, 2002.

P. R. Rosenbaum and D. B. Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55, 1983.

P. Rubenstein, S. Weichwald, S. Bongers, J. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf. Causal Consistency of Structural Equation Models. In *33rd Conference on Uncertainty in Artificial Intelligence (UAI 2017)*, pages 808–817. Curran Associates, Inc., 2017.

P. K. Rubenstein, S. Bongers, B. Schölkopf, and J. M. Mooij. From Deterministic ODEs to Dynamic Structural Causal Models. In *34th Conference on Uncertainty in Artificial Intelligence (UAI 2018)*, 2018.

D. B. Rubin. Causal Inference using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

A. Rudi, L. Carratino, and L. Rosasco. Falkon: An Optimal Large Scale Kernel Method. In *Advances in Neural Information Processing Systems*, pages 3888–3898, 2017.

F. Russo. *Causality and Causal Modelling in the Social Sciences*. Springer, 2010.

K. Sadeghi and T. Soo. Axiomatization of Interventional Probability Distributions. *arXiv preprint arXiv:2305.04479*, 2023.

A. Saha and B. Palaniappan. Learning with Operator-Valued Kernels in Reproducing Kernel Krein Spaces. *Advances in Neural Information Processing Systems*, 33, 2020.

P. Schenone. Causality: A Decision Theoretic Foundation. *arXiv preprint arXiv:1812.07414*, 2018.

B. Schölkopf. Causality for Machine Learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 765–804. ACM, 2022.

B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT press, 2001.

B. Schölkopf and J. von Kügelgen. From Statistical to Causal Learning. In *Proceedings of the International Congress of Mathematicians*, page 1, 2022.

B. Schölkopf, K. Muandet, K. Fukumizu, S. Harmeling, and J. Peters. Computing Functions of Random Variables via Reproducing Kernel Hilbert Space Representations. *Statistics and Computing*, 25(4):755–766, 2015.

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.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, K. Fukumizu, et al. Equivalence of Distance-based and RKHS-based Statistics in Hypothesis Testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge university press, 2014.

U. Shalit, F. D. Johansson, and D. Sontag. Estimating Individual Treatment Effect: Generalization Bounds and Algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

S. Shen. Estimation and Inference of Distributional Partial Effects: Theory and Application. *Journal of Business & Economic Statistics*, 37(1):54–66, 2019.

T. Sheng and B. K. Sriperumbudur. On Distance and Kernel Measures of Conditional Independence. *arXiv preprint arXiv:1912.01103*, 2019.

C. Shi, D. Blei, and V. Veitch. Adapting Neural Networks for the Estimation of Treatment Effects. In *Advances in Neural Information Processing Systems*, pages 2507–2517, 2019.

G. R. Shorack and J. A. Wellner. *Empirical Processes with Applications to Statistics.* SIAM, 2009.

C.-J. Simon-Gabriel, A. Scibior, I. O. Tolstikhin, and B. Schölkopf. Consistent Kernel Mean Estimation for Functions of Random Variables. In *Advances in Neural Information Processing Systems*, pages 1732–1740, 2016.

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

R. Singh, L. Xu, and A. Gretton. Kernel Methods for Policy Evaluation: Treatment Effects, Mediation Analysis, and Off-Policy Planning. *arXiv preprint arXiv:2010.04855*, 2020.

S. Smale and D.-X. Zhou. Learning Theory Estimates via Integral Operators and Their Approximations. *Constructive approximation*, 26(2):153–172, 2007.

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.

L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert Space Embeddings of Conditional Distributions with Applications to Dynamical Systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968, 2009.

L. Song, B. Boots, S. M. Siddiqi, G. Gordon, and A. Smola. Hilbert Space Embeddings of Hidden Markov Models. In *Proceedings of the 27th on International Conference on Machine Learning*, pages 991–998, 2010a.

L. Song, A. Gretton, and C. Guestrin. Nonparametric Tree Graphical Models via Kernel Embeddings. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 765–772, 2010b.

L. Song, K. Fukumizu, and A. Gretton. Kernel Embeddings of Conditional Distributions: A Unified Kernel Framework for Nonparametric Inference in Graphical Models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, Prediction, and Search.* MIT press, 2000.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet. Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.

M. D. Stasinopoulos, R. A. Rigby, G. Z. Heller, V. Voudouris, and F. De Bastiani. *Flexible Regression and Smoothing: Using GAMLSS in R.* CRC Press, 2017.

I. Steinwart. On the Influence of the Kernel on the Consistency of Support Vector Machines. *Journal of machine learning research*, 2(Nov):67–93, 2001.

I. Steinwart and A. Christmann. *Support Vector Machines.* Springer Science & Business Media, 2008.

I. Steinwart, D. R. Hush, C. Scovel, et al. Optimal Rates for Regularized Least Squares Regression. In *COLT*, pages 79–93, 2009.

W. Stute. Conditional U-Statistics. *The Annals of Probability*, 19(2):812–825, 1991.

X. Su, C.-L. Tsai, H. Wang, D. M. Nickerson, and B. Li. Subgroup Analysis via Recursive Partitioning. *Journal of Machine Learning Research*, 10(2), 2009.

Z. Szabó and B. K. Sriperumbudur. Characteristic and Universal Tensor Product Kernels. *The Journal of Machine Learning Research*, 18(1):8724–8752, 2017.

Z. Szabó, B. K. Sriperumbudur, B. Póczos, and A. Gretton. Learning Theory for Distribution Regression. *The Journal of Machine Learning Research*, 17 (1):5272–5311, 2016.

S. A. van de Geer. *Empirical Processes in M-Estimation*, volume 6. Cambridge university press, 2000.

A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: with Applications to Statistics.* Springer Science & Business Media, 1996.

V. Vapnik. *The Nature of Statistical Learning Theory.* Springer science & business media, 1998.

J. Von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

J. Von Kügelgen, A. Mohamed, and S. Beckers. Backtracking Counterfactuals. In *Conference on Causal Learning and Reasoning*, pages 177–196. PMLR, 2023.

V. Vovk. Kernel Ridge Regression. In *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pages 105–116. Springer, 2013.

V. Vovk and G. Shafer. Game-Theoretic Probability. *Introduction to Imprecise Probabilities*, pages 114–134, 2014.

S. Wager and S. Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

M. Waldmann. *The Oxford Handbook of Causal Reasoning.* Oxford University Press, 2017.

P. Walley. *Statistical Reasoning with Imprecise Probabilities*, volume 42. Springer, 1991.

X. Wang, W. Pan, W. Hu, Y. Tian, and H. Zhang. Conditional Distance Correlation. *Journal of the American Statistical Association*, 110(512):1726–1734, 2015.

Y. Wang and M. I. Jordan. Desiderata for Representation Learning: A Causal Perspective. *arXiv preprint arXiv:2109.03795*, 2021.

L. Wasserman. *All of Nonparametric Statistics.* Springer Science & Business Media, 2006.

J. Weidmann. *Linear Operators in Hilbert Spaces*, volume 68. Springer New York, 1980.

H. White and K. Chalak. Settable Systems: An Extension of Pearl's Causal Model with Optimization, Equilibrium, and Learning. *Journal of Machine Learning Research*, 10(8), 2009.

J. Woodward. *Making Things Happen: A Theory of Causal Explanation.* Oxford university press, 2005.

L. Wu, A. Ledent, Y. Lei, and M. Kloft. Fine-Grained Generalization Analysis of Vector-Valued Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10338–10346, 2021.

K. Xia and E. Bareinboim. Neural Causal Abstractions. *arXiv preprint arXiv:2401.02602*, 2024.

D. Xu, Y. Shi, I. W. Tsang, Y.-S. Ong, C. Gong, and X. Shen. Survey on Multi-Output Learning. *IEEE transactions on neural networks and learning systems*, 31(7):2409–2429, 2019.

J. Yoon, J. Jordon, and M. van der Schaar. GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets. In *International Conference on Learning Representations*, 2018.

N. Yousefi, Y. Lei, M. Kloft, M. Mollaghasemi, and G. C. Anagnostopoulos. Local Rademacher Complexity-based Learning Guarantees for Multi-Task Learning. *The Journal of Machine Learning Research*, 19(1):1385–1431, 2018.

O. Zatarain-Vera. A Vector-Contraction Inequality for Rademacher Complexities Using p-Stable Variables. *arXiv preprint arXiv:1912.10136*, 2019.

M. Zečević, M. Willig, F. P. Busch, and J. Seng. Continual Causal Abstractions. In *AAAI Bridge Program on Continual Causality*, pages 45–51. PMLR, 2023.

F. M. Zennaro. Abstraction between Structural Causal Models: A Review of Definitions and Properties. In *UAI 2022 Workshop on Causal Representation Learning*, 2022.

F. M. Zennaro, M. Drávucz, G. Apachitei, W. D. Widanage, and T. Damoulas. Jointly Learning Consistent Causal Abstractions over Multiple Interventional Distributions. *arXiv preprint arXiv:2301.05893*, 2023.

J. Zhu and T. Hastie. Kernel Logistic Regression and the Import Vector Machine. *Journal of Computational and Graphical Statistics*, 14(1):185–205, 2005.

Appendix A

Proofs

A.1 Proofs for Chapter 1

Theorem 1.2.5. *From Definition 1.2.2, $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}$ is indeed a measure on $(\Omega, \mathcal{H})$, and $\mathbb{K}^{\mathrm{do}(U,\mathbb{Q},\mathbb{L})}$ is indeed a valid causal mechanism on $(\Omega, \mathcal{H}, \mathbb{P}^{\mathrm{do}(U,\mathbb{Q})})$, i.e. they satisfy the axioms of Definition 1.2.1.*

Proof of Theorem 1.2.5. That $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}$ is a measure on $(\Omega, \mathcal{H})$ follows immediately from the usual construction of measures from measures and transition probability kernels, see e.g. Çınlar (2011, p.38, Theorem 6.3). It remains to check that $\mathbb{K}^{\mathrm{do}(U,\mathbb{Q},\mathbb{L})}$ is a valid causal mechanism in the sense of Definition 1.2.1.

(i) For all $A \in \mathcal{H}$ and $\omega \in \Omega$,

$$\begin{aligned} K_{\emptyset}^{\mathrm{do}(U,\mathbb{Q},\mathbb{L})}(\omega, A) &= \int L_{\emptyset}(\omega_{\emptyset}, d\omega'_U) K_U((\omega_{\emptyset}, \omega'_U), A) \\ &= \int \mathbb{Q}(d\omega') K_U(\omega', A) \\ &= \mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(A), \end{aligned}$$

where we applied Axiom 1.2.1(i) to $L_{\emptyset}$.

(ii) For all $A \in \mathcal{H}_S$ and $B \in \mathcal{H}$, we have, by Axiom 1.2.1(ii) using the fact that $A \in \mathcal{H}_S \subseteq \mathcal{H}_{S\cup U}$,

$$\begin{aligned} & K_S^{\mathrm{do}(U,\mathbb{Q},\mathbb{L})}(\omega, A \cap B) \\ &= \int L_{S\cap U}(\omega_{S\cap U}, d\omega'_U) K_{S\cup U}((\omega_{S\setminus U}, \omega'_U), A \cap B) \\ &= \int L_{S\cap U}(\omega_{S\cap U}, d\omega'_U) 1_A((\omega_{S\setminus U}, \omega'_U)) K_{S\cup U}((\omega_{S\setminus U}, \omega'_U), B) \\ &= \int L_{S\cap U}(\omega_{S\cap U}, d\omega'_U) 1_A((\omega_{S\setminus U}, \omega'_{S\cap U})) K_{S\cup U}((\omega_{S\setminus U}, \omega'_U), B), \end{aligned}$$

where, in going from the third line to the fourth, we split the ω'_U in $1_A((\omega_{S\setminus U}, \omega'_U))$ into components $(\omega'_{S\cap U}, \omega'_{U\setminus S})$ and notice that since $A \in \mathcal{H}_S$, 1_A does not depend on the component $\omega'_{U\setminus S}$. Here, the map $\omega'_{S\cap U} \mapsto 1_A((\omega_{S\setminus U}, \omega'_{S\cap U}))$ is $\mathcal{H}_{S\cap U}$-measurable, so we can write it as the limit of an increasing sequence of positive $\mathcal{H}_{S\cap U}$-simple functions (see Section 1.1.1), say $(f_n)_{n\in\mathbb{N}}$ with $f_n = \sum_{i_n=1}^{m_n} b_{i_n} 1_{B_{i_n}}$, where $B_{i_n} \in \mathcal{H}_{S\cap U}$. Likewise, the map $\omega'_U \mapsto K_{S\cup U}((\omega_{S\setminus U}, \omega'_U), B)$ is $\mathcal{H}_U$-measurable, so we can write it as the limit of an increasing sequence of positive $\mathcal{H}_U$-simple functions, say $(g_n)_{n\in\mathbb{N}}$ with $g_n = \sum_{j_n=1}^{l_n} c_{j_n} 1_{C_{j_n}}$, where $C_{j_n} \in \mathcal{H}_U$. Hence

$$\begin{aligned}
&K_S^{\mathrm{do}(U,\mathbb{Q},\mathbb{L})}(\omega, A\cap B)\\
&\qquad = \int L_{S\cap U}(\omega_{S\cap U}, d\omega'_U)\left(\lim_{n\to\infty} f_n(\omega'_{S\cap U})\right)\left(\lim_{n\to\infty} g_n(\omega'_U)\right).
\end{aligned}$$

Since, for each ω'_U, both of the limits exist by construction, namely the original measurable functions, we have that the product of the limits is the limit of the products:

$$K_S^{\mathrm{do}(U,\mathbb{Q},\mathbb{L})}(\omega, A\cap B) = \int L_{S\cap U}(\omega_{S\cap U}, d\omega'_U)\lim_{n\to\infty}\left(f_n(\omega'_{S\cap U}) g_n(\omega'_U)\right).$$

Here, since f_n and g_n were individually sequences of increasing functions, the pointwise products $f_n g_n$ also form an increasing sequence of functions. Hence, we can apply the monotone convergence theorem to see that

$$\begin{aligned}
&K_S^{\mathrm{do}(U,\mathbb{Q},\mathbb{L})}(\omega, A\cap B)\\
&= \lim_{n\to\infty}\int L_{S\cap U}(\omega_{S\cap U}, d\omega'_U) f_n(\omega'_{S\cap U}) g_n(\omega'_U)\\
&= \lim_{n\to\infty}\sum_{i_n=1}^{m_n}\sum_{j_n=1}^{l_n} b_{i_n} c_{j_n}\int L_{S\cap U}(\omega_{S\cap U}, d\omega'_U) 1_{B_{i_n}}(\omega'_{S\cap U}) 1_{C_{j_n}}(\omega'_U)\\
&= \lim_{n\to\infty}\sum_{i_n=1}^{m_n}\sum_{j_n=1}^{l_n} b_{i_n} c_{j_n} L_{S\cap U}(\omega_{S\cap U}, B_{i_n}\cap C_{j_n})\\
&= \lim_{n\to\infty}\sum_{i_n=1}^{m_n}\sum_{j_n=1}^{l_n} b_{i_n} c_{j_n} 1_{B_{i_n}}(\omega_{S\cap U}) L_{S\cap U}(\omega_{S\cap U}, C_{j_n})\\
&= \lim_{n\to\infty}\sum_{i_n=1}^{m_n} b_{i_n} 1_{B_{i_n}}(\omega_{S\cap U})\sum_{j_n=1}^{l_n} c_{j_n} L_{S\cap U}(\omega_{S\cap U}, C_{j_n})\\
&= \left(\lim_{n\to\infty}\sum_{i_n=1}^{m_n} b_{i_n} 1_{B_{i_n}}(\omega_{S\cap U})\right)\left(\lim_{n\to\infty}\sum_{j_n=1}^{l_n} c_{j_n} L_{S\cap U}(\omega_{S\cap U}, C_{j_n})\right)\\
&= \left(\lim_{n\to\infty} f_n(\omega_{S\cap U})\right)\left(\lim_{n\to\infty}\int L_{S\cap U}(\omega_{S\cap U}, d\omega'_U)\sum_{j_n=1}^{l_n} c_j 1_{C_{j_n}}(\omega'_U)\right)
\end{aligned}$$

$$\begin{aligned}
&= 1_A((\omega_{S\setminus U},\omega_{S\cap U}))\int L_{S\cap U}(\omega_{S\cap U},d\omega'_U)\lim_{n\to\infty} g_n(\omega'_U)\\
&= 1_A(\omega_S)\int L_{S\cap U}(\omega_{S\cap U},d\omega'_U)K_{S\cup U}((\omega_{S\setminus U},\omega'_U),B)\\
&= 1_A(\omega_S)K_S^{\mathrm{do}(U,\mathbb{Q},\mathbb{L})}(\omega_S,B)
\end{aligned}$$

where, from the fourth line to the fifth, we used Axiom 1.2.1(ii); from the sixth line to the seventh, we used that limit of the products is the product of the limits again, noting that both of the limits exist by construction; from the eighth line to the ninth, we used monotone convergence theorem again. This is the required result.

□

Theorem 1.6.3. *Let* $(\Omega,\mathcal{H},\mathbb{P},\mathbb{K}) = (\times_{t\in T}E_t,\otimes_{t\in T}\mathcal{E}_t,\mathbb{P},\mathbb{K})$ *be a causal space, and* $U\in\mathcal{P}(T)$ *and* $\mathbb{Q}$ *a probability measure on* $(\Omega,\mathcal{H}_U)$*. Then after a hard intervention on* $\mathcal{H}_U$ *via* $\mathbb{Q}$*, the intervention causal kernels* $K_S^{\mathrm{do}(U,\mathbb{Q},\mathrm{hard})}$ *are given by*

$$\begin{aligned}
K_S^{\mathrm{do}(U,\mathbb{Q},\mathrm{hard})}(\omega,A) &= K_S^{\mathrm{do}(U,\mathbb{Q},\mathrm{hard})}(\omega_S,A)\\
&= \int \mathbb{Q}(d\omega'_{U\setminus S})K_{S\cup U}((\omega_S,\omega'_{U\setminus S}),A).
\end{aligned}$$

Proof of Theorem 1.6.3. We decompose $\mathcal{H}_U$ as a product σ-algebra into $\mathcal{H}_{S\cap U}\otimes\mathcal{H}_{U\setminus S}$. Then events of the form $B\cap C$ with $B\in\mathcal{H}_{S\cap U}$ and $C\in\mathcal{H}_{U\setminus S}$ generate $\mathcal{H}_U$, so for fixed $\omega_{S\cap U}$, the measure $L_{S\cap U}(\omega_{S\cap U},\cdot)$ is completely determined by $L_{S\cap U}(\omega_{S\cap U},B\cap C)$ for all $B\in\mathcal{H}_{S\cap U}$, $C\in\mathcal{H}_{U\setminus S}$. But we have

$$\begin{aligned}
L_{S\cap U}(\omega_{S\cap U},B\cap C) &= \delta_{\omega_{S\cap U}}(B)L_{S\cap U}(\omega_{S\cap U},C) \qquad \text{by Axiom 1.2.1(ii)}\\
&= \delta_{\omega_{S\cap U}}(B)\mathbb{Q}(C),
\end{aligned}$$

since $L_{S\cap U}$ is trivial and $C\in\mathcal{H}_{U\setminus S}$. So the measure $L_{S\cap U}(\omega_{S\cap U},\cdot)$ is a product measure of $\delta_{\omega_{S\cap U}}$ and $\mathbb{Q}$. Hence, applying Fubini's theorem,

$$\begin{aligned}
&K_S^{\mathrm{do}(U,\mathbb{Q},\mathrm{hard})}(\omega,A)\\
&= \int L_{S\cap U}(\omega_{S\cap U},d\omega'_U)K_{S\cup U}((\omega_{S\setminus U},\omega'_U),A)\\
&= \int\int K_{S\cup U}((\omega_{S\setminus U},\omega'_{S\cap U},\omega'_{U\setminus S}),A)\delta_{\omega_{S\cap U}}(d\omega'_{S\cap U})\mathbb{Q}(d\omega'_{U\setminus S})\\
&= \int K_{S\cup U}((\omega_{S\setminus U},\omega_{S\cap U},\omega'_{U\setminus S}),A)\mathbb{Q}(d\omega'_{U\setminus S})\\
&= \int \mathbb{Q}(d\omega'_{U\setminus S})K_{S\cup U}((\omega_S,\omega'_{U\setminus S}),A),
\end{aligned}$$

as required. □

Lemma 1.6.4. *Let $(\Omega,\mathcal{H},\mathbb{P},\mathbb{K})=(\times_{t\in T}E_t,\otimes_{t\in T}\mathcal{E}_t,\mathbb{P},\mathbb{K})$ be a causal space, and $U\in\mathcal{P}(T)$, $\mathbb{Q}$ a probability measure on $(\Omega,\mathcal{H}_U)$ and $\mathbb{L}=\{L_V:V\in\mathcal{P}(U)\}$ a causal mechanism on $(\Omega,\mathcal{H}_U,\mathbb{Q})$. Suppose we intervene on $\mathcal{H}_U$ via $(\mathbb{Q},\mathbb{L})$.*

(i) For $A\in\mathcal{H}_U$ and $V\in\mathcal{P}(T)$ with $V\cap U=\emptyset$, $\mathcal{H}_V$ has no causal effect on A in the intervention causal space $(\Omega,\mathcal{H},\mathbb{P}^{\text{do}(U,\mathbb{Q})},\mathbb{K}^{\text{do}(U,\mathbb{Q},\mathbb{L})})$, i.e. events in the σ-algebra $\mathcal{H}_U$ on which intervention took place are not causally affected by σ-algebras outside $\mathcal{H}_U$.

(ii) Again, let $V\in\mathcal{P}(T)$ with $V\cap U=\emptyset$, and also let $A\in\mathcal{H}$ be any event. If, in the original causal space, $\mathcal{H}_V$ had no causal effect on A, then in the intervention causal space, $\mathcal{H}_V$ has no causal effect on A either.

(iii) Now let $V\in\mathcal{P}(T)$, $A\in\mathcal{H}$ any event and suppose that the intervention on $\mathcal{H}_U$ via $\mathbb{Q}$ is hard. Then if $\mathcal{H}_V$ had no causal effect on A in the original causal space, then $\mathcal{H}_V$ has no causal effect on A in the intervention causal space either.

Proof of Lemma 1.6.4. (i) Take any $S\in\mathcal{P}(T)$. See that

$$\begin{aligned}
&K_S^{\text{do}(U,\mathbb{Q},\mathbb{L})}(\omega,A)\\
&=\int L_{S\cap U}(\omega_{S\cap U},d\omega'_U)K_{S\cup U}((\omega_{S\setminus U},\omega'_U),A)\\
&=\int L_{S\cap U}(\omega_{S\cap U},d\omega'_U)1_A(\omega'_U)\\
&=\int L_{S\cap U}(\omega_{S\cap U},d\omega'_U)K_{(S\setminus V)\cup U}((\omega_{(S\setminus V)\setminus U},\omega'_U),A)\\
&=\int L_{(S\setminus V)\cap U}(\omega_{(S\setminus V)\cap U},d\omega'_U)K_{(S\setminus V)\cup U}((\omega_{(S\setminus V)\setminus U},\omega'_U),A)\\
&=K_{S\setminus V}^{\text{do}(U,\mathbb{Q},\mathbb{L})}(\omega,A)
\end{aligned}$$

where, in going from the first line to the second and from the second line to the third, we used the fact that $A\in\mathcal{H}_U$, and in going from the third line to the fourth, we applied the fact that $(S\setminus V)\cap U=S\cap U$ since $V\cap U=\emptyset$. Since $S\in\mathcal{P}(T)$ was arbitrary, $\mathcal{H}_V$ has no causal effect on A in the intervention causal space.

(ii) Take any $S\in\mathcal{P}(T)$. See that

$$\begin{aligned}
&K_S^{\text{do}(U,\mathbb{Q},\mathbb{L})}(\omega,A)\\
&=\int L_{S\cap U}(\omega_{S\cap U},d\omega'_U)K_{S\cup U}((\omega_{S\setminus U},\omega'_U),A)\\
&=\int L_{S\cap U}(\omega_{S\cap U},d\omega'_U)K_{(S\cup U)\setminus V}((\omega_{(S\setminus V)\setminus U},\omega'_U),A)\\
&=\int L_{(S\setminus V)\cap U}(\omega_{(S\setminus V)\cap U},d\omega'_U)K_{(S\setminus V)\cup U}((\omega_{(S\setminus V)\setminus U},\omega'_U),A)
\end{aligned}$$

$$= K_{S\setminus V}^{\mathrm{do}(U,\mathbb{Q},\mathbb{L})}(\omega, A)$$

where, in going from the first line to the second, we used the fact that $\mathcal{H}_V$ has no causal effect on A in the original causal space, and in going from the second line to the third, we used $U \cap V = \emptyset$, which gives us $S \cap U = (S \setminus V) \cap U$ and $(S \cup U) \setminus V = (S \setminus V) \cup U$. Since $S \in \mathcal{P}(T)$ was arbitrary, $\mathcal{H}_V$ has no causal effect on A in the intervention causal space.

(iii) Take any $S \in \mathcal{P}(T)$. Apply Theorem 1.6.3 to see that

$$\begin{aligned}
&K_S^{\mathrm{do}(U,\mathbb{Q},\mathrm{hard})}(\omega, A)\\
&= \int \mathbb{Q}(d\omega'_{U\setminus S}) K_{S\cup U}((\omega_S, \omega'_{U\setminus S}), A)\\
&= \int \mathbb{Q}(d\omega'_{U\setminus S}) K_{(S\cup U)\setminus V}((\omega_S, \omega'_{U\setminus S}), A) && \text{Def. 1.3.1(i)}\\
&= \int \mathbb{Q}(d\omega'_{U\setminus S}) K_{((S\setminus V)\cup U)\setminus V}((\omega_S, \omega'_{U\setminus S}), A)\\
&= \int \mathbb{Q}(d\omega'_{U\setminus S}) K_{(S\setminus V)\cup U}((\omega_S, \omega'_{U\setminus S}), A) && \text{Def. 1.3.1(i)}\\
&= \int \mathbb{Q}(d\omega'_{U\setminus (S\setminus V)}) K_{(S\setminus V)\cup U}((\omega_{S\setminus V}, \omega'_{U\setminus (S\setminus V)}), A)\\
&= K_{S\setminus V}^{\mathrm{do}(U,\mathbb{Q})}(\omega, A),
\end{aligned}$$

where, in going from the second line to the third, we used that $(S\cup U)\setminus V = ((S \setminus V) \cup U) \setminus V$. Since $S \in \mathcal{P}(T)$ was arbitrary, $\mathcal{H}_V$ has no causal effect on A in the intervention causal space.

□

Lemma 1.6.5. *Let* $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t\in T} E_t, \otimes_{t\in T} \mathcal{E}_t, \mathbb{P}, \mathbb{K})$ *be a causal space, and* $U \in \mathcal{P}(T)$*. For an event* $A \in \mathcal{H}$*, if* $\mathcal{H}_U$ *has a dormant causal effect on* A *in the original causal space, then there exists a hard intervention and a subset* $V \subseteq U$ *such that in the intervention causal space,* $\mathcal{H}_V$ *has an active causal effect on* A*.*

Proof of Lemma 1.6.5. That $\mathcal{H}_U$ has a dormant causal effect on A tells us that $K_U(\omega, A) = \mathbb{P}(A)$ for all $\omega \in \Omega$, but there exists some $S \in \mathcal{P}(T)$ and some $\omega_0 \in \Omega$ such that $K_S(\omega_0, A) \neq K_{S\setminus U}(\omega_0, A)$. We must have $S \cap U \neq \emptyset$, since otherwise $S \setminus U = S$ and we cannot possibly have $K_S(\omega_0, A) \neq K_{S\setminus U}(\omega_0, A)$. Then we hard-intervene on $\mathcal{H}_{S\setminus U}$ with the Dirac measure on ω_0. Then apply Theorem 1.6.3 to see that

$$\begin{aligned}
K_{S\cap U}^{\mathrm{do}(S\setminus U,\delta_{\omega_0},\mathrm{hard})}((\omega_0)_{U\cap S}, A) &= \int \delta_{\omega_0}(d\omega'_{S\setminus U}) K_S(((\omega_0)_{U\cap S}, \omega'_{S\setminus U}), A)\\
&= K_S(\omega_0, A)\\
&\neq K_{S\setminus U}(\omega_0, A)
\end{aligned}$$

Note that the intervention measure on A is equal to $K_{S\setminus U}(\omega_0, A)$:

$$\mathbb{P}^{\mathrm{do}(S\setminus U,\delta_{\omega_0})}(A) = \int \delta_{\omega_0}(d\omega'_{S\setminus U})K_{S\setminus U}(\omega', A) = K_{S\setminus U}(\omega_0, A).$$

Putting these together, we have

$$K_{S\cap U}^{\mathrm{do}(S\setminus U,\delta_{\omega_0},\mathrm{hard})}(\omega_0, A) \neq \mathbb{P}^{\mathrm{do}(S\setminus U,\delta_{\omega_0})}(A),$$

i.e. in the intervention causal space $(\Omega, \mathcal{H}, \mathbb{P}^{\mathrm{do}(S\setminus U,\delta_{\omega_0})}, K_{S\cap U}^{\mathrm{do}(S\setminus U,\delta_{\omega_0},\mathrm{hard})})$, the σ-algebra $\mathcal{H}_{S\cap U}$ has an active causal effect on A. □

Lemma 1.6.6. *Let $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t\in T}E_t, \otimes_{t\in T}\mathcal{E}_t, \mathbb{P}, \mathbb{K})$ be a causal space, and $U, V \in \mathcal{P}(T)$. For an event $A \in \mathcal{H}$, suppose that $\mathcal{H}_U$ has no causal effect on A given $\mathcal{H}_V$ (see Definition 1.3.4). Then after an intervention on $\mathcal{H}_V$ via any $(\mathbb{Q}, \mathbb{L})$, $\mathcal{H}_{U\setminus V}$ has no causal effect on A.*

Proof of Lemma 1.6.6. Take any probability measure $\mathbb{Q}$ on $(\Omega, \mathcal{H}_V)$ and any causal mechanism $\mathbb{L}$ on $(\Omega, \mathcal{H}_V, \mathbb{Q})$. Then see that, for any $S \in \mathcal{P}(T)$ and all $\omega \in \Omega$,

$$\begin{aligned}
&K_S^{\mathrm{do}(V,\mathbb{Q},\mathbb{L})}(\omega, A)\\
&= \int L_{S\cap V}(\omega_{S\cap V}, d\omega'_V)K_{S\cup V}((\omega_{S\setminus V}, \omega'_V), A)\\
&= \int L_{S\cap V}(\omega_{S\cap V}, d\omega'_V)K_{(S\cup V)\setminus(U\setminus V)}((\omega_{S\setminus(U\cup V)}, \omega'_V), A)\\
&= \int L_{(S\setminus(U\setminus V))\cap V}(\omega_{(S\setminus(U\setminus V))\cap V}, d\omega'_V)K_{(S\setminus(U\setminus V))\cup V}((\omega_{S\setminus(U\cup V)}, \omega'_V), A)\\
&= K_{S\setminus(U\setminus V)}^{\mathrm{do}(V,\mathbb{Q},\mathbb{L})}(\omega, A),
\end{aligned}$$

where, in going from the first line to the second, we used the fact that $\mathcal{H}_U$ has no causal effect on A given $\mathcal{H}_V$, and in going from the second line to the third, we used identities $S\cap V = (S\setminus(U\setminus V))\cap V$ and $(S\cup V)\setminus(U\setminus V) = (S\setminus(U\setminus V))\cup V$. Since $S \in \mathcal{P}(T)$ was arbitrary, we have that $\mathcal{H}_{U\setminus V}$ has no causal effect on A in the intervention causal space. □

Theorem 1.6.7. *Let $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t\in T}E_t, \otimes_{t\in T}\mathcal{E}_t, \mathbb{P}, \mathbb{K})$ be a causal space, where the index set T can be written as $T = W \times \tilde{T}$, with W representing time and $\mathbb{K}$ respecting time. Take any $U \in \mathcal{P}(T)$ and any probability measure $\mathbb{Q}$ on $\mathcal{H}_U$. Then the intervention causal mechanism $\mathbb{K}^{\mathrm{do}(U,\mathbb{Q},\mathrm{hard})}$ also respects time.*

Proof of Theorem 1.6.7. Take any $w_1, w_2 \in W$ with $w_1 < w_2$. Since $\mathbb{K}$ respects time, we have that $\mathcal{H}_{w_2\times\tilde{T}}$ has no causal effect on $\mathcal{H}_{w_1\times\tilde{T}}$ in the original causal space. To show that $\mathcal{H}_{w_2\times\tilde{T}}$ has no causal effect on $\mathcal{H}_{w_1\times\tilde{T}}$ after a hard intervention on $\mathcal{H}_U$ via $\mathbb{Q}$, take any $S \in \mathcal{P}(T)$ and any event $A \in \mathcal{H}_{w_1\times\tilde{T}}$. Then using Theorem 1.6.3,

$$K_S^{\mathrm{do}(U,\mathbb{Q},\mathrm{hard})}(\omega, A)$$

$$
\begin{aligned}
&= \int \mathbb{Q}(d\omega')K_{S\cup U}((\omega_S,\omega'_{U\setminus S}),A)\\
&= \int \mathbb{Q}(d\omega')K_{(S\cup U)\setminus \mathcal{H}_{w_2\times\tilde{T}}}((\omega_{S\setminus\mathcal{H}_{w_2\times\tilde{T}}},\omega'_{U\setminus(S\cup\mathcal{H}_{w_2\times\tilde{T}})}),A)\\
&= \int \mathbb{Q}(d\omega')\\
&\quad K_{((S\cup U)\setminus\mathcal{H}_{w_2\times\tilde{T}})\cup(U\cap\mathcal{H}_{w_2\times\tilde{T}})}((\omega_{S\setminus\mathcal{H}_{w_2\times\tilde{T}}},\omega'_{(U\setminus(S\cup\mathcal{H}_{w_2\times\tilde{T}}))\cup(U\cap\mathcal{H}_{w_2\times\tilde{T}})}),A)\\
&= \int \mathbb{Q}(d\omega')K_{(S\setminus\mathcal{H}_{w_2\times\tilde{T}})\cup U}((\omega_{S\setminus\mathcal{H}_{w_2\times\tilde{T}}},\omega'_{U\setminus(S\setminus\mathcal{H}_{w_2\times\tilde{T}})}),A)\\
&= K^{\text{do}(U,\mathbb{Q},\text{hard})}_{S\setminus\mathcal{H}_{w_2\times\tilde{T}}}(\omega,A)
\end{aligned}
$$

where, from the second line to the third, we used the fact that $\mathcal{H}_{w_2\times\tilde{T}}$ has no causal effect on A, from the third line to the fourth we used the fact that $U\cap\mathcal{H}_{w_2\times\tilde{T}}$ has no causal effect on A (by Remark 1.3.2(e)) and Remark 1.3.2(g), and from the fourth line to the fifth, we used that $((S\cup U)\setminus\mathcal{H}_{w_2\times\tilde{T}})\cup(U\cap\mathcal{H}_{w_2\times\tilde{T}})=(S\setminus\mathcal{H}_{w_2\times\tilde{T}})\cup U$ and $(U\setminus(S\cup\mathcal{H}_{w_2\times\tilde{T}}))\cup(U\cap\mathcal{H}_{w_2\times\tilde{T}})=U\setminus(S\setminus\mathcal{H}_{w_2\times\tilde{T}})$. Since $S\in\mathcal{P}(T)$ was arbitrary, we have that $\mathcal{H}_{w_2\times\tilde{T}}$ has no causal effect on A (Definition 1.3.1(i)). Since $A\in\mathcal{H}_{w_1\times\tilde{T}}$ was arbitrary, $\mathcal{H}_{w_2\times\tilde{T}}$ has no causal effect on $\mathcal{H}_{w_1\times\tilde{T}}$, and so $\mathbb{K}^{\text{do}(U,\mathbb{Q},\text{hard})}$ respects time. □

Theorem 1.7.2. *Let $(\Omega,\mathcal{H},\mathbb{P},\mathbb{K})=(\times_{t\in T}E_t,\otimes_{t\in T}\mathcal{E}_t,\mathbb{P},\mathbb{K})$ be a causal space, and let $U\in\mathcal{P}(T)$.*

(i) For any measure $\mathbb{Q}$ on $\mathcal{H}_U$ and any causal mechanism $\mathbb{L}$ on $(\Omega,\mathcal{H}_U,\mathbb{Q})$, the causal kernel $K^{\text{do}(U,\mathbb{Q},\mathbb{L})}_U=K_U$ is a version of $\mathbb{P}^{\text{do}(U,\mathbb{Q})}_{\mathcal{H}_U}$, which means that $\mathcal{H}_U$ is a global source σ-algebra of $(\Omega,\mathcal{H},\mathbb{P}^{\text{do}(U,\mathbb{Q})},\mathbb{K}^{\text{do}(U,\mathbb{Q},\mathbb{L})})$.

(ii) Suppose $V\in\mathcal{P}(T)$ with $V\subseteq U$. Suppose that the measure $\mathbb{Q}$ on $(\Omega,\mathcal{H}_U)$ factorises over $\mathcal{H}_V$ and $\mathcal{H}_{U\setminus V}$, i.e. for any $A\in\mathcal{H}_V$ and $B\in\mathcal{H}_{U\setminus V}$, $\mathbb{Q}(A\cap B)=\mathbb{Q}(A)\mathbb{Q}(B)$. Then after a hard intervention on $\mathcal{H}_U$ via $\mathbb{Q}$, the causal kernel $K^{\text{do}(U,\mathbb{Q},\text{hard})}_V$ is a version of $\mathbb{P}^{\text{do}(U,\mathbb{Q})}_V$, which means that $\mathcal{H}_V$ is a global source σ-algebra of $(\Omega,\mathcal{H},\mathbb{P}^{\text{do}(U,\mathbb{Q})},\mathbb{K}^{\text{do}(U,\mathbb{Q},\text{hard})})$.

Proof of Theorem 1.7.2. Suppose that $f=\sum_{i=1}^m b_i 1_{B_i}$ is a $\mathcal{H}_U$-simple function, i.e. with $B_i\in\mathcal{H}_U$ for $i=1,...,m$. Then for any $B\in\mathcal{H}_U$,

$$
\begin{aligned}
\int_B f(\omega)\mathbb{P}^{\text{do}(U,\mathbb{Q})}(d\omega) &= \int_B \sum_{i=1}^m b_i 1_{B_i}(\omega)\mathbb{P}^{\text{do}(U,\mathbb{Q})}(d\omega)\\
&= \sum_{i=1}^m b_i \mathbb{P}^{\text{do}(U,\mathbb{Q})}(B\cap B_i)\\
&= \sum_{i=1}^m b_i \int \mathbb{Q}(d\omega)K_U(\omega,B\cap B_i)\\
&= \sum_{i=1}^m b_i \int \mathbb{Q}(d\omega)1_{B\cap B_i}(\omega) \qquad \text{by Axiom 1.2.1(ii)}
\end{aligned}
$$

$$
\begin{aligned}
&= \int_B \sum_{i=1}^m b_i 1_{B_i}(\omega)\mathbb{Q}(d\omega) \\
&= \int_B f(\omega)\mathbb{Q}(d\omega).
\end{aligned}
$$

Now, for any $\mathcal{H}_U$-measurable map $g : \Omega \to \mathbb{R}$, we can write it as a limit of an increasing sequence of positive $\mathcal{H}_U$-simple functions f_n (see Section 1.1.1), so for any $B \in \mathcal{H}_U$, using the monotone convergence theorem,

$$
\begin{aligned}
\int_B g(\omega)\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(d\omega) &= \int_B \lim_{n\to\infty} f_n(\omega)\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(d\omega) \\
&= \lim_{n\to\infty} \int_B f_n(\omega)\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(d\omega) \\
&= \lim_{n\to\infty} \int_B f_n(\omega)\mathbb{Q}(d\omega) \qquad \text{by above} \\
&= \int_B \lim_{n\to\infty} f_n(\omega)\mathbb{Q}(d\omega) \\
&= \int_B g(\omega)\mathbb{Q}(d\omega).
\end{aligned}
$$

We use this fact in the proof of both parts of this theorem.

(i) First note that we indeed have $K_U^{\mathrm{do}(U,\mathbb{Q},\mathbb{L})} = K_U$, by Remark 1.6.1(a). For any $A \in \mathcal{H}$, the map $\omega \mapsto K_U(\omega, A)$ is $\mathcal{H}_U$-measurable, so for any $B \in \mathcal{H}_U$,

$$
\begin{aligned}
\int_B K_U(\omega, A)\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(d\omega) &= \int_B K_U(\omega, A)\mathbb{Q}(d\omega) \qquad \text{by the above fact} \\
&= \int 1_B(\omega)K_U(\omega, A)\mathbb{Q}(d\omega) \\
&= \int K_U(\omega, A \cap B)\mathbb{Q}(d\omega) \\
&= \mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(A \cap B) \\
&= \int 1_{A\cap B}(\omega)\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(d\omega) \\
&= \int 1_B(\omega)1_A(\omega)\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(d\omega) \\
&= \int_B 1_A(\omega)\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(d\omega).
\end{aligned}
$$

So $K_U(\cdot, A) = K_U^{\mathrm{do}(U,\mathbb{Q},\mathbb{L})}(\cdot, A)$ is indeed a version of the conditional probability $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}_{\mathcal{H}_U}(A)$, which means that $\mathcal{H}_U$ is a global source of the intervened causal space $(\Omega, \mathcal{H}, \mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}, \mathbb{K}^{\mathrm{do}(U,\mathbb{Q},\mathbb{L})})$.

(ii) For any $A \in \mathcal{H}$, the map $\omega \mapsto K_V^{\mathrm{do}(U,\mathbb{Q})}(\omega, A)$ is $\mathcal{H}_V$-measurable and hence $\mathcal{H}_U$-measurable, so for any $B \in \mathcal{H}_V \subseteq \mathcal{H}_U$,

$$\begin{aligned}
&\int_B K_V^{\mathrm{do}(U,\mathbb{Q})}(\omega_V, A)\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(d\omega_V) \\
&= \int_B K_V^{\mathrm{do}(U,\mathbb{Q})}(\omega_V, A)\mathbb{Q}(d\omega_V) && \text{by above fact} \\
&= \int K_V^{\mathrm{do}(U,\mathbb{Q})}(\omega_V, A \cap B)\mathbb{Q}(d\omega_V) && \text{by Axiom 1.2.1(ii)} \\
&= \int \int \mathbb{Q}(d\omega'_{U\setminus V})K_U((\omega_V, \omega'_{U\setminus V}), A \cap B)\mathbb{Q}(d\omega_V) \\
&= \int K_U(\omega_U, A \cap B)\mathbb{Q}(d\omega_U) \\
&= \int_B 1_A(\omega)\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(d\omega).
\end{aligned}$$

where, in going from the third line to the fourth, we used Theorem 1.6.3, and to go from the fourth line to the fifth, we used the hypothesis that $\mathbb{Q}$ factorises over $\mathcal{H}_V$ and $\mathcal{H}_{U\setminus V}$, meaning $\mathbb{Q}(d\omega_{U\setminus V})\mathbb{Q}(d\omega_V) = \mathbb{Q}(d\omega_U)$. So $K_V^{\mathrm{do}(U,\mathbb{Q})}(\omega, A)$ is indeed a version of the conditional probability $\mathbb{P}_{\mathcal{H}_V}^{\mathrm{do}(U,\mathbb{Q})}(A)$, which means that $\mathcal{H}_V$ is a global source of the intervened causal space $(\Omega, \mathcal{H}, \mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}, \mathbb{K}^{\mathrm{do}(U,\mathbb{Q})})$.

□

Lemma 1.7.3. *Let $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t\in T} E_t, \otimes_{t\in T}\mathcal{E}_t, \mathbb{P}, \mathbb{K})$ be a causal space. Let $A \in \mathcal{H}$ be an event, and $U \in \mathcal{P}(T)$. If there exists a sub-σ-algebra $\mathcal{G}$ of $\mathcal{H}$ (not necessarily of the form $\mathcal{H}_V$ for some $V \in \mathcal{P}(T)$) such that*

(i) the conditional probability $\mathbb{P}_{\mathcal{H}_U\vee\mathcal{G}}^{\mathrm{do}(U,\mathbb{Q})}(\cdot, A)$ can be written in terms of $\mathbb{P}$ and $\mathbb{Q}$;

(ii) the causal kernel $K_U(\cdot, B)$ can be written in terms of $\mathbb{P}$ for all $B \in \mathcal{G}$;

then $\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(A)$ can be written in terms of $\mathbb{P}$ and $\mathbb{Q}$.

Proof of Lemma 1.7.3. By law of total expectations, for any $V \in \mathcal{P}(T)$, we have

$$\begin{aligned}
\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(A) &= \int \mathbb{P}_{\mathcal{H}_U\vee\mathcal{G}}^{\mathrm{do}(U,\mathbb{Q})}(\omega, A)\mathbb{P}^{\mathrm{do}(U,\mathbb{Q})}(d\omega) \\
&= \int \mathbb{P}_{\mathcal{H}_U\vee\mathcal{G}}^{\mathrm{do}(U,\mathbb{Q})}(\omega, A)\int \mathbb{Q}(d\omega')K_U(\omega', d\omega).
\end{aligned}$$

Here, $\mathbb{P}_{\mathcal{H}_U\vee\mathcal{G}}^{\mathrm{do}(U,\mathbb{Q})}(\omega, A)$ can be written in terms of $\mathbb{P}$ and $\mathbb{Q}$ by condition (i). Moreover, note that it suffices to be able to write the restriction of $K_U(\omega', \cdot)$ to $\mathcal{H}_U\vee\mathcal{G}$ in terms of $\mathbb{P}$, since the integration is of a $\mathcal{H}_U\vee\mathcal{G}$-measurable function. Since the collection of intersections $\{D\cap B, D \in \mathcal{H}_U, B \in \mathcal{G}\}$ is a π-system that generates

$\mathcal{H}_U \vee \mathcal{G}$ (Çınlar, 2011, p.5, 1.18), it suffices to check that $K_U(\omega', D \cap B)$ can be written in terms of $\mathbb{P}$ for all $D \in \mathcal{H}_U$ and $B \in \mathcal{G}$. But by interventional determinism (Definition 1.2.1(ii)), we have $K_U(\omega', D \cap B) = 1_D(\omega')K_U(\omega', B)$. Since $K_U(\omega', B)$ can be written in terms of $\mathbb{P}$ by condition (ii), the restriction of $K_U(\omega', \cdot)$ to $\mathcal{H}_U \vee \mathcal{G}$ can be written in terms of $\mathbb{P}$, and hence $\mathbb{P}^{\text{do}(U,\mathbb{Q})}(A)$ can be written in terms of $\mathbb{P}$ and $\mathbb{Q}$. □

Corollary 1.7.5. *Let $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K}) = (\times_{t\in T} E_t, \otimes_{t\in T} \mathcal{E}_t, \mathbb{P}, \mathbb{K})$ be a causal space. Let $A \in \mathcal{H}$ be an event, and $U \in \mathcal{P}(T)$. If there exists a $V \in \mathcal{P}(T)$ such that condition (i) of Lemma 1.7.3 is satisfied with $\mathcal{G} = \mathcal{H}_V$ and one of the following conditions is satisfied:*

(a) $\mathcal{H}_U$ is a local source of $\mathcal{H}_V$; or

(b) $\mathcal{H}_U$ has no causal effect on $\mathcal{H}_V$; or

(c) $V \subseteq U$,

then $\mathbb{P}^{\text{do}(U,\mathbb{Q})}(A)$ can be written in terms of $\mathbb{P}$ and $\mathbb{Q}$.

Proof of Corollary 1.7.5. Condition (i) of Lemma 1.7.3 is satisfied by hypothesis. If one of (a), (b) or (c) is satisfied, then trivially, condition (ii) of Lemma 1.7.3 is also satisfied. The result now follows from Lemma 1.7.3. □

A.2 Proofs for Chapter 2

Lemma 2.1.2. *The product causal space $\mathcal{C}^1 \otimes \mathcal{C}^2$ as defined in Definition 2.1.1 is a causal space.*

Proof of Lemma 2.1.2. It is a standard fact that $\mathbb{K}^1 \otimes \mathbb{K}^2$ defines a family of probability kernels[1]. For the first axiom of causal kernels (Definition 1.2.1(i)), we observe that

$$\begin{aligned}(K^1 \otimes K^2)_\emptyset((\omega_1, \omega_2), A_1 \times A_2) &= K^1_\emptyset(\omega_1, A_1)K^2_\emptyset(\omega_2, A_2)\\ &= \mathbb{P}^1(A_1)\mathbb{P}^2(A_2)\\ &= \mathbb{P}^1 \otimes \mathbb{P}^2(A_1 \times A_2).\end{aligned}$$

By standard reasoning based on the monotone class theorem, this extends to $A \in \mathcal{H}^1 \otimes \mathcal{H}^2$ and therefore the first axiom of causal spaces is satisfied.

For the second axiom of causal spaces, for any $S = S^1 \cup S^2$, first fix arbitrary $A_1 \in \mathcal{H}^1_{S^1}$ and $A_2 \in \mathcal{H}^1_{S^2}$. Then, for all $B_1 \in \mathcal{H}^1$ and $B_2 \in \mathcal{H}^2$, we find that for all $\omega = (\omega_1, \omega_2)$,

$$\begin{aligned}L_S(\omega, (A_1 \times A_2) \cap (B_1 \times B_2)) &= K^1_{S_1}(\omega_1, A_1 \cap B_1)K^2_{S^2}(\omega_2, A_2 \cap B_2)\\ &= \mathbf{1}_{A_1}(\omega_1)K^1_{S^1}(\omega_1, B_1)\mathbf{1}_{A_2}(\omega_2)K^2_{S^2}(\omega_2, B_2)\end{aligned}$$

[1] See, e.g. math.stackexchange.com/questions/84078/product-of-two-probability-kernel-is-a-probability-kernel

$$= \mathbf{1}_{A_1\times A_2}(\omega)L_S(\omega, B_1 \times B_2).$$

Hence, for this fixed pair A_1, A_2 and this ω, the measures $B \mapsto L_S(\omega,(A_1 \times A_2)\cap B)$ and $B \mapsto \mathbf{1}_{A_1\times A_2}(\omega)L_S(\omega, B)$ are identical on the generating rectangles $B_1 \times B_2$, hence they are identical on all of $\mathcal{H}^1 \otimes \mathcal{H}^2$ by the standard monotone class theorem reasoning. Now, since this is true for arbitrary rectangles $A_1 \times A_2$ with $A_1 \in \mathcal{H}^1_{S^1}$ and $A_2 \in \mathcal{H}^2_{S^2}$, if we now fix $B \in \mathcal{H}^1 \otimes \mathcal{H}^2$, we have that the two measures $A \mapsto L_S(\omega, A \cap B)$ and $A \mapsto \mathbf{1}_A(\omega)L_S(\omega, B)$ on $\mathcal{H}^1_{S^1} \otimes \mathcal{H}^2_{S^2}$ are identical on the generating rectangles $A_1 \times A_2$, hence they are identical on all of $\mathcal{H}^1_{S^1} \otimes \mathcal{H}^2_{S^2}$. Now both A and B are arbitrary elements of $\mathcal{H}^1 \otimes \mathcal{H}^2$ and $\mathcal{H}^1_{S^1} \otimes \mathcal{H}^2_{S^2}$ respectively. To conclude, we have, for all ω, $A \in \mathcal{H}^1 \otimes \mathcal{H}^2$ and $B \in \mathcal{H}^1_{S^1} \otimes \mathcal{H}^2_{S^2}$,

$$L_S(\omega, A \cap B) = \mathbf{1}_A(\omega)L_S(\omega, B),$$

confirming the second axiom of causal spaces. □

Lemma 2.1.3. *Let $\mathcal{C}^1 = (\Omega^1, \mathcal{H}^1, \mathbb{P}^1, \mathbb{K}^1)$ and $\mathcal{C}^2 = (\Omega^2, \mathcal{H}^2, \mathbb{P}^2, \mathbb{K}^2)$ with $\Omega^1 = \times_{t\in T^1}E_t$ and $\Omega^2 = \times_{t\in T^2}E_t$ be two causal spaces. Then in $\mathcal{C}^1 \otimes \mathcal{C}^2$,*

(i) $\mathcal{H}_{T^1}$ has no causal effect on $\mathcal{H}_{T^2}$, and $\mathcal{H}_{T^2}$ has no causal effect on $\mathcal{H}_{T^1}$;

(ii) $\mathcal{H}_{T^1}$ and $\mathcal{H}_{T^2}$ are (local) sources of each other.

Proof of Lemma 2.1.3. (i) Denote the causal kernels on the product space by K^p. Take any event $A \in \mathcal{H}_{T^2}$, and any $S \subseteq T^1 \cup T^2$. Note that S can be written as a union $S = S^1 \cup S^2$ for some $S^1 \subseteq T^1$ and $S^2 \subseteq T^2$. Then see that, by writing $A = \Omega^1 \times A' \in \mathcal{H}_{T^1} \otimes \mathcal{H}_{T^2}$ with $A' \subseteq \Omega^2$,

$$\begin{aligned} K^p_S(\omega, A) &= K^p_{S^1\cup S^2}(\omega, A) \\ &= K^1_{S^1}(\omega, \Omega_1)K^2_{S^2}(\omega, A') \\ &= K^1_\emptyset(\omega, \Omega_1)K_{S^2}(\omega, A') \\ &= K_{S\setminus T^1}(\omega, A). \end{aligned}$$

Here we used that $K^1_{S^1}(\omega, \Omega_1) = 1 = K^1_\emptyset(\omega, \Omega_1)$ because $K(\omega, \cdot)$ is a probability measure for a probability kernel.

So $\mathcal{H}_{T^1}$ has no causal effect on A. Implication in the other direction follows the same argument.

(ii) Take any $A \in \mathcal{H}_{T^2}$. By (i), $\mathcal{H}_{T^1}$ has no causal effect on A, so

$$K_{T^1}(\omega, A) = K_{T^1\setminus T^1}(\omega, A) = K_\emptyset(\omega, A) = \mathbb{P}(A).$$

But since $\mathcal{H}_{T^1}$ and $\mathcal{H}_{T^2}$ are probabilistically independent, $\mathbb{P}_{T^1}(A) = \mathbb{P}(A)$. Hence, $\mathbb{P}_{T^1}(A) = K_{T^1}(\omega, A)$, meaning $\mathcal{H}_{T_1}$ is a source of A. Since $A \in \mathcal{H}_{T_2}$ was arbitrary, $\mathcal{H}_{T_1}$ is a source of $\mathcal{H}_{T_2}$. The implication in the other direction follows the same argument.

□

Lemma 2.4.1. *Let $(\kappa_1, \rho_1) : \mathcal{C}^1 \to \mathcal{C}^2$ and $(\kappa_2, \rho_2) : \mathcal{C}^2 \to \mathcal{C}^3$ be causal transformations. If (κ_1, ρ_1) is an abstraction then $(\kappa_3, \rho_3) = (\kappa_1 \circ \kappa_2, \rho_1 \circ \rho_2) : \mathcal{C}^1 \to \mathcal{C}^3$ is a causal transformation.*

Proof of Lemma 2.4.1. First, we claim that the pair $(\kappa_3, \rho_3) = (\kappa_1 \circ \kappa_2, \rho_1 \circ \rho_2)$ is admissible. We have to show that, for any $S^3 \subset \rho_3(T^1)$ and $A \in \mathcal{H}^3_{S^3}$, the map $\kappa_3(\cdot, A)$ is measurable with respect to $\mathcal{H}^1_{\rho_3^{-1}(S^3)}$.

Let us call $\rho_2^{-1}(S^3) = S^2$. Note that, since $(\kappa_2, \rho_2) : \mathcal{C}^2 \to \mathcal{C}^3$ is a causal transformation, $\kappa_2(\cdot, A)$ is measurable with respect to $\mathcal{H}^2_{S^2}$. Since we assume that the first map is an abstraction, we find that $S^2 \subset \rho^1(T^1) = T^2$, and thus by Definition 2.2.1 that for $B \in \mathcal{H}^2_{S^2}$ the function $\kappa_1(\cdot, B)$ is measurable with respect to $\mathcal{H}^1_{\rho_3^{-1}(S^3)}$, where we used $\rho_3^{-1}(S^3) = \rho_1^{-1}(S^2)$. We now use the relation $\kappa_3(\omega, A) = \int \kappa_1(\omega, d\omega')\kappa_2(\omega', A)$. Since $\kappa_2(\cdot, A)$ is measurable with respect to $\mathcal{H}^2_{S^2}$, we conclude that we can approximate $\kappa_2(\cdot, A)$ by a simple function $\sum \alpha_i \mathbf{1}_{B_i}(\cdot)$ with $B_i \in \mathcal{H}^2_{S^2}$. But for such a simple function, we find

$$\int \kappa_1(\omega, d\omega') \sum_i \alpha_i \mathbf{1}_{B_i}(\omega') = \sum_i \alpha_i \kappa_1(\omega, B_i),$$

which is measurable with respect to $\mathcal{H}^1_{S^1}$ as a sum of measurable functions because (κ_1, ρ_1) is admissible. By passing to the limit (κ_3, ρ_3) is admissible.

Next we show that distributional consistency holds, which follows directly from distributional consistency of (κ_1, ρ_1) and (κ_2, ρ_2):

$$\begin{aligned}\int \mathbb{P}^1(d\omega)\kappa_3(\omega, A) &= \int \mathbb{P}^1(d\omega)\kappa_1(\omega, d\omega_2)\kappa_2(\omega_2, A)\\ &= \int \mathbb{P}^2(d\omega_2)\kappa_2(\omega_2, A)\\ &= \mathbb{P}^3(A).\end{aligned}$$

Next we consider interventional consistency. Let $S^3 \subset \rho_3(T^1)$ and define $S^2 = \rho_2^{-1}(S^3)$ and $S^1 = \rho_3^{-1}(S^1) = \rho_1^{-1}(S^2)$. Note that, since (κ_1, ρ_1) is an abstraction, i.e., ρ_1 is surjective, we have $S^2 \subset \rho_1(T^1) = T^2$. Now we find that, for $\omega_1 \in \Omega^1$ and $A \in \mathcal{H}^3$,

$$\begin{aligned}\int \kappa_3(\omega, d\omega')K^3_{S_3}(\omega', A) &= \int \kappa_1(\omega, d\omega_2)\kappa_2(\omega_2, d\omega')K^3_{S_3}(\omega', A)\\ &= \int \kappa_1(\omega, d\omega_2)K^2_{S_2}(\omega_2, d\omega')\kappa_2(\omega', A)\\ &= \int K^1_{S_1}(\omega, d\omega')\kappa_1(\omega', d\omega_2)\kappa_2(\omega_2, A)\\ &= \int K^1_{S_1}(\omega, d\omega')\kappa_3(\omega', A).\end{aligned}$$

This ends the proof as we have shown that (κ_3, ρ_3) is a causal transformation. □

Lemma 2.4.3. *Consider an acyclic SCM on variables $(X_1, \dots, X_d) \in \mathbb{R}^d$ with observational distribution $\mathbb{P}$. Let $S \subset [d]$, $R = S^c = [d] \setminus S$ and consider causal spaces $\mathcal{C}^1 = (\Omega^1, \mathcal{H}^1, \mathbb{P}^S, \mathbb{K})$ and $\mathcal{C}^2 = (\Omega^2, \mathcal{H}^2, \mathbb{P}, \mathbb{L})$, where we have $(\Omega^1, \mathcal{H}^1) = (\mathbb{R}^{|S|}, \mathcal{B}(\mathbb{R}^{|S|}))$ and $(\Omega^2, \mathcal{H}^2) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Moreover, $\mathbb{P}^S$ is the marginal distribution on the variables in S, and the causal mechanisms $\mathbb{K}$ and $\mathbb{L}$ are derived from the SCM. In particular, $\mathbb{K}$ is a marginalisation of $\mathbb{L}$, namely, for any $\omega \in \Omega^2$, any event $A \in \mathcal{H}^1$ and any $S' \subseteq S$, we have that $K_{S'}(\omega, A) = L_{S'}(\omega, A)$.*

Consider the map $\rho : S \hookrightarrow [d]$ and $\kappa(\cdot, A) = \mathbb{P}_{\mathcal{H}^1}(A)$. Then (ρ, κ) is a causal transformation from $\mathcal{C}^1$ to $\mathcal{C}^2$.

Proof of Lemma 2.4.3. First we note that as in Example 2.2.5 it is clear that (κ, ρ) is admissible and

$$\int \kappa(x_S, A) \mathbb{P}^S(dx_S) = \int \mathbb{P}_{\mathcal{H}^1}(A) d\mathbb{P}^S = \mathbb{P}(A),$$

so we have distributional consistency.

For interventional consistency, let $A \in \mathcal{H}^1$, $S' \subseteq S$ and $\omega \in \Omega^1$ be arbitrary. Then see that

$$\begin{aligned}\int K_{S'}(\omega, d\omega')\kappa(\omega', A) &= \int K_{S'}(\omega, d\omega')\mathbb{P}_{\mathcal{H}^1}(\omega', A) \\ &= \int K_{S'}(\omega, d\omega')\mathbf{1}_A(\omega') \qquad \text{since } A \in \mathcal{H}^1 \\ &= K_{S'}(\omega, A).\end{aligned}$$

On the other hand, see that, since $L_{S'}(\cdot, A)$ is measurable with respect to $\mathcal{H}^1$,

$$\begin{aligned}\int \kappa(\omega, d\omega')L_{S'}(\omega', A) &= \int \mathbb{P}_{\mathcal{H}^1}(\omega, d\omega')L_{S'}(\omega', A) \\ &= \int \mathbf{1}_{d\omega'}(\omega)L_{S'}(\omega', A) \\ &= L_{S'}(\omega, A).\end{aligned}$$

But by the marginalisation condition on the causal mechanisms $\mathbb{K}$ and $\mathbb{L}$, we have that $L_{S'}(\omega, A) = K_{S'}(\omega, A)$ for all $\omega \in \Omega^1$. This proves interventional consistency. □

Lemma 2.4.4. *Let $\mathcal{C}^2 = (\Omega^2, \mathcal{H}^2, \mathbb{P}^2, \mathbb{K}^2)$ and $\tilde{\mathcal{C}}^2 = (\Omega^2, \mathcal{H}^2, \tilde{\mathbb{P}}^2, \tilde{\mathbb{K}}^2)$ be two causal spaces with the same underlying measurable space.*

Let (κ, ρ) be an admissible pair for the measurable spaces $(\Omega^1, \mathcal{H}^1)$ and $(\Omega^2, \mathcal{H}^2)$. Assume that the pair (κ, ρ) defines causal transformations $\varphi : \mathcal{C}^1 \to \mathcal{C}^2$ and $\tilde{\varphi} : \mathcal{C}^1 \to \tilde{\mathcal{C}}^2$ be a causal transformations.

Then $\mathbb{P}^2 = \tilde{\mathbb{P}}^2$, and for all $A \in \mathcal{H}^2_{\rho(T^1)}$ and any $S \subseteq T^2$

$$K^2_S(\omega, A) = \tilde{K}^2_S(\omega, A) \qquad \textit{for } \mathbb{P}^2 = \tilde{\mathbb{P}}^2\textit{-a. e. } \omega \in \Omega^2.$$

Proof of Lemma 2.4.4. Applying distributional consistency of φ and $\tilde{\varphi}$, we find, for all $A \in \mathcal{H}^2$,

$$\mathbb{P}^2(A) = \int \mathbb{P}^1(d\omega)\kappa(\omega, A) = \tilde{\mathbb{P}}^2(A)$$

and thus $\mathbb{P}^2 = \tilde{\mathbb{P}}^2$.

Next, we consider $A \in \mathcal{H}^2_{\rho(T^1)}$ and $S \subset \rho(T^1)$. Let us define

$$B = \{\omega \in \Omega^2 : K^2_S(\omega, A) < \tilde{K}^2_S(\omega, A)\}.$$

Since $K^2_S(\cdot, A)$ and $\tilde{K}^2_S(\cdot, A)$ are $\mathcal{H}^2_S$ measurable, we find that $B \in \mathcal{H}^2_S \subset \mathcal{H}^2_{\rho(T^1)}$. Then the definition of causal spaces (see Definition 1.2.1) implies that

$$K^2_S(\omega', A \cap B) = \mathbf{1}_B(\omega')K^2_S(\omega', A).$$

Note that $A \cap B \in \mathcal{H}^2_{\rho(T^1)}$, so we can apply interventional consistency (2.2) for $\mathcal{C}^2$ and $\tilde{\mathcal{C}}^2$ and obtain, for any ω,

$$\begin{aligned}
\int \kappa(\omega, d\omega')\mathbf{1}_B(\omega')K^2_S(\omega', A) &= \int \kappa(\omega, d\omega')\mathbf{1}_B(\omega')K^2_S(\omega', A \cap B) \\
&= \int K^1_{\rho^{-1}(S)}(\omega, d\omega')\kappa(\omega', A) \\
&= \int \kappa(\omega, d\omega')\mathbf{1}_B(\omega')\tilde{K}^2_S(\omega', A \cap B) \\
&= \int \kappa(\omega, d\omega')\mathbf{1}_B(\omega')\tilde{K}^2_S(\omega', A).
\end{aligned}$$

We integrate this relation with respect to $\mathbb{P}^1(d\omega)$ and then apply distributional consistency to get

$$\begin{aligned}
0 &= \int \mathbb{P}^1(d\omega)\kappa(\omega, d\omega')\mathbf{1}_B(\omega')(\tilde{K}^2_S(\omega', A) - K^2_S(\omega', A)) \\
&= \int \mathbb{P}^2(d\omega')\mathbf{1}_B(\omega')(\tilde{K}^2_S(\omega', A) - K^2_S(\omega', A)) \\
&= \int_B \mathbb{P}^2(d\omega')(\tilde{K}^2_S(\omega', A) - K^2_S(\omega', A)).
\end{aligned}$$

On B, the integrand is strictly positive by definition. Thus we conclude that $\mathbb{P}^2(B) = 0$ and thus $\tilde{K}^2_S(\omega', A) \le K^2_S(\omega', A)$ holds almost surely.

The same reasoning implies the reverse bound, and we conclude that, $\mathbb{P}^2$-almost surely, the relation

$$\tilde{K}^2_S(\omega', A) = K^2_S(\omega', A)$$

holds. □

Lemma 2.4.5. *Suppose (f, ρ) is an admissible pair for the causal space $\mathcal{C}^1 = (\Omega^1, \mathcal{H}^1, \mathbb{P}^1, \mathbb{K}^1)$ to the measurable space $X^2 = (\Omega^2, \mathcal{H}^2)$ and assume that ρ is* *surjective and $f : \Omega_1 \to \Omega_2$ measurable. If f is surjective, there exists at most one causal space $\mathcal{C}^2 = (\Omega^2, \mathcal{H}^2, \mathbb{P}^2, \mathbb{K}^2)$ such that $(f, \rho) : \mathcal{C}^1 \to \mathcal{C}^2$ is a causal transformation.*

If, in addition, $K^1_{\rho^{-1}(S^2)}(\cdot, A)$ is measurable with respect to $f^{-1}(\mathcal{H}^2_{S^2})$ for all $A \in f^{-1}(\mathcal{H}^2)$ and all $S^2 \subset T^2$ then a unique causal space $\mathcal{C}^2$ exists such that $(f, \rho) : \mathcal{C}^1 \to \mathcal{C}^2$ is a causal transformation.

Proof of Lemma 2.4.5. We first prove uniqueness. The relation $f_*\mathbb{P}^1 = \mathbb{P}^2$ that are necessarily true for deterministic maps (see Section 2.2) implies that $\mathbb{P}^2$ is predetermined. Moreover, we find that, by (2.3), for any $A \in \mathcal{H}^2$, $S \subset T^2$ and any $\omega \in \Omega^1$,

$$K^1_{\rho^{-1}(S)}(\omega, f^{-1}(A)) = K^2_S(f(\omega), A).$$

But since f is surjective we conclude that due to interventional consistency $K^2_S(\omega', A)$ for $\omega' \in \Omega^2$ is unique.

To prove the existence we note that by assumption for fixed $A \in \mathcal{H}^2$ the function $K^1_{\rho^{-1}(S^2)}(\cdot, f^{-1}(A))$ is measurable with respect to $f^{-1}(\mathcal{H}^2_{S^2})$. Now by the Factorisation Lemma (, p.76, Theorem II.4.4) there is a measurable function $g : (\Omega^2, \mathcal{H}^2_{S^2}) \to \mathbb{R}$ such that

$$K^1_{\rho^{-1}(S^2)}(\omega, f^{-1}(A)) = g \circ f(\omega).$$

We define $K^2_{S^2}(\omega', A) = g(\omega')$. By surjectivity this defines $K^2_{S^2}$ everywhere and this defines a probability kernel because g is measurable.

It remains to verify that the resulting $\mathcal{C}^2$ is indeed a causal space. Using interventional and distributional consistency we obtain

$$\begin{aligned} K^2_\emptyset(f(\omega), A) &= K^1_\emptyset(\omega, f^{-1}(A)) \\ &= \mathbb{P}^1(f^{-1}(A)) \\ &= f_*\mathbb{P}^1(A) \\ &= \mathbb{P}^2(A). \end{aligned}$$

This verifies the first property of causal spaces. For the second property we observe that, for $A \in \mathcal{H}^2_{S^2}$ and $S^1 = \pi^{-1}(S^2)$, using causal consistency,

$$\begin{aligned} K^2_{S^2}(f(\omega), A \cap B) &= K^1_{S^1}(\omega, f^{-1}(A \cap B)) \\ &= K^1_{S^1}(\omega, f^{-1}(A) \cap f^{-1}(B)) \\ &= \mathbf{1}_{f^{-1}(A)}(\omega) K^1_{S^1}(\omega, f^{-1}(B)) \\ &= \mathbf{1}_A(f(\omega)) K^2_{S^2}(f(\omega), B). \end{aligned}$$

Here we used that $\mathcal{C}^1$ is a causal space and $f^{-1}(A) \in \mathcal{H}^1_{S^1}$. Thus, we conclude that we obtained a causal space $\mathcal{C}^2$. □

Lemma 2.4.6. *Let $\mathcal{C}^1 = (\Omega^1, \mathcal{H}^1, \mathbb{P}^1, \mathbb{K}^1)$ with $(\Omega^1, \mathcal{H}^1)$ a product with index set T^1 and $\mathcal{C}^2 = (\Omega^2, \mathcal{H}^2, \mathbb{P}^2, \mathbb{K}^2)$ with $(\Omega^2, \mathcal{H}^2)$ a product with index set T^2 be causal spaces, and let $(f, \rho) : \mathcal{C}^1 \to \mathcal{C}^2$ be a perfect abstraction.*

Let $U^1 = \rho^{-1}(U^2) \subseteq T^1$ for some $U^2 \subseteq T^2$. Let $\mathbb{Q}^1$ be a probability measure on $(\Omega^1, \mathcal{H}^1_{U^1})$ and $\mathbb{L}^1$ a causal mechanism on $(\Omega^1, \mathcal{H}^1_{U^1}, \mathbb{Q}^1)$. Suppose that, for all $S \subseteq U^2$ and $A \in \mathcal{H}^1$, the map $L^1_{\rho^{-1}(S)}(\cdot, A)$ is measurable with respect to $f^{-1}(\mathcal{H}^2_S)$, and consider the intervened causal spaces

$$\mathcal{C}^1_I = (\Omega^1, \mathcal{H}^1, (\mathbb{P}^1)^{\mathrm{do}(U^1,\mathbb{Q}^1)}, (\mathbb{K}^1)^{\mathrm{do}(U^1,\mathbb{Q}^1,\mathbb{L}^1)}),$$
$$\mathcal{C}^2_I = (\Omega^2, \mathcal{H}^2, (\mathbb{P}^2)^{\mathrm{do}(U^2,\mathbb{Q}^2)}, (\mathbb{K}^2)^{\mathrm{do}(U^2,\mathbb{Q}^2,\mathbb{L}^2)}),$$

where $\mathbb{Q}^2 = f_*\mathbb{Q}^1$ and $\mathbb{L}^2$ is the unique family of kernels satisfying

$$L^2_S(f(\omega), A) = L^1_{\rho^{-1}(S)}(\omega, f^{-1}(A))$$

for all $\omega \in \Omega^1$, $A \in \mathcal{H}^2$, and $S \subseteq U^2$. Then $(f, \rho) : \mathcal{C}^1_I \to \mathcal{C}^2_I$ is a perfect abstraction.

Proof of Lemma 2.4.6. First, we note that by Lemma 2.4.5 $\mathbb{L}^2$ exists and is unique. Thus, we need to verify distributional consistency and interventional consistency.

Let us first show $f_*(\mathbb{P}^1)^{\mathrm{do}(U^1,\mathbb{Q}^1)} = (\mathbb{P}^2)^{\mathrm{do}(U^2,\mathbb{Q}^2)}$. Since (f, ρ) is a causal transformation (i.e., interventional consistency as in (2.2) holds), we find that, for $A \in \mathcal{H}^2$,

$$\begin{aligned} f_*(\mathbb{P}^1)^{\mathrm{do}(U^1,\mathbb{Q}^1)}(A) &= \int \mathbb{Q}^1(d\omega) K^1_{U^1}(\omega, f^{-1}(A)) \\ &= \int \mathbb{Q}^1(d\omega) K^2_{U^2}(f(\omega), A) \\ &= \int (f_*\mathbb{Q}^1)(d\omega') K^2_{U^2}(\omega', A) \\ &= \int \mathbb{Q}^2(d\omega') K^2_{U^2}(\omega', A) \\ &= (\mathbb{P}^2)^{\mathrm{do}(U^2,\mathbb{Q}^2)}(A). \end{aligned}$$

Here we used the change of variable for pushforward-measures.

Next, we show interventional consistency of $(f, \rho) : \mathcal{C}^1_I \to \mathcal{C}^2_I$. For this, we introduce the shorthand $f_S = \pi_S \circ f$. Note that since f_S is measurable with respect to $\mathcal{H}^1_{\rho^{-1}(S)}$ we can find $\tilde{f}_S$ such that $f_S(\omega) = \tilde{f}_S(\omega_{\rho^{-1}(S)})$. Note that, by the interventional consistency of $(f, \rho) : \mathcal{C}^1 \to \mathcal{C}^2$, we have

$$K^1_{\rho^{-1}(S)}(\omega, f^{-1}(S)) = K^2_S(f(\omega), A) = K^2_S(\tilde{f}_S(\omega_S), A).$$

We can now show for $A \in \mathcal{H}^2$ and $S^1 = \rho^{-1}(S^2)$ that

$$\begin{aligned} &(K^1)^{\mathrm{do}(U^1,\mathbb{Q}^1,\mathbb{L}^1)}_{S^1}(\omega, f^{-1}(A)) \\ &= \int L^1_{S^1\cap U^1}(\omega_{S^1\cap U^1}, d\omega'_{U^1}) K^1_{S^1\cup U^1}((\omega_{S^1\setminus U^1}, \omega'_{U^1}), f^{-1}(A)) \end{aligned}$$

$$
\begin{aligned}
&= \int L^1_{S^1\cap U^1}(\omega_{S^1\cap U^1}, d\omega'_{U^1}) K^2_{S^2\cup U^2}(\tilde{f}_{S^2\setminus U^2}(\omega_{S^1\setminus U^1}), \tilde{f}_{U^2}(\omega'_{U^1}), A) \\
&= \int \left((\tilde{f}_{U^1})_*(L^1_{S^1\cap U^1}(\omega_{S^1\cap U^1}, \cdot)\right)(d\overline{\omega}_{U^2}) K^2_{S^2\cup U^2}(\tilde{f}_{S^2\setminus U^2}(\omega_{S^1\setminus U^1}), \overline{\omega}_{U^2}, A) \\
&= \int L^2_{S^2\cap U^2}(f(\omega)_{S^2\cap U^2}), d\overline{\omega}_{U^2}) K^2_{S^2\cup U^2}(f(\omega)_{S^2\setminus U^2}), \overline{\omega}_{U^2}, A) \\
&= (K^2)^{\mathrm{do}(U^2,\mathbb{Q}^2,\mathbb{L}^2)}_{S^2}(f(\omega), A).
\end{aligned}
$$

This ends the proof. □

Lemma 2.4.7. *Let $(f,\rho) : \mathcal{C}^1 \to \mathcal{C}^2$ be a perfect abstraction. Consider two sets $U^2, V^2 \subset T^2$ and denote $U^1 = \rho^{-1}(U^2)$ and $V^1 = \rho^{-1}(V^2)$. If $\mathcal{H}^1_{U^1}$ has no causal effect on $\mathcal{H}^1_{V^1}$ in $\mathcal{C}^1$, then $\mathcal{H}^2_{U^2}$ has no causal effect on $\mathcal{H}^2_{V^2}$ in $\mathcal{C}^2$.*

Proof of Lemma 2.4.7. Consider $A \in \mathcal{H}^2_{V^2}$ and any $S^2 \subset T^2$. Then for any $\omega' \in \Omega^2$ we find an $\omega \in \Omega^1$ such that $f(\omega) = \omega'$. Using interventional consistency and $f^{-1}(A) \in \mathcal{H}^1_{V^1}$ we conclude

$$
\begin{aligned}
K^2_{S^2}(\omega', A) &= K^1_{\rho^{-1}(S^2)}(\omega, f^{-1}(A)) \\
&= K^1_{\rho^{-1}(S^2)\setminus\rho^{-1}(U^2)}(\omega, f^{-1}(A)) \\
&= K^1_{\rho^{-1}(S^2\setminus U^2)}(\omega, f^{-1}(A)) \\
&= K^2_{S^2\setminus U^2}(\omega', A).
\end{aligned}
$$

This ends the proof. □

Lemma 2.4.8. *Let $(f,\rho) : \mathcal{C}^1 \to \mathcal{C}^2$ be a perfect abstraction. Consider two sets $U^2, V^2 \subset T^2$ and denote $U^1 = \rho^{-1}(U^2)$ and $V^1 = \rho^{-1}(V^2)$. Assume that $\mathcal{H}^2_{U^2}$ has an active causal effect on $\mathcal{H}^2_{V^2}$ in $\mathcal{C}^2$. Then $\mathcal{H}^1_{U^1}$ has an active causal effect on $\mathcal{H}^1_{V^1}$ in $\mathcal{C}^1$.*

Proof of Lemma 2.4.8. Since $\mathcal{H}^2_{U^2}$ has an active causal effect on $\mathcal{H}^2_{V^2}$ in $\mathcal{C}^2$, we find that there is an $\omega' \in \Omega^2$ and an $A \in \mathcal{H}^2_{V^1}$ such that

$$
K^2_{U^2}(\omega', A) \neq \mathbb{P}^2(A).
$$

By surjectivity there is $\omega \in \Omega^1$ such that $\omega' = f(\omega)$ and thus

$$
\begin{aligned}
K^1_{U^1}(\omega, f^{-1}(A)) &= K^2_{U^2}(\omega', A) \\
&\neq \mathbb{P}^2(A) \\
&= \mathbb{P}^1(f^{-1}(A)).
\end{aligned}
$$

The claim follows because $f^{-1}(A) \in \mathcal{H}^1_{U^1}$. □

Lemma 2.4.10. *Let $(f,\rho) : \mathcal{C}^1 \to \mathcal{C}^2$ be a perfect abstraction. Consider two sets $U^2, V^2 \subset T^2$ and denote $U^1 = \rho^{-1}(U^2)$ and $V^1 = \rho^{-1}(V^2)$. Assume that $\mathcal{H}^1_{U^1}$ is a local source of $\mathcal{H}^1_{V^1}$ in $\mathcal{C}^1$. Then $\mathcal{H}^2_{U^2}$ is a local source of $\mathcal{H}^2_{V^2}$ in $\mathcal{C}^2$.*

In particular, this implies that if $\mathcal{H}^1_{U^1}$ is a global source then $\mathcal{H}^2_{U^2}$ also is a global source.

Proof of Theorem 2.4.10. Our goal is to show that $K^2_{U^2}(\cdot, A)$ is a version of the conditional probability $\mathbb{P}^2_{\mathcal{H}^2_{U^2}}(A)$ for $A \in \mathcal{H}^2_{V^2}$. It is sufficient to show that for all $B \in \mathcal{H}^2_{U^2}$ the following relation holds

$$\int \mathbb{P}^2(d\omega')\mathbf{1}_A(\omega')\mathbf{1}_B(\omega') = \int \mathbb{P}^2(d\omega')\mathbf{1}_B(\omega')K^2_{U^2}(\omega', A). \tag{A.1}$$

Using that (f, ρ) is a perfect abstraction, $f^{-1}(A) \in \mathcal{H}^1_{V^1}$, $f^{-1}(B) \in \mathcal{H}^1_{U^1}$, and that $\mathcal{H}^1_{U^1}$ is a local source of $\mathcal{H}^1_{V^1}$ we find

$$\begin{aligned}
\int \mathbb{P}^2(d\omega')\mathbf{1}_A(\omega')\mathbf{1}_B(\omega') &= \int f_*\mathbb{P}^1(d\omega')\mathbf{1}_A(\omega')\mathbf{1}_B(\omega') \\
&= \int \mathbb{P}^1(d\omega)\mathbf{1}_A(f(\omega))\mathbf{1}_B(f(\omega)) \\
&= \int \mathbb{P}^1(d\omega)\mathbf{1}_{f^{-1}(A)}(\omega)\mathbf{1}_{f^{-1}(B)}(\omega) \\
&= \int \mathbb{P}^1(d\omega)K^1_{U^1}(\omega, f^{-1}(A))\mathbf{1}_{f^{-1}(B)}(\omega) \\
&= \int \mathbb{P}^1(d\omega)K^2_{U^2}(f(\omega), A)\mathbf{1}_B(f(\omega)) \\
&= \int f_*\mathbb{P}^1(d\omega')K^2_{U^2}(\omega', A)\mathbf{1}_B(\omega') \\
&= \int \mathbb{P}^2(d\omega')K^2_{U^2}(\omega', A)\mathbf{1}_B(\omega').
\end{aligned}$$

Thus we have shown that (A.1) holds and the proof is completed. □

A.3 Proofs for Chapter 3

Before we prove Theorem 3.1.6, we state the following definition and theorems related to measurable functions for Banach-space valued functions.

Definition A.3.1 ((Dinculeanu, 2000, p.4, Definition 5))**.** A function $H : \Omega \to \mathcal{H}$ is called an $\mathcal{F}$-simple function if it has the form $H = \sum_{i=1}^n h_i \mathbf{1}_{B_i}$ for some $h_i \in \mathcal{H}$ and $B_i \in \mathcal{F}$.

A function $H : \Omega \to \mathcal{H}$ is said to be $\mathcal{F}$-measurable if there is a sequence (H_n) of $\mathcal{H}$-valued, $\mathcal{F}$-simple functions such that $H_n \to H$ pointwise.

Theorem A.3.2 ((Dinculeanu, 2000, p.4, Theorem 6))**.** *If $H : \Omega \to \mathcal{H}$ is $\mathcal{F}$-measurable, then there is a sequence (H_n) of $\mathcal{H}$-valued, $\mathcal{F}$-simple functions such that $H_n \to H$ pointwise and $|H_n| \leq |H|$ for every n.*

Theorem A.3.3 ((Dinculeanu, 2000, p.19, Theorem 48), Lebesgue Convergence Theorem)**.** *Let (H_n) be a sequence in $L^1_{\mathcal{H}}(P)$, $H : \Omega \to \mathcal{H}$ a P-measurable function, and $g \in L^1_+(P)$ such that $H_n \to H$ P-almost everywhere and $|H_n| \leq g$, P-almost everywhere, for each n. Then $H \in L^1_{\mathcal{H}}(P)$ and $H_n \to H$ in $L^1_{\mathcal{H}}(P)$, i.e. $\int_\Omega H_n dP \to \int_\Omega H dP$.*

Theorem 3.1.6 (Adapted from (Çınlar, 2011, p.150, Proposition 2.5))**.** *Suppose that $P(\cdot \mid \mathcal{E})$ admits a regular version Q. Then $QH : \Omega \to \mathcal{H}$ with $\omega \mapsto Q_\omega H = \int_\Omega H(\omega')Q_\omega(d\omega')$ is a version of $\mathbb{E}[H \mid \mathcal{E}]$ for every Bochner P-integrable H.*

Proof of Theorem 3.1.6. Suppose H is Bochner P-integrable. Since Q is a regular version of $P(\cdot \mid \mathcal{E})$, it is a probability transition kernel from $(\Omega, \mathcal{E})$ to $(\Omega, \mathcal{F})$.

We first show that QH is measurable with respect to $\mathcal{E}$. The map $Q : \Omega \to \mathcal{H}$ is well-defined, since, for each $\omega \in \Omega$, $Q_\omega H$ is the Bochner-integral of H with respect to the measure $B \to Q_\omega(B)$. Since H is $\mathcal{F}$-measurable, by Theorem A.3.2, there is a sequence (H_n) of $\mathcal{H}$-valued, $\mathcal{F}$-simple functions such that $H_n \to H$ pointwise. Then for each $\omega \in \Omega$, $Q_\omega H = \lim_{n\to\infty} Q_\omega H_n$ by Theorem A.3.3. But for each n, we can write $H_n = \sum_{j=1}^m h_j \mathbf{1}_{B_j}$ for some $h_j \in \mathcal{H}$ and $B_j \in \mathcal{F}$, and so $Q_\omega H_n = \sum_{j=1}^m h_j Q_\omega(B_j)$. For each B_j the map $\omega \mapsto Q_\omega(B_j)$ is $\mathcal{E}$-measurable (by the definition of transition probability kernel), and so as a linear combination of $\mathcal{E}$-measurable functions, QH_n is $\mathcal{E}$-measurable. Hence, as a pointwise limit of $\mathcal{E}$-measurable functions, QH is also $\mathcal{E}$-measurable, by (Dinculeanu, 2000, p.6, Theorem 10).

Next, we show that, for all $A \in \mathcal{E}$, $\int_A HdP = \int_A QHdP$. Fix $A \in \mathcal{E}$. By Theorem A.3.2, there is a sequence (H_n) of $\mathcal{H}$-valued, $\mathcal{F}$-simple functions such that $H_n \to H$ pointwise. For each n, we can write $H_n = \sum_{j=1}^m h_j \mathbf{1}_{B_j}$ for some $h_j \in \mathcal{H}$ and $B_j \in \mathcal{F}$, and

$$\begin{aligned}
\int_A QH_n dP &= \int_A \sum_{j=1}^m h_j Q(B_j) dP \\
&= \int_A \sum_{j=1}^m h_j P(B_j \mid \mathcal{E}) dP \qquad \text{since } Q \text{ is a version of } P(\cdot \mid \mathcal{E}) \\
&= \sum_{j=1}^m h_j \int_A \mathbb{E}[\mathbf{1}_{B_j} \mid \mathcal{E}] dP \\
&= \int_A \sum_{j=1}^m h_j \mathbf{1}_{B_j} dP \qquad \text{since } A \in \mathcal{E} \\
&= \int_A H_n dP.
\end{aligned}$$

We have $H_n \to H$ pointwise by assertion, and as before, $QH_n \to QH$ pointwise. Hence,

$$\begin{aligned}
\int_A QHdP &= \lim_{n\to\infty} \int_A QH_n dP \qquad \text{by Theorem A.3.3} \\
&= \lim_{n\to\infty} \int_A H_n dP \qquad \text{by above} \\
&= \int_A HdP \qquad \text{by Theorem A.3.3.}
\end{aligned}$$

Hence, by the definition of the conditional expectation, QH is a version of $\mathbb{E}[H \mid \mathcal{E}]$. □

Theorem 3.1.11 (Generalised Conditional Jensen's Inequality)**.** *Suppose $\mathcal{T}$ is a real Hausdorff locally convex (possibly infinite-dimensional) linear topological space, and let C be a closed convex subset of $\mathcal{T}$. Suppose $(\Omega, \mathcal{F}, P)$ is a probability space, and $V : \Omega \to \mathcal{T}$ a Pettis-integrable random variable such that $V(\Omega) \subseteq C$. Let $f : C \to [-\infty, \infty)$ be a convex, lower semi-continuous extended-real-valued function such that $\mathbb{E}[f(V)]$ exists. Suppose $\mathcal{E}$ is a sub-σ-algebra of $\mathcal{F}$. Then*

$$f(\mathbb{E}[V \mid \mathcal{E}]) \leq \mathbb{E}[f(V) \mid \mathcal{E}].$$

Proof of Theorem 3.1.11. Let $\mathcal{T}^*$ be the dual space of all real-valued continuous linear functionals on $\mathcal{T}$. The first part of the proof of (Perlman, 1974, Theorem 3.6) tells us that, for all $v \in \mathcal{T}$, we can write

$$f(v) = \sup\{m(v) \mid m \text{ affine, } m \leq f \text{ on } C\},$$

where an *affine* function m on $\mathcal{T}$ is of the form $m(v) = v^*(v) + \alpha$ for some $v^* \in \mathcal{T}^*$ and $\alpha \in \mathbb{R}$. If we define the subset Q of $\mathcal{T}^* \times \mathbb{R}$ as

$$Q := \{(v^*, \alpha) : v^* \in \mathcal{T}^*, \alpha \in \mathbb{R}, v^*(v) + \alpha \leq f(v) \text{ for all } v \in \mathcal{T}\},$$

then we can rewrite f as

$$f(v) = \sup_{(v^*,\alpha) \in Q} \{v^*(v) + \alpha\}, \qquad \text{for all } v \in \mathcal{T}. \tag{A.2}$$

See that, for any $(v^*, \alpha) \in Q$, we have

$$\begin{aligned} \mathbb{E}\left[f(V) \mid \mathcal{E}\right] &\geq \mathbb{E}\left[v^*(V) + \alpha \mid \mathcal{E}\right] && \text{almost surely, by assumption (*)} \\ &= \mathbb{E}\left[v^*\left(V\right) \mid \mathcal{E}\right] + \alpha && \text{almost surely, by linearity (**).} \end{aligned}$$

Here, (*) and (**) use the properties of conditional expectation of vector-valued random variables given in (Dinculeanu, 2000, pp.45-46, Properties 43 and 40 respectively).

We want to show that $\mathbb{E}\left[v^*(V) \mid \mathcal{E}\right] = v^*\left(\mathbb{E}\left[V \mid \mathcal{E}\right]\right)$ almost surely, and in order to so, we show that the right-hand side is a version of the left-hand side. The right-hand side is clearly $\mathcal{E}$-measurable, since we have a linear operator on an $\mathcal{E}$-measurable random variable. Moreover, for any $A \in \mathcal{E}$, using the linearity of the integration operation (Cohn, 2013, p.403, Proposition E.11),

$$\begin{aligned} \int_A v^*\left(\mathbb{E}\left[V \mid \mathcal{E}\right]\right) dP &= v^*\left(\int_A \mathbb{E}\left[V \mid \mathcal{E}\right] dP\right) \\ &= v^*\left(\int_A V dP\right) \\ &= \int_A v^*\left(V\right) dP \end{aligned}$$

(here, all the equalities are almost-sure equalities). Hence, by the definition of the conditional expectation, we have that $\mathbb{E}\left[v^*(V) \mid \mathcal{E}\right] = v^*\left(\mathbb{E}\left[V \mid \mathcal{E}\right]\right)$ almost surely. Going back to our above work, this means that

$$\mathbb{E}\left[f(V) \mid \mathcal{E}\right] \geq v^*\left(\mathbb{E}\left[V \mid \mathcal{E}\right]\right) + \alpha.$$

Now take the supremum of the right-hand side over Q. Then (A.2) tells us that

$$\mathbb{E}\left[f(V) \mid \mathcal{E}\right] \geq f\left(\mathbb{E}\left[V \mid \mathcal{E}\right]\right),$$

as required. □

Lemma 3.2.2. *For any $f \in \mathcal{H}_\mathcal{X}$, $\mathbb{E}[f(X) \mid Z] = \langle f, \mu_{P_{X|Z}} \rangle_{\mathcal{H}_\mathcal{X}}$ almost surely.*

Proof of Lemma 3.2.2. The left-hand side is the conditional expectation of the real-valued random variable $f(X)$ given Z. We need to check that the right-hand side is also that. Note that $\langle f, \mu_{P_{X|Z}} \rangle_{\mathcal{H}_\mathcal{X}}$ is clearly Z-measurable, and P-integrable (by the Cauchy-Schwarz inequality and the integrability condition (3.1)). Take any $A \in \sigma(Z)$. Then

$$\begin{aligned}
\int_A \langle f, \mu_{P_{X|Z}} \rangle_{\mathcal{H}_\mathcal{X}} dP &= \int_A \left\langle f, \mathbb{E}_{X|Z}[k_\mathcal{X}(\cdot, X) \mid Z] \right\rangle_{\mathcal{H}_\mathcal{X}} dP && \text{by definition} \\
&= \left\langle f, \int_A \mathbb{E}_{X|Z}[k_\mathcal{X}(\cdot, X) \mid Z] dP \right\rangle_{\mathcal{H}_\mathcal{X}} && (+) \\
&= \left\langle f, \int_A k_\mathcal{X}(\cdot, X) dP \right\rangle_{\mathcal{H}_\mathcal{X}} && \text{see Definition 3.1.5} \\
&= \int_A \langle f, k_\mathcal{X}(\cdot, X) \rangle_{\mathcal{H}_\mathcal{X}} dP && (+) \\
&= \int_A f(X) dP
\end{aligned}$$

by the reproducing property. Here, in (+), we used the fact that the order of a continuous linear operator and Bochner integration can be interchanged (Dinculeanu, 2000, p.30, Theorem 36). Hence $\langle f, \mu_{P_{X|Z}} \rangle_{\mathcal{H}_\mathcal{X}}$ is a version of the conditional expectation $\mathbb{E}_{X|Z}[f(X) \mid Z]$. □

Lemma 3.2.3. *For any pair $f \in \mathcal{H}_\mathcal{X}$ and $g \in \mathcal{H}_\mathcal{Y}$, $\mathbb{E}[f(X)g(Y) \mid Z] = \langle f \otimes g, \mu_{P_{XY|Z}} \rangle_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}}$ almost surely.*

Proof of Lemma 3.2.3. The left-hand side is the conditional expectation of the real-valued random variable $f(X)g(Y)$ given Z. We need to check that the right-hand side is also that. Note that $\langle f \otimes g, \mu_{P_{XY|Z}} \rangle_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}}$ is clearly Z-measurable, and P-integrable (by the Cauchy-Schwarz inequality and the integrability condition (3.2)). Take any $A \in \sigma(Z)$. Then

$$\begin{aligned}
&\int_A \langle f \otimes g, \mu_{P_{XY|Z}} \rangle_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}} dP \\
&= \int_A \left\langle f \otimes g, \mathbb{E}[k_\mathcal{X}(\cdot, X) \otimes k_\mathcal{Y}(\cdot, Y) \mid Z] \right\rangle_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}} dP \\
&= \left\langle f \otimes g, \int_A \mathbb{E}[k_\mathcal{X}(\cdot, X) \otimes k_\mathcal{Y}(\cdot, Y) \mid Z] dP \right\rangle_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}} \\
&= \left\langle f \otimes g, \int_A k_\mathcal{X}(\cdot, X) \otimes k_\mathcal{Y}(\cdot, Y) dP \right\rangle_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}}
\end{aligned}$$

$$= \int_A \langle f \otimes g, k_{\mathcal{X}}(\cdot, X) \otimes k_{\mathcal{Y}}(\cdot, Y) \rangle_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}} dP$$
$$= \int_A f(X) g(Y) dP.$$

So $\langle f \otimes g, \mu_{P_{XY|Z}} \rangle_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}}$ is a version of $\mathbb{E}[f(X)g(Y) \mid Z]$. $\square$

Theorem 3.3.1. *Denote the Borel σ-algebra of $\mathcal{H}_{\mathcal{X}}$ by $\mathcal{B}(\mathcal{H}_{\mathcal{X}})$. Then we can write $\mu_{P_{X|Z}} = F_{P_{X|Z}} \circ Z$, where $F_{P_{X|Z}} : \mathcal{Z} \to \mathcal{H}_{\mathcal{X}}$ is some deterministic function, measurable with respect to $\mathfrak{Z}$ and $\mathcal{B}(\mathcal{H}_{\mathcal{X}})$.*

Proof of Theorem 3.3.1. Let $\text{Im}(Z) \subseteq \mathcal{Z}$ be the image of $Z : \Omega \to \mathcal{Z}$, and let $\tilde{\mathfrak{Z}}$ denote the σ-algebra on $\text{Im}(Z)$ defined by $\tilde{\mathfrak{Z}} = \{A \cap \text{Im}(Z) : A \in \mathfrak{Z}\}$ (see (Çınlar, 2011, page 5, 1.15)). We will first construct a function $\tilde{F} : \text{Im}(Z) \to \mathcal{H}_{\mathcal{X}}$, measurable with respect to $\tilde{\mathfrak{Z}}$ and $\mathcal{B}(\mathcal{H}_{\mathcal{X}})$, such that $\mu_{P_{X|Z}} = \tilde{F} \circ Z$.

For a given $z \in \text{Im}(Z) \subseteq \mathcal{Z}$, we have $Z^{-1}(z) \subseteq \Omega$. Suppose for contradiction that there are two distinct elements $\omega_1, \omega_2 \in Z^{-1}(z)$ such that $\mu_{P_{X|Z}}(\omega_1) \neq \mu_{P_{X|Z}}(\omega_2)$. Since $\mathcal{H}_{\mathcal{X}}$ is Hausdorff, there are disjoint open neighbourhoods N_1 and N_2 of $\mu_{P_{X|Z}}(\omega_1)$ and $\mu_{P_{X|Z}}(\omega_2)$ respectively. By definition of a Borel σ-algebra, we have $N_1, N_2 \in \mathcal{B}(\mathcal{H}_{\mathcal{X}})$, and since $\mu_{P_{X|Z}}$ is $\sigma(Z)$-measurable,

$$\mu^{-1}_{P_{X|Z}}(N_1), \mu^{-1}_{P_{X|Z}}(N_2) \in \sigma(Z). \tag{A.3}$$

Furthermore, $\mu^{-1}_{P_{X|Z}}(N_1)$ and $\mu^{-1}_{P_{X|Z}}(N_2)$ are neighbourhoods of ω_1 and ω_2 respectively, and are disjoint.

(i) For any $B \in \tilde{\mathfrak{Z}}$ with $z \in B$, since $Z(\omega_1) = z = Z(\omega_2)$, we have $\omega_1, \omega_2 \in Z^{-1}(B)$. So $Z^{-1}(B) \neq \mu^{-1}_{P_{X|Z}}(N_1)$ and $Z^{-1}(B) \neq \mu^{-1}_{P_{X|Z}}(N_2)$, as $\omega_2 \notin \mu^{-1}_{P_{X|Z}}(N_1)$ and $\omega_1 \notin \mu^{-1}_{P_{X|Z}}(N_2)$.

(ii) For any $B \in \tilde{\mathfrak{Z}}$ with $z \notin B$, we have $\omega_1 \notin Z^{-1}(B)$ and $\omega_2 \notin Z^{-1}(B)$. So $Z^{-1}(B) \neq \mu^{-1}_{P_{X|Z}}(N_1)$ and $Z^{-1}(B) \neq \mu^{-1}_{P_{X|Z}}(N_2)$.

Since $\sigma(Z) = \{Z^{-1}(B) \mid B \in \tilde{\mathfrak{Z}}\}$ (see (Çınlar, 2011), page 11, Exercise 2.20), we can't have $\mu^{-1}_{P_{X|Z}}(N_1) \in \sigma(Z)$ nor $\mu^{-1}_{P_{X|Z}}(N_2) \in \sigma(Z)$. This is a contradiction to (A.3). We therefore conclude that, for any $z \in \mathcal{Z}$, if $Z(\omega_1) = z = Z(\omega_2)$ for distinct $\omega_1, \omega_2 \in \Omega$, then $\mu_{P_{X|Z}}(\omega_1) = \mu_{P_{X|Z}}(\omega_2)$.

We define $\tilde{F}(z)$ to be the unique value of $\mu_{P_{X|Z}}(\omega)$ for all $\omega \in Z^{-1}(z)$. Then for any $\omega \in \Omega$, $\mu_{P_{X|Z}}(\omega) = \tilde{F}(Z(\omega))$ by construction. It remains to check that $\tilde{F}$ is measurable with respect to $\tilde{\mathfrak{Z}}$ and $\mathcal{B}(\mathcal{H}_{\mathcal{X}})$.

Take any $N \in \mathcal{B}(\mathcal{H}_{\mathcal{X}})$. Since $\mu_{P_{X|Z}}$ is $\sigma(Z)$-measurable,

$$\mu^{-1}_{P_{X|Z}}(N) = Z^{-1}(\tilde{F}^{-1}(N)) \in \sigma(Z).$$

Since $\sigma(Z) = \{Z^{-1}(B) \mid B \in \tilde{\mathfrak{Z}}\}$, we have $Z^{-1}(\tilde{F}^{-1}(N)) = Z^{-1}(C)$ for some $C \in \tilde{\mathfrak{Z}}$. Since the mapping $Z : \Omega \to \text{Im}(Z)$ is surjective, $\tilde{F}^{-1}(N) = C$. Hence $\tilde{F}^{-1}(N) \in \tilde{\mathfrak{Z}}$, and so $\tilde{F}$ is measurable with respect to $\tilde{\mathfrak{Z}}$ and $\mathcal{B}(\mathcal{H}_{\mathcal{X}})$.

Finally, we can extend $\tilde{F} : \mathrm{Im}(Z) \to \mathcal{H}_{\mathcal{X}}$ to $F : \mathcal{Z} \to \mathcal{H}_{\mathcal{X}}$ by (Dudley, 2018, page 128, Corollary 4.2.7) (note that $\mathcal{H}_{\mathcal{X}}$ is a complete metric space, and assumed to be separable in this theorem). □

Theorem 3.3.2. *$F_{P_{X|Z}}$ minimises both $\tilde{\mathcal{E}}_{X|Z}$ and $\mathcal{E}_{X|Z}$ over $L^2(\mathcal{Z}, P_Z; \mathcal{H}_{\mathcal{X}})$. Moreover, it is almost surely equal to any other minimiser of the loss functionals.*

Proof of Theorem 3.3.2. Recall that we have

$$\mathcal{E}_{X|Z}(F) := \mathbb{E}_Z\left[\|F_{P_{X|Z}}(Z) - F(Z)\|^2_{\mathcal{H}_{\mathcal{X}}}\right].$$

So clearly, $\mathcal{E}_{X|Z}(F_{P_{X|Z}}) = 0$, meaning $F_{P_{X|Z}}$ minimises $\mathcal{E}_{X|Z}$ in $L^2(\mathcal{Z}, P_Z; \mathcal{H}_{\mathcal{X}})$. So it only remains to show that $\tilde{\mathcal{E}}_{X|Z}$ is minimised in $L^2(\mathcal{Z}, P_Z; \mathcal{H}_{\mathcal{X}})$ by $F_{P_{X|Z}}$.

Let F be any element in $L^2(\mathcal{Z}, P_Z; \mathcal{H}_{\mathcal{X}})$. Then we have

$$\begin{aligned}\tilde{\mathcal{E}}_{X|Z}(F) - \tilde{\mathcal{E}}_{X|Z}(F_{P_{X|Z}}) &= \mathbb{E}_{X,Z}[\|k_{\mathcal{X}}(X,\cdot) - F(Z)\|^2_{\mathcal{H}_{\mathcal{X}}}] \\ &\quad - \mathbb{E}_{X,Z}[\|k_{\mathcal{X}}(X,\cdot) - F_{P_{X|Z}}(Z)\|^2_{\mathcal{H}_{\mathcal{X}}}] \\ &= \mathbb{E}_Z[\|F(Z)\|^2_{\mathcal{H}_{\mathcal{X}}}] \\ &\quad - 2\mathbb{E}_{X,Z}[\langle k_{\mathcal{X}}(X,\cdot), F(Z)\rangle_{\mathcal{H}_{\mathcal{X}}}] \\ &\quad + 2\mathbb{E}_{X,Z}\left[\langle k_{\mathcal{X}}(X,\cdot), F_{P_{X|Z}}(Z)\rangle_{\mathcal{H}_{\mathcal{X}}}\right] \\ &\quad - \mathbb{E}_Z\left[\|F_{P_{X|Z}}(Z)\|^2_{\mathcal{H}_{\mathcal{X}}}\right].\end{aligned} \tag{A.4}$$

Here, by the reproducing property and Lemma 3.2.2,

$$\begin{aligned}\mathbb{E}_{X,Z}\left[\langle k_{\mathcal{X}}(X,\cdot), F(Z)\rangle_{\mathcal{H}_{\mathcal{X}}}\right] &= \mathbb{E}_Z\left[\mathbb{E}_{X|Z}\left[F(Z)(X) \mid Z\right]\right] \\ &= \mathbb{E}_Z\left[\langle F(Z), \mu_{P_{X|Z}}\rangle_{\mathcal{H}_{\mathcal{X}}}\right] \\ &= \mathbb{E}_Z\left[\langle F(Z), F_{P_{X|Z}}(Z)\rangle_{\mathcal{H}_{\mathcal{X}}}\right]\end{aligned}$$

and similarly,

$$\begin{aligned}\mathbb{E}_{X,Z}[\langle k_{\mathcal{X}}(X,\cdot), F_{P_{X|Z}}(Z)\rangle_{\mathcal{H}_{\mathcal{X}}}] &= \mathbb{E}_Z[\mathbb{E}_{X|Z}[F_{P_{X|Z}}(Z)(X) \mid Z]] \\ &= \mathbb{E}_Z\left[\langle F_{P_{X|Z}}(Z), F_{P_{X|Z}}(Z)\rangle_{\mathcal{H}_{\mathcal{X}}}\right] \\ &= \mathbb{E}_Z\left[\|F_{P_{X|Z}}(Z)\|^2_{\mathcal{H}_{\mathcal{X}}}\right].\end{aligned}$$

Substituting these expressions back into (A.4), we have

$$\begin{aligned}&\tilde{\mathcal{E}}_{X|Z}(F) - \tilde{\mathcal{E}}_{X|Z}(F_{P_{X|Z}}) \\ &\qquad = \mathbb{E}_Z[\|F(Z)\|^2_{\mathcal{H}_{\mathcal{X}}}] - 2\mathbb{E}_Z[\langle F(Z), F_{P_{X|Z}}(Z)\rangle_{\mathcal{H}_{\mathcal{X}}}] + \mathbb{E}_Z[\|F_{P_{X|Z}}(Z)\|^2_{\mathcal{H}_{\mathcal{X}}}] \\ &\qquad = \mathbb{E}_Z[\|F(Z) - F_{P_{X|Z}}(Z)\|^2_{\mathcal{H}_{\mathcal{X}}}] \\ &\qquad \geq 0.\end{aligned}$$

Hence, $F_{P_{X|Z}}$ minimises $\tilde{\mathcal{E}}_{X|Z}$ in $L^2(\mathcal{Z}, P_Z; \mathcal{H}_{\mathcal{X}})$. The minimiser is further more P_Z-almost surely unique; indeed, if $F' \in L^2(\mathcal{Z}, P_Z; \mathcal{H}_{\mathcal{X}})$ is another minimiser of $\tilde{\mathcal{E}}_{X|Z}$, then the calculation in (A.4) shows that

$$\mathbb{E}_Z\left[\|F_{P_{X|Z}}(Z) - F'(Z)\|^2_{\mathcal{H}_{\mathcal{X}}}\right] = 0,$$

which immediately implies that $\|F_{P_{X|Z}}(Z) - F'(Z)\|_{\mathcal{H}_\mathcal{X}} = 0$ P_Z-almost surely, which in turn implies that $F_{P_{X|Z}} = F'$ P_Z-almost surely. □

Theorem 3.3.4. *Suppose that $k_\mathcal{X}$ and $k_\mathcal{Z}$ are bounded kernels, i.e. there are $B_\mathcal{Z}, B_\mathcal{X} > 0$ with $\sup_{z\in\mathcal{Z}} k_\mathcal{Z}(z,z) \le B_\mathcal{Z}^2$, $\sup_{x\in\mathcal{X}} k_\mathcal{X}(x,x) \le B_\mathcal{X}^2$, and that the operator-valued kernel $l_{\mathcal{XZ}}$ is $\mathcal{C}_0$-universal. Let the regularisation parameter λ_n decay to 0 at a slower rate than $\mathcal{O}(n^{-1/2})$. Then the learning algorithm that yields $\hat{F}_{P_{X|Z},n,\lambda_n}$ is universally consistent, i.e. for any joint distribution P_{XZ}, $\epsilon > 0$ and $\delta > 0$, $P_{XZ}(\tilde{\mathcal{E}}_{X|Z}(\hat{F}_{P_{X|Z},n,\lambda_n}) - \tilde{\mathcal{E}}_{X|Z}(F_{P_{X|Z}}) > \epsilon) < \delta$ for sufficiently large n.*

Proof of Theorem 3.3.4. Follows immediately from Theorem 5.1.10. □

Theorem 3.3.5. *Assume further that $F_{P_{X|Z}} \in \mathcal{G}_{\mathcal{XZ}}$. Then with probability at least $1-\delta$,*

$$\begin{aligned}\tilde{\mathcal{E}}_{X|Z}(\hat{F}_{P_{X|Z},n,\lambda_n}) - \tilde{\mathcal{E}}_{X|Z}(F_{P_{X|Z}}) &\le \lambda_n \left\|F_{P_{X|Z}}\right\|^2_{\mathcal{G}_{\mathcal{XZ}}} \\ &+ \frac{2\ln\left(\frac{4}{\delta}\right)}{3n\lambda_n}\left(1+\sqrt{1+\frac{18n}{\ln\left(\frac{4}{\delta}\right)}}\right) \\ &\quad \left(\left(B_\mathcal{Z}\left\|F_{P_{X|Z}}\right\|_{\mathcal{G}_{\mathcal{XZ}}}+B_\mathcal{X}\right)^2\lambda_n + B_\mathcal{X}^2\left(B_\mathcal{Z}+\sqrt{\lambda_n}\right)^2\right)\end{aligned}$$

Proof of Theorem 3.3.5. Follows immediately from Theorem 5.1.11. □

Theorem 3.4.2. *Suppose that $k_\mathcal{X}$ is characteristic, that P_Z and $P_{Z'}$ are absolutely continuous with respect to each other, and that $P(\cdot \mid Z)$ and $P(\cdot \mid Z')$ admit regular versions. Then $M_{P_{X|Z},P_{X'|Z'}} = 0$ almost everywhere if and only if, for almost all $z \in \mathcal{Z}$, $P_{X|Z=z}(B) = P_{X'|Z'=z}(B)$ for all $B \in \mathfrak{X}$.*

Proof of Theorem 3.4.2. Write Q and Q' for some regular versions of $P(\cdot \mid Z)$ and $P(\cdot \mid Z')$ respectively, and assume without loss of generality that the conditional distributions $P_{X|Z}$ and $P_{X'|Z'}$ are given by $P_{X|Z}(\omega)(B) = Q_\omega(X \in B)$ and $P_{X'|Z'}(\omega)(B) = Q'_\omega(X' \in B)$ for $B \in \mathfrak{X}$. By the definition of regular versions, for each $B \in \mathfrak{X}$, the real-valued random variables $\omega \mapsto P_{X|Z}(\omega)(B)$ and $\omega \mapsto P_{X'|Z'}(\omega)(B)$ are measurable with respect to Z and Z' respectively, and so there are functions $R_B : \mathcal{Z} \to \mathbb{R}$ and $R'_B : \mathcal{Z} \to \mathbb{R}$ such that $P_{X|Z}(\omega)(B) = R_B(Z(\omega))$ and $P_{X'|Z'}(\omega)(B) = R'_B(Z'(\omega))$. Moreover, for each fixed $z \in \mathcal{Z}$, the mappings $B \mapsto P_{X|Z}(Z^{-1}(z))(B) = R_B(z)$ and $B \mapsto P_{X'|Z'}(Z'^{-1}(z))(B) = R'_B(z)$ are measures. We write $R_B(z) = P_{X|Z=z}(B)$ and $R'_B(z) = P_{X'|Z'=z}(B)$.

By Theorem 3.1.6, there exists an event $A_1 \in \mathcal{F}$ with $P(A_1) = 1$ such that for all $\omega \in A_1$,

$$\begin{aligned}\mu_{P_{X|Z}}(\omega) &:= \mathbb{E}_{X|Z}[k_\mathcal{X}(X,\cdot) \mid Z](\omega) \\ &= \int_\Omega k_\mathcal{X}(X(\omega'),\cdot)Q_\omega(d\omega')\end{aligned}$$

$$= \int_{\mathcal{X}} k_{\mathcal{X}}(x,\cdot)P_{X|Z}(\omega)(dx),$$

and an event $A_2 \in \mathcal{F}$ with $P(A_2) = 1$ such that for all $\omega \in A_2$,

$$\begin{aligned}\mu_{P_{X'|Z'}}(\omega) &:= \mathbb{E}_{X'|Z'}[k_{\mathcal{X}}(X',\cdot) \mid Z'](\omega) \\ &= \int_{\Omega} k_{\mathcal{X}}(X'(\omega'),\cdot)Q_{\omega}(d\omega') \\ &= \int_{\mathcal{X}} k_{\mathcal{X}}(x',\cdot)P_{X'|Z'}(\omega)(dx').\end{aligned}$$

Suppose for contradiction that there exists some $D \in \mathfrak{Z}$ with $P_Z(D) > 0$ such that for all $z \in D$, $F_{P_{X|Z}}(z) \neq \int_{\mathcal{X}} k_{\mathcal{X}}(x,\cdot)R_{dx}(z)$. Then $P(Z^{-1}(D)) = P_Z(D) > 0$, and hence $P(Z^{-1}(D) \cap A_1) > 0$. For all $\omega \in Z^{-1}(D) \cap A_1$, we have $Z(\omega) \in D$, and hence

$$\mu_{P_{X|Z}}(\omega) = F_{P_{X|Z}}(Z(\omega)) \neq \int_{\mathcal{X}} k_{\mathcal{X}}(x,\cdot)R_{dx}(Z(\omega)) = \int_{\mathcal{X}} k_{\mathcal{X}}(x,\cdot)P_{X|Z}(\omega)(dx).$$

This contradicts our assertion that $\mu_{P_{X|Z}}(\omega) = \int_{\mathcal{X}} k_{\mathcal{X}}(x,\cdot)P_{X|Z}(\omega)(dx)$ for all $\omega \in A_1$, hence there does not exist $D \in \mathfrak{Z}$ with $P_Z(D) > 0$ such that for all $z \in D$, $F_{P_{X|Z}}(z) \neq \int_{\mathcal{X}} k_{\mathcal{X}}(x,\cdot)R_{dx}(z)$. Therefore, there must exist some $C_1 \in \mathfrak{Z}$ with $P_Z(C_1) = 1$ such that for all $z \in C_1$, $F_{P_{X|Z}}(z) = \int_{\mathcal{X}} k_{\mathcal{X}}(x,\cdot)R_{dx}(z)$. Similarly, there must exist some $C_2 \in \mathfrak{Z}$ with $P_Z(C_2) = 1$ such that for all $z \in C_2$, $F_{P_{X'|Z'}}(z) = \int_{\mathcal{X}} k_{\mathcal{X}}(x,\cdot)R'_{dx}(z)$. Since P_Z and $P_{Z'}$ are absolutely continuous with respect to each other, we also have $P_Z(C_2) = 1 = P_{Z'}(C_1)$.

($\Longrightarrow$) Suppose first that $\text{MCMD}_{P_{X|Z},P_{X'|Z'}} = \|F_{P_{X|Z}} - F_{P_{X'|Z'}}\|_{\mathcal{H}_{\mathcal{X}}} = 0$ P_Z-almost everywhere, i.e. there exists $C \in \mathfrak{Z}$ with $P_Z(C) = 1$ such that for all $z \in C$, $\|F_{P_{X|Z}}(z) - F_{P_{X'|Z'}}(z)\|_{\mathcal{H}_{\mathcal{X}}} = 0$. Then for each $z \in C \cap C_1 \cap C_2$,

$$\begin{aligned}\int_{\mathcal{X}} k_{\mathcal{X}}(x,\cdot)R_{dx}(z) &= F_{P_{X|Z}}(z) && \text{since } z \in C_1 \\ &= F_{P_{X'|Z'}}(z) && \text{since } z \in C \\ &= \int_{\mathcal{X}} k_{\mathcal{X}}(x,\cdot)R'_{dx}(z) && \text{since } z \in C_2.\end{aligned}$$

Since the kernel $k_{\mathcal{X}}$ is characteristic, this means that $B \mapsto R_B(z)$ and $B \mapsto R'_B(z)$ are the same probability measure on $(\mathcal{X}, \mathfrak{X})$. By countable intersection, we have $P_Z(C \cap C_1 \cap C_2) = 1$, so P_Z-almost everywhere,

$$P_{X|Z=z}(B) = P_{X'|Z'=z}(B)$$

for all $B \in \mathfrak{X}$.

($\Longleftarrow$) Now assume there exists $C \in \mathfrak{Z}$ with $P_Z(C) = 1$ such that for each $z \in C$, $R_B(z) = R'_B(z)$ for all $B \in \mathfrak{X}$. Then for all $z \in C \cap C_1 \cap C_2$,

$$\left\|F_{P_{X|Z}}(z) - F_{P_{X'|Z'}}(z)\right\|_{\mathcal{H}_{\mathcal{X}}}$$

$$\begin{aligned}
&= \left\| \int_{\mathcal{X}} k_{\mathcal{X}}(x,\cdot)R_{dx}(z) - \int_{\mathcal{X}} k_{\mathcal{X}}(x,\cdot)R'_{dx}(z) \right\|_{\mathcal{H}_{\mathcal{X}}} && \text{since } z \in C_1 \cap C_2 \\
&= \left\| \int_{\mathcal{X}} k_{\mathcal{X}}(x,\cdot)R_{dx}(z) - \int_{\mathcal{X}} k_{\mathcal{X}}(x,\cdot)R_{dx}(z) \right\|_{\mathcal{H}_{\mathcal{X}}} && \text{since } z \in C \\
&= 0,
\end{aligned}$$

and since $P_Z(C \cap C_1 \cap C_2) = 1$, $\|F_{P_{X|Z}} - F_{P_{X'|Z'}}\|_{\mathcal{H}_{\mathcal{X}}} = 0$ P_Z-almost everywhere.

□

Theorem 3.4.4. *Suppose $k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$ is a characteristic kernel*[2] *on $\mathcal{X} \times \mathcal{Y}$, and that $P(\cdot \mid Z)$ admits a regular version. Then* $\text{HSCIC}(X, Y \mid Z) = 0$ *almost surely if and only if $X \perp\!\!\!\perp Y \mid Z$.*

Proof of Theorem 3.4.4. Write Q for a regular version of $P(\cdot \mid Z)$, and assume without loss of generality that the conditional distributions $P_{X|Z}$, $P_{Y|Z}$ and $P_{XY|Z}$ are given by $P_{X|Z}(\omega)(B) = Q_\omega(X \in B)$ for $B \in \mathcal{X}$, $P_{Y|Z}(\omega)(C) = Q_\omega(Y \in C)$ for $C \in \mathfrak{Y}$ and $P_{XY|Z}(\omega)(D) = Q_\omega((X,Y) \in D)$ for $D \in \mathfrak{X} \times \mathfrak{Y}$. By Theorem 3.1.6, there exists an event $A_1 \in \mathcal{F}$ with $P(A_1) = 1$ such that for all $\omega \in A_1$,

$$\begin{aligned}
\mu_{P_{X|Z}}(\omega) &:= \mathbb{E}_{X|Z}[k_{\mathcal{X}}(X,\cdot) \mid Z](\omega) \\
&= \int_\Omega k_{\mathcal{X}}(X(\omega'),\cdot)Q_\omega(d\omega') \\
&= \int_{\mathcal{X}} k_{\mathcal{X}}(x,\cdot)P_{X|Z}(\omega)(dx),
\end{aligned}$$

an event $A_2 \in \mathcal{F}$ with $P(A_2) = 1$ such that for all $\omega \in A_2$,

$$\begin{aligned}
\mu_{P_{Y|Z}}(\omega) &:= \mathbb{E}_{Y|Z}[k_{\mathcal{Y}}(Y,\cdot) \mid Z](\omega) \\
&= \int_\Omega k_{\mathcal{Y}}(Y(\omega'),\cdot)Q_\omega(d\omega') \\
&= \int_{\mathcal{Y}} k_{\mathcal{Y}}(y,\cdot)P_{Y|Z}(\omega)(dy),
\end{aligned}$$

and an event $A_3 \in \mathcal{F}$ with $P(A_3) = 1$ such that for all $\omega \in A_3$,

$$\mu_{P_{XY|Z}}(\omega) = \int_{\mathcal{X} \times \mathcal{Y}} k_{\mathcal{X}}(x,\cdot) \otimes k_{\mathcal{Y}}(y,\cdot)P_{XY|Z}(\omega)(d(x,y)).$$

This means that, for each $\omega \in A_1$, $\mu_{P_{X|Z}}(\omega)$ is the mean embedding of $P_{X|Z}(\omega)$, and for each $\omega \in A_2$, $\mu_{P_{Y|Z}}(\omega)$ is the mean embedding of $P_{Y|Z}(\omega)$.

[2] See (Szabó and Sriperumbudur, 2017) for a detailed discussion on characteristic tensor product kernels.

($\Longrightarrow$) Suppose first that $\text{HSCIC}(X, Y \mid Z) = 0$ almost surely, i.e. there exists $A \in \mathcal{F}$ with $P(A) = 1$ such that for all $\omega \in A$, $\|\mu_{P_{XY|Z}}(\omega) - \mu_{P_{X|Z}}(\omega) \otimes \mu_{P_{Y|Z}}(\omega)\|_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}} = 0$. Then for each $\omega \in A \cap A_1 \cap A_2 \cap A_3$,

$$\begin{aligned}
\int_{\mathcal{X}\times\mathcal{Y}} k_\mathcal{X}(x, \cdot) \otimes k_\mathcal{Y}(y, \cdot) P_{XY|Z}(\omega)(d(x, y)) &= \mu_{P_{XY|Z}}(\omega) \\
&= \mu_{P_{X|Z}}(\omega) \otimes \mu_{P_{Y|Z}}(\omega) \\
&= \int_\mathcal{X} k_\mathcal{X}(x, \cdot) P_{X|Z}(\omega)(dx) \otimes \int_\mathcal{Y} k_\mathcal{Y}(y, \cdot) P_{Y|Z}(\omega)(dy) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} k_\mathcal{X}(x, \cdot) \otimes k_\mathcal{Y}(y, \cdot) P_{X|Z}(\omega) P_{Y|Z}(\omega)(d(x, y)).
\end{aligned}$$

Since the kernel $k_\mathcal{X} \otimes k_\mathcal{Y}$ is characteristic, the distributions $P_{XY|Z}(\omega)$ and $P_{X|Z}(\omega)P_{Y|Z}(\omega)$ on $\mathcal{X} \times \mathcal{Y}$ are the same. By countable intersection, we have $P(A \cap A_1 \cap A_2 \cap A_3) = 1$, so $P_{XY|Z}$ and $P_{X|Z}P_{Y|Z}$ are the same almost surely, and we have $X \perp\!\!\!\perp Y \mid Z$.

($\Longleftarrow$) Now assume $X \perp\!\!\!\perp Y \mid Z$, i.e. there exists $A \in \mathcal{F}$ with $P(A) = 1$ such that for each $\omega \in A$, the distributions $P_{XY|Z}(\omega)$ and $P_{X|Z}(\omega)P_{Y|Z}(\omega)$ are the same. Then for all $\omega \in A \cap A_1 \cap A_2 \cap A_3$,

$$\begin{aligned}
\mu_{P_{XY|Z}}(\omega) &= \int_{\mathcal{X}\times\mathcal{Y}} k_\mathcal{X}(x, \cdot) \otimes k_\mathcal{Y}(y, \cdot) P_{XY|Z}(\omega)(d(x, y)) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} k_\mathcal{X}(x, \cdot) \otimes k_\mathcal{Y}(y, \cdot) P_{X|Z}(\omega)(dx) P_{Y|Z}(\omega)(dy) \\
&= \int_\mathcal{X} k_\mathcal{X}(x, \cdot) P_{X|Z}(\omega)(dx) \otimes \int_\mathcal{Y} k_\mathcal{Y}(y, \cdot) P_{Y|Z}(\omega)(dy) \\
&= \mu_{P_{X|Z}}(\omega) \otimes \mu_{P_{Y|Z}}(\omega).
\end{aligned}$$

and since $P(A \cap A_1 \cap A_2 \cap A_3) = 1$, $\text{HSCIC}(X, Y \mid Z) = 0$ almost surely.

□

A.4 Proofs for Chapter 4

Lemma 4.3.1. *For each $x \in \mathcal{X}$, we have*

$$\begin{aligned}
\hat{U}^2_{\text{MMD}}(x) = \boldsymbol{k}_0^T(x)\mathbf{W}_0\mathbf{L}_0\mathbf{W}_0^T\boldsymbol{k}_0(x)& \\
- 2\boldsymbol{k}_0^T(x)\mathbf{W}_0\mathbf{L}\mathbf{W}_1^T\boldsymbol{k}_1(x)& \\
+ \boldsymbol{k}_1^T(x)\mathbf{W}_1\mathbf{L}_1\mathbf{W}_1^T\boldsymbol{k}_1(x)&,
\end{aligned}$$

where $[\mathbf{L}_0]_{1\le i,j\le n_0} = l(y_i^0, y_j^0)$, $[\mathbf{L}]_{1\le i\le n_0, 1\le j\le n_1} = l(y_i^0, y_j^1)$ *and* $[\mathbf{L}_1]_{1\le i,j\le n_1} = l(y_i^1, y_j^1)$.

Proof of Lemma 4.3.1. We use the reproducing property of $\mathcal{H}$ and (4.2) to see that, for any $x \in \mathcal{X}$,

$$
\begin{aligned}
\hat{U}^2_{\text{MMD}}(x) &= \left\|\hat{\mu}_{Y_1|X=x} - \hat{\mu}_{Y_0|X=x}\right\|^2_{\mathcal{H}} \\
&= \left\|\boldsymbol{k}_0^T(x)\mathbf{W}_0\boldsymbol{l}_0 - \boldsymbol{k}_1^T(x)\mathbf{W}_1\boldsymbol{l}_1\right\|^2_{\mathcal{H}} \\
&= \left\langle \sum_{i,j=1}^{n_0} k_0(x, x_i^0)\mathbf{W}_{0,ij} l(y_j^0, \cdot), \sum_{p,q=1}^{n_0} k_0(x, x_p^0)\mathbf{W}_{0,pq} l(y_q^0, \cdot) \right\rangle_{\mathcal{H}} \\
&\quad - 2\left\langle \sum_{i,j=1}^{n_0} k_0(x, x_i^0)\mathbf{W}_{0,ij} l(y_j^0, \cdot), \sum_{p,q=1}^{n_1} k_1(x, x_p^1)\mathbf{W}_{1,pq} l(y_q^1, \cdot) \right\rangle_{\mathcal{H}} \\
&\quad\quad + \left\langle \sum_{i,j=1}^{n_1} k_1(x, x_i^1)\mathbf{W}_{1,ij} l(y_j^1, \cdot), \sum_{p,q=1}^{n_1} k_1(x, x_p^1)\mathbf{W}_{1,pq} l(y_q^1, \cdot) \right\rangle_{\mathcal{H}} \\
&= \sum_{i,j,p,q=1}^{n_0} k_0(x, x_i^0)\mathbf{W}_{0,ij} l(y_j^0, y_q^0)\mathbf{W}^T_{0,qp} k_0(x_p^0, x) \\
&\quad - 2\sum_{i,j=1}^{n_0}\sum_{p,q=1}^{n_1} k_0(x, x_i^0)\mathbf{W}_{0,ij} l(y_j^0, y_q^1)\mathbf{W}^T_{1,qp} k_1(x_p^1, x) \\
&\quad\quad + \sum_{i,j,p,q=1}^{n_1} k_1(x, x_i^1)\mathbf{W}_{1,ij} l(y_j^1, y_q^1)\mathbf{W}^T_{1,qp} k_1(x_p^1, x) \\
&= \boldsymbol{k}_0^T(x)\mathbf{W}_0\mathbf{L}_0\mathbf{W}_0^T\boldsymbol{k}_0(x) \\
&\quad - 2\boldsymbol{k}_0^T(x)\mathbf{W}_0\mathbf{L}\mathbf{W}_1^T\boldsymbol{k}_1(x) \\
&\quad\quad + \boldsymbol{k}_1^T(x)\mathbf{W}_1\mathbf{L}_1\mathbf{W}_1^T\boldsymbol{k}_1(x).
\end{aligned}
$$

□

Theorem 4.3.2 (Universal consistency)**.** *Suppose that k_0, k_1 and l are bounded, that Γ_0 and Γ_1 are universal, and that $\lambda^0_{n_0}$ and $\lambda^1_{n_1}$ decay at slower rates than $\mathcal{O}(n_0^{-1/2})$ and $\mathcal{O}(n_1^{-1/2})$ respectively. Then as $n_0, n_1 \to \infty$,*

$$
\psi_{\text{MMD}}(\hat{U}_{\text{MMD}}) = \mathbb{E}[(\hat{U}_{\text{MMD}}(X) - U_{\text{MMD}}(X))^2] \xrightarrow{p} 0.
$$

Proof of Theorem 4.3.2. The simple inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ holds in any Hilbert space. Using this, we see that

$$
\begin{aligned}
\psi_{\text{MMD}}\left(\hat{U}_{\text{MMD}}\right) &= \mathbb{E}\left[\left(\hat{U}_{\text{MMD}}(X) - U_{\text{MMD}}(X)\right)^2\right] \\
&= \mathbb{E}\left[\left(\left\|\hat{\mu}_{Y_1|X} - \hat{\mu}_{Y_0|X}\right\|_{\mathcal{H}} - \left\|\mu_{Y_1|X} - \mu_{Y_0|X}\right\|_{\mathcal{H}}\right)^2\right] \\
&\leq \mathbb{E}\left[\left\|\hat{\mu}_{Y_1|X} - \mu_{Y_1|X} - \hat{\mu}_{Y_0|X} + \mu_{Y_0|X}\right\|^2_{\mathcal{H}}\right] \\
&\leq 2\mathbb{E}\left[\left\|\hat{\mu}_{Y_1|X} - \mu_{Y_1|X}\right\|^2_{\mathcal{H}} + \left\|\hat{\mu}_{Y_0|X} - \mu_{Y_0|X}\right\|^2_{\mathcal{H}}\right].
\end{aligned}
$$

Hence, it suffices to know that

$$\mathbb{E}\left[\left\|\hat{\mu}_{Y_1|X} - \mu_{Y_1|X}\right\|_{\mathcal{H}}^2\right] \xrightarrow{p} 0 \qquad \text{and} \qquad \mathbb{E}\left[\left\|\hat{\mu}_{Y_0|X} - \mu_{Y_0|X}\right\|_{\mathcal{H}}^2\right] \xrightarrow{p} 0.$$

But this follows immediately from Chapter 5, so the proof is complete. □

Lemma 4.3.3. *If l is a characteristic kernel, $P_{Y_0|X} \equiv P_{Y_1|X}$ if and only if $t = 0$.*

Proof of Lemma 4.3.3. We can assume without loss of generality that $P_{Y_0|X}$ and $P_{Y_1|X}$ are obtained from a regular version of $P(\cdot \mid X)$. Then by Theorem 3.1.6, there exist $C_0, C_1 \in \mathcal{F}$ with $P(C_0) = P(C_1) = 1$ such that for all $\omega \in C_0$, $\mu_{Y_0|X}(\omega) = \int_{\mathcal{Y}} l(y,\cdot)dP_{Y_0|X}(\omega)(y)$ and for all $\omega' \in C_1$, $\mu_{Y_1|X}(\omega') = \int_{\mathcal{Y}} l(y,\cdot)dP_{Y_1|X}(\omega')(y)$.

Suppose for contradiction that there exists some measurable $A \subseteq \mathcal{X}$ with $P_X(A) > 0$ such that for all $x \in A$, $\mu_{Y_0|X=x} \neq \int_{\mathcal{Y}} l(y,\cdot)dP_{Y_0|X=x}(y)$. Then $P(X^{-1}(A)) = P_X(A) > 0$, and hence $P(X^{-1}(A) \cap C_0) > 0$. For all $\omega \in X^{-1}(A) \cap C_0$, we have $X(\omega) \in A$, and hence

$$\mu_{Y_0|X}(\omega) \neq \int_{\mathcal{Y}} l(y,\cdot)dP_{Y_0|X=X(\omega)}(y) = \int_{\mathcal{Y}} l(y,\cdot)P_{Y_0|X}(\omega)(dy) = \mu_{Y_0|X}(\omega).$$

This is a contradiction, hence there does not exist a measurable $A \subseteq \mathcal{X}$ with $P_X(A) > 0$ such that for all $x \in A$, $\mu_{Y_0|X=x} \neq \int_{\mathcal{Y}} l(y,\cdot)dP_{Y_0|X=x}(y)$. Therefore, there must exist some measurable $A_0 \subseteq \mathcal{X}$ with $P_X(A_0) = 1$ such that for all $x \in A_0$, $\mu_{Y_0|X=x} = \int_{\mathcal{Y}} l(y,\cdot)dP_{Y_0|X=x}(y)$. Similarly, there must exist some measurable $A_1 \subseteq \mathcal{X}$ with $P_X(A_1) = 1$ such that for all $x \in A_1$, $\mu_{Y_1|X=x} = \int_{\mathcal{Y}} l(y,\cdot)dP_{Y_1|X=x}(y)$.

($\Longrightarrow$) Suppose that $P_{Y_0|X} \equiv P_{Y_1|X}$. This means that there exists a measurable $A \subseteq \mathcal{X}$ with $P_X(A) = 1$ such that for all $x \in A$, the measures $P_{Y_0|X=x}(\cdot)$ and $P_{Y_1|X=x}(\cdot)$ are the same. Then for all $x \in A \cap A_0 \cap A_1$,

$$\begin{aligned} \mu_{Y_0|X=x} &= \int_{\mathcal{Y}} l(y,\cdot)dP_{Y_0|X=x}(y) && \text{since } x \in A_0 \\ &= \int_{\mathcal{Y}} l(y,\cdot)dP_{Y_1|X=x}(y) && \text{since } x \in A \\ &= \mu_{Y_1|X=x} && \text{since } x \in A_1. \end{aligned}$$

Now, we have $P_X(A) = P_X(A_0) = P_X(A_1) = 1$, so $P_X(A \cap A_0 \cap A_1) = 1$. Since $\mu_{Y_0|X=x} = \mu_{Y_1|X=x}$ for all $x \in A \cap A_0 \cap A_1$, we have $\mu_{Y_0|X=\cdot} = \mu_{Y_1|X=\cdot}$ P_X-almost everywhere. Hence,

$$t = \mathbb{E}\left[\left\|\mu_{Y_1|X} - \mu_{Y_0|X}\right\|_{\mathcal{H}}^2\right] = 0$$

($\Longleftarrow$) Now suppose that $t = 0$, i.e. $\mu_{Y_0|X=\cdot} = \mu_{Y_1|X=\cdot}$ P_X-almost everywhere, say on a measurable set $A \subseteq \mathcal{X}$ with $P_X(A) = 1$. Suppose $x \in A \cap A_0 \cap A_1$.

Then

$$\begin{aligned}
\int_{\mathcal{Y}} l(y,\cdot)dP_{Y_0|X=x}(y) &= \mu_{Y_0|X=x} && \text{since } x \in A_0 \\
&= \mu_{Y_1|X=x} && \text{since } x \in A \\
&= \int_{\mathcal{Y}} l(y,\cdot)dP_{Y_1|X=x}(y) && \text{since } x \in A_1.
\end{aligned}$$

Since $k_{\mathcal{Y}}$ is characteristic, this means that $P_{Y_0|X=x}$ and $P_{Y_1|X=x}$ are the same measure. As before, we have $P_X(A \cap A_0 \cap A_1) = 1$, hence $P_{Y_0|X} \equiv P_{Y_1|X}$.

□

Lemma 4.3.4. *We have*

$$\begin{aligned}
\hat{t} = &\frac{1}{n}\mathrm{Tr}\left(\tilde{\mathbf{K}}_0\mathbf{W}_0\mathbf{L}_0\mathbf{W}_0^T\tilde{\mathbf{K}}_0^T\right) \\
&- \frac{2}{n}\mathrm{Tr}\left(\tilde{\mathbf{K}}_0\mathbf{W}_0\mathbf{L}\mathbf{W}_1^T\tilde{\mathbf{K}}_1^T\right) \\
&+ \frac{1}{n}\mathrm{Tr}\left(\tilde{\mathbf{K}}_1\mathbf{W}_1\mathbf{L}_1\mathbf{W}_1^T\tilde{\mathbf{K}}_1^T\right),
\end{aligned}$$

where $\mathbf{L}_0, \mathbf{L}_1$ *and* $\mathbf{L}$ *are as defined in Lemma 4.3.1 and* $[\tilde{\mathbf{K}}_0]_{1\leq i\leq n, 1\leq j\leq n_0} = k_0(x_i, x_j^0)$ *and* $[\tilde{\mathbf{K}}_1]_{1\leq i\leq n, 1\leq j\leq n_1} = k_1(x_i, x_j^1)$.

Proof of Lemma 4.3.4. See that, using the reproducing property in $\mathcal{H}$ again,

$$\begin{aligned}
\hat{t} &= \frac{1}{n}\sum_{i=1}^{n}\left\|\hat{\mu}_{Y_1|X=x_i} - \hat{\mu}_{Y_0|X=x_i}\right\|_{\mathcal{H}}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\left\|\hat{\mu}_{Y_1|X=x_i}\right\|_{\mathcal{H}}^2 - 2\left\langle\hat{\mu}_{Y_1|X=x_i}, \hat{\mu}_{Y_0|X=x_i}\right\rangle_{\mathcal{H}} + \left\|\hat{\mu}_{Y_0|X=x_i}\right\|_{\mathcal{H}}^2\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\left\|\boldsymbol{k}_0^T(x_i)\mathbf{W}_0\boldsymbol{l}_0\right\|_{\mathcal{H}}^2\right. \\
&\quad -2\left\langle\boldsymbol{k}_0^T(x_i)\mathbf{W}_0\boldsymbol{l}_0, \boldsymbol{k}_1^T(x_i)\mathbf{W}_1\boldsymbol{l}_1\right\rangle_{\mathcal{H}} \\
&\qquad \left. + \left\|\boldsymbol{k}_1^T(x_i)\mathbf{W}_1\boldsymbol{l}_1\right\|_{\mathcal{H}}^2\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\langle\sum_{p,q=1}^{n_0} k_0(x_p^0, x_i)\mathbf{W}_{0,pq}l(y_q^0,\cdot), \sum_{r,s=1}^{n_0} k_0(x_r^0, x_i)\mathbf{W}_{0,rs}l(y_s^0,\cdot)\right\rangle_{\mathcal{H}} \\
&\quad - \frac{2}{n}\sum_{i=1}^{n}\left\langle\sum_{p,q=1}^{n_0} k_0(x_p^0, x_i)\mathbf{W}_{0,pq}l(y_q^0,\cdot), \sum_{r,s=1}^{n_1} k_1(x_r^1, x_i)\mathbf{W}_{1,rs}l(y_s^1,\cdot)\right\rangle_{\mathcal{H}} \\
&\qquad + \frac{1}{n}\sum_{i=1}^{n}\left\langle\sum_{p,q=1}^{n_1} k_1(x_p^1, x_i)\mathbf{W}_{1,pq}l(y_q^1,\cdot), \sum_{r,s=1}^{n_1} k_1(x_r^1, x_i)\mathbf{W}_{1,rs}l(y_s^1,\cdot)\right\rangle_{\mathcal{H}}
\end{aligned}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{p,q,r,s=1}^{n_0} k_0(x_i, x_p^0)\mathbf{W}_{0,pq} l(y_q^0, y_s^0)\mathbf{W}_{0,sr}^T k_0(x_r^0, x_i)$$
$$- \frac{2}{n}\sum_{i=1}^{n}\sum_{p,q=1}^{n_0}\sum_{r,s=1}^{n_1} k_0(x_i, x_p^0)\mathbf{W}_{0,pq} l(y_q^0, y_s^1)\mathbf{W}_{1,sr}^T k_1(x_r^1, x_i)$$
$$+ \frac{1}{n}\sum_{i=1}^{n}\sum_{p,q,r,s=1}^{n_1} k_1(x_i, x_p^1)\mathbf{W}_{1,pq} l(y_q^1, y_s^1)\mathbf{W}_{1,sr}^T k_1(x_r^1, x_i)$$
$$= \frac{1}{n}\left\{\mathrm{Tr}\left(\tilde{\mathbf{K}}_0\mathbf{W}_0\mathbf{L}_0\mathbf{W}_0^T\tilde{\mathbf{K}}_0^T\right)\right.$$
$$-2\mathrm{Tr}\left(\tilde{\mathbf{K}}_0\mathbf{W}_0\mathbf{L}\mathbf{W}_1^T\tilde{\mathbf{K}}_1^T\right)$$
$$\left.+\mathrm{Tr}\left(\tilde{\mathbf{K}}_1\mathbf{W}_1\mathbf{L}_1\mathbf{W}_1^T\tilde{\mathbf{K}}_1^T\right)\right\}$$

□

Theorem 4.3.5. *Under the same assumptions as in Theorem 4.3.2, we have* $\hat{t} \xrightarrow{p} t$ *as* $n_0, n_1 \to \infty$.

Proof of Theorem 4.3.5. We decompose $\left|\hat{t} - t\right|$ as follows using the triangle inequality:

$$\left|\hat{t} - t\right| = \left|\frac{1}{n}\sum_{i=1}^{n}\left\|\hat{\mu}_{Y_1|X=x_i} - \hat{\mu}_{Y_0|X=x_i}\right\|_{\mathcal{H}}^2 - \mathbb{E}\left[\left\|\mu_{Y_1|X} - \mu_{Y_0|X}\right\|_{\mathcal{H}}^2\right]\right|$$
$$\leq \left|\frac{1}{n}\sum_{i=1}^{n}\left\|\hat{\mu}_{Y_1|X=x_i} - \hat{\mu}_{Y_0|X=x_i}\right\|_{\mathcal{H}}^2 - \mathbb{E}\left[\left\|\hat{\mu}_{Y_1|X} - \hat{\mu}_{Y_0|X}\right\|_{\mathcal{H}}^2\right]\right|$$
$$+ \left|\mathbb{E}\left[\left\|\hat{\mu}_{Y_1|X} - \hat{\mu}_{Y_0|X}\right\|_{\mathcal{H}}^2\right] - \mathbb{E}\left[\left\|\mu_{Y_1|X} - \mu_{Y_0|X}\right\|_{\mathcal{H}}^2\right]\right|$$

Here, the first term converges to 0 in probability by the uniform law of large numbers. For the second term, see that

$$\left|\mathbb{E}\left[\left\|\hat{\mu}_{Y_1|X} - \hat{\mu}_{Y_0|X}\right\|_{\mathcal{H}}^2\right] - \mathbb{E}\left[\left\|\mu_{Y_1|X} - \mu_{Y_0|X}\right\|_{\mathcal{H}}^2\right]\right|$$
$$= \left|\mathbb{E}\left[\left\|\hat{\mu}_{Y_1|X} - \mu_{Y_1|X} + \mu_{Y_1|X} - \mu_{Y_0|X} + \mu_{Y_0|X} - \hat{\mu}_{Y_0|X}\right\|_{\mathcal{H}}^2\right.\right.$$
$$\left.\left. - \left\|\mu_{Y_1|X} - \mu_{Y_0|X}\right\|_{\mathcal{H}}^2\right]\right|$$
$$= \left|\mathbb{E}\left[\left\|\hat{\mu}_{Y_1|X} - \mu_{Y_1|X}\right\|_{\mathcal{H}}^2 + \left\|\mu_{Y_0|X} - \hat{\mu}_{Y_0|X}\right\|_{\mathcal{H}}^2\right]\right.$$
$$+2\left\langle\hat{\mu}_{Y_1|X} - \mu_{Y_1|X}, \mu_{Y_1|X} - \mu_{Y_0|X}\right\rangle_{\mathcal{H}}$$
$$+ 2\left\langle\hat{\mu}_{Y_0|X} - \mu_{Y_0|X}, \mu_{Y_1|X} - \mu_{Y_0|X}\right\rangle_{\mathcal{H}}$$
$$\left.+2\left\langle\hat{\mu}_{Y_1|X} - \mu_{Y_1|X}, \hat{\mu}_{Y_0|X} - \hat{\mu}_{Y_0|X}\right\rangle_{\mathcal{H}}\right|.$$

Here, we have

$$\mathbb{E}\left[\left\|\hat{\mu}_{Y_1|X} - \mu_{Y_1|X}\right\|_{\mathcal{H}}^2\right] \xrightarrow{p} 0 \qquad \text{and} \qquad \mathbb{E}\left[\left\|\hat{\mu}_{Y_0|X} - \mu_{Y_0|X}\right\|_{\mathcal{H}}^2\right] \xrightarrow{p} 0$$

as in the proof of Theorem 4.3.2, so we are done. □

Theorem 4.4.1. *The solution $\hat{F}_0$ to the problem in (4.4) is*

$$\hat{F}_0(x_1, ..., x_r) = \sum_{i_1,...,i_r}^{n_0} k_0(x^0_{i_1}, x_1)...k_0(x^0_{i_r}, x_r)c^0_{i_1,...,i_r}$$

where the coefficients $c^0_{i_1,...,i_r} \in \mathbb{R}$ are the unique solution of the n^r linear equations,

$$\sum_{j_1,...,j_r=1}^{n_0} \left(k_0\left(x^0_{i_1}, x^0_{j_1}\right)...k_0\left(x^0_{i_r}, x^0_{j_r}\right) + \binom{n_0}{r} \lambda^0_{n_0} \delta_{i_1 j_1}...\delta_{i_r j_r} \right) c^0_{j_1,...,j_r} \\ = h\left(y^0_{i_1}, ..., y^0_{i_r}\right).$$

Proof of Theorem 4.4.1. Recall from (4.4) that

$$\hat{F}_0 = \arg\min_{F \in \mathcal{H}^r_0} \left\{ \frac{1}{\binom{n_0}{r}} \sum \left(F\left(x^0_{i_1}, ..., x^0_{i_r}\right) - h\left(y^0_{i_1}, ..., y^0_{i_r}\right)\right)^2 + \lambda^0_{n_0} \|F\|^2_{\mathcal{H}^r_0} \right\},$$

where the summation is over the $\binom{n_0}{r}$ combinations of r distinct elements $\{i_1, ..., i_r\}$ from $1, ..., n_0$. Write

$$\hat{F}'_0\left(x_1, ..., x_r\right) = \sum_{i_1,...,i_r=1}^{n_0} k_0\left(x^0_{i_1}, x_1\right)...k_0\left(x^0_{i_r}, x_r\right) c_{i_1,...,i_r}$$

where the coefficients $c_{i_1,...,i_r} \in \mathbb{R}$ are the unique solution of the n^r linear equations

$$\sum_{j_1,...,j_r=1}^{n_0} \left(k_0\left(x^0_{i_1}, x^0_{j_1}\right)...k_0\left(x^0_{i_r}, x^0_{j_r}\right) + \binom{n_0}{r} \lambda^0_{n_0} \delta_{i_1 j_1}...\delta_{i_r j_r} \right) c_{j_1,...,j_r} \\ = h\left(y^0_{i_1}, ..., y^0_{i_r}\right).$$

Also, for any $F \in \mathcal{H}^r_0$, write $\hat{\mathcal{E}}_{\text{reg}}(F)$ for the empirical regularised least-squares risk of F:

$$\hat{\mathcal{E}}_{\text{reg}}(F) = \frac{1}{\binom{n_0}{r}} \sum \left(F\left(x^0_{i_1}, ..., x^0_{i_r}\right) - h\left(y^0_{i_1}, ..., y^0_{i_r}\right)\right)^2 + \lambda^0_{n_0} \|F\|^2_{\mathcal{H}^r_0},$$

so that $\hat{F}_0 = \arg\min_{F \in \mathcal{H}^r_0} \hat{\mathcal{E}}_{\text{reg}}(F)$. We will show that $\hat{F}'_0 = \hat{F}_0$. For any $F \in \mathcal{H}^r_0$, write $G = F - \hat{F}'_0$. Then

$$\begin{aligned} \hat{\mathcal{E}}_{\text{reg}}(F) &= \frac{1}{\binom{n_0}{r}} \sum \left(F\left(x^0_{i_1}, ..., x^0_{i_r}\right) - h\left(y^0_{i_1}, ..., y^0_{i_r}\right)\right)^2 + \lambda^0_{n_0} \|F\|^2_{\mathcal{H}^r_0} \\ &= \hat{\mathcal{E}}_{\text{reg}}\left(\hat{F}'_0\right) \end{aligned}$$

$$
\begin{aligned}
&\quad + \frac{1}{\binom{n_0}{r}} \sum G\left(x_{i_1}^0, ..., x_{i_r}^0\right)^2 \\
&\quad\quad + \frac{2}{\binom{n_0}{r}} \sum G\left(x_{i_1}^0, ..., x_{i_r}^0\right) \left(\hat{F}_0'\left(x_{i_1}^0, ..., x_{i_r}^0\right) - h\left(y_{i_1}^0, ..., y_{i_r}^0\right)\right) \\
&\quad\quad\quad + \lambda_{n_0}^0 \|G\|_{\mathcal{H}_0^r}^2 + 2\lambda_{n_0}^0 \left\langle G, \hat{F}_0' \right\rangle_{\mathcal{H}_0^r} \\
&\geq \hat{\mathcal{E}}_{\text{reg}}\left(\hat{F}_0'\right) \\
&\quad - \frac{2}{\binom{n_0}{r}} \sum G\left(x_{i_1}^0, ..., x_{i_r}^0\right) \left(h\left(y_{i_1}^0, ..., y_{i_r}^0\right) - \hat{F}_0'\left(x_{i_1}^0, ..., x_{i_r}^0\right)\right) \\
&\quad\quad + 2\lambda_{n_0}^0 \left\langle G, \hat{F}_0' \right\rangle_{\mathcal{H}_0^r} \\
&= \hat{\mathcal{E}}_{\text{reg}}\left(\hat{F}_0'\right) \\
&\quad - 2\lambda_{n_0}^0 \sum G\left(x_{i_1}^0, ..., x_{i_r}^0\right) c_{i_1,...,i_r} \\
&\quad\quad + 2\lambda_{n_0}^0 \sum_{i_1,...,i_r=1}^{n_0} G\left(x_{i_1}^0, ..., x_{i_r}^0\right) c_{i_1,...,i_r} \\
&\quad\quad\quad\quad \text{by the reproducing property and the definition of } c_{i_1,...,i_r} \\
&= \hat{\mathcal{E}}_{\text{reg}}\left(\hat{F}_0'\right)
\end{aligned}
$$

Hence, $\hat{F}_0'$ minimises $\hat{\mathcal{E}}_{\text{reg}}$ in $\mathcal{H}_0^r$, and so $\hat{F}_0' = \hat{F}_0$ as required. □

Theorem 4.4.2. *Suppose k_0^r is a bounded and universal kernel and that $\lambda_{n_0}^0$ decays at a slower rate than $\mathcal{O}(n_0^{-1/2})$. Then as $n_0 \to \infty$,*

$$\mathbb{E}\left[\left(\hat{F}_0\left(X_1, ..., X_r\right) - F_0\left(X_1, ..., X_r\right)\right)^2\right] \xrightarrow{p} 0.$$

Proof of Theorem 4.4.2. Define

$$F_{0,\lambda_{n_0}^0} = \underset{F \in \mathcal{H}_0^r}{\arg\min} \left\{\mathbb{E}\left[\left(F\left(X_1, ..., X_r\right) - F_0\left(X_1, ..., X_r\right)\right)^2\right] + \lambda_{n_0}^0 \|F\|_{\mathcal{H}_0^r}^2\right\}.$$

By the bias-variance decomposition, this also minimises

$$\mathcal{E}_{\lambda_{n_0}^0}(F) = \mathbb{E}\left[\left(F\left(X_1, ..., X_r\right) - h\left(Y_1, ..., Y_r\right)\right)^2\right] + \lambda_{n_0}^0 \|F\|_{\mathcal{H}_0^r}^2.$$

Denote the space of P_X^r-square-integrable $\mathcal{X}^r \to \mathbb{R}$ functions by $L^2(\mathcal{X}^r, P_X^r)$, and define the inclusion operator

$$\iota : \mathcal{H}_0^r \to L^2(\mathcal{X}^r, P_X^r).$$

Then we see that

$$F_{0,\lambda_{n_0}^0} = \underset{F \in \mathcal{H}_0^r}{\arg\min} \left\{\|\iota(F) - F_0\|_2^2 + \lambda_{n_0}^0 \|F\|_{\mathcal{H}_0^r}^2\right\}$$

$$\begin{aligned}
\implies \qquad & 0 = \iota^*(\iota(F_{0,\lambda^0_{n_0}}) - F_0) + \lambda^0_{n_0} F_{0,\lambda^0_{n_0}} \\
\implies \qquad & F_{0,\lambda^0_{n_0}} = \left(\iota^* \circ \iota + \lambda^0_{n_0} I\right)^{-1} \iota^* F_0
\end{aligned}$$

Now, for any $\mathbf{x}^0 = (x^0_1, ..., x^0_{n_0})^T \in \mathcal{X}^{n_0}$, define the sampling operator

$$\begin{aligned}
& S_{\mathbf{x}^0} : \mathcal{H}^r_0 \to \mathbb{R}^{\binom{n_0}{r}}, \\
& \left(S_{\mathbf{x}^0}(F)\right)_{i_1,...,i_r} = \frac{1}{\binom{n_0}{r}} F\left(x^0_{i_1}, ..., x^0_{i_r}\right), \{i_1, ..., i_r\} \subset \{1, ..., n_0\},
\end{aligned}$$

with adjoint

$$S^*_{\mathbf{x}^0}\left(\mathbf{h}\right) = \frac{1}{\binom{n_0}{r}} \sum k_0\left(x^0_{i_1}, \cdot\right) ... k_0\left(x^0_{i_r}, \cdot\right) h_{i_1,...,i_r}, \qquad \mathbf{h} \in \mathbb{R}^{\binom{n_0}{r}};$$

indeed, for any $F \in \mathcal{H}^r_0$ and $\mathbf{h} \in \mathbb{R}^{\binom{n_0}{r}}$,

$$\begin{aligned}
\left\langle S_{\mathbf{x}^0} F, \mathbf{h} \right\rangle_{\mathbb{R}^{\binom{n_0}{r}}} &= \frac{1}{\binom{n_0}{r}} \sum F\left(x^0_{i_1}, ..., x^0_{i_r}\right) h_{i_1,...,i_r} \\
&= \frac{1}{\binom{n_0}{r}} \sum \left\langle F, k_0\left(x^0_{i_1}, \cdot\right) ... k_0\left(x^0_{i_r}, \cdot\right)\right\rangle_{\mathcal{H}^r_0} h_{i_1,...,i_r} \\
&= \left\langle F, \frac{1}{\binom{n_0}{r}} \sum k_0\left(x^0_{i_1}, \cdot\right) ... k_0\left(x^0_{i_r}, \cdot\right) h_{i_1,...,i_r} \right\rangle_{\mathcal{H}^r_0}.
\end{aligned}$$

For $\mathbf{y}^0 \in \mathcal{Y}^{n_0}$, write

$$h\left(\mathbf{y}^0\right) \in \mathbb{R}^{\binom{n_0}{r}}, \qquad h\left(\mathbf{y}^0\right)_{i_1,...,i_r} = h\left(y^0_{i_1}, ..., y^0_{i_r}\right), \{i_1, ..., i_r\} \subset \{1, ..., n_0\}.$$

Then we see that

$$\begin{aligned}
& \hat{F}_0 = \underset{F \in \mathcal{H}^r_0}{\arg\min} \left\{ \binom{n_0}{r} \left\| S_{\mathbf{x}^0}(F) - \frac{1}{\binom{n_0}{r}} h\left(\mathbf{y}^0\right) \right\|^2 + \lambda^0_{n_0} \left\|F\right\|^2_{\mathcal{H}^r_0} \right\} \\
\implies \qquad & 0 = \binom{n_0}{r} S^*_{\mathbf{x}^0} \left(S_{\mathbf{x}^0}\left(\hat{F}_0\right) - \frac{1}{\binom{n_0}{r}} h\left(\mathbf{y}^0\right) \right) + \lambda^0_{n_0} \hat{F}_0 \\
\implies \qquad & \hat{F}_0 = \left(\binom{n_0}{r} S^*_{\mathbf{x}^0} \circ S_{\mathbf{x}^0} + \lambda^0_{n_0} I \right)^{-1} S^*_{\mathbf{x}^0} h\left(\mathbf{y}^0\right).
\end{aligned}$$

We consider the following decomposition:

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{F}_0\left(X_1, ..., X_r\right) - F_0\left(X_1, ..., X_r\right)\right)^2\right] &= \left\| \iota \hat{F}_0 - F_0 \right\|^2_2 \\
&\leq 2 \left\| \iota \hat{F}_0 - \iota F_{0,\lambda^0_{n_0}} \right\|^2_2 \qquad \text{(a)}
\end{aligned}$$

$$+ 2 \left\| \iota F_{0,\lambda^0_{n_0}} - F_0 \right\|_2^2 . \qquad \text{(b)}$$

We are done if we show that the terms (a) and (b) separately converge to 0 (in probability, for (a)).

(a) See that

$$\begin{aligned}\hat{F}_0 - F_{0,\lambda^0_{n_0}} &= \left(\binom{n_0}{r} S^*_{\mathbf{x}^0} \circ S_{\mathbf{x}^0} + \lambda^0_{n_0} I \right)^{-1} S^*_{\mathbf{x}^0} h\left(\mathbf{y}^0\right) - F_{0,\lambda^0_{n_0}} \\ &= \left(\binom{n_0}{r} S^*_{\mathbf{x}^0} \circ S_{\mathbf{x}^0} + \lambda^0_{n_0} I \right)^{-1} \\ &\quad \left(S^*_{\mathbf{x}^0} h\left(\mathbf{y}^0\right) - \binom{n_0}{r} S^*_{\mathbf{x}^0} \circ S_{\mathbf{x}^0} F_{0,\lambda^0_{n_0}} + \iota^* \left(\iota F_{0,\lambda^0_{n_0}} - F_0 \right) \right).\end{aligned}$$

By spectral theorem,

$$\begin{aligned}&\left\| \hat{F}_0 - F_{0,\lambda^0_{n_0}} \right\|_{\mathcal{H}} \\ &\leq \frac{1}{\lambda^0_{n_0}} \left\| S^*_{\mathbf{x}^0} h\left(\mathbf{y}^0\right) - \binom{n_0}{r} S^*_{\mathbf{x}^0} \circ S_{\mathbf{x}^0} F_{0,\lambda^0_{n_0}} + \iota^* \left(\iota F_{0,\lambda^0_{n_0}} - F_0 \right) \right\|_{\mathcal{H}} .\end{aligned}$$

Using this inequality and Chebyshev's inequality, for any $\epsilon > 0$,

$$\begin{aligned}&P\left(\left\| \hat{F}_0 - F_{0,\lambda^0_{n_0}} \right\|_{\mathcal{H}} \geq \epsilon \right) \\ &\leq P\left(\frac{1}{\lambda^0_{n_0}} \left\| S^*_{\mathbf{x}^0} h\left(\mathbf{y}^0\right) - \binom{n_0}{r} S^*_{\mathbf{x}^0} \circ S_{\mathbf{x}^0} F_{0,\lambda^0_{n_0}} - \iota^* \left(F_0 - \iota F_{0,\lambda^0_{n_0}} \right) \right\|_{\mathcal{H}} \right. \\ &\qquad \geq \epsilon) \\ &\leq \frac{1}{(\lambda^0_{n_0})^2 \epsilon^2} \mathbb{E}\left[\left\| S^*_{\mathbf{x}^0} h\left(\mathbf{y}^0\right) - \binom{n_0}{r} S^*_{\mathbf{x}^0} \circ S_{\mathbf{x}^0} F_{0,\lambda^0_{n_0}} \right.\right. \\ &\qquad \left.\left. - \iota^* \left(F_0 - \iota F_{0,\lambda^0_{n_0}} \right) \right\|_{\mathcal{H}}^2 \right] \\ &\leq \frac{1}{(\lambda^0_{n_0})^2 \epsilon^2 \binom{n_0}{r}} \mathbb{E}\left[\left\| k_0\left(x^0_{i_1}, \cdot\right) ... k_0\left(x^0_{i_r}, \cdot\right) \left(h\left(y^0_{i_1}, ..., y^0_{i_r}\right) \right.\right.\right. \\ &\qquad \left.\left.\left. - F_{0,\lambda^0_{n_0}}\left(x^0_{i_1}, ..., x^0_{i_r}\right) \right) \right\|_{\mathcal{H}}^2 \right] \\ &\to 0\end{aligned}$$

as $n \to \infty$, since the kernel is bounded.

(b) Take an arbitrary $\epsilon > 0$. By the denseness of $\mathcal{H}_0^r$ in $L^2(\mathcal{X}^r, P_X^r)$, there exists some $F_\epsilon \in \mathcal{H}_0^r$ with

$$\| \iota F_\epsilon - F_0 \|_2^2 = \mathcal{E}(F_\epsilon) - \mathcal{E}(F_0) \leq \frac{\epsilon}{2}.$$

Then

$$\begin{aligned}\left\|\iota F_{0,\lambda^0_{n_0}} - F_0\right\|_2^2 &= \mathcal{E}\left(F_{0,\lambda^0_{n_0}}\right) - \mathcal{E}\left(F_0\right)\\ &\le \mathcal{E}_{\lambda^0_{n_0}}\left(F_{0,\lambda^0_{n_0}}\right) - \mathcal{E}\left(F_0\right)\\ &= \mathcal{E}_{\lambda^0_{n_0}}\left(F_{0,\lambda^0_{n_0}}\right) - \mathcal{E}_{\lambda^0_{n_0}}(F_\epsilon) + \mathcal{E}_{\lambda^0_{n_0}}(F_\epsilon) - \mathcal{E}(F_\epsilon)\\ &\quad + \mathcal{E}(F_\epsilon) - \mathcal{E}\left(F_0\right)\\ &\le \lambda^0_{n_0}\left\|F_\epsilon\right\|^2_{\mathcal{H}^r_0} + \frac{\epsilon}{2}.\end{aligned}$$

Now let n be large enough for

$$\lambda^0_{n_0}\left\|F_\epsilon\right\|^2_{\mathcal{H}^r_0} \le \frac{\epsilon}{2}$$

to hold.

□

A.5 Proofs for Chapter 5

Lemma 5.1.3. *Suppose Assumption 5.1.2 holds. Then*

(i) For all $f \in \mathcal{H}$,

$$\sup_{x\in\mathcal{X}}\left\|f(x)\right\|_{\mathcal{Y}} \le \sqrt{B}\left\|f\right\|_{\mathcal{H}}.$$

(ii) For all $n \in \mathbb{N}$,

$$\sup_{\mathbf{x}\in\mathcal{X}^n}\left\|S_{\mathbf{x}}\right\|^2_{\mathrm{op}} \le \frac{B}{n}.$$

Proof of Lemma 5.1.3. (i) We use the reproducing property and the Cauchy-Schwarz inequality repeatedly to obtain:

$$\begin{aligned}\left\|f(x)\right\|^2_{\mathcal{Y}} &= \langle f(x), f(x)\rangle_{\mathcal{Y}}\\ &= \langle f, K(\cdot,x)(f(x))\rangle_{\mathcal{H}}\\ &\le \left\|f\right\|_{\mathcal{H}}\langle K(\cdot,x)(f(x)), K(\cdot,x)(f(x))\rangle^{1/2}_{\mathcal{H}}\\ &= \left\|f\right\|_{\mathcal{H}}\langle f(x), K(x,x)(f(x))\rangle^{1/2}_{\mathcal{Y}}\\ &\le \left\|f\right\|_{\mathcal{H}}\left\|f(x)\right\|^{1/2}_{\mathcal{Y}}\left\|K(x,x)\left(f(x)\right)\right\|^{1/2}_{\mathcal{Y}}\\ &\le \left\|f\right\|_{\mathcal{H}}\left\|f(x)\right\|_{\mathcal{Y}}\left\|K(x,x)\right\|^{1/2}_{\mathrm{op}}.\end{aligned}$$

Now divide both sides by $\left\|f(x)\right\|_{\mathcal{Y}}$ and apply the bound in Assumption 5.1.2.

(ii) We can apply (i) to obtain

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{X}^n}\|S_{\mathbf{x}}\|_{\mathrm{op}}^2 &= \sup_{\mathbf{x}\in\mathcal{X}^n}\sup_{f\in\mathcal{H},\|f\|_{\mathcal{H}}\leq 1}\|S_{\mathbf{x}}f\|_{\mathcal{Y}^n}^2\\ &= \sup_{\mathbf{x}\in\mathcal{X}^n}\sup_{f\in\mathcal{H},\|f\|_{\mathcal{H}}\leq 1}\frac{1}{n^2}\sum_{i=1}^n\|f(x_i)\|_{\mathcal{Y}}^2\\ &\leq \frac{B}{n}.\end{aligned}$$

□

Lemma 5.1.4. *We state and prove some results about the inclusion operator and its adjoint.*

(i) An explicit integral expression for $\iota^* : L^2(\mathcal{X}, P_X;\mathcal{Y}) \to \mathcal{H}$ can be given as

$$\iota^*(f)(\cdot) = \mathbb{E}\left[K(\cdot, X)\, f(X)\right] \qquad \textit{for } f \in L^2(\mathcal{X}, P_X;\mathcal{Y}).$$

(ii) For any $f \in L^2(\mathcal{X}, P_X;\mathcal{Y})$ and any $n\in\mathbb{N}$,

$$\iota^*(f) = \mathbb{E}\left[S_{\mathbf{X}}^*\left((f(X_1), ..., f(X_n))^T\right)\right].$$

(iii) For any $f\in\mathcal{H}$ and any $n\in\mathbb{N}$,

$$\iota^*\circ\iota(f) = \mathbb{E}\left[nS_{\mathbf{X}}^*\circ S_{\mathbf{X}}(f)\right].$$

Proof of Lemma 5.1.4. (i) Take any $f_1\in\mathcal{H}$ and $f_2\in L^2(\mathcal{X}, P_X;\mathcal{Y})$. Then the reproducing property gives

$$\begin{aligned}\langle \iota f_1, f_2\rangle_2 &= \mathbb{E}\left[\langle f_1(X), f_2(X)\rangle_{\mathcal{Y}}\right]\\ &= \mathbb{E}\left[\langle f_1, K(\cdot, X)(f_2(X))\rangle_{\mathcal{H}}\right]\\ &= \langle f_1, \mathbb{E}\left[K(\cdot, X)(f_2(X))\right]\rangle_{\mathcal{H}}.\end{aligned}$$

(ii) The fact that $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} X$ and (i) immediately gives

$$\begin{aligned}\mathbb{E}\left[S_{\mathbf{X}}^*\left((f(X_1), ..., f(X_n))^T\right)\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n K(\cdot, X_i)\, f(X_i)\right]\\ &= \mathbb{E}\left[K(\cdot, X)\, f(X)\right]\\ &= \iota^*(f).\end{aligned}$$

(iii) Applying (ii) and the definition of $S_{\mathbf{X}}$,

$$\begin{aligned}\iota^*\circ\iota(f) &= \mathbb{E}\left[S_{\mathbf{X}}^*\left((F(X_1), ..., f(X_n))^T\right)\right]\\ &= \mathbb{E}\left[S_{\mathbf{X}}^*\left(nS_{\mathbf{X}}(f)\right)\right]\\ &= \mathbb{E}\left[nS_{\mathbf{X}}^*\circ S_{\mathbf{X}}(f)\right].\end{aligned}$$

□

Lemma 5.1.6. *We formulate the minimisers in $\mathcal{H}$ of the regularised risks R_λ and $R_{n,\lambda}$ in terms of the inclusion and evaluation operators. Similar results can be found in many places in the literature, for example Micchelli and Pontil (2005, Section 4) or Engl et al. (1996, p.117, Theorem 5.1).*

(i) The minimiser f_λ of the risk R_λ in $\mathcal{H}$ is unique and is given by

$$f_\lambda := \operatorname*{arg\,min}_{f\in\mathcal{H}} R_\lambda(f) = \left(\iota^*\circ\iota + \lambda\mathrm{Id}_\mathcal{H}\right)^{-1}\iota^* f^* = \iota^*\left(\iota\circ\iota^* + \lambda\mathrm{Id}_2\right)^{-1} f^*,$$

where $\mathrm{Id}_\mathcal{H} : \mathcal{H} \to \mathcal{H}$ and $\mathrm{Id}_2 : L^2(\mathcal{X}, P_X; \mathcal{Y}) \to L^2(\mathcal{X}, P_X; \mathcal{Y})$ are the identity operators.

(ii) The minimiser $\hat{f}_{n,\lambda}$ of the risk $R_{n,\lambda}$ in $\mathcal{H}$ is unique and is given by

$$\begin{aligned}\hat{f}_{n,\lambda} &:= \operatorname*{arg\,min}_{f\in\mathcal{H}} R_{n,\lambda}(f)\\ &= \left(nS^*_\mathbf{X}\circ S_\mathbf{X} + \lambda\mathrm{Id}_\mathcal{H}\right)^{-1} S^*_\mathbf{X}\mathbf{Y}\\ &= S^*_\mathbf{X}\left(nS_\mathbf{X}\circ S^*_\mathbf{X} + \lambda\mathrm{Id}_{\mathcal{Y}^n}\right)^{-1}\mathbf{Y},\end{aligned}$$

where $\mathrm{Id}_{\mathcal{Y}^n} : \mathcal{Y}^n \to \mathcal{Y}^n$ is the identity operator.

Proof of Lemma 5.1.6. (i) By Lemma 5.1.5, we have $f_\lambda = \operatorname{arg\,min}_{f\in\mathcal{H}} \tilde{R}_\lambda(f)$, where, for any $f \in \mathcal{H}$,

$$\begin{aligned}\tilde{R}_\lambda(f) &= \mathbb{E}\left[\|f(X) - f^*(X)\|^2_\mathcal{Y}\right] + \lambda\|f\|^2_\mathcal{H}\\ &= \|\iota(f) - f^*\|^2_2 + \lambda\|f\|^2_\mathcal{H}.\end{aligned}$$

Then $\tilde{R}_\lambda$ is clearly continuously Fréchet differentiable, coercive (Precup, 2002, p.105, Definition 7.4) and strictly convex (Precup, 2002, p.105, Definition 7.5). So by (Precup, 2002, p.106, Theorem 7.4), there exists a unique critical point f_λ that minimises $\tilde{R}_\lambda$, and by (Precup, 2002, p.105, Proposition 7.2), at this critical point, we have $\tilde{R}_\lambda(f_\lambda) = 0$. Denote by $J : L^2(\mathcal{X}, L_X; \mathcal{Y}) \to \mathbb{R}$ the map $f \mapsto \|f - f^*\|^2_2$; then we have $J'(f) = 2(f - f^*)$ by (Precup, 2002, p.100, Example 7.2). Taking the Fréchet derivative using (Precup, 2002, p.100, Example 7.3), we have

$$\begin{aligned}\tilde{R}'_\lambda(f) &= \iota^*\circ J'\circ\iota(f) + 2\lambda f\\ &= 2\iota^*\left(\iota(f) - f^*\right) + 2\lambda f\\ \Longrightarrow \qquad \iota^*\left(\iota(f_\lambda) - f^*\right) + \lambda f_\lambda &= 0\\ \Longrightarrow \qquad \left(\iota^*\circ\iota + \lambda\mathrm{Id}_\mathcal{H}\right) f_\lambda &= \iota^* f^*\\ \Longrightarrow \qquad f_\lambda &= \left(\iota^*\circ\iota + \lambda\mathrm{Id}_\mathcal{H}\right)^{-1}\iota^* f^*,\end{aligned}$$

where $(\iota^*\circ\iota + \lambda\mathrm{Id}_\mathcal{H})$ is invertible since $\iota^*\circ\iota$ is positive and self-adjoint, and $\lambda > 0$. Now see that

$$\left(\iota^*\circ\iota + \lambda\mathrm{Id}_\mathcal{H}\right)\iota^*\left(\iota\circ\iota^* + \lambda\mathrm{Id}_2\right)^{-1} f^*$$

$$\begin{aligned}&= \iota^* \left(\iota \circ \iota^* + \lambda \mathrm{Id}_2\right) \left(\iota \circ \iota^* + \lambda \mathrm{Id}_2\right)^{-1} f^* \\ &= \iota^* f^*.\end{aligned}$$

Apply $(\iota^* \circ \iota + \lambda \mathrm{Id}_{\mathcal{H}})^{-1}$ to both sides to obtain

$$f_\lambda = \iota^* \left(\iota \circ \iota^* + \lambda \mathrm{Id}_2\right)^{-1} f^*.$$

(ii) We can write $R_{n,\lambda}(f)$ as

$$\begin{aligned}R_{n,\lambda}(f) &= \frac{1}{n}\sum_{i=1}^{n} \|f(X_i) - Y_i\|_{\mathcal{Y}}^2 + \lambda \|f\|_{\mathcal{H}}^2 \\ &= n\left\|S_{\mathbf{X}}(f) - \frac{1}{n}\mathbf{Y}\right\|_{\mathcal{Y}^n}^2 + \lambda \left\|f\right\|_{\mathcal{H}}^2.\end{aligned}$$

Then following the same steps as in (i), we take the Fréchet derivative of $R_{n,\lambda}$ and set it to 0 at $\hat{f}_{n,\lambda}$:

$$\begin{aligned}R'_{n,\lambda}(f) &= 2nS_{\mathbf{X}}^* \left(S_{\mathbf{X}}(f) - \frac{1}{n}\mathbf{Y}\right) + 2\lambda f \\ \implies \quad nS_{\mathbf{X}}^* \left(S_{\mathbf{X}}(\hat{f}_{n,\lambda}) - \frac{1}{n}\mathbf{Y}\right) + \lambda \hat{f}_{n,\lambda} &= 0 \\ \implies \quad \left(nS_{\mathbf{X}}^* \circ S_{\mathbf{X}} + \lambda \mathrm{Id}_{\mathcal{H}}\right) \hat{f}_{n,\lambda} &= S_{\mathbf{X}}^* \mathbf{Y} \\ \implies \quad \hat{f}_{n,\lambda} &= \left(nS_{\mathbf{X}}^* \circ S_{\mathbf{X}} + \lambda \mathrm{Id}_{\mathcal{H}}\right)^{-1} S_{\mathbf{X}}^* \mathbf{Y},\end{aligned}$$

where $(nS_{\mathbf{X}}^* \circ S_{\mathbf{X}} + \lambda \mathrm{Id}_{\mathcal{H}})$ is invertible since $nS_{\mathbf{X}}^* \circ S_{\mathbf{X}}$ is positive and self-adjoint, and $\lambda > 0$.

By the same argument as in (i), we also have

$$\hat{f}_{n,\lambda} = S_{\mathbf{X}}^* \left(nS_{\mathbf{X}} \circ S_{\mathbf{X}}^* + \lambda \mathrm{Id}_{\mathcal{Y}^n}\right)^{-1} \mathbf{Y}.$$

□

Proposition 5.1.7 (Approximation Error)**.** *If $\iota\mathcal{H}$ is dense in $L^2(\mathcal{X}, P_X; \mathcal{Y})$, then $\|f^* - \iota f_\lambda\|_2^2 \to 0$ as $\lambda \to 0$.*

Proof of Proposition 5.1.7. Take an arbitrary $\epsilon > 0$. By the denseness of $\iota\mathcal{H}$ in $L^2(\mathcal{X}, P_X; \mathcal{Y})$, there exists some $f_\epsilon \in \mathcal{H}$ such that $R(f_\epsilon) - R(f^*) = \|\iota f_\epsilon - f^*\|_2^2 \leq \frac{\epsilon}{2}$. Then see that

$$\begin{aligned}\|f^* - \iota f_\lambda\|_2^2 &= R(f_\lambda) - R(f^*) && \text{by Lemma 5.1.5} \\ &\leq R_\lambda(f_\lambda) - R(f^*) && \text{since } R_\lambda(f) \geq R(f) \forall f \in \mathcal{H} \\ &\leq R_\lambda(f_\epsilon) - R(f_\epsilon) + R(f_\epsilon) - R(f^*) && \text{since } f_\lambda \text{ minimises } R_\lambda \text{ in } \mathcal{H} \\ &\leq R_\lambda(f_\epsilon) - R(f_\epsilon) + \frac{\epsilon}{2} && \text{by the choice of } f_\epsilon\end{aligned}$$

$$= \lambda \|f_\epsilon\|_{\mathcal{H}}^2 + \frac{\epsilon}{2} \qquad \text{by the definition of } R_\lambda.$$

Now if $\lambda \leq \frac{\epsilon}{2\|f_\epsilon\|_{\mathcal{H}}^2}$, then

$$\|f^* - \iota f_\lambda\|_2^2 \leq \epsilon,$$

as required. □

Proposition 5.1.8 (Estimation Error)**.** *Take any $\delta > 0$. Then*

$$P\left(\left\|\hat{f}_{n,\lambda} - f_\lambda\right\|_{\mathcal{H}}^2 \geq \frac{B\mathbb{E}\left[\|Y\|_{\mathcal{Y}}^2\right]}{n\lambda^2\delta}\right) \leq \delta.$$

In particular, if $\lambda = \lambda_n$ depends on n and converges to 0 at a slower rate than $\mathcal{O}(n^{-1/2})$, then

$$\left\|\hat{f}_{n,\lambda_n} - f_{\lambda_n}\right\|_{\mathcal{H}}^2 \xrightarrow{P} 0.$$

Proof of Proposition 5.1.8. By Lemma 5.1.6, we can write

$$\begin{aligned}
\hat{f}_{n,\lambda} - f_\lambda &= (nS_{\mathbf{X}}^* \circ S_{\mathbf{X}} + \lambda \mathrm{Id}_{\mathcal{H}})^{-1} S_{\mathbf{X}}^* \mathbf{Y} \\
&\quad - (nS_{\mathbf{X}}^* \circ S_{\mathbf{X}} + \lambda \mathrm{Id}_{\mathcal{H}})^{-1}(nS_{\mathbf{X}}^* \circ S_{\mathbf{X}} + \lambda \mathrm{Id}_{\mathcal{H}}) f_\lambda \\
&= (nS_{\mathbf{X}}^* \circ S_{\mathbf{X}} + \lambda \mathrm{Id}_{\mathcal{H}})^{-1} \left(S_{\mathbf{X}}^* \mathbf{Y} - nS_{\mathbf{X}}^* \circ S_{\mathbf{X}} f_\lambda - \lambda f_\lambda\right) \\
&= (nS_{\mathbf{X}}^* \circ S_{\mathbf{X}} + \lambda \mathrm{Id}_{\mathcal{H}})^{-1} \left(S_{\mathbf{X}}^* \mathbf{Y} - nS_{\mathbf{X}}^* \circ S_{\mathbf{X}} f_\lambda - \iota^* \left(f^* - \iota f_\lambda\right)\right). \quad (*)
\end{aligned}$$

Write σ for the spectrum of $nS_{\mathbf{X}}^* \circ S_{\mathbf{X}}$. Then by the spectral theorem for (non-compact) self-adjoint operators (Hall, 2013, p.141, Theorem 7.12), there exists a unique projection-valued measure μ on the Borel σ-algebra of σ such that

$$nS_{\mathbf{X}}^* \circ S_{\mathbf{X}} = \int_\sigma \gamma d\mu(\gamma),$$

whence, using the properties of operator-valued integration (Hall, 2013, p.139, Proposition 7.11) and fact that $\sigma \subseteq [0, \infty)$ (Conway, 1990, p.242, Theorem 3.8), we can bound its operator norm by

$$\left\|(nS_{\mathbf{X}}^* \circ S_{\mathbf{X}} + \lambda \mathrm{Id}_{\mathcal{H}})^{-1}\right\|_{\mathrm{op}} = \left\|\int_\sigma \frac{1}{\gamma + \lambda} d\mu(\gamma)\right\|_{\mathrm{op}} \leq \sup_{\gamma \in \sigma} \left|\frac{1}{\gamma + \lambda}\right| \leq \frac{1}{\lambda}.$$

Then returning to (*) and taking the $\mathcal{H}$-norm of both sides, we have

$$\left\|\hat{f}_{n,\lambda} - f_\lambda\right\|_{\mathcal{H}} \leq \frac{1}{\lambda} \left\|S_{\mathbf{X}}^* \mathbf{Y} - nS_{\mathbf{X}}^* \circ S_{\mathbf{X}} f_\lambda - \iota^* \left(f^* - \iota f_\lambda\right)\right\|_{\mathcal{H}}.$$

Hence, for any arbitrary $\epsilon > 0$, by Chebyshev's inequality,

$$P\left(\left\|\hat{f}_{n,\lambda} - f_\lambda\right\|_{\mathcal{H}} \geq \epsilon\right) \leq P\left(\frac{1}{\lambda} \left\|S_{\mathbf{X}}^* \mathbf{Y} - nS_{\mathbf{X}}^* \circ S_{\mathbf{X}} f_\lambda - \iota^* \left(f^* - \iota f_\lambda\right)\right\|_{\mathcal{H}} \geq \epsilon\right)$$

$$\leq \frac{1}{\lambda^2 \epsilon^2} \mathbb{E}\left[\| S_{\mathbf{X}}^* \mathbf{Y} - n S_{\mathbf{X}}^* \circ S_{\mathbf{X}} f_\lambda - \iota^* \left(f^* - \iota f_\lambda \right) \|_{\mathcal{H}}^2 \right].$$

Here, letting $Z = S_X^* Y - S_X^* \circ S_X f_\lambda$ and $Z_i = S_{X_i}^* Y_i - S_{X_i}^* \circ S_{X_i} f_\lambda$, Lemma 5.1.4(ii) and (iii) tells us that the integral is in fact simply $\mathbb{E}[\| \frac{1}{n} \sum_{i=1}^n Z_i - \mathbb{E}[Z] \|_{\mathcal{H}}^2]$. Hence,

$$\begin{aligned} P\left(\left\| \hat{f}_{n,\lambda} - f_\lambda \right\|_{\mathcal{H}} \geq \epsilon \right) &\leq \frac{1}{n \lambda^2 \epsilon^2} \mathbb{E}\left[\| S_X^* Y - S_X^* \circ S_X f_\lambda - \iota^* \left(f^* - \iota f_\lambda \right) \|_{\mathcal{H}}^2 \right] \\ &\leq \frac{1}{n \lambda^2 \epsilon^2} \mathbb{E}\left[\| S_X^* Y - S_X^* \circ S_X f_\lambda \|_{\mathcal{H}}^2 \right] \\ &\leq \frac{B}{n \lambda^2 \epsilon^2} \mathbb{E}\left[\| Y - f_\lambda \left(X \right) \|_{\mathcal{Y}}^2 \right], \end{aligned}$$

by Lemma 5.1.3(ii). Here, we use the fact that f_λ minimises R_λ in $\mathcal{H}$, i.e. $R_\lambda \left(f_\lambda \right) \leq R_\lambda \left(0 \right)$, to see that

$$\mathbb{E}\left[\| f_\lambda(X) - Y \|_{\mathcal{Y}}^2 \right] \leq \mathbb{E}\left[\| f_\lambda(X) - Y \|_{\mathcal{Y}}^2 \right] + \lambda \left\| f_\lambda \right\|_{\mathcal{H}}^2 \leq \mathbb{E}\left[\| Y \|_{\mathcal{Y}}^2 \right].$$

Hence,

$$P\left(\left\| \hat{f}_{n,\lambda} - f_\lambda \right\|_{\mathcal{H}} \geq \epsilon \right) \leq \frac{B \mathbb{E}\left[\| Y \|_{\mathcal{Y}}^2 \right]}{n \lambda^2 \epsilon^2},$$

from which the result follows. □

Theorem 5.1.10. *Suppose $\iota \mathcal{H}$ is dense in $L^2(\mathcal{X}, P_X; \mathcal{Y})$. Suppose that $\lambda = \lambda_n$ depends on the sample size n, and converges to 0 at a slower rate than $\mathcal{O}(n^{-1/2})$. Then we have*

$$R\left(\hat{f}_{n,\lambda_n} \right) - R\left(f^* \right) = \mathbb{E}\left[\left\| \hat{f}_{n,\lambda_n}(X) - f^*(X) \right\|_{\mathcal{Y}}^2 \right] = \left\| \iota \hat{f}_{n,\lambda_n} - f^* \right\|_2^2 \xrightarrow{P} 0.$$

Proof of Theorem 5.1.10. The simple inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ holds in any Hilbert space. Using this, we see that

$$\begin{aligned} \left\| \iota \hat{f}_{n,\lambda_n} - f^* \right\|_2^2 &\leq 2 \left\| \iota \hat{f}_{n,\lambda_n} - \iota f_{\lambda_n} \right\|_2^2 + 2 \left\| \iota f_{\lambda_n} - f^* \right\|_2^2 \\ &\leq 2B \left\| \hat{f}_{n,\lambda_n} - f_{\lambda_n} \right\|_{\mathcal{H}}^2 + 2 \left\| \iota f_{\lambda_n} - f^* \right\|_2^2, \end{aligned}$$

where we used the discussion after Lemma 5.1.3 in the last inequality. Here, the second term converges to 0 as $\lambda_n \to 0$ by Proposition 5.1.7, and the first term converges in probability to 0 by Proposition 5.1.8. Hence,

$$\left\| \iota \hat{f}_{n,\lambda_n} - f^* \right\|_2^2 \xrightarrow{P} 0$$

as required. □

Theorem 5.1.11. *For constants $M, C > 0$, define $\mathcal{P}(M, C)$ as the class of measures such that* *(i)* $\mathbb{E}\left[\|Y\|_{\mathcal{Y}}^2\right] \leq M$, *and*

(ii) $f^* = \iota f_{\mathcal{H}}^*$ *for some* $f_{\mathcal{H}}^* \in \mathcal{H}$ *with* $\|f_{\mathcal{H}}^*\|_{\mathcal{H}}^2 \leq C$.

Let $\mathcal{H}$ *be dense in* $L^2(\mathcal{X}, P_X; \mathcal{Y})$ *for all* $P \in \mathcal{P}(M, C)$. *Then*

$$\sup_{P \in \mathcal{P}(M,C)} P\left(\left\|\iota \hat{f}_{n,\lambda} - f^*\right\|_2^2 \geq \frac{2B^2 M}{n\lambda^2\delta} + 2\lambda C\right) \leq \delta.$$

In particular, if $\lambda = \lambda_n$ *depends on the sample size* n *and converges to 0 at the rate of* $\mathcal{O}(n^{-1/4})$, *then* $R(\hat{f}_{n,\lambda_n}) - R(f^*) = \mathcal{O}_P(n^{-1/4})$ *uniformly over the class* $\mathcal{P}(M, C)$ *of measures.*

Proof of Theorem 5.1.11. First, see that the condition (ii) helps simplify the proof of Proposition 5.1.7:

$$\begin{aligned} \|\iota f_\lambda - f^*\|_2^2 &= R(f_\lambda) - R(f_{\mathcal{H}}^*) \\ &\leq R_\lambda(f_\lambda) - R_\lambda(f_{\mathcal{H}}^*) + R_\lambda(f_{\mathcal{H}}^*) - R(f_{\mathcal{H}}^*) \\ &\leq \lambda \|f_{\mathcal{H}}^*\|_{\mathcal{H}}^2 \\ &\leq \lambda C. \end{aligned} \tag{*}$$

Then using the inequality $\|\iota \hat{f}_{n,\lambda} - f^*\|_2^2 \leq 2B\|\hat{f}_{n,\lambda} - f_\lambda\|_{\mathcal{H}}^2 + 2\|\iota f_\lambda - f^*\|_2^2$ as in the proof of Theorem 5.1.10,

$$\begin{aligned} &\sup_{P \in \mathcal{P}(M,C)} P\left(\left\|\iota \hat{f}_{n,\lambda} - f^*\right\|_2^2 > \frac{2B^2 M}{n\lambda^2\delta} + 2\lambda C\right) \\ &\leq \sup_{P \in \mathcal{P}(M,C)} P\left(\left\|\hat{f}_{n,\lambda} - f_\lambda\right\|_{\mathcal{H}}^2 > \frac{BM}{n\lambda^2\delta}\right) \\ &\quad + \sup_{P \in \mathcal{P}(M,C)} P\left(\|f^* - \iota f_\lambda\|_2^2 > \lambda C\right) \\ &\leq \sup_{P \in \mathcal{P}(M,C)} P\left(\left\|\hat{f}_{n,\lambda} - f_\lambda\right\|_{\mathcal{H}}^2 \geq \frac{B\mathbb{E}\left[\|Y\|_{\mathcal{Y}}^2\right]}{n\lambda^2\delta}\right) && \text{by } (*) \\ &\leq \delta && \text{by Proposition 5.1.8,} \end{aligned}$$

as required. □

Lemma 5.2.3. *We have*

$$\mathbb{E}\left[\|P_n - P\|_{\mathcal{G}}\right] \leq \mathbb{E}\left[\|P_n - P_n'\|_{\mathcal{G}}\right].$$

Proof of Lemma 5.2.3. Denote by $\mathcal{F}_n$ the σ-algebra generated by $X_1, ..., X_n$. Then for each $g \in \mathcal{G}$, we have

$$\mathbb{E}\left[P_n g \mid \mathcal{F}_n\right] = P_n g \qquad \text{and} \qquad \mathbb{E}\left[P_n' g \mid \mathcal{F}_n\right] = Pg,$$

and so

$$(P_n - P)g = \mathbb{E}\left[(P_n - P'_n)\, g \mid \mathcal{F}_n\right].$$

Now see that

$$\begin{aligned}\|P_n - P\|_{\mathcal{G}} &= \sup_{g\in\mathcal{G}} \|\mathbb{E}\left[(P_n - P'_n)\, g \mid \mathcal{F}_n\right]\|_{\mathcal{Y}} \\ &\leq \sup_{g\in\mathcal{G}} \mathbb{E}\left[\|(P_n - P'_n)\, g\|_{\mathcal{Y}} \mid \mathcal{F}_n\right] \quad \text{by Jensen's inequality} \\ &\leq \mathbb{E}\left[\sup_{g\in\mathcal{G}} \|(P_n - P'_n)\, g\|_{\mathcal{Y}} \mid \mathcal{F}_n\right].\end{aligned}$$

Now take expectations on both sides and apply the law of iterated expectations arrive at the result. □

Lemma 5.2.4 (Symmetrisation with means)**.** *We have*

$$\mathbb{E}\left[\|P_n - P\|_{\mathcal{G}}\right] \leq 2\mathbb{E}\left[\|P_n^{\sigma}\|_{\mathcal{G}}\right]$$

Proof of Lemma 5.2.4. Note that $\|P_n - P'_n\|_{\mathcal{G}}$ has the same distribution as $\|P_n^{\sigma} - P_n'^{\sigma}\|_{\mathcal{G}}$, since, for each $i = 1, ..., n$ and $g \in \mathcal{G}$. $g(X_i) - g(X'_i)$ and $\sigma_i\left(g(X_i) - g(X'_i)\right)$ have the same distribution. Hence, the triangle inequality gives us

$$\mathbb{E}\left[\|P_n - P'_n\|_{\mathcal{G}}\right] = \mathbb{E}\left[\|P_n^{\sigma} - P_n'^{\sigma}\|_{\mathcal{G}}\right] \leq \mathbb{E}\left[\|P_n^{\sigma}\|_{\mathcal{G}} + \|P_n'^{\sigma}\|_{\mathcal{G}}\right] = 2\mathbb{E}\left[\|P_n^{\sigma}\|_{\mathcal{G}}\right].$$

Now apply Lemma 5.2.3. □

Lemma 5.2.5 (Symmetrisation with probabilities)**.** *Let $a > 0$. Suppose that for all $g \in \mathcal{G}$,*

$$\mathbb{P}\left(\|(P_n - P)\, g\|_{\mathcal{Y}} > \frac{a}{2}\right) \leq \frac{1}{2}.$$

Then

$$\mathbb{P}\left(\|P_n - P\|_{\mathcal{G}} > a\right) \leq 4\mathbb{P}\left(\|P_n^{\sigma}\|_{\mathcal{G}} > \frac{a}{4}\right).$$

Proof of Lemma 5.2.5. Denote again by $\mathcal{F}_n$ the σ-algebra generated by $X_1, ..., X_n$. If we have $\|P_n - P\|_{\mathcal{G}} > a$, then we know that for some random function g_* depending on $X_1, ..., X_n$, $\|(P_n - P)\, g_*\|_{\mathcal{Y}} > a$. Because $X'_1, ..., X'_n$ are independent of $\mathcal{F}_n$,

$$\mathbb{P}\left(\|(P'_n - P)\, g_*\|_{\mathcal{Y}} > \frac{a}{2} \mid \mathcal{F}_n\right) = \mathbb{P}\left(\|(P_n - P)\, g_*\|_{\mathcal{Y}} > \frac{a}{2}\right) \leq \frac{1}{2}. \qquad (*)$$

Then see that,

$$\begin{aligned}&\mathbb{P}\left(\|P_n - P\|_{\mathcal{G}} > a\right) \\ &\leq \mathbb{P}\left(\|(P_n - P)\, g_*\|_{\mathcal{Y}} > a\right) \\ &= \mathbb{E}\left[\mathbf{1}\left\{\|(P_n - P)\, g_*\|_{\mathcal{Y}} > a\right\}\right] \\ &\leq 2\mathbb{E}\left[\mathbb{P}\left(\|(P'_n - P)\, g_*\|_{\mathcal{Y}} \leq \frac{a}{2} \mid \mathcal{F}_n\right)\mathbf{1}\left\{\|(P_n - P)\, g_*\|_{\mathcal{Y}} > a\right\}\right] \text{ by } (*)\end{aligned}$$

$$\begin{aligned}
&= 2\mathbb{E}\left[\mathbb{P}\left(\|(P_n' - P)\, g_*\|_{\mathcal{Y}} \leq \frac{a}{2} \text{ and } \|(P_n - P)\, g_*\|_{\mathcal{Y}} > a \mid \mathcal{F}_n\right)\right] \\
&= 2\mathbb{P}\left(\|(P_n' - P)\, g_*\|_{\mathcal{Y}} \leq \frac{a}{2} \text{ and } \|(P_n - P)\, g_*\|_{\mathcal{Y}} > a\right).
\end{aligned}$$

But if the two inequalities in the probability on the last line hold, then the reverse triangle inequality gives us

$$\frac{a}{2} < \|(P_n - P)\, g_*\|_{\mathcal{Y}} - \|(P_n' - P)\, g_*\|_{\mathcal{Y}} \leq \|(P_n - P_n')\, g_*\|_{\mathcal{Y}},$$

so

$$\begin{aligned}
\mathbb{P}\left(\|P_n - P\|_{\mathcal{G}} > a\right) &\leq 2\mathbb{P}\left(\|(P_n - P_n')\, g_*\|_{\mathcal{Y}} > \frac{a}{2}\right) \\
&\leq 2\mathbb{P}\left(\|P_n - P_n'\|_{\mathcal{G}} > \frac{a}{2}\right) \\
&= 2\mathbb{P}\left(\|P_n^\sigma - P_n'^\sigma\|_{\mathcal{G}} > \frac{a}{2}\right) \\
&\leq 2\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{G}} > \frac{a}{4} \text{ or } \|P_n'^\sigma\|_{\mathcal{G}} > \frac{a}{4}\right) \\
&\leq 4\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{G}} > \frac{a}{4}\right).
\end{aligned}$$

□

Lemma 5.2.6. *Let $\mathcal{G} = \{g_1, ..., g_N\} \in L^1(\mathcal{X}, P; \mathcal{Y})$ be a finite class of functions with cardinality $N > 1$. Then we have*

$$\|P_n - P\|_{\mathcal{G}} \to 0.$$

Proof of Lemma 5.2.6. Take any $K > 0$. Define the function $G : \mathcal{X} \to \mathbb{R}$ by

$$G(x) = \max_{1 \leq j \leq N} \|g_j(x)\|_{\mathcal{Y}}.$$

Since each $\|g_j\|_{\mathcal{Y}}$ is integrable, and we have a finite collection, G is also integrable. Then, for each $j = 1, ..., N$, define the function $\tilde{g}_j : \mathcal{X} \to \mathcal{Y}$ by $\tilde{g}_j = g_j \mathbf{1}\{G \leq K\}$. Then for all $i = 1, ..., n$, letting σ_i be independent Rademacher variables again, we have

$$\mathbb{E}\left[\sigma_i \tilde{g}_j(X_i)\right] = 0 \qquad \text{and} \qquad \|\sigma_i \tilde{g}_j(X_i)\|_{\mathcal{Y}} \leq K \text{ almost surely.}$$

Hence, for each $j = 1, ..., N$, by Hoeffding's inequality, for any $t > 0$, we have

$$\mathbb{P}\left(\|P_n^\sigma \tilde{g}_j\|_{\mathcal{Y}} \geq 2K\sqrt{\frac{t}{n}}\right) = \mathbb{P}\left(\left\|\sum_{i=1}^n \sigma_i \tilde{g}_j(X_i)\right\|_{\mathcal{Y}} \geq 2K\sqrt{nt}\right) \leq 2e^{-t}.$$

By the union bound, for any $t > 0$, we have

$$\mathbb{P}\left(\max_{1 \leq j \leq N} \|P_n^\sigma \tilde{g}_j\|_{\mathcal{Y}} \geq 2K\sqrt{\frac{t + \log N}{n}}\right)$$

$$\begin{aligned}
&\le N \max_{1\le j\le N} \mathbb{P}\left(\|P_n^\sigma \tilde{g}_j\|_{\mathcal{Y}} \ge 2K\sqrt{\frac{t+\log N}{n}} \right)\\
&\le 2e^{-t}.
\end{aligned}$$

Now see that, for each $j = 1, ..., N$, Chebyshev's inequality gives

$$\begin{aligned}
\mathbb{P}\left(\|(P_n - P)\,\tilde{g}_j\|_{\mathcal{Y}} > 4K\sqrt{\frac{t+\log N}{n}} \right) &\le \frac{n\mathbb{E}\left[\|(P_n-P)\,\tilde{g}_j\|_{\mathcal{Y}}^2\right]}{16K^2\,(t+\log N)}\\
&\le \frac{1}{16\,(t+\log N)}\\
&\le \frac{1}{2},
\end{aligned}$$

where the last inequality follows since $8t + 8\log N \ge 8\log 2 \ge 1$. Now apply Lemma 5.2.5 to see that

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\le j\le N} \|(P_n - P)\,\tilde{g}_j\|_{\mathcal{Y}} > 8K\sqrt{\frac{t+\log N}{n}} \right)\\
&\le 4\mathbb{P}\left(\max_{1\le j\le N} \|P_n^\sigma \tilde{g}_j\|_{\mathcal{Y}} > 2K\sqrt{\frac{t+\log N}{n}} \right)\\
&\le 8e^{-t}.
\end{aligned}$$

This tells us that

$$\max_{1\le j\le N} \|(P_n - P)\,\tilde{g}_j\|_{\mathcal{Y}} \xrightarrow{P} 0.$$

Finally, see that

$$\|P_n - P\|_{\mathcal{G}} \le \max_{1\le j\le N} \|(P_n-P)\,\tilde{g}_j\|_{\mathcal{Y}} + \max_{1\le j\le N} \|(P_n-P)\,g_j\mathbf{1}\,\{G>K\}\|_{\mathcal{Y}}\,.$$

Here, the first term converges to 0 in probability for any $K > 0$, as shown above, and the second term decomposes as

$$\begin{aligned}
\max_{1\le j\le N} \|(P_n-P)\,g_j\mathbf{1}\,\{G>K\}\|_{\mathcal{Y}} &\le (P_n+P)\,G\mathbf{1}\,\{G>K\}\\
&= (P_n-P)\,G\mathbf{1}\,\{G>K\} + 2PG\mathbf{1}\,\{G>K\}\\
&\le (P_n-P)\,G + 2PG\mathbf{1}\,\{G>K\}\,.
\end{aligned}$$

Here, the first term converges to 0 in probability by the weak law of large numbers, and the second term converges to 0 as $K \to \infty$, by Çınlar (2011, p.71, Lemma 3.10). □

Theorem 5.2.8. *Suppose that*

$$G \in L^1(\mathcal{X}, P; \mathbb{R}) \qquad \textit{and} \qquad \frac{1}{n}H(\delta, \mathcal{G}, \|\cdot\|_{1,P_n}) \xrightarrow{P} 0 \;\; \textit{for each } \delta > 0.$$

Then $\mathcal{G}$ is a Glivenko Cantelli class, i.e. $\|P_n - P\|_{\mathcal{G}} \xrightarrow{P} 0.$

Proof of Theorem 5.2.8. Take any $K > 0$ and $\delta > 0$. Denote again by $\mathcal{F}_n$ the σ-algebra generated by $X_1, ..., X_n$, and define $\mathcal{G}_K = \{g\mathbf{1}\{G \leq K\} : g \in \mathcal{G}\}$. Let $g_1, ..., g_N$, with $N = N(\delta, \mathcal{G}, \|\cdot\|_{1,P_n})$, be a minimal δ-covering of $\mathcal{G}$. Then N is a random variable, that is measurable with respect to $\mathcal{F}_n$. Moreover, writing $\tilde{g}_j = g_j\mathbf{1}\{G \leq K\}$ for each $j = 1, ..., N$, $\tilde{g}_1, ..., \tilde{g}_N$ form a δ-covering of $\mathcal{G}_K$, since, for any $\tilde{g} = g\mathbf{1}\{G \leq K\} \in \mathcal{G}_K$ for $g \in \mathcal{G}$, there exists $j \in \{1, ..., N\}$ with $\|g - g_j\|_{1,P_n} \leq \delta$, so $\|\tilde{g} - \tilde{g}_j\|_{1,P_n} \leq \|g - g_j\|_{1,P_n} \leq \delta$.

Note that, when $\|\tilde{g} - \tilde{g}_j\|_{1,P_n} = P_n \|\tilde{g} - \tilde{g}_j\|_{\mathcal{Y}} \leq \delta$, we have

$$\begin{aligned}\|P_n^\sigma \tilde{g}\|_{\mathcal{Y}} &\leq \|P_n^\sigma \tilde{g}_j\|_{\mathcal{Y}} + \|P_n^\sigma \tilde{g} - P_n^\sigma \tilde{g}_j\|_{\mathcal{Y}} \\ &\leq \|P_n^\sigma \tilde{g}_j\|_{\mathcal{Y}} + P_n \|\tilde{g} - \tilde{g}_j\|_{\mathcal{Y}} \\ &\leq \|P_n^\sigma \tilde{g}_j\|_{\mathcal{Y}} + \delta.\end{aligned}$$

So for any $\tilde{g} \in \mathcal{G}_K$,

$$\|P_n^\sigma \tilde{g}\|_{\mathcal{Y}} \leq \max_{1\leq j\leq N} \|P_n^\sigma \tilde{g}_j\|_{\mathcal{Y}} + \delta. \tag{*}$$

By Hoeffding's inequality and union bound (as in the proof of Lemma 5.2.6, since N is measurable with respect to $\mathcal{F}_n$), for any $t > 0$, we have

$$\mathbb{P}\left(\max_{1\leq j\leq N} \|P_n^\sigma \tilde{g}_j\|_{\mathcal{Y}} \geq 2K\sqrt{\frac{t + \log N}{n}} \mid \mathcal{F}_n\right) \leq 2e^{-t}.$$

We then apply (*) and integrate both sides (to remove the conditioning on $\mathcal{F}_n$) to see that, for any $t > 0$,

$$\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{G}_K} \geq \delta + 2K\sqrt{\frac{t + \log N}{n}}\right) \leq 2e^{-t}.$$

Then see that, using the elementary inequality $\sqrt{a} + \sqrt{b} \geq \sqrt{a+b}$,

$$\begin{aligned}&\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{G}_K} \geq 2\delta + 2K\sqrt{\frac{t}{n}}\right) \\ &\leq \mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{G}_K} \geq \delta + 2K\sqrt{\frac{t}{n}} + 2K\sqrt{\frac{\log N}{n}}\right) + \mathbb{P}\left(2K\sqrt{\frac{\log N}{n}} \geq \delta\right) \\ &\leq \mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{G}_K} \geq \delta + 2K\sqrt{\frac{t + \log N}{n}}\right) + \mathbb{P}\left(2K\sqrt{\frac{1}{n}H(\delta, \mathcal{G}, \|\cdot\|_{1,P_n})} \geq \delta\right) \\ &\leq 2e^{-t} + \mathbb{P}\left(2K\sqrt{\frac{1}{n}H(\delta, \mathcal{G}, \|\cdot\|_{1,P_n})} \geq \delta\right).\end{aligned}$$

Also, by Chebyshev's inequality, for each $\tilde{g} \in \mathcal{G}_K$, we have, for any $t \geq \frac{1}{8}$

$$\mathbb{P}\left(\|(P_n - P)\,\tilde{g}\|_{\mathcal{Y}} > 4\delta + 4K\sqrt{\frac{t}{n}}\right) \leq \mathbb{P}\left(\|(P_n - P)\,\tilde{g}\|_{\mathcal{Y}} > 4K\sqrt{\frac{t}{n}}\right)$$

$$
\begin{aligned}
&\leq \frac{n\mathbb{E}\left[\left\|\left(P_n - P\right)\tilde{g}\right\|_{\mathcal{Y}}^2\right]}{16K^2 t} \\
&\leq \frac{1}{16t} \\
&\leq \frac{1}{2}.
\end{aligned}
$$

Hence, we can apply symmetrisation with probabilities again (Lemma 5.2.5) to see that, for any $t \geq \frac{1}{8}$,

$$
\begin{aligned}
\mathbb{P}\left(\left\|P_n - P\right\|_{\mathcal{G}_K} \geq 8\delta + 8K\sqrt{\frac{t}{n}}\right) &\leq 4\mathbb{P}\left(\left\|P_n^\sigma\right\|_{\mathcal{G}_K} \geq 2\delta + 2K\sqrt{\frac{t}{n}}\right) \\
&\leq 2^{-t} + \mathbb{P}\left(2K\sqrt{\frac{1}{n}H(\delta, \mathcal{G}, \|\cdot\|_{1,P_n})} \geq \delta\right).
\end{aligned}
$$

Here, since $\delta > 0$ was arbitrary and $\frac{1}{n}H(\delta, \mathcal{G}, \|\cdot\|_{1,P_n}) \xrightarrow{P} 0$ by hypothesis, we have that $\mathcal{G}_K$ is Glivenko Cantelli.

Finally, see that

$$
\left\|P_n - P\right\|_{\mathcal{G}} \leq \sup_{\tilde{g}\in\mathcal{G}_K} \left\|\left(P_n - P\right)\tilde{g}\right\|_{\mathcal{Y}} + \sup_{g\in\mathcal{G}} \left\|\left(P_n - P\right) g\mathbf{1}\left\{G > K\right\}\right\|_{\mathcal{Y}}.
$$

Here, the first term converges to 0 in probability for any $K > 0$, as shown above, and the second term decomposes as

$$
\begin{aligned}
\sup_{g\in\mathcal{G}} \left\|\left(P_n - P\right) g\mathbf{1}\left\{G > K\right\}\right\|_{\mathcal{Y}} &\leq \left(P_n + P\right) G\mathbf{1}\left\{G > K\right\} \\
&= \left(P_n - P\right) G\mathbf{1}\left\{G > K\right\} + 2PG\mathbf{1}\left\{G > K\right\} \\
&\leq \left(P_n - P\right) G + 2PG\mathbf{1}\left\{G > K\right\}.
\end{aligned}
$$

Here, the first term converges to 0 in probability by the weak law of large numbers, and the second term converges to 0 as $K \to \infty$, by Çınlar (2011, p.71, Lemma 3.10), since G is integrable by hypothesis. □

Proposition 5.2.9 (Chaining)**.** *We fix $S \in \mathbb{N}$. Define*

$$
J_n := \sum_{s=0}^{S} 2^{-s} R_n \sqrt{2H_{s+1}}.
$$

(i) For all $t > 0$,

$$
\mathbb{P}\left(\sup_{g\in\mathcal{G}} \left\|\sum_{s=0}^{S} P_n^\sigma\left(g^{s+1} - g^s\right)\right\|_{\mathcal{Y}} \geq \frac{\sqrt{2}J_n}{\sqrt{n}} + 6R_n\sqrt{\frac{1+t}{n}} \mid \mathcal{F}_n\right) \leq 2e^{-t}.
$$

(ii) Suppose that $\varepsilon_1, ..., \varepsilon_n$ are i.i.d. Gaussian random variables in $\mathcal{Y}$ with mean 0 and covariance operator Q. Without loss of generality (by rescaling

if necessary), assume $\mathrm{Tr}Q = 1$. *For each* $g \in \mathcal{G}$, *we can consider the following inner product:*

$$\langle \varepsilon, g \rangle_{2,P_n} = \frac{1}{n} \sum_{i=1}^{n} \langle \varepsilon_i, g(X_i) \rangle_{\mathcal{Y}}.$$

Then for all $t > 0$,

$$\mathbb{P}\left(\sup_{g \in \mathcal{G}} \sum_{s=0}^{S} \left\langle \varepsilon, g^{s+1} - g^{s} \right\rangle_{2,P_n} \geq \frac{J_n}{\sqrt{n}} + 4R_n \sqrt{\frac{1+t}{n}} \mid \mathcal{F}_n \right) \leq e^{-t}.$$

Proof of Proposition 5.2.9. (i) Fix $s \in \{0, 1, ..., S\}$ and $k \in \{1, ..., N_{s+1}\}$. Denote

$$g_k^{s+1,s} = \arg\min_{\{g_j^s\}_{j=1}^{N_s}} \left\{ \left\| g_k^{s+1} - g_j^s \right\|_{2,P_n} \right\}.$$

Then

$$\left\| P_n^{\sigma} \left(g_k^{s+1} - g_k^{s+1,s} \right) \right\|_{\mathcal{Y}} \leq \frac{1}{n} \sum_{i=1}^{n} \left\| g_k^{s+1}(X_i) - g_k^{s+1,s}(X_i) \right\|_{\mathcal{Y}},$$

where

$$\sqrt{\sum_{i=1}^{n} \left\| g_k^{s+1}(X_i) - g_k^{s+1,s}(X_i) \right\|_{\mathcal{Y}}^2} = \sqrt{n} \left\| g_k^{s+1} - g_k^{s+1,s} \right\|_{2,P_n} \leq \sqrt{n} 2^{-s} R_n,$$

since the $\{g_j^s\}_{j=1}^{N_s}$ form a $2^{-s}R_n$-covering of $(\mathcal{G}, \|\cdot\|_{2,P_n})$. Hence, noting that R_n is measurable with respect to $\mathcal{F}_n$, Hoeffding's inequality gives, for any $t > 0$,

$$\mathbb{P}\left(\left\| P_n^{\sigma} \left(g_k^{s+1} - g_k^{s+1,s} \right) \right\|_{\mathcal{Y}} \geq 2^{-(s-1)} R_n \sqrt{\frac{t}{n}} \mid \mathcal{F}_n \right) \leq 2e^{-t}.$$

Therefore (by the union bound), for each $s = 0, 1, ..., S$ and all $t > 0$,

$$\mathbb{P}\left(\max_{k \in \{1,...,N_{s+1}\}} \left\| P_n^{\sigma} \left(g_k^{s+1} - g_k^{s+1,s} \right) \right\|_{\mathcal{Y}} \geq 2^{-(s-1)} R_n \sqrt{\frac{H_{s+1} + t}{n}} \mid \mathcal{F}_n \right) \\ \leq 2e^{-t}.$$

Fix t and for $s = 0, 1, ..., S$, let

$$\begin{aligned} \alpha_s :&= 2^{-(s-1)} R_n \left(\sqrt{H_{s+1}} + \sqrt{(1+s)(1+t)} \right) \\ &\geq 2^{-(s-1)} R_n \left(\sqrt{H_{s+1} + (1+s)(1+t)} \right), \end{aligned}$$

using $\sqrt{a}+\sqrt{b} \geq \sqrt{a+b}$. Then using $\sum_{s=0}^{S} 2^{-(s-1)}\sqrt{1+s} \leq 6$,

$$\begin{aligned}\sum_{s=0}^{S} \alpha_s &= \sqrt{2}J_n + \sum_{s=0}^{S} 2^{-(s-1)} R_n \sqrt{(1+s)(1+t)} \\ &\leq \sqrt{2}J_n + 6R_n\sqrt{1+t}.\end{aligned}$$

Therefore

$$\begin{aligned}&\mathbb{P}\left(\sup_{g\in\mathcal{G}}\left\|\sum_{s=0}^{S} P_n^{\sigma}\left(g^{s+1}-g^{s}\right)\right\|_{\mathcal{Y}} \geq \frac{\sqrt{2}J_n}{\sqrt{n}} + 6R_n\sqrt{\frac{1+t}{n}} \mid \mathcal{F}_n\right) \\ &\leq \mathbb{P}\left(\sup_{g\in\mathcal{G}}\left\|\sum_{s=0}^{S} P_n^{\sigma}\left(g^{s+1}-g^{s}\right)\right\|_{\mathcal{Y}} \geq \frac{1}{\sqrt{n}}\sum_{s=0}^{S}\alpha_s \mid \mathcal{F}_n\right) \\ &\leq \mathbb{P}\left(\sum_{s=0}^{S}\sup_{g\in\mathcal{G}}\left\|P_n^{\sigma}\left(g^{s+1}-g^{s}\right)\right\|_{\mathcal{Y}} \geq \frac{1}{\sqrt{n}}\sum_{s=0}^{S}\alpha_s \mid \mathcal{F}_n\right) \\ &\leq \sum_{s=0}^{S}\mathbb{P}\left(\sup_{g\in\mathcal{G}}\left\|P_n^{\sigma}\left(g^{s+1}-g^{s}\right)\right\|_{\mathcal{Y}} \geq \frac{1}{\sqrt{n}}\alpha_s \mid \mathcal{F}_n\right) \\ &= \sum_{s=0}^{S}\mathbb{P}\left(\max_{k=1,...,N_{s+1}}\left\|P_n^{\sigma}\left(g_k^{s+1}-g_k^{s+1,s}\right)\right\|_{\mathcal{Y}} \geq \frac{1}{\sqrt{n}}\alpha_s \mid \mathcal{F}_n\right) \\ &\leq 2\sum_{s=0}^{S} e^{-(1+s)(1+t)} \\ &\leq 2e^{-t}.\end{aligned}$$

(ii) Fix $s \in \{0,1,...,S\}$ and $k \in \{1,...,N_{s+1}\}$. Denote

$$g_k^{s+1,s} = \underset{\{g_j^s\}_{j=1}^{N_s}}{\arg\min}\left\{\left\|g_k^{s+1}-g_j^{s}\right\|_{2,P_n}\right\}.$$

Let $\lambda > 0$ be arbitrary. Then Markov's inequality gives us, for any $t > 0$,

$$\begin{aligned}&\mathbb{P}\left(\left\langle \varepsilon, g_k^{s+1}-g_k^{s+1,s}\right\rangle_{2,P_n} \geq 2^{-s}R_n\sqrt{\frac{2t}{n}} \mid \mathcal{F}_n\right) \\ &\leq e^{-\lambda 2^{-s}R_n\sqrt{\frac{2t}{n}}}\mathbb{E}\left[e^{\frac{\lambda}{n}\sum_{i=1}^{n}\left\langle \varepsilon_i, g_k^{s+1}(X_i)-g_k^{s+1,s}(X_i)\right\rangle_{\mathcal{Y}}} \mid \mathcal{F}_n\right] \\ &= e^{-\lambda 2^{-s}R_n\sqrt{\frac{2t}{n}}}\prod_{i=1}^{n}\mathbb{E}\left[e^{\frac{\lambda}{n}\left\langle \varepsilon_i, g_k^{s+1}(X_i)-g_k^{s+1,s}(X_i)\right\rangle_{\mathcal{Y}}} \mid \mathcal{F}_n\right].\end{aligned}$$

Here, since ε_i is a $\mathcal{Y}$-valued Gaussian random variable with mean 0 and covariance operator Q for each $i = 1,...,n$, the distribution of the real variable $\frac{\lambda}{n}\left\langle \varepsilon_i, g_k^{s+1}(X_i) - g_k^{s+1,s}(X_i)\right\rangle_{\mathcal{Y}}$ conditioned on $\mathcal{F}_n$ is real Gaussian with mean 0 and variance

$$\begin{aligned}&\frac{\lambda^2}{n^2}\mathbb{E}\left[\left\langle g_k^{s+1}(X_i) - g_k^{s+1,s}(X_i), \varepsilon_i\right\rangle_{\mathcal{Y}}^2 \mid \mathcal{F}_n\right]\\ &\quad \leq \frac{\lambda^2}{n^2}\left\|g_k^{s+1}(X_i) - g_k^{s+1,s}(X_i)\right\|_{\mathcal{Y}}^2,\end{aligned}$$

which follows from the Cauchy-Schwarz inequality and the fact that we have $\mathbb{E}\left[\|\varepsilon_i\|_{\mathcal{Y}}^2\right] = \mathrm{Tr}Q = 1$. Hence,

$$\begin{aligned}&\mathbb{P}\left(\left\langle \varepsilon, g_k^{s+1} - g_k^{s+1,s}\right\rangle_{2,P_n} \geq 2^{-s}R_n\sqrt{\frac{2t}{n}} \mid \mathcal{F}_n\right)\\ &\leq e^{-\lambda 2^{-s}R_n\sqrt{\frac{2t}{n}}}\prod_{i=1}^n e^{\frac{\lambda^2}{2n^2}\left\|g_k^{s+1}(X_i)-g_k^{s+1,s}(X_i)\right\|_{\mathcal{Y}}^2}\\ &= e^{-\lambda 2^{-s}R_n\sqrt{\frac{2t}{n}}} e^{\frac{\lambda^2}{2n^2}\sum_{i=1}^n\left\|g_k^{s+1}(X_i)-g_k^{s+1,s}(X_i)\right\|_{\mathcal{Y}}^2}\\ &= e^{-\lambda 2^{-s}R_n\sqrt{\frac{2t}{n}}} e^{\frac{\lambda^2}{2n}\left\|g_k^{s+1}-g_k^{s+1,s}\right\|_{2,P_n}^2}\\ &\leq e^{-\lambda 2^{-s}R_n\sqrt{\frac{2t}{n}}} e^{\frac{\lambda^2}{2n}\left(2^{-s}R_n\right)^2}.\end{aligned}$$

Now let $\lambda = \frac{\sqrt{2nt}}{2^{-s}R_n}$ to see that

$$\mathbb{P}\left(\left\langle \varepsilon, g_k^{s+1} - g_k^{s+1,s}\right\rangle_{2,P_n} \geq 2^{-s}R_n\sqrt{\frac{2t}{n}} \mid \mathcal{F}_n\right) \leq e^{-t}.$$

Therefore, by the union bound, for each $s = 0, 1, ..., S$ and all $t > 0$,

$$\begin{aligned}&\mathbb{P}\left(\max_{k\in\{1,...,N_{s+1}\}}\left\langle \varepsilon, g_k^{s+1} - g_k^{s+1,s}\right\rangle_{2,P_n} \geq 2^{-s}R_n\sqrt{\frac{2(t+H_{s+1})}{n}} \mid \mathcal{F}_n\right)\\ &\leq e^{-t}.\end{aligned}$$

Fix t and for $s = 0, 1, ..., S$, let

$$\begin{aligned}\alpha_s &:= 2^{-s}R_n\left(\sqrt{2H_{s+1}} + \sqrt{2(1+s)(1+t)}\right)\\ &\geq 2^{-s}R_n\sqrt{2\left(H_{s+1} + (1+s)(1+t)\right)}\end{aligned}$$

using $\sqrt{a} + \sqrt{b} \geq \sqrt{a+b}$. Then using $\sum_{s=0}^\infty 2^{-s}\sqrt{2(1+s)} \leq 4$,

$$\sum_{s=0}^\infty \alpha_s = J_n + \sum_{s=0}^\infty 2^{-s}R_n\sqrt{2(1+s)(1+t)} \leq J_n + 4R_n\sqrt{1+t}.$$

Then

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}}\sum_{s=0}^S\left\langle \varepsilon, g^{s+1} - g^s\right\rangle_{2,P_n} \geq \frac{J_n}{\sqrt{n}} + 4R_n\sqrt{\frac{1+t}{n}} \mid \mathcal{F}_n\right)$$

$$
\begin{aligned}
&\leq \mathbb{P}\left(\sum_{s=0}^{S} \sup_{g\in\mathcal{G}} \left\langle \varepsilon, g^{s+1} - g^{s} \right\rangle_{2,P_n} \geq \frac{1}{\sqrt{n}} \sum_{s=0}^{S} \alpha_s \mid \mathcal{F}_n \right) \\
&\leq \sum_{s=0}^{S} \mathbb{P}\left(\sup_{g\in\mathcal{G}} \left\langle \varepsilon, g^{s+1} - g^{s} \right\rangle_{2,P_n} \geq \frac{1}{\sqrt{n}} \alpha_s \mid \mathcal{F}_n \right) \\
&= \sum_{s=0}^{S} \mathbb{P}\left(\max_{k=1,\ldots,N_{s+1}} \left\langle \varepsilon, g_k^{s+1} - g_k^{s+1,s} \right\rangle_{2,P_n} \geq \frac{1}{\sqrt{n}} \alpha_s \mid \mathcal{F}_n \right) \\
&\leq \sum_{s=0}^{S} e^{-(1+s)(1+t)} \\
&\leq e^{-t}.
\end{aligned}
$$

□

Theorem 5.2.10. *Suppose that $\mathcal{G}$ satisfies the "uniform entropy condition", i.e. there exists a decreasing function $H : \mathbb{R} \to \mathbb{R}$ satisfying*

$$\int_0^1 \sqrt{H(u)}du < \infty$$

such that, for all $u > 0$ and any probability distribution Q with finite support,

$$H(u \left\| G \right\|_{2,Q}, \mathcal{G}, \left\| \cdot \right\|_{2,Q}) \leq H(u).$$

Then the empirical process ν_n is asymptotically equicontinuous.

Proof of Theorem 5.2.10. Take any arbitrary $g_0 \in \mathcal{G}$. We will show that ν_n is asymptotically equicontinuous at g_0. Take arbitrary $\epsilon_1, \epsilon_2 > 0$, and fix $S \in \mathbb{N}$. Define, for $\delta > 0$, the closed δ-ball around the origin:

$$\mathcal{B}(\delta) := \left\{ g \in \mathcal{G} : \left\| g \right\|_{2,P} \leq \delta \right\}.$$

Then clearly, the theoretical radius of $\mathcal{B}(\delta)$ is $\sup_{g\in\mathcal{B}(\delta)} \left\| g \right\|_{2,P} = \delta$. Denote the empirical radius of $\mathcal{B}(\delta)$ by $R_{n,\delta} = \sup_{g\in\mathcal{B}(\delta)} \left\| g \right\|_{2,P_n}$, and analogously to the proof of Proposition 5.2.9, define

$$J_{n,\delta} := \sum_{s=0}^{S} 2^{-s} R_{n,\delta} \sqrt{2H\left(2^{-(s+1)} R_{n,\delta}, \mathcal{B}(\delta), \left\| \cdot \right\|_{2,P_n}\right)}.$$

Also define

$$\mathcal{J}(\rho) := 8 \int_0^{\rho} \sqrt{2H(u)}du, \qquad \rho > 0,$$

which is bounded for any finite $\rho > 0$, by the uniform entropy condition.

Define $A \in \mathcal{F}$ as the event on which $R_{n,\delta} \leq 2\delta$ and $\left\| G \right\|_{2,P_n} \leq 2 \left\| G \right\|_{2,P}$. Then on this event, we have

$$J_{n,\delta} = \sum_{s=0}^{S} 2^{-s} R_{n,\delta} \sqrt{2H\left(2^{-(s+1)} R_{n,\delta}, \mathcal{B}(\delta), \left\| \cdot \right\|_{2,P_n}\right)}$$

$$\begin{aligned}
&\leq 4\int_0^{R_{n,\delta}} \sqrt{2H(u,\mathcal{B}(\delta),\|\cdot\|_{2,P_n})}du \\
&\leq 4\int_0^{2\delta} \sqrt{2H\left(u,\mathcal{G},\|\cdot\|_{2,P_n}\right)}du \qquad \text{since } R_{n,\delta}\leq 2\delta \text{ on } A \text{ and } \mathcal{B}(\delta)\subseteq\mathcal{G} \\
&\leq 4\int_0^{2\delta} \sqrt{2H\left(\frac{u}{\|G\|_{2,P_n}}\right)}du \qquad \text{by the uniform entropy condition} \\
&\leq 4\int_0^{2\delta} \sqrt{2H\left(\frac{u}{2\,\|G\|_{2,P}}\right)}du \\
&\quad \text{since } \|G\|_{2,P_n}\leq 2\,\|G\|_{2,P} \text{ on } A \text{ and } H \text{ is decreasing.} \\
&= 8\,\|G\|_{2,P}\int_0^{\frac{\delta}{\|G\|_{2,P}}} \sqrt{2H(u)}du \qquad \text{by substitution} \\
&= \|G\|_{2,P}\,\mathcal{J}\left(\frac{\delta}{\|G\|_{2,P}}\right).
\end{aligned}$$

On A, we also have

$$\begin{aligned}
\sup_{g\in\mathcal{B}(\delta)}\left\|P_n^\sigma\left(g-g^{S+1}\right)\right\|_{\mathcal{Y}} &\leq \sup_{g\in\mathcal{B}(\delta)}\left\|g-g^{S+1}\right\|_{1,P_n} \\
&\leq \sup_{g\in\mathcal{B}(\delta)}\left\|g-g^{S+1}\right\|_{2,P_n} \\
&\leq 2^{-(S+1)}R_{n,\delta} \\
&\leq 2^{-S}\delta. \qquad (*)
\end{aligned}$$

So on A, noting that

$$\begin{aligned}
\|P_n^\sigma\|_{\mathcal{B}(\delta)} &= \sup_{g\in\mathcal{B}(\delta)}\left\|P_n^\sigma\left(g-g^{S+1}\right)+\sum_{s=0}^{S}P_n^\sigma\left(g^{s+1}-g^s\right)\right\|_{\mathcal{Y}} \\
&\leq \sup_{g\in\mathcal{B}(\delta)}\left\|P_n^\sigma\left(g-g^{S+1}\right)\right\|_{\mathcal{Y}} + \sup_{g\in\mathcal{B}(\delta)}\left\|\sum_{s=0}^{S}P_n^\sigma\left(g^{s+1}-g^s\right)\right\|_{\mathcal{Y}},
\end{aligned}$$

we have, for all $t>0$,

$$\begin{aligned}
&\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{B}(\delta)} \geq \frac{\sqrt{2}\,\|G\|_{2,P}\,\mathcal{J}\left(\frac{\delta}{\|G\|_{2,P}}\right)}{\sqrt{n}} + 12\delta\sqrt{\frac{1+t}{n}} + 2^{-S}\delta \mid \mathcal{F}_n\right) \\
&= \mathbb{P}\Bigg(\sup_{g\in\mathcal{B}(\delta)}\left\|P_n^\sigma\left(g-g^{S+1}\right)\right\|_{\mathcal{Y}} + \sup_{g\in\mathcal{B}(\delta)}\left\|\sum_{s=0}^{S}P_n^\sigma\left(g^{s+1}-g^s\right)\right\|_{\mathcal{Y}} \\
&\qquad\qquad \geq \frac{\sqrt{2}J_{n,\delta}}{n} + 6R_{n,\delta}\sqrt{\frac{1+t}{n}} + 2^{-S}\delta \mid \mathcal{F}_n\Bigg)
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{P}\left(\sup_{g\in\mathcal{B}(\delta)}\left\|\sum_{s=0}^{S} P_n^\sigma\left(g^{s+1}-g^s\right)\right\|_{\mathcal{Y}} \geq \frac{\sqrt{2}J_{n,\delta}}{\sqrt{n}} + 6R_{n,\delta}\sqrt{\frac{1+t}{n}} \mid \mathcal{F}_n\right)\\
&\leq 2e^{-t},
\end{aligned}
$$

where the term $\mathbb{P}\left(\sup_{g\in\mathcal{B}(\delta)}\left\|P_n^\sigma\left(g-g^{S+1}\right)\right\|_{\mathcal{Y}} \geq 2^{-S}\delta \mid \mathcal{F}_n\right)$ vanishes by (*) and the last inequality follows Proposition 5.2.9(i). Then we can de-symmetrise using Lemma 5.2.5:

$$
\begin{aligned}
&\mathbb{P}\left(\|P_n - P\|_{\mathcal{B}(\delta)} \geq \frac{4\sqrt{2}\,\|G\|_{2,P}\,\mathcal{J}\left(\frac{\delta}{\|G\|_{2,P}}\right)}{\sqrt{n}} + 48\delta\sqrt{\frac{1+t}{n}} + 2^{-(S-2)}\delta\right)\\
&\leq 4\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{B}(\delta)} \geq \frac{\sqrt{2}\,\|G\|_{2,P}\,\mathcal{J}\left(\frac{\delta}{\|G\|_{2,P}}\right)}{\sqrt{n}} + 12\delta\sqrt{\frac{1+t}{n}} + 2^{-S}\delta\right)\\
&\leq 8e^{-t} + 4\mathbb{P}\left(R_{n,\delta} > 2\delta \text{ or } \|G\|_{2,P_n} > 2\,\|G\|_{2,P}\right)\\
&= 8e^{-t} + 4\mathbb{P}\left(\sup_{g\in\mathcal{B}(\delta)\cup\{G\}}\|g\|_{2,P_n}^2 > 4\sup_{g\in\mathcal{B}(\delta)\cup\{G\}}\|g\|_{2,P}^2\right).
\end{aligned}
$$

Now let $t = \log\left(\frac{8}{\epsilon_2}\right)$ and S large enough such that $2^{-(S-2)} \leq \frac{1}{\sqrt{n}}$, and δ small enough such that

$$
4\sqrt{2}\,\|G\|_{2,P}\,\mathcal{J}\left(\frac{\delta}{\|G\|_{2,P}}\right) + 48\delta\sqrt{1+\log\left(\frac{8}{\epsilon_2}\right)} + \delta \leq \epsilon_1.
$$

Then

$$
\begin{aligned}
&\mathbb{P}\left(\sqrt{n}\,\|P_n - P\|_{\mathcal{B}(\delta)} > \epsilon_1\right)\\
&\quad\leq \epsilon_2 + 4\mathbb{P}\left(\sup_{g\in\mathcal{B}(\delta)\cup\{G\}}\|g\|_{2,P_n}^2 > 4\sup_{g\in\mathcal{B}(\delta)\cup\{G\}}\|g\|_{2,P}^2\right).
\end{aligned}
$$

Hence, for any $g \in \mathcal{G}$ such that $\|g - g_0\|_{2,P} \leq \delta$,

$$
\begin{aligned}
&\mathbb{P}\left(\|\nu_n(g) - \nu_n(g_0)\|_{\mathcal{Y}} > \epsilon_1\right)\\
&= \mathbb{P}\left(\sqrt{n}\,\|(P_n - P)(g - g_0)\|_{\mathcal{Y}} > \epsilon_1\right)\\
&\leq \mathbb{P}\left(\sqrt{n}\,\|P_n - P\|_{\mathcal{B}(\delta)} > \epsilon_1\right)\\
&\leq \epsilon_2 + 4\mathbb{P}\left(\sup_{g\in\mathcal{B}(\delta)\cup\{G\}}\|g\|_{2,P_n}^2 > 4\sup_{g\in\mathcal{B}(\delta)\cup\{G\}}\|g\|_{2,P}^2\right).
\end{aligned}
$$

Here, by the uniform law of large numbers on $\mathcal{B}(\delta)\cap\{G\}$ (Theorem 5.2.8), the second term converges to 0 as $n \to \infty$. Hence, as ϵ_1 and ϵ_2 were arbitrary, we have asymptotic equicontinuity. □

Theorem 5.2.11. *Suppose that $\varepsilon_1, ..., \varepsilon_n$ are i.i.d. with Gaussian distribution with mean 0 and covariance operator Q, and that $\operatorname{Tr} Q = 1$. Further, suppose that*

$$J(\delta) := 4 \int_0^\delta \sqrt{2H(u, \mathcal{B}_{2,P_n}(\delta), \|\cdot\|_{2,P_n})}du < \infty,$$

for each $\delta > 0$ and $\frac{J(\delta)}{\delta^2}$ is decreasing in δ where $\mathcal{B}_{2,P_n}(\delta) := \{g \in \mathcal{G} : \|g\|_{2,P_n} \leq \delta\}$. Then for all $t \geq \frac{3}{8}$ and all δ_n satisfying

$$\sqrt{n}\delta_n^2 \geq 8 \left(J(\delta_n) + 4\delta_n \sqrt{1+t} + \delta_n \sqrt{\frac{8}{3}t} \right),$$

we have

$$\mathbb{P}\left(\|\hat{g}_n - g_0\|_{2,P_n} > \delta_n \right) \leq \left(1 + \frac{2}{e-1}\right) e^{-t}.$$

Proof of Theorem 5.2.11. First, recall the notation

$$\langle \varepsilon, g \rangle_{2,P_n} = \frac{1}{n} \sum_{i=1}^n \langle \varepsilon_i, g(X_i) \rangle_{\mathcal{Y}}$$

from Proposition 5.2.9(ii), and note that we have the following basic inequality

$$\|\hat{g}_n - g_0\|_{2,P_n}^2 \leq 2 \langle \varepsilon, \hat{g}_n - g_0 \rangle_{2,P_n}, \tag{*}$$

which follows from the fact that $\hat{g}_n$ minimises $\|Y_i - g(X_i)\|_{2,P_n}^2$ over $g \in \mathcal{G}$, giving

$$\|\varepsilon_i - (g_0 - \hat{g}_n)\|_{2,P_n}^2 = \|Y_i - \hat{g}_n(X_i)\|_{2,P_n}^2 \leq \|Y_i - g_0(X_i)\|_{2,P_n}^2 = \|\varepsilon_i\|_{2,P_n}^2.$$

We use a technique called the "peeling device", first introduced by van de Geer (2000). See that

$$\begin{aligned}
&\mathbb{P}\left(\|\hat{g}_n - g_0\|_{2,P_n} > \delta_n \right) = \mathbb{P}\left(\bigcup_{j=1}^\infty \left\{ 2^{j-1}\delta_n < \|\hat{g}_n - g_0\|_{2,P_n} \leq 2^j \delta_n \right\} \right) \\
&\leq \sum_{j=1}^\infty \mathbb{P}\left(2^{j-1}\delta_n < \|\hat{g}_n - g_0\|_{2,P_n} \leq 2^j \delta_n \right) \qquad \text{by the union bound} \\
&= \sum_{j=1}^\infty \mathbb{P}\left(\left\{ 2^{j-1}\delta_n < \|\hat{g}_n - g_0\|_{2,P_n} \right\} \bigcap \left\{ \hat{g}_n - g_0 \in \mathcal{B}_n(2^j\delta_n) \right\} \right) \\
&\leq \sum_{j=1}^\infty \mathbb{P}\left(\left\{ \left(2^{j-1}\delta_n\right)^2 < 2 \langle \varepsilon, \hat{g}_n - g_0 \rangle_{2,P_n} \right\} \bigcap \left\{ \hat{g}_n - g_0 \in \mathcal{B}_n(2^j\delta_n) \right\} \right) \quad \text{by (*)} \\
&\leq \sum_{j=1}^\infty \mathbb{P}\left(\sup_{g \in \mathcal{B}_n(2^j\delta_n)} 2 \langle \varepsilon, g \rangle_{2,P_n} > \left(2^{j-1}\delta_n\right)^2 \right)
\end{aligned}$$

$$= \sum_{j=1}^{\infty} \mathbb{P}\left(\sup_{g \in \mathcal{B}_n(2^j \delta_n)} \langle \varepsilon, g \rangle_{2,P_n} > \frac{1}{8}\left(2^j \delta_n\right)^2 \right).$$

Now, applying the hypothesis on δ_n, we see that, for each j,

$$\begin{aligned} \frac{1}{8}\left(2^j \delta_n\right)^2 &\geq \frac{(2^j)^2 J(\delta_n)}{\sqrt{n}} + 4(2^j)^2 \delta_n \sqrt{\frac{1+t}{n}} + \frac{\sqrt{\frac{8}{3}t}(2^j)^2 \delta_n}{\sqrt{n}} \\ &\geq \frac{J(2^j \delta_n)}{\sqrt{n}} + 4(2^j \delta_n)\sqrt{\frac{1+t+j}{n}} + \frac{\sqrt{\frac{8}{3}t}(2^j)^2 \delta_n}{\sqrt{n}} \\ &\geq \frac{J_n}{\sqrt{n}} + 4(2^j \delta_n)\sqrt{\frac{1+t+j}{n}} + \frac{\sqrt{\frac{8}{3}t}(2^j)^2 \delta_n}{\sqrt{n}} \end{aligned}$$

where we used the fact that $\frac{J(\delta)}{\delta^2}$ is decreasing in δ and $\sqrt{1+t+j} \leq 2^j \sqrt{1+t}$, and J_n is defined as in Proposition 5.2.9 with $\mathcal{G} = \mathcal{B}_n(2^j \delta_n)$ and $R_n = 2^j \delta_n$. On the other hand, we can write, for any $S \in \mathbb{N}$,

$$\langle \varepsilon, g \rangle_{2,P_n} = \left\langle \varepsilon, g - g^{S+1} \right\rangle_{2,P_n} + \sum_{s=0}^{S} \left\langle \varepsilon, g^{s+1} - g^s \right\rangle_{2,P_n},$$

using the chaining notation in Section 5.2.3. Hence,

$$\begin{aligned} &\mathbb{P}\left(\|\hat{g}_n - g_0\|_{2,P_n} > \delta_n \right) \\ &\leq \sum_{j=1}^{\infty} \mathbb{P}\left(\sup_{g \in \mathcal{B}_n(2^j \delta_n)} \left\langle \varepsilon, g - g^{S+1} \right\rangle_{2,P_n} > \frac{\sqrt{\frac{8}{3}t}(2^j)^2 \delta_n}{\sqrt{n}} \right) \\ &\quad + \sum_{j=1}^{\infty} \mathbb{P}\left(\sup_{g \in \mathcal{B}_n(2^j \delta_n)} \sum_{s=0}^{S} \left\langle \varepsilon, g^{s+1} - g^s \right\rangle_{2,P_n} > \frac{J_n}{\sqrt{n}} + 4(2^j \delta_n)\sqrt{\frac{1+t+j}{n}} \right) \\ &\leq \sum_{j=1}^{\infty} \mathbb{P}\left(\frac{2^j}{2^{S+1}} \delta_n \|\varepsilon\|_{2,P_n} > \frac{\sqrt{\frac{8}{3}t} 2^{2j} \delta_n}{\sqrt{n}} \right) + \sum_{j=1}^{\infty} e^{-(t+j)} \quad \text{Proposition 5.2.9(ii)} \\ &= \sum_{j=1}^{\infty} \mathbb{P}\left(\|\varepsilon\|_{2,P_n} > 2^j \sqrt{\frac{8}{3}t} \right) + \frac{1}{e-1} e^{-t} \quad \text{letting } S \text{ such that } \sqrt{n} \leq 2^{S+1} \\ &\leq \sum_{j=1}^{\infty} \mathbb{P}\left(\|\varepsilon\|_{2,P_n} > 2^j + \sqrt{\frac{8}{3}t} \right) + \frac{1}{e-1} e^{-t} \quad \text{since } t \geq \frac{3}{8} \\ &\leq \sum_{j=1}^{\infty} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} \|\varepsilon_i\|_{\mathcal{Y}}^2 > 2^{2j} + \frac{8}{3}t \right) + \frac{1}{e-1} e^{-t} \\ &\leq \sum_{j=1}^{\infty} e^{-\frac{3}{8}2^{2j} - t} \mathbb{E}\left[e^{\frac{3}{8}\frac{1}{n}\sum_{i=1}^{n} \|\varepsilon_i\|_{\mathcal{Y}}^2} \right] + \frac{1}{e-1} e^{-t} \qquad \text{by Markov's inequality} \end{aligned}$$

$$\begin{aligned}
&\leq \sum_{j=1}^{\infty} e^{-\frac{3}{8}2^{2j}-t} \prod_{i=1}^{n} \mathbb{E}\left[e^{\frac{3}{8}\frac{1}{n}\|\varepsilon_i\|_{\mathcal{Y}}^2}\right] + \frac{1}{e-1}e^{-t} \qquad \text{by independence} \\
&\leq \sum_{j=1}^{\infty} e^{-\frac{3}{8}2^{2j}-t} \mathbb{E}\left[e^{\frac{3}{8}\|\varepsilon_1\|_{\mathcal{Y}}^2}\right] + \frac{1}{e-1}e^{-t} \qquad \text{by Jensen's inequality} \\
&\leq \sum_{j=1}^{\infty} e^{-\frac{3}{8}2^{2j}-t} + \frac{1}{e-1}e^{-t} \qquad \text{by Gaussian concentration} \\
&\leq e^{-t}\left(e^{-\frac{3}{4}} - e^{-1} + \sum_{j=1}^{\infty} e^{-j} + \frac{1}{e-1}\right) \\
&\leq e^{-t}\left(1 + \frac{2}{e-1}\right).
\end{aligned}$$

□

Theorem 5.2.12. *Suppose the following uniform entropy condition is satisfied: there exists some function $H : \mathbb{R} \to \mathbb{R}$ satisfying*

$$\mathcal{J}(1) := 4\int_0^1 \sqrt{2H(u)}du < \infty,$$

such that, for all $u > 0$ and any probability distribution Q with finite support,

$$H(uL, \mathcal{L} \circ \mathcal{G}, \|\cdot\|_{2,Q}) \leq H(u).$$

Then

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} \left|\mathcal{R}(g) - \hat{\mathcal{R}}(g)\right| > \frac{4\sqrt{2}L\mathcal{J}(1)}{\sqrt{n}} + 24L\sqrt{\frac{1+t}{n}} + \frac{L}{\sqrt{n}}\right) \leq 2e^{-t}.$$

Proof of Theorem 5.2.12. First, denote by $\mathcal{L}\circ\mathcal{G}$ the class of functions $\mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ given by $(x, y) \mapsto \mathcal{L}(y, g(x))$ for $g \in \mathcal{G}$. Also, by an abuse of notation, for each $g \in \mathcal{G}$, denote by $\mathcal{L} \circ g$ the function $(x, y) \mapsto \mathcal{L}(y, g(x))$. Then we have

$$P\mathcal{L} \circ g = \mathcal{R}(g), \qquad P_n\mathcal{L} \circ g = \hat{\mathcal{R}}(g).$$

Since the loss $\mathcal{L}$ is bounded above by L, the empirical radius and the theoretical radius of $\mathcal{L}\circ\mathcal{G}$ are both bounded above by L. In the chaining notation of Section 5.2.3, define

$$J_n = \sum_{s=0}^{S} 2^{-s}L\sqrt{2H(2^{-(s+1)}L, \mathcal{L} \circ \mathcal{G}, \|\cdot\|_{2,P_n})}.$$

Then from the very definition of the chains, we have

$$\sup_{g\in\mathcal{G}} \left|P_n^{\sigma}\left(\mathcal{L} \circ g - \mathcal{L} \circ g^{S+1}\right)\right| \leq \sup_{g\in\mathcal{G}} \left\|\mathcal{L} \circ g - \mathcal{L} \circ g^{S+1}\right\|_{1,P_n}$$

$$\begin{aligned} &\leq \sup_{g\in\mathcal{G}} \left\| \mathcal{L}\circ g - \mathcal{L}\circ g^{S+1} \right\|_{2,P_n} \\ &\leq 2^{-(S+1)}L. \qquad (*) \end{aligned}$$

First, see that

$$\begin{aligned} J_n &= \sum_{s=0}^{S} 2^{-s} L \sqrt{2H(2^{-(s+1)}L, \mathcal{L}\circ\mathcal{G}, \|\cdot\|_{2,P_n})} \\ &\leq 4\int_0^L \sqrt{2H(u, \mathcal{L}\circ\mathcal{G}, \|\cdot\|_{2,P_n})}du \\ &\leq 4\int_0^L \sqrt{2H\left(\frac{u}{L}\right)}du && \text{by uniform entropy condition} \\ &= 4L\int_0^1 \sqrt{2H(u)}du && \text{by substitution} \\ &= L\mathcal{J}(1). && (**) \end{aligned}$$

Then, by the symmetrisation lemma (Lemma 5.2.5) followed by chaining, we have

$$\begin{aligned} &\mathbb{P}\left(\sup_{g\in\mathcal{G}} \left|\mathcal{R}(g) - \hat{\mathcal{R}}(g)\right| > \frac{4\sqrt{2}L\mathcal{J}(1)}{\sqrt{n}} + 24L\sqrt{\frac{1+t}{n}} + 2^{-(S-1)}L\right) \\ &= \mathbb{P}\left(\|P - P_n\|_{\mathcal{L}\circ\mathcal{G}} > \frac{4\sqrt{2}J_n}{\sqrt{n}} + 24L\sqrt{\frac{1+t}{n}} + 2^{-(S-1)}L\right) \qquad \text{by } (**) \\ &\leq 4\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{L}\circ\mathcal{G}} > \frac{\sqrt{2}J_n}{\sqrt{n}} + 6L\sqrt{\frac{1+t}{n}} + 2^{-(S+1)}L\right) \qquad \text{by Lemma 5.2.5} \\ &\leq 4\mathbb{P}\left(\sup_{g\in\mathcal{G}} \left|P_n^\sigma\left(\mathcal{L}\circ g - \mathcal{L}\circ g^{S+1}\right)\right| + \sup_{g\in\mathcal{G}}\left|\sum_{s=0}^{S} P_n^\sigma\left(\mathcal{L}\circ g^{s+1} - \mathcal{L}\circ g^{s}\right)\right|\right. \\ &\qquad\qquad \left. > \frac{\sqrt{2}J_n}{\sqrt{n}} + 6L\sqrt{\frac{1+t}{n}} + 2^{-(S+1)}L\right) \\ &\leq 4\mathbb{P}\left(\sup_{g\in\mathcal{G}}\left|\sum_{s=0}^{S} P_n^\sigma\left(\mathcal{L}\circ g^{s+1} - \mathcal{L}\circ g^{s}\right)\right| > \frac{\sqrt{2}J_n}{\sqrt{n}} + 6L\sqrt{\frac{1+t}{n}}\right) \\ &\quad + 4\mathbb{P}\left(\sup_{g\in\mathcal{G}} \left|P_n^\sigma\left(\mathcal{L}\circ g - \mathcal{L}\circ g^{S+1}\right)\right| > 2^{-(S+1)}L\right) \qquad \text{by the union bound} \\ &\leq 2e^{-t}, \end{aligned}$$

where the second term disappears by (*) and the first term is bounded by Proposition 5.2.9(i). Now letting S be large enough such that $\sqrt{n} \leq 2^{S+1}$,

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} \left|\mathcal{R}(g) - \hat{\mathcal{R}}(g)\right| > \frac{4\sqrt{2}L\mathcal{J}(1)}{\sqrt{n}} + 24L\sqrt{\frac{1+t}{n}} + \frac{L}{\sqrt{n}}\right) \leq 2e^{-t}.$$

□

We now prove Theorems 5.4.1, 5.4.2 and 5.4.3. The idea is to approximate smooth functions by piecewise polynomials (Kolmogorov, 1955). We start with some development shared by the three Theorems.

Let $g \in \mathcal{G}_B^m$, $x, x+h \in \mathcal{X}$ and $p \in \mathbb{N}_0^d$ with $[p] \leq m-1$. Then $D^p g$ is $(m-[p])$-times differentiable, and

$$\begin{aligned}\|(D^p g)^{(m-[p])}(x)\|_{\text{op}} &= \| \sum_{[q]\leq m-[p]} \frac{(m-[p])!}{q!} D^{p+q} g(x) \|_{\mathcal{Y}} \\ &\leq K_B \sum_{[q]\leq m-[p]} \frac{(m-[p])!}{q!}.\end{aligned}$$

Hence,

$$D^p g(x+h) = \sum_{[q]\leq m-1-[p]} \frac{h^q}{q!} D^{p+q} g(x) + R_p(g,x,h) \tag{*}$$

by Taylor's Theorem (Theorem 5.3.1), where

$$\|R_p(g,x,h)\|_{\mathcal{Y}} \leq K_B \frac{\|h\|^{m-[p]}}{(m-[p])!} \sum_{[q]\leq m-[p]} \frac{(m-[p])!}{q!}.$$

So there is a constant $K_1 = K_1(K_B, m, d) \geq 1$ such that, for all $g \in \mathcal{G}_B^m$, $x \in \mathcal{X}$, $x+h \in \mathcal{X}$ and $p \in \mathbb{N}_0^d$ with $[p] \leq m-1$,

$$\|R_p(g,x,h)\|_{\mathcal{Y}} \leq K_1 \|h\|^{m-[p]}. \tag{**}$$

Let $\Delta := (\frac{\delta}{4K_1})^{\frac{1}{m}}$, and $x_{(1)}, ..., x_{(L)}$ a $\frac{\Delta}{2}$-net in $\mathcal{X}$, i.e. $\sup_{x\in\mathcal{X}}\{\inf_{1\leq l\leq L}\|x - x_{(l)}\|\} \leq \frac{\Delta}{2}$. By decomposing $\mathcal{X}$ into cubes of side $\left\lceil \frac{d^{1/2}}{\Delta} \right\rceil^{-1}$ and taking the $x_{(l)}$ as the centres thereof, we can take

$$L \leq K_2 \delta^{-\frac{d}{m}} \tag{\dagger}$$

for some constant $K_2 = K_2(d, K_1)$. Now, for each $k = 0, 1, ..., m-1$, define $\delta_k = \frac{\delta}{2\Delta^k e^d}$. We construct a cover of B as follows. First, to ease the notation, write $N_k = N(\frac{1}{2}\delta_k, B, \|\cdot\|_{\mathcal{Y}})$, and find a set $\{a_j^k, j = 1, ..., N_k\} \subset B$ such that $\mathcal{B}(a_j^k, \frac{1}{2}\delta_k)$ cover B. Then define

$$\begin{aligned}A_1^k = \mathcal{B}(a_1^k, \frac{1}{2}\delta_k), A_2^k \\ = \mathcal{B}(a_2^k, \frac{1}{2}\delta_k)\backslash\mathcal{B}(a_1^k, \frac{1}{2}\delta_k), ..., A_{N_k}^k \\ = \mathcal{B}(a_{N_k}^k, \frac{1}{2}\delta_k)\backslash \cup_{j=1}^{N_k-1} \mathcal{B}(a_j^k, \frac{1}{2}\delta_k).\end{aligned}$$

Then $\mathcal{A}_k := \{A_j^k, j = 1, ..., N_k\}$ is a cover of B of cardinality N_k, whose sets A_j^k have diameter at most δ_k and are disjoint. For each $l = 1, ..., L$, $g \in \mathcal{G}_B^m$ and $p \in \mathbb{N}_0^d$ with $[p] \leq m-1$, define $A_{l,p}(g)$ as the unique set in $\mathcal{A}_{[p]}$ such that $D^p g(x_{(l)}) \in A_{l,p}(g)$, and $a_{l,p}(g)$ as the centre of the ball from which $A_{l,p}(g)$ was created, so that $\|a_{l,p}(g) - D^p g(x_{(l)})\|_{\mathcal{Y}} \leq \frac{1}{2}\delta_{[p]}$. Then if $g_1, g_2 \in \mathcal{G}_B^m$ are such that $A_{l,p}(g_1) = A_{l,p}(g_2)$ for all $l = 1, ..., L$ and all $p \in \mathbb{N}_0^d$ with $[p] \leq m-1$, then

$$\|D^p(g_1 - g_2)(x_{(l)})\|_{\mathcal{Y}} \leq \delta_{[p]}, \qquad (***)$$

since the diameter of $A_{l,p}(g_1) = A_{l,p}(g_2)$ is at most $\delta_{[p]}$. For each $x \in \mathcal{X}$, take $x_{(l)}$ such that $\|x - x_{(l)}\| \leq \frac{\Delta}{2}$. Then we have, by putting $p = 0$ into (*),

$$\begin{aligned}
&\|(g_1 - g_2)(x)\|_{\mathcal{Y}} \\
&= \Bigg\|R_0(g_1, x_{(l)}, x - x_{(l)}) - R_0(g_2, x_{(l)}, x - x_{(l)}) \\
&\qquad + \sum_{[p]\leq m-1} \frac{(x - x_{(l)})^p}{p!} D^p(g_1 - g_2)(x_{(l)})\Bigg\|_{\mathcal{Y}} \\
&\leq 2K_1\|x - x_{(l)}\|^m + \sum_{[p]\leq m-1} \delta_{[p]} \frac{\|x - x_{(l)}\|^{[p]}}{p!} \qquad \text{by (**) with } p = 0 \text{ and (***)} \\
&\leq 2K_1\Delta^m + \sum_{k=0}^{m-1} \delta_k \Delta^k \left(\sum_{[p]=k} \frac{1}{p!}\right) \\
&\leq \frac{\delta}{2} + \left(\max_{k\leq m-1} \delta_k \Delta^k\right) \sum_{k=0}^{m-1} \frac{d^k}{k!} \\
&\leq \frac{\delta}{2} + \frac{\delta}{2e^d} e^d \\
&= \delta.
\end{aligned}$$

It follows that the δ-covering number $N(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty)$ with respect to the supremum norm is bounded by the number of distinct possibilities for $\{A_{l,p}(g) : l = 1, ..., L, g \in \mathcal{G}_B^m, p \in \mathbb{N}_0^d, [p] \leq m-1\}$.

Proof of Theorem 5.4.1. Let $x_{(l)}$ be ordered so that for $1 < l \leq L$, $\|x_{(l')} - x_{(l)}\| \leq \Delta$ for some $l' < l$. For each $l = 1, ..., L$ and $p \in \mathbb{N}_0^d$ with $[p] \leq m-1$, we write $\mathcal{A}_{l,p}$ for the number of possibilities of $A_{l,p}(g)$ for $g \in \mathcal{G}_B^m$, and for each $l = 1, ..., L$, we write $\mathcal{A}_l$ for the number of possibilities of $A_{l,p}(g)$ as $p \in \mathbb{N}_0^d$ varies with $[p] \leq m-1$. For $l = 1$, we have $D^p g(x_{(1)}) \in B$ for each $p \in \mathbb{N}_0^d$ with $[p] \leq m-1$. So

$$\mathcal{A}_{1,p} \leq N_{[p]} = N\left(\frac{1}{4e^d}\delta^{\frac{m-[p]}{m}}(4K_1)^{\frac{[p]}{m}}, B, \|\cdot\|_{\mathcal{Y}}\right) \leq N\left(\frac{\delta}{4e^d}, B, \|\cdot\|_{\mathcal{Y}}\right),$$

where the last upper bound follows since $N(\cdot, B, \|\cdot\|_{\mathcal{Y}})$ is a decreasing function, and we have $K_1 \geq 1$ and $0 < \delta < 1$. This upper bound has no dependence on p.

The number of different $p \in \mathbb{N}_0^d$ with $[p] \leq m-1$ is equal to $\binom{m+d-1}{d}$, which is bounded above by m^d, and so $\mathcal{A}_1 \leq N(\frac{\delta}{4e^d}, B, \|\cdot\|_\mathcal{Y})^{m^d}$. Since $B = B \cap \mathcal{B}(0, K_B)$ is (M, τ_{asd})-homogeneous, $N(\frac{\delta}{4e^d}, B, \|\cdot\|_\mathcal{Y}) \leq M(\frac{4e^d K_B}{\delta})^{\tau_{\text{asd}}}$, and so

$$\mathcal{A}_1 \leq M^{m^d} \left(\frac{4e^d K_B}{\delta}\right)^{\tau_{\text{asd}} m^d}. \tag{††}$$

Now, for $1 < l \leq L$, suppose that $A_{l',q}(g)$ is given for all $l' < l$ and all $q \in \mathbb{N}_0^d$ with $[q] \leq m-1$. Choose $l' < l$ such that $\|x_{(l')} - x_{(l)}\| \leq \Delta$, and write $y_{l,p}(g) := \sum_{[q] \leq m-1-[p]} \frac{(x_{(l')} - x_{(l)})^q}{q!} a_{l',p+q}(g)$. Then for any $p \in \mathbb{N}_0^d$ with $[p] \leq m-1$, (*) tells us that

$$\begin{aligned}
&\left\|D^p g(x_{(l)}) - y_{l,p}(g)\right\|_\mathcal{Y} \\
&= \left\|R_p(g, x_{(l')}, x_{(l)} - x_{(l')})\right\|_\mathcal{Y} \\
&\quad + \sum_{[q] \leq m-1-[p]} \frac{\|x_{(l')} - x_{(l)}\|^{[q]}}{q!} \left\|D^{p+q} g(x_{(l')}) - a_{l',p+q}(g)\right\|_\mathcal{Y} \\
&\leq K_1 \Delta^{m-[p]} + \sum_{[q] \leq m-1-[p]} \delta_{[p+q]} \frac{\Delta^q}{q!} \\
&= K_1 \frac{\Delta^m}{\Delta^{[p]}} + \delta_{[p]} \sum_{k=0}^{m-1-[p]} \delta_k \Delta^k \left(\sum_{[q]=k} \frac{1}{q!}\right) \\
&\leq \frac{e^d + 1}{2} \delta_{[p]}.
\end{aligned}$$

As $a_{l',p+q}(g)$ is given for all $[q] \leq m-1-[p]$, $y_{l,p}(g)$ is a fixed point in $\mathcal{Y}$. So $\mathcal{A}_{l,p}$ is bounded by the number of sets in $\mathcal{A}_{[p]}$ that intersect with $B_{l,p}(g) := B \cap \mathcal{B}\left(y_{l,p}(g), \frac{e^d+1}{2}\delta_{[p]}\right)$. Define $\mathcal{A}_{l,p}(g) := \{A \in \mathcal{A}_{[p]} : A \cap B_{l,p}(g) = \emptyset\}$ and $\mathcal{A}'_{l,p}(g) := \{A \in \mathcal{A}_{[p]} : A \cap B_{l,p}(g) \neq \emptyset\}$, so that $\mathcal{A}_{[p]} = \mathcal{A}_{l,p}(g) \cup \mathcal{A}'_{l,p}(g)$, $N_{[p]} = |\mathcal{A}_{[p]}| = |\mathcal{A}_{l,p}(g)| + |\mathcal{A}'_{l,p}(g)|$ and $\mathcal{A}_{l,p} \leq |\mathcal{A}'_{l,p}(g)|$. Now, write $B^+_{l,p}(g) := B \cap \mathcal{B}(y_{l,p}(g), \frac{e^d+3}{2}\delta_{[p]})$. Then we have $A \subset B^+_{l,p}(g)$ for all $A \in \mathcal{A}'_{l,p}(g)$. Let $\mathcal{A}^+_{l,p}(g)$ be a $\frac{1}{2}\delta_{[p]}$-cover of $B^+_{l,p}(g)$ with minimal cardinality $N(\frac{1}{2}\delta_{[p]}, B^+_{l,p}(g), \|\cdot\|_\mathcal{Y})$. Since B is (M, τ_{asd})-homogeneous, $N(\frac{1}{2}\delta_{[p]}, B^+_{l,p}(g), \|\cdot\|_\mathcal{Y}) \leq M(e^d+3)^{\tau_{\text{asd}}}$. By taking the union $\mathcal{A}^+_{l,p}(g)$ with $\mathcal{A}_{l,p}(g)$, we have a $\frac{1}{2}\delta_{[p]}$-cover of B with cardinality at most $|\mathcal{A}_{l,p}(g)| + M(e^d+3)^{\tau_{\text{asd}}}$. So if $|\mathcal{A}'_{l,p}(g)| > M(e^d+3)^{\tau_{\text{asd}}}$, then we have found a $\frac{1}{2}\delta_{[p]}$-cover of B with cardinality strictly less than $N_{[p]}$, contradicting its minimality. Hence, we must have $\mathcal{A}_{l,p} \leq |\mathcal{A}'_{l,p}(g)| \leq M(e^d+3)^{\tau_{\text{asd}}}$. But the latter quantity is a constant that does not depend on δ or p. Thus

$$\mathcal{A}_l \leq \prod_{[p] \leq m-1} \mathcal{A}_{l,p} \leq M^{m^d} \left(e^d + 3\right)^{\tau_{\text{asd}} m^d}. \tag{†††}$$

Putting together (†), (††) and († † †), we arrive at

$$\begin{aligned} N\left(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty\right) &\leq \prod_{l=1}^{L} \mathcal{A}_l \\ &\leq M^{m^d} \left(\frac{4e^d K_B}{\delta}\right)^{\tau_{\text{asd}} m^d} M^{m^d K_2 \delta^{-\frac{d}{m}}} \left(e^d+3\right)^{\tau_{\text{asd}} m^d K_2 \delta^{-\frac{d}{m}}}, \end{aligned}$$

and so

$$\begin{aligned} H\left(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty\right) &\leq \delta^{-\frac{d}{m}} \log\left(M^{m^d K_2} \left(e^d+3\right)^{\tau_{\text{asd}} m^d K_2}\right) \\ &\quad + m^d \log\left(M \left(\frac{4e^d K_B}{\delta}\right)^{\tau_{\text{asd}}}\right) \\ &\leq K\delta^{-\frac{d}{m}}, \end{aligned}$$

where K is a constant depending on $M, m, d, K_2, \tau_{\text{asd}}$ and K_B. With the second term, we bounded $\log\left(\frac{1}{\delta}\right)$ by a constant times $\delta^{-\frac{d}{m}}$. □

Proof of Theorem 5.4.2. Suppose $g \in \mathcal{G}_B^m$. With notation as in the proof of Theorem 5.4.1, for each $l = 1, ..., L$ and each $p \in \mathbb{N}_0^d$ with $[p] \leq m-1$, we have

$$\mathcal{A}_{l,p} \leq N_{[p]} = N\left(\frac{\delta}{2\Delta^{[p]} e^d}, B, \|\cdot\|_{\mathcal{Y}}\right) \leq N\left(\frac{\delta}{2e^d}, B, \|\cdot\|_{\mathcal{Y}}\right) \leq \left(\frac{\delta}{2e^d}\right)^{-(\tau_{\text{box}}+1)},$$

where the second last upper bound follows since $N(\cdot, B, \|\cdot\|_{\mathcal{Y}})$ is a decreasing function, and we have $K_1 \geq 1$ and $0 < \delta < 1$, and the last upper bound follows from Equation (box) in Section 5.3. This upper bound has no dependence on l or p. So for each $l = 1, ..., L$, $\mathcal{A}_l \leq \left(\frac{2e^d}{\delta}\right)^{(\tau_{\text{box}}+1)m^d}$. Putting this together with (†), we arrive at

$$N\left(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty\right) \leq \prod_{l=1}^{L} \mathcal{A}_l \leq \left(\frac{2e^d}{\delta}\right)^{(\tau_{\text{box}}+1)m^d K_2 \delta^{-\frac{d}{m}}},$$

and so

$$H\left(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty\right) \leq (\tau_{\text{box}}+1) m^d K_2 \delta^{-\frac{d}{m}} \log\left(\frac{2e^d}{\delta}\right) \leq K\delta^{-\frac{d}{m}} \log\left(\frac{1}{\delta}\right),$$

where K is a constant depending on m, d, K_2 and τ_{box}. □

Proof of Theorem 5.4.3. Suppose $g \in \mathcal{G}_B^m$. With notation as in the proof of Theorem 5.4.1, for each $l = 1, ..., L$ and each $p \in \mathbb{N}_0^d$ with $[p] \leq m-1$, we have

$$\mathcal{A}_{l,p} \leq N_{[p]}$$

$$
\begin{aligned}
&= N\left(\frac{\delta}{2\Delta^{[p]}e^d}, B, \|\cdot\|_{\mathcal{Y}}\right)\\
&\leq N\left(\frac{\delta}{2e^d}, B, \|\cdot\|_{\mathcal{Y}}\right)\\
&\leq \exp\left\{M\left(\frac{\delta}{2e^d}\right)^{-\tau_{\exp}}\right\},
\end{aligned}
$$

where the second last upper bound follows since $N(\cdot, B, \|\cdot\|_{\mathcal{Y}})$ is a decreasing function, and we have $K_1 \geq 1$ and $0 < \delta < 1$. This upper bound has no dependence on l or p. So for each $l = 1, ..., L$,

$$
\mathcal{A}_l \leq \exp\left\{M\left(\frac{\delta}{2e^d}\right)^{-\tau_{\exp}} m^d\right\}.
$$

Putting this together with (†), we arrive at

$$
N\left(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty\right) \leq \prod_{l=1}^{L} \mathcal{A}_l \leq \exp\left\{M\left(\frac{\delta}{2e^d}\right)^{-\tau_{\exp}} m^d K_2 \delta^{-\frac{d}{m}}\right\},
$$

and so

$$
H\left(\delta, \mathcal{G}_B^m, \|\cdot\|_\infty\right) \leq M\left(\frac{1}{2e^d}\right)^{-\tau_{\exp}} m^d K_2 \delta^{-\frac{d}{m}-\tau_{\exp}} \leq K\delta^{-\left(\frac{d}{m}+\tau_{\exp}\right)},
$$

where K is a constant depending on m, d, M, K_2 and $\tau_{\exp}$. □

Theorem 5.5.3. *Let $S \in \mathbb{N}$ be any (large) integer. The empirical Rademacher complexity is bounded as*

$$
\hat{\mathfrak{R}}_n(\mathcal{G}) \leq 2^{-(S+1)}R_n + \frac{2}{\sqrt{n}}J_n,
$$

where we recall that $R_n = \sup_{g\in\mathcal{G}}\|g\|_{2,P_n}$ is the empirical radius and $J_n = \sum_{s=0}^{S} 2^{-s}R_n\sqrt{2H_{s+1}}$ is the uniform entropy bound.

Proof. See that

$$
\begin{aligned}
\hat{\mathfrak{R}}_n(\mathcal{G}) &= \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(X_i)\right\|_{\mathcal{Y}} \mid \mathcal{F}_n\right]\\
&= \mathbb{E}\left[\sup_{g\in\mathcal{G}}\|P_n^\sigma g\|_{\mathcal{Y}} \mid \mathcal{F}_n\right]\\
&= \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\|P_n^\sigma\left(g - g^{S+1}\right) + \sum_{s=0}^{S} P_n^\sigma\left(g^{s+1} - g^s\right)\right\|_{\mathcal{Y}} \mid \mathcal{F}_n\right]\\
&\leq \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\|P_n^\sigma\left(g - g^{S+1}\right)\right\|_{\mathcal{Y}} \mid \mathcal{F}_n\right]
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \mathbb{E}\left[\sup_{g \in \mathcal{G}} \left\| \sum_{s=0}^{S} P_n^\sigma \left(g^{s+1} - g^s\right) \right\|_{\mathcal{Y}} \mid \mathcal{F}_n \right] \\
&\leq \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \left\| g(X_i) - g^{S+1}(X_i) \right\|_{\mathcal{Y}} \\
&\quad + \sum_{s=0}^{S} \mathbb{E}\left[\sup_{g \in \mathcal{G}} \left\| P_n^\sigma \left(g^{s+1} - g^s\right) \right\|_{\mathcal{Y}} \mid \mathcal{F}_n \right] \\
&\leq \sup_{g \in \mathcal{G}} \left\| g - g^{S+1} \right\|_{2, P_n} \\
&\quad + \sum_{s=0}^{S} \mathbb{E}\left[\max_{k \in \{1, \ldots, N_{s+1}\}} \left\| P_n^\sigma \left(g_k^{s+1} - g_k^{s+1,s}\right) \right\|_{\mathcal{Y}} \mid \mathcal{F}_n \right] \\
&\leq 2^{-(S+1)} R_n \\
&\quad + \sum_{s=0}^{S} \frac{1}{\lambda_s} \log \left(\mathbb{E}\left[\sum_{k=1}^{N_{s+1}} e^{\lambda_s \left\| P_n^\sigma \left(g_k^{s+1} - g_k^{s+1,s}\right) \right\|_{\mathcal{Y}}} \mid \mathcal{F}_n \right] \right) && (a) \\
&\leq 2^{-(S+1)} R_n \\
&\quad + \sum_{s=0}^{S} \frac{1}{\lambda_s} \log \left(\sum_{k=1}^{N_{s+1}} \mathbb{E}\left[2\cosh\left(\lambda_s \left\| P_n^\sigma \left(g_k^{s+1} - g_k^{s+1,s}\right) \right\|_{\mathcal{Y}} \right) \mid \mathcal{F}_n \right] \right) && (b) \\
&\leq 2^{-(S+1)} R_n + \sum_{s=0}^{S} \frac{1}{\lambda_s} \log \left(2 \sum_{k=1}^{N_{s+1}} e^{\frac{\lambda_s^2}{n} (2^{-s} R_n)^2} \right) && (c) \\
&= 2^{-(S+1)} R_n + \sum_{s=0}^{S} \frac{1}{\lambda_s} \log \left(2 N_{s+1} e^{\frac{\lambda_s^2}{n} (2^{-s} R_n)^2} \right) \\
&= 2^{-(S+1)} R_n + \sum_{s=0}^{S} \frac{1}{\lambda_s} \left(H_{s+1} + \log 2 \right) + \frac{\lambda_s}{n} \sum_{s=0}^{S} (2^{-s} R_n)^2 \\
&\leq 2^{-(S+1)} R_n + \sum_{s=0}^{S} \frac{1}{\lambda_s} 2 H_{s+1} + \frac{\lambda_s}{n} \sum_{s=0}^{S} (2^{-s} R_n)^2 && (d) \\
&= 2^{-(S+1)} R_n + \frac{2}{\sqrt{n}} \sum_{s=0}^{S} 2^{-s} R_n \sqrt{2 H_{s+1}} && (e) \\
&= 2^{-(S+1)} R_n + \frac{2}{\sqrt{n}} J_n
\end{aligned}
$$

where, in (a), we used Jensen's inequality and the fact that the sum of positive numbers is greater than their maximum; in (b), we used the basic fact $e^x \leq 2\cosh x$; in (c), we used Hoeffding's inequality in Hilbert spaces; in (d), we used the fact that $H_{s+1} \geq \log 2$; and in (e), we let

$$\lambda_s = \frac{\sqrt{2nH_{s+1}}}{2^{-s}R_n}.$$

□

Lemma 5.5.5. *Suppose that for each $Y \in \mathcal{Y}$, the $\mathcal{Y} \to \mathbb{R}$ map $y \mapsto \mathcal{L}(Y, y)$ is c-Lipschitz for some constant $c > 0$, i.e. for $y_1, y_2 \in \mathcal{Y}$, $|\mathcal{L}(Y, y_1) - \mathcal{L}(Y, y_2)| \leq c\|y_1 - y_2\|_{\mathcal{Y}}$. Then for any $\delta > 0$, we have*

$$H(c\delta, \mathcal{L} \circ \mathcal{G}, \|\cdot\|_{2,P_n}) \leq H(\delta, \mathcal{G}, \|\cdot\|_{2,P_n}).$$

Proof. To ease the notation, write $N = N(\delta, \mathcal{G}, \|\cdot\|_{2,P_n})$, and let $g_1, ..., g_N$ be a minimal δ-covering of $\mathcal{G}$. Then for any $\mathcal{L} \circ g \in \mathcal{L} \circ \mathcal{G}$, there exists some g_j, $j \in \{1, ..., N\}$ with $\|g - g_j\|_{2,P_n} = (\frac{1}{n}\sum_{i=1}^{n}\|g(X_i) - g_j(X_i)\|_{\mathcal{Y}}^2)^{1/2} \leq \delta$. Then by the Lipschitz condition on $\mathcal{L}$,

$$\begin{aligned}
\|\mathcal{L} \circ g - \mathcal{L} \circ g_j\|_{2,P_n} &= \left(\frac{1}{n}\sum_{i=1}^{n}|\mathcal{L}(Y_i, g(X_i)) - \mathcal{L}(Y_i, g_j(X_i))|^2\right)^{\frac{1}{2}} \\
&\leq \left(\frac{1}{n}\sum_{i=1}^{n}c^2\,\|g(X_i) - g_j(X_i)\|_{\mathcal{Y}}^2\right)^{\frac{1}{2}} \\
&= c\,\|g - g_j\|_{2,P_n} \\
&\leq c\delta.
\end{aligned}$$

Hence $\mathcal{L} \circ g_1, ..., \mathcal{L} \circ g_N$ is a $c\delta$-covering of $\mathcal{L} \circ \mathcal{G}$, i.e.

$$N(c\delta, \mathcal{L} \circ \mathcal{G}, \|\cdot\|_{2,P_n}) \leq N(\delta, \mathcal{G}, \|\cdot\|_{2,P_n}).$$

Now finish the proof by taking logarithms of both sides. □