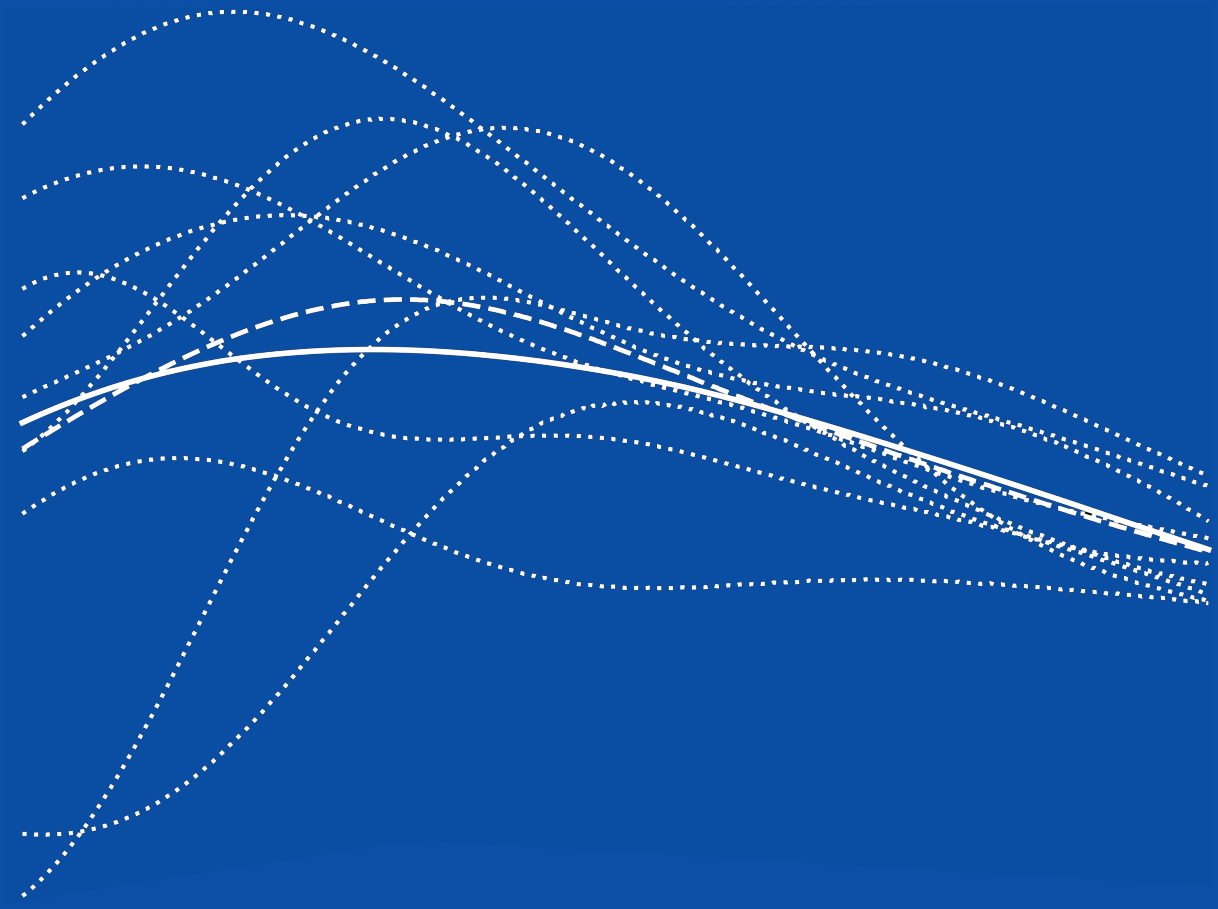
Proximal Causal Inference

Adjusting for the Unobserved



Author:
Francesco Di Giuseppe

Supervisors:
Prof. dr. A.W. van der Vaart

Committee member:
Dr. ir. G. F. Nane

TU Delft

.

"I like to understand simple things very well, because my brain is very slow"
-Michel Talagrand

Delft University of Technology
DIAM - Statistics

Master Thesis

Applied Mathematics

Proximal Causal Inference

Adjusting for the Unobserved

Author:
Francesco Di Giuseppe

Supervisors:
Prof. dr. A.W. van der Vaart

Committee member:
Dr. ir. G. F. Nane

to obtain the degree of Master of Science
at Delft University of Technology
on
Friday 30th August, 2024

Acknowledgments

The greatest acknowledgment goes to Professor Aad van der Vaart, without whom there would not have been this thesis. He consistently took the time to meet with me to discuss the many questions I had, as repetitive as they may have been. He allowed me the freedom to study whatever I found interesting, whilst still always pointing me in the right direction. His support extended beyond academics; when I felt my progress was slow, he reminded me that good things take time. I have grown a lot in the past months, both as a mathematician and a person, and he has played a significant role in both. I am very grateful to have had him as supervisor.
I would also like to thank Professor Tina Nane, who has supported me ever since I met her in the Decision Theory course. When other plans didn't work out as expected, she was there not only from an academic standpoint but also on a personal level. All the resources and support she has provided have helped getting me here.
On a personal note, a huge thank you goes to all those who have accompanied me throughout this journey. In particular, my second family at *Casa dello Studente* and the *friends from the* 4^{th} *floor*. You were the best company I could have asked for. A special mention goes to my partner Paula, who has endured all my ups and downs throughout these months and chose to spend many hours of her free time in the study rooms just to be by my side, holding my hand. Lastly, this thesis is dedicated to my family. I would not have been able to even begin studying mathematics without their support. Thank you to *i nonni e la zia* who have always been there for me, no matter how far, whether I was in Milan, Joensuu, or Delft.

To my mother, my brother Ale, and the memory of my father—whose smile and fist bumps have helped me through the toughest of times—this is for you.

Francesco Di Giuseppe

Delft, 2024

All errors are my own.

Abstract

Causal relationships are at the heart of the scientific method. The causal revolution of the 21^{st} century has opened the doors for many new approaches to quantify such relationships. In this thesis, we study the novel framework of proximal causal inference, which enables estimation of causal parameters even in the presence of unmeasured confounders, overcoming the limitations imposed by the Conditional Exchangeability assumption of the classic causal framework. In particular, we shall focus on determining and estimating the Causal Exposure Response Function (CERF) under this new set of assumptions. First, we introduce the problem and present a literature review of existing approaches to estimate bridge functions; then, we show theoretical results for extensions of classic linear results and a novel quasi-Bayesian method. This is then completed by showcasing performance on many simulated numerical examples and two real-world problems: Sustainable Causal Investing, and the effect of exercise on sleep.

Contents

1

Introduction

This thesis is aimed at graduate and undergraduate students in the field of mathematics and statistics. As such, it is structured in a total of 7 chapters. In this first introductory chapter, the history and concept of causality and the counterfactual framework are introduced. Furthermore, we highlight the pitfalls of the classic causal framework and introduce the advantages and shortcomings of the proximal approach. Next, preliminary concepts from functional analysis and mathematical statistics are introduced in preparation for the coming chapters. First we report a review of the state of the art and recent developments of the proximal inference literature. Afterwards, chapter 4 focuses on an extension to include higher order terms in two stage least square regression, standard error estimation using the delta method. Then, in chapter 5 we develop a novel application of the quasi-Bayesian approach to the problem of proximal inference. Afterwards we perform several numerical experiments to highlight and compare the new estimator. The thesis is then completed with an application to two concrete problems, causal sustainable investing and the effect of intense exercise on deep sleep. The work is completed with a conclusion and discussion on the topic. Throughout the thesis, we present results and proofs from various sources, reformulating and giving intuition behind them. Whenever the proofs are not informative or excessively technical, they will be omitted.

Introduction

Although one of statistics' initial interests was to determine causal relations between variables from observations, it was quickly dismissed in favor of studying correlation over causality. Sir Francis Galton is credited for introducing the concept of correlation to the field of statistics, which was later formalized by Karl Pearson. Both statisticians emphasized the fact that correlation is not causality, yet they did not provide a clear answer on how to define the latter. Due to its novel nature, correlation was then seen as the solution to all problems, and the search for causality was abandoned. Unfortunately, the opinion that correlation should be prioritized over causality quickly transformed into a dogma of the field, and statisticians avoided the concept altogether. At a similar time, in the 1920s, Sir Ronald Fisher was studying the effect of specific fertilizers on crop yield. Initially, he set up a basic experiment involving Latin squares, dividing the field into a grid, and placing one fertilizer into each grid. The results of this experiment were strongly biased, and Fisher recognized this. Due to the fact that the resulting outcome might be influenced by different factors such as preexisting fertility of the soil, sun exposure, or some other unaccounted for quantity, the experiments in each sub-square were not comparable. Repeating the experiment and placing the same fertilizer in the same grid would reduce the uncertainty of the effect of the fertilizer on the specific grid, but would it make the final result comparable with the ones in other grids? No, the interaction between fertilizers and their respective grids would remain and still make the final average results incomparable. In the end, Fisher carried out the experiment by repeatedly changing fertilizer and grid distribution. By doing so, he removed any potential confounders in the ground as each fertilizer was applied to a different plot of land, and also any personal bias in the fertilizer assigning process. The concept of randomizing, randomly assigning which grid receives which fertilizers, is the same method used today in modern experimental sciences: randomly decide which subject receives the control and which the treatment. Fisher had invented the randomized controlled trial. Unfortunately, as innovative as he was, Fisher would also propagate a new dogma, asserting that the randomized control trial was the only valid approach to determine causal relations between variables.

One of the most important discussions and debates revolving around the subject of health of the past century, is whether smoking is a significant cause of cancer or not. Although younger generations might

nowadays take this as an indisputable fact, up until half a century ago this question had no statistical, mathematical, or scientific tools to answer it. Many statisticians, whether because of personal belief or ulterior motives, supported the claims made by the large tobacco companies. Even Fisher himself defended the opinion that smoking and cancer were only spuriously correlated, stating even that the data could suggest a paradoxical protective effect [52].
At the beginning of the 20^{th} century, lung cancer was thought to be caused by environmental factors such as influenza, HPV, or various forms of air pollution and other chemicals and byproducts of the industrial revolution [45]. Although these exposures indeed contribute to a higher risk of cancerous cell development, overlooking the role of smoke in the lungs' health would be a significant oversight. Nonetheless, tobacco lobbies used other potential sources as means to discredit those who claimed that cigarette smoke increased the risk of cancer. As it is clear from this example, discerning the primary cause of an outcome can be difficult; many variables might influence and affect a result. As proposed in [54], diseases are multi-factorial and often have many causal sources. For these reasons, it is important to be able to objectively quantify the effects a certain variable might have on an outcome. Unfortunately, consumers were not made aware of the risks and dangers that smoking carries and were strongly deceived by the large tobacco companies. In this regard, the world 20 years ago was much different than it is today; Hollywood movies used cigarettes as product placement, teenagers were the primary target of advertisement, companies would receive support from physicians and statisticians alike, one could smoke on planes, and the list goes on. Nowadays, many countries in the world have placed bans on advertisements related to smoking, cigarette and other tobacco products carry pictures and texts to discourage the use, and smoking is not allowed in many public areas such as restaurants, airplanes, and hospitals. It was only in 1999 that the World Health Organisation (WHO), published a treaty to push nations to "protect present and future generations from the devastating health, social, environmental and economic consequences of tobacco".
Unfortunately the tools to prove the negative effects of smoking on health were not yet available. It should be of no surprise that the framework to answer questions regarding causality, even in non experimental settings, was born out of epidemiologists' and econometricians' needs. These needs would only be answered after the so called 'Causal Revolution' at the end of the century.
Causal inference is also a staple of modern econometric theory and it is often used to evaluate or even predict the effect of a certain policy, whether it be social or monetary. Guido Imbens and Joshua Angrist, two economists, were awarded the 2021 Sveriges Riksbank Prize in Economic Sciences[3], also referred to as the Nobel Prize in Economics. Their recognition stems from their contributions to the field of causality and its practical applications in real-world social advancements. In particular, their work includes determining the causal relationships between employment rates and minimum wage, as well as between education and income.
Returning to the smoking debate from a more econometric point of view, one might wonder the effectiveness of policies put in place to reduce tobacco use. In 1988 California passed Proposition 99, also known as the Tobacco Tax and Health Protection Act. This legislation aimed at reducing smoking rates by implementing various tobacco control measures such as funding anti-smoking education and prevention programs, and increasing taxes on tobacco products. In [2], the authors develop a causal method known as Synthetic Controls to determine whether or not the proposition had an effect on the sale and thus consumption of cigarettes. In particular, the Synthetic control method consists of constructing a least square copy of the test variable from various control variables pre-treatment. From this, one can compare reality to what would have happened had California not introduced Proposition 99. Assuming that the underlying machinery is correct, it is clear from figure 1.1 that the introduction of Proposition 99 helped to further reduce the already declining tobacco consumption in California. In this example, we used causal inference in a posteriori analysis, but it can also be used in a predictive manner, to decide whether or not to enact a certain policy in advance.

1.1 Counterfactuals, Experiments and Observations

There are many different taxonomies of data but the one of interest to the causal method is the observational/experimental categorization. The main objective of the scientific method is to understand and explain our surrounding world by studying *why* phenomena occur. This is done by proposing a hypothesis and then attempting to falsify it through experiments. Note that hypotheses are always of causal nature: *Gravity causes the apple to fall, smoking causes cancer,...*
Scientific theories aim to deduce causal relationships between events. Experiments, if carried out correctly, enable the collection of *controlled* data. The scientists perform experiments in a controlled environment or use randomization to reduce the influence of unaccounted variables called confounders. Unfortunately, experimental data can be extremely hard to obtain. The process of performing experiments is often an

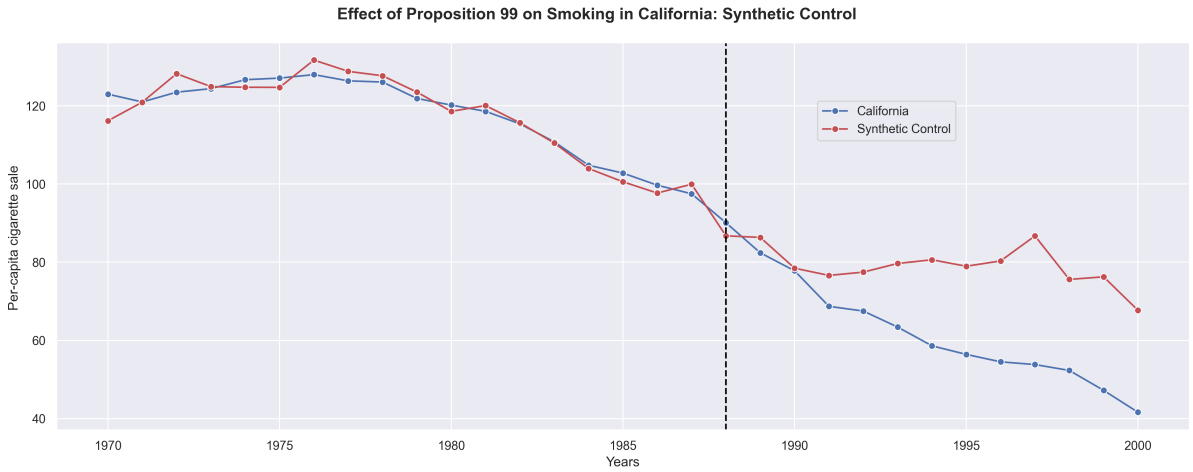


Figure 1.1: Synthetic control method to determine the effect of proposition 99 on smoking habits in California. Enactment date of Proposition 99 1988 (dashed line). Data from [2].

expensive one and, in certain situations, might not be feasible or even ethical. Returning to the smoking and cancer debate, tobacco companies and Fisher defended the opinion that smoking and cancer could be correlated through some unmeasured confounder [12]. In Fisher's eyes the only conclusive way to determine the effect of the former on the latter would be to set up an experiment and randomly assign a patient to either the smoking or control group. Coercing someone to consume a highly addictive and possibly carcinogenic substance such as tobacco or depriving them of it without their consent, but rather due to chance, would not be an ethical practice and thus the experiment becomes impossible. Similarly, in [43], a cohort of smelter workers was studied to determine the effect of sustained exposure to arsenic on mortality. A randomized control trial to test the true effect would be impractical as a random group of people cannot be reassigned occupation, for possibly many years, to determine causal relationships in the name of science. In situations such as these, one can only hope to collect observational data. People that smoke continue smoking, and smelter workers continue with their jobs. Unfortunately, the uncontrolled and non-experimental nature of the data generating process opens the door to the (possible) presence of confounders, carrying intrinsic biases. This is the reason Fisher was so dubious and reluctant to accept that smoke exposure had an effect on cancer, he believed that observational data was not valid.
The answer to causal questions lies behind questions *What if...?*. In [44], Rubin gives a clear and quantitative definition of causality:

> *"the causal effect of one treatment, E, over another, C, for a particular unit and an interval of time from t_1 to t_2 is the difference between what would have happened at time t_2."*

In the latter half of the 20^{th} century, philosophers David Lewis and Robert Stalnaker formalized the idea behind the *What if...?* question into the theory of counterfactuals[20]. This theory revolved around the concept of closest possible world, a 'potential' world most similar to ours except for one difference. For this reason, counterfactual outcomes are also referred to as potential outcomes. Returning to the agricultural experiments, from the closest world interpretation, the solution to the presence of confounders would be to place the same fertilizer on the entire field and compare the results between distinct possible worlds. Clearly this is infeasible, since one would need to have access to these other worlds to observe the results, counterfactual outcomes are key in deducing causal relations. The notation we adopt for counterfactual thinking is the following: Let Y be the outcome variable, then if treatment $A = E$ occurs the observed outcome is Y^E; otherwise, if the treatment assigned in C, the observed outcome would have been Y^C.
Fisher was not the only one interested in agricultural experiments, in a similar period, Jerzy Neyman also studied the problem. In his master's thesis, Neyman recognized the fact that comparing yields under different underlying soil conditions would pose a risk to the validity of the study [49]. Rather than asking about parallel worlds, Neyman was interested in statistically adjusting the observed data to somehow consider it as experimental. This theory starts to come to fruition in [44], where Rubin proposes a framework that matches measured confounders between treatment groups to estimate potential outcomes. Thus, if one were able to account for the confounding variables, then the observed data could be considered *controlled* and finally yield causal conclusions. This framework is known as Neyman-Rubin causal model; its assumptions and their violation will be the main topic of this thesis.

Many causally related events, especially in the clinical setting, are not deductive in nature but rather probabilistic. Not all smokers will eventually develop lung cancer and not all non-smokers will not develop lung cancer. Thus one would assume that the language of probabilities would be an excellent tool to measure causal relationships, but it is not so simple. From a probabilistic point of view, two events (A,B) are said to be independent if the joint law factorizes into the marginals $P(A, B) = P(A)P(B)$. When this is not the case, the two events are said to be dependent. Conditional probabilities quantify the level of association between the events by 'normalizing' the remaining probability of event A.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

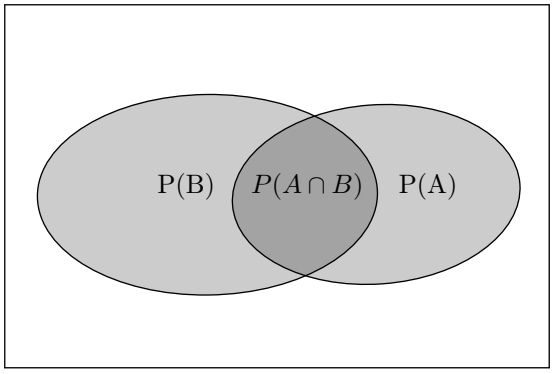


Figure 1.2: Graphical representation of conditional probabilities.

In the case of independent events($A \perp\!\!\!\perp B$), the joint probability factorizes into the two marginals and simplifies with the denominator, thus $P(A|B) = P(A)$. Similarly, the concept of independence can be extended to random variables, which are said to be independent if the joint distribution factorizes for all possible combinations of events. If two random variables X, Y are independent then the same rules as before apply and thus $P(Y, X) = P(Y)P(X)$. Notice that conditioning on a variable coincides with the act of observing since it answers the question:

What is the probability of $Y \in A$, having observed $X \in B$?

Although useful, observation is not the main approach used by the scientific method to deduce causal claims. Observation is only correlation. When performing experiments, scientists do not simply *observe* events but rather perform or *intervene* on the possible variables.

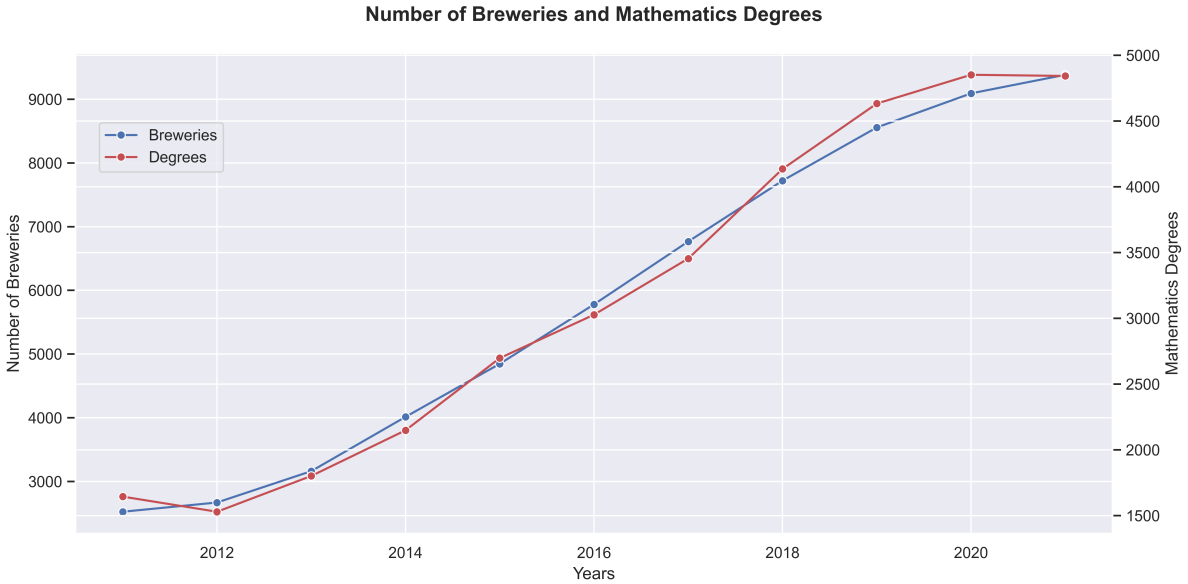


Figure 1.3: Number of breweries(blue) and number of associate degrees in Mathematics and Statistics (red) in the U.S. over the past 13 years. Data from [5],[14]

To highlight that correlation is not the same as causal relationships, data regarding the number of breweries and the number of associate degrees in mathematics was collected over the past 10 years ([5],[14]). The values are plotted in figure 1.3. There seems to exist an almost perfect correlation between number of mathematics degrees and number of breweries. Even deploying 'statistical tests', one observes an almost perfect correlation coefficient. Although mathematicians are known to often socialize in pubs, it is very unlikely that the relationship between the two events is a causal one and no scientist would conclude that either causes the other.
Unfortunately the language of probability is unable to capture the subtle difference between observation and intervention. There is a need for a new tool to account for interventions and then capture causal claims. From a probability point of view, given a treatment A and outcome Y, we say that A causes Y if the potential outcome Y^A is more likely than when it is not intervened on $Y^{\overline{A}}$. Thus a treatment has a causal effect on outcome if it increases the likelihood of the outcome occurring. Using the notation of both counterfactuals and probability, A causes outcome B $\iff P(Y^A = B) > P(Y^{\overline{A}} = B)$. Reading this in terms of possible worlds: A causes B if, the probability of observing B in the closest possible world where A occurs, is higher than in the one where it does not.

Problems

Although the 'Causal Revolution' has drastically changed the way causal quantities are estimated, it is not the end-all solution. The assumptions required by the counterfactual model are not always plausible and their violations would invalidate any drawn conclusions. Many have defined the current period in time as the 'Era of Big Data' due to the overwhelming abundance of it in every day life. It is hard to find an electronic device that does not collect and utilize data. The use of this data is also very vast. Doctors use patient information for personalized medicine, consumers are supported in the choice of which products to buy by target advertisements, factories are able to perform predictive maintenance on their equipment. The interest in data has been ever increasing, but in the words of Judea Pearl, 'data are profoundly dumb'[42]. The example presented in figure 1.3 highlights this. Passively collected data is only observational data and is nothing but a representation of the underlying mechanisms. If one is unable to describe the process which generates such data, it is of no use. The belief that data is the solution to every problem is very misleading. Data without structure is just a pile of numbers on paper (or computers). With structure here, we mean underlying dependencies between the data generating processes. Currently, the most common method of imposing such structures is through expert elicitation methods, such as interviews, queries, or structured expert judgement. Unfortunately, human intervention is not reasonable when the underlying process is not clear or the number of covariates is large. There has also been a large interest in developing 'automatic' structure learning algorithms. Algorithms such as the PC algorithm, the FCI or other variations blindly attempt to learn structural relations from correlation and specific structures that arise from the requirement that an event cannot cause itself. The philosophical debate is still open on the feasibility of learning structures directly from data, or rather can only be captured by some innate human capacity. Obviously, the separation between the two methods is not so harsh but rather there are in-betweens where expert opinion is either aided or combined with algorithms. Nonetheless, the proposal of wrong structural assumptions may lead to possibly wrong conclusions. A second problem, often used as reason to discredit the causal claims formulated using the Neyman-Rubin causal model, is the validity of the Conditional Exchangeability assumption (assumption 1.3). This assumption will be discussed in greater mathematical depth later on, in essence it requires that all confounders between treatment and effect are captured. This assumption enables us to adjust observational data and ensure 'artificial randomization' and use it as experimental data.
Proximal causal inference, a new framework inspired by experimental biology, is the central topic of this thesis. It aims to relax the conditional exchangeability assumption to enable the presence of certain unmeasured confounding. As is always the case, there is no free lunch. The ability to include quantities that are not observed when adjusting for potential outcomes comes as trade-off for a more stringent structural assumptions. In particular, it requires the existence and correct classification of negative controls or proxies, a particular type of confounders.

1.2 Counterfactual Framework

As previously introduced, the Nyeman-Rubin Causal framework enables us to convert observational data into experimental one. Although it is not the only one used to define causality, it is the most commonly adopted one. For this reason and ease of exposition, we shall refer to such model and its assumptions as the 'classic causal framework'. Before introducing the necessary assumptions of the model, consider the following example which shows that the naive approach of considering observational data as experimental can often lead to seriously wrong conclusions.

Example 1.1 (Naive Treatment Effect and Average Treatment Effect)

Consider an experiment with binary treatment $A \in \{0,1\}$ and binary outcome $Y \in \{0,1\}$. Suppose that all outcomes are available to us in table 1.1, both the observed (Highlighted in yellow) and true counterfactual ones.

	Potential Outcome										Sum
Y^0	0	0	0	0	0	1	0	1	0	0	2
Y^1	1	1	1	0	0	1	0	1	0	1	6

Table 1.1: Data for toy example to highlight difference between observations(highlighted in yellow) and counterfactuals.

If one were to only observe the highlighted values and calculate the NATE (Naive Average Treatment Effect) instead of the ATE by simply taking the difference of the observed conditional outcome:

$$NATE = \mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=0] = \frac{1}{5} - \frac{2}{5} = -\frac{1}{5}$$

If this were to be taken as the causal effect, the conclusion would be that the treatment has a negative effect on the outcome.

The true effect of the treatment is given by the Average Treatment Effect:

$$ATE = \mathbb{E}\left[Y^1 - Y^0\right] = \frac{6}{10} - \frac{2}{10} = +\frac{2}{5}$$

In this case, the conclusion is that the treatment has a positive (average) effect on the outcome.

1.2.1 Causal Quantities of Interest

In the socio-economic sphere, causality is often invoked to analyse and answer questions regarding policies' effectiveness, evaluating past impacts or future effects. The questions of interest are often along the lines of:

"What *if* policy A had not been enforced?",
"What will happen *if* policy A is enforced?"

Similarly, in the medical setting, one is interested in the effectiveness of a drug. In this case the question is:

"What *if* drug A is taken?"

In the rest of the work, without loss of generality, we will adopt the nomenclature from the clinical medicine setting and the variable that is the direct object of the proposition *What if...* will be denoted as the treatment.
Having already previously introduced it in Example example 1.1, the most commonly studied causal quantity is the average treatment effect(ATE):

Definition 1.1 (Average Treatment Effect)
Given a binary treatment taking values in $\mathcal{A} = \{0,1\}$, the Average Treatment Effect is defined as:

$$ATE = \mathbb{E}\left[Y^1 - Y^0\right]$$

Although denoted binary treatment, the random variable is often intended as a *TAKEN*(1) or *NOT TAKEN* (0). The ATE represents the expected improvement when taking the treatment with baseline reference no treatment(0). Similarly in the presence of a single treatment A with multiple treatment values $\mathcal{A} = \{1, 2, \dots\}$ is conventionally used as a set of labels and there is no ordinal nature to it. In this case the average effect of treatment k with respect to baseline 0 will be denoted as $ATE^k = \mathbb{E}\left[Y^k - Y^0\right]$. Alternatively, one may be interested in the causal effect of only the treatment $\mathbb{E}\left[Y^k\right]$ without comparison to baseline.

Definition 1.2 (Causal Exposure Response Function)
Given treatment A taking values in $\mathcal{A}$ and outcome Y, the Causal Exposure Response Function (CERF) is defined as:

$$\begin{aligned} \chi :& \mathcal{A} \to \mathbb{R} \\ & a \mapsto \mathbb{E}\left[Y^a\right] \end{aligned}$$

It is of particular interest the case where the treatment takes continuous values in $\mathcal{A} \subseteq \mathbb{R}$. Estimation of the CERF, under classic or proximal assumptions, will be of central relevance during this thesis. Although the CERF is not a parameter in the classic sense ($\theta \in \mathbb{R}$) but rather a function, we will often refer to it as a parameter of interest.

1.2.2 Classic Assumptions

The assumptions of the Neyman-Rubin causal model are the following:

Assumption 1.1 (Counterfactual Consistency)
Given the observed treatment A the counterfactual outcome Y^A is the observed outcome, i.e.

$$Y = Y^A$$

In case of binary treatment, this can also be written as $Y = AY + (1 - A)Y$.

Assumption 1.2 (Positivity)
Given a treatment A taking values in $\mathcal{A}$ the probability of observing treatment A is bounded away from 0 and 1 in each strata of L. In other words:

$$0 < P(A = a|L) < 1 \qquad \forall a \in \mathcal{A}$$

In the case of continuous treatment, $\mathcal{A} \subseteq \mathbb{R}$, the cumulative distribution function must be bounded away from 0 and 1.

Assumption 1.3 (Conditional Exchangeability)
All possible confounders between treatment A and outcome Y are captured in the set of variables L. This entails the following conditional independence:

$$Y^a \perp\!\!\!\perp A|L \quad \forall a$$

Assumption 1.1 ensures that the observed values are correctly defined. Interpreting it from Possible World theory, the world closest to the one where treatment $A = a$ is administered is the one we have observed and thus the outcome is the observed one. Assumption 1.2 allows us to use covariates L to match and thus transform observational data into counterfactual one. In the experimental setting, there would be no problem to begin with as observed outcome would controlled by laboratory techniques or random treatment assignment. Assumption 1.3 is the most important yet fragile pillar of the Neyman-Rubin causal model. Given the values of L, the outcome and treatment are independent, acting as a 'virtual randomisation'. The assumption is also known as no-unmeasured confounding since its presence could invalidate the conditional independence.
Now that the necessary assumptions are clear, we can return to Example example 1.1 where the bias found when wrongly estimating the ATE with the NATE is given by the following result.

Lemma (NATE bias)
In the case of binary treatment, under assumption 1.1 the following holds:

$$\begin{aligned} ATE = NATE + \left(\mathbb{E}\left[Y^1 - Y^0|A = 1\right] - \mathbb{E}\left[Y^1 - Y^0|A = 0\right]\right)\mathbb{E}\left[A\right] + \\ + \mathbb{E}\left[Y^1|A = 0\right] - \mathbb{E}\left[Y^1|A = 1\right] \end{aligned} \tag{1.1}$$

Proof. Since the treatment is binary, $\mathbb{E}[A] = P(A=1)$. Denote $\Delta = Y^1 - Y^0$

$$
\begin{aligned}
ATE &= \mathbb{E}[\Delta] \\
&= \mathbb{E}[\Delta|A=1]\,P(A=1) + \mathbb{E}[\Delta|A=0]\,P(A=0) && \text{(Total Prob.)} \\
&= (\mathbb{E}[\Delta|A=1] - \mathbb{E}[\Delta|A=0])\,\mathbb{E}[A=1] + \mathbb{E}[\Delta|A=0] \\
&= (\mathbb{E}[\Delta|A=1] - \mathbb{E}[\Delta|A=0])\,\mathbb{E}[A=1] + \mathbb{E}\left[Y^1|A=0\right] - \mathbb{E}\left[Y^0|A=0\right] \\
&= (\mathbb{E}[\Delta|A=1] - \mathbb{E}[\Delta|A=0])\,\mathbb{E}[A=1] + \mathbb{E}\left[Y^1|A=0\right] - \mathbb{E}\left[Y^0|A=0\right] \pm \mathbb{E}\left[Y^1|A=1\right] \\
&= (\mathbb{E}[\Delta|A=1] - \mathbb{E}[\Delta|A=0])\,\mathbb{E}[A=1] + \underbrace{\mathbb{E}\left[Y^1|A=1\right] - \mathbb{E}\left[Y^0|A=0\right]}_{\text{NATE}} + \mathbb{E}\left[Y^1|A=0\right] - \mathbb{E}\left[Y^1|A=1\right]
\end{aligned}
$$

□

The second term $\left(\mathbb{E}\left[Y^1 - Y^0|A=1\right] - \mathbb{E}\left[Y^1 - Y^0|A=0\right]\right)\mathbb{E}[A]$ in equation (1.1) is also known as differential effect bias. Differential effect bias refers to the systematic difference between how the treatment affects the treated and how it would affect the the control group. The third term $\mathbb{E}\left[Y^1|A=0\right] - \mathbb{E}\left[Y^1|A=1\right]$ is the selection bias. This occurs when the two groups are systematically different, even if the entire cohort had been treated.

Adjustment

Under the classic causal assumptions, observational data can be considered experimental by adjusting for confounders. Intuitively, if the underlying effect of other variables on treatment and outcome were known, then this effect could be statistically accounted for. The key theorem in this identification is the following:

Theorem 1.1 (g-formula)
Under assumption 1.1,assumption 1.3 the Causal Exposure Response Function (CERF) can be identified as:

$$\chi(a) = \mathbb{E}[Y^a] = \mathbb{E}_L\left[\mathbb{E}[Y|A=a,L]\right]$$

Proof.

$$
\begin{aligned}
\mathbb{E}[Y^a] &= \mathbb{E}_L\left[\mathbb{E}[Y^a|L]\right] && \text{(Tower Property)} \\
&= \mathbb{E}_L\left[\mathbb{E}[Y^a|L,A]\right] && \text{(Conditional Exchangeability)} \\
&= \mathbb{E}_L\left[\mathbb{E}[Y|L,A=a]\right] && \text{(Counterfactual Consistency)}
\end{aligned}
$$

□

Theorem 1.1 is also sometimes referred to as backdoor adjustment.

1.3 Proximal Framework

Theorem 1.1 shows the importance for assumption 1.3, yet there is a lot of scepticism around its validity. From a practical and experimental point of view, it might seem impossible to capture all common confounders without omitting some. Moreover, which variables should be collected? Once again, the results depend on the capacity of experts to correctly specify which items to focus on. Note that failing to include some confounder violates the randomisation within strata, returning to the starting steps and possibly lead to large biases in the estimation of the causal quantities of interest.
In parallel, one may argue that due to instrument recording precision or difficulty in capturing the variable, the *true* confounders are never measured but rather some (possibly) inaccurate representation of them. For example, there is currently no universal test to diagnose dementia. Often the testing procedure proceeds by steps, initially through cognitive tests, such as the Mini Mental State Examination (MMSE) or the Alzheimer's disease assessment scale-Cognitive (ADAS-Cog) and later through brain imaging techniques. The cognitive tests consists of questionnaires, the MMSE for example consists in 11 questions. Although indicative, the cognitive tests on their own cannot fully capture the current state of the disease but rather imperfect proxies that indicate the overall progression of the disease. Similarly, scans of the brain are not conclusive to determine the presence nor the level of the disease in patients, but can still be insightful for treatment.
The conditional exchangeability assumption is untestable as no statistical test can be carried out to verify that all confounders have been captured. To relax this assumption and accept the presence of unmeasured confounders, mathematicians and epidemiologists looked at other applied sciences for new methods. In particular, taking inspiration from biology and the experimental step of the scientific method, the proximal causal learning framework leverages what is known as the negative control method. This approach consists in deploying the experiment in such a way to ensure that no result is obtained. Under these specific conditions, if a result is detected then the presence of unmeasured confounders cannot be ruled out. In particular, the existence of two negative controls is required: Negative Control Outcome, and Negative Control Exposure.

Definition 1.3 (Negative Control Outcome)
A variable W is a Negative Control Outcome, or NCO, if it satisfies the following conditions:

- *W is known not to be caused by the treatment A.*
- *The association between treatment A and Negative Control Outcome W is subject to the same confounding mechanism as A and Y.*

Often in the observational medical studies, patients are required to self-administer the treatment. The effectiveness of the treatment depends on the consistency that a patient has in taking the treatment. While this might be perfectly controlled in an experimental environment, if not accounted for, might introduce bias into the observational study. As suggested by the Neyman-Rubin causal model, if one were able to account for this quantity, statistical adjustments could help remove the bias. Quantifying this parameter is no easy feat. One might argue that consistency in taking medication is related with the level of health consciousness an individual has. Highly health conscious individuals are more likely to stick to the medical regiment whereas non-health conscious individuals are more likely to forget. There are many methods to quantify health consciousness of an individual such as self reported levels, questionnaires, number of visits to the doctor and so on. Nonetheless, all of these methods are not truly quantifying the health consciousness of an individual but are in reality proxies, imperfect representations of the true unmeasured underlying factors.
All these variables can considered negative control outcome variables since it is an imperfect proxy for the overall health-consciousness of the individual but does not have a direct effect on whether the treatment is assigned or not.

Definition 1.4 (Negative Control Exposure)
A variable Z is a Negative Control Exposure, or NCE, if it satisfies the following conditions:

- *Z is known not to be a direct cause of the outcome Y.*
- *The association between outcome Y and Negative Control Exposure Z is subject to the same confounding mechanism as the treatment A and outcome Y.*

The negative control exposure variables are factors that only have a direct causal association with confounders and treatment. From a clinical perspective, negative control exposures are possible values or parameters that aid a doctor in the choice to administer the treatment, but are not direct causes of

outcome. A recurring example of a negative exposure in the literature [53][13] is CD4 cell count in assignment of HIV treatment. CD4 cells are a type of white blood cells essential to the immune system in fighting off diseases. HIV attacks these cells and prevents them from duplicating, thus leading to a lower count over time and a compromised immune system. This parameter is used by practitioners to decide whether or not to administer treatment, higher CD4 count patients are less likely to receive it. The value of a CD4 count test has no direct causal effect on the outcome but, acts as a representation for the overall health of the immune system, which confounds both treatment and outcome. Thus, CD4 count is a valid negative control[53].

1.3.1 Proximal Assumptions

The proximal inference framework leverages the existence and correct classification of the Negative Control Outcome and Exposure variables to admit certain types of unmeasured confounding. Assumptions 1.1 and 1.2 remain from the classic setting but now assumption 1.3 is exchanged for the following:

Assumption 1.4 (Proximal Exchangeability)
There exist negative control exposure variables Z, negative control outcome variables W and measured confounder X. Moreover the type of variable is correctly identified. This is equivalent to the following independence assumptions:

$$Y \perp\!\!\!\perp Z | A, X, U$$
$$W \perp\!\!\!\perp (A, Z) | X, U$$

Note that in the literature this assumption is not often called *Proximal Exchangeability*, but rather Negative Control Assumption or some other variation. Since it somewhat replaces the conditional exchangeability assumption we shall refer to it as such.

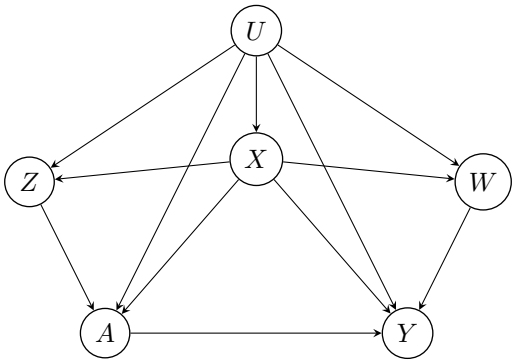


Figure 1.4: General Proximal Structure.

Figure 1.4 is a graphical representation of the proximal assumptions since it guarantees the negative controls independence requirements:

- Given A, X, U all paths between Y and Z are closed, thus Z is a valid NCE.
- Given X, U all paths between A, Z and W are closed, making W a valid NCO.

Although figure 1.4 gives a very intuitive interpretation of the structure needed for the problem, it is by no means the only one.

Example
The following graphical models are examples of potential proximal structures.

Additional assumptions, which we shall refer to as technical assumptions, are often made to ensure existence or regularity of the problem. A recurring assumption in the literature is the following.

Definition 1.5 (Completeness)
A set of distributions $\{\mathfrak{L}(U)\}$ is said to be complete if for every integrable function g:

$$\mathbb{E}[g(U)] = 0 \iff g(U) = 0 \quad a.s.$$

Similarly, given a set of conditional distributions $\{\mathfrak{L}(U|S = s)\ s \in \mathcal{S}\}$ is complete if for all measurable g and $\forall s \in \mathcal{S}$:

$$\mathbb{E}[g(U)|S = s] = 0 \iff g(U) = 0 \quad a.s.$$

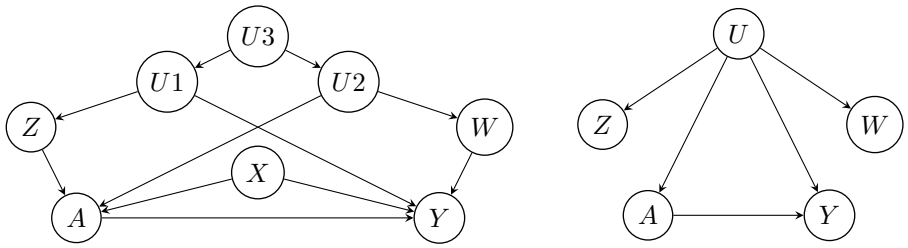


Figure 1.5: Examples of valid other proximal structures.

Whenever a set of conditional distributions $\mathfrak{L}(U|X)$ is complete we shall simply refer to $U|X$-completeness. In the proximal setting we will often require $U|Z, A, X$ or $U|W, A, X$ completeness. This assumption is often interpreted as a condition relating the range of U and the range of the proxies. In the sense that the proxies can capture the variability of the unmeasured confounder. In the case of categorical variables U,W,Z taking values on $\mathcal{U}, \mathcal{W}, \mathcal{Z}$ respectively, the condition is equivalent to requiring $\min(|\mathcal{Z}|, |\mathcal{W}|) \geq |\mathcal{U}|$.

Example 1.2 ([36]Independence $\nRightarrow$ Completeness)
Independence is not enough for completeness. Suppose that $U = (X_1, X_2, \cdots X_N)$ *and* $Z = X_1$ *with* $X_i \sim \mathcal{N}(0,1)$ *i.i.d. Then* $g(U) = \sum_{i=2}^{N} X_i$ *is measurable and* $g(U) \neq 0$ *a.s but*

$$\begin{aligned}
\mathbb{E}\left[g(U)|Z\right] &= \mathbb{E}\left[\sum_{i=2}^{N} X_i | Z\right] \\
&= \mathbb{E}\left[\sum_{i=2}^{N} X_i | X_1\right] \\
&= \mathbb{E}\left[\sum_{i=2}^{N} X_i\right] \\
&= \sum_{i=2}^{N} \mathbb{E}\left[X_i\right] = 0
\end{aligned}$$

In general, the conditional completeness assumption $U|S$ is violated if the space of functions $\mathcal{F}$ orthogonal to the density $f(u|s) \quad \forall s \in \mathcal{S}$ is not the zero element. Suppose $\mathcal{F} \perp f(u|s)$, and there exists $\mathcal{F} \ni g \neq 0$ then $\int_{\mathcal{U}} g(u) dP(u|s) = 0$ since the expectation can be seen as an inner product (and thus the notion of orthogonality is induced).

1.3.2 Bridge Functions

Recalling theorem 1.1, if the variable U were measured, the CERF could be identified by simply iterated conditioning as:

$$\chi(a) = \mathbb{E}\left[Y^a\right] = \mathbb{E}_{U,X}\left[\mathbb{E}\left[Y|A=a, X, U\right]\right]$$

Unfortunately, the unmeasured nature of U makes it impossible to apply such a formula. Nonetheless, the proximal exchangeability assumption 1.4 enables us to use specific conditional independencies to overcome this problem. In particular, so-called bridge functions enable us to estimate causal quantities.

Definition 1.6 (Outcome Bridge function)
The function $h : \mathcal{W} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$ *is an Outcome Bridge Function if it solves the following Fredholm integral equation of the first kind:*

$$\mathbb{E}\left[Y|Z, A, X\right] = \mathbb{E}\left[h(W, A, X)|Z, A, X\right] \tag{1.2}$$

The following theorem shows that, under an additional completeness assumption, the outcome bridge function above enables the identification of the CERF.

Theorem 1.2 (Outcome g-Formula)
Assume that proximal exchangeability, counterfactual consistency and positivity hold. Moreover assume $U|Z, A, X$*-completeness (definition 1.5), then if there exists an outcome bridge function* h*, the CERF is identified as:*

$$\chi(a) = \mathbb{E}\left[Y^a\right] = \mathbb{E}\left[h(W, a, X)\right]$$

Proof.

$$\mathbb{E}\left[Y|Z, A=a, X\right] = \mathbb{E}_U\left[\mathbb{E}\left[Y|Z, A=a, X, U\right]|Z, A=a, X\right]$$

$$\mathbb{E}\left[h(W,a,X)|Z, A=a, X\right] = \mathbb{E}_U\left[\mathbb{E}\left[h(W,a,X)|Z, A=a, X, U\right]|Z, A=a, X\right]$$

Subtracting the two terms and using the definition of outcome bridge function h:

$$\begin{aligned}
\mathbb{E}\left[Y|Z, A=a, X\right] - \mathbb{E}\left[h(W,a,X)|Z, A=a, X\right] &= 0 \\
\mathbb{E}_U\left[\mathbb{E}\left[Y - h(W,a,X)|Z, A=a, X, U\right]|Z, A=a, X\right] &= 0 \\
\mathbb{E}_U\left[\mathbb{E}\left[Y - h(W,a,X)|A=a, X, U\right]|Z, A=a, X\right] &= 0 \qquad \text{(NCE)} \\
\iff & \\
\mathbb{E}\left[Y - h(W,a,X)|A=a, X, U\right] &= 0
\end{aligned}$$

The proof is complete by applying theorem 1.1

$$\begin{aligned}
\mathbb{E}\left[Y^a\right] &= \mathbb{E}_{U,X}\left[\mathbb{E}\left[Y|A=a, X, U\right]\right] && \text{(g-formula)} \\
&= \mathbb{E}_{U,X}\left[\mathbb{E}\left[h(W,a,X)|A=a, X,\right]\right] && \text{(previous)} \\
&= \mathbb{E}\left[h(W,a,X)\right]
\end{aligned}$$

□

Due to the similarity to theorem 1.1, this theorem is often referred to as the proximal g-formula. Alternatively, exchanging the roles of NCO and NCE suggests the existence of another bridge function that enables identification, the Exposure Bridge function:

Definition 1.7 (Exposure Bridge Function)
The function $q : \mathcal{Z} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$ is an Exposure Bridge Function if it solves the following Fredholm integral equation of the first kind:

$$\mathbb{E}\left[q(Z, A, X)|W, A, X\right] = \frac{1}{f(A|WX)} \tag{1.3}$$

where $f(A|WX)$ is the conditional density function of A given W,X.

Once again, the following result proves to be a proximal version of the g-formula and enables identification of the CERF.

Theorem 1.3 (Exposure g-Formula)
Assume that proximal exchangeability, counterfactual consistency and positivity hold. Moreover assume $U|W, A, X$-completeness (definition 1.5), then if there exists an exposure bridge function q, the CERF is identified as:

$$\chi(a) = \mathbb{E}\left[Y^a\right] = \mathbb{E}\left[Y \cdot 1_{A=a} \cdot q(Z, a, X)\right]$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[q(Z, A, X)|W, A, X\right] &= \mathbb{E}_U\left[\mathbb{E}\left[q(Z, A, X)|W, A, X, U\right]|W, A, X\right] && \text{(Tower)} \\
&= \mathbb{E}_U\left[\mathbb{E}\left[q(Z, A, X)|A, X, U\right]|W, A, X\right] && (W \perp\!\!\!\perp (Z, A)|U, X)
\end{aligned}$$

$$\mathbb{E}\left[q(Z, a, X)|W, A, X\right] = \frac{1}{f(A|W, X)} \qquad \text{(Definition)}$$

$$= \mathbb{E}_U\left[\frac{1}{f(A|W, X, U)}|W, A, X\right] \qquad \text{(Tower* property)}$$

Subtracting the two, and by the assumed $U|W, A, X$-completeness, we have that:

$$\mathbb{E}_U\left[\frac{1}{f(A|W, X, U)} - \mathbb{E}\left[q(Z, A, X)|A, X, U\right]|W, A, X\right] = 0$$

$$\iff$$

$$\frac{1}{f(A|W,X,U)} = \mathbb{E}\left[q(Z,A,X)|A,X,U\right]$$

Applying theorem 1.1 to the CERF:

$$\begin{aligned}
\mathbb{E}\left[Y^a\right] &= \mathbb{E}_{X,U}\left[\mathbb{E}\left[Y \cdot 1_{A=a}|X,U\right]\right] \cdot \frac{1}{f(A=a|X,U)} \\
&= \mathbb{E}_{X,U}\left[\mathbb{E}\left[Y \cdot 1_{A=a}|X,U\right]\right] \cdot \mathbb{E}\left[q(Z,a,X)|A=a,X,U\right] && \text{(previous result)} \\
&= \mathbb{E}_{X,U}\left[\mathbb{E}\left[Y \cdot 1_{A=a} q(Z,a,X)|X,U\right]\right] && \text{(lemma A.11)} \\
&= \mathbb{E}\left[Y \cdot 1_{A=a} \cdot q(Z,a,X)\right]
\end{aligned}$$

□

The above theorems suggest that, if such bridge functions exist, their estimation is equivalent to the estimation of the CERF. Thus, the presence of unmeasured confounders is circumnavigated by using a specific data structure and the bridge functions as an intermediate step. Unfortunately, the proximal method has its downsides. Even when the problem permits it, estimation of the causal quantities is much more complicated than in the classic case, as will be explored in the next sections.

1.3.3 Classic vs. Proximal

Although it can be argued that both the conditional exchangeability and proximal exchangeability assumptions are untestable and must rely on expert judgment for correct identification, the latter is often considered less stringent. Obviously, the comparison is not so black and white since it depends on the data that can be collected an the structure of the problem.

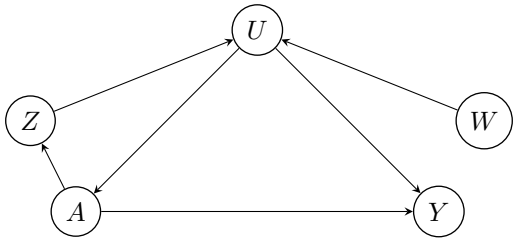


Figure 1.6: Graphical model where proximal exchangeability is not valid [53]

For example, figure 1.6 shows an invalid graphical model and structure. In this case, proximal exchangeability does not hold since we do not have $W \perp\!\!\!\perp (Z,A)|U$. Notice that if $W \to U$ was not present, the the arrow $Z \to U$ would not be a problem. The presence of $W \to U$ makes it such that U becomes a collider and conditioning on it opens the path $Z - U - W$.
Additionally, classic causal methods relying on the conditional exchangeability assumption are more developed and, as we will see later, often more relaxed on other technical assumptions. Nonetheless, the ability to overcome the presence of unmeasured confounding without direct modeling can be seen as strong advantage.

2

Background Mathematics

2.1 Hilbert Spaces

Definition 2.1 (Hilbert Space)
Let $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ be an inner-product on the vector space $\mathcal{H}$. $\mathcal{H}$ is a Hilbert Space if it is complete with respect to the norm $\|\cdot\|_{\mathcal{H}}$ induced by the above inner-product.

Hilbert spaces are vector spaces with an inner product, a generalization of euclidean vector spaces. With the additional requirement for completeness with respect to the norm induced by the inner product, the classic concepts of limits and convergence in terms of Cauchy sequences are well defined. In the rest of this thesis we shall focus on separable Hilbert spaces, these contain a countable dense subset, meaning we can always determine a countable orthonormal basis. If $\mathcal{H}$ is a finite dimensional vector space, then separability comes automatically.

Example 2.1
Let $\mathcal{H}$ be a vector space over $\mathbb{R}^2$, the 2-dimensional euclidean vector space. Let $a = (x_a, y_a), b = (x_b, y_b)$ be two vectors in $\mathcal{H}$. With the inner product $\langle a, b \rangle = x_a x_b + y_a y_b$, $\mathcal{H}$ is a Hilbert space.

The definition is general enough to include (possibly) infinite-dimensional objects such as functions. The points or elements of such spaces are functions, and since Hilbert spaces are vector spaces, scaling and countable sums are closed operations. Just as in elementary linear algebra, given a basis of the space, one can identify any element with a set of weights of said basis; the same holds for functions in Hilbert spaces[1]. Similarly to the finite-dimensional case, the notion of inner product as geometric projections remains, along with all the related nomenclature. Two elements of a Hilbert space are said to be orthogonal if their inner product is zero. Furthermore classic results such as the spectral theorem can be extended to linear operators, maps between function spaces that generalize the concept of matrices. All these considerations make Hilbert spaces of functions a natural working space.

Example 2.2
The space of Lebesgue-square integrable functions $L_2([0,1], \lambda) = \{f : \int_0^1 f(x)^2 d\lambda(x) < \infty\}$ equipped with the inner product $\langle f_1, f_2 \rangle = \int_0^1 f_1(x) f_2(x) d\lambda(x)$ is a Hilbert Space.

Important Results from Functional Analysis

As will be discussed in section 2.3, the Fredholm integral equations used to characterize bridge functions of proximal inference can be reformulated as operator problems. The following are important results from functional analysis regarding Hilbert spaces, and linear operators between them that will be re-occurring throughout this thesis. They are presented here along with some intuition.

Theorem 2.1 (Riesz Representation)
Let $\mathcal{H}$ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$. For every bounded linear functional $\Gamma \in \mathcal{H}^*$ there exists a unique element $\alpha \in \mathcal{H}$, called Riesz representer of Γ such that:

$$\Gamma f = \langle f, \alpha \rangle_{\mathcal{H}} \qquad \forall f \in \mathcal{H}$$

[1] Separable Hilbert spaces

The Riesz representer theorem tells us that every linear continuous functional on $\mathcal{H}$ can be represented by an element in $\mathcal{H}$ itself. Applying such functional to any element in $\mathcal{H}$ will be equivalent to projecting onto the representer itself. Large emphasis will be placed on the evaluation functional, due to their close relationship with reproducing kernel Hilbert spaces.
Just as function spaces can be seen as generalizations of Euclidean vector spaces $\mathbb{R}^n$, linear operators can be interpreted as generalizations of matrices, i.e. linear maps.

Definition 2.2 (Linear Operator)
Let $K : \mathcal{H}_1 \to \mathcal{H}_2$ be a linear map between two Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2$, i.e. $K(\lambda f) = \lambda \cdot K(f)$ and $K(f+g) = K(f) + K(g) \quad \forall f, g \in \mathcal{H}_1, \forall \lambda \in \mathbb{R}$. Such map is referred to as a Linear Operator.

Definition 2.3 (Adjoint Operator)
Let $K : \mathcal{H}_1 \to \mathcal{H}_2$ be a bounded linear operator between two Hilbert spaces, the adjoint of K is the operator $K^* : \mathcal{H}_2 \to \mathcal{H}_1$ such that:

$$\langle Kf, g \rangle_{\mathcal{H}_2} = \langle f, K^* g \rangle_{\mathcal{H}_2} \qquad \forall f \in \mathcal{H}_1, \forall g \in \mathcal{H}_2$$

One might now wonder if classic results for matrices also hold for these generalizations. This is often the case in Hilbert spaces for a particular type of operators, known as compact operators.

Definition 2.4 ([56]Compact Operator)
Let $K : \mathcal{H}_1 \to \mathcal{H}_2$ be a linear operator. K is said to be compact if it maps the unit ball of $\mathcal{H}_1$ into a relatively compact subset of $\mathcal{H}_2$, a set with compact closure. In particular, for Hilbert spaces K is the limit of a sequence of finite rank operators $\{K_i\}$, i.e., $Im(K_i)$ is a finite dimensional subspace in $\mathcal{H}_2$.

Compact operators can thus be seen as the *limit* of matrices. A classic result of interest is the spectral theorem, which guarantees that a $n \times n$ positive definite matrix A admits non-negative eigenvalues. A similar theorem exists for linear compact operators.

Proposition 2.1 (Spectral theorem for Operators)
Let $K : \mathcal{H}_1 \to \mathcal{H}_2$ be a compact linear operator and $K^* : \mathcal{H}_2 \to \mathcal{H}_1$ be its adjoint. By the spectral theorem, the self-adjoint compact operator $K^*K : \mathcal{H}_1 \to \mathcal{H}_1$ admits non-negative eigen-decomposition $\{\lambda_i^2, \varphi_i\}$ where $\{\varphi_i\}$ is a orthonormal basis of $\mathcal{H}_1$. Similarly, $KK^* : \mathcal{H}_2 \to \mathcal{H}_2$ admits eigen-decomposition $\{\lambda_i^2, \phi_i\}$ and $\{\phi_i\}$ is a orthonormal basis of $\mathcal{H}_2$. The values $\{\lambda_i^2\}$ are the eigenvalues of the operator and its adjoint.

By definition of compact operators, it follows that for infinite dimensional Hilbert spaces 0 is an accumulation point of the eigenvalues. As such, we always choose to order them in a decreasing order such that $\lambda_i \to 0$. We can also define the operator equivalent of singular value decomposition, the singular system.

Proposition 2.2 (Singular System)
Let $K : \mathcal{H}_1 \to \mathcal{H}_2$ be a compact linear operator. Then there exists $\{\lambda_i, \varphi_i, \phi_i\}$, where the quantities are defined as in proposition 2.1. The latter defined as the singular system of K and the values $\{\lambda_i\}$ are defined as its singular values.

In linear algebra, the singular values of a matrix inform us in which directions applying the matrix will *stretch* vectors. This can also be seen as a *loss* or *compression* of information in the direction of smaller singular values. A similar train of thought can be applied to linear operators, seeing them as 'infinite' dimensional matrices; larger singular values will carry more information.
Matrices can be represented as an expansion of their singular value decomposition, and since compact operators admit a singular system we might wonder if a similar result holds. An operator can be represented as a series expansion of the basis $\phi_i \otimes \varphi_i$ due to the following result. This highlights that the application of an operator is simply a (infinite) repeated projection onto its eigenfunctions and re-scaling.

Theorem 2.2 ([40] Series representation)
Let $K : \mathcal{H}_1 \to \mathcal{H}_2$ be a compact linear operator and $\{\lambda_i, \varphi_i, \phi_i\}$ be its singular system. Then K admits series representation:

$$K = \sum_i \lambda_i \phi_i \otimes \varphi_i$$

where $\forall f \in \mathcal{H}_1, g \in \mathcal{H}_2$ $(g \otimes f)z = \langle f, z \rangle g$. In other words, if $f = \sum f_i \varphi_i$:

$$Kf = \sum_i \lambda_i (\phi_i \otimes \varphi_i)(f) = \sum_i \lambda_i \langle f, \varphi_i \rangle \;\; \phi_i = \sum_i \lambda_i f_i \phi_i$$

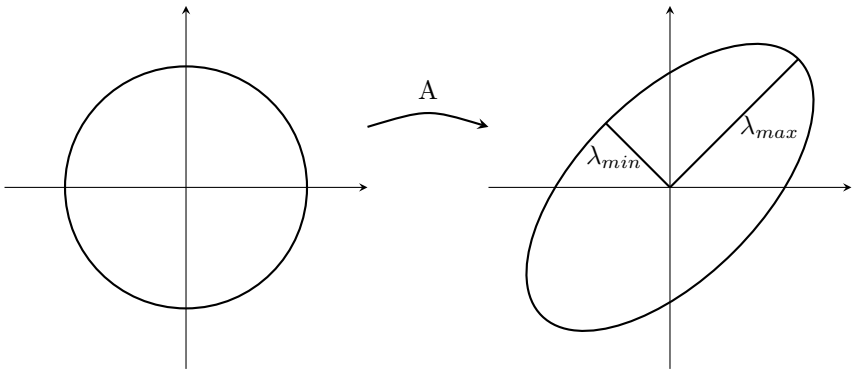


Figure 2.1: Two dimensional vector space before and after applying a matrix A with singular values $\lambda_{min}, \lambda_{max}$. Vectors along the direction of the eigenvector associated with the smallest eigen value are *compressed.*

Example 2.3
Just like in the two dimensional example of figure 2.1, theorem 2.2 highlights that the operator gives more importance to the higher valued eigenvalues λ_i^2. If function f is defined as $f_i = 1 \quad i = 1, 2$ and $f_i = 0$ otherwise, and $\lambda_1 > \lambda_2$, the previous representation highlights how operator K stretches the function more in the direction ϕ_1 than ϕ_2. The same idea could be applied for all trailing singular values if $f_i \neq 0$.

In the case of integral operators, proposition 2.1 is often referred to as Mercer's theorem. Additionally, this result enables to express the kernel of its integral operator as a series of eigenfuctions and eigenvalues.

Theorem 2.3 ([39] Mercer's theorem)
Let X be a compact space with strictly positive Borel measure μ. Let k be a symmetric, continuous, positive definite function, i.e. $\forall f \in L_2(X, \mu)$:

$$\int k(u, v) f(u) f(v) du dv \geq 0$$

Then the integral operator defined as:

$$T : L_2(X, \mu) \to L_2(X, \mu) \tag{2.1}$$

$$f \mapsto \int_X k(\cdot, x) f(x) d\mu(x) \tag{2.2}$$

admits eigen decomposition $\{\lambda_i^2, \varphi_i\}$ and its kernel can be written as:

$$k(x_1, x_2) = \sum \lambda_i^2 \varphi_i(x_1) \varphi_i(x_2)$$

Example 2.4 ([62])
Let $k(x, y) = (1 + xy)^2$ be the polynomial kernel over $[-1, 1] \times [-1, 1]$. For any function $f : [-1, 1] \to \mathbb{R}$:

$$\int_{-1}^{1} k(x, y) f(y) dy = \int_{-1}^{1} (1 + 2xy + x^2 y^2) f(y) dxy = \int_{-1}^{1} f(y) dy + 2x \int_{-1}^{1} y f(y) dy + x^2 \int_{-1}^{1} y^2 f(y) dy$$

Thus, the eigenfunctions of k must be polynomials of at most order 2, i.e. $f(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$.

The previous results rely on using compact integral operators. The following is a sufficient condition on the kernel to guarantee compactness of the associated operator.

Theorem 2.4 (Compactness of integral operators)
Let $T : L_2(X, \mu) \to L_2(X, \nu)$ be an integral operator with kernel k such that $Tf(x) = \int k(x, y) f(y) d\mu(y)$. A sufficient condition for T to be compact is:

$$\int \int |k(x, y)|^2 d\mu d\nu < +\infty$$

Moreover, by Proposition 7.8 of [40] T is compact $\iff$ the adjoint operator T^* is compact.

Since the majority of the results will revolve around the conditional expectation operator, it will always implicitly be considered compact. Using theorem 2.4, we prove mild conditions under which the operator is compact.

Proposition 2.3 (Compactness of Conditional Expectation)
Consider the conditional expectation operator:

$$\begin{aligned} E : L_2(X, P_X) &\to L_2(Y, P_Y) \\ f &\mapsto \mathbb{E}\left[f(X)|Y = \cdot\right] \end{aligned}$$

E is compact if the kernel is square integrable. If (X, Y) admits a density f_{XY} relative to a product measure $\mu \otimes \nu$, the kernel defined as

$$k(x, y) = \frac{f_{XY}(x, y)}{f_X(x) f_Y(y)}$$

characterizes the conditional expectation operator. If

$$\int \int f_{X|Y}(x|y) f_{Y|X}(y|x) d\mu(x) d\nu(y) < +\infty$$

then the operator E is compact.

Proof. By definition of conditional expectation and Bayes' theorem:

$$Eg(y) = \mathbb{E}\left[g(X)|Y = y\right] = \int_{\mathcal{X}} g(x) f_{X|Y}(x|y) d\mu(x) = \int_{\mathcal{X}} g(x) \frac{f_{XY}(x, y)}{f_Y(y)} d\mu(x) = \int_{\mathcal{X}} g(x) k(x, y) \underbrace{f_X(x) d\mu(x)}_{dP_X(x)}$$

Using Bayes theorem notice that:

$$\begin{aligned} \int \int |k(x, y)|^2 dP_X(x) dP_Y(y) &= \int \int |k(x, y)|^2 f_X(x) d\mu(x) f_Y(y) d\nu(y) \\ &= \int \int \left(\frac{f_{XY}(x, y)}{f_X(x) f_Y(y)} \right)^2 f_X(x) dx f_Y(y) d\nu(y) \\ &= \int \int \left(\frac{f_{X|Y}(x|y) \cdot f_{Y|X}(y|x)}{f_X(x) f_Y(y)} \right) f_X(x) d\mu(x) f_Y(y) d\nu(y) \\ &= \int \int f_{X|Y}(x|y) \cdot f_{Y|X}(y|x) d\mu(x) d\nu(y) \end{aligned}$$

Then, theorem 2.4 implies that E is compact. □

Finally, the following result enables us to talk about solutions to the operator problems that will be discussed in section 2.3. This result involves the singular systems of the operator. In particular, it requires that the source condition can be *reached* by the operator K and that the solution is well defined in $L_2(X, \mu)$.

Theorem 2.5 (Picard's Theorem [11])
Let $K : L_2(X, \mu) \to L_2(Y, \nu)$ be a compact linear operator and let $\{\lambda_i, \varphi_i, \phi_i\}$ be the associated singular system. The problem:

$$Kh = f_0$$

admits solution $\iff$ $f_0 \in \ker(K^*)^{\perp}$[2] *and* $\sum_i \frac{1}{\lambda_i^2} \langle f_0, \phi_i \rangle^2 < \infty$. *In such case, a solution is given by:*

$$h = \sum_i \frac{1}{\lambda_i} \varphi_i \langle f_0, \phi_i \rangle$$

Proof. Without loss of generality suppose that $h = \sum h_i \varphi_i$. The problem of determining h is equivalent to determining the weights $h_i = \langle h, \varphi_i \rangle$. By theorem 2.2:

$$\begin{aligned} Kh = \sum \lambda_i h_i \phi_i = \sum \langle f_0, \phi_i \rangle \phi_i \\ \iff \lambda_i h_i = \langle f_0, \phi_i \rangle \\ \iff \begin{cases} h_i = \frac{\langle f_0, \phi_i \rangle}{\lambda_i} & \lambda_i \neq 0 \\ \langle f_0, \phi_i \rangle = 0 & \lambda_i = 0 \end{cases} \end{aligned} \tag{2.3}$$

[2]From linear algebra we remember that $\ker(A)^{\perp} = Range(A^T)$. This is similar in infinite dimensional spaces except that we have to take the closure to ensure that the limit points are also included.

$f_0 \in \ker(K^*)^\perp \iff \langle f_0, w\rangle = 0 \quad \forall w \in \ker(K^*)$. A function w belongs to $\ker(K^*) \iff K^*w = \sum \lambda_i \langle w, \phi_i\rangle \varphi_i = \underline{0} \iff \lambda_i \langle w, \phi_i\rangle = 0 \quad \forall i$. By generality of w, when $\lambda_i = 0$ necessarily $\langle f_0, \phi_i\rangle = 0$.

□

Reproducing Kernel Hilbert Spaces

In applied mathematics, particular interest is placed on so called reproducing kernel Hilbert spaces. These spaces are interesting because they enable the use of Riesz representation (theorem 2.1) on the evaluation functionals. This property enables to transform optimization problems on RKHS to the kernel matrix and its empirical counterpart. Moreover, the functions found in RKHS automatically have certain regularity properties that might be desired when solving empirical risk minimization problems. In this section we will discuss reproducing kernel Hilbert spaces and their properties.

Definition 2.5 (Reproducing Kernel Hilbert Space)
Let k be a symmetric, positive definite function on $\mathcal{X} \times \mathcal{X}$. Define:

$$\mathcal{H}_0 = \left\{ f : f = \sum_i f_i k(x_i, \cdot) \quad f_i \in \mathbb{R}, x_i \in \mathcal{X} \right\}$$
$$\langle f, g\rangle_{\mathcal{H}} = \sum_{i,j} f_i g_j k(x_i, x_j)$$

Then $\mathcal{H} = \overline{\mathcal{H}_0}$ where the completion is taken with respect to the norm induced by the above inner product. $\mathcal{H}$ is a Reproducing Kernel Hilbert Space.

For a fixed kernel k, identifying a function in $\mathcal{H}$ is equivalent to determining the set of weights α_i. This is exactly the same as determining an element in a vector space by the coefficients of a given basis.

Proposition 2.4 (Bounded Evaluation Functional)
Let $\mathcal{H}$ be a reproducing kernel Hilbert space, then the evaluation functionals $\Gamma_x : \mathcal{H} \to \mathbb{R}$ are bounded, i.e. $\forall x \in \mathcal{X} \quad |\Gamma_x f| = |f(x)| \leq C_x \|f\|_{\mathcal{H}}$ where $\|\Gamma_x\| = \|k(x, \cdot)\|_{\mathcal{H}}$.

Proof. $k(x, \cdot)$ is also a function in $\mathcal{H}$ and as such it can be written as $k(x, \cdot) = \sum a_i k(x_i, \cdot)$. Without loss of generality consider $x = x_1$, then $k(x, \cdot)$ is uniquely identified by $a = (a_1, a_2, a_3, \cdots) = (1, 0, 0, \cdots)$. Thus:

$$\langle f, k(x, \cdot)\rangle = \sum_j \sum_i f_i a_j k(x_i, x_j) = \sum_i f_i k(x_i, x_1) = f(x_1)$$

For linear operators, continuity and boundedness coincide.

$$|\Gamma_x f| = |f(x)| = |\langle f, k(x, \cdot)\rangle| \leq \|k(x, \cdot)\|_{\mathcal{H}} \|f\|_{\mathcal{H}}$$

□

Boundedness of evaluation functionals is sometimes used to directly define RKHSs[39]. This condition ensures that the Riesz representer theorem (theorem 2.1) can be applied to guarantee the existence of the representer. Interestingly, the value of any function in the RKHS at point x is given by the 'infinite dimensional' projection of f onto the representer. Since Hilbert spaces are vector spaces, this consists in applying the operator to each element of the orthonormal basis $\{\varphi_i\}$ of $\mathcal{H}$ or, equivalently[3], projecting each element of the basis onto the Riesz representer:

$$\Gamma_x f = \Gamma_x \left(\sum_{i=1} f_i \varphi_i \right) = \sum_{i=1} f_i \left(\Gamma_x \varphi_i \right) = \sum_{i=1} f_i \varphi_i(x)$$
$$\Gamma_x f = \langle f, \alpha(x)\rangle_{\mathcal{H}} = \langle \sum_{i=1} f_i \varphi_i, \alpha(x)\rangle_{\mathcal{H}} = \sum_{i=1} f_i \langle \varphi_i, \alpha(x)\rangle_{\mathcal{H}} = \sum_{i=1} f_i \varphi_i(x)$$

In the first case, we use the continuity(boundedness) of Γ_x, whereas in the second case the continuity of the inner product. Due to definition 2.5 and proposition 2.4, one might notice that the defining kernel coincides with the representer $\alpha(x)$ of the evaluation kernel. This is often referred to as the *reproducing trick* of RKHS.

[3] The operator is continuous and thus $\lim_{n\to\infty} A\left(\sum^n \phi_i\right) = A\left(\lim_{n\to\infty} \sum^n \phi_i\right)$

Definition 2.6 (Representing Kernel and Feature Maps)
A bivariate function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called representing kernel if $\forall x \in \mathcal{X}$, the function $k(x, \cdot) \in \mathcal{H}$ is the Riesz representer of the evaluation functional Γ_x. The map $\varphi : \mathcal{X} \to \mathcal{H}$ such that $\varphi(x) = k(x, \cdot)$ is known as canonical feature map.

This raises the question if the same RKHS $\mathcal{H}$ can be generated by different kernels. The following theorem highlights that the correspondence between $\mathcal{H}$ and its reproducing kernel is one-to-one.

Theorem 2.6 ([39])
For every symmetric positive definite function k on $\mathcal{X} \times \mathcal{X}$ there exists a unique RKHS $\mathcal{H}$ with k as its reproducing kernel. Conversely, the reproducing kernel of a RKHS is unique and positive definite

Furthermore, since a linear operator's boundedness is equivalent to its continuity, functions living in a RKHS can be seen as being *sufficiently* regular. This *regularity* is determined by the norm of the evaluation operator in the following sense:

Lemma 2.1
Let f_1, f_2 be real-valued functions with support $\mathcal{X}$ in the Reproducing Kernel Hilbert Space $\mathcal{H}$, then:

$$|f_1(x) - f_2(x)| \leq \|k(x, \cdot)\| \, \|f - g\|_{\mathcal{H}}$$

Proof. Let Γ_x be the evaluation functional on the RKHS. By definition Γ_x is linear and since f_1, f_2 live in the RKHS Γ_x is bounded. Thus:

$$|f_1(x) - f_2(x)| = |\Gamma_x f_1 - \Gamma_x f_2| = |\Gamma_x (f_1 - f_2)| \leq \|\Gamma_x\| \, \|f_1 - f_2\|_{\mathcal{H}}$$

□

Lemma 2.1 shows that functions that are close in RKHS are also pointwise close. This property is extremely useful in function approximation since it ensures that minimizing RKHS distances also minimizes pointwise discrepancies, in other words f_1 that well approximates f_2 in RKHS will also be close when evaluated at x. If my function belongs to $\mathcal{H}$, I can approximate it as a linear combination of *simpler* functions. A possible idea is that of approximating $\mathcal{H}$ with a combination of $k(x_i, \cdot)$ might achieve sufficient accuracy in $\mathcal{H}$ and, in light of the lemma 2.1 on $\mathbb{R}$. Intuitively, increasing the number of basis elements will increase the precision in generating the space $\mathcal{H}^n$ that approximates the 'true' underlying space $\mathcal{H}$. Unfortunately, considering a fixed kernel also fixes the space taken into consideration and might preclude the presence of certain functions, more will be discussed later on.
The fact that RKHS are particular types of Hilbert spaces raises the question whether or not that all Hilbert spaces are RKHS. Unfortunately this is not the case. The following example shows that the set of square integrable functions on $[0, 1]$ is not a RKHS.

Example 2.5 ([62] Not all Hilbert spaces are Reproducing)
Although as shown in example 2.2, $L_2([0, 1], \lambda)$ is a Hilbert space, it is not a reproducing kernel Hilbert space. Consider the sequence $f_n(x) = x^n$. Since $\int_0^1 f_n^2(x)dx = \int_0^1 x^{2n} dx = \frac{x^{2n+1}}{2n+1}\Big|_0^1 = \frac{1}{2n+1}$ $f_n \in L_2$ $\forall n \in \mathbb{N}$ and $\|f_n\|_2 \to 0$. However we do not have pointwise convergence since $\forall n \in \mathbb{N}$ $f_n(1) = 1$. This contradicts lemma 2.1 and thus $L_2([0, 1], \lambda)$ is not a reproducing kernel Hilbert space.

Integral Definition

Associated with each kernel is its integral operator. The RKHS $\mathcal{H}$ can be characterized by the spectral decomposition of such integral operator. The regularity of the functions in $\mathcal{H}$ can be expressed in terms of such decomposition.

Definition 2.7 (Kernel Operator)
Let (X, μ) be a metric space, then for every positive definite kernel k, we define the kernel operator of k as:

$$\begin{aligned} T : L_2(X, \mu) &\to L_2(X, \mu) \\ f &\mapsto \int f(x) k(x, \cdot) d\mu(x) \end{aligned}$$

Moreover, the adjoint of a kernel operator is the integral operator in the other variable, i.e.:

$$\begin{aligned} T^* : L_2(X, \mu) &\to L_2(X, \mu) \\ g &\mapsto \int g(y) k(\cdot, y) d\mu(y) \end{aligned}$$

Since the kernel operator of $\mathcal{H}$ is defined by a symmetric positive definite function, T is self-adjoint. Under additional mild restrictions on k, T is compact and satisfies the requirements for *theorem* 2.3*and proposition* 2.1, and admits eigendecomposition $\{\lambda_i^2, \varphi_i\}$. There exists[4] an operator S such that $S^2 = T$. Defining $T^{1/2} = S$ the square root of the operator T, it will admit series expansion with its singular values rather than the eigenvalues. It can be shown[48] that $T^{1/2}$ is an isometry between $L_2(X, \mu)$ to $\mathcal{H}$. This enables us to write:

$$\begin{aligned} \langle f, g \rangle_{\mathcal{H}} &= \left\langle T^{-1/2} f, T^{-1/2} g \right\rangle_{L_2} \\ \|f\|_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}} &= \left\langle T^{-1/2} f, T^{-1/2} f \right\rangle_{L_2} = \left\| T^{-1/2} f \right\|_2^2 \end{aligned} \tag{2.4}$$

Since $T^{-1/2}$ is an isometry, the set of functions $\{\lambda_i \varphi_i\}$ form an orthonormal basis of our RKHS $\mathcal{H}$. By theorem 2.3, this space will be the same as the RKHS associated with the kernel k and it can be characterized in terms of the spectral decomposition of T as:

$$\mathcal{H} = \{ f = \sum f_i \varphi_i : \|f\|_{\mathcal{H}}^2 = \sum \frac{f_i^2}{\lambda_i^2} < +\infty \} \tag{2.5}$$

Note that the kernel operator is defined over $L_2(X, \mu)$ and thus both the choice of kernel k and the underlying measure μ play a role in its eigendecomposition, and thus in the corresponding RKHS $\mathcal{H}$. The formulation in terms of the eigenvalues of T highlights the implicit regularization effect of working in said RKHS. Larger eigenvalues, permit larger or *rougher* functions to be included in $\mathcal{H}$. If the eigenvalues were exactly equal to one, i.e. $\lambda_i = 1$ then the space would coincide with L_2 directly. This is not possible since the infinite dimensional identity operator is not compact.

Power Spaces

Various square matrix operations such as the square root are defined in terms of the spectral decomposition by simply applying the operation to the eigenvalues. Similarly, powers of the kernel operator T can also be defined. The reproducing kernel Hilbert spaces associated to the powers T will then contain functions that have finite norm with respect to the adjusted eigenvalues.

Definition 2.8 ([50]Power Spaces)
Let $\{\lambda_i^2\}$ be the eigenvalues of the kernel operator T that characterizes $\mathcal{H}$. The Power Space

$$\mathcal{H}^\alpha \quad \forall \alpha \in (0, 1]$$

associated with T^α is the subspace of L_2:

$$\mathcal{H}^\alpha = \{ f \in \mathcal{H} : \sum \frac{f_i^2}{\lambda_i^{2\alpha}} < +\infty \}$$

Functions contained in the power spaces depend on the value α. Intuitively, small values of α translate eigenvalues of T^α will decay more quickly. This means that power spaces will include *rougher* or larger functions than $\mathcal{H}$. Similarly, large values(close to 1) will not greatly affect the eigenvalues and thus contain similar regularity functions. As $\alpha \to 0$ the eigenvalues of T^α tend to 1 and thus the power space tends to L_2.

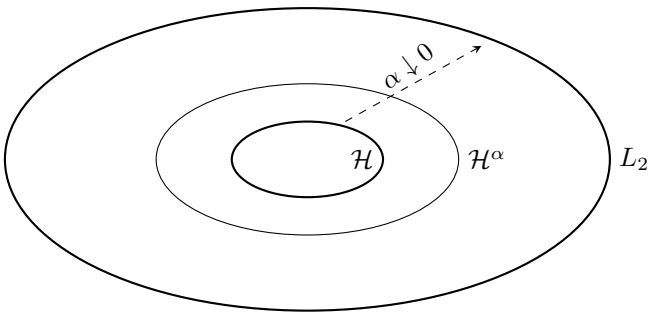


Figure 2.2: Graphical representation of Power spaces inclusions.

[4]Proposition 10.58 of [40]

Measures

From a measure theoretic perspective, the evaluation functional coincides with expectation with respect to a point mass measure. Let δ_{x_0} be the Dirac measure at $x = x_0$, then the evaluation of a function at point $x = x_0$ can be interpreted as applying the operator $\iota_{x_0} f = \int_{\mathcal{X}} f(t) d\delta_{x_0}(t) = f(x_0)$. Thus, if the function f lives in the RKHS, the operator of $\iota_{x_0} : \mathcal{H} \to \mathbb{R}$ such that $\iota_{x_0}(\cdot) = \int_{\mathcal{X}} \cdot d\delta_{x_0}(t)$ coincides with the evaluation operator Γ_{x_0}, and thus the Riesz representer is identical for both:

$$f(x_0) = \int_{\mathcal{X}} f(t) d\delta_{x_0}(t) = \int_{\mathcal{X}} \langle f, k(t, \cdot) \rangle d\delta_{x_0}(t) = \langle f, \int_{\mathcal{X}} k(t, \cdot) d\delta_{x_0}(t) \rangle = \langle f, k(x_0, \cdot) \rangle_{\mathcal{H}}$$

More generally, a measure on $\mathcal{X}$ is simply a weighted average of certain subsets of $\mathcal{X}$ where the weights are assigned by the 'measure' of the set. If the support $\mathcal{X}$ is finite with elements $\{x_1, x_2, \ldots, x_n\}$, then

$$\mu = \sum_{i=1}^{n} \alpha_i \delta_{x_i} \quad \alpha_i \geq 0$$

defines a measure. Now, if $f : \mathcal{X} \to \mathbb{R}$ is a measurable function, what about the map $\int f(x) d\mu(x)$? By linearity of integration:

$$\int_{\mathcal{X}} f(t) d\mu = \int_{\mathcal{X}} \sum \alpha_i f(t) d\delta_{x_i}(t) = \sum \alpha_i \int_{\mathcal{X}} f(t) d\delta_{x_i}(t) = \sum \alpha_i f(x_i) \tag{2.6}$$

A probability measure is just a particular type of measure such that $\sum_{i=1}^{n} w_i = 1$. In such case, equation (2.6) coincides with the common notion of expected value. This idea can be extended to more complicated objects, in which case for a random variable X taking values on $\mathcal{X}$, we will require that the kernel on $\mathcal{X}$ to be measurable.

Definition 2.9 (Mean Embedding)
Let X be a random variable taking values on $\mathcal{X}$ with distribution P_X, and $\mathcal{H}$ be a RKHS with measurable reproducing kernel k. The mean embedding of P_X in $\mathcal{H}$ is the map

$$\begin{aligned} \mu : & \mathfrak{L}(X) \to \mathcal{H} \\ & P \mapsto \int_{\mathcal{X}} k(x, \cdot) dP(x) \end{aligned}$$

A kernel k is said to be characteristic if the above map is injective.

In light of the following theorem, by a slight abuse of notation given a distribution P we also refer to the element of RKHS $\mu_P = \mu[P]$ as the mean embedding of P.

Lemma 2.2
Let $\mathcal{H}$ be a RKHS with reproducing kernel k. Let X be a random variable taking values in $\mathcal{X}$, with $X \sim P_X$. If $\mathbb{E}\left[\sqrt{k(X, X)}\right] < \infty$, then $\mu_{P_X} \in \mathcal{H}$ and

$$\mathbb{E}[f(X)] = \langle f, \mu_{P_X} \rangle$$

Proof. The expectation functional $\mathbb{E}_{P_X}[f(X)]$ is linear by definition. If $\mathbb{E}_{P_X}\left[\sqrt{k(X, X)}\right] < \infty$, it is also a bounded functional for all functions f in RKHS $\mathcal{H}$ since:

$$\begin{aligned} |\mathbb{E}[f(X)]| &\leq \mathbb{E}[|f(X)|] = \mathbb{E}[|\langle f, k(X, \cdot) \rangle|] \\ &\leq \mathbb{E}\left[|\sqrt{\langle f, f \rangle} \sqrt{\langle k(X, \cdot), k(X, \cdot) \rangle}|\right] = \mathbb{E}\left[|\sqrt{\langle f, f \rangle} \sqrt{k(X, X)}|\right] = \mathbb{E}\left[|\, \|f\|_{\mathcal{H}} \sqrt{k(X, X)}|\right] \end{aligned}$$

Where the first inequality follows from Jensen's and the second is Cauchy Schwarz. Since this operator is linear and bounded, it satisfies the requirements of the Riesz representation theorem and thus admits representer. This Riesz representer α must satisfy:

$$\begin{aligned} \langle \alpha, f \rangle &= \mathbb{E}[f(X)] \\ \langle \mu_{P_X}, f \rangle = \langle \int_{\mathcal{X}} k(x, \cdot) dP_X(x), f \rangle &= \int_{\mathcal{X}} \langle k(x, \cdot), f \rangle dP_X(x) = \int_{\mathcal{X}} f(x) dP_X(x) = \mathbb{E}[f(X)] \end{aligned}$$

Thus $\mu_{P_X} = \alpha \in \mathcal{H}$ and is the Riesz representer of $\mathbb{E}[\cdot]$ □

In other words, for sufficiently regular distributions, the Riesz representer of the expectation functional with respect to P_X is uniquely identified by the mean embedding. This means that given any function living in a RKHS, calculating the expectation of a function is equivalent to projecting it onto the mean embedding.

Conditional Mean Embedding

Since bridge functions of proximal inference definition 1.6 are characterized by equations involving conditional expectation, the marginal mean embedding might not be the correct object of interest. What is needed is a similar mathematical tool for the conditional expectation. Note that whereas marginal expectation is a functional, it takes in an element of $\mathcal{H}$ and returns a number, conditional expectation is an operator, a map between two function spaces $L_2(X, \mu)$ and $L_2(Y, \nu)$. The function that is returned is the least squares projection onto the space of square integrable functions in Y. To find the conditional mean embedding, we must first introduce some mathematical machinery that enables us to work with the covariance operators, these topics will also resurface in section 3.4.

$$\mathcal{H} \xrightarrow{\mathbb{E}\left[\cdot|Y=\cdot\right]} \mathcal{G} \longrightarrow \mathbb{R}$$

$$\mathcal{H} \xrightarrow{\mathbb{E}_X\left[\cdot\right]} \mathbb{R}$$

Figure 2.3: Representation of expectation acting between spaces.

Definition 2.10 (Covariance Operator)
Let (X,Y) be random variables taking values in $\mathcal{X}, \mathcal{Y}$ respectively. Moreover let $\mathcal{H}$ and $\mathcal{G}$, be two RKHS on $\mathcal{X}, \mathcal{Y}$ respectively. The Cross Covariance Operator is the unique bounded operator $C_{YX} : \mathcal{H} \to \mathcal{G}$ such that:

$$cov(f(X), g(Y)) = \langle g, C_{YX} f\rangle_{\mathcal{G}}$$

The Covariance Operator[5] is defined when $Y = X$ as the unique bounded operator C_{XX}:

$$cov(f_1(X), f_2(X)) = \langle f_1, C_{XX} f_2\rangle_{\mathcal{H}} = \langle C_{XX} f_1, f_2\rangle_{\mathcal{H}}$$

The covariance operator is similar to the kernel operator definition 2.7, whose eigenfunctions span the RKHS $\mathcal{H}$, since $cov(f_1(X), f_2(X)) = \int f_1(x) f_2(x) dP_X(x)$. The two should not be confused since T is an operator from $L_2(X, P_X)$ whereas C_{XX} is only well defined over $\mathcal{H}$. Nonetheless the two operators share eigenvalues and are equal on the the restriction of T to the reproducing kernel Hilbert Space $\mathcal{H}$.

Theorem 2.7 ([39])
If the function $\mathbb{E}\left[g(Y)|X = \cdot\right] \in \mathcal{H}$ for $g \in \mathcal{G}$ then:

$$C_{XX}\mathbb{E}\left[g(Y)|X = \cdot\right] = C_{XY} g$$

The condition that $\mathbb{E}\left[g(Y)|X = \cdot\right] \in \mathcal{H}$ is commonly referred to as the well-specified assumption and, as will be discussed in section 2.3, it is a strong requirement. Nonetheless, the conditional mean embedding is defined as:

Definition 2.11 (Conditional Mean Embedding)
Let $C_{XX} : \mathcal{H} \to \mathcal{H}$ and $C_{YX} : \mathcal{H} \to \mathcal{G}$ be the covariance and cross-covariance operators defined in definition 2.10, the Conditional Mean Embedding is the operator

$$\begin{aligned} U_{Y|X} :& \mathcal{H} \to \mathcal{G} \\ & f \mapsto C_{YX} C_{XX}^{-1} f \end{aligned}$$

By slight abuse of notation, we also denote $\mu_{Y|x} = C_{YX} C_{XX}^{-1} k(x, \cdot) \in \mathcal{G}$ as conditional mean embedding.

Theorem 2.7 is implicitly used to define the conditional mean embedding of definition 2.11. This definition implicitly carries the well-specified assumption. The conditional mean embedding $U_Y|X$ is an operator between spaces $\mathcal{H}, \mathcal{G}$. It can also be seen as an element of the tensor product space $\mathcal{H} \otimes \mathcal{G}$. By properties of reproducing of the reproducing kernel Hilbert spaces, this is isomorphic to the space of bivariate functions $X \times Y \mapsto \mathbb{R}$. Fixing either $X = x$ or $Y = y$ makes it such that the $U_y|X \in \mathcal{H}$ or $U_Y|X \in \mathcal{G}$. Definition 2.11 enables us to use the conditional mean embedding like the marginal mean embedding. The following corollary shows these properties.

[5] These are the non-centered covariance operators. This specification is necessary because some authors define the covariance operators as $\mathbb{E}\left[[(k_X(X, \cdot) - \mu_X) \otimes (k_Y(Y, \cdot) - \mu_Y)]\right]$

Corollary 2.1 ([39])
If the requirements of theorem 2.7 are satisfied, then the conditional mean embedding of definition 2.11 respects the following properties:

$$\mathbb{E}\left[g(Y)|X=x\right] = \langle g, \mu_{Y|x}\rangle_{\mathcal{G}}$$

Proof. If $\mathbb{E}\left[g(Y)|X=\cdot\right] \in \mathcal{H}$ then we can use the reproducing property of $\mathcal{H}$.

$$\mathbb{E}\left[g(Y)|X=x\right] = \langle \mathbb{E}\left[g(Y)|X=\cdot\right], k(x,\cdot)\rangle = \langle C_{XX}^{-1}C_{XY}g, k(x,\cdot)\rangle = \langle g, C_{YX}C_{XX}^{-1}k(x,\cdot)\rangle$$

□

In this corollary we have shown that the conditional mean embedding behaves just like the marginal mean embedding. The application of the conditional expectation operator to any element coincides with the projection of the element onto the conditional mean embedding.

Problems

The mean embedding of definition 2.9 is the Riesz representer of the expectation functional, the only requirement for it to be well defined is that the kernel is P_X measurable. On the contrary, the conditional mean embedding is not directly justified by Riesz theorem but rather theorem 2.7. For it to be defined we need well-specification which requires that given any function $g \in \mathcal{G}$ its conditional expectation belongs to $\mathcal{H}$ as a function of x. Theorem 2.6 enables us to identify the RKHS given a positive definite kernel, thus for a fixed kernel the associated RKHS only contains functions that can be written as the limit of the linear span of said kernel. Even simple functions such as constants might not belong to certain RKHS.

Example 2.6 (Corollary 4.44 of [50])
Suppose that $\mathcal{H}$ and $\mathcal{G}$ are RKHS generated by Gaussian Kernels $k(x,y) = e^{-(x-y)^2}$. Moreover assume that X and Y are independent. Then, fixing $g \in \mathcal{G}$, $\mathbb{E}\left[g(Y)|X=\cdot\right] = \mathbb{E}\left[g(Y)\right]$ is a constant function in x. The only constant function in the RKHS generated from the Gaussian Kernel is the null function.

Additionally, the conditional mean embedding does not always act as expected. Consider two independent random variables X, Y, one would expect the conditional mean embedding $\mu_{Y|x}$ to coincide with the marginal mean embedding μ_Y. Unfortunately this is not the case since:

$$\begin{aligned} X \perp\!\!\!\perp Y &\implies cov(f(X), g(Y)) = 0 \quad \forall f \in \mathcal{H} \quad \forall g \in \mathcal{G} \\ &\iff \langle g, C_{YX} f\rangle_{\mathcal{G}} = 0 \quad \forall f \in \mathcal{H} \quad \forall g \in \mathcal{G} \\ &\iff C_{YX} = 0 \end{aligned}$$

Estimation

Although useful, the embedding quantities are not readily available without parametric assumptions on the distributions, and need to be estimated from data. As such we are interested in determining the rate of convergence of the empirical embedding to the true ones. The empirical kernel mean embedding is obtained by taking sample averages of the feature maps. The latter are still functions, which means that the estimator. will still be a function.

Proposition 2.5 ([57])
Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel bounded by κ and $\mathcal{H}$ its associated RKHS. Let μ_X be marginal mean embedding and $\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} \varphi(x_i)$ its empirical estimate. Then for $\delta > 0$ with probability $1-\delta$ we have:

$$\|\mu_X - \hat{\mu}_X\|_{\mathcal{H}} \leq \sqrt{\frac{4\kappa^2 \ln(2/\delta)}{n}}$$

This means that we can estimate the conditional mean embedding at a rate of $n^{-\frac{1}{2}}$. This makes sense as it is nothing but a sample average in an infinite dimensional space. As the number of samples increase to ∞, we expect the sample law to converge to the true law at such rate, and regular enough maps of it should as well. The conditional mean embedding can be estimated using so called sampling operators which associate a function of $\mathcal{H}$ with a random vector of samples and viceversa through the adjoint operator.

Definition 2.12 (Sampling operator)
Let $\mathcal{H}$ be a RKHS space of functions from $\mathcal{X} \to \mathbb{R}$ generated from the kernel associate with the feature map φ. Consider $x \subseteq \mathcal{X}$ a discrete subset of $\mathcal{X}$ with cardinality n. The sampling operator associated with x is defined as:

$$\begin{aligned} S_x :&\mathcal{H} \to \mathbb{R}^n \\ &g \mapsto [g(x_1), \ldots, g(x_n)] \end{aligned}$$

The adjoint of the sampling operator is defined as:

$$\begin{aligned} S_x^* :&\mathbb{R}^n \to \mathcal{H} \\ &v \to \sum_{i=1}^{n} v_i \varphi(x_i) \end{aligned}$$

The sample equivalent of C_{XX} will then be constructed as $\hat{C}_{XX} = \frac{1}{n} S_x^* S_x$ and of C_{YX} as $\hat{C}_{YX} = \frac{1}{n} S_y^* S_x$.

2.2 A primer on Rademacher Complexity

Since L_2 might be too large, a common approach in statistics is to restrict the search to a specific function class $\mathcal{F} \subset L_2$. Intuitively the difficulty in empirical estimation of a function f depends on the complexity of the function class $\mathcal{F}$ in which it lies [62]. There are many different ways to measure the complexity of a function class, but we shall focus on Rademacher complexity. This quantity is useful bounding uniform laws in the non-asymptotic setting, enabling us to bound excess risks involved in empirical minimization procedures and help us obtain finite sample estimation error rates for RKHS hypothesis spaces.

Empirical Risk

To determine parameters of interest, we will have to optimize some sort of loss function. Since we are never able to observe the full distribution but rather a finite number of samples, such minimization must occur on a data dependent loss. The procedure will remain valid only if the error between the *true* parameter and its empirically determined counterpart is always small, no matter what sample we observe. We begin by defining risks, and their empirical counterparts. The notation used is that of [62], and although we refer to it as a parameter $\theta \in \Theta$ can be a (possibly) infinite dimensional object such as a function and function space.

Definition 2.13
Let L_θ be some loss function and $\{X_i\}$ observed i.i.d. samples. For a parameter θ of interest, define:

$$\hat{R}_n(\theta, \theta_0) = \frac{1}{n}\sum^{n} L_\theta(X_i) \qquad \text{(Empirical Risk)}$$

$$R(\theta, \theta_0) = \mathbb{E}_{\theta_0}\left[L_\theta(X)\right] \qquad \text{(Population Risk)}$$

$$E(\hat{\theta}, \theta_0) = R(\hat{\theta}, \theta_0) - \inf_{\theta} R(\theta, \theta_0) \qquad \text{(Excess Risk)}$$

An empirical estimation procedure aims to determine $\hat{\theta}$ that minimizes the empirical risk $\hat{R}_n$. Both this risk and its minimizer are random since they are data dependent, on a different sample one would obtain different results. We are then interested in determining the excess risk committed using $\hat{\theta}$ as a plug-in estimator; of course if $\hat{\theta}$ well approximates the *true* θ_0 then such value will be small. Suppose that the infimum of R can be attained by some parameter θ_{inf}, then the excess risk can be decomposed into three terms:

$$E(\hat{\theta}, \theta_0) = \underbrace{R(\hat{\theta}, \theta_0) - \hat{R}_n(\hat{\theta}, \theta_0)}_{T_1} + \underbrace{\hat{R}_n(\hat{\theta}, \theta_0) - \hat{R}_n(\theta_{inf}, \theta_0)}_{T_2} + \underbrace{\hat{R}_n(\theta_{inf}, \theta_0) - R(\theta_{inf}, \theta_0)}_{T_3} \tag{2.7}$$

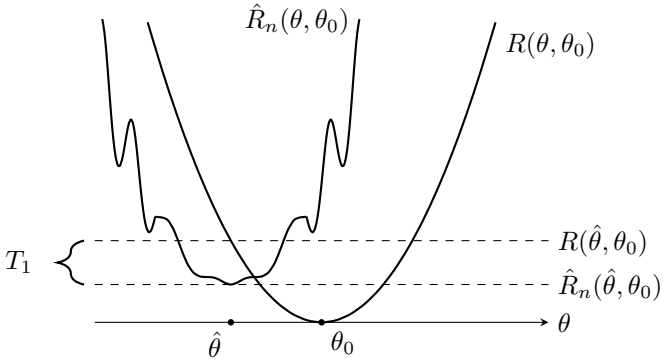


Figure 2.4: Representation of terms in the excess risk. The $\hat{R}_n$ empirical risk is a noisy, data dependent function. Resampling X_i would result in a different $\hat{R}_n$.

Note that T_2 is always non-negative since $\hat{\theta}$ minimizes the empirical risk. The terms we are interested in controlling are T_1, T_3 which are similar:

$$T_1 = \mathbb{E}\left[L_{\hat{\theta}(X)}\right] - \frac{1}{n}\sum L_{\hat{\theta}}(X_i) \tag{2.8}$$

Remember that since $\hat{\theta}$ is a data dependent quantity, T_1 is also random. To bound this, we need a uniform law of large numbers over all possible losses parameterized by θ, i.e. the set $\{x \mapsto L_\theta(x) \quad \theta \in \Theta\}$. This

can be done by using Rademacher complexities. Rather than searching for a uniform bound over the entire class, i.e. $\sup_{\mathcal{F}}$ we consider the complexity by focusing on small neighborhoods of the observed data, this is called localization. The reason why localization is important is that the chosen hypothesis class $\mathcal{F}$ might still be too large and algorithms will concentrate only on subsets with small errors.

Definition 2.14 ([62]Localized Rademacher Complexity)
For a given $\delta > 0$ and a function class $\mathcal{F}$, the local Rademacher complexity of $\mathcal{F}$ is defined as:

$$\mathcal{R}_n(\delta, \mathcal{F}) = \mathbb{E}_{\epsilon, X}\left[\sup_{f \in \mathcal{F}: \|f\|_2 \leq \delta} |\frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i)|\right]$$

where $\{X_i\}$ are i.i.d. samples from the underlying distribution and $\{\epsilon_i\}$ are i.i.d. Rademacher variables, i.e. $\epsilon_i \in \{-1, 1\}$ i.i.d. with probability $\frac{1}{2}$ independent of X_i.
Similarly, the localized empirical Rademacher complexity is defined as the data dependent quantity:

$$\hat{\mathcal{R}}_n(\delta) = \mathbb{E}_{\epsilon}\left[\sup_{f \in \mathcal{F}: \|f\|_2 \leq \delta} |\frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(x_i)|\right]$$

The Rademacher complexity of a function class $\mathcal{F}$ is the average maximum correlation between functions in $\mathcal{F}$ and a random noise vector. Essentially, if the Rademacher complexity is high the function class contains functions that can interpolate any random noise. At first, this might seem to be a positive quality of a function class since any set of points can be interpolated. Clearly, this means that the function class $\mathcal{F}$ will overfit to any observed data. Such class is thus too large. On the other hand, if the complexity of $\mathcal{F}$ is not large enough functions therein will not adapt well to the data. The localized Rademacher complexity further restricts to the function class with *small* L_2 norm. The sweet spot between the Rademacher complexity and class size trade off is given by the critical radius.

Definition 2.15 ([62]Critical Radius)
The critical radius of a uniformly bounded function class $\mathcal{F}_B = \{\|f\|_\infty \leq B\}$ is defined as:

$$\delta_n = \arg\min_{\delta}\left\{\mathcal{R}_n(\delta, \mathcal{F}) \leq \frac{\delta^2}{B}\right\}$$

The empirical critical radius $\hat{\delta}_n$ is the solution to the above definition by replacing the localized Rademacher complexity with its empirical counterpart.

Theorem 2.8 ([62])
Given $\mathcal{F}_B$ let δ_n be the critical radius of $\mathcal{F}_B$. Then with probability $1 - c_1 e^{-c_2 \frac{nt^2}{B^2}}$ we have that $\forall t \geq \delta_n$:

$$|\,\|f\|_n^2 - \|f\|_2^2\,| \leq \frac{1}{2}\|f\|_2^2 + \frac{t^2}{2} \quad \forall f \in \mathcal{F}_B$$

where $\|f\|_n = \frac{1}{n}\sum_{i=1}^{n} f(x_i)^2$. Additionally if $n\delta_n^2 \geq \frac{2}{c_2}\log\left(4\log\left(\frac{1}{\delta_n}\right)\right)$ then:

$$|\,\|f\|_n - \|f\|_2\,| \leq c_0\delta_n$$

Similar result holds with empirical counterparts of the above.

This result enables us to uniformly bound quantities such as equation (2.8) over the function class in terms of critical radii. Another advantage of using Reproducing Kernel Hilbert Spaces as working classes is that the localized Rademacher complexities can be determined by the eigendecay of the integral operator that characterize them.

Proposition 2.6 ([62])
Let $\mathcal{H}_1$ be the unit ball of an RKHS associated with kernel k and let the operator T have eigenvalues $\{\lambda_i^2\}$. Then the localized Rademacher complexity is upper bounded by:

$$\mathcal{R}_n(\delta) \leq \sqrt{\frac{2}{n}}\sqrt{\sum_{i=1}^{\infty}\min\{\lambda_i^2, \delta^2\}}$$

Equivalently for the empirical equivalent, let $\left\{\hat{\lambda}_i^2\right\}$ be the eigenvalues of the normalized kernel matrix $K(x_i, x_j) = \frac{k(x_i, x_j)}{n}$. Then the localized empirical Rademacher complexity is upper bounded by:

$$\hat{\mathcal{R}}_n(\delta) \leq \sqrt{\frac{2}{n}}\sqrt{\sum_{i=1}^{n}\min\{\hat{\lambda}_i^2, \delta^2\}}$$

2.3 Conditional Moment Restriction and Ill-Posedness

The Fredholm integrals of equations (1.2) and (1.3) that define the bridge functions can also be interpreted as constraints on function spaces. Since we are considering integrals with respect to probability measures, these coincide with conditional expectation and as such, they are often termed as conditional moment restrictions. The rest of the thesis primarily focuses on outcome bridge functions as they are the most studied in the literature, but the same approaches can be taken for the exposure bridge function $q(Z, A, X)$. Particular attention to the latter will resurface when discussing doubly robust approaches in section 3.2.

Finding the bridge function is an ill-posed problem

Denote the conditional expectation operator of the law $W|Z, A, X$:

$$\begin{aligned} E : L_2(W, A, X) &\to L_2(Z, A, X) \\ h &\mapsto \mathbb{E}\left[h(W, A, X)|Z, A, X = \cdot\right] \end{aligned}$$

The operator E is linear by linearity of expectation and, under the mild conditions of proposition 2.3, it is also compact. This ensures that the singular system of E is well defined. Moreover the adjoint operator of E is the conditional expectation operator in the conditioning variable defined as $E^* : L_2(Z, A, X) \to L_2(W, A, X)$.

Proposition 2.7 ([64])
The adjoint operator E^* of E is such that:

$$\begin{aligned} E^* : L_2(Z, A, X) &\to L_2(W, A, X) \\ g &\mapsto \mathbb{E}\left[g(Z, A, X)|W, A, X = \cdot\right] \end{aligned}$$

Moreover, if E is a compact linear operator then E^* is also compact.

Proof. Denote $V_h = (W, A, X)$ and $V_q = (Z, A, X)$

$$\begin{aligned} &\mathbb{E}\left[(Eh)(V_h) \cdot (E^*g)(V_h)\right] = \mathbb{E}\left[h(V_h) \cdot (E^*g)(V_h)\right] = \langle h, E^*g\rangle_2 \\ &= \langle Eh, g\rangle_2 = \mathbb{E}\left[(Eh)(V_h) \cdot g(V_q)\right] = \mathbb{E}\left[(Eh)(V_h) \cdot \mathbb{E}\left[g(V_q)|V_h\right]\right] \end{aligned}$$

Compactness of the adjoint is a necessary and sufficient condition for the compactness of E (theorem 2.4). □

Using this notation, the Fredholm integral equations that characterize the bridge functions can be rewritten as the following operator problems:

$$\begin{aligned} &\textit{Determine h,q respectively such that:} \\ &Eh = g_0 \\ &E^*q = f_0 \end{aligned} \tag{2.9}$$

where $g_0 = \mathbb{E}\left[Y|Z, A, X\right], f_0 = \frac{1}{f(a|W,X)}$. Adopting the language of PDEs, we shall refer to f_0, g_0 as the source terms or source conditions of the problem. If both the singular system of E, and the source terms $g_0 = \sum \alpha_i \phi_i$ were known, under additional assumptions[6], one would be able to apply Picard's theorem (theorem 2.5) to obtain a solution. This will be discussed slightly more in depth in the Existence section later. In practice the source conditions g_0, f_0 cannot be observed, but rather noisy observations $\hat{g}_n, \hat{f}_n$ can be estimated from the data. This inconvenience complicates the problem as it now becomes:

$$\begin{aligned} &\textit{Determine h,q respectively such that:} \\ &Eh = \hat{g}_n \\ &E^*q = \hat{f}_n \end{aligned}$$

The problem becomes much more difficult if the operator has to also be estimated from the data. In the mathematics literature, in particular inverse problems, the concept of *ill-posed* problem arises. The definition first appeared in Hadamard's work regarding partial differential equations but the idea is general enough that it is used in various other fields.

[6] $g_0 \in L_2(Z)$ and fast enough decay of the coefficients α_i with respect to the singular values λ_i.

Definition 2.16 (Ill-Posed problem [35])
A problem is well-posed if it satisfies the following semi-formal criteria for the problem data:

- *Existence of a solution.*
- *Uniqueness of the solution.*
- *Continuous dependence of the solution on the data.*

A problem that is not well-posed is termed ill-posed.

The conditions of the above definition are natural requirements. Existence of a solution is necessary for the problem to be well posed, otherwise one would search for a solution in vain. In the proximal setting, the existence of bridge functions is either directly assumed or guaranteed by requiring additional conditions on the behavior of the conditional expectation operator E and source conditions. Some consider the existence requirement to be a weak one since a solution might exists for a slightly relaxed problem through the use of regularization on the source term. Similarly, the uniqueness requirement of definition 2.16, is also weak, since the problem can always be reformulated into the search for a 'minimal' or 'regularized' solution. The issues posed by non-uniquess are mostly related to the empirical estimation from data procedure. Multiple bridge functions are not problematic for identification of the CERF. Nonetheless, many authors directly require uniqueness of a bridge function to avoid working with possible bias introduced by regularization terms.

Example 2.7 (Multiple Solutions)
Let $\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}\right)$. *Suppose that the source condition is* $g_0 = y$. *It can be shown that* $X|Y = y \sim \mathcal{N}\left(\frac{1}{2}y, \frac{3}{4}\right)$. *Then the operator problem of determining* f *such that* $Ef = g_0$ *admits infinite solutions. Any function* $f(x, y) = ax + by$ *is a solution if* $\frac{a}{2} + b = 1$. *The solution becomes unique if we*

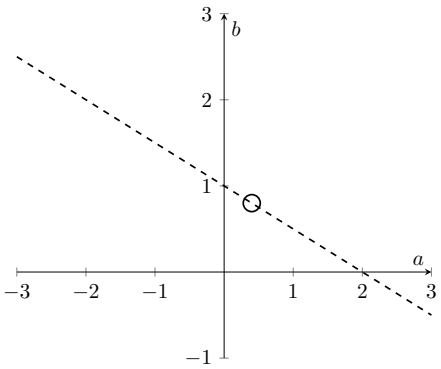


Figure 2.5: Set of valid coefficients of $ax + by$ for a solution to the above operator problem. Coefficients of the *minimum* solution are circled.

restrict ourselves to the class of function $\mathcal{F} = \{f : f(x, y) = ax + by\}$ *and search for the solution with the smallest* $a^2 + b^2$. *Then the regularized solution is given by* $f(x, y) = 0.4x + 0.8y$

We have seen that the previous requirements are weak requirements since they can be remedied by algebraic methods[35]. The most critical requirement for a problem to be well-posed is the third point, continuous dependency on the problem data. This is closely related to the idea of stability in partial differential equations, slight changes in initial conditions should not greatly affect my solution. This depends on the invertibiility of the operator E. In general, the problem of determining solutions to Fredholm integral equations does not satisfy this requirement and thus such problems are often ill-posed [25, 65].

Existence of Bridge functions

As previously mentioned, many works in the proximal inference literature directly assume the existence of a function that solve the conditional moment restriction[36, 18]. Alternatively, additional conditions on the conditional expectation operator or on the laws of the random variables can automatically imply the existence of bridge functions. The most common requirements [13, 53, 36] are the ones necessary for compactness of the conditional expectation operator. In particular [37, 13] make use of Picard's theorem (theorem 2.5), to ensure that the solution to the problem exists. The requirements are the following:

- $\int_{\mathcal{W}} \int_{\mathcal{Z}} dP(w|z,a,x)dP(z|w,a,x) < +\infty$
- $\sum_i \frac{1}{\nu_i^2} \langle g_0, \phi_i \rangle^2 < +\infty$

The first requirement is a sufficient condition for the compactness of the operator E (as given by proposition 2.3) and guarantees the existence of a singular system $\{\nu_i, \varphi_i, \phi_i\}$. The second point ensures that the source function respects the requirements of Picard's theorem. The conditions of theorem 2.5 are now satisfied and we are guaranteed the existence of a solution of the form:

$$\sum_i \frac{1}{\nu_i} \langle g_0, \phi_i \rangle \varphi_i$$

The solution belongs to the space of interest $L_2(W, A, X)$ since condition three guarantees that its norm is finite. Similar requirements can be made for the exposure bridge function by exchanging g_0 with f_0.

Non-Uniqueness and Weak Identification

All that is required from the proximal g-formula to estimate the causal parameters is the completeness assumption and a function that satisfies the conditional moment restriction. Albeit the discussion on existence of a solution to the Fredholm integral problem through Picard's theorem, there has not been a requirement for the uniqueness of bridge functions. In truth, the Fredholm integral equations might admit multiple solutions and then the conditional moment restrictions would identify a subset $\mathcal{H}_0$ of L_2 or some function class of interest $\mathcal{H}$. The number of points(functions) in such subset, which we will refer to as set of bridge functions, might not be one.

Definition 2.17 (Weak Identification)
A parameter of interest θ is said to be weakly identified if it is set identified.

Nonetheless, non-uniqueness of the solutions does not pose a problem to the identification of the causal parameters of interest. This is because the proximal g-formula weakly identifies the CERF. This is shown in the following result:

Proposition 2.8 ([6]Weak Identification of the CERF)
Suppose that the conditions necessary for the proximal g-formula (theorem 1.2) hold. The CERF and ATE are weakly identified for any function in the set of outcomes $\mathcal{H}_0$. A similar result holds for the outcome bridge function q, requiring analogous assumptions.

Proof. We will only consider the problem for the outcome bridge function, the result for the exposure bridge function is exactly the same. Without loss of generality suppose that there exist two different functions $h_1, h_2 \in \mathcal{H}_0$ such that $h_1 \neq h_2$ satisfy the conditional moment restriction equation (2.9). Then $\mathbb{E}[h_1(W, A, X)|Z, A, X] = \mathbb{E}[Y|Z, A, X] = \mathbb{E}[h_2(W, A, X)|Z, A, X]$ and by the proximal g-formula theorem 1.2 we have that $\chi(a) = \mathbb{E}[h_1(W, a, X)] = \mathbb{E}[h_2(W, a, X)]$ □

Proposition 2.8 shows that both CERF and ATE are weakly identified for any solution to equation (2.9). Any point in $\mathcal{H}_0, \mathcal{Q}_0$ recovers the causal quantity of interest, avoiding any theoretical problem in causal identification generated by non-uniqueness of the bridge functions. However, problems arise during estimation procedures. Most estimation procedures aim to minimize some risk functional involving the conditional moment restrictions. When minimizing risks, convexity ensures that the procedure to reach the minima behaves well. In the case of distinct solutions to the conditional moment restrictions, the risk will not be convex and one might encounter difficulties in converging to the true solution such as oscillating between valid solutions. This issue is resolved in one of two ways: requiring uniqueness as an assumption by other uniqueness requirements [53, 13], or including additional restrictions on the function sets. Proposition 2.8 clearly highlights that bridge functions are equivalent $Z|W, A, X$ or $W|Z, A, X$ almost surely. This suggests that imposing completeness of $Z|W, A, X$ or $W|Z, A, X$ ensures that the set of bridge functions only consists of a single point h_0 or q_0. Some might argue that the first set of assumptions is highly technical, as it is additional completeness requirements that are not testable. The latter solution often involves the search for a 'minimal' or regularized solution through the use of Tikhonov regularization. In the statistical learning literature, this approach prevents over fitting and aligns with Occam's razor principle, which suggests choosing the simpler solution when having to choose between two valid solutions. This principle also applies to bridge function estimation, even though bridge functions have no direct interpretation and serve only as a means to an end and the concept of 'simplicity' often becomes that of 'smallest norm'.

Continuous dependence on the data

The main difficulty in solving an ill-posed problem is that the conditional expectation operator compresses the information of the higher order terms, yet we are expected to have higher precision of the source in those directions. The following example shows that for *imprecise* source conditions, the solutions can vary greatly.

Example 2.8
Consider a compact operator K with singular system $\{\lambda_i, \varphi_i, \phi_i\}$ and suppose we are interested in determining the solution to $Kh = f_0$. Applying theorem 2.5 we obtain a solution h_1 such that $h_{1_i} = \frac{\langle f_0, \varphi_i \rangle}{\lambda_i}$. Supposed now that the source is infinitesimally perturbed by δ in the i^{th} direction, i.e. $f_1 = f_0 + \delta\varphi_i$. The problem to be solved is now $Kh = f_1$. Always invoking theorem 2.5, the solution h_2 has coefficients $h_{2_i} = \frac{\langle f_1, \varphi_i \rangle}{\lambda_i}$. In other words, h_2 is identical to h_1 except in the i^{th} direction.
Since the source condition was only slightly perturbed one would expect h_2 to remain close to h_1, but:

$$\|h_1 - h_2\|_2^2 = \sum_{i=1} (h_{1_i} - h_{2_i})^2 = \frac{\delta^2}{\lambda_i^2}$$

Compact operators have decreasing eigenvalues in i and thus for an arbitrarily small perturbation δ, the distance from the original solution is blown up by λ_i^2.

It is clear from example 2.8, that the problem is caused by the trailing eigenvalues of the operator, smaller λ_i lead to a larger mistake of the solution. For this reason, the *degree* or level of ill-posedness can be expressed with respect to the rates at which the eigenvalues of the conditional expectation operator E decay.

Definition 2.18 (Degree of ill-posedness)
Let $K : \mathcal{H}_1 \to \mathcal{H}_2$ be a compact linear operator, and $\{\nu_i, \varphi_i, \phi_i\}$ its singular system. The problem of determining h such that $Kh = f_0$ is said to be mildly ill-posed if the decay of the singular values of K is polynomial, i.e. $\nu_i \asymp i^{-p}$ for some $p > 0$. If the decay is faster, the problem is said to be severely ill-posed.

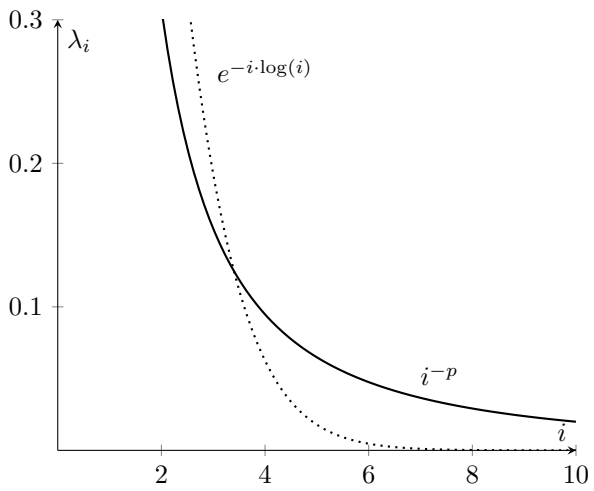


Figure 2.6: Mildly(dark) and severely(dashed) ill-posed eigen decay.

Figure 2.1 shows the *compression rate* of the operator E. For the mildly ill-posed problem, higher values of p means higher compression of information in those directions and thus difficulty in recovering stable solutions.

Example 2.9 (RKHS)
Let $K : \mathcal{H}_1 \to \mathcal{H}_2$ be a compact linear operator with singular system $\{\nu_i, \varphi_i, \phi_i\}$. Suppose that the problem is mildly ill-posed $\nu_i \asymp i^{-p} \quad p > 0$. Then, by theorem 2.5 the solution of the problem $Kh = f_0$ for $f_0 \in \mathcal{H}_2$ is given by:

$$h = \sum_i \frac{1}{\nu_i} \varphi_i \langle f_0, \phi_i \rangle$$

Moreover if $\mathcal{H}_2$ is a RKHS characterized by the same eigen decomposition of K, in other words the associated kernel T is diagonalized by $\{\phi_i\}$ and the eigenvalues match. $f_0 \in \mathcal{H}_2$ can now be written as $f_0 = \sum_i \alpha_i \phi_i$. Then, the norm of the found solution is $\|h\|^2_{\mathcal{H}_2} = \sum_i \frac{\alpha_i^2}{\nu_i^2}$. This function will belong to $\mathcal{H}_2 \iff \sum_i \frac{\alpha_i^2}{\nu_i^2} < \infty$. A possible condition is to require $\alpha_i^2 i^{2p}$ to go to zero faster than i^{-1}, such that the sum converges. In this case, the coefficients α_i must decay faster than $i^{-\frac{1}{2}-p}$. If the coefficients of f_0 are also polynomial, say $\alpha_i \asymp i^{-\eta}$, then we must require $\eta \geq \frac{1}{2} + p$.

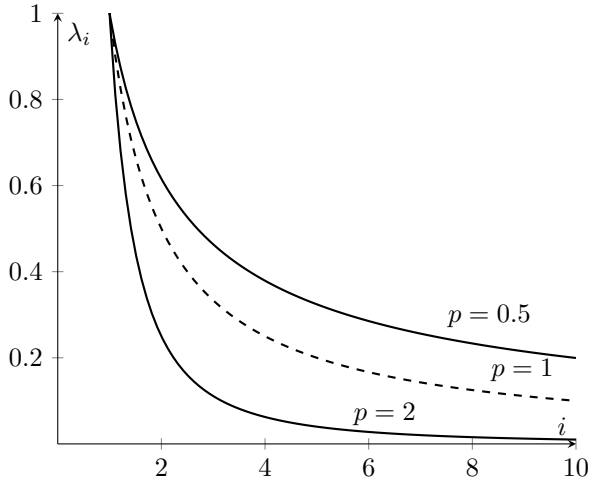


Figure 2.7: Necessary decay for polynomial convergence.

A solution often adopted in the literature to overcome issues such as those that arise from equation (2.15) is to define a measure of the ill-posedness of the operator E and only focus on specific function classes $\mathcal{H} \subset L_2(W)$ where such measure is contained.

Definition 2.19 ([10, 18]Ill-posedness measure)
Let E be the conditional expectation operator and let $\mathcal{H}$ be an hypothesis class. Then the ill-posedness measure of E with respect to $\mathcal{H}$ is given by:

$$\tau(\mathcal{H}) = \sup_{h \in \mathcal{H}} \frac{\|h - h_0\|_2}{\|E(h - h_0)\|_2}$$

The local measure of Ill-Posedness is defined as:

$$\tau_{\mathcal{H}}(\delta) = \sup_{\mathcal{H}} \left\{ \|h - h_0\|_2 : \mathbb{E}\left[(E(h - h_0))^2)\right] \leq \delta^2 \right\}$$

Note that since projections always decrease L^2 distances, τ always takes values between 1 and $+\infty$, and will take on $+\infty$ due to the ill posedness of the problem. This measure relates the distance pre and post application of the operator E. One would hope that functions that are far, and thus *distinguishable* in $L_2(W, A, X)$ are also far or *distinguishable* in the arrival space $Im(E)$. In such case, the value of τ will be small. The more the operator E shrinks or compresses the information from $L_2(W)$, the higher the value τ will take on. Since the operator E is defined using the kernel in proposition 2.3, the ill-posedness measure can also be understood as the informativeness of the conditioning variables. Highly informative variables will have a low ill-posedness measure since the smoothing effect will be lower, whereas poorly informative variables will lead to high values of τ. The reason why we are interested in bounding the ill-posedness measure of an hypothesis class $\mathcal{H}$ is that the L_2 error between any function $h \in \mathcal{H}$ and the true bridge function h_0 would then be bounded as:

$$\|h - h_0\|_2 \leq \tau \|E(h - h_0)\|_2 \tag{2.10}$$

Notice that if a completeness assumption such as $W|Z, A, X$ is made, then the measure of ill posedness is automatically finite (definition 2.19). In general, other restrictions on the distribution of the proxies can translate bounds on mean square violation to bounds on the stronger norm $\|h - h_0\|_2$[15].

Example 2.10
Consider the conditional expectation operator $E : L_2(\mathcal{X}) \to L_2(\mathcal{Y})$ and suppose that the $X = Y$. Then,

the eigenvalues of E are 1, the smoothing effect is null and the ill-posedness measure is[7]*:*

$$\tau = \sup_{h \in L_2} \frac{\|h - h_0\|_2}{\|E(h - h_0)\|_2} = \sup_{h \in L_2} \frac{\|h - h_0\|_2}{\|h - h_0\|_2} = 1$$

Example 2.11
Let $\tau_\delta(\mathcal{H}_J)$ be the local measure of ill posedness of definition 2.19 and suppose that operator E has spectral decomposition $E = \sum \nu_i \phi_i \otimes \varphi_i$. Consider the J dimensional sieve $\mathcal{H}_J = span\{\varphi_{i_1} \dots \varphi_{i_J}\}$. Then for the hypothesis class $\mathcal{H}_J$ has

$$\frac{\tau^2_{\mathcal{H}_J}(\delta)}{\delta^2} = \max\left\{\frac{1}{\nu^2_{i_1}}, \dots, \frac{1}{\nu^2_{i_J}}\right\}$$

Any function in $\mathcal{H}_J$ can be expressed as $h = \sum_{k=1}^J h_k \varphi_{ik}$. Then the applying E to any function in $\mathcal{H}_J$ we obtain $Eh = \sum_{k=1}^J \nu_{i_k} h_k \phi_{ik}$. The norm is then $\|Eh\|_2^2 = \nu^2_{i_k} h_k^2$. The supremum is obtained when $h_k^2 = \frac{\delta^2}{\nu^2_{i_k}}$ is largest, in other words for the smallest singular value ν_{i_k}. Setting all other h_k terms to 0. We have that for this choice $\|h\|_2^2 = \frac{\delta^2}{\nu^2_{i_k}}$ and the required ratio is $\frac{1}{\min_k \nu^2_{i_k}} = \max_k \frac{1}{\nu^2_{i_k}}$. Thus, as the size of the sieve grows, so does the measure of ill-posedness.

In [15], the authors relate the critical radius of the function class taken into consideration with the errors committed using the sample analogue of $\|E(h - h_0)\|_2$. The following result enables us to bound the ill-posedness measure for RKHS hypothesis classes in terms of the critical radius:

Lemma 2.3 ([15] Ill-posedness measure bound for RKHS)
Let $\mathcal{H}_B = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq B\}$ be a bounded subspace of a reproducing kernel Hilbert space $\mathcal{H}$ associated to a kernel operator T with eigendecomposition $\{\lambda_i^2, \varphi_i\}$. Let E be the conditional expectation operator previously defined and $\{\nu_i, \varphi_i, \phi_i\}$ be its singular system. Let

$$V^m_{ij} = \mathbb{E}\left[E\varphi_i \cdot E\varphi_j\right] \quad i, j \in \{1, \dots, m\} \tag{2.11}$$

Suppose that the minimum eigenvalue of V^m is bounded away from zero, i.e. $\min\{EIG(V^m)\} = \mu_m \geq \tau_m$ and $\exists c > 0$ such that $\forall i \leq m < k$:

$$|\mathbb{E}\left[E\varphi_i \cdot E\varphi_k\right]| \leq c\tau_m \tag{2.12}$$

Then it holds that:

$$\tau^2_{\mathcal{H}_B}(\delta) \leq \min_{m \in \mathbb{N}_+} \left\{\frac{4\delta^2}{\tau_m^2} + (4c^2 + 1)B\mu^2_{m+1}\right\} \tag{2.13}$$

When working in RKHS $\mathcal{H}$, one has to hope that the kernel operator T and the conditional expectation operator E have *similar* eigenfunctions. Equation (2.11) quantifies the level of smoothing that the conditional expectation has over the eigenfunctions of T. If E were such as in example 2.10, i.e. its eigenvalues are one, then the eigenfucntions would all remain the same and V^m would coincide with the identity matrix. Requiring that the smallest eigenvalue of V_m is bounded away from 0 ensures that the eigenbasis remains independent, in other words no information is lost. Similarly, equation (2.12) ensures that the smoothing of the conditional expectation maintains orthogonality of the first m eigenfunctions of T. In chapter 5 we will assume that the eigenbasis of E also diagonalizes the operator T associated to $\mathcal{H}$, thus satisfying such requirements.

Example 2.12 ([15])
Suppose that the hypothesis class $\mathcal{H}$ is a Reproducing Kernel Hilbert Space with polynomial eigen-decay, i.e. T_k has eigenvalues $\lambda_i^2 \asymp i^{-p}$ for some $p > 0$ and the problem is mildly ill-posed(definition 2.18), i.e. the operator E has eignevalues with polynomial decay $\nu_i^2 \asymp i^{-q}$ for some $q > 0$. Equation (2.13) can be rewritten, up to multiplicative constants, as $4\delta^2 i^p + (4k^2 + 1)Bi^{-q}$. The optimal i will satisfy first order conditions, i.e.:

$$\begin{aligned} 4\delta^2 p i^{p-1} - (4k^2 + 1)Bqi^{-q-1} &= 0 \\ \frac{4\delta^2}{(4k^2 + 1)B} = \frac{qi^{-q-1}}{pi^{p-1}} &\asymp \lambda_i^2 \nu_i^{-2} \end{aligned} \tag{2.14}$$

This shows that, at the optimal value i, the eigenvalues approximately solve equation (2.14). Thus $i \asymp \delta^{\frac{2}{p+q}}$.

[7] The identity matrix is not a compact operator.

Regularization

Determining a solution to the operator problem with $K : L_2(X) \to L_2(Y)$ might be ill-posed but what about the well-posedness relative to new spaces $\mathcal{H} \subset L_2(X), \mathcal{G} \subset L_2(Y)$?[9]. If the either the source or the true solution is assumed to be in a regular enough subspace, the problem might be well posed. In particular, all that is needed is for the true bridge function h_0 to belong to some power space of the range of the operator (K^*K). Using the notation of power spaces and definition 2.8 $Range(K^*K^\beta) \quad \beta > 0$. It can also be possible to dampen the trailing eigenvalues of the conditional expectation operator [9]. This can be done in various ways, but we shall focus primarily on Tikhonov regularization. The solution we are looking for must satisfy $K^*Kh = E^*f_0$. The Tikhonov regularized solution searches for $(K^*K + R\mathit{I})h = E^*f_0$. Applying theorem 2.5 and proposition A.6, the solution is now given by:

$$h_R = \sum_{i=1} \frac{\nu_i}{\nu_i^2 + R} \langle f_0, \phi_i \rangle \varphi_i$$

Notice that this type of regularization introduces bias on each estimated term.

Proposition 2.9
The bias on each term of the solution of a Tikhonov regularized problem is given by:

$$b_i(R) = \frac{R}{\nu_i^2 + R}$$

Proof. Assume that h_0 is the *true* solution to the operator problem given by theorem 2.5, i.e. $h_0 = \sum_{i=1} \nu_i \langle f_0, \phi_i \rangle \varphi_i$ and let h_R be the solution to the regularized problem. Then:

$$\|h_0 - h_R\|_2^2 = \sum_{i=1} \langle f_0, \phi_i \rangle^2 \left(\frac{1}{\nu_i} - \frac{\nu_i}{\nu_i^2 + R} \right)^2 = \sum_{i=1} \frac{\langle f_0, \phi_i \rangle^2}{\nu_i^2} b_i(R)^2 = \sum_{i=1} \langle h_0, \varphi_i \rangle^2 b_i(R)^2$$

Thus $b_i(R) = \nu_i \left(\frac{1}{\nu_i} - \frac{\nu_i}{\nu_i^2 + R} \right) = \frac{R}{\nu_i^2 + R}$ □

This proposition shows that for the directions associated with larger singular values the regularization have a smaller bias, which is expected since more *information* is known about them. Returning to example 2.8, the direction would now only be blown up by $\frac{\nu_i}{\nu_i^2 + R}$, which is not arbitrarily large for a fixed value of R. This in turn enables the errors to be controlled by the regularization (see corollary A.2). In practice, the regularization coefficient will be sent to zero as more and more data and information comes in so that the population risk can be sent to zero. Although a more regular function, i.e. $h_0 \in \mathcal{H}^\beta$ for large β might make the problem easier to solve since less information will be place on the trailing coefficient, using Tikhonov regularization the benefits do not increase indefinitely with the growing of β.

Example 2.13 ([9]Saturation of Tikhonov regularization)
Suppose that $h_0 \in \mathcal{H}^\beta$. Let us consider the slack between the true solution and the regularized one:

$$\|h_0 - h_R\|_2^2 = \sum_{i=1} b_i^2(R) \langle h_0, \phi_i \rangle^2 = \sum_{i=1} b_i^2(R) \nu_i^{2\beta} \frac{\langle f_0, \phi_i \rangle^2}{\nu_i^{2\beta}} \leq \sup_i \left\{ b_i^2(R) \nu_i^{2\beta} \right\} \cdot \|h_0\|_\beta^2 = C_R \|h_0\|_\beta^2$$

Since we assume $h_0 \in \mathcal{H}^\beta$, it must have finite $\|\cdot\|_\beta$ and thus convergence of the regularized solution will depend solely on the rate of C_R to zero in R. Notice that:

$$\frac{\partial}{\partial \nu_i} b_i^2(R) \nu_i^{2\beta} = \frac{2R^2 \nu_i^{2\beta - 1} \left((\beta - 2) \nu_i^2 + \beta R \right)}{\left(\nu_i^2 + R \right)^3}$$

For $\beta < 2$ the sup is reached for $\nu_i^2 = \frac{R\beta}{2-\beta}$ and thus we have that $C_R \asymp R^\beta$. Whereas if $\beta \geq 2$ $C_R = R \sup_i \nu_i^{2(\beta-2)} = k \cdot R^2$ for some finite $k > 0$. This suggests that when estimating the function h_0, the maximum regularity we can take advantage of is 2. For regularities below 2, the rate R^β kicks in and R^2 otherwise.

Approximate bridge functions enable approximate estimation

Since empirical estimation procedures never achieve the *truth* with a finite number of samples but rather hope to approach the truth in some limit, one might wonder if 'approximate' bridge functions can still

enable us to recover 'approximate' CERF estimates. First of all, one needs to define the concept of approximate bridge functions; since they are characterized by conditional moment restrictions, an approximate bridge function could either be a L_2 function that well approximates the *true* bridge function h_0 or one that almost satisfies the conditional moment restriction. Let us consider first a result by minimizing the violation of the worst case moment restriction

Lemma 2.4
Let $h_0 \in \mathcal{H}_0$ be an outcome bridge function, thus satisfying the conditional moment restriction equation (1.2)*. Moreover let $\hat{h}$ be a function that has a small worst case violation of the conditional moment restriction equation* (2.9)[8]*:*

$$\left\| E\hat{h} - g_0 \right\|_\infty^2 = \sup_{Z,A,X} \mathbb{E}\left[Y - \hat{h}(W,A,X) | Z,A,X \right]^2 \leq \epsilon^2$$

Then using $\hat{h}$ in the proximal g-formula theorem 1.2 instead of the true bridge function h_0, the following will hold:

$$|\chi(a) - \hat{\chi}(a)|^2 \leq \epsilon^2$$

Proof.

$$\begin{aligned}
&|\chi(a) - \hat{\chi}(a)|^2 = |\mathbb{E}\left[h_0(W,a,X)\right] - \mathbb{E}\left[\hat{h}(W,a,X)\right]|^2 = |\mathbb{E}\left[h_0(W,a,X) - \hat{h}(W,a,X)\right]|^2 \\
&= |\mathbb{E}\left[\mathbb{E}\left[h_0(W,a,X) - \hat{h}(W,a,X)|Z,a,X\right]\right]|^2 \leq \mathbb{E}\left[\mathbb{E}\left[h_0(W,a,X) - \hat{h}(W,a,X)|Z,a,X\right]^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[h_0(W,a,X) - Y + Y - \hat{h}(W,a,X)|Z,a,X\right]^2\right] = \mathbb{E}\left[\mathbb{E}\left[Y - \hat{h}(W,a,X)|Z,a,X\right]^2\right] \\
&\leq \sup_{Z,A,X} \mathbb{E}\left[\mathbb{E}\left[Y - \hat{h}(W,A,X)|Z,A,X\right]^2\right] \leq \mathbb{E}\left[\sup_{Z,A,X} \mathbb{E}\left[Y - \hat{h}(W,A,X)|Z,A,X\right]^2\right] \leq \epsilon^2
\end{aligned}$$

□

The idea of controlling the worst case scenario is known as maximum moment restriction, and will later be exploited by various methods to transform conditional moments into regular moment equations to estimate the bridge function. Does using a function $L_2(W,A,X)$ close to the true bridge function enable us to recover the CERF? Under an additional assumption, we will show such result in theorem 5.2. Alternatively, one might wonder if functions that only perform small *average* violations of conditional moment restriction still enable approximate recovery of the CERF.

$$\|Eh - g_0\|_2^2 = \mathbb{E}\left[\mathbb{E}\left[h(W,A,X) - Y|Z,A,X\right]^2\right] \tag{2.15}$$

Clearly true bridge functions have null average deviations from the CMR, and thus this quantity equal to zero. Unfortunately this is not immediatly the case since:

$$\begin{aligned}
\|h - h_0\|_2^2 &= \mathbb{E}\left[(h(W,A,X) - h_0(W,A,X))^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(h(W,A,X) - h_0(W,A,X))^2 | Z,A,X\right]\right] \geq \mathbb{E}\left[\mathbb{E}\left[(h(W,A,X) - h_0(W,A,X)|Z,A,X\right]^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[h(W,A,X) - Y|Z,A,X\right]^2\right] = \|Eh - g_0\|_2^2
\end{aligned}$$

Nonetheless, this quantity is observable and could be used as a minimization criteria to identify the bridge function. With additional assumptions controlling the ill-posedness measure(definition 2.19), we can relate the weaker metric to the stronger.

[8] $\sup(f^2) = (\sup f)^2$

3

Adjusting for the unobserved

In the literature, methods are often developed to identify the ATE in the case of binary treatment, whilst CERF estimation methods are of secondary interest. The aim of this section is to introduce existing methods and approaches used in the literature under the proximal exchangeability assumptions to estimate the CERF. Moreover, the focal point of proximal inference is the proximal g-formula which suggests that identifying bridge functions in turn identifies the causal quantities of interest. Thus, most methods redirect focus on the problem to identifying the bridge functions instead of the causal quantities themselves.

Notation

Whenever the problem need not involve the individual variables or notation becomes cluttered, we will interchangeably adopt the following notation:

$$V_h = (W, A, X) \qquad V_q = (Z, A, X)$$

This will also translate to the samples, i.e. $V_{h_i} = (w_i, a_i, x_i)$ and $V_{q_i} = (z_i, a_i, x_i)$. This is similar to what is done in other works[26, 18].

3.1 Foregoing the bridge function

The first idea one has to quantify the effect of a variable on another is to perform a simple linear regression. As already discussed in section 1.3, including all the variables in the regression problem as measured confounders will render the obtained results biased. As such, different regression approaches that leverage the assumed structural independencies have been developed. The following two stage linear regression approaches forgo the estimation of the bridge function altogether but rather leverage the independence structure given by assumption 1.4 to identify the causal exposure. Moreover since the proximal g-formula (theorem 1.2) is not invoked, no completeness assumption is necessary. Note that choosing to work with a parametric model might avoid the ill-posedness of the problem by a form of implicit regularization. By enforcing some parametric constraint, we are automatically imposing regularity conditions about the solution and its trailing coefficients. We will study the classic proximal two stage least squares approach, which relies on linearity assumptions, and a Bayesian extension that only requires a weaker assumption to capture potentially non linear response functions.

3.1.1 Proximal Two Stage Least Squares

The method of Proximal Two Stage Least Squares (P2SLS) was first introduced in [46] and later [53]. It is eerily similar to the methods often used in econometrics, in particular the parametric instrumental variable literature. The regression problem is divided in two steps, the first aims to determine the coefficients of the dependencies between the proxy variables and treatment. This first regression will inherently be biased due to the presence of the unmeasured confounder. Afterwards, using the fitted values in the second stage regression, we are able to account for the bias from the unmeasured confounder. For sake of presentation the measured confounder X is omitted but results remain valid even in its presence. The Proximal Two Stage Least Squares makes stringent assumptions in the structural form of the data, requiring linearity in their underlying relationships:

Assumption 3.1 (Linearity)
The data satisfies the following linear relationships:

$$\begin{aligned}\mathbb{E}[Y|W,A,U] &= \beta_{0Y} + \beta_{AY}A + \beta_{WY}W + \beta_{UY}U \\ \mathbb{E}[W|A,Z,U] &= \beta_{0W} + \beta_{UW}U \\ \mathbb{E}[U|Z,A] &\text{ is linear in } Z, A. \\ \beta_{UW} &\neq 0\end{aligned}$$

Rather than requiring a completeness assumption, which relates the variability of the unmeasured confounding and proxy, the last point of assumption 3.1 ensures that the former has an effect on the latter. Under these linear models, the ATE per unit of A is constant across all possible treatments and coincides with the slope of $\frac{\partial}{\partial a}\chi(a) = \mathbb{E}\left[Y^{a+1} - Y^a\right] = \beta_{AY} \quad \forall a$ as shown in the following result.

Proposition 3.1
Under assumptions 1.4 and 3.1 then:

$$\chi(a) = \tilde{\beta}_0 + \beta_{AY}a$$

where $\tilde{\beta}_0 = \beta_{0Y} + \mathbb{E}[\beta_{WY}W + \beta_{UY}U]$

Proof. By applying the standard g-formula:

$$\begin{aligned}\chi(a) = \mathbb{E}[Y^a] &= \mathbb{E}[\mathbb{E}[Y|W, A=a, U]] \\ &= \mathbb{E}[\beta_{0Y} + \beta_{AY}a + \beta_{WY}W + \beta_{UY}U] \\ &= \beta_{0Y} + \beta_{AY}a + \mathbb{E}[\beta_{WY}W + \beta_{UY}U]\end{aligned}$$

□

Determining the CERF will require identifying an additional constant $\tilde{\beta}_0$ derived from integrating the above mentioned quantity. This constant is the intercept of the CERF and represents the baseline $\mathbb{E}\left[Y^0\right]$ effect under no treatment.
The linearity assumption makes the model particularly simple, since it is assumed that as the treatment A increases or decreases, the effect on outcome stays constant. Nonetheless it can be an effective estimation method in situations where the data is concentrated around a linear response and outliers are trimmed out. On the other hand, assuming linearity enables us to easily scale the approach to problems with a large number of samples due to an asymptotic computational complexity of $O\left(C^2N\right)$ which is quadratic in the number of features C, and only linear in the number of samples N. This scalability capacity will not always be the norm for all other methods. Why do these simple regressions return unbiased coefficients even in the presence of unmeasured confounding?

Fitted Values

As previously introduced, the use of fitted first stage regression into the second enables us to obtain unbiased regression coefficients. Why would this be? By taking conditional expectation on both sides of the assumed structure and law of iterated expectation, it follows that

$$\begin{aligned}\mathbb{E}[Y|Z,A] &= \beta_{0Y} + \beta_{AY}A + \beta_{WY}\mathbb{E}[W|ZA] + \beta_{UY}\mathbb{E}[U|ZA] \\ \mathbb{E}[W|ZA] &= \beta_{0W} + \beta_{UW}\mathbb{E}[U|ZA]\end{aligned}$$

and thus, substituting the second equation into the first:

$$\mathbb{E}[Y|ZA] = \beta_{0Y} + \beta_{AY}A + \beta_{WY}\mathbb{E}[W|ZA] + \frac{\beta_{UY}}{\beta_{UW}}\left(\mathbb{E}[W|ZA] - \beta_{0W}\right) \tag{3.1}$$

Equation (3.1) and the assumed linear relation between U and Z, A suggests that regressing Y on A and $\hat{W}$, an unbiased estimator of $\mathbb{E}[W|Z,A]$, will lead to an unbiased estimator of β_{AY}. Notice that we are only really interested in the coefficient β_{AY}. The immediate choice of $\hat{W}$ is the fitted values found after the ordinary least square regression $W|Z,A$ is carried out. The recovered slope coefficient $\hat{\beta}_{AY}$ is then an unbiased estimator of the ATE per unit and thus the slope of the CERF $\frac{\partial}{\partial a}\chi(a)$.
The additional quantity of $\tilde{\beta}_0$ is necessary to account for $\beta_{0Y} + \beta_{WY}\mathbb{E}[W] + \beta_{UY}\mathbb{E}[U]$ and obtain complete identification of the CERF $\chi(a)$. Directly substituting the estimated intercept coefficient from equation (3.1) will result in large biases due to the presence of extra terms $\frac{\beta_{UY}}{\beta_{UW}}\beta_{0W}$. Notice that by iterated expectation we have that:

$$\beta_{0Y} + \beta_{WY}\mathbb{E}[W] + \beta_{UY}\mathbb{E}[U] = \mathbb{E}[Y] - \beta_{AY}\mathbb{E}[A]$$

Thus a valid unbiased parametric estimator for the CERF under assumption 3.1 is given by the following.

Definition 3.1 (P2SLS Estimator)
Let $\hat{\beta}_{AY}$ be the coefficient estimated with the P2SLS procedure then the P2SLS estimator for the CERF is given by:

$$\hat{\chi}(a) = \mathbb{E}[Y] + \hat{\beta}_{AY}(a - \mathbb{E}[A]) \tag{3.2}$$

Correction method

The previous result can also be obtained by another route which was first proposed in [46]. Here, rather than using the fitted coefficients in the second stage regression, the correction happens after the entire estimation procedure. In this case both regressions are performed on the set of observed variables $Z, A,(\mathrm{X})$. The *biased* coefficients are here denoted by θ, γ rather than the true β:

$$Y \sim \theta_{0Y} + \theta_{AY}A + \theta_{ZY}Z$$
$$W \sim \gamma_{0W} + \gamma_{AW}A + \gamma_{ZW}Z$$

Clearly the coefficients found by performing such regression do not coincide with the true structural ones because the above regressions do not respect assumption 3.1, translating to $\theta_{AY} \neq \beta_{AY}$. Nonetheless, under the above assumptions the following must hold:

$$\theta_{0Y} + \theta_{AY}A + \theta_{ZY}Z = \beta_{0Y} + \beta_{AY}A + \left(\frac{\beta_{UY}}{\beta_{UW}} + \beta_{WY}\right) \cdot (\gamma_{0W} + \gamma_{AW}A + \gamma_{ZW}Z) - \frac{\beta_{UY}}{\beta_{UW}}\beta_{0W}$$

From the above display we are able identify the amount of bias for each term in the hope that it can be adjusted for:

$$\theta_{ZY} = \left(\frac{\beta_{UY}}{\beta_{UW}} + \beta_{WY}\right)\gamma_{ZW} \tag{3.3}$$

$$\theta_{AY} = \beta_{AY} + \left(\frac{\beta_{UY}}{\beta_{UW}} + \beta_{WY}\right) \cdot \gamma_{AW} \tag{3.4}$$

Although the ATE or slope coefficient in equation (3.4) is biased, the ratio between the estimated coefficients found in equation (3.3) enables us to estimate the ratio between the unobserved effect coefficients of U and correct for the bias. Thus, a valid unbiased estimator for $\frac{\partial}{\partial a}\chi(a)$ is given by $\hat{\beta}_{AY} = \theta_{AY} - \frac{\theta_{ZY}}{\gamma_{ZW}}\gamma_{AW}$. The intercept estimation remains the same as before. A similar reasoning can be applied to include measured confounders X.

3.1.2 Bayesian Proximal 2SLS

An extension of the regular P2SLS method was recently proposed in [23]. Rather than requiring a global linearity restriction, only a form of *local linearity* is required. This enables us to work with possibly non linear CERFs. The authors of the paper adopt a specific proximal structure, whose DAG is given in figure 3.1 where the NCO and NCE do not affect the treatment and outcome respectively. As such, the method is more similar to an instrumental variable method with two instruments. Nonetheless, the two causal paths could be added without any problems like in the previous section. For presentation purposes the measured confounder is omitted but the results can easily be adjusted.
The main assumption is the following:

Assumption 3.2
The data is generated from a mixture of Gaussians. The effect of A and U on the outcome is linear within each mixture component, in particular:

$$Y|A, Z, U, S = k \sim \mathcal{N}\left(\beta_{0k} + \beta_{Ak}A + \beta_{Uk}U, \sigma^2_{yk}\right)$$
$$W|A, Z, U \sim \mathcal{N}\left(\beta_{0W} + \beta_{UW}U, \sigma^2_{w}\right)$$

where the probability for each cluster are given by:

$$P(S = k) = \omega_k(A) = \Phi(\alpha_k(A)) \prod_{r<k} (1 - \Phi(\alpha_r(A))) \tag{3.5}$$

$$\alpha_k(A) \sim N(\mu_{\alpha,k}(A), 1) \tag{3.6}$$

Moreover assume that the effect of U on W is non-zero, i.e. $\beta_{UW} \neq 0$.

Assumption 3.2 can be interpreted as requiring local linearity, on each cluster S, rather than global linearity. Taking an infinite number of clusters, we can hope to approximate a smooth function well enough. Additionally, due to the presence of mixtures, the assumption 1.4 must hold for each cluster k and thus $Y|A,Z,U,S=k \sim Y|A,U,S=k$ and $W|A,Z,U \sim W|U$. Moreover, due to equation (3.5) and the proof, there is an additional *implicit*[1] assumption, the conditional independence: $S \perp\!\!\!\perp (Z,W,U)\,|A$. Equation (3.5) highlights that the weights for each cluster are determined through a stick breaking process.
The graphical representation of the model satisfying assumption 3.2 is the following:

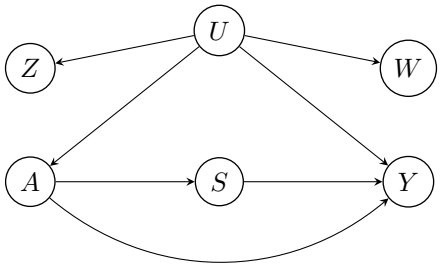


Figure 3.1: Graphical model with additional latent variable node

The proximal assumption is satisfied by the structure of the graph, and the new cluster variable S satisfies the additional independence assumption $S \perp\!\!\!\perp (Z,W,U)\,|A$ since:

- $U \to A \to S$ conditioning on a non-collider closes the path.
- $U \to Y \leftarrow S$ Not conditioning on a collider closes the path.
- All other paths starting from W,Z are through variations of the previous.

Now the structural equations are identical to the ones the linear proximal two stage least squares on each cluster $S=k$. Similarly to before, taking conditional expectations on both sides, the *first stage* is given by:

$$\begin{aligned}\mathbb{E}\left[W|A,Z,S=k\right] &= \mathbb{E}\left[W|A,Z\right] && (S \perp\!\!\!\perp W|A)\\ &= \mathbb{E}_U\left[\mathbb{E}\left[W|A,Z,U\right]|A,Z\right]\\ &= \beta_{0W} + \beta_{UW}\mathbb{E}\left[U|A,Z\right] && (3.7)\end{aligned}$$

Whereas the second is found as:

$$\begin{aligned}\mathbb{E}\left[Y|A,Z,S=k\right] &= \mathbb{E}_U\left[\mathbb{E}\left[Y|A,Z,U,S=k\right]|A,Z,S=k\right]\\ &= \mathbb{E}_U\left[\mathbb{E}\left[Y|A,Z,U,S=k\right]|A,Z\right] && (S \perp\!\!\!\perp U|A)\\ &= \beta_{0k} + \beta_{Ak}A + \beta_{Uk}\mathbb{E}\left[U|A,Z\right] && (3.8)\end{aligned}$$

Substituting equation (3.7) into equation (3.8) we finally obtain:

$$\mathbb{E}\left[Y|A,Z,S=k\right] = \beta_{0k} + \beta_{Ak}A + \frac{\beta_{Uk}}{\beta_{UW}}\left(\mathbb{E}\left[W|A,Z,S=k\right] - \beta_{0W}\right) \tag{3.9}$$

Notice that this is the same result as in the *regular* P2SLS case within each cluster k.

Proposition 3.2
Let assumption 3.2 hold. Moreover, let $\{\gamma\}$ be the coefficients of the first stage regression and $\{\theta_k\}$ the coefficients of the second stage in cluster k. Then:

$$\chi(a) = \sum \omega_k(a)\left[\left(\theta_{Ak} - \frac{\theta_{Zk}}{\gamma_{ZW}}\gamma_{AW}\right)\cdot a + \left(\theta_{0,k} + \theta_{Z,k}\mathbb{E}\left[Z\right] + \theta_{Z,k}\frac{\gamma_{AW}}{\gamma_{ZW}}\mathbb{E}\left[A\right]\right)\right] \tag{3.10}$$

[1] This assumption does not appear in the original paper but is needed for the validity of the results.

Proof.

$$
\begin{aligned}
\chi(a) &= \mathbb{E}\left[Y^a\right] \\
&= \mathbb{E}_U\left[\mathbb{E}\left[Y|A=a,U\right]\right] && \text{(Backdoor)} \\
&= \mathbb{E}_U\left[\mathbb{E}_S\left[\mathbb{E}\left[Y|A=a,U,S\right]|A=a,U\right]\right] && \text{(Tower property)} \\
&= \mathbb{E}_U\left[\mathbb{E}_S\left[\mathbb{E}\left[Y|A=a,U,S\right]|A=a\right]\right] && (S \perp\!\!\!\perp U|A) \\
&= \mathbb{E}_U\left[\sum P\left(S=k|A=a\right)\mathbb{E}\left[Y|A=a,U,S=k\right]\right] \\
&= \sum P\left(S=k|A=a\right)\mathbb{E}_U\left[\mathbb{E}\left[Y|A=a,U,S=k\right]\right] && (S \perp\!\!\!\perp U|A) \\
&= \sum \omega_k\left(a\right)\mathbb{E}_U\left[\mathbb{E}\left[Y|A=a,U,S=k\right]\right] && \text{(Mixture of normals)} \\
&= \sum \omega_k\left(a\right)\mathbb{E}_U\left[\mathbb{E}\left[Y|Z,A=a,U,S=k\right]\right] && (Y \perp\!\!\!\perp Z|A,U) \\
&= \sum \omega_k\left(a\right)\mathbb{E}_U\left[\beta_{0k}+\beta_{Ak}a+\beta_{Uk}U\right] && \text{(assumption 3.2)} \\
&= \sum \omega_k\left(a\right)\left[\left(\theta_{Ak}-\frac{\theta_{Zk}}{\gamma_{ZW}}\gamma_{AW}\right)\cdot a+\left(\beta_{0k}+\beta_{Uk}\mathbb{E}_U\left[U\right]\right)\right] \\
&= \sum \omega_k\left(a\right)\left[\left(\theta_{Ak}-\frac{\theta_{Zk}}{\gamma_{ZW}}\gamma_{AW}\right)\cdot a+\left(\theta_{0,k}+\theta_{Z,k}\mathbb{E}\left[Z\right]+\theta_{Z,k}\frac{\gamma_{AW}}{\gamma_{ZW}}\mathbb{E}\left[A\right]\right)\right]
\end{aligned}
$$

□

The Bayesian regression to estimate the coefficients of interest θ, γ is carried out with a MCMC procedure, in particular Gibbs sampling. These are methods that are applied when a closed form of the complete joint distribution is not available. Gibbs sampling consists in repeatedly sampling one parameter conditional on all the others from simple distributions. If implemented correctly, in the limit the sampled chain is a representative sample of the posterior distribution in which case various estimators such as the posterior mean can be constructed as the average of the observed parameters in the chain. The particular choice for the prior distributions are Normal distributions for the regression coefficients θ, γ and Inverse Gamma distributions for the variances $\sigma^2_{yk}, \sigma^2_w$ of assumption 3.2. The original α of equation (3.6) were then modeled as a linear function of the treatment $\alpha_k = \eta_{0,k} + \eta_{\nu,k}a$, where $\eta \sim \mathcal{N}\left(\mu_\eta, \sigma^2_\eta\right)$.

Discussion

Overall, the linearity requirements of assumption 3.1 are extremely tight. Indeed the Bayesian approach increases flexibility but requires proper specification of the prior distributions. In the original draft of the paper, the construction of α above-mentioned sampled $\eta_{\nu,k}$ for each quantile ν, essentially losing the ability for points to pass between different clusters. The outcome was quick convergence to a piece-wise linear estimate of the CERF withing each quantile. More of this will be discussed in section 4.3. It is also important to note that the last requirement of assumption 3.1 ensures that both equation (3.1) and equation (3.3) are well defined. Nonetheless, if W is a weak instrument, or weak proxy in our case, i.e. β_{UW} is very small, then the ratio might be ill-conditioned. An additional discussion analysing the relation in terms of the true coefficients with all Gaussian random variables can be found in appendix A.1.1.

3.2 Semiparametric Proximal Inference

Semiparametric statistics is the collection of methods at the intersection of non-parametric and parametric statistics. This is because the former involve unknown infinite dimensional parameters(functions), whereas the latter only deal with finite dimensional ones (subsets of $\mathbb{R}^d$). Semiparametrics deals with both, but places most interest on the finite dimensional parameter, whereas the infinite dimensional one is considered a nuisance component. As the name suggests these are parameters (functions) that are not of direct interest but are necessary to identify lower dimensional parameters of interest. The authors of [13] develop a theory for semiparametric estimation of the ATE in proximal inference. This work was the first work that proposed the exposure bridge function as up until then only the outcome bridge function was considered. The authors determine the influence function for the ATE in and prove that it possesses the doubly robust property. Additionally they show that the found influence function is the efficient influence function, enabling the use of classical results such as the Cramér-Rao bound to claim that an estimator built using it is the most efficient asymptotically linear estimator. This section aims to first introduce basic concepts from semiparametric statistics and then the results from the aforementioned paper.

3.2.1 Background in semiparametrics

A model $\mathcal{P}$ is a collection of probability distributions. Parametric models essentially restrict the search of the density P to a family of distributions $\mathcal{P}_\Theta = \{P_\theta : \theta \in \Theta\}$ indexed by a parameter θ that lives in $\Theta \subseteq \mathbb{R}^d$. Parametric estimators will then be functionals from $\mathcal{P}_\Theta$ to $\mathbb{R}$. As one can imagine, this is restrictive since $\mathcal{P}_\theta$ is only a slice of the set of all possible distributions $\mathcal{P}$. On the other hand, finite dimensional parameters are much more simple to deal with and there exist many useful results from parametric theory such as Fisher information and the Cramér Rao lower bound. Differently, nonparametrics makes no restrictions about the underlying distributions and as such the model $\mathcal{P}$ might be very large. The lack of assumptions and restrictions make it hard to have general results like in the parametric setting.
The idea of semiparametric information theory is to consider sub-models of the large $\mathcal{P}$ on which it is possible to apply results from parametric statistics. In particular we are primarily interested in defining something similar to the Fisher information for an infinite dimensional model $\mathcal{P}$ rather than $\mathcal{P}_\theta$. In the spirit of differential geometry, this is done by considering subparametric models $\mathcal{P}_t$, models parameterized by a finite dimensional parameter, and taking the infimum over all possible models. Since we are interested in defining Fisher information, for a submodel to be considered it must contain the true distributions P_0 and be quadratic mean differentiable at $t = 0$[59], i.e. there exists a measurable function g such that:

$$\int \left(\frac{dP_t^{\frac{1}{2}} - dP_0^{\frac{1}{2}}}{t} - \frac{1}{2} g dP^{\frac{1}{2}} \right)^2 \to 0 \tag{3.11}$$

Similarly to the parametric setting, the function g is the score function of $\mathcal{P}_t$. The collection of all score functions varying over all parametric sub-models gains a particular name, the tangent set.

Definition 3.2 (Tangent Set)
The Tangent Set $\mathcal{T}$ of $\mathcal{P}$ at P_0 is the collection of score functions g of all possible parametric sub-models satisfying equation (3.11)*. Moreover define $\overline{\mathcal{T}}$ as the closure of the linear span of the score functions in $\mathcal{T}$.*

Now, for each submodel we can define the Fisher information in the usual manner as $\mathbb{E}\left[g^2\right]$. For the entire model $\mathcal{P}$, the information will not be larger than *worst case* scenario. Such scenario is not necessarily connected to a specific submodel but could be *reached* as a linear combination of score function of various sub-models, which is why we are interested in $\overline{\mathcal{T}}$ rather than $\mathcal{T}$. The information for $\mathcal{P}$ is then defined as the worst case scenario by taking the infimum.

Example 3.1 (Parametric Model)
Let $\sigma^2 \in \mathbb{R}^+$ be fixed and let $\mathcal{P}_\theta = \left\{\mathcal{N}\left(\theta, \sigma^2\right) \quad \theta \in \mathbb{R}\right\}$ be our parametric model. We are interested in determining the Fisher information of $\mathcal{P}_\theta$. The score function is:

$$g_\theta(x) := \frac{\partial \log dP_\theta(x)}{\partial \theta} = \frac{\partial \log f_\theta(x)}{\partial \theta} = \frac{1}{f_\theta(x)} \cdot \frac{\partial f_\theta(x)}{\partial \theta} \tag{3.12}$$

Notice that:

$$\frac{\partial}{\partial \theta} g_\theta = \frac{\partial}{\partial \theta}\left(\frac{1}{f_\theta}\frac{\partial}{\partial \theta} f_\theta\right) = \frac{1}{f_\theta}\left(\frac{\partial}{\partial \theta} f_\theta\right)^2 - \frac{1}{f_\theta^2}\left(\frac{\partial^2}{\partial \theta^2} f_\theta\right) = \frac{1}{f_\theta}\left(\frac{\partial}{\partial \theta} f_\theta\right)^2 - g_\theta^2$$

Since the Fisher information of $\mathcal{P}_\theta$ is the expected value of the score squared:

$$I(\theta) = \mathbb{E}\left[g_\theta^2(x)\right] = \underbrace{\mathbb{E}\left[\frac{1}{f_\theta}\left(\frac{\partial}{\partial\theta}f_\theta\right)^2\right]}_{=0} - \mathbb{E}\left[\frac{\partial}{\partial\theta}g_\theta\right]$$

Since our parametric model is a collection of Gaussians indexed by the mean, we obtain:

$$g_\theta(x) = \frac{\partial}{\partial\theta}\log\left[\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\theta)^2}{2\sigma^2}}\right] = \frac{\partial}{\partial\theta}\left(-\frac{1}{2}\log\left(2\pi\sigma\right) - \frac{(x-\theta)^2}{2\sigma^2}\right) = \frac{x-\theta}{\sigma^2}$$

$$\frac{\partial}{\partial\theta}g_\theta(x) = -\frac{1}{\sigma^2}$$

Combining the preceding, it quickly follows that the Fisher information of $\mathcal{P}_\theta$ is given by:

$$I(\theta) = \frac{1}{\sigma^2}$$

This result gives some intuition towards the concept of information in models. For small values of σ, the observed data will concentrate around the mean and the information will be large; each sample is very informative about the underlying model and thus very sensitive to observations. On the contrary, large values of σ are associated with large dispersion of the data, leading to small Fisher information and thus more resistant to 'unlikely' observations. This suggests that having high information enables me to have less uncertainty when observing data.

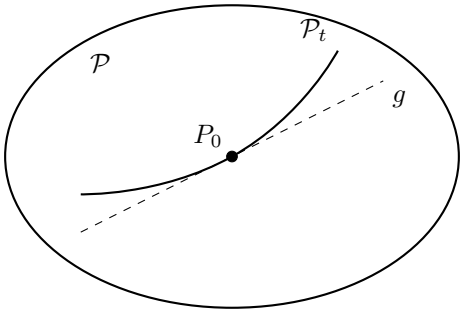


Figure 3.2: Graphical representation of a parametric sub-model and score.

In practice we never observe the true underlying distribution P_0 but only samples from it. Empirical procedures use observed data to produce an estimate, the empirical distribution $\hat{P}_n$. If the procedure is consistent, $\hat{P}_n$ will asymptotically approximate the true underlying distribution P_0. Since an estimator is a map from $\mathcal{P}$ to $\mathbb{R}$ it is reasonable to wonder about the rate of change of $\psi(P)$ while varying P. Since we are working with large spaces of infinite dimensional objects, these are called Fréchet derivatives but the intuition is the same as for *regular* derivatives. Even in the infinite dimensional case, it is immediate that this rate of change with respect to P is closely tied to the concept of Fisher information just as in example 3.1.

Definition 3.3 ([59]Differentiable Functional)
A functional $\psi : \mathcal{P} \to \mathbb{R}$ is said to be differentiable at P relative to a given tangent set $\mathcal{T}$ if there exists a continuous linear map $\dot{\psi} : L_2(P) \to \mathbb{R}$ such that $\forall g \in \mathcal{T}$:

$$\frac{\partial}{\partial t}\psi_t\Big|_{t=0} := \lim_{t\to 0}\frac{\psi(P_t) - \psi(P)}{t} = \dot{\psi}g$$

$\dot{\psi}$ is the Influence Function of ψ.

Thus, if ψ is differentiable its derivative $\dot{\psi}$ is a continuous and thus bounded, linear functional. By theorem 2.1 there exists a representer $\varphi \in L_2$ such that $\dot{\psi}g = \langle\varphi, g\rangle \quad \forall f \in L_2$. Notice that definition 3.3 is only specified for functions in $\mathcal{T}$ and as such, the behavior of $\dot{\psi}$ is arbitrary for functions outside the tangent space. Thus the representer is not necessarily unique in L_2. Any of these representers are influence functions of ψ. Nonetheless, since $\overline{\mathcal{T}}$ is closed by construction, with the same inner product as L_2, it is itself a Hilbert space and thus admits a unique representer within $\overline{\mathcal{T}}$. This particular influence function has particular properties and as such it gains its own definition.

Definition 3.4 (Efficient Influence Function)
Let ψ be a functional of interest and $\dot{\psi}$ its derivative, then there $\exists!\ \varphi_{EIF} \in \overline{\mathcal{T}}$, called Efficient Influence Function such that

$$\dot{\psi} g = \langle \varphi_{EIF}, g \rangle \quad \forall g \in \mathcal{T}$$

By construction and properties of Hilbert spaces, this representer is the projection of any other influence function onto $\overline{\mathcal{T}}$. Since the Fisher information of $\mathcal{P}$ is now introduced, it is now possible to show that the optimal asymptotic variance of an estimator depends on its efficient influence function.

Lemma 3.1 ([59]Asymptotic Variance)
Suppose $\psi : \mathcal{P} \to \mathbb{R}$ is differentiable at P relative to $\mathcal{T}$ and let φ_{EIF} be its efficient influence function then the optimal asymptotic variance is given by $\mathbb{E}\left[\psi^2_{EIF}\right]$.

Proof. This follows from the Cramér-Rao minimum variance unbiased estimator lower bound. It states that the variance of any unbiased estimator must be greater or equal to than $\left(\frac{\partial \psi_t}{\partial t}\right)^2 \cdot \frac{1}{\mathbb{E}[g^2]}$. As we have defined the information on $\mathcal{P}$ as the $\inf_{g \in \overline{\mathcal{T}}} \mathbb{E}\left[g^2\right]$, the Cramér-Rao lower bound is equal to[2] $\sup_{g \in \overline{\mathcal{T}}} \frac{\langle \varphi_{EIF}, g \rangle}{\langle g, g \rangle} \leq \langle \varphi_{EIF}, \varphi_{EIF} \rangle$ where the inequality is due to Cauchy-Schwarz. Since $\varphi_{EIF} \in \overline{\mathcal{T}}$, it is linearly dependent from g and thus the inequality is tight. □

3.2.2 Semiparametric Proximal Inference

Having introduced the concepts from semiparametric theory, it is clear why the influence function is important and even more so, determining the efficient influence function. Lemma 3.1 shows that if we are interested in determining the most efficient regular asymptotically linear estimator, in other words the lowest variance unbiased estimator of ψ, we can do so by identifying the efficient influence function. The theorem from [13] proves that the determined influence function is efficient. We restrict ourselves to showing that it is an influence function and an additional discussion about its efficiency is left to appendix A.

Theorem 3.1 ([13] Semiparametric Proximal Inference)
Consider a binary treatment A taking values in $\mathcal{A} = \{0, 1\}$. Suppose that proximal exchangeability (assumption 1.4) holds. Moreover assume that there exist unique bridge functions h_0, q_0. Then:

$$IF(\psi) = (-1)^{1-A} q_0(Z, A, X)\left[Y - h_0(W, A, X)\right] + h_0(W, 1, X) - h_0(W, 0, X) - \psi \tag{3.13}$$

is an influence function. Moreover, if the conditional expectation operator and its adjoint are surejective, then IF is also efficient and its local efficiency bound is $\mathbb{E}\left[EIF^2(\psi)\right]$[3].

Proof. Introducing the notation $\Delta_A h = h(W, 1, X) - h(W, 0, X)$, apply the proximal g-formula to determine $ATE = \mathbb{E}\left[\Delta_A h\right]$. Let $\mathcal{D} = (Y, A, W, X, Z)$. Using the definition of definition 3.3, we are interested in determining a function φ such that $\frac{\partial}{\partial t}\psi_t\big|_{t=0} = \mathbb{E}\left[\varphi \cdot S(\mathcal{D}; t)\right]\big|_{t=0}$. Applying the chain rule:

$$\frac{\partial}{\partial t}\psi_t\bigg|_{t=0} = \frac{\partial}{\partial t}\mathbb{E}_t\left[\Delta_A h_t\right]\bigg|_{t=0} = \underbrace{\mathbb{E}\left[\Delta_A h \cdot S(W, X)\right]}_{T_1} + \underbrace{\mathbb{E}\left[\frac{\partial}{\partial t}\Delta_A h_t\bigg|_{t=0}\right]}_{T_2}$$

For T_1, notice that:

- $\mathbb{E}\left[\Delta_A h \cdot S(Z, Y, A|W, X)\right] = S(Z, Y, A|W, X)\mathbb{E}\left[\Delta_A h\right] = S(Z, Y, A|W, X) \cdot \psi$
- $\mathbb{E}\left[\left(\Delta_A h - \psi\right) S(W, X)\right] = \mathbb{E}\left[\Delta_A h \cdot S(W, X)\right] - \psi \cdot \underbrace{\mathbb{E}\left[S(W, X)\right]}_{=0} = \mathbb{E}\left[\Delta_A h \cdot S(W, X)\right]$

The above and properties of scores implies that T_1 can be rewritten as:

$$T_1 = \mathbb{E}\left[\left(\Delta_A h_t - \psi\right) \cdot \left(S(Z, Y, A|W, X) + S(W, X)\right)\right] = \mathbb{E}\left[\left(\Delta_A h_t - \psi\right) \cdot S(\mathcal{D})\right]$$

[2] $\frac{1}{\inf A} = \sup\left(\frac{1}{A}\right)$
[3] [13, 26]

Similarly for T_2 notice that using the chain rule, the conditional moment restriction for the outcome function implies:

$$\frac{\partial}{\partial t}\mathbb{E}_t\left[Y - h_t(W,A,X)|Z,A,X\right]\Big|_{t=0} = 0$$

$$\mathbb{E}\left[\underbrace{(Y - h_t(W,A,X))}_{=\epsilon}\cdot S(W,Y|Z,A,X)|Z,A,X\right] = \mathbb{E}\left[\frac{\partial}{\partial t}h_t(W,A,X)\Big|_{t=0}|Z,A,X\right]$$

$$\begin{aligned}
T_2 &= \mathbb{E}\left[\frac{(-1)^{1-A}}{f(A|W,X)}\frac{\partial}{\partial t}h_t(W,A,X)\Big|_{t=0}\right] \\
&= \mathbb{E}\left[(-1)^{1-A}q(Z,A,X)\cdot\frac{\partial}{\partial t}h_t(W,A,X)\Big|_{t=0}\right] \\
&= \mathbb{E}\left[(-1)^{1-A}q(Z,A,X)\cdot\mathbb{E}\left[\frac{\partial}{\partial t}h_t(W,A,X)\Big|_{t=0}|Z,A,X\right]\right] \\
&= \mathbb{E}\left[(-1)^{1-A}q(Z,A,X)\cdot\epsilon\cdot S(W,Y|Z,A,X)\right] \\
&= \mathbb{E}\left[(-1)^{1-A}q(Z,A,X)\cdot\epsilon\cdot S(W,Y|Z,A,X)\right] + \underbrace{\mathbb{E}\left[(-1)^{1-A}q(Z,A,X)\cdot t\cdot S(Z,A,X)\right]}_{=0} \\
&= \mathbb{E}\left[(-1)^{1-A}q(Z,A,X)\cdot\epsilon\cdot S(\mathcal{D})\right]
\end{aligned}$$

Recombining the two terms:

$$\frac{\partial}{\partial t}\psi_t\Big|_{t=0} = \mathbb{E}\left[\left((-1)^{1-A}q(Z,A,X)\cdot(Y - h(W,A,X)) + \Delta_A h - \psi\right)S(\mathcal{D})\right] \tag{3.14}$$

This shows that the following is an influence function for ψ:

$$IF(\psi) = (-1)^{1-A}q(Z,A,X)\cdot(Y - h(W,A,X)) + \Delta_A h - \psi$$

□

The uniqueness of the bridge functions required by theorem 3.1 can be guaranteed by additional completeness assumptions like $W|Z,A,X$ completeness as discussed in section 2.3. The influence function of equation (3.13) can be reformulated highlighting the dependence on the true bridge functions h, q as:

$$IF(\psi;\mathcal{D},h,q) = -\mathbf{1}_{A=a}\cdot q(Z,A,X)h(W,A,X) + \mathbf{1}_{A=a}\cdot Yq(Z,A,X) + h(W,a,X) - \psi \tag{3.15}$$

This particular formulation will come useful in the upcoming sections.

Robustness

Robustness is a property often desired in estimators, as it implies reduced sensitivity to slight changes in the data. Therefore, robustness is closely related to the rate of change of the estimator with respect to changes in P and thus the influence function. A robust estimator is such that for slight variations in $P \in \mathcal{P}, \psi(P)$ also does not vary much.

Example 3.2 (Mean Estimation and Huber Estimators)
Suppose we are interested in estimating the mean μ_0 of some distribution of X with heavy tails. The first reasonable estimator is the sample average $\hat{\mu}$. In particular, if x_i are i.i.d. samples, $\hat{\mu}$ solves:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu) = 0$$

Unfortunately this estimator is not very robust to outliers, which will often occur when sampling from a distribution with heavy tails. A possible solution is to consider so called Huber estimators. Rather than estimating, the sample average of the data, consider $\hat{\mu}$ solving the following:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)\mathbf{1}_{|x_i-\mu|\leq k} - k\mathbf{1}_{(x_i-\mu\leq -k)} + k\mathbf{1}_{(x_i-\mu\geq k)} = 0$$

In this case, the moment equation is constructed such that the outliers do not have much weight in the estimating equation as they will be capped off at k.

In our situation, we have two nuisances: the exposure and outcome bridge functions. The fact that we are leveraging both bridges in the construction of the influence function of equation (3.15) raises the question whether or not there is some robustness to misspecification of at most one of the two.

Corollary 3.1 ([18] Double Robustness)
For all choices of nuisance functions h, q:

$$\mathbb{E}\left[IF(\psi;\mathcal{D},h_0,q_0)\right] = \mathbb{E}\left[IF(\psi;\mathcal{D},h_0,q)\right] = \mathbb{E}\left[IF(\psi;\mathcal{D},h,q_0)\right] = 0$$

Proof. Suppose that $h = h_0$, with a slight abuse of notation foregoing the dependency on the variables, notice that:

$$\begin{aligned} IF(\psi;\mathcal{D},h_0,q_0) - IF(\psi;\mathcal{D},h_0,q) &= -\mathbf{1}_{A=a}q_0h_0 + \mathbf{1}_{A=a}Yq_0 + h_0 - \psi + \mathbf{1}_{A=a}qh_0 - \mathbf{1}_{A=a}Yq - h_0 + \psi \\ &= \mathbf{1}_{A=a}h_0\left(q - q_0\right) - \mathbf{1}_{A=a}Y\left(q - q_0\right) \\ &= \mathbf{1}_{A=a}\left(q - q_0\right)\left(h_0 - Y\right) \end{aligned}$$

By linearity of expectation and tower property:

$$\begin{aligned} \mathbb{E}\left[IF(\psi;\mathcal{D},h_0,q_0)\right] - \mathbb{E}\left[IF(\psi;\mathcal{D},h_0,q)\right] &= \mathbb{E}\left[\mathbf{1}_{A=a}\left(q - q_0\right)\left(h_0 - Y\right)\right] \\ &= \mathbb{E}\left[\mathbf{1}_{A=a}\left(q - q_0\right)\underbrace{\mathbb{E}\left[h_0 - Y|Z,A,X\right]}_{=0}\right] = 0 \end{aligned}$$

Similarly, assume that $q = q_0$ then for the exposure bridge function notice that:

$$\begin{aligned} IF(\psi;\mathcal{D},h_0,q_0) - IF(\psi;\mathcal{D},h,q_0) &= -\mathbf{1}_{A=a}q_0h_0 + \mathbf{1}_{A=a}Yq_0 + h_0 - \psi + \mathbf{1}_{A=a}q_0h - \mathbf{1}_{A=a}Yq_0 - h + \psi \\ &= \mathbf{1}_{A=a}q_0(h - h_0) - (h - h_0) \\ &= \left(\mathbf{1}_{A=a}q_0 - 1\right)\left(h - h_0\right) \end{aligned}$$

Once again by linearity of expectation and tower property:

$$\begin{aligned} \mathbb{E}\left[IF(\psi;\mathcal{D},h_0,q_0)\right] - \mathbb{E}\left[IF(\psi;\mathcal{D},h_0,q)\right] &= \mathbb{E}\left[\left(\mathbf{1}_{A=a}q_0 - 1\right)\left(h - h_0\right)\right] \\ &= \mathbb{E}\left[\underbrace{\mathbb{E}\left[\mathbf{1}_{A=a}q_0 - 1|W,A,X\right]}_{=0}(h - h_0)\right] = 0 \end{aligned}$$

If both nuisances are correctly specified, then it is immediate that $\mathbb{E}\left[IF(\psi;\mathcal{D},h_0,q_0)\right] = 0$ completing the proof. □

To rephrase corollary 3.1, the influence function constructed with only one bridge correctly sepcified is first order bias *resistant* to one misspecification while only losing efficiency.

3.2.3 An estimator built upon the influence function

The double robustness property of the influence function and the efficiency results from semiparametrics suggest that an estimator with the same form as equation (3.15) would inherit such properties. The construction of the estimator is similar to that of M-estimators. Suppose we have n data points, $\mathcal{D}^n = \{(y_i, a_i, w_i, z_i, x_i\}$. To avoid bias generated from overfitting, split the data into L subsets $\mathcal{D}_1, \dots \mathcal{D}_L$. For each $k \in \{1, \dots L\}$ the bridge functions $\hat{g}, \hat{h}$ are estimated using the data from $\mathcal{D}^n \setminus \mathcal{D}_k$. Afterwards $\hat{\psi}_k$ is determined as

$$\hat{\psi}_k = \frac{1}{|\mathcal{D}_k|}\sum_{i\in\mathcal{D}_k}(-1)^{1-a_i}\hat{q}(z_i,a_i,x_i)\left[y_i - \hat{h}(w_i,a_i,x_i)\right] + \hat{h}(w_i,1,x_i) - \hat{h}(w_i,0,x_i)$$

Ultimately, the estimator for the average treatment effect is the average of the previously found $\hat{\psi}_k$s.

$$\hat{\psi} = \frac{1}{L}\sum_{k=1}^{L}\hat{\psi}_k \tag{3.16}$$

In [18], the authors show the asymptotic behaviour of such estimator on conditions regarding the ill posedness measure of the conditional expectation operator E and the convergence rates of the projected distances.

Theorem 3.2 ([18])
Suppose that the true bridge functions h_0, q_0 are bounded. Moreover assume that $\hat{q}, \hat{h}$ are consistent estimators for the bridge functions q, h, i.e. $\left\|h - \hat{h}\right\|_2 = o_P(1), \|q - \hat{q}\|_2 = o_P(1)$. Additionally suppose that $\left\|E(h - \hat{h})\right\|_2 = O(r_h(n))$ and $\|E^*(q - \hat{q})\|_2 = O(r_q(n))$ such that:

$$\min\{r_h \cdot \tau_{\mathcal{Q}}(r_q(n)), r_q \cdot \tau_{\mathcal{H}}(r_h(n))\} = o(n^{-\frac{1}{2}})$$

where E, E^* are the conditional expectation operators defined in section 2.3 and $\tau_{\mathcal{H}}(\delta)$ is the local ill-posedness measure of definition 2.19. Then the estimator of equation (3.16) *is asymptotically linear and:*

$$\sqrt{n}\left(\hat{\psi} - \psi)\right) \rightsquigarrow \mathcal{N}\left(0, var(IF(V; \psi, h_0, q_0)\right)$$

The proof of the theorem is omitted but it consists of splitting the errors into the three terms of equation (2.7): empirical, population, and excess risk and show that under such assumption they all vanish. This theorem shows that the estimator built leveraging corollary 3.1, only possesses second order bias. This in turn implies the Neyman orthogonality property of the estimator, a property recently popularized in the field of machine learning. On top of this, it is important that the rate of convergence of one bridge function and the measure of ill posedness of the class for the other combine to attain a rate of $\sqrt{n}$; only the product bias between the two slacks needs to go to zero at such rate for the estimator to obtain parametric rate. This means that in practice, slow estimation rates of one of the bridge functions can be overcome by faster rates for the other.
The advantages brought from using both bridge functions do not come as free lunches. First of all, there is now the problem of estimating two bridge functions and the theorem does not tell us how, but merely requires it. Moreover, for each one to correctly identify the ATE the respective completeness assumptions of the proximal g-formula (theorems 1.2 and 1.3) need to be satisfied. As shown in example 1.2, this is not a trivial requirement. Furthermore, the derived influence function and the associated methods are restricted to binary treatments. A recent extension to continuous setting and CERF estimation relying on kernel smoothing techniques is presented in section 3.2.4 . Nonetheless, this entails many additional technical assumptions. Additionally, the authors of [13] require the conditional expectation operator and its adjoint to be surjective for the influence function to be efficient, the authors of [26] require bijectivity. These are slightly peculiar requirements and a lengthier discussion on this is included in appendix A.2 .

3.2.4 The influence function as estimating equations

In [18], the influence function of equation (3.15) is used to construct estimating equations for the bridge functions. Consider perturbing the influence function of equation (3.15) to obtain:

$$prt(h_0, q_0; \dot{h}, \dot{q}) = IF(\psi; V, h_0 + \dot{h}, q_0 + \dot{q}) - IF(\psi; V, h_0, q_0)$$

Corollary 3.1 ensures that *varying* the influence function along the axis of figure 3.3, the expected perturbation must remain null. This highlights the fact that the expected perturbation in each direction can be used as estimating equations for the bridge functions:

$$\mathbb{E}\left[prt(h_0, q; 0, \dot{q})\right] = 0 \qquad \forall q, \dot{q} \in \mathcal{Q} \qquad\qquad \mathbb{E}\left[prt(h, q_0; \dot{h}, 0)\right] = 0 \qquad \forall h, \dot{h} \in \mathcal{H}$$

The perturbations in ecah direction result in:

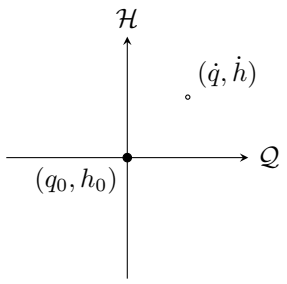


Figure 3.3: Along the axis the expected perturbation is zero.

$$prt(h_0, q; 0, \dot{q}) = \mathbf{1}_{A=a}\dot{q}(V_q)(h_0(V_h) - Y)$$
$$prt(h, q_0; \dot{h}, 0) = \dot{h}(V_h)\left(-\mathbf{1}_{A=a}q_0(V_q) + 1\right)$$

Assumption 3.3 (Well-Posedness)
The true bridge functions q_0, h_0 belong to the hypothesis classes $\mathcal{Q}, \mathcal{H}$ respectively.

Under assumption 3.3 then the bridge function will necessarily be the minimizers of the worst case expected perturbation, in other words, under norm constraints on $\dot{h}, \dot{q}$, they solve the following optimization problem [18]:

$$\begin{aligned} q_0 &= \arg\min_{q}\max_{\dot{h}} \mathbb{E}\left[prt(h, q; \dot{h}, 0)\right] \\ h_0 &= \arg\min_{h}\max_{\dot{q}} \mathbb{E}\left[prt(h, q; 0, \dot{q})\right] \end{aligned} \tag{3.17}$$

Moreover, noticing that the bridge functions are variationally independent due to assumption 1.4, equation (3.17) can be optimized individually. Alternatively, adversarial methods could be implemented to directly optimize in one step:

$$(\hat{h}, \hat{q}) = \arg\min_{\mathcal{H},\mathcal{Q}}\left[\max_{\dot{h}\in\mathcal{H}} \mathbb{E}\left[prt(h, q; \dot{h}, 0)\right] + \max_{\dot{q}\in\mathcal{Q}} \mathbb{E}\left[prt(h, q; 0, \dot{q})\right]\right] \tag{3.18}$$

Optimizing equation (3.18) in one step has the additional difficulties of simultaneous optimization and reaching a stable equilibrium. In [33] the authors study this from a game theoretic perspective. We focus on the individual formulation of proposition 3.3. The following result shows the estimating equations for problem:

Proposition 3.3 ([18])
Under norm constraints on $\dot{h}, \dot{q}$, the solution to the optimization problem in equation (3.17) *is given by:*

$$\begin{aligned} \hat{q} &= \arg\min_{\mathcal{Q}}\max_{\dot{h}} \mathbb{E}\left[\dot{h}(V_h)\cdot(\mathbf{1}_{A=a}q(V_q) - 1)\right] \\ \hat{h} &= \arg\min_{\mathcal{H}}\max_{\dot{q}} \mathbb{E}\left[\mathbf{1}_{A=a}\dot{q}(V_q)\left(h(V_h) - Y\right)\right] \end{aligned} \tag{3.19}$$

Assumption 3.3 ensures that the solutions to the above problem coincide with the true bridge functions. Note that the norm constraint is necessary, otherwise the outer minimization could not be well posed. The other possibility is to impose penalization on norms of said functions as:

$$\begin{aligned} \hat{q} &= \arg\min_{\mathcal{Q}}\max_{\dot{h}} \mathbb{E}\left[\dot{h}(V_h)\cdot(\mathbf{1}_{A=a}q(V_q) - 1) - \dot{h}^2(V_h)\right] \\ \hat{h} &= \arg\min_{\mathcal{H}}\max_{\dot{q}} \mathbb{E}\left[\mathbf{1}_{A=a}\dot{q}(V_q)\left(h(V_h) - Y\right) - \dot{q}^2(V_q)\right] \end{aligned} \tag{3.20}$$

The authors of [18] simply state that these terms are added for stability. These terms are necessary unless norm constraints on the function classes are imposed. Note that these are not the same as regularizers but rather stabilizers as coined in [26]. In estimation, we send the regularization parameter to zero so that our solution is less and less biased(at least for Tikhonov regularization). Stabilizers do not carry bias but can be seen as a different characterization of conditional moment equations[26]. We formalize this in the following result:

Proposition 3.4 (Supremum and Norm Penalization)
Let $\mathcal{G}$ be a Hilbert space, then:

$$\frac{\|f\|^2}{4\lambda} = \sup_{g\in\mathcal{G}}\left[\langle f, g\rangle - \lambda\|g\|^2\right]$$

Proof. Any element of $\mathcal{G}$ can be decomposed into $g = af + g^{\perp}$ with $\langle f, g^{\perp}\rangle = 0$. Then the above display gives:

$$\sup_{a\in\mathbb{R}, g\in\mathcal{G}}\left[a\|f\|^2 - a^2\lambda\|f\|^2 - \lambda\|g^{\perp}\|^2\right]$$

Differentiating in a it is easy to see that this is maximized at $a = \frac{1}{2\lambda}$ and $g^{\perp} = 0$. This gives $\left(\frac{1}{2\lambda} - \frac{\lambda}{(2\lambda)^2}\right)\|f\|^2 = \frac{\|f\|^2}{4\lambda}$ □

The expectation is an inner product on L_2 and thus we can apply proposition 3.4 to a general $f \in L_2$ as:

$$\frac{1}{4\lambda}\mathbb{E}\left[\mathbb{E}\left[Y - h(V_h)|V_q\right]\right]^2 = 0 \iff$$
$$\sup_{g \in L_2} \mathbb{E}\left[g(V_q)\cdot(Y - h(V_h))\right] - \lambda \left\|g\right\|^2 = 0$$

We will find something similar in our own Quasi Bayesian approach in chapter 5. In this work, the authors set $\lambda = 1$ whereas we will require $\lambda = \frac{1}{2}$. For the value of $\lambda = \frac{1}{2}$, a similar result can be found by using a square loss and invoking so called Fenchel duality (appendix A.3.1). The solution to the stabilized optimization is the same as the one without, but with noom constraints.

Corollary 3.2 ([18])
Under assumption 3.3, the solutions of equation (3.20) *satisfy equation* (3.19).

Proof. The proof is given for the first estimating equation, the second follows similarly.

$$\begin{aligned}\mathbb{E}\left[\dot{h}(V_h)\cdot(\mathbf{1}_{A=a}q(V_q) - 1) - \dot{h}(V_h)^2\right] &= \mathbb{E}\left[-\dot{h}(V_h)^2 + \dot{h}(V_h)\left(\mathbf{1}_{A=a}q(V_q) - 1\right) \pm \frac{1}{4}\left(\mathbf{1}_{A=a}q(V_q) - 1\right)^2\right] \\ &= \mathbb{E}\left[-\left(\dot{h}(V_h) - \frac{1}{2}\left(\mathbf{1}_{A=a}q(V_q) - 1\right)\right)^2 + \frac{1}{4}\left(\mathbf{1}_{A=a}q(V_q) - 1\right)^2\right]\end{aligned}$$

The optimization problem can be rewritten as:

$$\hat{q} = \arg\min_{\mathcal{Q}} - \min_{\dot{h}} \mathbb{E}\left[\left(\dot{h}(V_h) - \frac{1}{2}\left(\mathbf{1}_{A=a}q(V_q) - 1\right)^2\right)^2 - \frac{1}{4}\left(\mathbf{1}_{A=a}q(V_q) - 1\right)^2\right]$$

Conditional expectation is the function (in the conditioning variable) that minimizes the MSE. For this reason, the above is optimized at

$$\dot{h}(V_h) = \frac{1}{2}\mathbb{E}\left[\mathbf{1}_{A=a}q(V_q) - 1|V_h\right]$$

Note that this is true only if the above function is within our hypothesis space $\mathcal{H}$. The authors do not require this directly but rather impose a the hypothesis classes $\mathcal{H}, \mathcal{Q}$ to be dense in $L_2(V_h), L_2(V_q)$. Replacing this in the previous display:

$$\begin{aligned}\hat{q} &= \arg\min_{\mathcal{Q}} - \left(\mathbb{E}\left[\left(\frac{1}{2}\mathbb{E}\left[\mathbf{1}_{A=a}q(V_q) - 1|V_h\right] - \frac{1}{2}\left(\mathbf{1}_{A=a}q(V_q) - 1\right)^2\right)^2 - \frac{1}{4}\left(\mathbf{1}_{A=a}q(V_q) - 1\right)^2\right]\right) \\ &= \arg\min_{\mathcal{Q}} - \left(\mathbb{E}\left[\left(\frac{1}{2}\mathbb{E}\left[\mathbf{1}_{A=a}q(V_q) - 1|V_h\right] - \frac{1}{2}\left(\mathbf{1}_{A=a}q(V_q) - 1\right)^2\right)^2 - \frac{1}{4}\left(\mathbf{1}_{A=a}q(V_q) - 1\right)^2\right]\right) \\ &= \arg\min_{\mathcal{Q}} -\frac{1}{4}\left(\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{A=a}q(V_q) - 1|V_h\right]^2 - 2\left(\mathbf{1}_{A=a}q(V_q) - 1\right)\mathbb{E}\left[\mathbf{1}_{A=a}q(V_q) - 1|V_h\right]\right]\right) \\ &= \arg\min_{\mathcal{Q}} -\frac{1}{4}\left(\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{A=a}q(V_q) - 1|V_h\right]^2 - 2\left(\mathbf{1}_{A=a}q(V_q) - 1\right)\mathbb{E}\left[\mathbf{1}_{A=a}q(V_q) - 1|V_h\right]\right]\right) \\ &= \arg\min_{\mathcal{Q}} \frac{1}{4}\left(\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{A=a}q(V_q) - 1|V_h\right]^2\right]\right)\end{aligned}$$

where in the second to last step we used tower property as: $\mathbb{E}\left[f(X)\mathbb{E}\left[f(X)|Y\right]\right] = \mathbb{E}\left[\mathbb{E}\left[f(X)|Y\right]\mathbb{E}\left[f(X)|Y\right]\right] = \mathbb{E}\left[\mathbb{E}\left[f(X)|Y\right]^2\right]$. Since this minimization is non-negative, the solution will be obtained when equality to zero(if possible). Invoking assumption 3.3, then equality is possible and it occurs at q_0 by construction of the exposure bridge function. □

Focusing on RKHSs hypothesis classes for optimization of equation (3.20) , it is possible to control the local ill-posedness measure in terms of Rademacher complexity and critical radii by invoking lemma 2.3. An additional advantage of using such hypothesis classes is that the empirical solutions admit a closed form expression, and can be made to accommodate for additional Tikhonov regularization. In such case the empirical optimization becomes:

$$\begin{aligned}\hat{h} &= \arg\min_{\mathcal{H}} \sup_{\mathcal{Q}} \frac{1}{n}\sum_{i=1}^{n} q_i(h_i - y_i) - h_i^2 - \lambda_{\mathcal{Q}}^h \left\|q\right\|_{\mathcal{Q}}^2 + \lambda_{\mathcal{H}}^h \left\|h\right\|_{\mathcal{H}}^2 \\ \hat{q} &= \arg\min_{\mathcal{Q}} \sup_{\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} h_i\left(\mathbf{1}_{a_i=a}q_i - 1\right) - q_i^2 - \lambda_{\mathcal{H}}^q \left\|h\right\|_{\mathcal{H}} + \lambda_{\mathcal{Q}}^q \left\|h\right\|_{\mathcal{H}}\end{aligned} \tag{3.21}$$

where $h_i = h(V_{h_i})$ and $q_i = q(V_{q_i})$. The solution to the above optimization has a closed form solution given by the following result.

Proposition 3.5 ([18])
Let k_h, k_q be kernel associated to RKHSs $\mathcal{H}, \mathcal{Q}$ respectively. The solution to equation (3.21) is given by:

$$\hat{h} = \sum_{i}^{n} \alpha_i k_h(\mathcal{D}_i^h, \cdot)$$

where $\mathcal{D}_i^h = (w_i, a_i, x_i) \quad i \in 1, \ldots, n$ and

$$\begin{aligned}
\alpha &= \left(K_{h,n} I_A \Gamma I_A K_{h,n} + n^2 \lambda_{\mathcal{H}}^h K_{h,n}\right)^{\dagger} K_{h,n} I_A \Gamma I_{AY} && \in \mathbb{R}^n \\
K_{h,n} &= \left[k_h\left(\mathcal{D}_i^h, \mathcal{D}_j^h\right)\right]_{ij} \quad \forall i,j \in \{0,\ldots,n\} && \in \mathbb{R}^{n \times n} \\
\Gamma &= \frac{1}{4} K_{q,n} \left(\frac{1}{n} K_{q,n} + \lambda_{\mathcal{Q}}^h I\right) && \in \mathbb{R}^n \\
I_A &= \left[\mathbf{1}_{a_i = a} \cdot \delta_{ij}\right]_{ij} \quad \forall i,j \in \{0,\ldots,n\} && \in \mathbb{R}^{n \times n} \\
I_{AY} &= \left[Y_i \cdot \mathbf{1}_{a_i = a} \cdot \delta_{ij}\right]_{ij} \quad \forall i,j \in \{0,\ldots,n\} && \in \mathbb{R}^{n \times n} \\
I &= \left[\delta_{ij}\right]_{ij} \quad \forall i,j \in \{0,\ldots,n\} && \in \mathbb{R}^{n \times n}
\end{aligned}$$

The same result holds for $\hat{q}$ by inverting the roles of k_h and k_q.

Proposition 3.5 shows that to estimate the bridge functions, one simply needs calculate the inverse of two matrices. Implementation wise, this much simpler than other iterated procedures. Computationally, matrix inversion quickly becomes intractable since the computational complexity is $O(n^3)$[8]. In general, this makes kernel methods non scalable. Nonetheless it is possible to implement other approximation techniques such as Nystrom approximations to lighten the computational cost.

3.3 Proximal Kernel Doubly Robust Estimator

The semiparametric proximal approach introduced in section 3.2 has several advantages amongst which the double robustness of the influence function to misspecification of at most one bridge function h_0, q_0. If one is interested in estimating the ATE between two treatments, then the estimator in equation (3.16) is an appropriate estimator. Moreover, if the influence function is efficient, then the estimator is the best asymptotically linear unbiased estimator possible. Unfortunately, due to the indicator function present in the optimizations of equation (3.15), such estimator is only really valid for binary treatment. If one were in the presence of categorical treatment, i.e. the cardinality of $\mathcal{A}$ is finite, one could implement the same approach and compare each treatment to some baseline treatment 0 to obtain $\chi(a)$.

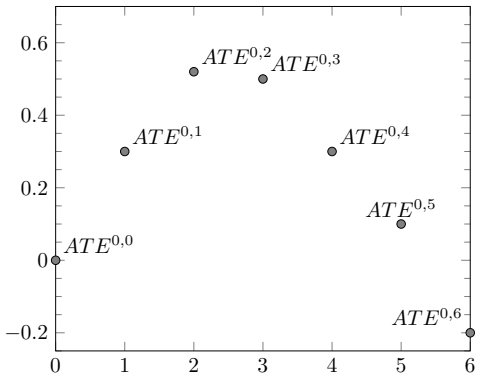


Figure 3.4: Categorical treatment with baseline treatment 0. Ordering is for presentation purposes only. *Connecting* the dots is only natural if the set of treatments $\mathcal{A}$ is an ordered set.

Clearly, this approach becomes computationally inefficient as the cardinality of $\mathcal{A}$ grows. Taking this to the limit, the approach is obviously infeasible to determine the continuous treatment effect and the CERF. The authors of [67] study a method to extend the framework to enable double robust estimation in the continuous setting.

3.3.1 An extension of the semiparametric estimator to the continuous

As previously introduced, the main issue with the semiparametric proposed by [13, 18] is the presence of the indicator or Dirac delta function in equation (3.15). The indicator places a mass of one in an area of infinitesimal size. If one expects some regularity of the bridge functions function, it is possible to borrow a some clever tricks from nonparametric statistics and to spread this mass onto a larger area.

Example 3.3 (Non-parametric density estimation)
A naive approach to estimate the density function of a random variable X is to consider the outline of the histogram of the sampled data. The histograms are nothing but a collection of indicator functions over an interval. In this case the Dirac delta coincides with an indicator function of a bin with zero width. Considering bins of equal width h, a plausible estimator for the density $f(x)$ becomes:

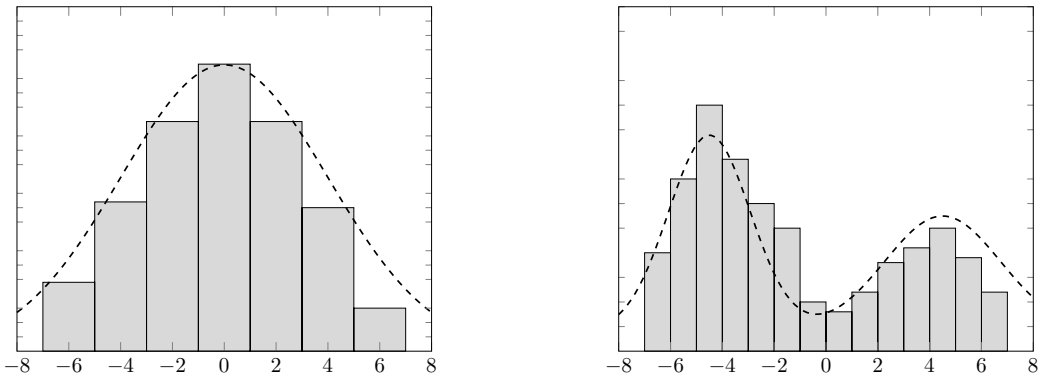


Figure 3.5: Estimating the denisty as the outline of the histogram. On the left data sampled from a $\mathcal{N}(0, 4)$. On the right a more *exotic* bimodal distribution.

$$\hat{f}_\sigma(x) = \frac{1}{2n\sigma} \sum_{i=1}^{n} \mathbf{1}_{(x-\sigma, x+\sigma)}(X_i)$$

This estimator is a particular case of a more general class of estimators known as Kernel Density Estimators(KDE):

$$\hat{f}_\sigma(x) = \frac{1}{n\sigma} \sum_{i=1}^{n} K\left(\frac{x - X_i}{\sigma}\right) = \frac{1}{n} \sum_{i=1}^{n} K_\sigma\left(x - X_i\right)$$

In particular the estimator of figure 3.5 adopts the kernel $K(x - X_i) = \frac{1}{2}\mathbf{1}_{x-\sigma, x+\sigma}(X_i)$, which entails that the samples of X are uniformly distributed within each bin. By changing the kernel function, a different distribution is expected over each interval and it is possible to achieve different results. The most commonly adopted kernel is the so called Gaussian kernel, i.e. $K_\sigma\left(x - X_i\right) = e^{-\frac{(x-X_i)^2}{2\sigma^2}}$. The

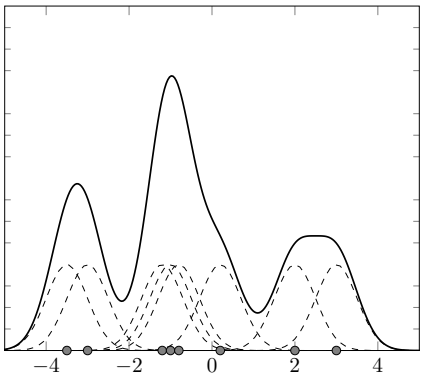


Figure 3.6: Example of KDE with Gaussian kernel. Note that here the plot of each kernel is *normalized* such that the total area equals one.

Gaussian kernel entails a normal distribution centered at the sampled point with variance σ^2. As is highlighted in figure 3.6, the bandwidth parameter σ controls the level of 'influence' that close points have. Large values of σ translate into far away points sharing lots of information whereas small values of σ means distant points remain independent.

A similar approach to KDE is used in the PKDR method. The indicator function in the influence function of equation (3.15) is replaced by a suitable kernel function K_σ:

$$\chi_\sigma(a) = \mathbb{E}\left[K_\sigma\left(A - a\right) q_0(Z, a, X) \cdot Y\right] \tag{3.22}$$

By a suitable kernel we intend one that behaves somewhat similar to the Dirac delta, it has total mass equal to one and it is an even function. Such conditions are presented in the following assumption:

Assumption 3.4 (Kernel conditions)
The kernel function K satisfies the following requirements:

$$\int K(t)dt = 1 \qquad \int tK(t)dt = 0 \qquad \kappa_2(K) = \int t^2 K(t)dt < +\infty$$

Example 3.4 (Epanechnikov Kernel)
The Epanechnikov kernel defined as:

$$K(t) = \frac{3}{4}\left(1 - t^2\right) \qquad |t| \leq 1$$

satisfies the requirements of assumption 3.4.

- $\int K(t)dt = \int_1^1 \frac{3}{4}(1 - t^2)dt = \frac{3}{4}\left(t - \frac{t^3}{3}\right)|_{-1}^{1} = \frac{3}{4}\left(1 - \frac{1}{3} + 1 - \frac{1}{3}\right) = 1$
- *$K(t)$ is even w.r.t. 0 and t is odd* $\implies$ *$t \cdot K(t)$ is odd.*
- $\int K(t)dt = \int_1^1 \frac{3}{4}(t^2 - t^4)dt = \frac{3}{4}\left(\frac{t^3}{3} - \frac{t^5}{5}\right)|_{-1}^{1} = \frac{9}{20}$

In this case, rather than computing one to one comparisons for discrete variables, information is automatically *shared* by the smoothing of the kernel K. If the bandwidth parameter of the kernel were zero, then one would expect the estimator built upon equation (3.22) to perfectly match the one achieved by the proximal g-formula.

Proposition 3.6 ([67])
Suppose that $\chi(a) = \mathbb{E}\left[\mathbf{1}_{A=a} q_0(Z,a,X)Y\right]$ is continuous and bounded in a. Moreover, suppose that the kernel K_σ satisfies assumption 3.4 then:

$$\chi(a) = \lim_{\sigma \to 0} \chi_\sigma(a)$$

where $\chi_\sigma(a)$ is defined in equation (3.22).

The above result is directly taken from [67]. It is not evident what is meant by "$\mathbb{E}\left[\mathbf{1}_{A=a} q_0(Z,a,X)Y\right]$ *is continuous in* a". If the treatment indeed takes values in $\mathcal{A} \subset \mathbb{R}$, then the above quantity is always equal to zero. This is because the set $A = a \in \mathbb{R}$ has measure 0. If the $\mathcal{A}$ contains atoms then the treatment cannot be continuous. To somewhat make sense of the above, a possibility is to consider $\mathbb{E}\left[Y q(Z,a,X) f(a|X) | A = a\right]$ and require continuity of the conditional density in a; but it is ultimately not clear if this is what is implied or true. In any case, we give the proof as if such object was well defined (and not simply 0).

Proof. Since χ is bounded, we can invoke the dominated convergence theorem to switch order of the limit and integral. Using variable substitution $u = \frac{s-a}{h}$, the definition of $K_\sigma(t) = \frac{K(t/\sigma)}{\sigma}$ and assumption 3.4 we have that $\forall a$:

$$\lim_{\sigma \to 0} \int K_\sigma(s-a)\chi(s)ds = \int \lim_{\sigma \to 0} \frac{K\left(\frac{s-a}{\sigma}\right)}{\sigma} \chi(s)ds = \int \lim_{\sigma \to 0} K(u)\chi(\sigma u + a)du = \int K(u)\chi(a)du = \chi(a)$$

□

A direct consequence of proposition 3.6 is that an estimator with the form:

$$\chi_\sigma(a) = \mathbb{E}\left[K_\sigma\left(A-a\right)\left(Y - h_0(W,a,X)\right) \cdot q_0(Z,a,X) + h_0(W,a,X)\right] \tag{3.23}$$

gains the advantages previously discussed as the bandwidth parameter tends to zero. This suggests its empirical equivalent:

Definition 3.5 (PKDR Estimator)
Let $\mathcal{D}^t$ be a test set of size n_t, the PKDR Estimator for the CERF is defined as:

$$\hat{\chi}_\sigma(a) = \frac{1}{n_t \sigma} \sum_{i \in \mathcal{D}^t} K\left(\frac{a_i - a}{\sigma}\right)\left(y_i - \hat{h}(w_i,a,x_i)\right) \cdot \hat{q}(z_i,a,x_i) + \hat{h}(w_i,a,x_i) \tag{3.24}$$

Kernel smoothing does carry some downsides. Even though it can be shown that the KDE of example 3.3 has bias going to zero as the bandwidth σ decreases, this comes at the cost of second order bias or variance which increases. This is the well known bias-variance tradeoff. The same happens for the PKDR estimator. The following theorem determines the conditions necessary and the optimal rates for the aforementioned tradeoff.

Theorem 3.3 ([67])
Suppose that $\hat{h}, \hat{q}$ well approximate the true bridge functions h_0, q_0, i.e. $\left\|\hat{h} - h_0\right\|_2 = o(1), \left\|\hat{q} - q_0\right\|_2 = o(1)$. Moreover suppose that $\left\|\hat{h} - h_0\right\|_2 \left\|\hat{q} - q_0\right\|_2 = O\left((n\sigma)^{-\frac{1}{2}}\right)$

- $n\sigma^5 < +\infty, n\sigma \to \infty$
- $h_0, f(z,a|w,x), f(w,a|z,x) \in C_a^2(\mathcal{A})$
- $\sup\{|h_0|, |q_0|, |\hat{h}|, |\hat{q}|\} < +\infty$

Then the PKDR estimator of equation (3.24) *has:*

$$Bias\left(\hat{\chi}(a)\right) = \frac{\sigma^2}{2}\kappa_2(K)B + o\left((n\sigma)^{-\frac{1}{2}}\right) \qquad B = \mathbb{E}\left[q_0 \cdot \left(\frac{\partial}{\partial a} h_0(W,a,X) \cdot \frac{\partial}{\partial a} f(W,a|Z,X) + \frac{\partial^2}{\partial a^2} h_0(W,a,X)\right)\right]$$

$$Var\left(\hat{\chi}(a)\right) = \frac{\Omega_2(K)}{n\sigma}\kappa_2(K)\left(V + o(1)\right) \qquad V = \mathbb{E}\left[\mathbf{1}_{A=a} q_0^2(Z,a,X) \cdot \left(Y - h_0(W,a,X)\right)\right]$$

From the requirements of the above theorem, it is immediate that the optimal bias-variance trade-off is achieved by a bandwidth $\sigma = o(n^{-\frac{1}{5}})$. This leads to an optimal rate for $\|\chi - \hat{\chi}\|_2 = O(n^{-\frac{2}{5}})$. Whereas continuity assumptions on the conditional densities of treatments might be reasonable for a given problem, the same requirement becomes highly technical in the case of the outcome bridge function. As discussed in section 1.3, the bridge functions do not have a clear interpretation and as such assumptions of their regularities must always be made with care.

3.4 Kernel Methods

In this next section we focus on estimation of the outcome bridge functions in the case it lies within some RKHS. The main advantage of these methods is that the solution to the optimization problem comes in closed form. The authors of [36] develop two kernel methods for the continuous estimation of the CERF. The first, KPV, is a two stage method that learns the conditional expectation operator with its conditional mean embedding. The second method considers the maximum violation of the conditional moment restriction over the test function class $\mathcal{G}$. In this section we will often use the feature maps of the kernels associated with the reproducing kernel Hilbert spaces. Then notation we adopt is the same of definition 2.6, $\varphi_X(x_i) = k_X(x_i, \cdot)$. It is important to remember that even though $\varphi_X(x_i)$ might seem like the evaluation of a function, and thus a number, it is not. A feature map φ_X evaluated $\varphi_X(x_i)$ is an element of $\mathcal{H}_\mathcal{X}$. In particular it is the function associated to fixing the an input (the kernel is symmetric) to x_i and letting the remaining one roam free. By the *reproducing trick*, this is inline with the Machine learning notion of feature maps: $\langle \varphi_X(x_i), \varphi_X(x_j) \rangle_{\mathcal{H}_\mathcal{X}} = k_X(x_i, x_j)$. Moreover, since the bridge function takes in three arguments, we will work with reproducing kernel Hilbert spaces associated to kernels that take two inputs, each of dimension 3. We will choose to use the tensor kernel as the spaces $\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}$ and $\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}$ are isometrically isomorphic [36]. This is because we work in separable Hilbert spaces[4] and the reproducing trick:

$$\begin{aligned}\langle f \otimes g, \varphi_X(x_i) \otimes \varphi_Y(y_i) \rangle_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}} &= \langle f \otimes g, k_X(x_i, \cdot) \otimes k_Y(y_i, \cdot) \rangle_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}} \\ &= \langle f, k_X(x_i, \cdot) \rangle_{\mathcal{H}_\mathcal{X}} \langle g, k_Y(y_i, \cdot) \rangle_{\mathcal{H}_\mathcal{Y}} = f(x_i) g(y_i)\end{aligned}$$

To lighten notation, the feature map will always remain φ and which RKHS it is associated to will be immediate by its argument: $\varphi(x_i) = k_X(x_i, \cdot) \in \mathcal{H}_\mathcal{X}, \varphi(a_i) = k_A(a_i, \cdot) \in \mathcal{H}_\mathcal{A}, \cdots$ and so on. For these reasons, we will use the notation $\varphi(a_i, x_i)$ and $\varphi(a_i) \otimes \varphi(x_i)$ interchangeably.

3.4.1 Kernel Proxy Variable

In the Proxy Two Stage Least Squares approach, the first stage regression attempts to learn the conditional distribution $W|Z, A, X$ under the assumption that the true model is relegated to a linear distribution. The learnt conditional distribution is used to solve for the CERF, the parameter β_A, in the second stage. Clearly, the linear assumption is often excessively stringent. The Kernel Proxy Variable approach is a two stage non-parametric approach that weakens the linearity assumption by projecting the features into reproducing kernel Hilbert spaces and *perform* the two regressions there. The main assumptions of the method ensure that the problem we are solving is well posed with respect to the hypothesis classes considered, similar by introducing regularizes and source assumptions as discussed in section 2.3.

First Stage

The first stage aims to learn the conditional dependency between the outcome proxy and the treatment, its proxy and the measured confounder. In particular, this is done by searching for the conditional mean embedding of the conditional expectation operator E. Remember that the conditional mean embedding $\mu_{W|Z,A,X}$ is an operator between $\mathcal{H}_\mathcal{W}$ and $\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{X}}$ (definition 2.11). By the isometry between Hilbert spaces and the space of operators, $\mu_{W|Z,A,X}$ can also be seen as an element of $\mathcal{H}_{\mathcal{W},\mathcal{Z},\mathcal{A},\mathcal{X}}$. As such, we can evaluate it the conditioning argument:

$$\left\langle \mu_{W|Z,A,X}, \varphi(z_i, a_i, x_i) \right\rangle_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{X}}} = \mu_{W|z_i,a_i,x_i} \in \mathcal{H}_\mathcal{W}$$

. This is better explained in section 2.1. We would then have that:

$$\mathbb{E}\left[h(W, a_i, x_i) | z_i, a_i, x_i\right] = \left\langle h, \mu_{W|z_i,a_i,x_i}) \right\rangle_{\mathcal{H}_\mathcal{W}} \tag{3.25}$$

If one were to determine such $\mu_{W|Z,A,X}$, then by corollary 2.1 it would be possible to determine the conditional expectation of any function h by simply projecting onto $\mu_{W|Z,A,X}$, enabling us to minimize some risk involving the characterizing conditional moment restriction without having to work with the conditional expectation operator directly. This first stage is assumed to be well-posed by requiring that the conditional mean embedding lies within some power space of the chosen reproducing kernel Hilbert space $\mathcal{G} \subset L_2(Z, A, X)$, similar to the discussion previously held on regularizers.

[4] If the reproducing kernel k_X is continuous and its domain $\mathcal{X}$ is Polish then $\mathcal{H}_\mathcal{X}$ is separable [34].

Assumption 3.5
The conditional mean embedding is well defined within some power space of the covariance operator: there $\exists c_1 \in (0,1] : E\varphi(W) \in \mathcal{H}^{c_1}_{\mathcal{Z}\mathcal{A}\mathcal{X}}$*, where E is the conditional expectation operator.*

This assumption guarantees the well posedness of the first stage regression and assumes the regularity of $\mu_{W|z,a,x}$ as c_1. The rates of convergence are then going to depend on this parameter. In practice, given a set of samples we solve the ridge regression problem on a first partition of the dataset $\mathcal{D}_1$ of size n_1:

$$\hat{\mu}_{W|Z,A,X} = \underset{\mu_{W|Z,A,X} \in \mathcal{H}_{\mathcal{W}\mathcal{Z}\mathcal{A}\mathcal{X}}}{\arg\min} \frac{1}{n_1} \sum_{i \in \mathcal{D}_1} \left\| \varphi(w_i) - \mu_{W|z_i,a_i,x_i} \right\|^2_{\mathcal{H}_\mathcal{W}} + \lambda_1 \left\| \mu_{W|Z,A,X} \right\|^2_{\mathcal{H}_{\mathcal{W}\mathcal{Z}\mathcal{A}\mathcal{X}}} \tag{3.26}$$

Although it might not seem like it, we are trying to learn a matrix of weights. By the properties of RKHS and seeing the operator by its spectral decomposition, we can rewrite the inner part of above as:

$$\begin{aligned}
&\frac{1}{n_1} \sum_{i \in \mathcal{D}_1} \left\| k(w_i, \cdot) - \left\langle \mu_{W|Z,A,X}, k((z_i,a_i,x_i), \cdot) \right\rangle_{\mathcal{H}_{Z,A,X}} \right\|^2_{\mathcal{H}_W} \\
=&\frac{1}{n_1} \sum_{i \in \mathcal{D}_1} \left\| k(w_i, \cdot) - \sum_{j=1}^{\infty} \alpha_{i,j} k(w_j, \cdot) \cdot \langle k((z_i,a_i,x_i), \cdot), k((z_j,a_j,x_j), \cdot) \rangle_{\mathcal{H}_{Z,A,X}} \right\|^2_{\mathcal{H}_W} \\
=&\frac{1}{n_1} \sum_{i \in \mathcal{D}_1} \left\| k(w_i, \cdot) - \sum_{j=1}^{\infty} \alpha_{i,j} k(w_j, \cdot) \cdot k((z_i,a_i,x_i), (z_j,a_j,x_j)) \right\|^2_{\mathcal{H}_W}
\end{aligned}$$

In practice, the inner infinite sum is replaced by the sum over $\mathcal{D}_1$. This is because we will search in the span of the n_1 terms $k(w_i, \cdot) \quad w_i \in \mathcal{D}_1$. Thus, for fixed kernel, $\mu_{W|Z,A,X}$ will be an $n_1 \times n_1$ matrix consisting of the above α_{ij}. The formulation of the reproducing kernel Hilbert space enables us to have a closed form solution to equation (3.26) in terms of the kernel matrices containing the basis functions evaluated at the sampled data.
The extra regularization term adds additional bias to the solution as previously discussed, and must decay to zero as the data increases. Under the optimal choice of such decay, the estimated conditional mean embedding is consistent and has a rate given by the following result:

Proposition 3.7 ([36]Stage 1 Consistency)
Let

$$\lambda_1 \asymp \left(\frac{\ln(2/\delta)}{\sqrt{n_1 c_1}} \right)^{\frac{2}{c_1+2}}$$

Then, for any $z, a, x \in \mathcal{Z} \times \mathcal{A} \times \mathcal{X}$ *and any* $\delta \in (0,1)$*, the following holds with probability* $1-\delta$*:*

$$|\hat{\mu}_{W|z,a,x} - \mu_{W|z,a,x}|_{\mathcal{H}} \leq \kappa^3 r_C(\delta, n_1, c_1) \asymp (c_1+2) \cdot \left(\frac{\ln(2/\delta)}{\sqrt{n_1} c_1} \right)^{\frac{c_1}{c_1+2}}$$

The proof for this result can be found in the appendix of [36]. It is technical and does not provide any additional insight, so we choose to omit it.

Stage 2

The second stage regression now attempts to estimate the outcome bridge function by minimizing a risk functional. In this case, the authors decide to minimize:

$$R(f) = \mathbb{E}\left[(Y - \mathbb{E}\left[f(W,A,X) | Z,A,X \right])^2 \right] \tag{3.27}$$

If the conditional expectation operator were perfectly known, then it could be used directly to calculate the second term in the risk. In practice we will plug in the estimate for the conditional mean embedding $\mu_{W|Z,A,X}$ obtained in the first stage. However, the following result shows that, in search for the bridge function, this is a valid risk in the sense that the true bridge function minimizes it.

Proposition 3.8 ([36])
The functional of equation (3.27) *is minimized by the true outcome bridge function* h_0*.*

Proof.

$$\begin{aligned} R(f) &= \mathbb{E}\left[(Y - \mathbb{E}\left[f(W,A,X)|Z,A,X\right])^2\right] \\ &\geq \mathbb{E}\left[(Y - \mathbb{E}\left[Y|Z,A,X\right])^2\right] \\ &= \mathbb{E}\left[(Y - \mathbb{E}\left[h_0(W,A,X)|Z,A,X\right])^2\right] \end{aligned}$$

The inequality follows from the fact that the conditional expectation is the function that minimizes the MSE, i.e. $\mathbb{E}\left[Y|Z,A,X\right] = \underset{g \in L_2(Z,A,X)}{\arg\min}\ \mathbb{E}\left[(Y - g(Z,A,X))^2\right]$. Then, by definition of outcome bridge function $\mathbb{E}\left[h_0(W,A,X)|Z,A,X\right] = \mathbb{E}\left[Y|Z,A,X\right]$ we obtain the last equality. Moreover, since the expectation of a non-negative function (x^2) remains non-negative, any function that is not the true bridge function will obtain a non-zero positive risk. □

Applying the law of iterated expectation and Jensen's inequality, it is clear that this risk is an upper-bound for the mean violation of the CMR[36] :

$$\mathbb{E}_{Y,Z,A,X}\left[(Y - \mathbb{E}\left[f|Z\right])^2\right] = \mathbb{E}_{Z,A,X}\left[\mathbb{E}_{Y|Z,A,X}\left[(Y - \mathbb{E}\left[f|Z\right])^2\right]\right] \geq \mathbb{E}_{Z,A,X}\left[(\mathbb{E}\left[Y|Z,A,X\right] - \mathbb{E}\left[f|Z,A,X\right])^2\right]$$

The method is restricted to consider the risk over a RKHS class $\mathcal{H}_{W,A,X}$. Moreover, the authors assume $W|Z,A,X$ completeness over $\mathcal{H}_{W,A,X}$. This makes it such that the distance to the true bridge function h_0 in $\mathcal{H}_{W,A,X}$ can be by minimizing the aforementioned risk. Similarly to the first stage, we must impose some well posedness restrictions, in particular that the solution is contained within some power space of the range of operator

Assumption 3.6
The minimizer of equation (3.27) *belongs to* $\mathcal{H}_{WAX}^{c_2}$ *for some* $c_2 \in (0,1]$. *Assume the problem to be mildly ill posed with* $\nu_i^2 \asymp i^{-b}$.

To estimate the outcome bridge function, the second stage substitutes the true conditional expectation operator by the previously learnt conditional mean embedding $\hat{\mu}_{W|z,a,x}$. To avoid bias from over fitting, the second stage is estimated on a new partition of the dataset $\mathcal{D}_2$.

$$\hat{R}(h) = \sum_{i \in \mathcal{D}_2} \left(y_i - \left\langle h, \varphi(a_i,x_i) \otimes \hat{\mu}_{W|z_i,a_i,x_i} \right\rangle\right)^2 + \lambda_2 \left\|h\right\|_{\mathcal{H}_{W,A,X}}^2 \tag{3.28}$$

As previously mentioned, one of the main advantages of working in a RKHS is the computational simplicity in achieving a closed form solution. The regularized problem admits the following solution.

Theorem 3.4 ([36])
For any $\lambda_2 > 0$, *there exists a unique solution to the empirical risk minimization problem in equation* (3.28) *and it is given by* $\hat{V} = (\hat{T}_2 + \lambda_2)^{-1}\hat{g}$ *with:*

$$\begin{aligned} \hat{T}_2 &= \frac{1}{n_2}\sum_{i=1}^{n_2}\left[\mu_{W|z_i,a_i,x_i} \otimes \varphi(a_i,x_i)\right] \otimes \left[\mu_{W|z_i,a_i,x_i} \otimes \varphi(a_i,x_i)\right] \\ \hat{g} &= \frac{1}{n_2}\sum_{i=1}^{n_2}\left[\mu_{W|z_i,a_i,x_i} \otimes \varphi(a_i,x_i)\right] y_i \end{aligned}$$

The convergence rate of final estimated bridge function is terms of optimal regularization coefficients and the dimensions of the two partitions used $\mathcal{D}_1, \mathcal{D}_2$.

Theorem 3.5 ([36]Stage 2 Convergence)
Under the previous assumptions, fix $\zeta > 0$ *and choose* $\lambda_1 = n_1^{\frac{1}{c_1+2}}$ *and* $n_1 = n_2^{\frac{\zeta(c_1+2)}{(c_1)}}$, *then:*

1. *If* $\zeta \leq \frac{b(c_2+2)}{b(c_2+2)}$) *choose* $\lambda_2 = n_2^{-\frac{\zeta}{c_2+2}}$. *Then:*

$$\hat{R}(\hat{h}) - \hat{R}(h_0) = O_p\left(n_2^{-\frac{\zeta(c_2+1)}{c_2+2}}\right)$$

2. *If* $\zeta \geq \frac{b(c_2+2)}{b(c_2+2)}$) *choose* $\lambda_2 = n_2^{-\frac{b}{b(c_2+1)+1}}$. *Then:*

$$\hat{R}(\hat{h}_{WAX}) - \hat{R}(h_0) = O_p\left(n_2^{-\frac{b(c_2+1)}{b(c_2+1)+1}}\right)$$

The KPV estimator is then built by projecting the solution of stage 2 onto an estimator for the marginal mean embedding μ_{WX}.

Definition 3.6 (KPV Estimator)
Let $\mathcal{D}^t$ be a test set of unobserved samples of size n_t. Let $\hat{h}$ be the estimate of the outcome bridge function from the KPV two stage approach. The KPV Estimator for the CERF is defined as:

$$\hat{\chi}(a) = \left\langle \hat{h}, \hat{\mu}_{W,X} \otimes \varphi(a) \right\rangle_{\mathcal{H}_{WX}}$$

where $\hat{\mu}_{WX}$ is an estimator for the marginal mean embedding $\hat{\mu}_{WX} = \frac{1}{n_t} \sum_{i \in \mathcal{D}^t} \varphi(w_i) \otimes \varphi(x_i)$

The authors of [36], then determine the estimation error at test time on $\mathcal{D}^t$. Just as before, the rate depends on the previous regularities.

Corollary 3.3 ([36]Causal Consistency)
Consider a sample at test time $\mathcal{D}^t$. Let λ_1, n_1 be as in theorem 3.5 and $\hat{\mu}_{XW}$ as in definition 3.6. Then $\forall a \in \mathcal{A}$, under the assumptions of theorem 3.5 the KPV estimator is such that:

1. *If $\zeta \leq \frac{b(c_2+2)}{b(c_2+1)+1}$ then:*
$$|\hat{\chi}(a) - \chi(a)| \leq O_p\left(n_t^{-\frac{1}{2}} + n_2^{-\frac{\zeta(c_2)}{c_2+2}}\right)$$
2. *If $\zeta \geq \frac{b(c_2+2)}{b(c_2+1)+1}$ then:*
$$|\hat{\chi}(a) - \chi(a)| \leq O_p\left(n_t^{-\frac{1}{2}} + n_2^{-\frac{bc_2}{b(c_2+1)+1}}\right)$$

3.4.2 Proximal Maximum Moment Restriction

As already discussed in section 2.3 the bridge functions are characterized by conditional moment restrictions. The Proxy Maximum Moment Restriction(PMMR) exploits this restriction by focusing on RKHSs to estimate the *best* bridge function in a one step procedure.
By tower property of expectation and conditional moment restriction of the bridge function, we have that for all measurable $g \in L_2(V_q)$:

$$\mathbb{E}\left[(Y - h_0) \cdot g(V_q)\right] = \mathbb{E}\left[\underbrace{\mathbb{E}\left[Y - h_0 | V_q\right]}_{=0} \cdot g(V_q)\right] = 0$$

If the conditional moment restriction is imposed uniformly over all functions in a certain class $\mathcal{G} \subset L_2$, i.e.

$$\forall g \in \mathcal{G} \quad \mathbb{E}\left[(Y - h(V_h))\, g(V_q)\right] = 0$$

we refer to it as a maximum moment restriction over $\mathcal{G}$. As shown above, the true bridge function is such that it respects a maximum moment restriction uniformly over all of L_2. Thus, the PMMR method leverages this by constructing the risk as:

$$R(h) = \sup_{g \in \mathcal{G}} \mathbb{E}\left[(Y - h(V_h))\, g(V_q)\right]^2 \tag{3.29}$$

Let $k : V_q \times V_q \to \mathbb{R}$ be a reproducing kernel associated to $\mathcal{G}$ to be a reproducing kernel Hilbert space, then the risk can nicely be reformulated in terms of the inner products on $\mathcal{G}$:

Lemma 3.2
Assume that $\mathbb{E}\left[(Y - h(V_h))^2 \cdot k(V_q, V_q))\right] < +\infty$ and let V' be an independent copy of the random variable V. Then, the risk of equation (3.29) can be reformulated as:

$$R(h) = \mathbb{E}\left[(Y - h(V_h))(Y' - h(V_h'))k(V_q, V_q')\right] \tag{3.30}$$

Proof.
Notice that by requiring $\mathbb{E}\left[(Y - h(V_h))^2 \cdot k(V_q, V_q))\right] < +\infty$, we have that $\mathbb{E}\left[(Y - h(V_h))^2 \cdot k(V_q, \cdot))\right] \in$

$\mathcal{G}$. By the reproducing trick on $\mathcal{G}$ we have that $g(V_q) = \langle g, k(V_q, \cdot)\rangle$ and thus:

$$\begin{aligned} R(h) &= \sup_{g\in\mathcal{G}} \mathbb{E}\left[(Y - h(V_h))\,\langle g, k(V_q,\cdot)\rangle\right]^2 \\ &= \sup_{g\in\mathcal{G}} \mathbb{E}\left[\langle g, (Y - h(V_h))\cdot k(V_q,\cdot)\rangle\right]^2 \\ &= \sup_{g\in\mathcal{G}} \left(\langle g, \mathbb{E}\left[(Y - h(V_h))\cdot k(V_q,\cdot)\right]\rangle\right)^2 \\ &= \left\|\mathbb{E}\left[(Y - h(V_h))\cdot k(V_q,\cdot)\right]\right\|_{\mathcal{G}}^2 \end{aligned}$$

where the last equality follows from the fact that the inner product is a projection and, since its element is withing $\mathcal{G}$ by assumption, its supremum must be the projection against itself. Then, by reproducing property once more:

$$\begin{aligned} R(h) &= \langle \mathbb{E}\left[(Y - h(V_h))\cdot k(V_q,\cdot)\right], \mathbb{E}\left[(Y - h(V_h))\cdot k(V_q,\cdot)\right]\rangle \\ &= \mathbb{E}\left[\left\langle (Y - h(V_h))\cdot k(V_q,\cdot), (Y' - h(V_h'))\cdot k(V_q',\cdot)\right\rangle\right] \\ &= \mathbb{E}\left[(Y - h(V_h))\,(Y' - h(V_h'))\,k(V_q, V_q')\right] \end{aligned}$$

□

The expectation of equation (3.30) can be approximated with the sample average as a V-statistic:

$$\hat{R}(h) = \frac{1}{n^2}\sum_{i,j\in\mathcal{D}^n} (y_i - h(V_{h_i}))\left(y_j - h(V_{h_j})\right) k(V_{q_i}, V_{q_j}) \tag{3.31}$$

Notice that this estimator is actually a biased estimator for R since when $i = j$ the samples used are not independent. Additionally, unlike in the KPV counterpart, working in a single stage we do not need to split the data into non-overlapping subsets. The PMMR estimate for the outcome bridge function can then be determined as a solution to the Tikhonov regularized ridge regression:

$$\hat{h}_\lambda = \arg\min_{\mathcal{H}} \hat{R}_V(h) + \lambda \|h\|_{\mathcal{H}}^2 \tag{3.32}$$

Let $T : \mathcal{H} \to \mathcal{G}$ be the operator such that $Th = \mathbb{E}\left[h(V_h)k_{\mathcal{G}}(V_q,\cdot)\right]$. The regularized risk and its empirical counterpart are given by:

$$R_\lambda(h) = \|g - Th\|_{\mathcal{G}}^2 + \lambda\|h\|_{\mathcal{H}}^2 \qquad\qquad \hat{R}_\lambda(h) = \left\|\hat{g} - \hat{T}h\right\|_{\mathcal{G}}^2 + \lambda\|h\|_{\mathcal{H}}^2$$

The minimizers of these risks can be found by expanding as:

$$\begin{aligned} R_\lambda(h) &= \langle g - Th, g - Th\rangle_{\mathcal{G}} + \lambda\langle h, h\rangle_{\mathcal{H}} \\ &= \langle g, g\rangle_{\mathcal{G}} - 2\langle g, Th\rangle_{\mathcal{G}} + \langle Th, Th\rangle_{\mathcal{G}} + \lambda\langle h, h\rangle_{\mathcal{H}} \\ &= \|g\|_{\mathcal{G}}^2 - 2\langle T^*g, h\rangle_{\mathcal{H}} + \langle T^*Th, h\rangle_{\mathcal{H}} + \lambda\langle h, h\rangle_{\mathcal{H}} \\ &= \|g\|_{\mathcal{G}}^2 + \langle h, (T^*T + \lambda I)h - 2T^*g\rangle_{\mathcal{H}} \end{aligned} \tag{Δ}$$

Differentiating[5] in h and setting equal to zero, we find that the minima is obtained at:

$$\begin{aligned} &\frac{\partial}{\partial h}R_\lambda(h) = (T^*T + \lambda I)h - 2T^*g + (T^*T + \lambda I)h = 0 \\ &\iff h_\lambda = (T^*T + \lambda I)^{-1}T^*g = \Gamma_\lambda T^*g \end{aligned}$$

Assumption 3.7 (PMMR Assumption)
Assume that the true bridge function is $h_0 \in Range(T^*T^\beta)$ for $\beta > 0$. Moreover, the kernel associated with $\mathcal{G}$ is bounded and integrally strictly positive definite($\iint f(V_q)k(V_q, V_q')f(V_q')dV_q, dV_q' = 0 \iff f = 0$).

Note that this assumption removes the ill-posedness of the problem by imposing smoothness requirements onto the outcome bridge function h_0. The problem of determining h such that $T^*Th = h_0$ is now automatically well posed because it admits solution by assumption. The smaller the β the more irregular h_0 and also the solution. Similarly to the discussion on power spaces (section 2.1), for $\beta \to 0$ the smoothness requirements diminish as more and more functions are included in T^*T^β, limiting to L_2. Although $\beta > 0$, when solving a regularized problem with Tikhonov regularization, the benefit of having regularities larger than 2 is lost due to the saturation of the regularization (example 2.13).

[5] Frechet derivative

Theorem 3.6 ([36]Convergence rate)
Let $\hat{h}_\lambda$ be the solution to equation (3.32). If $\lambda = n^{-\frac{1}{2}max\left(\frac{2}{\beta+2},\frac{1}{2}\right)}$ then,

$$\left\|\hat{h}_\lambda - h\right\|_{\mathcal{H}} = O_p\left(n^{-min\left(\frac{\beta}{2\beta+4},\frac{1}{4}\right)}\right)$$

We give a concise proof of the result. This to show how the risk must be split into its empirical and population counterpart. Moreover, it is interesting to observe how the Tikhonov regularization saturation affects the rate of convergence.

Proof.

$$\left\|\hat{h}_\lambda - h_0\right\| \leq \underbrace{\left\|\hat{h}_\lambda - h_\lambda\right\|}_{A} + \underbrace{\|h_\lambda - h_0\|}_{B}$$

Let us focus on A. By Δ we have that:

$$\begin{aligned}
\hat{h}_\lambda - h_\lambda &= \hat{\Gamma}_\lambda \hat{T}^* \hat{g} - \Gamma_\lambda T^* g \\
&= \hat{\Gamma}_\lambda \hat{T}^* \left(\hat{g} - \hat{T}h_0\right) + \hat{\Gamma}_\lambda \hat{T}^* \hat{T} h_0 - \Gamma_\lambda T^* T h_0 \\
&= \hat{\Gamma}_\lambda \hat{T}^* \left(\hat{g} - \hat{T}h_0\right) + \hat{\Gamma}_\lambda \left(\hat{T}^*\hat{T} - T^*T\right) h_0 - \left(\Gamma_\lambda - \hat{\Gamma}_\lambda\right) T^* T h_0 \\
&= \hat{\Gamma}_\lambda \hat{T}^* \left(\hat{g} - \hat{T}h_0\right) + \hat{\Gamma}_\lambda \left(\hat{T}^*\hat{T} - T^*T\right) h_0 + \hat{\Gamma}_\lambda \left(\hat{T}^*\hat{T} - T^*T\right) \underbrace{\Gamma_\lambda T^* T h_0}_{=h_\lambda} \\
&= \hat{\Gamma}_\lambda \hat{T}^* \left(\hat{g} - \hat{T}h_0\right) + \hat{\Gamma}_\lambda \left(\hat{T}^*\hat{T} - T^*T\right) (h_0 - h_\lambda)
\end{aligned}$$

Thus it follows that:

$$A \leq \left\|\hat{\Gamma}_\lambda\right\| \left(\left\|\hat{T}^*\hat{g} - \hat{T}^*\hat{T}h_0\right\| + \left\|\hat{T}^*\hat{T} - T^*T\right\| \|h_0 - h_\lambda\|\right)$$

And then:

$$\left\|\hat{h}_\lambda - h_0\right\| \leq \left\|\hat{\Gamma}_\lambda\right\| \left\|\hat{T}^*\hat{g} - \hat{T}^*\hat{T}h_0\right\| + \left(\left\|\hat{\Gamma}_\lambda\right\| \left\|\hat{T}^*\hat{T} - T^*T\right\| + 1\right) \|h_0 - h_\lambda\|$$

Under proposition 3.11 of [9] and applying Bennett's inequality [36], we have:

$$\begin{aligned}
\left\|\hat{\Gamma}_\lambda\right\| &= O\left(\lambda^{-1}\right) \\
\left\|\hat{T}^*\hat{g} - \hat{T}^*\hat{T}h_0\right\| &= O(n^{-\frac{1}{2}}) \\
\left\|\hat{T}^*\hat{T} - T^*T\right\| &= O(n^{-\frac{1}{2}}) \\
\|h_\lambda - h_0\| &= O(\lambda^{\frac{1}{2}\min(\beta,2)})
\end{aligned}$$

For this to converge we must have that $\frac{\|h_\lambda - h_0\|}{\lambda n^{\frac{1}{2}}} \to 0$. Thus we -must impose that $\lambda n^{\frac{1}{2}} \geq \lambda^{\frac{1}{2}\min(\beta,2)}$. This is satisfied by choosing the regularization parameter $\lambda \geq n^{-\max\left(\frac{1}{\beta+2},\frac{1}{4}\right)}$. The optimal regularization parameter is found when it is equal to the RHS of the above inequality, which then gives an optimal rate (in λ) of:

$$n^{-\frac{1}{2}+\max\left(\frac{1}{\beta+2},\frac{1}{4}\right)} = n^{\max\left(-\frac{\beta}{2\beta+4},-\frac{1}{4}\right)} = n^{-\min\left(\frac{\beta}{2\beta+4},\frac{1}{4}\right)}$$

□

The best case scenario is the rate of $n^{-\frac{1}{4}}$ for the bridge function to converge, although the norm is stronger than the L_2 norm. Once the PMMR estimate for the outcome bridge function is found, it can be used to build the PMMR estimator for the CERF.

Definition 3.7 (PMMR Estimator)
Let $\mathcal{D}^t$ be a test set of unobserved samples of size n_t. Let $\hat{h}$ be the estimated solution from the previous procedure. The PMMR Estimator for the CERF is defined as:

$$\hat{\chi}(a) = \frac{1}{n_t} \sum_{\mathcal{D}^t} \hat{h}(w_i, a, x_i) = \left\langle \hat{h}, \hat{\mu}_{WX} \otimes \varphi(a) \right\rangle_{\mathcal{H}_{WX}}$$

where $\hat{\mu}_{WX}$ is an estimator for the marginal mean embedding $\hat{\mu}_{WX} = \frac{1}{n_t} \sum_{i \in \mathcal{D}^t} \varphi(w_i) \otimes \varphi(x_i)$

Since we are projecting against an estimated marginal mean embedding, there is a result similar to the KPV for the PMMR estimator.

Corollary 3.4 ([36])
Consider a sample at test time $\mathcal{D}^t$ of size n_t. Let $\hat{h}$ be the estimator of definition 3.7. Then $\forall a \in \mathcal{A}$, under the assumptions of theorem 3.6 and assumption 3.9 the PMMR estimator is such that:

$$|\hat{\chi}(a) - \chi(a)| \leq O_p\left(n_t^{-\frac{1}{2}} + n^{-\min\left(\frac{\beta}{2\beta+4}, \frac{1}{4}\right)}\right)$$

Proof.

$$\begin{aligned}
|\hat{\chi}(a) - \chi(a)| &= |\left\langle \hat{h}, \hat{\mu}_{WX} \otimes \varphi(a) \right\rangle - \langle h_0, \mu_{WX} \otimes \varphi(a) \rangle| \\
&= |\left\langle \hat{h}, \hat{\mu}_{WX} \otimes \varphi(a) \right\rangle - \langle h_0, \mu_{WX} \otimes \varphi(a) \rangle \pm \left\langle \hat{h}, \mu_{WX} \otimes \varphi(a) \right\rangle| \\
&= |\left\langle \hat{h}, \hat{\mu}_{WX} \otimes \varphi(a) \right\rangle - \langle h_0, \mu_{WX} \otimes \varphi(a) \rangle \pm \left\langle \hat{h}, \mu_{WX} \otimes \varphi(a) \right\rangle| \\
&= |\left\langle \hat{h}, \hat{\mu}_{WX} - \mu_{WX} \otimes \varphi(a) \right\rangle - \left\langle h_0 - \hat{h}, \mu_{WX} \otimes \varphi(a) \right\rangle| \\
&\leq \underbrace{\left\|\hat{h}\right\|_{\mathcal{H}} \|\varphi(a)\|_{\mathcal{H}} \|\hat{\mu}_{WX} - \mu_{WX}\|_{\mathcal{H}}}_{A} + \underbrace{\left\|\hat{h} - h_0\right\|_{\mathcal{H}} \|\varphi(a)\|_{\mathcal{H}} \|\mu_{WX}\|_{\mathcal{H}}}_{B}
\end{aligned}$$

A converges because $\hat{h} \in \mathcal{H}$ by construction, and the empirical mean embedding converges to the true mean embedding (proposition 2.5). The latter controls the rate of convergence with $O(n_t^{-\frac{1}{2}})$.
B converges because assumption 3.7 guarantees that $\|\mu_{WX}\|_{\mathcal{H}} < +\infty$ (lemma 2.2). □

The rate of convergence is once more dominated by the the rate of convergence of the method to the bridge function. Notice that, just like KPV, this point-wise norm follows from convergence in $\mathcal{H}$ and is stronger than a L_2 norm convergence result.

3.5 Flexible Approaches

The authors of [69, 30] develop methods similar to the ones presented in section 3.4 but use flexible parametric models to solve the minimization problems. The overall trend in applied mathematics is that flexible parametric models is a synonym of Neural Networks, in fact the previously mentioned papers are developed to use Neural Networks.

3.5.1 Neural Networks

In the past twenty years, Neural Networks have been at the center of attention for their flexibility, and possibly their interesting name. Neural networks are essentially a linear combination of many simple partially linear or non linear functions. For a fixed architecture, the weights of these linear combinations are the parameters of the model. A Neural Network with N layers and input x of dimension can be written as:

$$\sigma_N\left(\omega_N \cdot \sigma_{N-1}\left(\dots\left(\sigma_2\left(\omega_i \cdot x + b_1\right) + b_2\right)\dots\right) + b_{N-1}\right)$$

The functions σ_i are also known as activation functions and are responsible for adding the non-linearity to the system. ω_i and b_i are respectively the weights and bias used in the linear combination of the nonlinear terms.
The most commonly used activation functions are the ReLU $\sigma(x_i) = \max(x_i, 0)$, a *broken* linear function and the soft-max $\sigma(x)_i = \frac{e^{x_i}}{\sum_{i=1} e^{x_i}}$, a smooth exponential function. The idea behind this machinery is to repeatedly perform a linear combination of the previous layer, and subsequently apply the non-linear activation functions until the outcome is reached. Then, using a loss function that compares the predicted outcome to the true outcome of a test set, adjust the weights and biases to minimize the loss through back propagation. The hope is that after enough training has occurred, the optimal weights have been determined to capture the non-linearity of the underlying function.
A common issue with neural networks is achieving a local minima of the loss function. To overcome this, the most commonly used solution is to introduce regularization to penalize larger solutions, solutions with large weights.

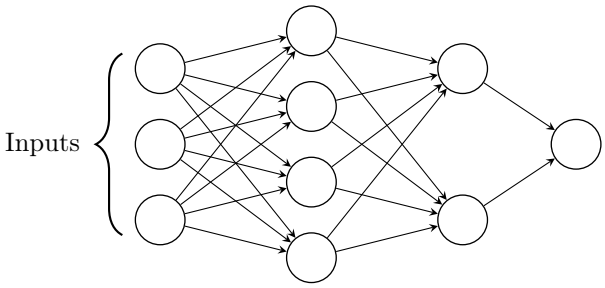


Figure 3.7: Graphical representation of Neural network with 3 inputs, 2 hidden layers and 1 output.

If neural networks are so flexible, why can one not be implemented to estimate the bridge function? The problem with this is that the loss is constructed on the observable data. In practice, neural networks usually minimize the mean squared error which is not the correct loss due to the presence of the unmeasured confounder U. The next methods adopt them to solve the problem of bridge functions.

3.5.2 Deep Feature Proxy Variable

The Deep Feature Proxy Variable approach presented in [69] is the flexible analogue of the KPV[36] method introduced earlier. The latter assumes the RKHS to be an infinite dimensional space associated with a fixed kernel. The canonical feature maps are then derived from the eigenvalues of the Kernel integral operator. Any function therein, in particular the conditional mean embedding, is uniquely identified by the weights of the infinite linear combination of aforementioned basis. The downside of this approach is that the having a fixed kernel also fixes the eigenfunctions and consequently the working Hilbert space $\mathcal{H}$. Alternatively, one could decide to learn adaptive basis functions from the data. To do this, we must consider a finite number of basis functions that can be parameterized by some flexible models to learn the optimal configurations in a data-dependent manner. In such a way, the space generated from the linear combination of the latter can in some way be seen as *optimal* within the class of parametric models.
For ease of exposition we omit the measured confounders X. We are interested in simultaneously learning the feature maps $\varphi_{N_{A_1}}, \varphi_{N_Z}, \varphi_{N_{A_2}}, \varphi_{N_W}$ and the optimal weights ω_h, ω_μ. Each feature map is parameterized as a Neural Net, and thus we are interested in learning the optimal weight and biases of the individual network, which we denote

$$N_{A_1}, N_{A_2}, N_Z, N_W$$

Similarly to the previously discussed methods, the data is split into two non overlapping subsets $\mathcal{D}_1, \mathcal{D}_2$ of dimensions n_1, n_2 respectively to ensure that no over-fitting biases are induced. Moreover, a $W|Z, A$ assumption ensures that the ill-posedness measure is finite and the risk ensures minimization in $\mathcal{H}$.

Stage 1

The regularized empirical risk of equation (3.26) now becomes the loss for the neural net:

$$\hat{\mathcal{L}}_1(\omega_\mu, N_{A_1}, N_Z) = \frac{1}{n_1} \sum_{i \in \mathcal{D}_1} \left\| \varphi_{N_W}(w_i) - \omega_\mu \left[\varphi_{N_{A_1}}(a_i) \otimes \varphi_{N_Z}(z_i) \right] \right\|^2 + \lambda_1 \left\| \omega_\mu \right\|^2 \tag{3.33}$$

This loss can be formulated in a more adversarial manner as:

$$\min_{\mathcal{G}} \hat{\mathcal{L}}_1(g) = \min_{\mathcal{G}} \frac{1}{n_1} \sum_{i \in \mathcal{D}_1} \left\| \varphi_{N_W}(w_i) - g(a_i, z_i) \right\|^2 + \lambda_1 \left\| \omega_\mu \right\|^2$$

where the adversarial is specified to be the set of all functions that can be constructed from the outer product of the two Neural Net features $\mathcal{G} = \{g : \omega \left(\varphi_{N_{A_1}}(a) \otimes \varphi_{N_Z}(z) \right)\}$. Similarly, let $\mathcal{H}_{N_W}$ be the space associated to the feature map φ_{N_W}. Both of these spaces are the spaces that can be reached as linear combinations of the Neural Network generated feature maps. The conditional mean embedding, even though it only acts as one since it is not associated to a fixed space, is estimated to be $\hat{\mu}_{W|z,a} = \hat{\omega}_\mu \left(\hat{\varphi}_{N_{A_1}}(a) \otimes \hat{\varphi}_{N_Z}(z) \right)$ where $\hat{\omega}_\mu$ are the optimum weights(matrix as discussed in KPV) and $\hat{\varphi}_{N_{A_1}}, \hat{\varphi}_{N_Z}$ are the trained neural nets.

Assumption 3.8 (DFPV Assumptions)
The outcome variables is assumed to be bounded, i.e. $|Y| \leq M$. *The adversarial function class* $\mathcal{G}$ *is*

assumed to contain only bounded functions in both a, z, i.e. $\|g(a,z)\|_\infty \leq 1 \quad \forall g \in \mathcal{G}$. The feature map φ_{N_W} is also assumed to be bounded and $\forall a \in \mathcal{A} \forall v : \|v\| \leq 1 \quad \omega_\mu \left[\varphi_{N_{A_1}} \otimes v\right] \leq 1$

Lemma 3.3 ([69] Stage 1 Consistency)
Let $\hat{\mu}_{W|Z,A}$ be the solution to the stage one minimization procedure. Under assumption 3.8 and given $\mathcal{D}_1$ of size n_1 then $\forall f \in \mathcal{H}_{N_W} \otimes \mathcal{H}_{N_{A_1}}$, $\delta > 0$, with at least probability $1 - 2\delta$ we have:

$$\left\|\left\langle f, \hat{\mu}_{W|Z,A}\right\rangle - \mathbb{E}\left[f(W,A)|Z,A\right]\right\|^2_{L_2(A,Z)} \leq M \sqrt{\left(\kappa_1 + 4\hat{\mathcal{R}}(\mathcal{H}_1) + 24\sqrt{\frac{\log 2/\delta}{2n_1}}\right)}$$

where $\mathcal{H}_1 = \left\{\mathcal{W} \times \mathcal{A} \times \mathcal{Z} \mapsto \|\varphi_{N_W}(w) - g(a,z)\|^2 \quad \forall g \in \mathcal{G} \; \forall \varphi_{N_W} \in \mathcal{H}_{N_W}\right\}$, $\hat{\mathcal{R}}$ is the empirical Rademacher complexity of $\mathcal{H}_1$ on $\mathcal{D}_1$, and $\kappa_1 = \max_{\varphi_{N_W}} \min_{\mathcal{G}} \mathbb{E}\left[\|\varphi_{N_W}(W) - g(Z,A)\|^2\right]$.

It is not very clear what the space of functions $\mathcal{H}_1$ is, this makes it even more difficult to decide if the Radamacher complexities indeed vanish. Moreover, notice that it is not possible to optimize the feature map φ_{N_W} since there is no observable loss that can inform us of its performance.

Stage 2

The second stage stage loss is identical to the one of the KPV method from equation (3.28) with the fixed feature maps replaced by the neural network parameterized feature maps.

$$\hat{\mathcal{L}}_2(\omega_h, N_{A_2}, N_W) = \frac{1}{n_2} \sum_{i \in \mathcal{D}_2} \left(y_i - \omega_h \left(\varphi_{N_{A_2}}(a_i) \otimes \hat{\mu}_{W|z_i,a_i}\right)\right)^2 + \lambda_2 \|\omega_h\|^2 \tag{3.34}$$

Plugging in the optimizers of the above loss, the estimated bridge function is then constructed as:

$$\hat{h}(a,w) = \hat{\omega}_h \left(\hat{\varphi}_{N_{A_1}}(a) \otimes \hat{\varphi}_{N_W}(w)\right) \tag{3.35}$$

Using this in combination with the minimizer determined in the first stage, the entire method is shown to be consistent if the Rademacher complexities vanish.

Lemma 3.4 ([69] Bridge function consistency)
Let $\mathcal{D}_1, \mathcal{D}_2$ be training sets of size n_1, n_2 respectively. Let $\hat{\mu}_{W|ZA}$ be the solution first stage on $\mathcal{D}_1$ and $\hat{h}$ be the DFPV bridge estimate of equation (3.35) constructed using the weights of stage 2 on $\mathcal{D}_2$. Let κ_1 be as in lemma 3.3 and $\kappa_2 = \min_{\mathcal{H}} \|\mathbb{E}\left[h(W,A)|Z,A\right] - \mathbb{E}\left[h_0(W,A)|Z,A\right]\|_{L_2(Z,A)}$. Under assumption 3.8 for any $\delta > 0$, with at least probability $1 - 2\delta$ we have:

$$\left\|\left\langle \hat{\mu}_{W|ZA}, \hat{h}\right\rangle - \mathbb{E}[h_0(W,A)|Z,A]\right\|_{L_2(A,Z)} \leq \kappa_2 + \sqrt{M\left(\kappa_1 + 4\hat{\mathcal{R}}(\mathcal{H}_1) + 24\sqrt{\tfrac{\log 2/\delta}{2n_1}}\right)} + \sqrt{\left(4\hat{\mathcal{R}}(\mathcal{H}_2) + 24M^2\sqrt{\tfrac{\log 2/\delta}{2n_2}}\right)}$$

where $\mathcal{H}_1$ is as in lemma 3.3, $\mathcal{H}_2 = \left\{\mathcal{Y} \times \mathcal{A} \times \mathcal{Z} \mapsto (y - \omega(\varphi_{N_{A_2}}(a) \otimes g(a,z)))^2 \quad \forall g \in \mathcal{G}, \; \forall \omega\right\}$ and $\hat{\mathcal{R}}$ is the empirical Rademacher complexity of the respective hypothesis classes.

The dependency of the two losses on the weights ω_h, ω_μ makes it difficult to use these losses to train the hyper parameters of the neural networks with gradient descent techniques. Remember from section 3.4.1 that the minimization problem above has a closed form solution if the feature were known. At each iteration, the optimal weights are determined for given feature maps. As such, the losses can be reformulated only in terms of the network parameters $N_{A_1}, N_{A_2}, N_Z, N_W$ enabling us to train the networks.

Assumption 3.9 ([69])
Let $f_W(w), f_{W|A}|(w|a)$ be the density distributions of P_W and $P_{W|A}$. Assume that:

$$\eta_a = \left\| \frac{f_W(W)}{f_{W|A}(W|a)} \right\|^2_{P(W|A=a)} < +\infty$$

Under this additional assumption, if the Rademacher complexities of the function classes vanish, it can be shown that the DFPV estimator is consistent.

Definition 3.8 (DFPV Estimator)
Let $\mathcal{D}^t$ be a test set of unobserved samples of size n_t. Let $\hat{h}$ be the bridge function found through the previous two stage estimation procedure. The DFPV Estimator is defined as:

$$\hat{\chi}(a) = \frac{1}{n_t} \sum_{i \in \mathcal{D}^t} \hat{h}(w_i, a)$$

Theorem 3.7 ([69] Causal Consistency)
Let $\hat{\chi}$ be the DFPV estimator, and $\eta = sup_{a \in \mathcal{A}} \eta_a$ from assumption 3.9, then with probability at least $1 - 7\delta$:

$$\|\hat{\chi} - \chi\|_{L_2(A)} \leq \sqrt{\frac{2 \log 2/\delta}{n_t}} + \eta \tau B_1$$

where B_1 is the bound from lemma 3.4 and τ is the measure of ill-posedness.

3.5.3 Neural Maximum Moment Restriction

NMMR is a one stage approach that can be seen as the neural network equivalent of the PMMR. Rather than determining h within a RKHS, the function is parameterized as a neural network and the risk in equation (3.27) is used as a loss to train it. Additionally, the authors consider an unbiased estimator of equation (3.27) in the U-statistic:

$$\hat{R}_U(h) = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} (y_i - h(V_{h_i})(y_j - h(V_{h_j}) k(V_{h_i}, V_{h_j}) \tag{3.36}$$

The difference with the V-statistic previously used is that the when $i = j$ the terms are not considered thus removing the bias. Moreover, this is the minimum variance unbiased estimator of the risk [30]. As we shall also notice in our numerical experiments in chapter 6, the performance of the U-statistic is worse than when employing the V-statistic.
The empirical loss function now introduces $\Lambda : \mathcal{H} \times \Theta_h \to [0, M_\lambda]$, a regularizer that penalizes the network size, the size of the weights and biases:

$$\hat{\mathcal{L}} = (Y - h(V_h))^T K (Y - h(V_h)) + \Lambda(h, \theta_h) \tag{3.37}$$

$K_{ij} = k(V_{h_i}, V_{h_j})$ is the kernel matrix and the diagonal is included or excluded depending on whether the U or V statistic risk is used. Similarly to the DFPV method, the authors forego any source or regularity assumptions and directly assume *that the procedure is sufficient to recover the bridge function.* Moreover it is left in vague term depending on the Rademacher complexities.

Theorem 3.8 ([30]NMMR consistency)
Let $\tilde{h}_k$ be the minimizer of the regularized population risk $R_k(h)$ and $\hat{h}_{k,U,\lambda,n}$ be the minimizer of the empirical regularized risk $\hat{R}_{k,U,\lambda,n}(h)$ for $h \in \mathcal{H}$. Let k kernel associated with $\mathcal{G} \subset L_2(Z, A, X)$ bounded by M_k, and let h_0 be the true bridge function. Also let,

$$\begin{aligned} d_k^2(h, h') = \mathbb{E}\big[&(h(V_h) - h'(V_h))(h(V_h') - h'(V_h')) \\ &\times k\left((V_q), (V_q')\right)\big] \end{aligned}$$

Then, $d_k^2(h_0, h) = R_k(h)$ and, with probability at least $1 - \delta$,

$$\begin{aligned} d_k^2\left(h_0, \hat{h}_{k,U,\lambda,n}\right) \leq\ & d_k^2\left(h_0, \tilde{h}_k\right) + \lambda M_\lambda + 8M\left(\mathcal{R}_{n-1}(\mathcal{F}) + \mathcal{R}_n(\mathcal{F})\right) \\ &+ 16M^2 M_k \left(\frac{2}{n} \log \frac{2}{\delta}\right)^{\frac{1}{2}} + 10\left(2 \log 2\right)^{\frac{1}{2}} M^2 M_k n^{-\frac{1}{2}} \end{aligned}$$

Further, if k is Integral Strictly Positive Definite, then d_k is a metric on $L^2_{\mathcal{A}\mathcal{X}\mathcal{Z}}$ and, if the right hand side of the inequality goes to zero as n goes to infinity
$d_k\left(\mathbb{E}[h_0|A, X, Z] - \mathbb{E}\left[\hat{h}_{k,\lambda,n}\middle|A, X, Z\right]\right) \xrightarrow{\mathrm{P}} 0$ *so* $\mathbb{E}\left[\hat{h}_{k,\lambda,n}\middle|A, X, Z\right] \xrightarrow{\mathrm{P}} \mathbb{E}[h_0|A, X, Z]$ *in* d_k.

$$\mathcal{F}' = \left\{ f \mid \exists h \in \mathcal{H}, \exists V_q \in \mathcal{Z} \times \mathcal{A} \times \mathcal{X}, \forall V_q \in \mathcal{Z} \times \mathcal{A} \times \mathcal{X} :\ f(w', V_q') = h(V_h')\, k\left((V_q'), (V_q)\right) \right\}$$

The proof is long and not particularly insightful and is therefore omitted completely. The result mainly follows from concentration inequalities and symmetrization. An almost identical result can be shown when using the regularized V-statistic risk.
This bound has a lot of terms to consider. Foregoing the regularization term λ and the Rademacher complexities, the controlling term is $\sqrt{n}$. Thus, the best rate achievable by the method is $O(n^{-\frac{1}{4}})$. Moreover, the theorem is stated with the vague requirement of 'if the RHS of the inequality goes to zero'. Since the first term $d_k^2\left(h_0, \tilde{h}_k\right)$ must also vanish, the authors suggest increasing the network's complexity with the number(in practice increasing the number of layers and nodes). This has to be done slowly enough that the Rademacher complexities do not increase. Moreover, the authors do not state any convergence results for the Rademacher complexities, a discussion will follow at the end of the chapter, but state that they expect these terms to vanish with sample size. A similar discussion is held for the regularization coefficient which is expected to decay to send the bias term to zero without compromising the convergence of the complexities. This discussion is purely qualitative as there are no proven results.
After training, the NMMR estimator is constructed on a held out dataset as the sample average of the outcome of the network.

Definition 3.9 (NMMR Estmimator)
Let $\mathcal{D}^t$ be a test set of unobserved samples of size n_t. Let $\hat{h}$ be the neural network trained on $\mathcal{D}^n$. The NMMR Estimator is constructed as:

$$\hat{\chi}(a) = \frac{1}{n_t} \sum_{\mathcal{D}^t} \hat{h}(w_i, a, x_i)$$

Although the authors of the paper do not prove consistency of the estimator so constructed, it can be shown in a similar manner as the DFPV estimator.

Note

While both neural network based methods prove *theoretical* guarantees for convergence of the bridge functions, and in practice represent the state of the art, the assumption that the Rademacher complexities of the relative function classes vanish is in no way justified. Although there exist certain results on the complexities of fixed networks vanishing at a rate $\sqrt{n}$, they are for simple architectures and losses. Moreover, the authors want various terms to go to zero whilst increasing network complexity[30] which complicates the problem even further. The main difficulty in determining rates for these Rademacher complexities is also the structure of the function classes considered, especially in the outer product of feature maps of two separate neural nets $\varphi_{N_{A_1}}(a) \otimes \varphi_{N_Z}(z)$. The authors of both papers recognize this gap in the literature [68, 31].
Moreover, the methods and the associated algorithms do not guarantee that the optimization procedures recover the correct parameters. Rather, both methods assume that it is indeed sufficient to do so. On the other hand neural networks are more salable than matrix inversion since the training occurs in smaller batches of the entire dataset. Overall, the theoretical guarantees are not as *strong* as the other methods previously introduced. Nonetheless, this does not discredit their effectiveness in the practical setting. Both PMMR and KPV quickly become computationally impractical due to the need to invert large matrices. Differently, their neural network approaches are much more scalable due to the batch learning aspect of Neural Nets.

A discussion on two stage methods

In [38] the authors of discuss the general problems with two stage regressions. It is no wonder that in econometrics, the two stage regression is often referred to as the *forbidden regression* when assumption 3.1 does not hold[66, 38]. The first critique raised is the incompatibility with a general principle by Vapnik[61]:

> *When solving a problem of interest, do not solve a more general problem as an intermediate step.*

With the two stage approach, we attempt to learn the more general dependency of $W|Z, A$ to then solve for the true dependency of $Y|Z, A$. The second problem is that while both stage regression are asymptotically valid, the second stage requires unbiased estimators of the conditional expectation. By determining the coefficients or parameters of the first regression, thus effectively *stopping* it at a finite number of data points, the estimator might present some residual bias. This is particularly true in problems where the amount of data is relatively small. On the other hand, two stage methods have the advantage that they can use finite amount of data more efficiently. This is because the identifying

the conditional expectation, or the coefficients in the first stage of P2SLS, only requires (W, Z, A, X). Similarly, the second stage only requires (Y, Z, A, X). Thus, the partitions $\mathcal{D}_1, \mathcal{D}_2$ do not have to contain all observation for the two stage methods to be successfully applied. If one has an incomplete dataset with missing data, assuming the data is missing at random, such approach enables use of all available data. It is important to note that the data must be missing at random, otherwise the violation of i.i.d. occurs and the learnt conditional distribution is not valid for the second stage.
When observing the numerical results (chapter 6), the one stage approaches seem to fare much better than their two stage counterparts.

4

Linear Extensions

In this chapter, we introduce some new approaches to estimate the CERF. The first approach introduces higher order polynomial terms in the two stage regression of section 3.1.1. This enables correct estimation of causal parameters of structural models with higher order treatment and higher order confounding as long as all the variables, except the outcome, are Gaussian. The second approach, ProxySplines, is purely practical. It is born from the need to estimate the the range of the best treatment value rather than the exact CERF.

4.1 Errors of P2SLS

We are interested in determining standard errors of the P2SLS method. When applying the P2SLS method, the common consensus is to use bootstrap to estimate the standard errors [53]. We give a possibly new characterization of the standard errors of the simple P2SLS regression by combining the coefficient adjustment, sample splitting and the Delta Method.

Theorem 4.1 (Delta Method)
If there exists $\hat{\underline{B}}_n$ a sequence of random variable such that:

$$\sqrt{n}\left(\hat{\underline{B}}_n - \underline{B_0}\right) \xrightarrow{D} \mathcal{N}(0, \Sigma)$$

then, for any continuously differentiable function f we have that:

$$\sqrt{n}\left(f(\hat{\underline{B}}_n) - f(\underline{B_0})\right) \xrightarrow{D} \mathcal{N}(0, \nabla f(\underline{B_0})^T \Sigma \nabla f(\underline{B_0}))$$

Under assumption 3.1, the OLS solutions of the coefficients of linear regression are sample means, and by the central limit theorem the asymptotic distribution of the regression parameters is a centered Gaussian. Remember now that the *true* treatment coefficient β_A of assumption 3.1 can be found with the correction method as:

$$\beta_A = \theta_{AY} - \frac{\gamma_{AW}}{\gamma_{ZW}}\theta_{ZY} \tag{4.1}$$

Setting $\underline{B_0}$ equal to the true structural coefficients of the parametric model in assumption 3.1, and $\hat{B}$ to the estimated regression coefficients, we can apply the delta method to derive standard errors for the CERF coefficients. The main issue is that since we are performing two separate regressions, no *complete* covariance matrix is available to us. Nonetheless, under the assumption of i.i.d. samples and using sample splitting to perform each OLS regression we are able to guarantee that the coefficients are independent. Like this the *complete* covariance matrix can be built as the block diagonal matrix $\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}$ with Σ_1, Σ_2 being the covariance of the first and second stage regressions.

Theorem 4.2 (Standard Errors for P2SLS)
Under assumption 3.1 the P2SLS estimator for the ATE is such that:

$$\sqrt{n}\left(\hat{\beta}_A - \beta_A\right) \xrightarrow{D} \mathcal{N}(0, \tilde{\sigma})$$

where $\tilde{\sigma} = \nabla\left(\theta_{AY} - \frac{\gamma_{AW}}{\gamma_{ZW}}\theta_{ZY}\right)^T \Sigma \nabla\left(\theta_{AY} - \frac{\gamma_{AW}}{\gamma_{ZW}}\theta_{ZY}\right)$ and Σ is the complete covariance matrix previously defined.

Proof. We are interested in applying theorem 4.1 to the function $f(\underline{B_0}) = \theta_{AY} - \frac{\gamma_{AW}}{\gamma_{ZW}}\theta_{ZY}$. As already discussed, the OLS regressions converge do their true parameters by the CLT. Taking the partial derivatives of f we obtain:

$$\frac{\partial}{\partial\theta_{AY}}f(\underline{B_0}) = 1 \qquad \frac{\partial}{\partial\theta_{ZY}}f(\underline{B_0}) = -\frac{\gamma_{AW}}{\gamma_{ZW}} \qquad \frac{\partial}{\partial\gamma_{AW}}f(\underline{B_0}) = -\frac{\theta_{ZY}}{\gamma_{ZW}} \qquad \frac{\partial}{\partial\gamma_{ZW}}f(\underline{B_0}) = -\frac{\theta_{ZY}\gamma_{AW}}{\gamma_{ZW}^2}$$

$\nabla f(B_0) = \left[1, -\frac{\gamma_{AW}}{\gamma_{ZW}}, \frac{\theta_{ZY}}{\gamma_{ZW}}, -\frac{\theta_{ZY}\gamma_{AW}}{\gamma_{ZW}^2}\right]$ As previously discussed, using sample splitting we obtain that the coefficients of the two regressions are independent and the covariance matrix is the block matrix:

$$\Sigma = \begin{bmatrix} \sigma_{\theta_A} & \rho_\theta & 0 & 0 \\ \rho_\theta & \sigma_{\theta_Z} & 0 & 0 \\ 0 & 0 & \sigma_{\gamma_A} & \rho_\gamma \\ 0 & 0 & \rho_\gamma & \sigma_{\gamma_Z} \end{bmatrix}$$

Following the computation in lemma A.1

$$\sigma^2 = \sigma_{\theta_A}^2 - 2\frac{\gamma_{AW}}{\gamma_{ZW}}\rho_\theta + (\frac{\gamma_{AW}}{\gamma_{ZW}})^2\sigma_{\theta_Z}^2 + (\frac{\theta_{ZY}}{\gamma_{ZW}})^2\sigma_{\gamma_A} - 2\frac{\theta_{ZY}^2\gamma_{AW}}{\gamma_{ZW}^3}\rho_\gamma + (\frac{\theta_{ZY}\gamma_{AW}}{\gamma_{ZW}^2})^2\sigma_{\gamma_Z} \tag{4.2}$$

□

This result suggests that we can estimate the standard errors of the P2SLS estimator by replacing the covariance matrix with the sample covariance matrix and sample coefficients. As guaranteed by assumption 3.1, the coefficient γ_{ZW} and its estimate must be bounded away from zero. As soon as this coefficient nears zero, the estimates will deteriorate and the variance even more so due to the higher order term in the denominator.

4.2 Higher Order Proximal 2SLS

The proximal two stage least squares approach discussed in section 3.1.1 might be too simple to capture the true CERF in real-world applications since rarely responses of natural processes are linear. For example, in the agricultural setting, fertilizers tend to increase crop yield but can also deteriorate soil quality by influencing various quantities such as ph-level and and soil microbiome [70, 24]. Thus, after a certain threshold the fertilizer starts to have a negative effect on crop yield. This might suggest an upside parabolic CERF for quantity of fertilizer(treatment) and crop yield(outcome). To account for such situations, we introduce Higher Order Proximal 2SLS. This method introduces higher order estimation terms in the regression equations to account for possible non-linearities both in treatment and confounding. Note that this can be seen as a kernel regression with polynomial type kernel. Although some works implement 'quadratic feature' P2SLS in their numerical experiments section, it is unclear if this is what they do.

Higher Order Treatment with Linear Confounding

Rather than requiring assumption 3.1, which constrains the models to be completely linear, suppose that the effect of treatment on outcome is of higher order up to K, while both the *first stage* and $U|ZA$ remain linear:

$$\begin{aligned} Y &= \beta_{0Y} + \sum_1^K \beta_{AY_i}A^i + \beta_{UY_i}U + \epsilon_Y \\ W &= \beta_{0W} + \beta_{UW}U + \epsilon_W \qquad\qquad \epsilon_\cdot \sim \mathcal{N}(0,\sigma^2) \text{ independent of } Y, A, W, Z, U \\ U &= \nu_{0U} + \nu_{AU}A + \nu_{ZU}Z + \epsilon_U \end{aligned}$$

Then, the first stage coefficients can easily be estimated with ordinary least squares just as in the regular P2SLS by fitting the following first stage regression on the observed A, Z:

$$\begin{aligned} W &\sim \beta_{0W} + \beta_{UW}(\nu_{0U} + \nu_{AU}A + \nu_{ZU}Z + \epsilon_U) \\ \mathbb{E}[W|Z,A] &= \gamma_{0W} + \gamma_{AW}A + \gamma_{ZW}Z \end{aligned}$$

Using the *correction* framework introduced in section 3.1.1, this shows exactly what underlying parameters each δ coefficient is estimating

$$\gamma_{0W} = \beta_{0W} + \beta_{UW}\nu_{0U} \quad \gamma_{AW} = \beta_{UW}\nu_{AU} \quad \gamma_{ZW} = \beta_{UW}\nu_{ZU} \tag{4.3}$$

Just as before, we fit an OLS regression containing the higher order terms for A in the second stage:

$$Y = \beta_{0Y} + \sum_{1}^{K} \beta_{AY_i} A^i + \beta_{UY}(\nu_{0U} + \nu_{AU}A + \nu_{ZU}Z + \epsilon_U) + \epsilon_Y$$
$$\mathbb{E}\left[Y|ZA\right] = \theta_0 + \sum_{1}^{K} \theta_{A_i} A^i + \theta_{ZY} Z$$

Since the unmeasured confounder appears only at the first order, it only has a first order bias of the treatment on the outcome, in particular the coefficients identify:

$$\theta_0 = \beta_{0Y} + \beta_{UY}\nu_{0U} \qquad \theta_{AY_1} = \beta_{AY_1} + \beta_{UY}\nu_{AU} \qquad \theta_{ZY} = \beta_{UY}\nu_{ZU} \quad \theta_{A_i} = \beta_{AY_i} \quad \forall i \in \{2, \dots K\} \tag{4.4}$$

Similarly to the first order linear case, the coefficient β_{UY} can be recovered by adjusting the coefficients of equation (4.4). Then taking the ratio like before we can recover the true parameter of interest $\beta_{AY_1} = \theta_{AY_1} - \frac{\theta_{ZY}}{\gamma_{ZW}}\gamma_{AW}$. From equation (4.4) it can be noticed that of all recovered coefficients for A, the first order is the only one biased by U. Thus, we wonder what would happen if the measured confounding in U is of higher order.

Higher Order Treatment and Confounding

Now we want to include higher order terms for both treatment and unmeasured confounder U. Suppose that the effect of treatment on outcome is of order up to K, whereas of the unmeasured confounder U is also up to order K. The assumption that $U|ZA$ is linear remains the same:

$$Y = \beta_{0Y} + \sum_{1}^{K} \beta_{AY_i} A^i + \sum^{K} \beta_{UY_i} U^i + \epsilon_Y$$
$$W = \beta_{0W} + \beta_{UW}U + \epsilon_W \qquad \epsilon_\cdot \sim \mathcal{N}(0, \sigma^2) \text{ independent of } Y, A, W, Z, U$$
$$U = \nu_{0U} + \nu_{AU}A + \nu_{ZU}Z + \epsilon_U$$

The CERF is given to us by the adjustment formula and remains a polynomial of order K:

$$\begin{aligned}
\mathbb{E}\left[Y^a\right] &= \mathbb{E}\left[\mathbb{E}\left[Y|A=a,U\right]\right] \\
&= \mathbb{E}\left[\beta_{0Y} + \sum_{1}^{K} \beta_{AY_i} a^i + \sum^{K} \beta_{UY_i} U^i\right] \\
&= \beta_{0Y} + \sum_{1}^{K} \beta_{AY_i} a^i + \sum^{K} \beta_{UY_i} \mathbb{E}\left[U^i\right] \\
&= \tilde{\beta_{0Y}} + \sum_{1}^{K} \beta_{AY_i} a^i
\end{aligned}$$

The first stage regression $W|Z, A$ does not change and returns coefficients that have identifications given in equation (4.3). To now include all the unmeasured confounding in the second stage, the higher order terms and their cross products the since binomial expansion of each $U^k = (\nu_{0U} + \nu_{AU}A + \nu_{ZU}Z + \epsilon)^k$ can be written as:

$$U^k = \sum_{j}^{k} \binom{k}{j} (\nu_{0U} + \epsilon)^{k-j} (\nu_{AU}A + \nu_{ZU}Z)^j$$

Noticing that this is a '*symmetric*' problem for the terms $\nu_{AU}^i A^i$ and $\nu_{ZU}^i Z^i$ so they will have the same coefficients $\binom{k}{i} (\nu_{0U} + \epsilon)^{k-i}$. This means that taking the ratio between them will cancel out the terms and result in being equal to the usual adjustment ratio found in the first stage regression: $\left(\frac{\nu_{AU}}{\nu_{ZU}}\right)^i = \left(\frac{\gamma_{AW}}{\gamma_{ZW}}\right)^i$. Thus the bias term of each coefficient can be adjusted by using this ratio.

Example 4.1
For second order expansion:

$$U^2 = \nu_{AU}^2 A^2 + 2\nu_{AU}A\nu_{0U} + 2\nu_{AU}A\epsilon_U + 2\nu_{AU}A\nu_{ZU}Z + \nu_{0U}^2 + 2\nu_{0U}\epsilon_U + \nu_{ZU}^2 Z^2 + 2\nu_{ZU}\nu_{0U}Z + 2\nu_{ZU}Z\epsilon_U + \epsilon_U^2$$

which, in conditional expectation would become:

$$\begin{aligned}\mathbb{E}\left[U^2|ZA\right] &= \nu_{AU}^2 A^2 + 2\nu_{AU}A\nu_{0U} + 2\nu_{AU}A\mathbb{E}\left[\epsilon_U\right] + 2\nu_{AU}A\nu_{ZU}Z \\ &\quad + \nu_{0U}^2 + 2\nu_{0U}\mathbb{E}\left[\epsilon_U\right] + \nu_{ZU}^2 Z^2 + 2\nu_{ZU}\nu_{0U}Z + 2\nu_{ZU}Z\mathbb{E}\left[\epsilon_U\right] + \mathbb{E}\left[\epsilon_U^2\right] \\ &= \nu_{AU}^2 A^2 + 2\nu_{AU}A\nu_{0U} + 2\nu_{AU}A\nu_{ZU}Z + \nu_{0U}^2 + \nu_{ZU}^2 Z^2 + 2\nu_{ZU}\nu_{0U}Z + \sigma^2\end{aligned}$$

Thus, if the effect of the unmeasured confounder is of order up to two, one would fit the following regression:

$$\mathbb{E}\left[Y|ZA\right] = \theta_0 + \theta_{AY_1}A + \theta_{AY_2}A^2 + \theta_{ZY_1}Z + \theta_{ZY_2}Z^2 + \theta_{A_1Z_1}AZ$$

The coefficients will have the following bias:

$$\begin{aligned}\theta_0 &= \beta_{0Y} + \beta_{UY_1}\nu_{0U} + \beta_{UY_2}\left(\nu_{0U}^2 + \sigma^2\right) \\ \theta_{AY_1} &= \beta_{AY_1} + \beta_{UY_1}\gamma_A + 2\beta_{UY_2}\gamma_A\nu_{0U} = \beta_{AY_1} + \nu_{AU}\left(\beta_{UY_1} + 2\beta_{UY_2}\nu_{0U}\right) && (4.5) \\ \theta_{AY_2} &= \beta_{AY_2} + \beta_{UY_2}\left(\gamma_A^2\right) && (4.6) \\ \theta_{ZY_1} &= \beta_{UY_1}\gamma_Z + 2\beta_{UY_2}\gamma_Z\nu_{0U} = \nu_{ZU}\left(\beta_{UY_1} + 2\beta_{UY_2}\nu_{0U}\right) && (4.7) \\ \theta_{ZY_2} &= \beta_{UY_2}\left(\gamma_Z^2\right) && (4.8)\end{aligned}$$

The first stage regression remains the same and thus, the coefficients remain the same as in the classic P2SLS. This implies that the ratio $\frac{\gamma_{AW}}{\gamma_{ZW}} = \frac{\nu_{AU}}{\nu_{ZU}}$. *For the first order bias, we notice from equation* (4.7) *that* $\frac{\theta_{ZY_1}}{\nu_{ZU}} = (\beta_{UY_1} + 2\beta_{UY_2}\nu_{0U})$. *Replacing this into equation* (4.5)

$$\theta_{AY_1} = \beta_{AY_1} + \frac{\nu_{AU}}{\nu_{ZU}}\theta_{Z1} = \beta_{AY_1} + \frac{\gamma_{AW}}{\gamma_{ZW}}\theta_{Z1}$$

To remove the second order bias we do the same thing:

$$\theta_{AY_2} = \beta_{AY_2} + \frac{\theta_{ZY_2}}{\gamma_Z^2}\gamma_A^2 = \beta_{AY_2} + \left(\frac{\gamma_{AW}}{\gamma_{ZW}}\right)^2\theta_{ZY_2}$$

Notice, the fact that the variance from ϵ_U *deos not bother us and will be included in the estimate of the intercept, which is not of primary interest to us.*

Example 4.2
For third order expansion:

$$\begin{aligned}U^3 =& \nu_{AU}^3 A^3 + 3\nu_{AU}^2 A^2\nu_{0U} + 3\nu_{AU}^2 A^2\epsilon_U + 3\nu_{AU}^2 A^2\nu_{ZU}Z + 3\nu_{AU}A\nu_{0U}^2 \\ &+ 6\nu_{AU}A\nu_{0U}\epsilon_U + 3\nu_{AU}A\epsilon_U^2 + 3\nu_{AU}A\nu_{ZU}^2 Z^2 + 6\nu_{AU}A\nu_{ZU}\nu_{0U}Z \\ &+ 6\nu_{AU}A\nu_{ZU}Z\epsilon_U + \nu_{0U}^3 + 3\nu_{0U}^2\epsilon_U + 3\nu_{0U}\epsilon_U^2 + \nu_{ZU}^3 Z^3 \\ &+ 3\nu_{ZU}^2\nu_{0U}Z^2 + 3\nu_{ZU}^2 Z^2\epsilon_U + 3\nu_{ZU}\nu_{0U}^2 Z + 6\nu_{ZU}\nu_{0U}Z\epsilon_U + 3\nu_{ZU}Z\epsilon_U^2 + \epsilon_U^3\end{aligned}$$

Denote the third moment of the error as $\mathbb{E}\left[\epsilon_U^3\right] = \kappa$ *In expectation:*

$$\begin{aligned}\mathbb{E}\left[U^3|Z,A\right] =& \nu_{AU}^3 A^3 + 3\nu_{AU}^2 A^2\nu_{0U} + +3\nu_{AU}A\nu_{0U}^2 + 3\nu_{AU}A\sigma^2 + \\ &+ 6\nu_{AU}A\nu_{ZU}\nu_{0U}Z + 3\nu_{AU}A\nu_{ZU}^2 Z^2 + 3\nu_{AU}^2 A^2\nu_{ZU}Z \\ &+ \nu_{ZU}^3 Z^3 + 3\nu_{ZU}^2\nu_{0U}Z^2 + 3\nu_{ZU}\nu_{0U}^2 Z + 3\nu_{ZU}Z\sigma^2 \\ &+ \nu_{0U}^3 + 3\nu_{0U}\sigma^2 + \kappa\end{aligned}$$

In this case one would fit the second stage regression:

$$\mathbb{E}\left[Y|ZA\right] = \theta_0 + \theta_{AY_1}A + \theta_{AY_2}A^2 + \theta_{AY_3}A^3 + \theta_{ZY_1}Z + \theta_{ZY_2}Z^2 + \theta_{ZY_3}Z^3 + + \theta_{A_1Z_1}AZ + + \theta_{A_2Z_1}A^2Z + + \theta_{A_1Z_2}AZ^2$$

The bias generated from the higher order terms will be the previously found plus the new terms:

$$\begin{aligned}
\theta_0 &= \beta_{0Y} + \beta_{UY_1}\nu_{0U} + \beta_{UY_2}\left(\nu_{0U}^2 + \sigma^2\right) + \beta_{UY_3}\left(\nu_{0U}^3 + 3\nu_{0U}\sigma^2 + \kappa\right) \\
\theta_{AY_1} &= \beta_{AY_1} + \beta_{UY_1}\gamma_A + \beta_{UY_2}\gamma_A\nu_{0U} + \beta_{UY_3}(3\gamma_A\nu_{0U}^2 \\
&\quad + 3\nu_{AU}\sigma^2) = \beta_{AY_1} + \nu_{AU}\left(\beta_{UY_1} + 2\beta_{UY_2}\nu_{0U} + 3\beta_{UY_3}(\nu_{0U}^2 + \sigma^2)\right) \qquad (4.9) \\
\theta_{ZY_1} &= \beta_{UY_1}\gamma_Z + \beta_{UY_2}\gamma_Z\nu_{0U} + \beta_{UY_3}(3\gamma_Z\nu_{0U}^2 + 3\nu_{ZU}\sigma^2) \\
&= \nu_{ZU}\left(\beta_{UY_1} + 2\beta_{UY_2}\nu_{0U} + 3\beta_{UY_3}(\nu_{0U}^2 + \sigma^2)\right) \qquad (4.10) \\
\theta_{AY_2} &= \beta_{AY_2} + \beta_{UY_2}\gamma_A^2 + 3\beta_{UY_3}\nu_{AU}^2\nu_{0U} = \beta_{AY_2} + \nu_{AU}^2\left(\beta_{UY_2} + 3\beta_{UY_3}\nu_{0U}\right) \qquad (4.11) \\
\theta_{ZY_2} &= \beta_{UY_2}\gamma_Z^2 + 3\beta_{UY_3}\nu_{ZU}^2\nu_{0U} = \nu_{ZU}^2\left(\beta_{UY_2} + 3\beta_{UY_3}\nu_{0U}\right) \\
\theta_{AY_3} &= \beta_{AY_3} + \beta_{UY_3}\gamma_A^3 \qquad (4.12) \\
\theta_{ZY_3} &= \beta_{UY_3}\gamma_Z^3
\end{aligned}$$

Taking the same approach as before the ratio helps us correct the unmeasured confounder and unbias the estimates:

$$\begin{aligned}
\beta_{AY_1} &= \theta_{AY_1} - \theta_{AY_1}\frac{\nu_{AU}}{\nu_{ZU}} \\
\beta_{AY_2} &= \theta_{AY_2} - \theta_{ZY_2}\left(\frac{\nu_{AU}}{\nu_{ZU}}\right)^2 = \beta_{AY_2} - \theta_{ZY_2}\left(\frac{\gamma_{AW}}{\gamma_{ZW}}\right)^2 \\
\beta_{AY_3} &= \theta_{AY_3} - \theta_{ZY_3}\left(\frac{\nu_{AU}}{\nu_{ZU}}\right)^3 = \theta_{AY_3} - \theta_{ZY_3}\left(\frac{\gamma_{AW}}{\gamma_{ZW}}\right)^3
\end{aligned}$$

Once again, the fact that the second or third moment of the error do not disappear does not influence estimation of other parameters.

Note that if the order A is assumed to be greater than that of the unmeasured confounder, by the same reasoning in the previous section, the higher order terms will not be biased. On the other hand if the order of A is assumed to be lower than the unmeasured confounder lower order terms will be affected by the cross product terms of $(\nu_{0U} + \nu_{AU}A + \nu_{ZU}Z)^i$, nonetheless these terms can still be adjusted for by fitting the correct regression and using the symmetry of the problem in ν_{AU}, ν_{ZU}.

First Stage Higher Order, Second Stage Linear

Now we change things around, we are interested in the case where $U|ZA$ is still linear but the relationship between W and U is not.

$$\begin{aligned}
Y &= \beta_{0Y} + \beta_{AY}A + \beta_{UY}U + \epsilon_Y \\
W &= \beta_{0W} + \beta_{UW}U^2 + \epsilon_W \\
U &= \nu_{0U} + \nu_{AU}A + \nu_{ZU}Z + \epsilon_U
\end{aligned}$$

Observing the dependency of W on A and Z, now one needs to fit the correct first stage regression and include A^2, Z^2 and the cross-product terms AZ. We fit the following regression based on the observables:

$$\begin{aligned}
W &= \beta_{0W} + \beta_{UW}\left(\nu_{0U} + \nu_{AU}A + \nu_{ZU}Z + \epsilon_U\right)^2 + \epsilon_W \\
&= \beta_{0W} + \beta_{UW}(\nu_{0U}^2 + 2\nu_{0U}\epsilon_U + A^2\nu_{AU}^2 + 2\nu_{0U}A\nu_{AU} + 2A\nu_{ZU}\nu_{AU}Z \\
&\quad + 2A\nu_{AU}\epsilon_U + \nu_{ZU}^2Z^2 + 2\nu_{0U}\nu_{ZU}Z + 2\nu_{ZU}Z\epsilon_U + \epsilon_U^2) + \epsilon_W
\end{aligned}$$

In expectation the additive errors out and obtain:

$$\mathbb{E}\left[W|ZA\right] = \beta_{0W} + \beta_{UW}(\nu_{0U}^2 + A^2\nu_{AU}^2 + 2\nu_{0U}A\nu_{AU} + 2A\nu_{ZU}\nu_{AU}Z + \nu_{ZU}^2Z^2 + 2\nu_{0U}\nu_{ZU}Z + \sigma^2) + \epsilon_W$$

This suggests fitting the following regression:

$$\mathbb{E}\left[W|ZA\right] = \gamma_{0W} + \delta_{A_1}A + \delta_{A_2}A^2 + \delta_{Z_1}Z + \delta_{Z_2}Z^2 + \delta_{A1Z1}AZ \qquad (4.13)$$

In this case these fitted δs are estimating:

$$\begin{aligned}
\gamma_{0W} &= \beta_{0W} + \beta_{UW}(\nu_{0U}^2 + \sigma^2) \\
\delta_{A_1} &= 2\beta_{UW}\nu_{0U}\nu_{AU} \qquad (4.14) \\
\delta_{Z_1} &= 2\beta_{UW}\nu_{0U}\nu_{ZU} \qquad (4.15) \\
\delta_{A_1 Z_1} &= 2\beta_{UW}\nu_{AU}\nu_{ZU} \\
\delta_{A_2} &= \beta_{UW}\nu_{AU}^2 \\
\delta_{Z_2} &= \beta_{UW}\nu_{ZU}^2
\end{aligned}$$

Once again, the bias on Y by fitting a regression on the observables:

$$\mathbb{E}[Y|ZA] = \theta_0 + \theta_A A + \theta_Z Z$$

is given by:

$$\theta_A = \beta_{AY} + \beta_{UY}\nu_{AU} = \beta_{AY} + \frac{\nu_{AU}}{\nu_{ZU}}\theta_{ZY} \qquad (4.16)$$

$$\theta_Z = \beta_{UY}\nu_{ZU} \qquad (4.17)$$

The ratio $\frac{\nu_{AU}}{\nu_{ZU}}$ can be recovered from the ratio of equation (4.14) equation (4.15).

4.3 ProxySplines

In real world applications, the entire CERF is rarely of interest but rather only a range within which the maxima(minima) of the function falls in. While re-implementing the original paper of Bayesian 2SLS, the Markov chain quickly converged to locally linear polynomials on each quantiles of A[1]. This sparked the idea for an approach that is common in various sub-fields of mathematics, from Finite Element Methods methods in PDEs to computer graphics; splitting a non linear objects into smaller locally linear approximations. This approach consists in splitting the samples into smaller different subsets, and locally fitting an unbiased two stage least square linear regression. The hope is that the true CERF is smooth enough that in the limit, as the interval length goes to zero, the function is linear. Note that some smoothness condition is imposed on the CERF itself, which possesses real-world interpretability, as opposed to the bridge function, which does not have any interpretation.
In general, estimation of a linear function $f(x)$ is much easier than non-linear due to the linearity of the expectation $\mathbb{E}[f(X)|Z] = f(\mathbb{E}[X|Z])$. As such, one could fit Higher order regressions on local parts of A. The problem is that whereas in the BP2SLS, the subsets are automatically decided by the Bayesian hierarchical method, ProxySplines needs manual specification of the nodes. In such case, one would have to correctly identify these intervals a priori. Then, if the linear or higher order model holds on each specified subset we would be able to obtain a consistent estimate for the CERF.

A word of caution

Although polynomials are dense in L_2 and thus can theoretically approximate any continuous function, they are not the ideal choice. Polynomials are extremely prone to overfitting and outliers can heavily skew them. This is slightly remedied if one considers local polynomial as in ProxySplines. The Proxysplines assumption that the selected subsets indeed satisfy the linearity condition is very unrealistic. Nonetheless, in practical application such as the agriculture and fertilizer example, it could be reasonable that the expert has some idea of the ranges where we can expect a certain behaviour of the CERF.

[1] This was because the original posterior distributions were defined per quantile, thus automatically assigning the same mean leading to a quick convergence to the quantiles.

5

Non-Parametric Bayesian Proximal Inference

This chapter aims to introduce the notions of the Bayesian paradigm, in particular the non-parametric aspect of working with large models. The background information is mostly derived from [19], in particular Chapter 11. By no means does this chapter constitute an in depth analysis of the topic, but rather an attempt at summarizing and giving intuition behind the key elements of the field. We then proceed to introduce the Proxy Quasi-Bayes method, which leverages new results from the dual instrumental variable literature to develop a quasi Bayesian approach to estimate the outcome bridge function and the CERF.

5.1 Bayesian Non Parametrics

In statistics, there are two main philosophical approaches: frequentist and Bayesian. The frequentist approach focuses on the frequency or proportion of data, relying on the long-run behavior of estimates. Frequentists believe in the existence of a *true* underlying parameter and expect their estimators to converge to it at a certain rate. The other side of the coin is the Bayesian paradigm. Rather than focusing solely on a parameter pointwise, Bayesians are interested in the distribution of possible parameter values given the observed data. The initial *guess* or prior enables the incorporation of pre-existing knowledge about the parameter space rather than remaining agnostic about it. As more data becomes available, the initial belief is gradually updated to reflect the new information. Both of these advantages are achieved by leveraging Bayes' theorem to update beliefs, resulting in what is known as the posterior distribution:

$$\underbrace{P(\Theta|D^n)}_{Posterior} \propto \underbrace{P(D^n|\Theta)}_{Likelihood}\underbrace{P(\Theta)}_{Prior}$$

The posterior represents the degree of belief the Bayesian has over a parameter set and can be used for inference. One would hope, that as the amount of data increases, the effect of the prior to diminish and the posterior tends to concentrate around the *true* parameter, similar to what happens in the frequentist case. This is known as posterior consistency. In recent years, the Bayesian paradigm has become more widespread due to increased computational power. This has enabled the implementation of demanding approaches such as Markov Chain Monte Carlo (MCMC) methods, which allow for the approximations of complex posterior distributions that were previously intractable.

5.1.1 Gaussian Process Priors

The Bayesian approach of incorporating prior belief onto the parameter space is also applicable to infinite dimensional spaces such as function spaces, in such case we talk about Bayesian Non-parametrics. The most common approach is with the use of Gaussian Processes. These objects can be interpreted as the equivalent of the normal distribution over a large parameter space, a space of functions.

Definition 5.1 ([19] Gaussian Process)
A Gaussian Process is a stochastic process $G = \{G_t : t \in T\}$ indexed by an arbitrary set T such that:

$$[G_{t_1}, \dots G_{t_k}] \sim \mathcal{N}(\mu, \Sigma) \quad \forall t_1, \dots, t_k \in T \quad \forall k \in \mathbb{N}$$

For a given t, the marginals $[G_{t_1}, \dots G_{t_k}]$ are normally distributed. A multivariate normal distribution is identified by two parameters, the mean and the covariance matrix. By varying $t \in T$ we obtain the mean and covariance function:

$$\mu(t) = \mathbb{E}[G_t] \qquad k(s,t) = cov(G_s, G_t)$$

For convenience and without loss of generality the mean function is set to $\underline{0}$, this enables us to rewrite the covariance as $k(s,t) = \mathbb{E}[G_s G_t]$[1]. The sample paths of the process, $G_t(\omega)$ for ω in the sample space Ω, are then functions $t \mapsto G_t \in \mathbb{R}$. The *sampled function* depend on the sampled ω and as such are random functions. The process can thus be interpreted as a map between the sample space and some function space, i.e. $G_t(\cdot) : \Omega \to \mathcal{F}$ [19]. This suggests that a Gaussian Processes can be used as *distributions over functions*. Since the only choice of *parameter* is the covariance kernel, this is how prior knowledge is embedded into the prior distribution. In particular, as discussed in section 2.1, the kernel identifies a set of function thus, choosing the kernel somewhat chooses the space that one works on.

Series Expansion

In line with the presentation of section 2.1 where we highlighted the series expansion of functions in function spaces, we will primarily focus on Gaussian Process series priors. Any Gaussian random element in a separable $\mathcal{H}$ can be represented as its *Karhunen-Loève* expansion [19]. Let $\{\lambda_i, \varphi_i\}$ be the singular system of the integral operator associated with kernel k, then:

$$G = \sum_{i=1}^{+\infty} Z_i \lambda_i \varphi_i \qquad Z_i \sim \mathcal{N}(0,1) \tag{5.1}$$

This representation further highlights how Gaussian processes can be interpreted as priors on functions. This representation moves the randomness from G to the individual basis coefficients Z_i while $\lambda_i \varphi_i$ are non-random. In other words, sampling $G(\omega)$ consists in sampling an infinite sequence of coefficients Z_i. Heuristically, the object represented by G in equation (5.1) is a random Gaussian object in a function space since linear combinations of Gaussians *remain* Gaussian. Notice that while $Z_i \sim \mathcal{N}(0,1)$ the presence of the eigenvalues of k, rescale the variance of the coefficients to $\mathcal{N}\left(0, \lambda_i^2\right)$. Thus the *noise* in sampling these coefficients decreases with λ_i, which means that the trailing eigenfunctions will bring a smaller contribution. Connecting this with the remarks made on proposition 2.2 highlights that in $\mathcal{H}$, the coefficients G is distributed on a *infinite dimensional* ellipsoid with axis of diminishing length λ_i.

Reproducing Kernel Hilbert Space of a Gaussian Process

Since the covariance kernel k is the function that determines the covariance of the Gaussian process, one might wonder about the properties of the Reproducing Kernel Hilbert Space associated with such k and tus with the Gaussian Process.

Definition 5.2 ([19] RKHS of Gaussian Process)
Let $G = \{G_t : t \in T\}$ be a mean zero Gaussian Process. The Reproducing Kernel Hilbert Space of G is the set of functions:

$$\mathcal{H} = \left\{ f : T \to \mathbb{R} : f_H(t) = \mathbb{E}[G_t H] \quad H \in \overline{lin}(G) \right\}$$

where $\overline{lin}(G)$ is the first order chaos of G, the closure in L_2 of the set of linear combinations $\sum \alpha_i G_{t_i}$ $\alpha_i \in \mathbb{R}$. The inner product on $\mathcal{H}$ is:

$$\langle f_{H_1}, f_{H_2} \rangle_{\mathcal{H}} = \mathbb{E}[H_1 H_2]$$

This construction is clearly closely tied with the covariance of the GP as the defined inner product is just like the covariance. Indeed we notice that given a a function in the first order chaos $H = \sum \alpha_i G_{s_i}$, the associated function in $\mathcal{H}$ is given by:

$$f_H(t) = \mathbb{E}[H G_t] = \mathbb{E}\left[\sum_i \alpha_i G_{s_i} G_t\right] = \sum_i \alpha_i \mathbb{E}[G_{s_i} G_t] = \sum_i \alpha_i k(s_i, t)$$

This is same form of a generic RKHS definition 2.5 and since the Gaussian process is identified by its covariance kernel, the GP's RKHS is the one *generated* by the eigen-functions of the covariance kernel k(more precisely the singular system of the associated kernel operator). One might automatically assume that samples from it will belong to $\mathcal{H}$ with probability one. Counterintuitively, this is not the case[51, 27][2].

[1] $cov(A,B) = \mathbb{E}[AB] - \mathbb{E}[A]\mathbb{E}[B]$

[2] This is known as Driscoll's theorem.

Example 5.1 ([21]Brownian Motion)
Suppose that G is Brownian motion on the interval $[0,T]$. The covariance kernel of G is $k(s,t) = s \wedge t$. The RKHS of G is:

$$\mathcal{H} = \{f : f(0) = 0 \quad \int_0^T \left(\frac{df(t)}{dt}\right)^2 < +\infty\}$$

It can be shown that the sample paths of Brownian Motion are nowhere differentiable with probability one, thus its sample paths do not belong to $\mathcal{H}$ with probability one.

Example 5.2 (Heuristic Explanation[27])
Let $f \sim GP(0,k)$ be a sample from a Gaussian process and $\mathcal{H}$ its reproducing kernel Hilbert space. Define $f_n = \sum_{i=1}^n Z_i \lambda_i \varphi_i$. Then $f = \lim_{n\to\infty} f_n$. The expected $\mathcal{H}$ norm squared of f_n is given by:

$$\mathbb{E}\left[\|f_n\|_{\mathcal{H}}^2\right] = \mathbb{E}\left[\left\|\sum_{i=1}^n Z_i\lambda_i\varphi_i\right\|_{\mathcal{H}}^2\right] = \sum_{i=1}^n \mathbb{E}\left[Z_i^2\right]\lambda_i^2\|\varphi_i\|_{\mathcal{H}}^2 = \sum_{i=1}^n \mathbb{E}\left[Z_i^2\right] = \sum_{i=1}^n 1$$

Taking the limit this sum diverges and it the expected norm is infinite, thus not belonging to the RKHS. This is only for intuitive purposes since the convergence of the KL expansion $f = \lim_{n\to\infty} f_n$ is in the mean square sense.

In general, the samples from a Gaussian Process are rougher than the functions that lie within its reproducing kernel Hilbert space. Nonetheless, it is possible to show that there exists some power space of $\mathcal{H}$ such that the paths are there contained with probability one[51]. Many arguments in the Bayesian nonparametrics literature proceed by decomposing samples of GP in two parts, the main part which belongs inside $\mathcal{H}$ and an error term that remains in the power space.

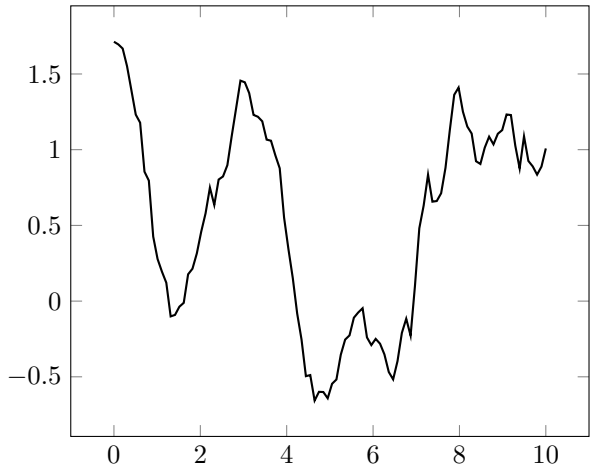


Figure 5.1: Sample from Gaussian Process with Matern Kernel.

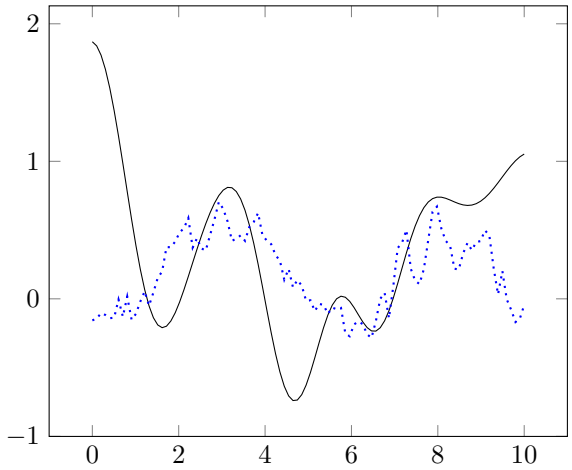


Figure 5.2: Decomposition in two terms: *smooth* in $\mathcal{H}$(full) and *rough* in $\mathcal{H}^\alpha$(dotted).

5.2 Quasi-Bayesian Methods

Returning to estimation of causal quantities under the proximal framework, the only Bayesian method developed so far is the one described in section 3.1.2. As discussed it forgoes the bridge functions altogether. It would be interesting to possibly consider a Bayesian non parametric approach to estimate the bridge function. Moreover, since CERF is weakly identified by the set of bridge functions, with a Bayesian approach we could find a set with high probability rather than a point solution. In the limit, one would hope that all elements of said set approximately satisfy the conditional moment restriction and thus enable us to find estimates for the CERF using some posterior estimator. This could overcome the issues of non-uniqueness.
The problem with directly implementing Gaussian process regression in the proximal causal learning setting is that it solves a *regular* regression problem:

Determine f such that:

$$\mathbb{E}[Y|X] = \mathbb{E}[f(X)|X] = f(X) \tag{5.2}$$

In such case, with additional assumptions on the distribution of $Y|X$, it is possible to obtain and estimate the likelihood from the observed data. If one appropriately chooses the likelihood, the posterior might admit a closed form. Even if this is not the case, it is possible to apply MCMC approaches that enable estimation of the posteriror through sampling schemes.
Clearly the approach of equation (5.2) is not compatible with the proximal inference problem due to the presence of the unmeasured confounder, meaning that the estimated $\hat{h}$ will not coincide with the true bridge h_0. Recent developments in the non-parametric literature have placed emphasis on so called Quasi-Bayesian methods. These methods can be used in situations where the likelihood itself is not readily available but rather other types of restrictions characterise the problem. The Quasi-Bayes approach substitutes the true likelihood with a function $Q\left(D^n, h\right)$, known as the quasi-likelihood. This function needs to behave like a likelihood, take values such that around neighbourhoods of the '*true*' parameter(function) Q has high values and lower further away from the truth. In practice, due to its resemblance to the Gaussian density, the most commonly chosen quasi-likelihood is constructed as an exponential with some data dependent loss function $\ell_n(h)$:

$$Q\left(D^n, h\right) = \exp\left(-\frac{n}{\lambda}\ell_n(h)\right) \tag{5.3}$$

This loss function needs to be appropriately chosen for the problem at hand as long as $\ell_n(h) \downarrow 0$ in appropriate neighbourhoods of the truth and $\ell_n(h) \uparrow$ as we move further away. Bayes' theorem can then be used to construct:

$$\frac{\Pi_n\left(dh|\mathcal{D}^n\right)}{\Pi(dh)} \propto Q\left(D^n, h\right) \tag{5.4}$$

where the $\propto$ is up to a normalization constant such that $\Pi_n(dh|\mathcal{D}^n)$ integrates to one to be a distribution. Although equation (5.4) resembles Bayes theorem, the quasi-posterior Π_n is not a posterior in the classic sense since it does not involve the true likelihood but rather a quasi-posterior.

Example 5.3 (Gaussian Process Regression)
Suppose that $f \sim GP$ and consider the quasi likelihood of equation (5.3) *with the loss $l_n(f) = \frac{\lambda}{n}\sum_i^n (f(x_i) - y_i)^2$. Then the quasi-likelihood coincides with the likelihood. Moreover if $Y \sim \mathcal{N}\left(f(X), \sigma^2\right)$ then the quasi-likelihood is the 'regular' likelihood and the quasi-Bayes and Bayes methods are the same.*

5.2.1 Proximal Quasi Bayesian

The authors of [63] developed a quasi Bayesian approach to estimate the structural function f_0 in the non-parametric instrumental variable setting. Instrumental variables are also models characterized by conditional moment restrictions. In particular, they respect the following graphical model:

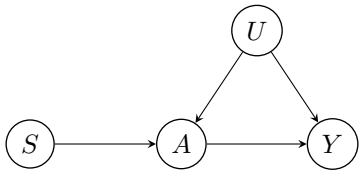


Figure 5.3: Graphical model for instrumental variables.

The conditional moment restriction that characterizes this model is given by $\mathbb{E}\left[Y|S\right] = \mathbb{E}\left[f_0(A)|S\right]$. Intuitively, the effect that S has on A enables us to adjust for the effect of U. This conditional moment restriction is extremely similar to the one present in proximal inference case where we replace S with (V_q) and (V_h) as arguments of the outcome bridge function h_0 rather than the structural function f_0. The procedure is one stage and thus the chosen loss for the quasi likelihood involves the mean violation of the conditional moment restriction. As already discussed in section 2.3, minimization of this loss does not guarantee the minimization of $\|h - h_0\|$ due to the ill-posedness nature of the problem. Nonetheless, as seen in other works, the ill posedness can be controlled on certain function classes. The population loss is the following:

$$\mathcal{L}(h) = \max_{g \in \mathcal{G}} \mathbb{E}\left[(h(V_h) - Y) \cdot g(V_q) - \frac{1}{2}g^2(V_q)\right] \tag{5.5}$$

The authors of [63] justify this choice due to Fenchel duality(see proposition A.1). We prefer to justify it in terms of the stabilizers also discussed in section 3.2. These stabilizers do not add bias to our estimate

equation but are a reformulation of the conditional expectation as shown in proposition 3.4:

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[(Y-h(V_h))|V_q\right]^2\right] &= \sup_{g\in L_2} 2\mathbb{E}\left[g(V_q)\mathbb{E}\left[(Y-h(V_h))|V_q\right]\right] - \mathbb{E}\left[g(V_q)^2\right] \\ &= \sup_{g\in L_2} 2\mathbb{E}\left[\mathbb{E}\left[g(V_q)(Y-h(V_h))|V_q\right]\right] - \mathbb{E}\left[g(V_q)^2\right] \\ &= \sup_{g\in L_2} 2\mathbb{E}\left[g(V_q)(Y-h(V_h))\right] - \mathbb{E}\left[g(V_q)^2\right]\end{aligned}$$

Although obtained by different means, the final objective is the same as in the Quasi Bayes dual IV setting but with the additional presence of the treatment variable on both sides of conditioning. Similarly to all other methods studied, regularization terms are added to avoid issues discussed in section 2.3. The objective can thus be reformulated as an adversarial problem:

$$\min_{h\in\mathcal{H}}\max_{g\in\mathcal{G}} 2\mathbb{E}\left[(h(V_h)-Y)\cdot g(V_q) - \frac{g^2(V_q)}{2}\right] - \nu\left\|g\right\|_{\mathcal{G}}^2 \tag{5.6}$$

Notice that in the end, this loss function of equation (5.9) is similar to the one used in equation (3.20).

Empirical loss

The empirical equivalent of equation (5.6) is a specific case of a general loss function $d_n^2\left(\hat{E}_n h - \hat{b}\right)$. In our case we choose $d_n^2(g) = \|g\|_n^2 + \nu\|g\|_{\mathcal{G}}^2 = \frac{1}{n}\sum_{i=1}^n g(V_{q_i})^2 + \nu\|g\|_{\mathcal{G}}^2$ and $\hat{E}_n(h), \hat{b}$ are solutions to a ridge regression problem.

Proposition 5.1
Let $\hat{E}_n h, \hat{b}$ be the solutions to regularized ridge regression problems:

$$\hat{E}_n h = \arg\min_{\mathcal{G}} \frac{1}{n}\sum (h(V_{h_i}) - g(V_{q_i}))^2 + \nu\|g\|_{\mathcal{G}}^2$$
$$\hat{b} = \arg\min_{\mathcal{G}} \frac{1}{n}\sum (y_i - g(V_{q_i}))^2 + \nu\|g\|_{\mathcal{G}}^2$$

Moreover let $d_n^2(f) = \frac{1}{n}\sum^n f(x_i)^2 + \nu\|f\|_{\mathcal{H}_x}^2$. Then we have that:

$$\frac{1}{2}d_n^2(\hat{E}_n h - \hat{b}) = \max_{\mathcal{G}} \frac{1}{n}\left[\sum_{}^{n}(h(V_{h_i}) - y_i)g(V_{q_i}) - \frac{g(V_{q_i})^2}{2}\right] - \frac{\nu}{2}\|g\|_{\mathcal{G}}^2$$

Proof. Let S_{V_h} and S_{V_q} be the sampling operators of V_h, V_q respectively (definition 2.12). We start off with determining the solutions to the ridge regression problems:

$$\begin{aligned}\hat{E}_n h &= \arg\min_{\mathcal{G}} \frac{1}{n}\sum (h(V_{h_i}) - g(V_{q_i}))^2 + \nu\|g\|_{\mathcal{G}}^2 \\ &= \arg\min_{\mathcal{G}} \frac{1}{n}\left\langle S_{V_h}h - S_{V_q}g, S_{V_h}h - S_{V_q}g\right\rangle + \nu\left\langle g,g\right\rangle_{\mathcal{G}} \\ &= \arg\min_{\mathcal{G}} \frac{1}{n}\left[\left\langle S_{V_h}^* S_{V_h}h, h\right\rangle_{\mathcal{H}} + \left\langle S_{V_q}^* S_{V_q}g, g\right\rangle_{\mathcal{G}} - 2\left\langle S_{V_q}^* S_{V_h}h, g\right\rangle_{\mathcal{G}}\right] + \nu\left\langle g,g\right\rangle_{\mathcal{G}}\end{aligned}$$

Deriving in g we obtain the minimizer as the solution to:

$$\frac{1}{n}S_{V_q}^* S_{V_q}g + \nu g = \frac{1}{n}S_{V_q}^* S_{V_h}h$$
$$\left(S_{V_q}^* S_{V_q} + n\nu I\right)g = S_{V_q}^* S_{V_h}h$$

Similarly for $\hat{b}$:

$$\begin{aligned}\hat{b} &= \arg\min_{\mathcal{G}} \frac{1}{n}\sum (y_i - g(V_{q_i}))^2 + \nu\|g\|_{\mathcal{G}}^2 \\ &= \arg\min_{\mathcal{G}} \frac{1}{n}\left\langle Y - S_{V_q}g, Y - S_{V_q}g\right\rangle + \nu\left\langle g,g\right\rangle_{\mathcal{G}} \\ &= \arg\min_{\mathcal{G}} \frac{1}{n}\left[\left\langle Y,Y\right\rangle_2 + \left\langle S_{V_q}^* S_{V_q}g, g\right\rangle_{\mathcal{G}} - 2\left\langle S_{V_q}^* Y, g\right\rangle_{\mathcal{G}}\right] + \nu\left\langle g,g\right\rangle_{\mathcal{G}}\end{aligned}$$

$$\frac{1}{n}S_{V_q}^* S_{V_q} g + \nu g = \frac{1}{n}S_{V_q}^* Y$$
$$\left(S_{V_q}^* S_{V_q} + n\nu I\right) g = S_{V_q}^* Y$$

Thus $\hat{E}_n h - \hat{b}$ must satisfy:

$$\left(S_{V_q}^* S_{V_q} + n\nu I\right) \hat{E}_n h - \hat{b} = \left(S_{V_q}^* S_{V_h} h - S_{V_q}^* Y\right)$$

Let us now focus on the general expression of $d_n^2(f)$:

$$\begin{aligned} d_n^2(f) &= \frac{1}{n}\sum^n f(x_i)^2 + \nu \|f\|_{\mathcal{H}_x}^2 \\ &= \frac{1}{n}\langle S_x f, S_x f\rangle + \nu \langle f, f\rangle_{\mathcal{H}_x} \\ &= \frac{1}{n}\langle S_x^* S_x f, f\rangle_{\mathcal{H}_x} + \nu \langle f, f\rangle_{\mathcal{H}_x} \\ &= \frac{1}{n}\langle (S_x^* S_x + n\nu I) f, f\rangle_{\mathcal{H}_x} \end{aligned}$$

If we plug in $\hat{E}_n h - \hat{b}$ we obtain:

$$\begin{aligned} d_n^2(\hat{E}_n h - \hat{b}) &= \frac{1}{n}\left\langle \left(S_{V_q}^* S_{V_q} + n\nu I\right)\left(S_{V_q}^* S_{V_q} + n\nu I\right)^{-1}\left(S_{V_q}^* S_{V_h} h - S_{V_q}^* Y\right), \left(S_{V_q}^* S_{V_q} + n\nu I\right)^{-1}\left(S_{V_q}^* S_{V_h} h - S_{V_q}^* Y\right)\right\rangle_{\mathcal{G}} \\ &= \frac{1}{n}\left\langle \left(S_{V_q}^* S_{V_h} h - S_{V_q}^* Y\right), \left(S_{V_q}^* S_{V_q} + n\nu I\right)^{-1}\left(S_{V_q}^* S_{V_h} h - S_{V_q}^* Y\right)\right\rangle_{\mathcal{G}} \end{aligned} \tag{5.7}$$

I now want to show that: $\max_{\mathcal{G}} \frac{1}{n}\left[\sum^n (h(V_{h_i}) - y_i) g(V_{q_i}) - \frac{g(V_{q_i})^2}{2}\right] - \frac{\nu}{2}\|g\|_{\mathcal{G}}^2$ is equivalent to $\frac{1}{2} d_n^2(\hat{E}_n h - \hat{b})$:

$$\begin{aligned} \frac{1}{n}\left[\sum^n (h(V_{h_i}) - y_i) g(V_{q_i}) - \frac{g(V_{q_i})^2}{2}\right] - \frac{\nu}{2}\|g\|_{\mathcal{G}}^2 &= \frac{1}{n}\langle S_{V_h} h - Y, S_{V_q} g\rangle_2 - \frac{1}{2n}\langle S_{V_q} g, S_{V_q} g\rangle_2 - \frac{\nu}{2}\langle g, g\rangle_{\mathcal{G}} \\ &= \frac{1}{n}\left\langle S_{V_q}^* S_{V_h} h - S_{V_q}^* Y, g\right\rangle_{\mathcal{G}} - \frac{1}{2n}\left\langle S_{V_q}^* S_{V_q} g, g\right\rangle_{\mathcal{G}} - \frac{\nu}{2}\langle g, g\rangle_{\mathcal{G}} \\ &= \frac{1}{n}\left\langle S_{V_q}^* S_{V_h} h - S_{V_q}^* Y - \frac{1}{2} S_{V_q}^* S_{V_q} g - \frac{n\nu}{2} g, g\right\rangle_{\mathcal{G}} \\ &= \frac{1}{n}\left\langle S_{V_q}^* S_{V_h} h - S_{V_q}^* Y - \frac{1}{2}\left(S_{V_q}^* S_{V_q} + n\nu I\right) g, g\right\rangle_{\mathcal{G}} \end{aligned}$$

This quantity is maximized by choosing g as:

$$g = \left(S_{V_q}^* S_{V_q} + n\nu I\right)^{-1}\left(S_{V_q}^* S_{V_h} h - S_{V_q}^* Y\right)$$

Substituting this in the previous we obtain:

$$\frac{1}{2n}\left\langle S_{V_q}^* S_{V_h} h - S_{V_q}^* Y, \left(S_{V_q}^* S_{V_q} + n\nu I\right)^{-1}\left(S_{V_q}^* S_{V_h} h - S_{V_q}^* Y\right)\right\rangle_{\mathcal{G}}$$

which is exactly half of Equation 5.7.

□

Choosing $l_n(h) = \frac{1}{2} d_n^2(\hat{E}_n h - \hat{b})$ and plugging this choice of loss into the quasi-likelihood of equation (5.3) we obtain:

$$\frac{\Pi_n\left(dh|\mathcal{D}^n\right)}{\Pi(dh)} \propto exp\left(-\frac{n}{2\lambda} d_n^2(\hat{E}_n(h) - \hat{b})\right) \tag{5.8}$$

The restriction on the function class $\mathcal{G}$ is to contain $\mathbb{E}\left[h(V_h) - h'(V_h)|V_q\right] \quad \forall h, h' \in \mathcal{H}$ and assumption A.7. In particular, we focus on $\mathcal{G}$ being a RKHS due to both the computational aspect and the nice properties involving the ill posedness to guarantee the latter assumption. The empirical equivalent of the loss in equation (5.6) is found by simply replacing the expectation with the sample mean:

$$\min_{\mathcal{H}} \hat{\mathcal{L}}(h) = \min_{h\in\mathcal{H}} \max_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \left((h(V_{h_i}) - y_i) g(V_{q_i}) - \frac{g(V_{q_i})^2}{2}\right) - \frac{\nu}{2}\|g\|_{\mathcal{G}}^2 \tag{5.9}$$

5.2.2 Results

A Bayesian procedure, or Quasi-Bayesian in this case, is only really of interest if in the limit it *identifies* the correct set of parameters. This is equivalent to the concept of consistency of a frequentist estimator; if the entire distribution is *observed*, the estimator converges to an area around the *true* one.

Definition 5.3 ([19]Posterior Consistency)
The posterior distribution $\Pi_n(\cdot|\mathcal{D}^n)$ *is said to be weakly consistent at* $\theta_0 \in \Theta$ *if*

$$\forall \epsilon > 0 \quad \Pi_n(\{\theta, d(\theta, \theta_0) > \epsilon | \mathcal{D}^n\} \to 0$$

in P_{θ_0} *probability as* $n \to \infty$*, where* $d(\cdot, \cdot)$ *is a distance.*

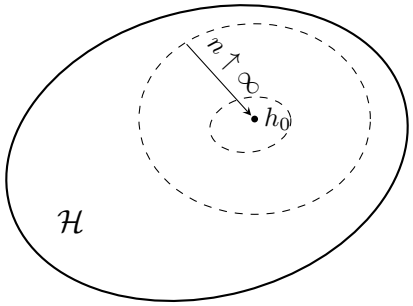


Figure 5.4: Graphical representation of desired behaviour of posterior distribution as number of observation increases.

Figure 5.4 shows what we would like to achieve with the Quasi Bayesian procedure, the posterior shrinking around the true bridge function h_0. If the posterior is consistent, then it concentrates its mass in smaller and smaller areas of the true parameter as more data is observed. Consistency can also be interpreted as the posterior distribution converging weakly to a Dirac delta at the true parameter in probability [19]. The following result shows that, under the assumptions discussed in appendix A.3.2 the quasi-posterior built using equation (A.2) places its mass around the true bridge function.

Theorem 5.1 ([63]Quasi-Bayes Posterior Consistency)
Let $p \geq 0, b > 1$ *be constants as in appendix A.3.2. Under the assumptions in appendix A.3.2, fix* $\lambda = 1, \overline{\nu} = C\delta_n^2 \asymp n^{-\frac{b+2p}{b+2p+1}}, C > 0$*. Then there exists a constant* $M > 0$ *such that for* $\epsilon_n^2 = n^{-\frac{b}{b+2p+1}}$ *we have:*

$$\begin{aligned} &\Pi_n(\{h : \|h - h_0\|_2^2 > M\epsilon_n^2\}|D^n) \xrightarrow{P_{\mathcal{D}^n}} 0 \\ &\Pi_n(\{h : \|E(h - h_0)\|_2^2 > M\delta_n^2\}|D^n) \xrightarrow{P_{\mathcal{D}^n}} 0 \end{aligned} \tag{5.10}$$

where E *is the conditional expectation operator,* $Eh = \mathbb{E}[h(V_h)|V_q]$*,* b *dictates the regularity of* h_0 *and* p *the level of ill-posedness of* E*.*

Inference

Theorem 5.1 shows that the quasi-posterior concentrates around the true structural function or in our case the bridge functions h_0. This enables weak identification of the CERF. We are now interested in performing inference on the CERF using the posterior distribution $\Pi_n(dh|\mathcal{D}^n)$
A commonly adopted estimator in Bayesian statistics is the Maximum A Posteriori (MAP) estimator, $\hat{\theta}_{MAP} = \arg\max_\theta \Pi_n(\theta|\mathcal{D}^n)$.

Example 5.4 (MLE and MAP)
Using Bayes' theorem we have that:

$$P(\Theta|D^n) \propto P(D^n|\Theta)P(\Theta)$$

The MAP is the θ *that maximizes the posterior distribution* $P(\cdot|\mathcal{D}^n)$ *whereas the MLE maximizes* $P(\mathcal{D}^n|\cdot)$*. This shows that when the prior is uniform over* Θ*, the MAP and MLE coincide. In general, the MAP can be interpreted as a maximum likelihood reweighed by initial belief. If, as the amount of data increases, the prior effect washes out then the two coincide.*

A MAP-like estimator could be built using the quasi-posterior. If the parameter of interest lives in a subset of $\mathbb{R}^d$ then this will coincide with the mode, or peak, of the posterior density. In the infinite dimensional case it becomes slightly more complicated, the MAP coincides withe the center of the smallest ball that contains half of the posterior mass. Clearly, the MAP does not have much computational appeal. A second and possibly more computational friendly point estimator is the posterior mean.

Definition 5.4 (Posterior mean)
Let $\Pi_n(\cdot|\mathcal{D}^n)$ be the posterior mean, the Posterior Mean is defined as:

$$\hat{h}_n = \int h\Pi_n(dh|\mathcal{D}^n)$$

The appeal of the posterior mean as an estimator is its computational simplicity. In practice, if one has conjugate prior, a closed form solution can be derived. If one carries out an MCMC procedure, it can be simply calculated as the average of the chain. Similar ideas can be adapted for the quasi-bayesian approach. Unfortunately simple does not always coincide with practical. Countrerintuitively, just because the posterior distribution concentrates around the optimal parameter, it does not automatically guarantee that the posterior mean also adopts the same property.

Note

Although the authors of [63] state that they are able to obtain consistency of the posterior mean over the entire parameter set, it is not clear how. The main problem is that the set of functions where the ill-posedness of the problem is finite, and thus the error sets of theorem 5.1 are equivalent up to multiplicative constants, has a small measure but unfortunately not necessarily small enough. Even though from a prior point of view that set is negligible (corollary A.1), the L_2 norm of the GP samples might be infinitely larger and blow up said convergence.

Nonetheless a possible solution is to construct the Θ_m posterior mean, i.e. $\hat{h}_{\Theta_m} = \mathbb{E}\left[\mathbf{1}_{\Theta_m} \cdot h|\mathcal{D}^n\right]$ with $\Theta_m = \{h = h_\rho + h_e : \|h_\rho\|^2_{\mathcal{H}} \leq cn^{\frac{1}{b+2p+1}},\ \|h_e\|_2^2 \leq cn^{-\frac{b}{b+2p+1}},\ \|h_e\|_\infty \leq c\}$. This is a set of functions that, as n increases more and more RKHS mass and smaller L_2 mass. This enables us to use the results from the proof of theorem 5.1 to ensure that this estimator is arbitrarily L_2 close to the true h_0. The downside is that this estimator does not have a closed form solution like the regular quasi-posterior mean and it is not clear how it could be calculated.

Proposition 5.2 (Consistency of the Θ_m Posterior Mean)
With high probability, there exists a constant $\tilde{M} > 0$ such that the posterior mean $\hat{h}_{\Theta_m}$ of $\Pi_n(\cdot|\mathcal{D}^n)$ satisfies:

$$\left\|\hat{h}_{\Theta_m} - h_0\right\|_2^2 \leq \tilde{M}\epsilon_n^2$$

Proof. We are interested in the convergence to zero of the probability of $err_{\hat{h}_{\Theta_m}} = \left\{\mathcal{D}^n : \left\|\hat{h}_{\Theta_m} - h_0\right\|_2^2 > \tilde{M}\epsilon_n^2\right\}$.
By the law of total probabilities.

$$\begin{aligned}
P\left(err_{\hat{h}_{\Theta_m}}\right) &= P\left(err_{\hat{h}_{\Theta_m}} \cap E_n(\mathcal{D}^n)\right) + P\left(err_{\hat{h}_{\Theta_m}} \cap \overline{E_n}(\mathcal{D}^n)\right) \\
&\leq P\left(err_{\hat{h}_{\Theta_m}} \cap E_n(\mathcal{D}^n)\right) + P\left(\overline{E_n}(\mathcal{D}^n)\right) \\
&= P\left(err_{\hat{h}_{\Theta_m}} \cap E_n(\mathcal{D}^n) \cap A\right) + P\left(err_{\hat{h}_{\Theta_m}} \cap E_n(\mathcal{D}^n) \cap \overline{A}\right) + P\left(\overline{E_n}(\mathcal{D}^n)\right) \\
&\leq P\left(err_{\hat{h}_{\Theta_m}} \cap E_n(\mathcal{D}^n) \cap A\right) + P\left(\overline{A}\right) + P\left(\overline{E_n}(\mathcal{D}^n)\right) \\
&\leq P\left(err_{\hat{h}_{\Theta_m}} \cap E_n(\mathcal{D}^n) \cap A\right) + o_P(1) + o_P(1)
\end{aligned}$$

By the proof of theorem 5.1 in appendix A.3.2, the last two terms go to zero. We are only concerned on the deviations of the posterior mean on the set $E_n(\mathcal{D}^n) \cap A$. On this set we can now focus on the

deviation from the truth and use the results and bounds from the proof:

$$\begin{aligned}
\left\|h_0 - \hat{h}\right\|_2^2 &= \|h_0 - \mathbb{E}\left[h|D^n\right]\|_2^2 \\
&\leq \mathbb{E}\left[\|h_0 - h\|_2^2 \,|D^n\right]^2 \\
&= \int \|h_0 - h\|_2^2 d\Pi_n(h|D^n) \\
&= \int_{\|h_0-h\|_2^2 \geq M\epsilon_n^2} \|h_0 - h\|_2^2 d\Pi_n(h|D^n) + M\epsilon_n^2 \int_{\|h_0-h\|_2^2 < M\epsilon_n^2} d\Pi_n(h|D^n) \\
&\leq \int_{\|h_0-h\|_2^2 \geq M\epsilon_n^2} \|h_0 - h\|_2^2 d\Pi_n(h|D^n) + M\epsilon_n^2
\end{aligned}$$

The first term is the one that can cause *problems* because of the possibly large norm of h. We overcome this by decomposing the errors with a shelling argument as follows:

$$\begin{aligned}
\int_{\|h_0-h\|_2^2 \geq M\epsilon_n^2} \|h_0 - h\|_2^2 d\Pi_n(h|D^n) &= \sum_{i=1}^{\infty} \int_{Mi\epsilon_n^2 < \|h_0-h\|_2^2 \leq M(i+1)\epsilon_n^2} \|h_0 - h\|_2^2 d\Pi_n(h|D^n) \\
&\leq \sum_{i=1}^{\infty} M(i+1)\epsilon_n^2 \int_{Mi\epsilon_n^2 < \|h_0-h\|_2^2 \leq M(i+1)\epsilon_n^2} d\Pi_n(h|D^n) \\
&\leq \sum_{i=1}^{\infty} M(i+1)\epsilon_n^2 \int_{\|h_0-h\|_2^2 \geq Mi\epsilon_n^2} d\Pi_n(h|D^n) \\
&\leq \epsilon_n^2 \sum_{i=1}^{\infty} M(i+1)\mathrm{e}^{-M^2 i^2 n\delta_n^2} \lesssim \epsilon_n^2 \frac{\mathrm{e}^{-n\delta_n^2}}{n\delta_n^2}
\end{aligned}$$

Where the last two inequalities follows from the last step in the proof of theorem 5.1 and the fact that a sum can be upper bounded by the integral. In our case we have:

$$\sum_{t=2}^{+\infty} te^{-t^2 n\delta_n^2} \leq \int_1^{\infty} te^{-t^2 n\delta_n^2} dt = -\frac{e^{-t^2 n\delta_n^2}}{2n\delta_n^2}\Bigg|_1^{+\infty} = \frac{e^{-n\delta_n^2}}{2n\delta_n^2}$$

This shows that on the given subset $err_{\hat{h}_{\Theta_m}} \cap E_n(\mathcal{D}^n) \cap A$, the error is smaller than $\left\|\hat{h}_{\Theta_m} - h_0\right\|_2^2 \leq 2M\epsilon_n^2 \leq \tilde{M}\epsilon_n^2$. Applying the Continuous Mapping Theorem[3] to the sum of o_P converging terms we obtain convergence of the whole sum completing the proof. □

Now that we have shown that the posterior mean comes arbitrarily close to the true bridge function with high probability, all that is left to show is that the sample estimator built with it is consistent and nears the true CERF. To do this, we require the following assumption:

Assumption 5.1
Let f_W, f_A, f_X be the marginal distribution of W, A, X and $f_{W,A,X}$ be its joint density. Assume that $w \in \mathcal{W}, \forall a \in \mathcal{A}, x \in \mathcal{X}$:

$$f_W(w) \cdot f_A(a) \cdot f_X(x) \lesssim f_{W,A,X}(w, a, x)$$

Notice that this assumption is similar to the one presented in section 3.5.2. It is a restriction on the distributions of the variables which intuitively requires that observing one of the two does not *rule out* the chance of observing the other. We now formulate our CERF estimator constructed on the quasi-posterior Θ_m mean. To avoid the complications from using the same samples used in determining $\hat{h}$, we use sample-splitting.

Definition 5.5 (PQB Estimator)
Let $\mathcal{D}^t$ be a test set of unobserved samples of size n_t. Let $\hat{h}_{\Theta_m}$ be the posterior mean of the Quasi Bayesian procedure. The Proxy Quasi Bayesian (PQB) estimator is defined as:

$$\hat{\chi}(a) = \frac{1}{n}\sum_{i=1}^{n} \hat{h}_{\Theta_m}(w_i, a, x_i) \tag{5.11}$$

[3] If $(X_n, Y_n) \to 0$ in probability then for all continuous functions f we have that $f(X_n, Y_n) \to f(X, Y)$ in probability.

We now need to show that this choice of estimator is indeed a good choice as it will asymptotically approximate the true CERF.

Theorem 5.2 (CERF Estimator Consistency)
Let $\hat{\chi}$ be the PQB estimator constructed in definition 5.5. Let b, p be the regularities of the true bridge function defined in assumption A.3. Then:

$$\|\chi - \hat{\chi}\|_{L_2(A)} = O({n_t}^{-\frac{1}{2}} + n^{-\frac{b}{2(b+2p+1)}})$$

Proof.

$$\begin{aligned}
\|\chi - \hat{\chi}\|_2 &= \|\mathbb{E}_{W,X}\left[h_0(W,\cdot,X)\right] - \hat{\chi}\|_2 \\
&= \left\|\mathbb{E}_{W,X}\left[h_0(W,\cdot,X) - \hat{h}_{\Theta_m}(W,\cdot,X) + \hat{h}_{\Theta_m}(W,\cdot,X)\right] - \hat{\chi}\right\|_2 \\
&= \underbrace{\left\|\mathbb{E}_{W,X}\left[h_0(W,\cdot,X) - \hat{h}_{\Theta_m}(W,\cdot,X)\right]\right\|_2}_{A} + \underbrace{\left\|\mathbb{E}_{W,X}\left[\hat{h}_{\Theta_m}(W,\cdot,X)\right] - \hat{\chi}\right\|_2}_{B}
\end{aligned}$$

We first focus on term A:

$$\begin{aligned}
\left\|\mathbb{E}_{W,X}\left[h_0(W,\cdot,X) - \hat{h}_{\Theta_m}(W,\cdot,X)\right]\right\|_2^2 &\leq \int\int \left(h_0(w,a,x) - \hat{h}_{\Theta_m}(w,a,x)\right)^2 f(w)f(a)dadw \\
&\lesssim \int\int \left(h_0(w,a,x) - \hat{h}_{\Theta_m}(w,a,x)\right)^2 f(a,w)dadw \\
&= \left\|h_0 - \hat{h}_{\Theta_m}\right\|_2^2 \lesssim \epsilon_n^2 \asymp n^{-\frac{b}{b+2p+1}}
\end{aligned}$$

This is by theorem 5.1. Now for the second term B, using the fact that $\mathcal{D}^t$ is unobserved (sample splitting):

$$\begin{aligned}
\left\|\mathbb{E}_{W,X}\left[\hat{h}_{\Theta_m}(W,\cdot,X)\right] - \mathbb{E}\left[\frac{1}{n_t}\sum_{i=1}^{n_t}\hat{h}_{\Theta_m}(w_i,\cdot)\right]\right\|_{L_2(A)}^2 &\leq \int \mathbb{E}\left(\mathbb{E}_W\left[\hat{h}_{\Theta_m}(W,a)\right] - \frac{1}{n_t}\sum_{i=1}^{n_t}\hat{h}_{\Theta_m}(w_i,a,x_i)\right)^2 f(a)da \\
&= \int Var\left(\frac{1}{n_t}\sum_{i=1}^{n_t}\hat{h}_{\Theta_m}(w_i,a,x_i)\right) f(a)da \\
&= \int \frac{1}{n_t} Var\left(\hat{h}_{\Theta_m}(W,a)\right) f(a)da \\
&\leq \int \frac{1}{n_t}\mathbb{E}_{W,X}\left[\hat{h}_{\Theta_m}(W,a,X)^2\right] f(a)da \\
&= \int\int \frac{1}{n_t}\hat{h}_{\Theta_m}(w,a,x)^2 f(w)f(a)f(x)dwdadx \\
&\lesssim \frac{1}{n_t}\int \hat{h}_{\Theta_m}(w,a,x)^2 f(w,a,x)dwdadx \\
&= \frac{1}{n_t}\left\|\hat{h}_{\Theta_m}\right\|_{L_2(V_h)}^2 \lesssim \frac{1}{n_t}
\end{aligned}$$

Notice that we assume that $h_0 \in \overline{\mathcal{H}} \subset L_2(V_h)$ which means that $\|h_0\|_{L_2(V_h)} < +\infty$. By theorem 5.1 we have that with high probability $\left\|h_0 - \hat{h}_{\Theta_m}\right\|_2^2 \leq M\epsilon_n^2 < +\infty$. Applying the triangle inequality $\left\|\hat{h}_{\Theta_m}\right\|_2 = \left\|\hat{h}_{\Theta_m} - h_0 + h_0\right\|_2 \leq \left\|\hat{h}_{\Theta_m} - h_0\right\|_2 + \|h_0\|_2 < +\infty$. Taking the square roots and recombining the two terms we obtain the desired result. □

The main rate of convergence of this estimator is the one adopted from the convergence of the posterior mean to the bridge function $\epsilon_n^2 = n^{-\frac{b}{b+2p+1}}$. Per assumption A.5, the parameter p decides the eigen-decay of the conditional expectation operator, in other words the level of ill-posedness. If $p = 0$ then there would be no ill-posedness and the best rate one could strive for is parametric $O(n^{-\frac{1}{2}})$ for $b \to \infty$. Clearly this is only possible for a kernel with very quickly decaying eigenvalues, and thus very smooth *true* bridge function h_0. Moreover, the problem is ill-posed and thus $p \neq 0$. Nonetheless, depending on the p, b of the problem it might surpass other rates mentioned so far.

5.2.3 Closed Form Estimator

As suggested by proposition 5.1, the operator $\hat{E}_n$ can be constructed using the empirical covariance matrices as $\hat{E}_n = \hat{C}^{-1}_{V_q,V_q}\hat{C}_{V_q,V_h}$. To include regularization and match the above definition, we consider the regularized version of the covariance operator $\hat{C}_{V_q,V_q,\nu} = \left(\hat{C}_{V_q,V_q} + n\nu \cdot I\right)$. The regularization term improves the conditioning of the estimated matrix. In terms of the sampling operators and their adjoints $\hat{C}_{V_q,V_h} = \frac{1}{n}S^*_{V_q}S_{V_h}$ (definition 2.12). The source condition is estimated in a similar way as $\mathbb{E}[Y|V_q] \approx \hat{b} = \frac{1}{n}S^*_{V_q}Y$.

Lemma 5.1 ([63])
The loss function of equation (5.9) *can be rewritten as:*

$$\hat{\mathcal{L}}(h) = \left\| \hat{C}^{-\frac{1}{2}}_{V_q,V_q,\nu}\left(\hat{C}_{V_q,V_h}h - \frac{S^*_{V_q}Y}{n}\right)\right\|^2_{\mathcal{G}} \tag{5.12}$$

$$\tag{5.13}$$

and it admits closed form minimizer:

$$\tilde{h} = \hat{C}^{-1}_{V_q,V_h}\frac{S^*_{V_q}}{n}Y$$

Moreover, if a regularization term $\lambda\|h\|^2_{\mathcal{H}}$ is added to the the above loss, the solution becomes:

$$\tilde{h} = \left(\hat{C}_{V_h,V_q}\hat{C}^{-1}_{V_q,V_q,\nu}\hat{C}_{V_q,V_h} + \lambda I\right)^{-1}\frac{S^*_{V_q}}{n}Y = \left(S^*_{V_h}LS^*_{V_h} + \lambda I\right)^{-1}\frac{S^*_{V_q}}{n}Y$$

Proof. From the proof of proposition 5.1 we have that:

$$\begin{aligned}
\frac{1}{2}d^2_n(\hat{E}_nh - \hat{b}) &= \frac{1}{n}\left\langle \left(S^*_{V_q}S_{V_h}h - S^*_{V_q}Y\right), \left(S^*_{V_q}S_{V_q} + n\nu I\right)^{-1}\left(S^*_{V_q}S_{V_h}h - S^*_{V_q}Y\right)\right\rangle_{\mathcal{G}} \\
&= \left\langle C_{V_q,V_h}h - \frac{S^*_{V_q}}{n}Y, \left(C_{V_q,V_q} + \nu I\right)^{-1}\left(C_{V_q,V_h}h - \frac{S^*_{V_q}Y}{n}\right)\right\rangle_{\mathcal{G}} \\
&= \left\langle C_{V_q,V_h}h - \frac{S^*_{V_q}}{n}Y, C^{-1}_{V_q,V_q,\nu}\left(C_{V_q,V_h}h - \frac{S^*_{V_q}Y}{n}\right)\right\rangle_{\mathcal{G}} \\
&= \left\|\hat{C}^{-1/2}_{V_q,V_q,\nu}\left(\hat{C}_{V_q,V_h}h - \frac{S^*_{V_q}}{n}Y\right)\right\|^2_{\mathcal{G}}
\end{aligned} \tag{L.1}$$

Where the last step is by the fact that the covariance is self adjoint and the first statement is shown. We now search for the optimum in h which is reached when such norm is zero:

$$\tilde{h} = \hat{C}^{-1}_{V_q,V_h}\frac{S^*_{V_q}}{n}Y$$

If one were to include an extra penalization term on the norm of h as $\lambda\|h\|^2_{\mathcal{H}} = \lambda\langle h,h\rangle_{\mathcal{H}}$, then equation (L.1) would become:

$$\left\|\hat{C}^{-1/2}_{V_q,V_q,\nu}\left(\hat{C}_{V_q,V_h}h - \frac{S^*_{V_q}}{n}Y\right)\right\|^2_{\mathcal{G}} + \lambda\|h\|^2_{\mathcal{H}} \tag{L.2}$$

By deriving in h we obtain:

$$2\left[\hat{C}_{V_h,V_q}\hat{C}^{-1}_{V_q,V_q,\nu}\left(\hat{C}_{V_q,V_h}h - \frac{S^*_{V_q}}{n}Y\right) + \lambda h\right] = 0 \tag{L.3}$$

Subsequently, equation (L.3) is optimized by choosing:

$$\tilde{h} = \left(\hat{C}_{V_h,V_q}\hat{C}^{-1}_{V_q,V_q,\nu}\hat{C}_{V_q,V_h} + \lambda I\right)^{-1}\frac{S^*_{V_q}}{n}Y = \left(S^*_{V_h}LS^*_{V_h} + \lambda I\right)^{-1}\frac{S^*_{V_q}}{n}Y$$

The loss can finally be rewritten as:

$$\begin{aligned}
\left\|\hat{C}_{V_q,V_q,\nu}^{-1/2}\left(\hat{C}_{V_q,V_h}h-\frac{S_{V_q}^*}{n}Y\right)\right\|_{\mathcal{G}}^2+\lambda\|h\|_{\mathcal{H}}^2 &= \left(\hat{C}_{V_q,V_h}h-\frac{S_{V_q}^*}{n}Y\right)^T\hat{C}_{V_q,V_q,\nu}^{-1}\left(\hat{C}_{V_q,V_h}h-\frac{S_{V_q}^*}{n}Y\right)+\lambda\|h\|_{\mathcal{H}}^2 \\
&= \left(\frac{S_{V_q}^*S_{V_h}}{n}h-\frac{S_{V_q}^*}{n}Y\right)^T\hat{C}_{V_q,V_q,\nu}^{-1}\left(\frac{S_{V_q}^*S_{V_h}}{n}h-\frac{S_{V_q}^*}{n}Y\right)+\lambda\|h\|_{\mathcal{H}}^2 \\
&= (S_{V_h}h-Y)^T\frac{S_{V_q}}{n}\hat{C}_{V_q,V_q,\nu}^{-1}\frac{S_{V_q}^*}{n}(S_{V_h}h-Y)+\lambda\|h\|_{\mathcal{H}}^2 \\
&= (h(V_h)-Y)^T\frac{L}{n}(h(V_h)-Y)+\lambda\|h\|_{\mathcal{H}}^2 \\
&= \lambda\left((h(V_h)-Y)^T\frac{L}{n\lambda}(h(V_h)-Y)+\|h\|_{\mathcal{H}}^2\right)
\end{aligned}$$

□

Remember now that a multivariate normal distribution $X\sim\mathcal{N}(\mu,\Sigma)$ has as density proportional to:

$$\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

Notice that this is similar to the construction of the quasi-likelihood as $\exp\left(-\frac{n}{2\lambda}d_n^2(\hat{E}_nh-\hat{b}\right))$ with the exponential function is then similar to a normal $\mathcal{N}\left(h(V_h),\lambda L^{-1}\right)$. This enables us to to make approximate inference on the posterior by considering a fictitious data generating process:

$$h\sim\Pi=GP(0,k_{V_h})\quad Y|h,\mathcal{D}^n\sim P_{\text{fic}}\sim\mathcal{N}\left(h(V_h),\lambda L^{-1}\right)$$

Using the fact that GP priors are conjugate to P_{fic} we can find a closed form solution for the posterior mean. Let $V_{h_t}=(_{V_h t})$ be test points, then the posterior follows

$$\begin{aligned}
\Pi_n(h(V_{h_t})|\mathcal{D}^n) = P_{\text{fic}}\left(h(V_{h_t})|Y\right) &= P_{\text{fic}}\left(m,S\right) \\
m &= K_{V_{h\,t},V_h}\left(\lambda I+LK_{V_h,V_h}\right)^{-1}LY \\
S &= K_{V_{h\,t},V_{h\,t}}-K_{V_{h\,t},V_h}L\left(\lambda I+LK_{V_h,V_h}\right)^{-1}K_{V_h,V_{h\,t}} \\
L &= K_{V_q,V_q}\left(n\nu I+K_{V_q,V_q}\right)^{-1}
\end{aligned}$$

We then use m as an estimator for the posterior mean and use it to construct the CERF estimator as in equation (5.11). Note that the fact that the quasi-likelihood has a form similar to a Gaussian does not mean that the data is distributed as such. Moreover, this is not the going to give us the $\hat{h}_{\Theta_m}$ mean but the quasi Bayesian posterior mean over the entire parameter space. Nonetheless, we use it in practice as it produces goodd numerical results(chapter 6).

6

Numerical Experiments

In this chapter we apply the proximal inference method to detect and quantify various causal relationships. First we showcase, on simple synthetic examples, the need for the causal framework and even more so one that accounts for unmeasured confounders. Later, we apply the methods studied in this thesis to two types of data.

6.1 The need to adjust for the unobserved

Here we introduce some examples with simulated data to showcase that regular regression does not return the desired parameters which justifies the need for a causal framework capable of adjusting for confounders. Afterwards, to go a step further, we apply the studied methods framework that accounts for unobserved confounders. Consider the following data generating process, where the data is generated sequentially and all errors are sampled independently $\epsilon_i \sim \mathcal{N}(0,1)$:

$$\begin{aligned} U &= 1 + \epsilon_U \\ W &= \frac{1}{2} \cdot U + \epsilon_W \\ Z &= -1 \cdot U + \epsilon_Z \\ A &= 10 + 0.8 \cdot Z - U + \epsilon_A \\ Y &= 3 + 2 \cdot A + 4 \cdot W + 4 \cdot U + \epsilon_Y \end{aligned} \tag{6.1}$$

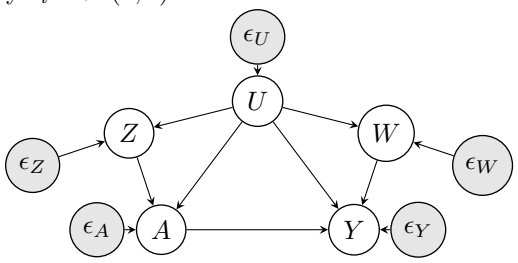


Figure 6.1: Graphical structure for additive linear model.

The average treatment effect, and thus the slope of the CERF, is given by the coefficient of the A term in equation (6.1). This becomes clear by applying proposition 3.1 to the given *true* model:

$$\begin{aligned} \chi(a) = \mathbb{E}\left[Y^a\right] &= \mathbb{E}\left[\mathbb{E}\left[Y|A=a,W,U\right]\right] \\ &= 3 + 2 \cdot a + 4 \cdot \mathbb{E}\left[W\right] + 4 \cdot \mathbb{E}\left[U\right] \\ &= \underbrace{9}_{\beta_0} + \underbrace{2}_{\beta_A} \cdot a \end{aligned}$$

We sample 10,000 data points from the model of equation (6.1) to obtain the following marginal distributions: Notice that by simply '*looking*' at figure 6.2, no clear trend is distinguishable. By assuming that

	$\hat{\beta}_0 \pm 1.96 \cdot SE(\hat{\beta}_0)$	$\hat{\beta}_A \pm 1.96 \cdot SE(\hat{\beta}_A)$
$Y\|A$	27.14 ± 0.41	-0.211 ± 0.047
$Y\|A,W,Z$	16.49 ± 0.37	0.777 ± 0.039

Table 6.1: Estimated coefficients for naive and classic causal regression.

only A affects Y and applying a naive regression one obtains the red slope. As presented in table 6.1, the estimated slope not only is far from the true value but also presents the opposite sign. One would

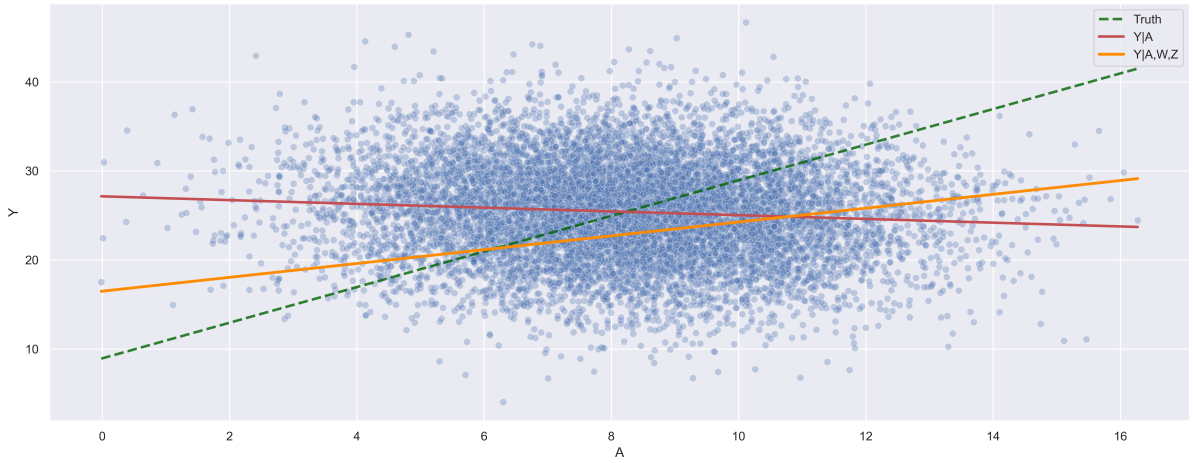


Figure 6.2: Samples from model the above linear confounded model and estimated regressions. Vertical axis is the outcome (Y), and horizontal axis is the treatment (A). The regresssion functions: Naive regression $Y|A$ (red), *classic* causal regression $Y|W, Z, A$(orange), and the truth(green).

then conclude that higher doses of the treatment A would lead to a worse outcome. Since we are able to observe the proxy variables, the next step is to assume that (Z, W) are measured confounders. In this case the results are slightly more promising since the correct sign is well identified. Nonetheless, the slope is severely underestimated as the true β_A does not fall within the 95% confidence interval. This simple example highlights the need for a framework that adjusts for both observed and unobserved confounders.

6.2 Simulated Data

Here we present numerical results on synthetic generated data. In the first case, the data is determined as linear combination of Gaussians with potential nonlinearities arising in the structural equation of Y. For a discussion on how the parameters interact when sequentially generated, see appendix A.1.1. In the second case, we explore data generated from more exotic non-linear processes.

6.2.1 Gaussian Models

Linear

Similarly to the previous example of figure 6.6, consider the following data generating process:

$$\begin{aligned} U &= 1 + \epsilon_U \\ W &= 1 - 4 \cdot U + \epsilon_W \\ Z &= -1 + 2 \cdot U + \epsilon_Z \\ A &= 6 + 2 \cdot Z - U + \epsilon_A \\ Y &= \frac{2 + 2 \cdot A + 5 \cdot W - 2 \cdot U + \epsilon_Y}{10} \end{aligned} \tag{6.2}$$

The graphical model remains the same as the one in figure 6.1. We sample two sets of data: one of size 1,000 and one of size 10,000. Simply observing figure 6.3, the trend seems to be negative and thus one would estimate that the treatment has a negative effect on the outcome. Fitting a naive regression, i.e. $Y|A$ results in a similar conclusion. Considering the proximal structure associated with the data generating process, we fit P2SLS and the PQB method. Both methods well approximate the true CERF. The P2SLS(orange) has an easier time as it is constrained to a linear parametric model but PQB also performs well in both situations. Both methods improve with the increase of number of samples.

Parabolic

It is often much more realistic to model natural phenomena as parabolic, or some other polynomial rather than simply linear. To test the performance of the models in such scenarios, data is generated as a linear combination of Gaussians, except for the outcome structural equation which becomes quadratic in U A

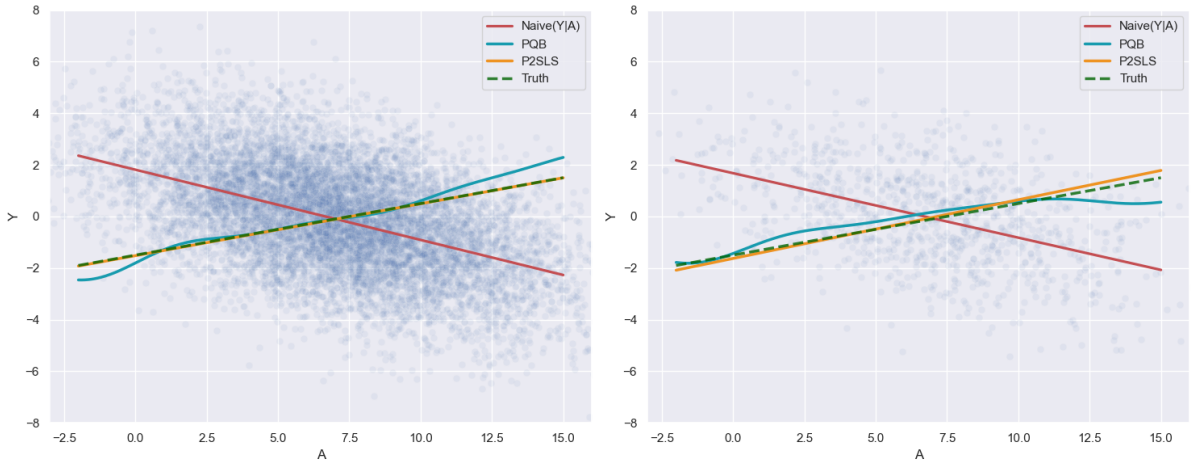


Figure 6.3: Linear example with fitted P2SLS and PQB estimators. 10,000 data points on the left and 1,000 on the right

and X. The graphical model becomes *fully proximal* as it now also contains measured confounders and is pictured in figure 6.1. In this case, we set all additive noise $\epsilon_\cdot \sim \mathcal{N}(0, 0.2)$.

$$
\begin{aligned}
U &= 1 + \epsilon_U \\
X &= \frac{1}{2} \cdot U + \epsilon_X \\
W &= -X + 2 \cdot U + \epsilon_W \\
Z &= 2 \cdot X - 2.5 \cdot U + \epsilon_Z \\
A &= 6 - 3 \cdot X + 2 \cdot Z + U + \epsilon_A \\
Y &= \frac{12 + 4 \cdot A - A^2 - 6 \cdot X^2 - 4 \cdot U - 3 \cdot U^2}{10} + \epsilon_Y
\end{aligned}
$$

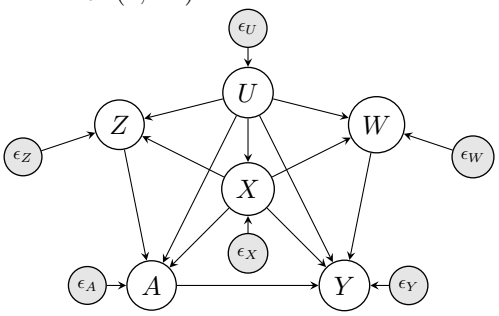


Figure 6.4: Graphical structure with additive errors and measured confounder.

The presence of the measured confounder, and that of higher confounding terms in the structural equation of Y makes the model much more complicated than that of the first linear example. Notice that by the g-formula, the True CERF is given by:

$$
\mathbb{E}\left[Y^a\right] = \mathbb{E}\left[\mathbb{E}\left[Y|A=a,W,U\right]\right] = \frac{12 + 4 \cdot a - a^2 - 6 \cdot \left(\frac{1}{4} + \frac{5}{4}\sigma^2\right) - 4 - 3 \cdot \left(1 + \sigma^2\right)}{10} = \frac{2.6 + 4 \cdot a - a^2}{10}
$$

We satisfy the assumptions needed for the Higher Order P2SLS and, in this situation, both the order of confounding and of treatment effect is quadratic. For this reason we fit a higher order proximal 2SLS with order 2 treatment and order 2 confounding. Additionally, we fit the proximal quasi bayesian method. The results are similar to the linear example. The naive regression $Y|A, A^2$ misses the truth by quite a large margin. In both samples the HigherOrder estimator is the better one due to the innate regularization of the parametric model. Nonetheless, the PQB estimator manages to match the shape and curvature of the true CERF in data dense neighborhoods while faltering at the edges. Notice that the P2SLS method is not as *precise* as it was in the linear case. This is because fitting a polynomial is more difficult, as it is much more sensitive to outliers. Nonetheless, as the sample size increase, both methods approach the truth.

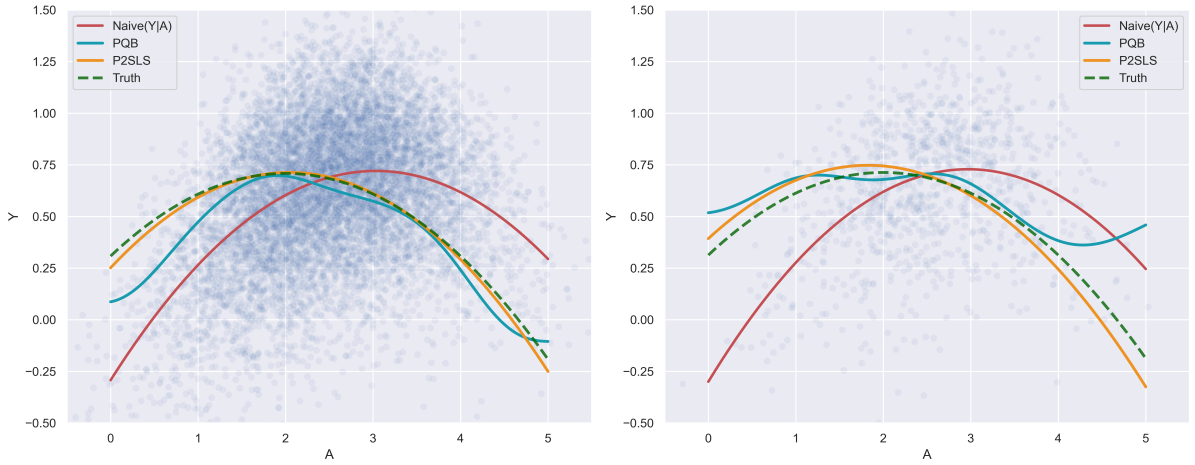


Figure 6.5: Scatter plot of sampled data points, naive regression with quadratic terms(red), higher order P2SLS(orange), PQB(teal) and the true CERF(green). On the left 10,000 data points and on the right 1,000.

6.2.2 Non-Linear Simulation: Demand Experiment

The Demand experiment was first introduced in the DFPV paper[69] and has since become a staple benchmark for proximal causal experiments. The fictitious problem is that of determining the causal relationship between airplane ticket sales(Y) and ticket price (A), where these are confounded by a potential demand (U). The negative control exposure are two fuel prices (Z_1, Z_2), one would expect that as the price of fuel costs increase, so must the ticket price. The latter are modeled using periodic functions to simulate a seasonal trend. The negative control outcome is instead chosen to be the number of website views the airplane company receives(W). Although nowadays companies adjust prices based on this, which would generate a direct path from W to A that violates assumption 1.4, in a fair market this would be a valid assumption. The underlying data generating process is the following:

$$
\begin{aligned}
U &\sim Unif(0, 10) \\
W &= 45 + 7 \cdot g(U) + \epsilon_W \\
Z_1 &= \sin(2\pi U/10) + \epsilon_{Z_1} \\
Z_2 &= \cos(2\pi U/10) + \epsilon_{Z_2} \\
A &= 35 + (Z_1 + 3) \cdot g(U) + Z_2 + \epsilon_A \\
Y &= A \cdot \min(e^{\frac{W-A}{10}}, 5) - 5 \cdot g(U) + \epsilon_Y
\end{aligned}
\tag{6.3}
$$

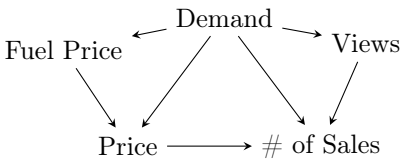


Figure 6.6: Graphical representation for the data generating process of the Demand problem.

where $\epsilon_\cdot \sim \mathcal{N}(0, 1)$ and $g(u) = -4 + 2 \cdot \left(\frac{(u-5)^4}{600} + e^{-4 \cdot (u-5)^2} + \frac{u}{10} \right)$. Notice that while there are additive errors, they are not propagated through the *network* in a linear manner as each structural function is highly non-linear. The data so generated gives rise to a peculiar joint distribution.
The true causal exposure response function is found using the backdoor adjustment formula, but due to the complex non-linearities determining the closed form of χ is extremely complicated. Rather, a Monte Carlo approach is carried out on a large simulated dataset such that:

$$
\chi(a) = \mathbb{E}\left[\mathbb{E}\left[Y|W, a, U\right]\right] \approx \frac{1}{n} \sum_{i=1}^{n} a \cdot \min(e^{\frac{w_i - a}{10}}, 5) - 5 \cdot g(u_i)
$$

The only discernible trend between price and number of sales from examining the scatter plots in figures 6.7 and 6.8 is a negative slope as the price increases, which is to be expected. In practice one would be interested in determining the price that causes the most sales, the peak of the CERF in figure 6.8. Although much of the data is concentrated around the mean of A (≈ 27.3), the authors of [69, 30] only focus on the interval $[10, 30]$. There is no clear reason why since the estimated median lies around 28.5, and thus almost half the data will be lying outside the considered interval. More results on the extended

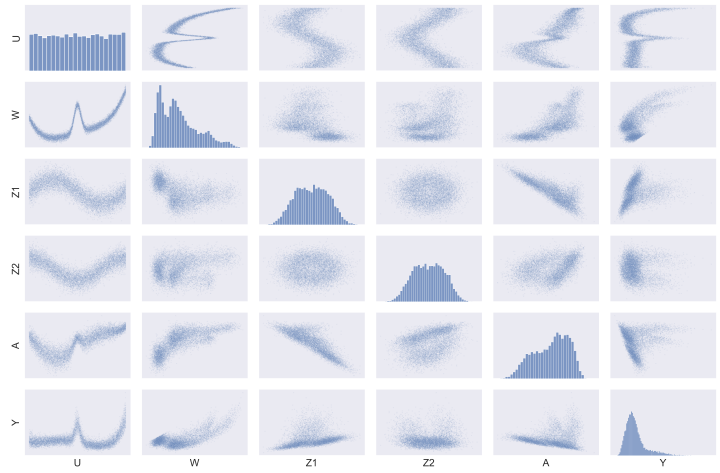


Figure 6.7: Pairwise joint distribution between 10,000 variables sampled from the Demand data generating process.

Figure 6.8: True CERF χ of the Demand problem and scatter plot of sampled data.

interval are included in the Appendix. The ProxySplines method is not considered since interest is placed only on the first half of the distribution. The evaluation metric for the problem is the causal mean square error (c-MSE).

Definition (Causal Mean Square Error)
The causal mean square erroris defined as:

$$c\text{-}MSE = \|\chi - \hat{\chi}\|_2^2 = \mathbb{E}\left[(\chi(A) - \hat{\chi}(A))^2\right]$$

Given a test dataset $\mathcal{D}^t$ of size n_t, the sample equivalent is:

$$c\text{-}MSE_t = \frac{1}{n_t} \sum_{i \in \mathcal{D}^t} (\chi(a_i) - \hat{\chi}(a_i))^2$$

In practice the authors of [69, 30] only evaluate their function on 10 points within the range [10,30]. Apart from being a conceptual mistake to assume that the 'clipped' distribution is the same as the *unclipped* one, 10 points is by no means a valid test set. Maybe this was done to simulate a *real-life* scenario, although it might as well have been because some methods are not as performant away from the mean as will be seen in the upcoming simulations. For fairness of comparison we compute both the c-MSE$_{10}$ error, the one used in the previous papers, and the c-MSE$_{1k}$, the empirical c-MSE on 1,000 points randomly sampled from the *unclipped* distribution. The interquantile range of the errors is also considered as a measure of performance for the models in appendix A.4.

Hyper Parameter Choice

The choice of hyper parameter for all existing methods is taken from the original papers. This is assumed to be optimal hyper parameter selection. Additional choices such as kernel, and network architectures for the deep learning methods are directly taken from the papers. [69, 30] both run grid search for the selection of the optimum architecture and network parameters. The NMRR network architecture is a fully connected network with ReLU activation, 4 layers deep for the U-statistic, 3 layers for the V-statistic, and 80 neurons per layer. The inputs of the network are the sample $(Y - V_{h_i}), (Y - V_{h_j})$ and the evaluation of $k(V_{q_i}, V_{q_j})$ where k is the Gaussian kernel. The optimizer is Adam[28] with weight decay regularization fixed to 3e-6, learning rate 3e-3 and the network is trained for 3,000 epochs even though convergence is often reached much earlier.
The DFPV method assigns a neural network to each variable. All networks share a similar structure with only two layers with 32 and 16 nodes respectively. Since there are two negative controls, the specific network presents two inputs while all others have 1 input. The output is 8 dimensional. Linear combinations are then used to perform the linear regressions. The regularization parameters for each regression are both set to 0.01. The learning rate is not specified and automatically decided by the Adam optimizer[28]. The weight decay is instead set to 0.1 for all networks.
Moreover, it is important to notice that methods that require a held-out dataset to evaluate the CERF, do so by splitting the data in half. This means that if the number of samples considered is 10,000 the P2SLS method will use all data points to estimate the CERF whereas the KPV method will use 5,000 to fit the model and 5,000 to evaluate $\hat{\chi}$. When working with kernels involving multiple variables, we will specify the kernels for each individual variable and combine them as the product of the marginal kernels:

$$k\left((x_1, y_1), (x_2, y_2)\right) = k(x_1, x_2) \cdot k(y_1, y_2)$$

This is possible because the Hilbert space $\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}$ is isometrically isomorphic to $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$[36].
For the Quasi-Bayesian method, the kernels used are the Gaussian Kernels with the parameters selected using the method suggested in [63] which is taken from the KPV method[36]. This resulted in regularization parameters $\nu = 0.001$ and $\lambda = 0.004$. For the Higher Order Proximal 2SLS, we decide to fit a linear in A . Although the structural function g in equation (6.3) contains an exponential term in U, which could be interpreted as '*infinite order*' when seeing it in terms of Taylor expansions; we decide to stop at 4^{th} since it is also present in g and e^{-x^2} quickly vanishes.

Results

The experiments were carried out over three sample sizes to evaluate the model performances in different situations. The tested sample sizes were 10,000 5,000 and 1,000. The graphs below show the mean of the model and the $5\% - 95\%$ covered in the 20 runs over different seeds. From the figures 6.9 to 6.11, it is possible to notice that all models perform similarly where the data is highly concentrated (around the median of 28). Moving away from that area we notice that all models start to deteriorate, as expected. The best result is given by our model PQB, which not only comes closer to the truth in data-dense areas but also further from the median.

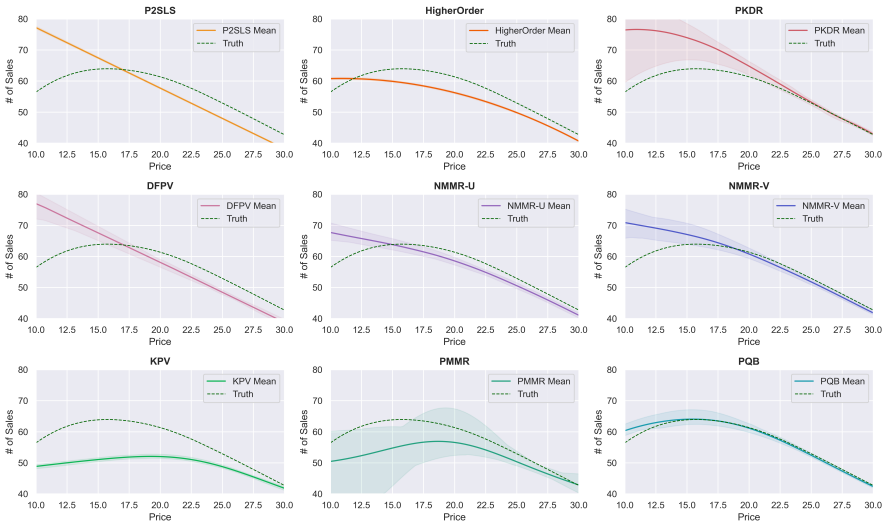


Figure 6.9: Individual comparison of the different methods trained on 10,000 data points from the Demand problem over 20 simulated runs.

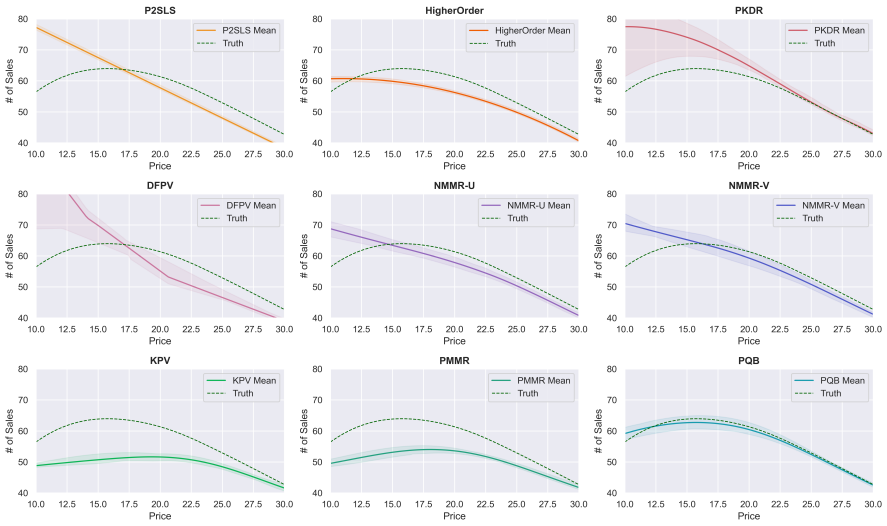


Figure 6.10: Individual comparison of the different methods trained on 5,000 data points from the Demand problem over 20 simulated runs.

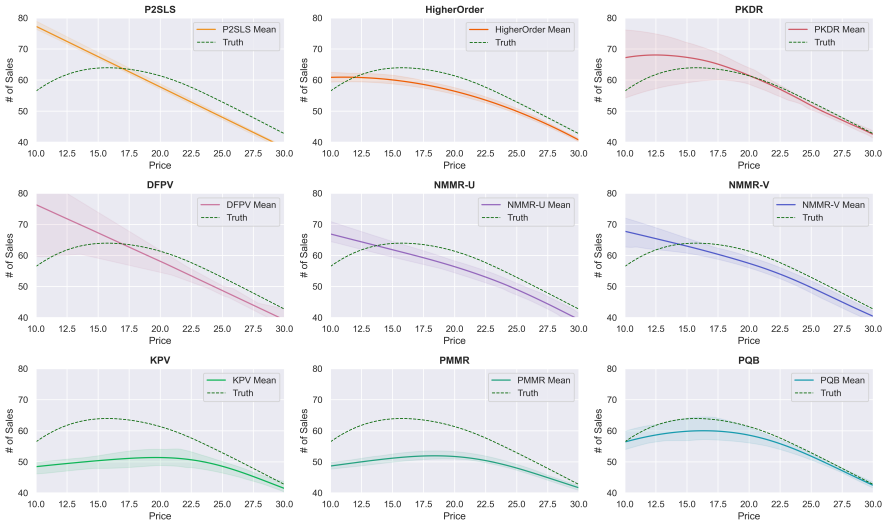


Figure 6.11: Individual comparison of the different methods trained on 1,000 data points from the Demand problem over 20 simulated runs.

The method also manages to capture the curvature of the true χ in places where the data is also less concentrated, whereas the second best performing method NMMR does not. The performance of PQB over other models is confirmed by both CMSE metrics and the sampled error quantiles reported in table 6.2 (and for c-MSE$_{10}$ in table A.1). The PQB method has the lowest mean of the CMSE errors and also the lowest variance. The graphs also highlight how the method has the tightest quantile bounds around the 20-30 range and only deteriorates in the tail. Another error metric is included in appendix A.4.

Model	**c-MSE**$_{1k}$ ($train_n = 10,000$)	**c-MSE**$_{1k}$ ($train_n = 5,000$)	**c-MSE**$_{1k}$ ($train_n = 1,000$)
P2SLS	24.93 ± 2.89	24.88 ± 3.49	25.06 ± 6.51
HigherOrder	17.46 ± 2.37	17.48 ± 4.23	17.69 ± 5.15
PKDR	16.58 ± 15.83	16.96 ± 16.69	5.96 ± 9.30
DFPV	21.95 ± 10.61	19.91 ± 15.10	25.96 ± 18.17
NMMR-U	5.46 ± 4.32	8.17 ± 7.86	19.29 ± 28.12
NMMR-V	4.05 ± 3.48	5.29 ± 4.36	10.11 ± 10.73
KPV	34.83 ± 6.55	37.99 ± 10.93	41.31 ± 25.12
PMMR	39.12 ± 96.09	26.20 ± 12.07	38.45 ± 11.02
PQB	$\mathbf{0.75 \pm 1.13}$	$\mathbf{1.10 \pm 1.27}$	$\mathbf{4.36 \pm 6.68}$

Table 6.2: Model performance using Causal Mean Error on 1,000 test points.

Decreasing the number of samples used to train shows that performance deteriorates substantially on the outliers of the distribution. The odd result of PKDR is that it performed better . This is likely due to the automatic hyperparameter selection occurring at each run. The authors do not specify the procedure and might suffer from outliers. Similarly, PMMR has odd results in the large training setting, likely due to numerical problems while inverting a large matrix. Nonetheless the PQB method recovers the shape of the curve, as can be observed in figures 6.9 to 6.11.

6.3 Case Study - Sustainable Causal Investing

In recent years, two key investment strategies have gained significant traction: Sustainable Investing and Causal Investing. Sustainable or green investing has seen a notable increase in popularity as both individual investors and institutions recognize the importance of addressing global challenges such as climate change, resource depletion, and social inequality. Many investment funds have started integrating ESG (Environmental, Social, and Governance) criteria into their investment strategies. ESGs are used to evaluate the sustainability and societal impact of investments in companies or assets. Unfortunately, there are no universally adopted standards on how these ratings are produced and evaluated, leading to inconsistencies between different rating providers. Nonetheless, green assets are seen as more profitable in the short term when compared to regular investments [32]. This has led many companies to 'greenwash' their image to appear more environmentally friendly.

Similarly, following the causal revolution, a lot of interest has circled around quantifying causal relationship between financial variables rather than mere correlations. In traditional financial analysis, correlation is often used to predict asset performance. Causal investing seeks to quantify the true causal relationships in the market to achieve better performance. By identifying these causal links, investors can make more informed choices and targeted investments.

Similar to the previously discussed Demand experiment, modeling financial markets involves many unobserved confounders such as demand and sentiments. This makes it unreasonable to apply the conditional exchangeability assumption of *classic* causal inference. Instead, we might collect proxies to adjust for these unobserved factors. For this reason, the proximal inference framework is a perfect fit. The research question of interest combines causal and sustainable investing to answer the following question:

Does investing in fossil fuel companies have a causal effect on the return?

To answer this question, we will consider data from investment funds between 31-01-2024 and 31-05-2024. The data, obtained from [17], contains detailed information on how the capital of these funds is allocated along with various sustainability scores. We choose the treatment to be the percentage of fund assets invested in fossil fuel holdings (A) at the beginning of the study and the outcome is percentual increase in market asset value(Y) at the end of the study. The underlying market conditions and sentiments are considered the unmeasured confounder(U). The negative treatment variable (Z) is chosen to be the rating assigned to the ticker at the beginning of the study (American grading system A-F). The idea behind this is that it represents the sentiment or propensity of the asset manager to invest in fossil fuel holdings. The outcome proxy (W) is chosen to be the average return for the funds in the same category. In practice these assumptions might be too simplistic as the period of the study is long and the funds alter their capital consistency over said period. Moreover, market dynamics are complicated and unlikely to be well modeled by just a few variables. Nonetheless, it is interesting to apply the proximal method to such a problem.

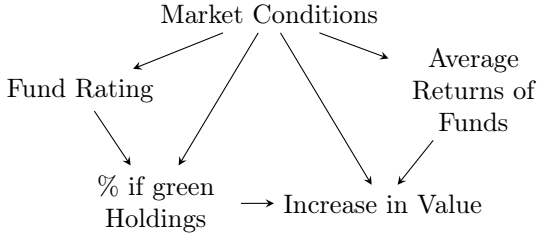


Figure 6.12: Graphical model for the Sustainable Causal Investing example.

The Data

We have a total of 3546 samples. The distribution of the funds within the categories and fossil fuel rating is the following:

Category	Count
U.S. Equity Fund	1655
International Equity Fund	831
Allocation Fund	763
Sector Equity Fund	297

Table 6.3: Number of funds by category.

Category	Count
A	324
B	281
C	722
D	1246
F	973

Table 6.4: Number of funds by category.

Having seen that the PQB method works well in the low number of samples case for the Demand experiment, we decide to split the data in three parts: 1000 samples to decide the hyper parameters of the model, 1273 for training and the remaining 1273 for the prediction. Applying the same procedure as before we obtain regularization parameters of $\lambda = 0.1$ $\nu = 20.3$. We also fit a simple P2SLS and a Higher order P2SLS with order 2 treatment and linear confounding.

Results

In this case we do not have an estimate of the true CERF as the data is real observational data. The estimated CERF by the three considered methods is reported in figure 6.13. The simpler P2SLS, and

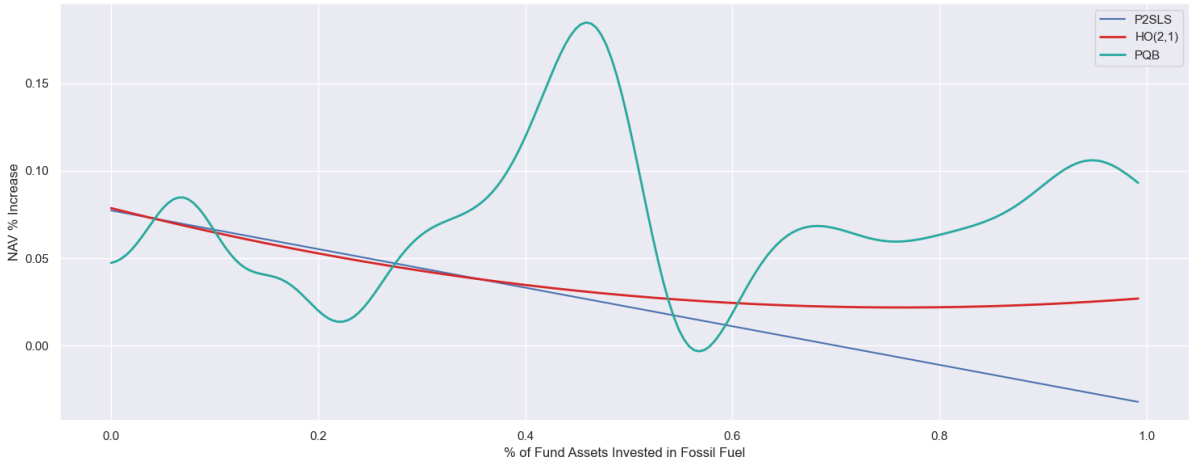


Figure 6.13: Various method on the considered ESG dataset.

Higher Order model indeed recognize a trend but might be too simple. With respect to these results, it seems that investing more into fossil fuels is not as profitable as staying green. The second order treatment recognizes some curvature in the later part of the CERF. The PQB method also captures an initial downward tren but then captures a large peak at the center and trails off at the sides. This seems to suggest that the CERF achieves its maxima around the half way point, and thus the optimal choice is to allocate half of funds to fossil fuel.

6.4 Case Study - Effect of Exercise on Sleep

In 1997, the authors of [47] performed a 10 week randomized control trial on a group of elderly participants to determine the effect of exercise on subjective sleep quality and depression. Results highlighted that physical exercise had a significant positive effect on the latter. A more recent study of 2020 [58] also shows similar results on a middle age demographic. Unfortunately, as the authors of [16] point out, stress and anxiety can act as confounders between treatment and outcome. The first two studies only measure subjective sleep quality assessed with a questionnaire. Various studies [1, 7] have found that longer Slow Wave Sleep(SWS), commonly known as deep sleep, is highly correlated with better subjective sleep. Moreover, the scientific literature seems to agree that exercise leads to increased amounts of deep sleep [71]. The best way to measure this quantity is through a polysomnography, a sleep test. Unfortunately these laboratory experiments involve complicated measuring tools and could introduce anxiety, which would in turn generate biased results. Nowadays, wearable technology enables users to measure and evaluate both sleep quality and physical activity. This generates an abundance of observational data that overcomes the previously mentioned laboratory induced confounding, which could be used to study the causal effect of physical exercise on sleep. Additionally, users can easily log their psychological well being through various apps. As discussed in section 1.3, questionnaires about psychological states are not direct measures, but rather proxies, for the underlying conditions.

Similarly to other studies [16, 22, 47, 58] we aim to answer the following research question:

What is the causal effect of exercise on deep sleep?

The main difference from the existing studies is that this is a study based on purely observational data, we quantify the effect of exercise on deep sleep and not a subjective measure of it, and the data is collected without affecting daily routines.

Data and Structure

The Simula PMData sports dataset[55], is an open real-life dataset which has collected daily health data from 16 participants over the period of 5 months. The participants collected data in two ways: Fitbit Versa 2 smartwatch wristband to measure heart rate, sleep and exercise, and questionnaires to evaluate the lifestyle choices. This dataset is a perfect example of observational data; there is no experimental set-up to guarantee any randomization or guarantee the capture of all confounders.

The treatment variable (A) is set to the *intense exercise time* and outcome (Y) the *minutes of deep sleep*. The unmeasured confounder (U) directly affecting both is *Stress*. This may be either physiological or psychological stress. Both types of stress have an effect on both sleep time and whether or not one exercises. For negative control exposure (Z) we consider the answers associated with the physical stress questionnaire: *soreness* and *fatigue*. Notice that this variable is the perception of physical stress and does not quantify the true stress. As such, it will influence the amount of exercise performed but should not have an effect on total amount of time spent sleeping. Moreover, since the questionnaire is answered at the beginning of the day, it is guaranteed that the exercise time (A) does not have an effect on fatigue(Z) since a effect cannot occur before a cause. For negative control outcome (W), we consider the *average amount of deep sleep* in the previous week. We expect this variable to be a proxy for overall psychological stress and also account for personal basis of deep sleep. The only effect (W) could have on (A) is with a backdoor path through (U).

Additionally, to characterize the baseline fitness level of the individual, we decide to include a measured confounding (X). This variable will consist of the *average weekly exercise*. In practice we expect this to affect (A), as higher fitness individuals are more likely to train; but also (Z) (W) and (Y).

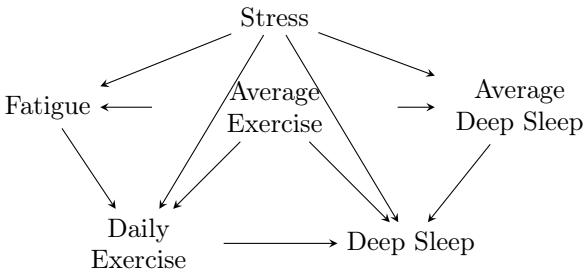


Figure 6.14: Graphical representation for the sleep case study.

Figure 6.15: Joint density of observed variables for the sleep experiment.

After combining the various data repositories, we obtain a total of 1478 samples. The (Z) variables have discrete support as they are answers to a 5 point likert scale. Even though the the smartwatch only measures the remaining variables with a precision of 1 minute, we assume that the support of (W),(X),(A) and (Y) is $\mathbb{R}^+$. The joint distribution of the observable variables can be found in figure 6.15.

Results

To answer our question of interest, we fit two models: Higher Order P2SLS and Proxy Quasi Bayesian. Seen the success of PQB on small datasets, we decide to train the hyper parameters of the model on a dataset of size 478 and use the remaining 1,000 points to fit the model. The optimum parameters found are $\lambda = 1.2$ and $\nu = 0.67$. For the Higher Order P2SLS we decide to fit a model with $(A^i, X^j, U^k) \quad (i, j, k) = (2, 2, 2)$. This is because we would expect the effect to be somewhat parabolic, an increase until a certain point and then detrimental.

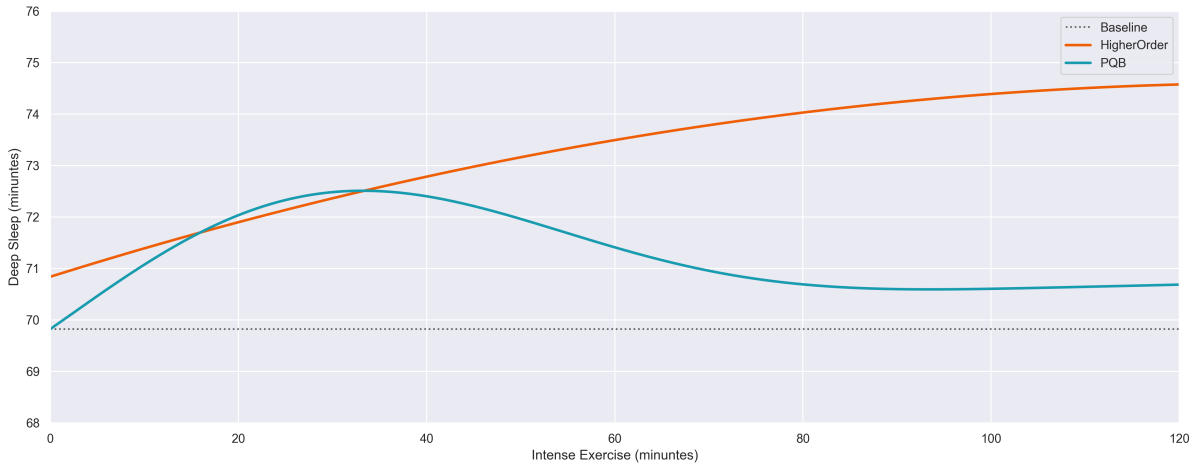


Figure 6.16: Estimated CERF for the sleep experiment.

Figure 6.16 contains the results of the two estimated CERFs. The PQB estimator highlights that intense physical exercise has a total positive effect on SWS time, but this effect deteriorates after a certain amount and plateaus as the exercise session becomes longer. This can be explained by the fact that too much exercise tires out the the subject too much, in other words excessive exercise hinders sleep quality [4]. The results form the HigherOrder model also show a positive effect that then seems to plateau, but keeps increasing without stopping. The result may be less reliable as it does not have regularization to penalize and might overfit to areas where data is more scarce. Nonetheless, both results confirm the existing knowledge on the effect of physical exercise on deep sleep[16, 71].

A small note on the assumptions

The case study should not be taken as definitive medical evidence, but rather showcase a potential application of the proximal inference framework. The decisions made here were simply based on common sense and personal experience and do not have expert medical justification. As observed by [29], the

relationship between sleep and exercise seems to be bidirectional. Indeed higher amounts of exercise lead to improved sleep and higher amounts of SWS, but the converse seems to also have an effect. Low quality sleep leads to lower physical activity levels. The chosen structure for this experiment does not consider this. Moreover, it completely ignores the fact that the data also has a temporal component to it. There are methods in the proximal inference framework that would allow for estimation of causal quantities in such a setting; for more details [53, 41]. Additionally, we assume that the data is i.i.d sampled even though this is not exactly the case due to the fact that we are using the measurements of the same individuals over different days. This is somewhat corrected by the fact that W and X are the individual average results. Interestingly, we started off wanting to measure sleep quality and ended up using amount of SWS as outcome. Note that this quantity is not sleep quality in it of itself, but rather one possible cause. Further investigations could explore a different structure where SWS time is a cause of sleep quality. Another key refinement would be to measure both exercise time but also intensity. In the current experiment, the intensity is measured by the heart rate at which it is performed but some exercises may be more exhausting than others.

7

Conclusion

The causal revolution along with the greater availability of data have opened many doors for causal quantities to be determined. In chapter 1, we introduced the history of causality in statistics and the scientific disciplines. The g-formula enables us to transform observational data into experimental and make inference on the causal quantities. The most questionable assumption of the classic framework is the conditional exchangeability assumption, which ensures that treatment is assigned at random given the covariates. This is where the topic of proximal causal inference comes in, which allows more flexibility compared to the classical framework by requiring proximal exchangeability. The specific structure of negative control then enables causal identification even in the presence of unmeasured confounders. Nonetheless, we must be critical as no approach is an all end all solution, but rather a delicate balance of assumptions and objectives. The proximal framework trades off the additional complexity of bridge functions for the capacity to adjust for the unobserved. The extra requirements of completeness and structure are not always guaranteed to hold, and depend on the availability of the data. Even so, as discussed in chapter 2, the characterization of the bridge functions as solutions to ill-posed problems makes it troublesome to correctly estimate them. The non-invertibility of the conditional expectation operator makes it non-trivial to determine the solution. Additionally, the operator needs to be estimated from the data.
In chapter 3 we studied the state of the art, and found that it is often difficult to balance theoretical complexity with practical needs. If one prefers the latter, one should consider linear parametric models, which can become more flexible and better adapt to real world situations with the Higher Order methods introduced in chapter 4. Otherwise, if the proximal structure of the variables is the only knowledge available about the problem, then a different non-parametric approach studied might be more suitable. One could also choose to use the more novel neural network based approach but, as of currently, would lose the stronger theoretical guarantees associated with KPV,PMMR,PKDR. The first two require well posedness in terms of powers of RKHS spaces and uniqueness assumptions and result in point wise error bounds for the CERF estimator. The PKDR approach leverages both the outcome and exposure bridge function to construct an estimator. Under additional assumptions on the critical radii of the function class considered and the rate of convergence of the bridge estimates, the ATE estimator converges at rate $\sqrt{n}$. As seen in section 3.3, the ATE estimator can be smeared over continuous data using kernels as in kernel density estimation. In such case, at the optimal trade-off between bias and variance, the best obtainable rate is $n^{-\frac{2}{5}}$.
In chapter 5 we introduced the proxy quasi-Bayes method, which presents strong theoretical guarantees under minimal assumptions and good empirical performance as later observed. The rate of convergence of the estimator depends on the ill-posedness measure of the problem. Additionally, if it were reasonable to make smoothness assumptions on the bridge function, the PQB method can match them by choice of an appropriate kernel (often Matérn to match order of continuity). In chapter 6, the newly formulated approaches were tested on synthetic data, generated to match the assumptions and on real world cases. In the latter case, the findings are in line with the existing literature on the topic.

7.1 Future work

In the middle of writing of this thesis, computer scientists became extremely excited about a *new* network architecture that would, in their opinion, revolutionize the world, Kolmogorov Arnold Networks. The authors attempted to provide a theoretical framework to justify the new architecture but, unfortunately, was not well theoretically sound, and interest in the topic quickly died down. Nonetheless, an attempt

was made to deploy the architecture for NMMR and DFPV procedures. Unfortunately the results were not successful and were outperformed by the existing architectures. However, it is worth noticing that the architecture in the numerical experiments was chosen through cross validation. For a fully connected network with the same width at each layer, the variables at play are *network depth*, *network width*, *learning rate* and *weight-decay*. This means if each *grid* of testing parameters is n, we would have to check a total of n^4 iterations. This is already computationally impractical. If additionally we want to test for a non-fully connected architecture, or uneven width within the layers, the problem becomes much worse.

The proximal framework is heavily reliant on the specific structure of the negative controls. One might argue that its ability to adjust the unobserved is similar to that of instrumental variables, the difference being the structure. Thus, other more exotic structures might enable identification in the presence of unmeasured confounders under different assumptions. Moreover, the majority of the work has been done for binary treatment and Average Treatment Effect rather than the CERF studied in this thesis. The door is open for many new causal approaches.

Bibliography

[1] Torbjoorn Aakerstedt et al. "Good sleep—its timing and physiological sleep characteristics". In: *Journal of sleep research* 6.4 (1997), pp. 221–229.

[2] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program". In: *Journal of the American Statistical Association* 105 (June 2010), pp. 493–505.

[3] *All Prizes in Economic Sciences*. NobelPrize.org, 2018. URL: https://www.nobelprize.org/prizes/lists/all-prizes-in-economic-sciences/.

[4] Mohammad A. Alnawwar et al. "The Effect of Physical Activity on Sleep Quality and Sleep Disorder: A Systematic Review". In: *Cureus* ().

[5] Brewers Association. *National Beer Sales and Production Data | Brewers Association*. Brewers Association, 2022. URL: https://www.brewersassociation.org/statistics-and-data/national-beer-stats/.

[6] Andrew Bennett et al. *Inference on Strongly Identified Functionals of Weakly Identified Functions*. 2023. arXiv: 2208.08291 [stat.ME].

[7] Michael H Bonnet. "Sleep restoration as a function of periodic awakening, movement, or electroencephalographic change". In: *Sleep* 10.4 (1987), pp. 364–373.

[8] Peter Burgisser, Michael Clausen, and Mohammad Amin Shokrollahi. "Problems Related to Matrix Multiplication". In: *Algebraic Complexity Theory: With the Collaboration of Thomas Lickteig*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, pp. 425–453. ISBN: 978-3-662-03338-8. DOI: 10.1007/978-3-662-03338-8_16. URL: https://doi.org/10.1007/978-3-662-03338-8_16.

[9] Marine Carrasco, Jean-Pierre Florens, and Eric Renault. "Chapter 77 Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization". In: ed. by James J. Heckman and Edward E. Leamer. Vol. 6. Handbook of Econometrics. Elsevier, 2007, pp. 5633–5751. DOI: https://doi.org/10.1016/S1573-4412(07)06077-1. URL: https://www.sciencedirect.com/science/article/pii/S1573441207060771.

[10] Xiaohong Chen and Demian Pouzo. "Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals". In: *Econometrica* 80.1 (2012), pp. 277–321. DOI: https://doi.org/10.3982/ECTA7888. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA7888.

[11] David Colton. "Inverse Acoustic and Electromagnetic Scattering Theory, Second Edition". In: *Inverse Problems* 47 (Jan. 2003).

[12] Jerome Cornfield et al. "Smoking and lung cancer: recent evidence and a discussion of some questions". In: *Journal of the National Cancer Institute* 22.1 (Jan. 1959), pp. 173–203.

[13] Yifan Cui et al. *Semiparametric proximal causal inference*. 2023. arXiv: 2011.08411 [stat.ME].

[14] *Digest of Education Statistics, 2022*. nces.ed.gov. URL: https://nces.ed.gov/programs/digest/d22/tables/dt22_321.10.asp.

[15] Nishanth Dikkala et al. "Minimax Estimation of Conditional Moment Models". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 12248–12262. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/8fcd9e5482a62a5fa130468f4cf641ef-Paper.pdf.

[16] Helen S. Driver and Sheila R. Taylor. "Exercise and sleep". In: *Sleep Medicine Reviews* 4.4 (2000), pp. 387–402. ISSN: 1087-0792. DOI: https://doi.org/10.1053/smrv.2000.0110. URL: https://www.sciencedirect.com/science/article/pii/S1087079200901102.

[17] Fossil Free Funds. https://fossilfreefunds.org/. (Visited on 07/26/2024).

[18] AmirEmad Ghassami et al. *Minimax Kernel Machine Learning for a Class of Doubly Robust Functionals with Application to Proximal Causal Inference*. 2022. arXiv: 2104.02929.

[19] Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017.

[20] Franco Di Giuseppe. "IL CONDIZIONALE PROBABILISTICO DI R. STALNAKER". Tesi di laurea, Corso di laurea in Filosofia. Relatore: Prof. Maria Luisa Dalla Chiara.LTTL1990000000025. Firenze, Italy: Università degli Studi di Firenze, 1990.

[21] Kjetil B. Halvorsen. *Gaussian processes, sample paths and associated Hilbert space*. MathOverflow post. 2011. URL: https://mathoverflow.net/questions/59739/gaussian-processes-sample-paths-and-associated-hilbert-space.

[22] J.A. Horne. "The effects of exercise upon sleep: A critical review". In: *Biological Psychology* 12.4 (1981), pp. 241–290. ISSN: 0301-0511. DOI: https://doi.org/10.1016/0301-0511(81)90001-6. URL: https://www.sciencedirect.com/science/article/pii/0301051181900016.

[23] Jie Kate Hu, Dafne Zorzetto, and Francesca Dominici. *A Bayesian Nonparametric Method to Adjust for Unmeasured Confounding with Negative Controls*. 2023. arXiv: 2309.02631 [stat.ME].

[24] Impello Biosciences. *Does Conventional Fertilizer Harm Soil Health?* 2023. URL: https://impellobio.com/blogs/inoculants/does-conventional-fertilizer-harm-soil-health.

[25] S. Kabanikhin et al. "Definitions and examples of inverse and ill-posed problems". In: *Journal of Inverse and Ill-posed Problems* 16 (Jan. 2008), pp. 317–357.

[26] Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. *Causal Inference Under Unmeasured Confounding With Negative Controls: A Minimax Learning Approach*. 2022. arXiv: 2103.14029 [stat.ML]. URL: https://arxiv.org/abs/2103.14029.

[27] Motonobu Kanagawa et al. *Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences*. 2018. arXiv: 1807.02582 [stat.ML]. URL: https://arxiv.org/abs/1807.02582.

[28] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG]. URL: https://arxiv.org/abs/1412.6980.

[29] Christopher E Kline. "The bidirectional relationship between exercise and sleep: Implications for exercise adherence and sleep improvement". In: *American Journal of Lifestyle Medicine* 8.6 (Nov. 2014), pp. 375–379. DOI: 10.1177/1559827614544437.

[30] Benjamin Kompa et al. *Deep Learning Methods for Proximal Inference via Maximum Moment Restriction*. 2022. arXiv: 2205.09824 [stat.ML].

[31] Benjamin Kompa et al. *Deep Learning Methods for Proximal Inference via Maximum Moment Restriction*. OpenReview, Oct. 2022. URL: https://openreview.net/forum?id=fRWwcgfXXZ&referrer=%5Bthe%20profile%20of%20Benjamin%20Kompa%5D(%2Fprofile%3Fid%3D~Benjamin_Kompa1) (visited on 06/13/2024).

[32] Roman Kräussl, Tobi Oladiran, and Denitsa Stefanova. "A review on ESG investing: Investors' expectations, beliefs and perceptions". In: *Journal of Economic Surveys* 38.2 (2024), pp. 476–502. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/joes.12599.

[33] Greg Lewis and Vasilis Syrgkanis. *Adversarial Generalized Method of Moments*. 2018. arXiv: 1803.07164 [econ.EM]. URL: https://arxiv.org/abs/1803.07164.

[34] Zhu Li et al. *Optimal Rates for Regularized Conditional Mean Embedding Learning*. 2023. arXiv: 2208.01711 [stat.ML]. URL: https://arxiv.org/abs/2208.01711.

[35] Oliver J. Maclaren and Ruanui Nicholson. *What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems*. 2020. arXiv: 1904.02826 [math.ST].

[36] Afsaneh Mastouri et al. *Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restriction*. 2023. arXiv: 2105.04544 [cs.LG].

[37] Wang Miao, Zhi Geng, and Eric Tchetgen Tchetgen. *Identifying Causal Effects With Proxy Variables of an Unmeasured Confounder*. 2018. arXiv: 1609.08816 [stat.ME].

[38] Krikamol Muandet et al. *Dual Instrumental Variable Regression*. 2020. arXiv: 1910.12358 [stat.ML]. URL: https://arxiv.org/abs/1910.12358.

[39] Krikamol Muandet et al. “Kernel Mean Embedding of Distributions: A Review and Beyond”. In: *Foundations and Trends® in Machine Learning* 10.1–2 (2017), pp. 1–141. ISSN: 1935-8245. DOI: 10.1561/2200000060. URL: http://dx.doi.org/10.1561/2200000060.

[40] Jan van Neerven. *Functional Analysis*. Cambridge University Press, June 2022. ISBN: 9781009232470. DOI: 10.1017/9781009232487. URL: http://dx.doi.org/10.1017/9781009232487.

[41] Chan Park and Eric Tchetgen Tchetgen. *Single Proxy Synthetic Control*. 2023. arXiv: 2307.16353 [stat.ME]. URL: https://arxiv.org/abs/2307.16353.

[42] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Penguin Books, 2019.

[43] James Robins. “A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect”. In: *Mathematical Modelling* 7.9 (1986), pp. 1393–1512. ISSN: 0270-0255. DOI: https://doi.org/10.1016/0270-0255(86)90088-6. URL: https://www.sciencedirect.com/science/article/pii/0270025586900886.

[44] Donald B. Rubin. “Assignment to Treatment Group on the Basis of a Covariate”. In: *Journal of Educational Statistics* 2.1 (1977), pp. 1–26. DOI: 10.2307/1164933. URL: https://doi.org/10.2307/1164933 (visited on 04/01/2024).

[45] Tracy A. Ruegg. “Historical Perspectives of the Causation of Lung Cancer”. In: *Global Qualitative Nursing Research* 2 (May 2015), p. 233339361558597. DOI: 10.1177/2333393615585972.

[46] Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. *A Selective Review of Negative Control Methods in Epidemiology*. 2022. arXiv: 2009.05641 [stat.ME].

[47] Nalin A. Singh, Karen M. Clements, and Maria A. Fiatarone. “A Randomized Controlled Trial of the Effect of Exercise on Sleep”. In: *Sleep* 20.2 (Feb. 1997), pp. 95–101. ISSN: 0161-8105. DOI: 10.1093/sleep/20.2.95. URL: https://doi.org/10.1093/sleep/20.2.95.

[48] S. Smale and F. Cucker. “On the mathematical foundations of learning”. In: *Bulletin of the American Mathematical Society* 39.1 (2001), pp. 1–49.

[49] Jerzy Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9”. In: *Statistical Science* 5.4 (1923), pp. 465–472. ISSN: 08834237, 21688745. URL: http://www.jstor.org/stable/2245382 (visited on 04/12/2024).

[50] I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer New York, 2008. ISBN: 9780387772424. URL: https://books.google.nl/books?id=HUnqnrpYt4IC.

[51] Ingo Steinwart. *Convergence Types and Rates in Generic Karhunen-Loeve Expansions with Applications to Sample Path Properties*. 2017. arXiv: 1403.1040 [math.PR]. URL: https://arxiv.org/abs/1403.1040.

[52] Paul D. Stolley. “When genius errs: R.A. Fisher and the lung cancer controversy”. In: *American Journal of Epidemiology* 133.5 (1991), pp. 416–425. DOI: 10.1093/oxfordjournals.aje.a115904.

[53] Eric J Tchetgen Tchetgen et al. *An Introduction to Proximal Causal Learning*. 2020. arXiv: 2009.10982 [stat.ME].

[54] Paul Thagard. “Explaining Disease: Correlations, Causes, and Mechanisms”. In: *Minds and Machines* 8 (1998), pp. 61–78. DOI: 10.1023/a:1008286314688.

[55] Vajira Thambawita et al. “PMData: A Sports Logging Dataset”. In: *Proceedings of the 11th ACM Multimedia Systems Conference*. MMSys ’20. Istanbul, Turkey: Association for Computing Machinery, 2020, pp. 231–236. DOI: 10.1145/3339825.3394926.

[56] José Carlos Santos Todd Rowland. *Compact Operator*. URL: https://mathworld.wolfram.com/CompactOperator.html.

[57] Ilya Tolstikhin, Bharath Sriperumbudur, and Krikamol Muandet. *Minimax Estimation of Kernel Mean Embeddings*. 2017. arXiv: 1602.04361 [math.ST]. URL: https://arxiv.org/abs/1602.04361.

[58] Tseng-Hau Tseng et al. “Effects of exercise training on sleep quality and heart rate variability in middle-aged and older adults with poor sleep quality: a randomized controlled trial”. In: *Journal of Clinical Sleep Medicine* 16.9 (2020), pp. 1483–1492. DOI: 10.5664/jcsm.8560. eprint: https://jcsm.aasm.org/doi/pdf/10.5664/jcsm.8560. URL: https://jcsm.aasm.org/doi/abs/10.5664/jcsm.8560.

[59] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[60] Aad van der Vaart. *Causality and Graphical Models*. Version 31-05-2024. May 2024.

[61] Vladimir Vapnik. *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Berlin, Heidelberg: Springer-Verlag, 1982. ISBN: 0387907335.

[62] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[63] Ziyu Wang et al. *Quasi-Bayesian Dual Instrumental Variable Regression*. 2021. arXiv: 2106.08750 [stat.ML]. URL: https://arxiv.org/abs/2106.08750.

[64] Ziyu Wang et al. *Spectral Representation Learning for Conditional Moment Models*. 2022. arXiv: 2210.16525 [stat.ML].

[65] Abdul-Majid Wazwaz. "The regularization method for Fredholm integral equations of the first kind". In: *Computers & Mathematics with Applications* 61.10 (2011), pp. 2981–2986. ISSN: 0898-1221. DOI: https://doi.org/10.1016/j.camwa.2011.03.083. URL: https://www.sciencedirect.com/science/article/pii/S0898122111002628.

[66] Jefrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, 2010. ISBN: 9780262232586. URL: http://www.jstor.org/stable/j.ctt5hhcfr (visited on 07/14/2024).

[67] Yong Wu et al. *Doubly Robust Proximal Causal Learning for Continuous Treatments*. 2024. arXiv: 2309.12819 [stat.ME]. URL: https://arxiv.org/abs/2309.12819.

[68] Liyuan Xu, Heishiro Kanagawa, and Arthur Gretton. *Deep Proxy Causal Learning and its Application to Confounded Bandit Policy Evaluation*. openreview.net, Nov. 2021. URL: https://openreview.net/forum?id=0FDxsIEv9G (visited on 08/03/2024).

[69] Liyuan Xu, Heishiro Kanagawa, and Arthur Gretton. *Deep Proxy Causal Learning and its Application to Confounded Bandit Policy Evaluation*. 2024. arXiv: 2106.03907 [cs.LG]. URL: https://arxiv.org/abs/2106.03907.

[70] Zhichao Yin et al. "Nitrogen, phosphorus, and potassium fertilization to achieve expected yield and improve yield components of mung bean". In: *PLOS ONE* 13 (Oct. 2018), e0206285. DOI: 10.1371/journal.pone.0206285.

[71] Shawn D Youngstedt. "Effects of exercise on sleep". In: *Clinics in Sports Medicine* 24.2 (2005), pp. 355–365, xi. URL: https://tudelft.on.worldcat.org/oclc/112080371.

A

Appendix

A.1 Linear-Extras

Lemma A.1 (Delta Method Calculations)
Let

$$\Sigma = \begin{bmatrix} \sigma^2_{\theta_A} & \rho_\theta & 0 & 0 \\ \rho_\theta & \sigma^2_{\theta_Z} & 0 & 0 \\ 0 & 0 & \sigma_{\gamma_A} & \rho_\gamma \\ 0 & 0 & \rho_\gamma & \sigma_{\gamma_Z} \end{bmatrix}$$

and $\nabla f(B_0) = \left[1, -\frac{\gamma_{AW}}{\gamma_{ZW}}, \frac{\theta_{ZY}}{\gamma_{ZW}}, -\frac{\theta_{ZY}\gamma_{AW}}{\gamma^2_{ZW}}\right]$. *Then*

$$\begin{aligned}\nabla f(B_0)\Sigma\nabla f(B_0)^T = \sigma^2_{\theta_A} - \frac{\gamma_{AW}}{\gamma_{ZW}}\rho_\theta - \frac{\gamma_{AW}}{\gamma_{ZW}}\rho_\theta + (\frac{\gamma_{AW}}{\gamma_{ZW}})^2\sigma^2_{\theta_Z} \\ + (\frac{\theta_{ZY}}{\gamma_{ZW}})^2\sigma_{\gamma_A} - \frac{\theta^2_{ZY}\gamma_{AW}}{\gamma^3_{ZW}}\rho_\gamma - \frac{\theta^2_{ZY}\gamma_{AW}}{\gamma^3_{ZW}}\rho_\gamma + (\frac{\theta_{ZY}\gamma_{AW}}{\gamma^2_{ZW}})^2\sigma_{\gamma_Z}\end{aligned}$$

Proof.

$$\nabla f(B_0)\Sigma = \left[\sigma^2_{\theta_A} - \tfrac{\gamma_{AW}}{\gamma_{ZW}}\rho_\theta, \rho_\theta - \tfrac{\gamma_{AW}}{\gamma_{ZW}}\sigma^2_{\theta_Z}, \tfrac{\theta_{ZY}}{\gamma_{ZW}}\sigma_{\gamma_A} - \tfrac{\theta_{ZY}\gamma_{AW}}{\gamma^2_{ZW}}\rho_\gamma, \tfrac{\theta_{ZY}}{\gamma_{ZW}}\rho_\gamma - \tfrac{\theta_{ZY}\gamma_{AW}}{\gamma^2_{ZW}}\sigma_{\gamma_Z}\right]$$

$$\begin{aligned}\nabla f(B_0)\Sigma\nabla f(B_0)^T = (\sigma^2_{\theta_A} - \frac{\gamma_{AW}}{\gamma_{ZW}}\rho_\theta)\cdot 1 + (\rho_\theta - \frac{\gamma_{AW}}{\gamma_{ZW}}\sigma^2_{\theta_Z})\cdot(-\frac{\gamma_{AW}}{\gamma_{ZW}}) \\ + (\frac{\theta_{ZY}}{\gamma_{ZW}}\sigma_{\gamma_A} - \frac{\theta_{ZY}\gamma_{AW}}{\gamma^2_{ZW}}\rho_\gamma)\cdot\frac{\theta_{ZY}}{\gamma_{ZW}} + (\frac{\theta_{ZY}}{\gamma_{ZW}}\rho_\gamma - \frac{\theta_{ZY}\gamma_{AW}}{\gamma^2_{ZW}}\sigma_{\gamma_Z})\cdot(-\frac{\theta_{ZY}\gamma_{AW}}{\gamma^2_{ZW}})\end{aligned}$$

Simplifying:

$$\nabla f(B_0)\Sigma\nabla f(B_0)^T = \sigma^2_{\theta_A} - 2\frac{\gamma_{AW}}{\gamma_{ZW}}\rho_\theta + (\frac{\gamma_{AW}}{\gamma_{ZW}})^2\sigma^2_{\theta_Z} + (\frac{\theta_{ZY}}{\gamma_{ZW}})^2\sigma_{\gamma_A} - 2\frac{\theta^2_{ZY}\gamma_{AW}}{\gamma^3_{ZW}}\rho_\gamma + (\frac{\theta_{ZY}\gamma_{AW}}{\gamma^2_{ZW}})^2\sigma_{\gamma_Z}$$

□

A.1.1 Gaussian Models

Suppose I sequentially generate the data as:

$$\begin{aligned} U &= \beta_0 + \epsilon_U \\ W &= \varphi_0 + \varphi_U \cdot U + \epsilon_W \\ Z &= \gamma_0 + \gamma_U \cdot U + \epsilon_Z \\ A &= \alpha_0 + \alpha_U \cdot U + \alpha_Z \cdot Z + \epsilon_A \end{aligned}$$

Where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. and independent from the other variables, This implies that $\mathbb{E}[X\epsilon_Y] = \sigma^2\delta_{X=Y}$
This way we have the following expectations:

$\mathbb{E}[U] = \beta_0$
$\mathbb{E}[W] = \varphi_0 + \varphi_U\beta_0$
$\mathbb{E}[Z] = \gamma_0 + \gamma_U\beta_0$
$\mathbb{E}[A] = \alpha_0 + \alpha_U\beta_0 + \alpha_Z(\gamma_0 + \gamma_U\beta_0)$
$\mathbb{E}\left[U^2\right] = \beta_0^2 + \sigma^2$

$$\mathbb{E}[UW] = \mathbb{E}[U \cdot (\varphi_0 + \varphi_U \cdot U + \epsilon_W)] = \varphi_0\beta_0 + \varphi_U(\beta_0^2 + \sigma^2)$$

$$\mathbb{E}[UZ] = \mathbb{E}[U \cdot (\gamma_0 + \gamma_U \cdot U + \epsilon_Z)] = \gamma_0\beta_0 + \gamma_U(\beta_0^2 + \sigma^2)$$

$$\mathbb{E}[UA] = \mathbb{E}[U \cdot (\alpha_0 + \alpha_U \cdot U + \alpha_Z \cdot Z + \epsilon_A] = \alpha_0\beta_0 + \alpha_U(\beta_0^2 + \sigma^2) + \alpha_Z\left(\gamma_0\beta_0 + \gamma_U(\beta_0^2 + \sigma^2)\right)$$

$$\mathbb{E}[ZW] = \mathbb{E}[(\gamma_0 + \gamma_U \cdot U + \epsilon_Z)(\varphi_0 + \varphi_U \cdot U + \epsilon_W)] = \gamma_0\varphi_0 + \beta_0(\gamma_0\varphi_U + \gamma_U\varphi_0) + \gamma_U\varphi_U(\beta_0^2 + 1)$$

$$\mathbb{E}\left[W^2\right] = \mathbb{E}[W \cdot (\varphi_0 + \varphi_U \cdot U + \epsilon_W)])] = \varphi_0(\varphi_0 + \varphi_U\beta_0) + \varphi_U(\varphi_0\beta_0 + \varphi_U(\beta_0^2 + \sigma^2)) + \sigma^2$$

$$\mathbb{E}\left[Z^2\right] = \mathbb{E}[Z \cdot (\gamma_0 + \gamma_U \cdot U + \epsilon_Z)])] = \gamma_0(\gamma_0 + \gamma_U\beta_0) + \gamma_U(\gamma_0\beta_0 + \gamma_U(\beta_0^2 + \sigma^2)) + \sigma^2$$
$$\mathbb{E}[AZ] = \mathbb{E}[Z \cdot (\alpha_0 + \alpha_U \cdot U + \alpha_Z \cdot Z + \epsilon_A)])]$$
$$= \alpha_0(\gamma_0 + \gamma_U\beta_0) + \alpha_U(\gamma_0\beta_0 + \gamma_U(\beta_0^2 + \sigma^2)) + \alpha_Z(\gamma_0(\gamma_0 + \gamma_U\beta_0) + \gamma_U(\gamma_0\beta_0 + \gamma_U(\beta_0^2 + \sigma^2)) + \sigma^2)$$

$$\mathbb{E}\left[A^2\right] = \mathbb{E}[A \cdot (\alpha_0 + \alpha_U \cdot U + \alpha_Z \cdot Z + \epsilon_A)]$$
$$= \alpha_0(\alpha_0 + \alpha_U\beta_0 + \alpha_Z(\gamma_0 + \gamma_U\beta_0)) + \alpha_U \cdot (\alpha_0\beta_0 + \alpha_U(\beta_0^2 + \sigma^2) + \alpha_Z\left(\gamma_0\beta_0 + \gamma_U(\beta_0^2 + \sigma^2)\right))$$
$$+ \alpha_Z \cdot (\alpha_0(\gamma_0 + \gamma_U\beta_0) + \alpha_U(\gamma_0\beta_0 + \gamma_U(\beta_0^2 + \sigma^2)) + \alpha_Z(\gamma_0(\gamma_0 + \gamma_U\beta_0) + \gamma_U(\gamma_0\beta_0 + \gamma_U(\beta_0^2 + \sigma^2)) + \sigma^2)) + \sigma^2$$

$cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
$cov(aX + b, Y) = acov(X, Y)$
$cov(U, U) = \beta_0^2 + \sigma^2 - \beta_0^2 = \sigma^2$
$cov(U, W) = \varphi_U cov(U, U) = \varphi_U\sigma^2$
$cov(U, Z) = \gamma_U cov(U, U) = \gamma_U\sigma^2$
$cov(U, A) = \alpha_U cov(U, U) + \alpha_Z cov(U, Z) = \sigma^2(\alpha_U + \alpha_Z\gamma_U)$
$cov(Z, Z) = \gamma_U^2 cov(U, U) + cov(\epsilon_Z, \epsilon_Z) = \sigma^2(\gamma_U^2 + 1)$
$cov(W, W) = \varphi_U^2 cov(U, U) + cov(\epsilon_W, \epsilon_W) = \sigma^2(\varphi_U^2 + 1)$
$cov(Z, W) = \varphi_U\gamma_U\sigma^2$
$cov(A, Z) = \alpha_Z cov(Z, Z) + \alpha_U cov(Z, U) = \alpha_Z\sigma^2(\gamma_U^2 + 1) + \alpha_U\gamma_U\sigma^2$
$cov(A, W) = \alpha_Z cov(Z, W) + \alpha_U cov(U, W) = \alpha_Z\varphi_U\gamma_U\sigma^2 + \alpha_U\varphi_U\sigma^2$
$cov(A, A) = \alpha_Z cov(A, Z) + \alpha_U cov(A, U) + \sigma^2 = \sigma^2\left(\alpha_Z^2(\gamma_U^2 + 1) + \alpha_U^2 + 2\alpha_U\alpha_Z\gamma_U + 1\right)$

$$\Sigma = \begin{bmatrix} \Sigma_{UU} & \Sigma_{UZ} & \Sigma_{UA} \\ \Sigma_{ZU} & \Sigma_{ZZ} & \Sigma_{ZA} \\ \Sigma_{AU} & \Sigma_{AZ} & \Sigma_{AA} \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \gamma_U & (\alpha_U + \alpha_Z\gamma_U) \\ \gamma_U & (\gamma_U^2 + 1) & \alpha_Z(\gamma_U^2 + 1) + \alpha_U\gamma_U \\ (\alpha_U + \alpha_Z\gamma_U) & \alpha_Z(\gamma_U^2 + 1) + \alpha_U\gamma_U & \alpha_Z^2(\gamma_U^2 + 1) + \alpha_U^2 + 2\alpha_U\alpha_Z\gamma_U + 1 \end{bmatrix}$$

Letting $S = (Z, A)$ conditional distribution of $U|ZA$ is then:

$$U|ZA \sim \mathcal{N}(\tilde{\mu}, \tilde{\Sigma})$$
$$\tilde{\mu} = \mu_U - \Sigma_{SU}\Sigma_{SS}^{-1}(S - \mu_S)$$
$$\tilde{\Sigma} = \Sigma_{UU}$$

Letting $S^c = S - \mu_S$ we have that:

$$\tilde{\mu} = \mu_U + \frac{Z^c \left(\Sigma_{AA}\Sigma_{UZ} - \Sigma_{UA}\Sigma_{AZ}\right) + A^c \left(\Sigma_{ZZ}\Sigma_{UA} - \Sigma_{UZ}\Sigma_{ZA}\right)}{\Sigma_{ZA}^2 - \Sigma_{AA}\Sigma_{ZZ}} \tag{A.1}$$

The coefficients for $\mathbb{E}\left[U|Z,A\right] = \eta_0 + \eta_Z Z + \eta_A A$ are then found as:

$$\eta_Z = \frac{\left(\Sigma_{AA}\Sigma_{UZ} - \Sigma_{UA}\Sigma_{AZ}\right)}{\Sigma_{ZA}^2 - \Sigma_{AA}\Sigma_{ZZ}} = \frac{\gamma_U - \alpha_U \cdot \alpha_Z}{\alpha_U^2 + \gamma_U^2 + 1}$$
$$\eta_A = \frac{\left(\Sigma_{ZZ}\Sigma_{UA} - \Sigma_{UZ}\Sigma_{AZ}\right)}{\Sigma_{ZA}^2 - \Sigma_{AA}\Sigma_{ZZ}} = \frac{\alpha_U}{\alpha_U^2 + \gamma_U^2 + 1}$$

The ratio we wish to estimate is thus:

$$\frac{\theta_{AY}}{\theta_{ZY}} = \frac{\eta_A}{\eta_Z} = \frac{\alpha_U}{\gamma_U - \alpha_U \cdot \alpha_Z}$$

This is ill conditioned if $\gamma_U - \alpha_U\alpha_Z$ is close to zero.

A.2 Efficiency of the influence function in Semiparametric Proximal Causal Inference

Theorem 3.1 shows that the found function is an influence function for the model characterized by the existence of an outcome bridge function. The authors of [13] then state that, under the additional assumption that both E, E^* are surjective, the influence function of equation (3.13) is efficient. This is done by showing that such function is within the tangent space.

Proof. From the proof of Theorem 3.1 we have that the tangent set is:

$$\begin{aligned}\mathcal{T} = \{&S(Y,W,Z,A,X) = S(Z,A,X) + S(Y,W|Z,A,X) : \\ &S(Z,A,X) \in L_2(Z,A,X) : \mathbb{E}\left[S(Z,A,X)\right] = 0 \\ &S(Y,W|Z,A,X) \in L_2(Z,A,X)^{\perp} : \mathbb{E}\left[S(Y,W|Z,A,X)\right] = 0 \\ &\mathbb{E}\left[(Y - h_0(W,A,X))S(Y,W|Z,A,X)|Z,A,X\right] \in Range(E)\}\end{aligned}$$

Now notice that:

$$\begin{aligned}IF &= (-1)^{1-A}q(Z,A,X)\cdot(Y - h(W,A,X)) + \Delta_A h - \psi \\ &= (-1)^{1-A}q(Z,A,X)\cdot(Y - h(W,A,X)) + \Delta_A h - \psi \pm \mathbb{E}\left[\Delta_A h - \psi|Z,A,X\right] \\ &= \mathbb{E}\left[\Delta_A h - \psi\right]|Z,A,X + (-1)^{1-A}q(Z,A,X)\cdot(Y - h(W,A,X)) + \Delta_A h - \psi - \mathbb{E}\left[\Delta_A h - \psi|Z,A,X\right]\end{aligned}$$

By the proof of Theorem 3.1, the first term is in $\mathcal{T}$. For the remaining term:

$$\begin{aligned}&\mathbb{E}\left[(-1)^{1-A}q(Z,A,X)\cdot\left(Y - h(W,A,X)^2\right)|Z,A,X\right] = 0 \\ &\mathbb{E}\left[\Delta_A h - \psi - \mathbb{E}\left[\Delta_A h - \psi|Z,A,X\right]|Z,A,X\right] = \mathbb{E}\left[\Delta_A h - \psi|Z,A,X\right] - \mathbb{E}\left[\Delta_A h - \psi|Z,A,X\right] = 0\end{aligned}$$

Moreover, by the assumption on the surjectivity of E we have that both are within the closure of Range(E) and thus the IF is in $\mathcal{T}$. □

Surjectivity is an odd requirement as the image of the operator would have to coincide with all of L_2. Moreover, conditional expectation operators are often compact and thus their range cannot coincide with all of L_2[60]. Even more so, the bijectivity of the conditional expectation operator required in [26] is an even stranger requirement since, if this was the case, invertibility of E would not be a problem and the problem would not be ill posed.

A.3 Quasi-Bayesian Proximal Inference

A.3.1 Fenchel Duality

The authors of the paper justify the loss function using so called Fenchel Duality. This was the main approach also used in the original dual instrumental variable paper[38]. For sake of completion it is also reported here.

Definition A.1 (Proper Function)
A function h is said to be proper if the domain of h is not the empty set and it is always greater than $-\infty$.

Definition A.2 (Semicontinuity)
A function is said to be upper semicontinuous at x_0 if $\forall \epsilon > 0$ there exists $\delta > 0$ such that:

$$\forall x \in B(x_0, \delta) \quad f(x) - f(x_0) < \epsilon$$

. Similarly, a function is said to be lower semicontinuous at x_0 if: $\forall \epsilon > 0$ there exists $\delta > 0$ such that:

$$\forall x \in B(x_0, \delta) \quad f(x) - f(x_0) > \epsilon$$

.

From these definitions, it clearly follows that a continuous function is both upper and lower semicontinuous.

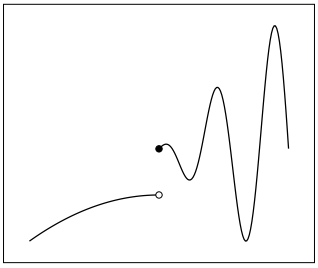


Figure A.1: Example of upper semicontinuous function.

Theorem A.1 ([38] Interchangeability)
Let W be a random variable on Ω. For all $W \in \Omega$ let $h(\cdot, W) : \mathbb{R} \to (-\infty, +\infty)$ be proper and upper semicontinuous and concave. Then:

$$\mathbb{E}_W \left[\max_{u \in \mathbb{R}} f(u, W) \right] = \max_{u \in U(\Omega)} \mathbb{E} \left[f(u(W), W) \right]$$

where $U(\Omega) : \{u : \Omega \to \mathbb{R}\}$.

Definition A.3 (Convex Conjugate)
For any function $h : \mathbb{R}^n \to \mathbb{R}$ the convex conjugate to h is defined by:

$$h^*(s) = \sup_x \{\langle s, x \rangle - h(x)\}$$

where $\langle \cdot, \cdot \rangle$ is the classic inner product on $\mathbb{R}^n$.

Definition A.4 ([38]Fenchel Duality[1])
Let $l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ be a proper, convex, and lower semi-continuous loss function for any value in its first argument and $l_y^* = l^*(y, \cdot)$ a convex conjugate of $l_y = l(y, \cdot)$ which is also proper, convex, and lower semi-continuous w.r.t. the second argument. Then, $l_y(v) = \max_u \{uv - l_y^*(u)\}$.

Proposition A.1 (Conjugate of the squared loss)
Let $l_v(t) = \frac{1}{2}(v - t)^2 \quad v, t \in \mathbb{R}$ be the squared loss function. The convex conjugate of l_v is given by:

$$l_v^*(s) = sv + \frac{1}{2}s^2$$

[1]The following resource provides interactive graphs to gain an intuition: https://remilepriol.github.io/dualityviz/

Proof. First observe that $l_v(t)$ is proper since it its domain is $\mathbb{R}$, its convex and continuous and thus lower-semicontoniuous. By definition of convex conjugate

$$\begin{aligned} l_v^*(s) &= \sup_{t\in\mathbb{R}} \{st - l_v(t)\} \\ &= \sup_{t\in\mathbb{R}} \left\{ st - \frac{1}{2}(v-t)^2 \right\} \end{aligned}$$

l_v^* is proper,concave and upper-semi continuous in s by the same reasoning as for l_v. Taking the derivative we can determine its zeros:

$$\frac{d}{dt}\left[st - \frac{1}{2}(v-t)^2 \right] = s + (v-t)$$

Setting this equal to zero, the critical point is given by $t = s + v$. Since the function is concave in t this is necessarily a maximum. Plugging this back into the definition of l_v^* we obtain:

$$l_v^*(s) = s(s+v) - \frac{1}{2}\left(v - (s+v)\right)^2 = s(s+v) - \frac{1}{2}s^2 = sv + \frac{1}{2}s^2$$

□

Proposition A.2
If the function class $\mathcal{G}$ contains all functions $\mathbb{E}\left[h(W,A) - h'(W,A)|Z,A\right] \quad \forall h, h' \in \mathcal{H}$ then the loss in equation (5.5) is equivalent to:

$$\mathcal{L}(h) = \max_{u\in\mathcal{G}} \mathbb{E}_{Z,A}\left[(h(W,A) - Y)\cdot u(Z,A) - \frac{1}{2}u^2(Z,A) \right] \tag{A.2}$$

Proof. Consider the squared loss function $l = \frac{1}{2}\left(\mathbb{E}\left[Y|Z,A\right] - \mathbb{E}\left[h(W,A)|Z,A\right]\right)^2$. Then using the results from appendix A.3.1, in particular Interchangeability and Fenchel Duality:

$$\begin{aligned} \mathcal{L}(h) &= \mathbb{E}_{Z,A}\left[l\left(\mathbb{E}\left[Y|Z,A\right], \mathbb{E}\left[h(W,A)|Z,A\right]\right)\right] \\ &= \mathbb{E}_{Z,A}\left[\max_{u\in\mathbb{R}} \mathbb{E}\left[h(W,A)|Z,A\right]\cdot u - l^*_{\mathbb{E}[Y|Z,A]}(u) \right] \\ &= \max_{g\in\mathcal{G}(Z,A)} \mathbb{E}_{Z,A}\left[\mathbb{E}\left[h(W,A)|Z,A\right]\cdot g(Z,A) - l^*_{\mathbb{E}[Y|Z,A]}(g(Z,A)) \right] \\ &= \max_{g\in\mathcal{G}(Z,A)} \mathbb{E}_{Z,A}\left[h(W,A)\cdot g(Z,A) - l^*_{\mathbb{E}[Y|Z,A]}(g(Z,A)) \right] \end{aligned}$$

By proposition A.1 we have that $l_v^*(s) = sv + \frac{1}{2}s^2$. Substituting this in the previous completes the proof:

$$\begin{aligned} \mathcal{L}(h) &= \max_{g\in\mathcal{G}(Z,A)} \mathbb{E}_{Z,A}\left[h(W,A)\cdot g(Z,A)\right] - \mathbb{E}_{Z,A}\left[\mathbb{E}\left[Y|Z,A\right] g(Z,A) + \frac{1}{2}g^2(Z,A) \right] \\ &= \max_{g\in\mathcal{G}(Z,A)} \mathbb{E}_{Z,A}\left[(h(W,A) - Y)\cdot g(Z,A) - \frac{1}{2}g^2(Z,A) \right] \end{aligned}$$

□

Notice that the assumption on the function class $\mathcal{G}$ is similar to the one of *Corollary* 3.2 in the semi-parametric setting.

A.3.2 Non-parametric Bayesian results

Here are the assumptions and technical results used in [63] to achieve the posterior consistency result in the Instrumental Variable setting. They are presented with the notation adapted to the proximal inference problem.

Assumption A.1
$Y - h_0(V_h)$ is bounded, i.e. $\exists B : \|Y - h_0\|_\infty \leq B$.

Assumption A.2 ensures that the associated kernel operator admits a Mercer decomposition theorem 2.3. Intuitively, the third point of this assumption requires every point in $\mathcal{X}$ to have some probability of happening.

Assumption A.2 *1. $\mathcal{W}, \mathcal{A}, \mathcal{X}, \mathcal{Z}$ are Polish*[2]*.*

2. k_{V_h} is measurable continuous, bounded and L_2 universal

3. P_{V_h} has full support.

Assumption A.3 enforces the eigenvalues of $\mathcal{H}$ to have a polynomial decay, where b determines the regularity of the functions in it. As discussed in definition 2.8, lower b translates to rougher or less regular functions. On the contrary if b is large, then the functions in $\mathcal{H}$ are smoother. Note that assumption A.3 excludes the Gaussian kernel since it has exponentially, and thus much faster decaying eigenvalues. Nonetheless, we implement it in practice due to its computational simplicity and can heuristically consider it as a Matérn with very very large b. If some other regularities were known about h_0 then the choice of b becomes immediate.

Assumption A.3 *1. The eigenvalues of the kernel operator defining $\mathcal{H}$ are such that $\lambda_i^2 \asymp i^{-(b+1)}$ for some $b > 1$.*

Since the samples of a gaussian process do not belong to its reproducing kernel hilbert space, assumption A.4 ensures that the true bridge function h_0 is in some power space of the RKHS or at least that there is an element in $\mathcal{H}$ arbitrarily close to h_0. This enables us to have a rougher or more irregular true bridge function than in the RKHS of the GP.

Assumption A.4
Define $\overline{\mathcal{H}} = \mathcal{H}^{\frac{b}{b+1}}$. Then we require that:

*1. $h_0 \in \overline{\mathcal{H}}$ **OR** $\exists h_i^* \in \mathcal{H}$ such that $\|h_i^*\|_{\mathcal{H}} \lesssim i^{\frac{1}{b+1}}$ and $\|h_0 - h_i^*\|_\infty^2 \lesssim 1$*

The problem is assumed to be mildly ill posed, i.e. the eigenvalues of the conditional expectation operator are polynomially decaying. The parapmeter p decides the level of ill-posedness. Higher values of p translate to more compression of information and thus a higher level of ill-posedness.

Assumption A.5 (Mildly-ill posed)
The conditional expectation operator E has singular values $\nu_i \asymp i^{-p}$ where $p \geq 0$.

Definition A.5 (L_2 Projection)
Let $\{\varphi_i\}$ be the previously defined ONB of $L_2(V_h, P_{V_h})$. Then the L_2 projection on the first j elements of $\{\varphi_i\}$ is defined as:

$$Proj_J h = \sum_{i=1}^{J} \langle h, \varphi_i \rangle_2 \varphi_i$$

The following assumptions ensures that the eigenfunctions of the kernel operator K and E^*E are aligned up to scaling constants. Although the authors use assumption A.6, we can also use a stronger condition in that $\{\varphi_i\}$ diagonalizes E^*E. This implies their assumption A.6 as is shown in proposition A.3.

Assumption A.6 ([63]Link Conditions)
For all $h \in L_2(_{V_h}, P_{V_h})$ we assume that $\forall J$:

$$\|Eh\|_2^2 \gtrsim J^{-2p} \|Proj_J h\|_2^2 \qquad \text{(Reverse Link)}$$

$$\|Eh\|_2^2 \lesssim \sum_{i=1}^{\infty} i^{-2p} \langle h, \varphi_i \rangle_2^2 \qquad \text{(Link)}$$

Proposition A.3
If $\{\varphi_i\}$ diagonalizes E^*E where E is the conditional expectation operator, then assumption A.6 is satisfied.

Proof. Let $\{\varphi_i\}$ be the ONB of the kernel operator T associated with the RKHS $\mathcal{H}$. By assumption this basis diagonalizes E^*E. This means that both E and its adjoint E^* can be written as series expansion:

$$Eh = \sum_{i=1}^{\infty} \nu_i \langle h, \varphi_i \rangle_2 \phi_i$$

[2]Separable completely metrizable topological space. In other words, nicely behaving spaces.

It is immediate by assumption A.5 that:

$$\|Eh\|_2^2 = \sum_{i=1}^{\infty} \nu_i^2 \langle h, \varphi_i \rangle_2^2 \asymp \sum_{i=1}^{\infty} i^{-2p} \langle h, \varphi_i \rangle_2^2 = \sum_{i=1}^{J} i^{-2p} \langle Proj_J h, \varphi_i \rangle_2^2 + \sum_{i=J+1}^{\infty} i^{-2p} \langle Proj_{>J} h, \varphi_i \rangle_2^2$$

The Link assumption is satisfied since $A \asymp B \equiv \exists c > 0 : A = cB \implies A \leq cB \equiv A \lesssim B$. Similarly for the Reverse link assumption:

$$\|Eh\|_2^2 \geq \sum_{i=1}^{J} i^{-2p} \langle Proj_J h, \varphi_i \rangle_2^2 \geq J^{-2p} \sum_{i=1}^{J} \langle Proj_J h, \varphi_i \rangle_2^2$$

since i^{-2p} is a decreasing sequence in i. □

The following assumption defines the critical radius of the adversarial function class $\mathcal{G}$. This enables us to use various concentration inequalities.

Assumption A.7 (Complexities)
Denote by $\mathcal{G}_1$ the unit norm ball of $\mathcal{G}$. Let δ_n be the critical radius of the local Rademacher complexity of $\mathcal{G}_1$. Then

1. *Functions in $_{\mathcal{I}}$ are uniformly bounded and $\delta_n^2 \lesssim n^{-\frac{b+2p}{b+2p+1}}$*
2. *The restriction of E to $\overline{\mathcal{H}}$ has image in $\mathcal{G}$, i.e. $E(\overline{\mathcal{H}}) \subset \mathcal{G}$ and is bounded $\|Eh\|_{\mathcal{G}} \leq \|h\|_{\overline{\mathcal{H}}}$*

This lemma shows that the powerspace of a RKHS is the RKHS associated to the kernel operator with kernel equal to the power placed on its eigenvalues. This is very intuitive since $K = \int k d\mu$ can be seen as a very large matrix, any power K^α is equivalent to applying it to the 'diagonal matrix' consisting of its eigenvalues λ_i.

Lemma A.2 ([63]Lemma 10)
Let $\mathcal{H}$ satisfy assumptions assumptions A.2 and A.3. Then $\forall \alpha \in [\frac{b_e}{b+1}, \infty) \supset [\frac{b}{b+1}, \infty)$ the power space:

$$\mathcal{H}^\alpha = \left\{ \sum_{i=1}^{\infty} a_j \varphi_j : \sum_{i=1} \lambda_i^{-2\alpha} a_i^2 < \infty \right\}$$

- *is a reproducing kernel Hilbert space with kernel given by*

$$k_\alpha(x_1, x_2) = \sum_{i=1}^{\infty} \lambda_i^{2\alpha} \varphi_j(x_1) \varphi_j(x_2)$$

- *k_α is a bounded kernel*

This lemma is used to find upperbounds for the norms $\|\cdot\|_{\overline{\mathcal{H}}}, \|\cdot\|_\infty$ in terms of the L_2 and $\mathcal{H}$ norm. As $b \to \infty$ $\overline{\mathcal{H}} \to \mathcal{H}$ and the $\|\cdot\|_{\overline{\mathcal{H}}}$ norm will mostly be similar to $\|\cdot\|_{\mathcal{H}}$.

Lemma A.3 ([63]Lemma 11)
Let $\mathcal{H}$ satisfy assumptions assumptions A.2 and A.3, and let $\overline{\mathcal{H}} = \mathcal{H}^{\frac{b}{b+1}}$. Then $\forall f \in \mathcal{H}$ we have the following:

$$\|h\|_{\overline{\mathcal{H}}}^2 \lesssim (\|h\|_{\mathcal{H}}^2)^{\frac{b}{b+1}} (\|h\|_2^2)^{\frac{1}{b+1}}$$
$$\|h\|_\infty^2 \lesssim (\|h\|_{\mathcal{H}}^2)^{\frac{b}{b+1}} (\|h\|_2^2)^{\frac{1}{b+1}}$$

This result tells me two things: the first is that the probability that a sample from a GP W is small in $\mathcal{H}^{\frac{b'}{b+1}}$ decays exponentially. The second states that W can be decomposed (with high probability) into two terms, a principal one h_ρ in $\mathcal{H}$ and an errror term h_e in $\mathcal{H}^{\frac{b'}{b+1}}$.
If $b' \to b$ then my τ will decrease very slowly and remain close to one much longer.

Lemma A.4 ([63]Lemma 12)
Let $W \sim \Pi$ be a Gaussian Process with covariance kernel k_{V_h}. Then $\forall b' \in [0, b)$ $m > 0$ $\exists c_1, c_2, c_3 > 0$ and ${\tau'}_m^2 = m^{-\frac{b-b'}{b+1}}$ such that:

$$\Pi(\{\|W\|^2_{\mathcal{H}^{\frac{b'}{b+1}}} \leq {\tau'}_m^2\}) \geq \exp(-c_1 m^{\frac{1}{b+1}})$$
$$\Pi(\Theta'_m) \geq 1 - \exp(-c_3 m^{\frac{1}{b+1}})$$

where $\Theta'_m = \{W = h_\rho + h_e : \|h_\rho\|_{\mathcal{H}}^2 \leq c_2 m^{\frac{1}{b+1}},\ \|h_e\|^2_{\mathcal{H}^{\frac{b'}{b+1}}} \leq {\tau'}_m^2\}$

The following lemma shows that a projection in L_2 is also a projection in a power space $\mathcal{H}^\alpha$ but with the basis rescaled by the singular values of the space.

Lemma A.5 ([63]Lemma 15: Projection Equivalences)
Let $\mathcal{H}$ satisfy assumptions A.2 and A.3 and $\overline{\mathcal{H}} = \mathcal{H}^\alpha$ with $\alpha \in [0,1]$ such that $\tilde{b} = (b+1)\alpha$ and let $\tilde{\lambda_i}^2 := \lambda_i^{2\alpha} \asymp i^{-\tilde{b}} = i^{-\alpha(b+1)}$. Then $\forall f \in \overline{\mathcal{H}}$ we have:

1. $Proj_J h = \sum_{i=1}^{j} \left\langle h, \tilde{\lambda_i}\varphi_i \right\rangle_{\overline{\mathcal{H}}} \tilde{\lambda_i}\varphi_i$
2. $Proj_J$ *is the orthogonal projection onto span of the first* J $\{\tilde{\lambda_i}\varphi_i\}$
3. $\|Proj_{>J} h\|_2 \lesssim \|h\|_{\overline{\mathcal{H}}} J^{-\frac{\tilde{b}}{2}}$

Proof. $\{\varphi_i\}$ are the eigenfunctions of L_2 and λ_i^2 are the eigenvalues of the kernel operator associated with $\mathcal{H}$. $\{\varphi_i\}$ forms an O.N.B. of L_2 and thus h can be rewritten as $h = \sum_{i=1} h_i\varphi_i = \sum_{i=1} \langle h, \varphi_i\rangle_2 \varphi_i$, then:

$$\begin{aligned}
\|h\|_2^2 &= \sum_{i=1} \langle h, \varphi_i\rangle_2^2 = h_i^2 \\
\|h\|_{\mathcal{H}^\alpha} &= \sum_{i=1} \frac{1}{\lambda_i^{2\alpha}} \langle h, \varphi_i\rangle_2^2 = \sum_{i=1} \frac{h_i^2}{\lambda_i^{2\alpha}} \\
&= \langle h, h\rangle_{\mathcal{H}^\alpha} = \sum_{i=1} h_i^2 \langle \varphi_i, \varphi_i\rangle_{\mathcal{H}^\alpha}
\end{aligned}$$

Thus, $\|\varphi_i\|_{\mathcal{H}^\alpha}^2 = \frac{1}{\lambda_i^{2\alpha}}$ and then $\{\lambda_i^\alpha \varphi_i\}$ is ONB of $\mathcal{H}^\alpha$. Thus we can rewrite the inner product of L_2 as:

$$\begin{aligned}
\langle h, \varphi\rangle_2 \varphi_i &= \left\langle \sum_{i=1} h_i\varphi_i, \varphi_i \right\rangle_2 \varphi_i = \sum_{i=1} h_i \langle \varphi_i, \varphi_i\rangle_2 \varphi_i = \sum_{i=1} h_i \lambda_i^{2\alpha} \langle \varphi_i, \varphi_i\rangle_{\mathcal{H}^\alpha} \varphi_i \\
&= \sum_{i=1} h_i \langle \varphi_i, \lambda_i^\alpha \varphi_i\rangle_{\mathcal{H}^\alpha} \lambda_i^\alpha \varphi_i = \left\langle \sum_{i=1} h_i\varphi_i, \lambda_i^\alpha \varphi_i \right\rangle_{\mathcal{H}^\alpha} \lambda_i^\alpha \varphi_i = \langle h, \lambda_i^\alpha \varphi_i\rangle_{\mathcal{H}^\alpha} \lambda_i^\alpha \varphi_i
\end{aligned}$$

This means that $Proj_J h = \sum_{i=1}^{J} \langle h, \varphi_i\rangle_2 \varphi_i = \sum_{i=1}^{J} \langle h, \lambda_i^\alpha \varphi_i\rangle_{\mathcal{H}^\alpha} \lambda_i^\alpha \varphi_i$. For the last item, notice that:

$$\begin{aligned}
\|Proj_{>J} h\|_2^2 &= \left\| \sum_{i=J+1} \langle h, \lambda_i^\alpha \varphi_i\rangle_{\mathcal{H}^\alpha} \lambda_i^\alpha \varphi_i \right\|_2^2 = \sum_{i=J+1} \langle h, \lambda_i^\alpha \varphi_i\rangle_{\mathcal{H}^\alpha}^2 \lambda_i^{2\alpha} \|\varphi_i\|_2^2 \\
&\leq \lambda_J^{2\alpha} \sum_{i=J+1} \langle h, \lambda_i^\alpha \varphi_i\rangle_{\mathcal{H}^\alpha}^2 = \lambda_J^{2\alpha} \|Proj_{>J} h\|_{\mathcal{H}^\alpha}^2 \leq \lambda_J^{2\alpha} \|h\|_{\mathcal{H}^\alpha}^2 \asymp J^{-\tilde{b}} \|h\|_{\mathcal{H}^\alpha}^2
\end{aligned}$$

Where the inequality follows from $\lambda_i^{2\alpha}$ being a decreasing positive sequence and as such is always less than or equal to its first term, and $\|\varphi_i\|_2 = 1$. □

This result is similar to that of lemma A.4. The W can be decomposed in two parts: a primary term h_ρ and an 'error' term h_e. Here the error is shown to have both small L_2 norm and $\mathcal{H}_{\frac{b'}{b+1}}$ norm. Moreover, under the assumptions previously introduced the 'error' term is bounded (supnorm is bounded). Since my sample is not in $\mathcal{H}$ with probability one, as m grows, the more functions will be contain in $\mathcal{H}$, and the smaller the error term.

Corollary A.1 ([63])
Let $W \sim \Pi$ be a Gaussian Process with covariance kernel k_{V_h}. Then we have that:

1. $\forall b' \in [0, b)\ \exists c_{b'}, c'_{b'} > 0\ \forall m \in \mathbb{N}\ \tau_m^2 = m^{-\frac{b}{b+1}}$:

$$\Pi(\Theta_{m,b'}) \geq 1 - \exp\left(-c'_{b'} m \tau_m^2\right)$$
$$\Theta_{m,b'} = \{h = h_\rho + h_e : \|h_\rho\|_{\mathcal{H}}^2 \leq c_{b'} m \tau_m^2,\ \|h_e\|_2^2 \leq c_{b'} \tau_m^2,\ \|h_e\|_{\mathcal{H}^{\frac{b'}{b+1}}} \leq c_{b'} m^{-\frac{b-b'}{b+1}}\}$$

2. *Under Assumption assumption A.3 (in particular 3.2)* $\implies \exists c, c' > 0\ \forall m \in \mathbb{N}$:

$$\Pi(\Theta_m) \geq 1 - \exp\left(-c' m \tau_m^2\right)$$
$$\Theta_m = \{h = h_\rho + h_e : \|h_\rho\|_{\mathcal{H}}^2 \leq c m \tau_m^2,\ \|h_e\|_2^2 \leq c \tau_m^2,\ \|h_e\|_\infty \leq c\}$$

Proof. We just need to prove that $\Theta_{m,b'}, \Theta_m$ is the same as the set described in lemma A.4. We first define two intermediate quantities:

$$\tilde{h}_\rho = h_\rho + Proj_J h_e \quad \tilde{h}_e = Proj_{>J} h_e$$

Projections always decrease norms and thus if $h_e \in \Theta'_m \implies \|h_e\|^2_{\mathcal{H}^{\frac{b'}{b+1}}} \leq \tau'^2_m$

$$\tau'^2_m \geq \|h_e\|_{\mathcal{H}^{\frac{b'}{b+1}}} \geq max\left\{\left\|\tilde{h}_e\right\|_{\mathcal{H}^{\frac{b'}{b+1}}}, \|Proj_{>J} h_e\|_{\mathcal{H}^{\frac{b'}{b+1}}}\right\}$$

By lemma A.5 we have that:

$$\left\|\tilde{h}_e\right\|_2^2 \leq \|h_e\|^2_{\mathcal{H}^{\frac{b'}{b+1}}}\ J^{-\bar{b}}$$

Using the triangle inequality:

$$\left\|\tilde{h}_\rho\right\|_{\mathcal{H}}^2 \leq \left(\|h_\rho\|_{\mathcal{H}}^2 + \|Proj_J h_e\|_{\mathcal{H}}^2\right)$$
$$\|Proj_J h_e\|_{\mathcal{H}}^2 = \sum_{i=1}^{J} \frac{\langle h, \varphi_i\rangle_2^2}{\lambda_i^2} \asymp \sum_{i=1}^{J} \langle h, \varphi_i\rangle_2^2\ i^{(b+1)}$$
$$= \sum_{i=1}^{J} \langle h, \varphi_i\rangle_2^2\ i^{(b \pm b' + 1)} \leq J^{b-b'+1} \sum_{i=1}^{J} \langle h, \varphi_i\rangle_2^2\ i^{b'} \asymp J^{b-b'+1} \sum_{i=1}^{J} \langle h, \varphi_i\rangle_2^2\ {\lambda_i}^{2b'}$$
$$= J^{b-b'+1} \sum_{i=1}^{J} \left\langle h, \lambda_i^{b'} \varphi_i \right\rangle_2^2 = C\, \|Proj_{>J} h\|_{\mathcal{H}^{\frac{b'}{b+1}}}$$

where we are able to pull put $J^{b-b'+1}$ because $i^{(b \pm b'+1)}$ is a increasing sequence. Such value can be upperbounded by $C \geq J^{b-b'+1}$. □

This result is key. If a function can be decomposed in primary term and error term with a certain precision, by assumption A.6 the error will also have bounded 2 norm post application of conditional expectation E. In other words, controlling the error in $\mathcal{H}$ enables me to control the error in $Im(E)$, i.e. having a finite measure of ill-posedness.

Lemma A.6 ([63]Lemma 16)
Let m be the first integer larger than $n^{\frac{b+1}{b+2p+1}}$ then for any function $h = \tilde{h}_\rho + \tilde{h}_e$ such that:

$$\left\|\tilde{h}_\rho\right\|_{\mathcal{H}}^2 \lesssim m\tau_m^2 \qquad \left\|\tilde{h}_e\right\|_2^2 \lesssim \tau_m^2 \qquad \left\|\tilde{h}_e\right\|_\infty \lesssim 1$$

we have the decomposition $h = h_\rho + h_e$ such that:

$$\|h_\rho\|_{\mathcal{H}}^2 \lesssim m\tau_m^2 \qquad \|h_e\|_2^2 \lesssim \tau_m^2 \qquad \|Eh_e\|_2^2 \lesssim \delta_n^2 \qquad \|h_e\|_\infty \lesssim 1$$

Proof. $\tilde{h}_e \in L_2$ and for any finite J, the projection $Proj_J\tilde{h}_e$ is well defined and is in $\mathcal{H}$. Let m' be the first integer bigger than $m^{\frac{1}{b+1}}$, then:

$$\left\|\mathrm{Proj}_{m'}\tilde{h}_e\right\|_{\mathcal{H}}^2 = \sum_{i=1}^{m'} \frac{1}{\lambda_i^2}\left\langle \tilde{h}_e, \varphi_i\right\rangle_2^2 \leq \frac{1}{\lambda_{m'}^2}\sum_{i=1}^{m'}\left\langle \tilde{h}_e, \varphi_i\right\rangle_2^2 \leq \frac{1}{\lambda_{m'}^2}\left\|\tilde{h}_e\right\|_2^2 \asymp m'^{b+1}\left\|\tilde{h}_e\right\|_2^2 \lesssim m'^{b+1}\tau_m^2 \asymp m\tau_m^2$$

Where we used the fact that $\frac{1}{\lambda_i}$ is an <u>increasing</u> positive sequence in i, assumption A.5 and $b > 1$. Defining:

$$h_\rho := \tilde{h}_\rho + \mathrm{Proj}_{m'}\tilde{h}_e \qquad h_e := \mathrm{Proj}_{>m'}\tilde{h}_e$$

using the triangle inequality we have the needed norm:

$$\|h_\rho\|_{\mathcal{H}}^2 = \left\|\tilde{h}_\rho + \mathrm{Proj}_{m'}\tilde{h}_e\right\|_{\mathcal{H}}^2 \leq 2\left\|\tilde{h}_\rho\right\|_{\mathcal{H}}^2 + 2\left\|\mathrm{Proj}_{m'}\tilde{h}_e\right\|_{\mathcal{H}}^2 \lesssim m\tau_m^2$$

Similarly for h_e, we use the fact that projections always decrease L_2 norm, assumption A.6 and lemma A.3:

$$\begin{aligned}
\|h_e\|_2^2 &\leq \left\|\tilde{h}_e\right\|_2^2 \lesssim \tau_m^2 \\
\|Eh_e\|_2^2 &\lesssim m'^{-2p}\|h_e\|_2^2 \lesssim n^{-\frac{2p}{b+2p+1}} \cdot n^{-\frac{b}{b+2p+1}} = n^{-\frac{b+2p}{b+2p+1}} \asymp \delta_n^2 \\
\|h_e\|_\infty = \left\|\tilde{h}_e - \mathrm{Proj}_{m'}\tilde{h}_e\right\|_\infty &\leq \left\|\tilde{h}_e\right\|_\infty + \left\|\mathrm{Proj}_{m'}\tilde{h}_e\right\|_\infty \lesssim 1 + \underbrace{\left\|\mathrm{Proj}_{m'}\tilde{h}_e\right\|_{\mathcal{H}}^{\frac{b}{b+1}}\left\|\mathrm{Proj}_{m'}\tilde{h}_e\right\|_2^{\frac{1}{b+1}}}_{\asymp \underbrace{m^{\frac{1}{2}\frac{b}{(b+1)^2}} \cdot m^{-\frac{1}{2}\frac{b}{(b+1)^2}}}_{=1}} \lesssim 1
\end{aligned}$$

□

This is a concentration inequality relating the probability of the sample average deviating from the true mean.

Lemma A.7 ([63]Lemma 17)
Let L be a random variable bounded by B_L and with finite variance $var(L) < \sigma_L^2$. Let $L_i \sim P_L$ be iid samples and $\eta > 0$ then:

$$P\left(|\frac{1}{n}\sum_{i=1} L_i - \mathbb{E}\left[L\right]| \geq (\sqrt{2}\eta + \frac{2\eta}{3})\frac{\sigma_L^2}{B_l}\right) \leq \exp\left(-\frac{n\sigma_L^2}{B_L^2}\eta\right)$$

In particular they will use this inequality to define an event where the the empirical expectation is greater than the true with high probability. Define:

$$\Psi_n(f,g) = \frac{1}{n}\sum_{i=1}^{n} g(V_{q_i})\left(h(V_{h_i}) - y_i\right) \quad \Psi(f,g) = \mathbb{E}\left[g(V_q)\left(h(V_h) - Y\right)\right]$$
$$L := \left(h(V_h) - Y\right)g(V_q) - g^2(V_q) \quad L_i = (h(V_{h_i}) - y_i)g(V_q) - g^2(V_q)$$

Notice that:

$$|L| = |\left(h(V_h) - Y - g(V_q)\right)g(V_q)| = |\left(h(V_h) \pm h_0(X) - Y - g(V_q)\right)g(V_q)| \leq (\|h - h_0\|_\infty + \underbrace{\|h_0 - Y\|_\infty}_{\leq B} + \|g\|_\infty)\|g\|_\infty$$
$$\leq (\|h - h_0\|_\infty + B + \|g\|_\infty)\|g\|_\infty \leq (\|h - h_0\|_\infty + B + \|g\|_\infty)^2$$

$$Var(L) \leq \mathbb{E}\left[L^2\right] \leq (\|h - h_0\|_\infty + B + \|g\|_\infty)^2 \|g\|_2^2$$

$$\Psi_n - \|g\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} g(V_{q_i})\left(h(V_{h_i}) - y_i\right) - g(V_{q_i})^2 = \frac{1}{n}\sum_{i=1}^{n} L_i$$
$$\Psi - \|g\|_2^2 = \mathbb{E}\left[g(V_q)\left(h(V_h) - Y\right)\right] - \mathbb{E}\left[g^2(V_q)\right] = \mathbb{E}\left[L\right]$$

Moreover, since $P(|A| > \xi) \geq P(A \geq \xi)$ lemma A.7 can be applied to obtain:

$$\Psi_n - \|g\|_n^2 - \Psi + \|g\|_2^2 \geq (\sqrt{2}\eta + \frac{2\eta}{3})\frac{\sigma_L^2}{B_l} = (\sqrt{2}\eta + \frac{2\eta}{3})\frac{(\|h - h_0\|_\infty + B + \|g\|_\infty)^2 \|g\|_2^2}{(\|h - h_0\|_\infty + B + \|g\|_\infty)^2} = (\sqrt{2}\eta + \frac{2\eta}{3})\|g\|_2^2$$
$$\Psi_n - \|g\|_n^2 \geq \Psi - \|g\|_2^2 \cdot \left(1 - (\sqrt{2}\eta + \frac{2\eta}{3})\right)$$

Small Note

They choose η such that the coefficient of $\|g\|_2^2$ is equal to $\frac{3}{2}$. The η that satisfies this would be negative and thus not a viable solution. Moreover due to a small computational mistake in the proof of theorem A.2, the coefficient of $\|g\|_2^2$ should be equal to $\frac{3}{4}$. This would solve the proof in theorem A.2 and enables solving for a valid $\eta = \frac{12}{100}$. Nonetheless, this does not change the RHS used by the authors[3] but only the LHS. The event, probability and its respective bound that we will use are:

$$P\left(\Psi_n - \|g\|_n^2 \geq \Psi - \frac{3}{4}\|g\|_2^2\right) \leq \exp\left(-\frac{n\|g\|_2^2}{16(\|h - h_0\|_\infty + B + \|g\|_\infty)^2}\right) \tag{A.3}$$

[3] $\exp -\frac{1}{16}x \geq \exp -\frac{12}{100}x$

This is the most important result. It finds a bound for the prior probability of deviating from the primary component of the function. As $n \uparrow \infty$ then $\delta_n \downarrow 0$ and $\Pi \downarrow 0$. As more samples are observed, the higher the probability. Note that $\delta_n \downarrow 0$ much more slowly than Π. It is unclear how they use *assumption A.6* to 'revert the inequalities'.

Lemma A.8 ([63]Lemma 18)
Let $W \sim \Pi$ be a Gaussian Process, $h_\rho \in \mathcal{H}$ an arbitrary function. Then

$$\log \Pi(\{\|E(W-h_\rho)\|_2^2 \leq \underbrace{n^{-\frac{b+2p}{b+2p+1}}}_{\delta_n^2}\}) \lesssim -n^{\frac{1}{b+2p+1}} \asymp -n\delta_n^2$$

Moreover if $\|h_\rho\|_{\mathcal{H}}^2 \lesssim n^{\frac{1}{b+2p+1}}$

$$\implies \log \Pi(\{\|E(W-h_\rho)\|_2^2 \leq \underbrace{n^{-\frac{b+2p}{b+2p+1}}}_{\asymp \delta_n^2},\ \|W-h_\rho\|_2^2 \leq \underbrace{n^{-\frac{b}{b+2p+1}}}_{\asymp \tau_n^2}\}) \gtrsim -n^{\frac{1}{b+2p+1}} \asymp -n\delta_n^2$$

Proof. We can rewrite $W = \sum_{i=1} \lambda_i \epsilon_i \varphi_i$ where $\{\lambda_i, \varphi\}$ is the eigen decomposition of the kernel operator associated with $\mathcal{H}$.

$$\|EProj_J(W-h_\rho)\|_2^2 \lesssim \sum_{i=1}^{J} i^{-2p} \langle W-h_i, \varphi_i\rangle_2^2 = \sum_{i=1}^{J} i^{-2p}(\lambda_i\epsilon_i - \langle h_\rho, \varphi_i\rangle_2)^2$$
$$\|EProj_{>J}(W-h_\rho)\|_2^2 \lesssim (J+1)^{-2p} \|Proj_{>J}(W-h_\rho)\|_2^2$$

By the triangle inequality:$\|E(W-h_\rho)\|_2 \leq \|EProj_J(W-h_\rho)\|_2 + \|EProj_{>J}(W-h_\rho)\|_2$ and thus as the number of 'sieves' grows $\|EProj_{>J}(W-h_\rho)\| \to 0$ and

$$\|E(W-h_\rho)\|_2^2 \lesssim \lim_{J\to\infty} \|EProj_J(W-h_\rho)\|_2^2 \leq \sum_{i=1}^{J=\infty} i^{-2p}(\lambda_i\epsilon_i - \langle h_\rho, \varphi_i\rangle_2)^2 \tag{A.4}$$
$$= \sum_{i=1}^{J=\infty} (\lambda_i\epsilon_i i^{-p} - \left\langle \tilde{h}_\rho, \varphi_i \right\rangle_2)^2 = C \left\| \tilde{W} - \tilde{h}_\rho \right\|_2^2 \tag{A.5}$$

where $\tilde{h}_\rho = \sum_{i=1} i^{-p} h_i \varphi_i$ if $h_\rho = \sum_{i=1} h_i \varphi_i$. This can be seen as a more *regular* version of h_ρ which lives in an appropriate power space $\tilde{\mathcal{H}}$. Similarly, $\tilde{W} = \sum_{i=1} \lambda_i \epsilon_i i^{-p} \varphi_i$ is a Gaussian Process on the same power space. equation (A.4) shows that $\{\|E(W-h_\rho)\|_2 \leq \epsilon\} \supset \{||C\tilde{W} - \tilde{h}_\rho||_2 \leq \epsilon\}$.The authors then state: "By assumption A.6(RL) we can reverse the inequalities above with a different constant C' such that the inclusion also goes the other way":

$$\{||C'\tilde{W} - \tilde{h}_\rho||_2 \leq \epsilon\} \supset \{\|E(W-h_\rho)\|_2 \leq \epsilon\} \supset \{||C\tilde{W} - \tilde{h}_\rho||_2 \leq \epsilon\} \tag{A.6}$$

Then using Decentered Small Ball Lemma [19]:

$$\Pi(\left\{\left\|\tilde{W} - \tilde{h}_\rho\right\|_2 \leq \delta_n\right\}) \leq \underbrace{e^{-\frac{1}{2}\|\tilde{h}_j\|_{\tilde{\mathcal{H}}}^2}}_{\leq 1} \Pi\left(\left\{\left\|\tilde{W}\right\|_2 \leq \delta_n\right\}\right) \leq \Pi\left(\left\{\left\|\tilde{W}\right\|_2 \leq \delta_n\right\}\right) \leq e^{-C'n\delta_n^2}$$

Where the last inequality is due to corollary A.1. The decentered small ball lemma holds up to constants in both directions and thus:

$$e^{-C_2 n\delta_n^2} \leq \Pi(\left\{\left\|\tilde{W} - \tilde{h}_\rho\right\|_2\right) \leq e^{-C_2' n\delta_n^2}$$

Using the set inclusions of equation (A.6) and decentered small ball theorem we have proven the first statement:

$$\Pi(\{\|E(W-h_\rho)\|_2^2 \leq n^{-\frac{b+2p}{b+2p+1}}\}) \leq \Pi(\{\left\|\tilde{W} - \tilde{h}_\rho)\right\|_2^2 \leq n^{-\frac{b+2p}{b+2p+1}}\}) \leq e^{-C\cdot n^{-\frac{b+2p}{b+2p+1}}\cdot n}$$
$$\implies \log(\Pi(\{\|E(W-h_\rho)\|_2^2 \leq n^{-\frac{b+2p}{b+2p+1}}\})) \lesssim -n^{\frac{1}{b+2p+1}}$$

The second statement requires us to determine the prior probability of: $\{\|E(W-h_\rho)\|_2^2 \leq \underbrace{n^{-\frac{b+2p}{b+2p+1}}}_{\delta_n^2},\ \|W-h_\rho\|_2^2 \leq \underbrace{n^{-\frac{b}{b+2p+1}}}_{\epsilon_n^2}\}$. This will be done by determining the probability of $\Theta_{dh} \cap \Theta_{dz}$ which will imply the desired event.

1. $\Theta_{dz} = \left\{ \left\| \text{Proj}_{m'}(\tilde{W} - \tilde{h}_\rho) \right\|_2 \leq \frac{\delta_n}{2C}, \left\| \text{Proj}_{>m'}(\tilde{W} - \tilde{h}_\rho) \right\|_2 \leq \frac{\delta_n}{2C} \right\}$
2. $\Theta_{dh} = \left\{ \|\text{Proj}_{>m'}(W - h_\rho)\|_2 \leq \tau_m \right\}$

Due to equation (A.6), Θ_{dz} implies the first inequality of the set and the second is immediate from Θ_{dh}. For this reason, we aim to bound the probability of each.

$$\begin{aligned}
\log \Pi(\Theta_{dz}) :&= \log \Pi(\Theta_{dzl} \cap \Theta_{dzh}) \\
&= \log \Pi\left(\left\{ \left\| \text{Proj}_{m'}(\tilde{W} - \tilde{h}_\rho) \right\|_2 \leq \frac{\delta_n}{2C}, \left\| \text{Proj}_{>m'}(\tilde{W} - \tilde{h}_\rho) \right\|_2 \leq \frac{\delta_n}{2C} \right\} \right) \\
&\overset{4}{\geq} \log \Pi\left(\left\{ \left\| (\tilde{W} - \tilde{h}_\rho) \right\|_2 \leq \frac{\delta_n}{2C} \right\} \right) \gtrsim -n\delta_n^2
\end{aligned}$$

Moreover, by equation (A.6) on Θ_{dz}:

$$\max\left\{ \|E(\text{Proj}_{m'}(W - h_\rho))\|_2^2, \|E(W - h_\rho)\|_2^2 \right\} \leq \delta_n^2$$

Using (RL) assumption A.6:

$$\begin{aligned}
\delta_n^2 \geq \|E(W - h_\rho)\|_2^2 &\gtrsim m'^{-2p} \|\text{Proj}_{m'}(W - h_\rho)\|_2^2 \\
\implies \|\text{Proj}_{m'}(W - h_\rho)\|_2^2 &\lesssim m'^{2p} \delta_n^2 \asymp \tau_m^2
\end{aligned}$$

Once more, projections decrease distances thus $\{\|Proj_J h\| \leq \kappa\} \supset \{\|f\| \leq \kappa\}$. Combining this with the decentered small ball lemma:

$$\log \Pi(\Theta_{dh}) = \log \Pi\left(\left\{ \|\text{Proj}_{>m'}(W - h_\rho)\|_2 \leq \tau_m \right\} \right) \geq \log \Pi\left(\left\{ \|W - h_\rho\|_2 \leq \tau_m \right\} \right) \gtrsim -n^{\frac{1}{b+2p+1}} \asymp -n\delta_n^2$$

Notice that on the event Θ_{dh}:

$$\begin{aligned}
\left\| \text{Proj}_{>m'}(\tilde{W} - \tilde{h}_\rho) \right\|_2^2 &= \sum_{i=m'+1}^{\infty} i^{-2p} \left(\epsilon_i \lambda_i - \langle h_\rho, \varphi_i \rangle_2 \right)^2 \\
&\leq m'^{-2p} \sum_{i=m'+1}^{\infty} \left(\epsilon_i \lambda_i - \langle h_\rho, \varphi_i \rangle_2 \right)^2 \\
&= m'^{-2p} \underbrace{\|\text{Proj}_{>m'}(W - h_\rho)\|_2^2}_{\leq \tau_m^2} \\
&\lesssim n^{-\frac{2p}{b+2p+1} - \frac{b}{b+2p+1}} \asymp \delta_n^2
\end{aligned}$$

This shows that after rescaling $\Theta_{dh} \subset \Theta_{dzh} \implies \Theta_{dh} \cap \Theta_{dzh} = \Theta_{dh}$

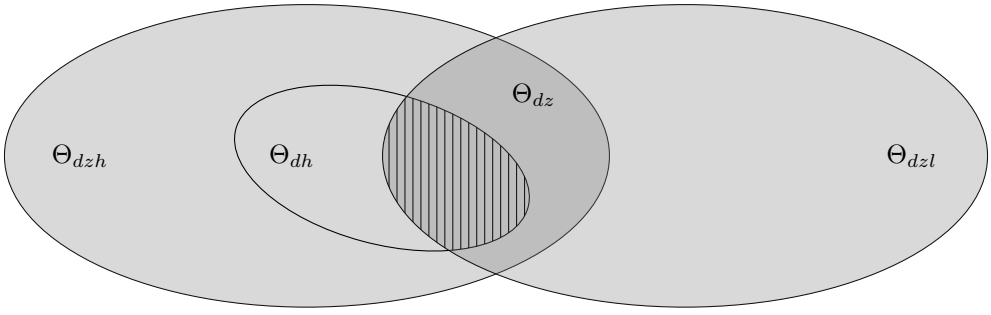


Figure A.2: The various sets visualized. $\Theta_{dh} \subset \Theta_{dzh}$ up to scaling constant. The barred area corresponds to $\Theta_{dh} \cap \Theta_{dz}$.

We are interested in $\Theta_{dh} \cap \Theta_{dz} = \Theta_{dh} \cap \Theta_{dzh} \cap \Theta_{dzl} = \Theta_{dh} \cap \Theta_{dzl}$. Moreover, Θ_{dh} only involves the first m' elements whereas the Θ_{dzl} the trailing ones, and thus have independent probabilities and factorize.

$$\log \Pi(\Theta_{dh} \cap \Theta_{dz}) = \underbrace{\log \Pi(\Theta_{dzl})}_{\gtrsim -n\delta_n^2} + \underbrace{\log \Pi(\Theta_{dh})}_{\gtrsim -n\delta_n^2} \gtrsim -n\delta_n^2$$

□

Lemma A.9 ([63]Lemma 19)
Let $\delta_n^2 = n^{-\frac{b+2p}{b+2p+1}}, \epsilon_n^2 = n^{-\frac{b}{b+2p+1}}$. *There exists constants* $c_1, \dots, c_4 > 0$ *such that:*

$$\Pi(\Theta_{d0}) \geq \exp\left(-c_4 n \delta_n^2\right)$$
$$\Theta_{d0} = \{h : \|E(h-h_0)\|_2 \leq c_1\delta_n \quad \|h-h_0\|_2 \leq c_2\epsilon_n \quad \|h-h_0\|_\infty \leq c_3\}$$

Proof. By assumption A.4, we can rewrite the true bridge function as the sum of two elements: $h_0 = \tilde{h}_\rho + \tilde{h}_e$ with:

$$\left\|\tilde{h}_\rho\right\|_{\mathcal{H}}^2 \lesssim m\tau_m^2 \quad \left\|\tilde{h}_e\right\|_2^2 \lesssim \tau_m^2 \quad \left\|\tilde{h}_e\right\|_\infty \leq 1$$

By lemma A.6 there is a decomposition $h_0 = h_\rho + h_e$ with

$$\|h_\rho\|_{\mathcal{H}}^2 \lesssim m\tau_m^2 \quad \|h_e\|_2^2 \lesssim \tau_m^2 \quad \|Eh_e\|_2^2 \lesssim \delta_n^2 \quad \|h_e\|_\infty \lesssim 1$$

Applying part two of lemma A.8 to h_ρ:

$$\|h-h_0\|_2 \lesssim \epsilon_n \quad \|E(h-h_0)\|_2 \lesssim \delta_n$$

have the required probability.
Define $C\Theta_m = \{h : \frac{h}{C} \in \Theta_m :\}$ $\Pi\left(\overline{C\Theta_m}\right) \leq \exp\left(C'' n\delta_n^2\right)$, we can choose C so that Θ_{d0} has the required probability. By corollary A.1, for any $h \in \Theta_{d0} \subseteq C\Theta_m$ we can write $h = \tilde{h}_\rho + \tilde{h}_e$ such that:

$$\left\|\tilde{h}_\rho\right\|_{\mathcal{H}}^2 \lesssim n^{\frac{1}{b+2p+1}} \quad \left\|\tilde{h}_e\right\|_2^2 \lesssim n^{-\frac{b}{b+2p+1}} \quad \left\|\tilde{h}_e\right\|_\infty^2 \lesssim 1$$

Always by assumption A.4 h_0 also admits such decomposition, thus the difference also admits such decomposition, i.e. $h - h_0 = h'_\rho + h'_e$. By the reverse triangle inequality[5]:

$$\left\|h'_\rho\right\|_2 \leq \underbrace{\left\|h'_\rho - h'_e\right\|_2}_{\lesssim \epsilon_n} + \underbrace{\|h'_e\|_2}_{\lesssim \epsilon_n} \lesssim \epsilon_n$$

Moreover, applying lemma A.3:

$$\|h-h_0\|_\infty^2 \lesssim \left\|h'_\rho\right\|_\infty^2 + \|h'_e\|_\infty^2 \lesssim \left\|h'_\rho\right\|_{\mathcal{H}}^{\frac{2b}{b+1}} \left\|h'_\rho\right\|_2^{\frac{2}{b+1}} + \|h'_e\|_\infty^2 \lesssim n^{\frac{1}{b+2p+1}\cdot\frac{b}{b+1}} \cdot n^{-\frac{b}{b+2p+1}\cdot\frac{1}{b+1}} + 1 \leq 2$$

□

[5] If $|A| \geq |B| \implies |A-B| \geq |A| - |B|$.

Here we define the $\mathcal{D}^n$ event on which the $\|E(h-h_0)\|_2^2$ is controlled. This enables us to have a lower bound bound for the loss. This in turn lets us control the normalizing term of the posterior (marginal over the entire parameter set).

Proposition A.4 ([63]Proposition 20)
Let $\lambda = 1$ then for all $\overline{\nu} \geq 2c\delta_n^2$ there exist $c_1, \dots c_3 > 0$ and a $\mathcal{D}^n$ measurable event $E_n(\mathcal{D}^n$ with probability tending to 1 in which:

$$\Pi(\{h \in \Theta_{d0} : l_n(f) \leq c_1 \|E(h-h_0)\|_2^2 + c_2\delta_n^2\}) \geq \frac{1}{2}\Pi(\Theta_{d0}) \tag{A.7}$$

Proof. Suppose that there exists constants $C_1, C_2 > 0$, and a sequence $\eta_n \to 0$ such that for any fixed $h \in \Theta_{d0}$ the event $I(h, \mathcal{D}^n) = \left\{h \in \Theta_{d0} : l_n(f) \leq C_1 \|E(h-h_0)\|_2^2 + C_2\delta^2\right\}$:

$$\mathbb{E}\left[I(h, \mathcal{D}^n)\right] \geq 1 - \eta_n$$

The claim is that if such conditions indeed hold, equation (A.7) holds with $\mathcal{D}^n$ probability greater than $1 - 4\eta_n$. The proof is by contradiction: Suppose that equation (A.7) holds with $\mathcal{D}^n$ probability less that $1 - 4\eta_n$. Define $\Pi_d(dh) = \frac{\Pi(dh \cap \Theta_{d0})}{\Pi(\Theta_{d0})}$, then:

$$\mathbb{E}\left[\Pi_d(I(h, \mathcal{D}^n))\right] = \iint \Pi_d(dh) dP^n = \iint dP^n \Pi_d(dh) \geq \int (1 - \eta_n)\Pi_d(dh) = 1 - \eta_n$$

Similarly, by law of total probabilities we can write:

$$\begin{aligned}
\mathbb{E}\left[\Pi_d(I(h, \mathcal{D}^n)))\right] &= \mathbb{E}\left[\Pi_d(I(h, \mathcal{D}^n)) \cdot \mathbf{1}_{34}\right] + P(\Pi_d(I(h, \mathcal{D}^n)) \cdot \mathbf{1}_{\overline{34}}) \\
&\overset{1}{\leq} \mathbb{E}\left[\mathbf{1}_{34}\right] + \frac{1}{2}\mathbb{E}\left[\mathbf{1}_{\overline{34}}\right] \\
&= 1 - 4\eta_n + \frac{1}{2} \cdot 4\eta_n = 1 - 2\eta_n
\end{aligned}$$

Where the last holds because if *equation* (A.7) does not hold it must be less than half. This contradicts the assumption. We must now show that such constants and a sequence η_n exist.
Let $\hat{\delta}_n$ be the empirical critical radius of $3\mathcal{G}_1$. Then, $P(\hat{\delta}_n \leq c\delta_n) \to 1$ and there exists $c' > 0$ such that with probability greater than $1 - c'e^{-c'n\delta_n^2} = 1 - \eta_n \to 1$ for a fixed $h \in \Theta_{d0}$ and all $g \in \mathcal{G}$ by theorem 2.8 we have:

$$\begin{aligned}
\|g\|_n^2 &\geq \frac{1}{2}\|g\|_2^2 - C\delta_n^2(1 + \|g\|_{\mathcal{G}}^2) \\
\Psi_n(f, g) &\leq \Psi(f, g) + 10LC\delta_n(\|g\|_2 + \delta_n(1 + \|g\|_{\mathcal{G}}))
\end{aligned}$$

The $\mathcal{D}^n$ event where this holds is defined as $E_n(\mathcal{D}^n)$. On this event, if $\overline{\nu} > 2C\delta_n^2$, the loss is upperbounded by:

$$l_n(f) \leq 2\|E(h-h_0)\|_2^2 + C_n'^2\delta_n^2$$

where $C_n'^2 \lesssim B^2$.
This proves that there exists an event $E_n(\mathcal{D}^n)$ with probability $1 - c'e^{-c'n\delta_n^2}$ on which equation (A.7) holds. □

[51] $P(A \cap B) \leq P(A)$

The authors show that on a subset Θ_m of the parameter space, the errors $\|E(h-h_0)\|_2$ and $\|h-h_0\|_2$ are equivalent. Since the loss estimates the first term, this shows that minimizing l_n also ensures minimization $\|h-h_0\|_2$.

Lemma A.10 ([63]Lemma 21)
There exist constants $M_0, C > 0$ such that $\forall n \in \mathbb{N},\ m = n^{-\frac{b+1}{b+2p+1}},\ f \in \Theta_m$:

$$\|h-h_0\|_2 \geq M_0\epsilon_n \quad \textit{only if } \|E(h-h_0)\|_2 \geq \frac{\|h-h_0\|_2}{C\epsilon_n}\delta_n$$
$$\|E(h-h_0)\|_2 \geq M_0\delta_n \quad \textit{only if } \|h-h_0\|_2 \geq \frac{\|E(h-h_0)\|_2}{C\delta_n}\epsilon_n$$

Proof. We can apply lemma A.6 to $\Delta h = h - h_0$ since $h \in \Theta_m$ by assumption of the lemma, and h_0 satisfies assumption A.4. Thus there exists $\tilde{\Delta h} \in \mathcal{H}$ such that:

$$\left\|\tilde{\Delta h}\right\|_{\mathcal{H}} \lesssim n\delta_n^2 \quad \left\|E(\Delta h - \tilde{\Delta h})\right\|_2^2 \lesssim \delta_n^2 \quad \left\|\Delta h - \tilde{\Delta h}\right\|_2^2 \lesssim \epsilon_n^2 \quad \left\|\Delta h - \tilde{\Delta h}\right\|_\infty^2 \lesssim 1$$

Let J be the first integer larger than $n^{\frac{1}{b+2p+1}}$, and using the following notation:

$$g = E\Delta h \quad \tilde{g} = E\tilde{\Delta h} \quad g_J = EProj_J\tilde{\Delta h}$$

I can choose $M_0 \geq \sqrt{C} + \frac{1}{2}$ by the triangle inequality,

$$\left\|\tilde{\Delta h}\right\|_2 \geq \|\Delta h\|_2 - \left\|\Delta h - \tilde{\Delta h}\right\|_2 \geq \|\Delta h\|_2 - \sqrt{C}\epsilon \geq \frac{1}{2}\|\Delta h\|_2$$
$$\left\|Proj_J\tilde{\Delta h}\right\|_2^2 = \left\|\tilde{\Delta h}\right\|_2^2 - \left\|Proj_{>J}\tilde{\Delta h}\right\|_2^2 \geq \frac{1}{4} - \|\Delta h\|_2^2 - \left\|\tilde{\Delta h}\right\|_{\mathcal{H}}^2 J^{-(b+1)} \geq \frac{1}{5}\|\Delta h\|_2^2$$

where the penultimate inequality follows from lemma A.5. Now using assumption A.6

$$\begin{aligned}
\|E\Delta h\|_2^2 &\geq \frac{1}{2}\left\|E\tilde{\Delta h}\right\|_2^2 - \left\|E\left(\Delta h - \tilde{\Delta h}\right)\right\|_2^2 \\
&\geq C_1' J^{-2p}\left\|Proj_J\tilde{\Delta h}\right\|_2^2 - \left\|E\left(\Delta h - \tilde{\Delta h}\right)\right\|_2^2 \\
&\geq \frac{1}{5}C_1' J^{-2p}\|\Delta h\|_2^2 - \left\|E\left(\Delta h - \tilde{\Delta h}\right)\right\|_2^2 \\
&= \frac{\delta_n^2}{\epsilon_n^2}\|\Delta h\|_2^2 - \delta_n^2 \gtrsim \frac{\delta_n^2}{\epsilon_n^2}\|\Delta h\|_2^2
\end{aligned}$$

This proves the first statement. To now prove the second: For any $h \in \Theta_m$, using assumption A.6 and lemma A.5:

$$\begin{aligned}
\|g - g_j\|_2^2 \leq 2\|g-\tilde{g}\|_2^2 + \|\tilde{g} - g_J\|_2^2 &\lesssim \delta_n^2 + \left\|EProj_{>J}\tilde{\Delta h}\right\|_2^2 \\
&\lesssim \delta_n^2 + J^{-2p}\left\|Proj_{>J}\tilde{\Delta h}\right\|_2^2 \leq \delta_n^2 + J^{-2p}\cdot J^{-(b+1)}\left\|\tilde{\Delta h}\right\|_{\mathcal{H}}^2 \\
&\lesssim \delta_n^2 + n^{-\frac{2p-1}{b+2p+1}-\frac{b+1}{b+2p+1}} \asymp \delta_n^2
\end{aligned}$$

By the assumption $\|g\|_2 \geq M_0\delta_n$, I can choose $M_0 \geq$

$$\left\|\tilde{\Delta h}\right\|_2 \gtrsim J^p\|g_J\|_2 \geq J^p\left(\|g\|_2 - \|g-g_J\|_2\right) \geq \frac{\|g\|_2}{2\delta_n}\epsilon_n \tag{A.8}$$
$$\left\|\tilde{\Delta h}\right\|_2 \geq \left\|\tilde{\Delta h}\right\|_2 - \left\|\Delta h - \tilde{\Delta h}\right\|_2 \gtrsim \left(\frac{\|g\|_2}{2\delta_n} - \sqrt{C}\right)\epsilon_n \geq \frac{\|g\|_2}{3\delta_n} \tag{A.9}$$

□

[5] $|A-B|^2 \leq 2A^2 + 2B^2$

This is shown by proving a stronger result: convergence of in L^1. By Markov's inequality this automatically implies convergence in probability.

Theorem A.2 ([63])
Fix $\lambda = 1, \overline{\nu} = C\delta_n^2 \asymp n^{-\frac{b+2p}{b+2p+1}}, C > 0$. *Then there exists a constant* $M > 0$ *such that for* $\epsilon_n^2 = n^{-\frac{b}{b+2p+1}}$ *we have:*

$$\begin{aligned} &\Pi(\{h : \|h - h_0\|_2^2 \geq M\epsilon_n^2\}|\mathcal{D}^n) \xrightarrow{P^{\mathcal{D}^n}} 0 \\ &\Pi(\{h : \|E(h - h_0)\|_2^2 \geq M\delta_n^2\}|\mathcal{D}^n) \xrightarrow{P^{\mathcal{D}^n}} 0 \end{aligned} \tag{A.10}$$

Proof. Define the set of functions where the violation from the true bridge function is greater than $M\epsilon_n^2$ as:

$$err_{fn} = \left\{h : \|h - h_0\|_2^2 \geq M\epsilon_n^2\right\}$$

As shown in the proof of *proposition A.4* there exists an event $E_n(\mathcal{D}^n)$ on which, with probability $\to 1$:

$$\Pi(\{h : l_n(f) \leq c_1 \|E(h - h_0)\|_2^2 + c_2\delta_n^2\}) \geq \frac{1}{2}\Pi(\Theta_{d0}) \geq \frac{1}{2}e^{-c_3 n\delta_n^2}$$

Using the law of total probabilities we can further decompose equation (A.10) as:

$$\begin{aligned} \mathbb{E}_{\mathcal{D}^n}\left[\Pi(err_{h,n}|\mathcal{D}^n)\right] &= \mathbb{E}_{\mathcal{D}^n}\left[\mathbf{1}_{E_n(\mathcal{D}^n)} \cdot \Pi(err_{h,n}|\mathcal{D}^n)\right] + \mathbb{E}_{\mathcal{D}^n}\left[\mathbf{1}_{\overline{E_n(\mathcal{D}^n)}} \cdot \Pi(err_{h,n}|\mathcal{D}^n)\right] \\ &\leq \mathbb{E}_{\mathcal{D}^n}\left[\mathbf{1}_{E_n(\mathcal{D}^n)} \cdot \Pi(err_{h,n}|\mathcal{D}^n))\right] + \mathbb{E}_{\mathcal{D}^n}\left[\mathbf{1}_{\overline{E_n(\mathcal{D}^n)}})\right] \\ &= \mathbb{E}_{\mathcal{D}^n}\left[\mathbf{1}_{E_n(\mathcal{D}^n)} \cdot \Pi(err_{h,n}|\mathcal{D}^n))\right] + o_{P^{\mathcal{D}^n}}(1) \end{aligned}$$

Since the quasi bayesian loss $Q\left(D^n, h\right) = e^{-nl_n(f)}$ might not integrate to one, we can rewrite $\Pi(err_{h,n}|\mathcal{D}^n)$ as:

$$\Pi_n(err_{h,n}|\mathcal{D}^n) = \frac{\int_{err_{h,n}} Q\left(D^n, h\right)\Pi(dh)}{\int_\Theta Q\left(D^n, h\right)\Pi(dh)}$$

By proposition A.4 we already have a lower bound on the denominator: $e^{-C_{den}n\delta_n^2}$ with $C_{den} \lesssim B^2$ and thus we only need to focus on the numerator:

$$\mathbb{E}_{\mathcal{D}^n}\left[\mathbf{1}_{E_n(\mathcal{D}^n)} \cdot \int_{err_{h,n}} e^{-nl_n(f)}\Pi(dh)\right] = \int_{E_n(\mathcal{D}^n)}\left[\int_{err_{h,n}} Q\left(D^n, h\right)\Pi(dh)\right] dP_{\mathcal{D}^n}$$

We need to find a bound that depends on M such that $C(M) \geq C_{den}$. Using the fact that $Q\left(D^n, h\right) \leq 1$, the law of total probabilities and Fubini:

$$\begin{aligned} \int_{E_n(\mathcal{D}^n)}\left[\int_{err_{h,n}} Q\left(D^n, h\right)\Pi(dh)\right] dP_{\mathcal{D}^n} &= \int_{E_n(\mathcal{D}^n)}\left[\int_{err_{h,n}\cap\Theta_m} Q\left(D^n, h\right)\Pi(dh) + \int_{err_{h,n}\cap\overline{\Theta}_m} Q\left(D^n, h\right)\Pi(dh)\right] dP_{\mathcal{D}^n} \\ &\leq \int_{E_n(\mathcal{D}^n)}\left[\int_{err_{h,n}\cap\Theta_m} Q\left(D^n, h\right)\Pi(dh) + \int_{err_{h,n}\cap\overline{\Theta}_m} \Pi(dh)\right] dP_{\mathcal{D}^n} \\ &\leq \int_{E_n(\mathcal{D}^n)}\left[\int_{err_{h,n}\cap\Theta_m} Q\left(D^n, h\right)\Pi(dh) + \int_{\overline{\Theta}_m} \Pi(dh)\right] dP_{\mathcal{D}^n} \\ &\leq \int_{E_n(\mathcal{D}^n)}\left[\int_{err_{h,n}\cap\Theta_m} Q\left(D^n, h\right)\Pi(dh)\right] dP_{\mathcal{D}^n} + \Pi(\overline{\Theta}_m) \\ &\leq \int_{err_{h,n}\cap\Theta_m}\left[\int_{E_n(\mathcal{D}^n)} Q\left(D^n, h\right) dP_{\mathcal{D}^n}\right]\Pi(dh) + \Pi(\overline{\Theta}_m) \end{aligned}$$

The second term goes to zero at a rate of $e^{-Cn\delta_n^2}$. Using the law of total probabilities once more, the first term will be divided over an event $A(h, \mathcal{D}^n)$ and its complement such that:

- On $A(h, \mathcal{D}^n)$ the loss can be lower bounded.

- $\overline{A}(h, \mathcal{D}^n)$ has small probability.

$$\int_{err_{h,n} \cap \Theta_m} \left[\int_{E_n(\mathcal{D}^n)} Q\left(D^n, h\right) dP_{\mathcal{D}^n} \right] \Pi(dh) \leq \underbrace{\int_{err_{h,n} \cap \Theta_m} \int_{E_n(\mathcal{D}^n) \cap A(h,\mathcal{D}^n)} Q\left(D^n, h\right) dP_{\mathcal{D}^n} \Pi(dh)}_{T_1} + \underbrace{\int_{err_{h,n} \cap \Theta_m} \int_{E_n(\mathcal{D}^n) \cap \overline{A}(h,\mathcal{D}^n)} dP_{\mathcal{D}^n} \Pi(dh)}_{T_2} \tag{A.11}$$

The goal is to use the concentration inequality of equation (A.3).

$$l_n(f) = \sup_{g \in \mathcal{G}} \Psi_n 2(h, g) - \|g\|_n^2 - \overline{\nu} \|g\|_{\mathcal{G}}^2$$

and:

$$g = E(h - h_0) \quad \tilde{g} = E(\tilde{\Delta} h) \quad g_J = E(Proj_J \tilde{\Delta} h)$$

For $h \in \Theta_m$ we define $A(h, \mathcal{D}^n) = \left\{ \Psi_n(f, g_J) - \|g_J\|_n^2 \geq \Psi(f, g_J) - \frac{3}{4} \|g_J\|_2^2 \right\}$. On A:

$$\begin{aligned}
l_n(f) &= \sup_{g \in \mathcal{G}} 2\Psi_n(f, g) - \|g\|_n^2 - \overline{\nu} \|g\|_{\mathcal{G}}^2 \\
&\geq 2\Psi_n(f, g_J) - \|g_J\|_n^2 - \overline{\nu} \|g_J\|_{\mathcal{G}}^2 && (g_J \in \mathcal{G}) \\
&\geq 2\Psi(f, g_J) - \frac{3}{2} \|g_J\|_2^2 - \overline{\nu} \|g_J\|_{\mathcal{G}}^2 && \text{(We are in A)} \\
&= 2\mathbb{E}\left[(h(V_h) - Y) g_J(Z)\right] - \frac{3}{2} \|g_J\|_2^2 - \overline{\nu} \|g_J\|_{\mathcal{G}}^2 && \text{(Definition)} \\
&\overset{6}{\geq} 2 \|g\|_2^2 - 2 \|g\|_2 \|g - g_J\|_2 - \frac{3}{2} \|g_J\|_2^2 - \overline{\nu} \|g_J\|_{\mathcal{G}}^2 && \text{(A.12)} \\
&\geq 2 \|g\|_2^2 - 2 \|g\|_2 \|g - g_J\|_2 - \frac{3}{2}(\|g\|_2^2 + \|g - g_J\|_2^2 + 2 \|g\|_2 \|g - g_J\|_2) - \overline{\nu} \|g_J\|_{\mathcal{G}}^2 \\
&\geq C_1 \|g\|_2^2 - C_2 \|g - g_J\|_2^2 - \overline{\nu} \|g_J\|_{\mathcal{G}}^2 && \text{(A.13)}
\end{aligned}$$

Now we define $U = \frac{\|h - h_0\|_2}{\epsilon_n}$ so that on $err_{h,n}$ $U \geq \sqrt{M}$. For sufficiently large M, i.e. $M > 1$ then $U > 1$:

$$U\delta_n = \|h - h_0\|_2 \frac{\delta_n}{\epsilon_n} \lesssim \|g\|_2 \leq \|h - h_0\|_2 = U\epsilon_n \leq U \tag{A.14}$$

This is because the sets $\left\{ h : \|h - h_0\|_2^2 \geq M\epsilon_n^2 \right\}$ and $\left\{ h : \|E(h - h_0)\|_2^2 \geq M\delta_n^2 \right\}$ are equivalent up to constants and both δ_n, ϵ_n are decreasing.

$$\left\| \tilde{\Delta} h \right\|_2 = \left\| \tilde{\Delta} h \pm h \pm h_0 \right\|_2 \leq \underbrace{\left\| h - h_0 - \tilde{\Delta} h \right\|_2}_{\leq C\epsilon_n \ \forall h \in \Theta_m} + \underbrace{\|h - h_0\|_2}_{= U\epsilon_n} \leq 2U\epsilon_n \tag{A.15}$$

To apply equation (A.3), we need to find the bounds for $\|h - h_0\|_\infty, \|g_J\|_\infty$. To bound the first we apply the triangle inequality and remember that in Θ_m the error is bounded by one. Then apply lemma A.2

$$\begin{aligned}
\|h - h_0\|_\infty^2 &\leq 2 \left\| \tilde{\Delta} h \right\|_\infty + 2 \left\| h - h_0 - \tilde{\Delta} h \right\|_\infty^2 \\
&\lesssim \left\| \tilde{\Delta} h \right\|_\infty + 1 && \text{(In } \Theta_m\text{)} \\
&\lesssim \left(\left\| \tilde{\Delta} h \right\|_{\mathcal{H}}^2 \right)^{\frac{b}{b+1}} \cdot \left(\left\| \tilde{\Delta} h \right\|_2^2 \right)^{\frac{1}{b+1}} + 1 && \text{(lemma A.3)} \\
&\lesssim n^{\frac{1}{b+2p+1} \cdot \frac{b}{b+1}} \cdot U^{\frac{2}{b+1}} n^{-\frac{b}{b+2p+1} \cdot \frac{1}{b+1}} + 1 \leq 2U^{\frac{2}{b+1}}
\end{aligned}$$

To bound the conditional expectation of $\tilde{\Delta h}$ in $\mathcal{G}$ we make use of assumption A.7, which ensures that the image of $\overline{\mathcal{H}}$ under E is bounded and lemma A.2

$$\begin{aligned}
\|g_J\|_{\mathcal{G}}^2 = \left\|EProj_J\tilde{\Delta h}\right\|_{\mathcal{G}}^2 &\lesssim \left\|Proj_J\tilde{\Delta h}\right\|_{\overline{\mathcal{H}}}^2 && \text{(assumption A.7)}\\
&\lesssim \left\|Proj_J\tilde{\Delta h}\right\|_{\mathcal{H}}^{\frac{2b}{b+1}} \cdot \left\|Proj_J\tilde{\Delta h}\right\|_2^{\frac{2}{b+1}}\\
&\leq \left\|Proj_J\tilde{\Delta h}\right\|_{\mathcal{H}}^{\frac{2b}{b+1}} \cdot \left\|\tilde{\Delta h}\right\|_2^{\frac{2}{b+1}}\\
&\lesssim \left\|Proj_J\tilde{\Delta h}\right\|_{\mathcal{H}}^{\frac{2b}{b+1}} \cdot (U\epsilon_n)^{\frac{2}{b+1}}\\
&\lesssim n^{\frac{1}{b+2p+1}\cdot\frac{b}{b+1}} \cdot U^{\frac{2}{b+1}} n^{-\frac{b}{b+2p+1}\cdot\frac{1}{b+1}} \leq 2U^{\frac{2}{b+1}} = U^{\frac{2}{b+1}}
\end{aligned}$$

Bounding T_1

The previous lower bound for the loss (equation (A.13)) can be written in terms of U as:

$$l_n(h) \geq C_1U^2\delta_n^2 - C_2\delta_n^2 - C_3U^{\frac{2}{b+1}} \geq C'U^2\delta_n^2 \geq C'M^2\delta_n^2$$

This enables me to upperbound T_1 as:

$$T_1 = \int_{E_n(\mathcal{D}^n)\cap\overline{A}(h,\mathcal{D}^n)} Q\left(D^n, h\right) dP_{\mathcal{D}^n} \leq \int_{\overline{A}(h,\mathcal{D}^n)} Q\left(D^n, h\right) dP_{\mathcal{D}^n} \leq \exp\left(-C'M^2n\delta_n^2\right)$$

Bounding T_2

Notice that $T_2 \leq P(\overline{A})$. Replacing all of the above bounds in equation (A.3):

$$-\log\ P(\overline{A}) \leq \frac{n\left\|g_J\right\|_2^2}{16\left(\left\|h-h_0\right\|_\infty + B + \left\|g_J\right\|_\infty\right)^2} \gtrsim \frac{U^2}{(U^{\frac{1}{b+1}}+B)^2} n\delta_n^2$$

Using the fact that $P(\overline{A}\cap E_n(\mathcal{D}^n)) \leq P(\overline{A})$ and the Continuous Mapping theorem, the proof is complete. □

A.4 Numerical-Experiments

In the simulated Demand experiment, around 40% of the data is larger than the 30 cut off chosen by the authors and around 99% of it is contained in the [10,40] interval.[7] For this reason it is very odd that

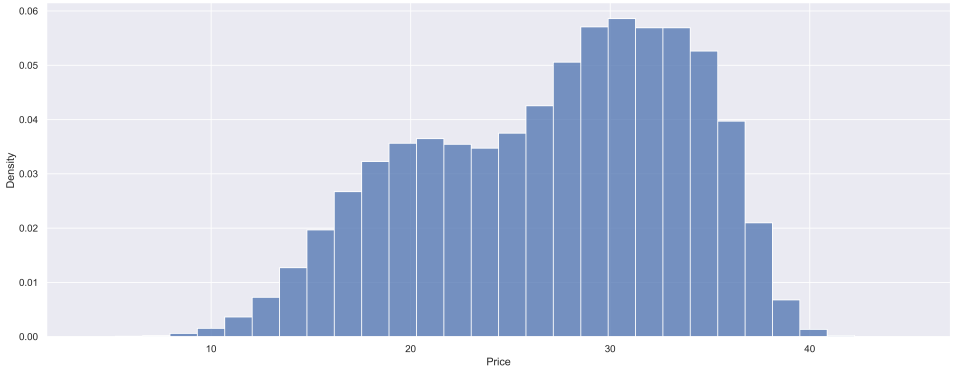


Figure A.3: Marginal distribution of the Price variable in the Demand simulated experiment.

the previous papers only consider the [10,30] interval. The CMSE of table A.1, are indeed calculated on data sampled from the entire distribution but figures 6.9 to 6.11 only show the predictions on the [10,30] interval. To highlight the behavior of the models on the support of the Price variable, we include Figure A.4. Here it is possible to observe that the performance of KPV,PMMR,DFPV deviates from the truth on the latter half of the graph. The other methods are instead able to correctly approximate the true CERF.

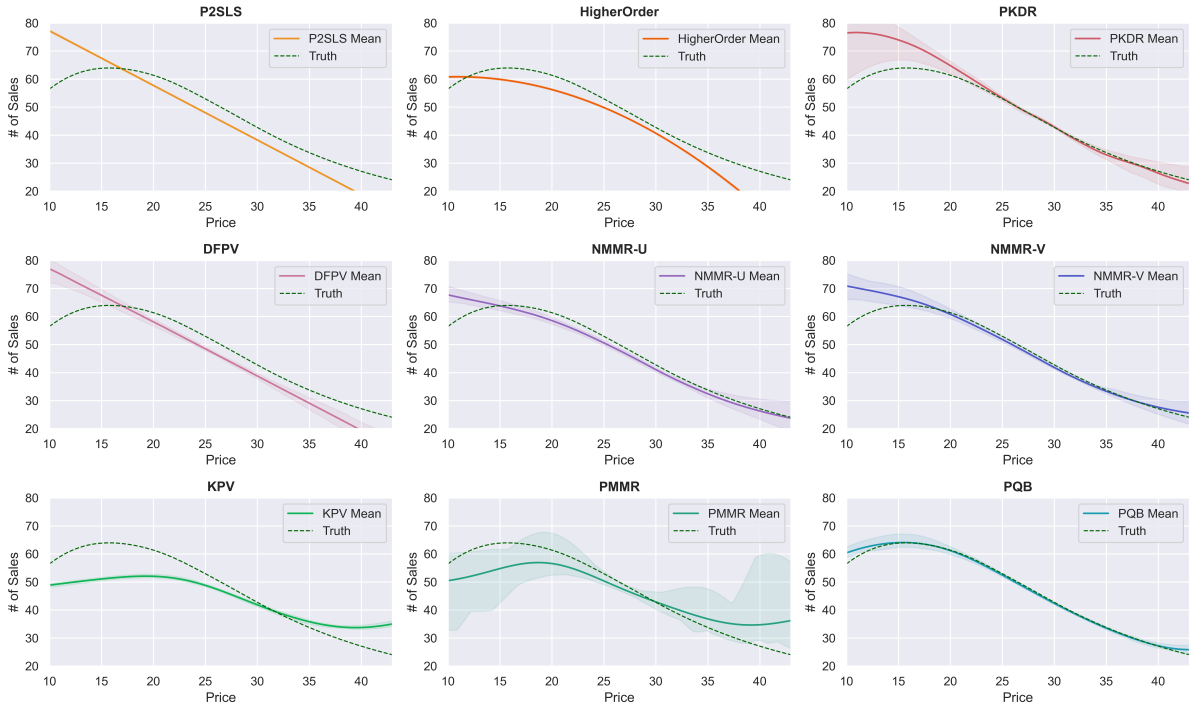


Figure A.4: Model predictions on the entire support of the price variable of the Demand experiment. Trained on 10,000 data points.

In Figure A.5 it is possible to observe the behavior of the individual runs of each model in the different sample regimes.

[7] Simulated 1,000,000 data points.

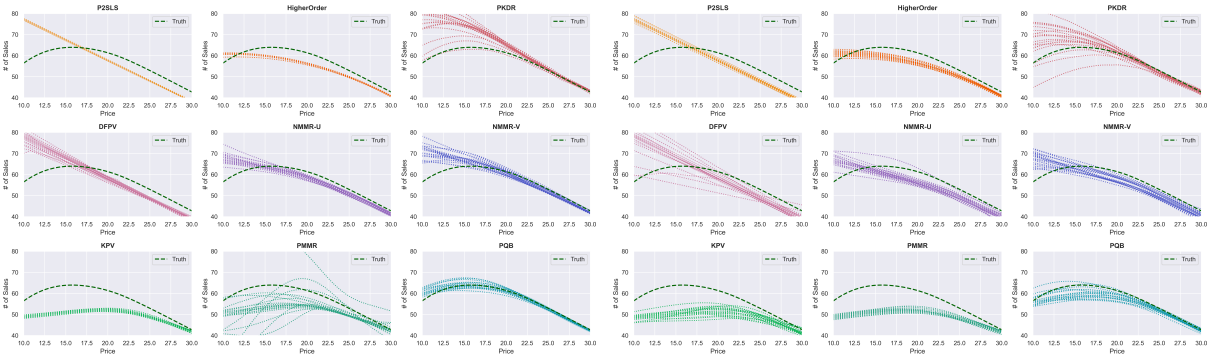


Figure A.5: Individual model predictions for each seed. Trained on 10,000 data points.

Figure A.6: Individual model predictions for each seed. Trained on 1,000 data points.

When evaluating a model, the average error the model makes is a possible measure of error. The mean and standard deviations are only two statistics. By the central limit theorem, these characterize the limiting distribution of the average error. In reality, the model is only fitted once and thus the distribution of the error is much more important than its expected error.[8]. In reality, a model is better performing than another if it has a *'tighter'* distribution of the error. A better measure for this is the error quantile range. We note that the sample size of 20 is still small for this metric to be definitive but still highlights the model's ability to consistently perform well. The interquantile ranges $[0.25, 0.5, 0.75]$ are reported in Table A.2. The PQB also outperforms other methods in this metric.

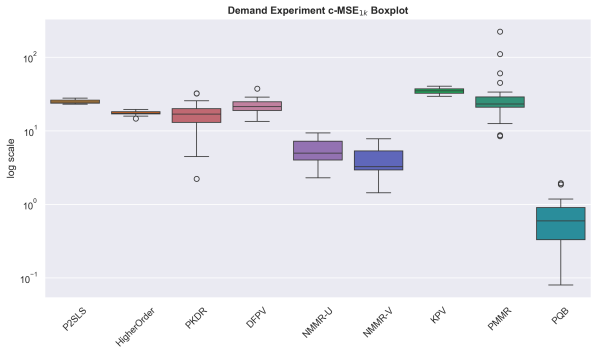


Figure A.7: Boxplot of the c-MSE$_{1k}$ for the Demand experiment with(lower is better).

[8]Thank you Matthias for the discussion on this.

Model	**c-MSE**$_{10}$ ($train_n = 10,000$)	**c-MSE**$_{10}$ ($train_n = 5,000$)	**c-MSE**$_{10}$ ($train_n = 1,000$)
P2SLS	24.89 ± 14.00	24.68 ± 13.35	24.56 ± 16.73
HigherOrder	17.38 ± 13.53	17.22 ± 13.37	16.57 ± 13.18
PKDR	13.36 ± 26.51	17.91 ± 39.06	4.99 ± 12.63
DFPV	22.99 ± 23.65	71.42 ± 439.45	26.55 ± 40.81
NMMR-U	4.88 ± 5.11	7.55 ± 8.56	17.83 ± 26.62
NMMR-V	3.28 ± 5.57	5.13 ± 7.05	9.35 ± 10.85
KPV	34.45 ± 31.92	37.80 ± 32.45	41.62 ± 46.69
PMMR	36.02 ± 77.47	26.19 ± 23.80	37.46 ± 27.60
PQB	$\mathbf{0.59 \pm 0.90}$	$\mathbf{0.99 \pm 1.35}$	$\mathbf{3.61 \pm 6.00}$

Table A.1: Model performance using Causal Mean Error on 10 test points.

Model	**QR** ($train_n = 10,000$)	**QR** ($train_n = 5,000$)	**QR** ($train_n = 1,000$)
P2SLS	$[23.79, 24.65, 26.22]$	$[23.75, 24.74, 26.21]$	$[22.51, 24.02, 26.29]$
HigherOrder	$[16.96, 17.36, 18.20]$	$[16.18, 17.74, 18.94]$	$[16.30, 16.78, 19.37]$
PKDR	$[13.07, 16.90, 20.03]$	$[11.40, 17.46, 22.24]$	$[2.78, 4.01, 9.06]$
DFPV	$[18.92, 21.39, 24.92]$	$[17.75, 22.18, 22.96]$	$[19.63, 26.01, 33.01]$
NMMR-U	$[4.00, 4.97, 7.25]$	$[5.63, 7.18, 9.82]$	$[10.13, 14.40, 23.20]$
NMMR-V	$[2.95, 3.25, 5.35]$	$[3.68, 4.43, 6.81]$	$[5.14, 10.34, 12.87]$
KPV	$[32.40, 34.92, 37.40]$	$[35.79, 38.99, 41.89]$	$[30.91, 42.86, 48.01]$
PMMR	$[20.96, 23.18, 28.87]$	$[22.56, 25.86, 28.18]$	$[35.64, 39.42, 42.04]$
QB	$[\mathbf{0.33}, \mathbf{0.60}, \mathbf{0.91}]$	$[\mathbf{0.62}, \mathbf{1.08}, \mathbf{1.52}]$	$[\mathbf{1.75}, \mathbf{2.96}, \mathbf{6.46}]$

Table A.2: Model performance using the quantile range for the c-MSE$_{1k}$ of the various models in the Demand Experiment.

A.5 Various Results

Lemma A.11
Let X,Y,Z be random variables taking values on $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ respectively. If $X \perp\!\!\!\perp Y|Z$, then:

$$\mathbb{E}\left[Y \cdot X \cdot 1_{Z=z}\right] = \mathbb{E}\left[Y \cdot 1_{Z=z}\right] \cdot \mathbb{E}\left[X|Z=z\right]$$

Proof.

$$\begin{aligned} dP(x,y,z) &= dP(x,y|z)dP(z) \\ &= dP(x|z)dP(y|z)dP(z) && (X \perp\!\!\!\perp Y|Z) \\ &= dP(x|z)dP(y|z)dP(z) \end{aligned}$$

$$\begin{aligned} \mathbb{E}\left[Y \cdot 1_{Z=z}\right] &= \int_{\mathcal{Y}\times\{Z=z\}} y dP(y,z) = \int_{\mathcal{Y}\times\{Z=z\}} y dP(y|z)dP(z) \\ \mathbb{E}\left[X|Z=z\right] &= \int_{\mathcal{X}} x dP(x|z) \end{aligned}$$

$$\begin{aligned} \mathbb{E}\left[Y \cdot 1_{Z=z}\right] \cdot \mathbb{E}\left[X|Z=z\right] &= \int_{\mathcal{Y}\times\{Z=z\}} y dP(y|z)dP(z) \cdot \int_{\mathcal{X}} x dP(x|z) \\ &= \int_{\mathcal{X}\times\mathcal{Y}\times\{Z=z\}} xy dP(x,y,z) \\ &= \int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} xy 1_{Z=z} dP(x,y,z) = \mathbb{E}\left[Y \times X \times 1_{Z=z}\right] \end{aligned}$$

□

Proposition A.5
Under the assumptions of subsection 3.4.2 we have that: $\left\|\hat{T}^*\hat{T} - T^*T\right\| = O(n^{-\frac{1}{2}})$ and $\left\|\hat{T}^*\hat{g} - \hat{T}^*\hat{T}h_0\right\| = O(n^{-\frac{1}{2}})$

This proof is adapted from [36].

Proof. Applying the triangle inequality multiple times:

$$\begin{aligned} \left\|\hat{T}^*\hat{T} - T^*T\right\| &= \left\|\hat{T}^*\hat{T} \pm \hat{T}^*T - T^*T\right\| \leq \left\|\hat{T}^*\hat{T} - \hat{T}^*T\right\| + \left\|\hat{T}^*T - T^*T\right\| \\ &= \left\|\hat{T}^*(\hat{T}-T)\right\| + \left\|(\hat{T}^*-T^*)T\right\| \\ &\leq \left\|(\hat{T}^*-T^*)(\hat{T}-T)\right\| + \left\|T^*(\hat{T}-T)\right\| + \left\|(\hat{T}^*-T^*)T\right\| \\ &\leq \left\|(\hat{T}-T)\right\|^2 + 2\left\|T\right\|\left\|(\hat{T}-T)\right\| = O\left(\frac{1}{n} + \frac{\|T\|}{\sqrt{n}}\right) \end{aligned}$$

where the last equality follows from Proposition 2.5. By Assumption 3.7 we have that $\|T\| < \infty$ and thus the over-all rate of convergence is $\sqrt{n}$. Similarly the second result can be decomposed in three terms:

$$\begin{aligned} \left\|\hat{T}^*\hat{g} - \hat{T}^*\hat{T}h_0\right\| &\leq \left\|\hat{T}^*\hat{g} - \hat{T}^*g\right\| + \left\|\hat{T}^*g - \hat{T}*g\right\| + \left\|\hat{T}^*Th_0 - \hat{T}^*\hat{T}h_0\right\| \\ &\leq \left\|\hat{T}^*\right\|\|\hat{g}-g\| + \left\|\hat{T}^*-\hat{T}\right\|\|g\| + \left\|\hat{T}^*T - \hat{T}^*\hat{T}\right\|\|h_0\| \end{aligned}$$

By similar reasoning as above and invoking Assumption 3.7 all above terms converge at rate $\sqrt{n}$. □

Proposition A.6
Let K be a compact self-adjoint operator and $\{\lambda_i^2, \varphi_i\}$ be its eigen-decomposition. Then:

$$\|K\| = \sup_i\{\lambda_i^2\}$$

Proof. By definition of $\|K\| = \sup\left\{\frac{\|Kv\|}{\|v\|} \quad v \in \mathcal{H}\right\}$, $Kv = \sum \lambda_i^2 \langle v, \varphi_i \rangle \varphi_i$ and its norm is $\sum \lambda_i^2 \langle v, \varphi_i \rangle$. The sup is reached by choosing v such that it is parallel to the eigenfunction associated with the largest eigenfunction and orthogonal to all others. This completes the proof. □

Corollary A.2
Let K be a compact self-adjoint operator and $\{\lambda_i^2, \varphi_i\}$ its eigen decomposition. For $R > 0$, if $(K + RI)$ is invertible, we have:

$$\left\|(K + RI)^{-1}\right\| \leq \frac{1}{R}$$

Proof. The spectrum of $K + RI$ is $\{\lambda_i^2 + R\}$. Thus if $(K + RI)$ is invertible, its spectrum is given by $\{\frac{1}{\lambda_i^2 + R}\}$. Then:

$$\left\|(K + RI)^{-1}\right\| = \sup_i \frac{1}{\lambda_i^2 + R} \leq \frac{1}{R}$$

□