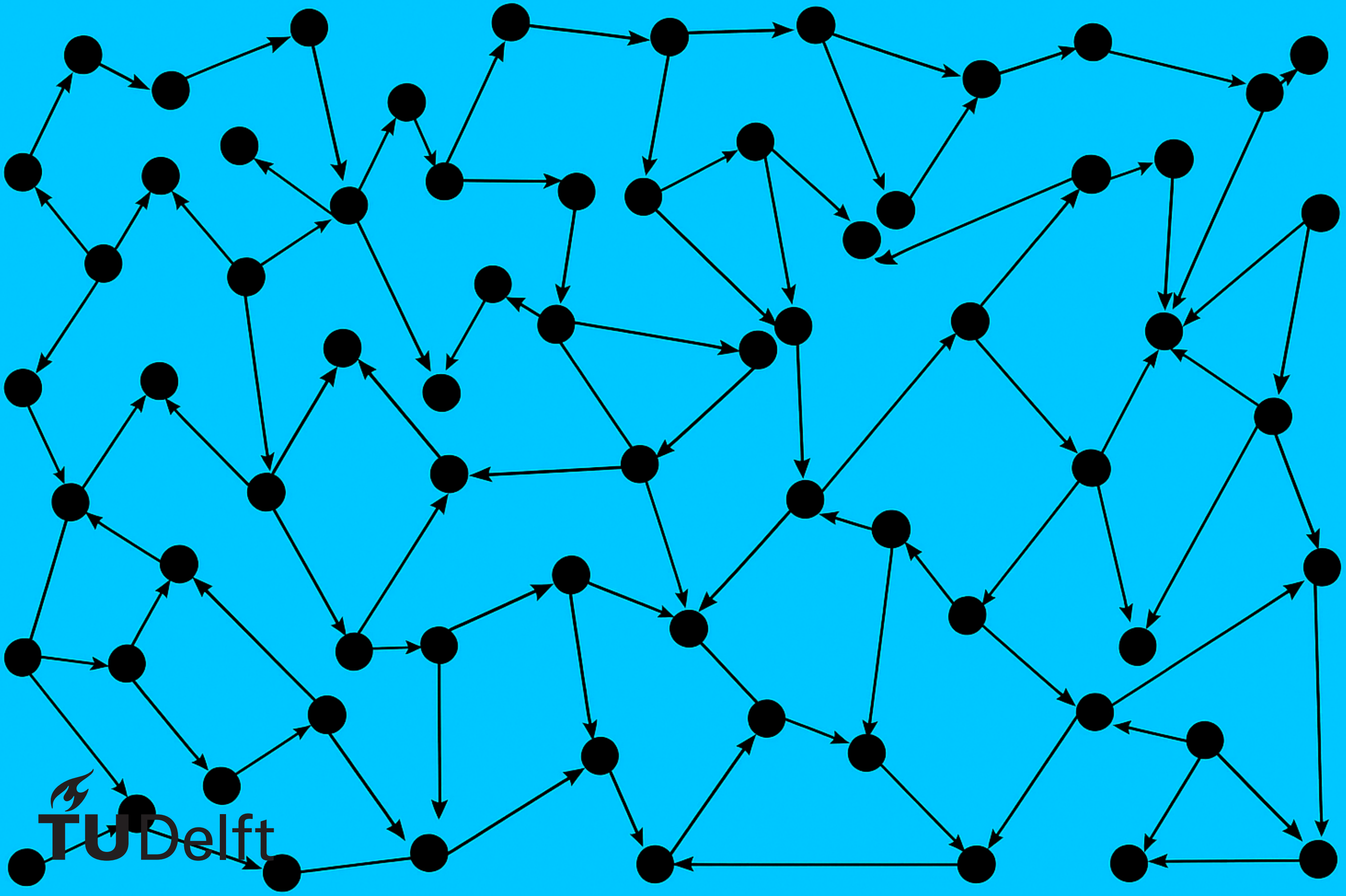
Estimating the treatment confounding bridge function through a dual kernel embedding method

Wei Liu

Student number: 5912075



Estimating the treatment confounding bridge function through a dual kernel embedding method

by

Wei Liu

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Monday August 25, 2025 at 13:00 PM.

Student number:	5912075	
Project duration:	December 17, 2024 – August 25, 2025	
Thesis committee:	Prof. dr. A.W. van der Vaart,	TU Delft, supervisor
	Dr. ir. J.H. Krijthe	TU Delft

Abstract

In proximal causal inference framework, the identification of average treatment effect (ATE) depends on finding the bridge functions. The bridge functions are functions about proxy variables used in the proximal standardization formulae. They are the solutions to two Fredholm integral equations of the first kind, whose existence is determined by Picard's conditions about the singular systems of two conditional expectation operators. However, since singular systems required by Picard's conditions are hard to determine, it is an extremely tough task to solve the bridge functions directly from the integral equations. Therefore, people turn to find estimators of the bridge functions. Many literatures have provided approaches to the estimators under certain assumptions although which inevitably restrict the feasibility of their application. In this thesis, we propose a kernel embedded estimator for the treatment confounding bridge function (q-bridge function) based on a dual kernel embedding method, under the assumption that there exist at least one bounded continuous q-bridge function for each treatment. In addition, we show the consistency of the q-bridge function estimator and give a consistent ATE estimator based on the proximal inverse probability weighted estimator.

Key words: Estimator; q-bridge function; ATE; Reproducing kernel Hilbert space; Fenchel duality; interchangeability.

Acknowledgement

First and foremost, I wish to express my deepest gratitude to my supervisor, Professor Aad van der Vaart. His course Advanced Topics in Statistics first opened the door for me into the field of causal inference, and he offered many novel and inspiring directions that helped me identify the subject I truly wished to pursue for my graduation project. Over the countless discussions during the eight months, he patiently answered my questions, pointed out the details I had overlooked, even though as simple and repetitive as they are. I couldn't have finished the writing without his guidance. Furthermore, he constantly reminded me to maintain a spirit of critical thinking, and his heuristic style of guidance greatly nurtured my ability to learn independently. It has been both an honor and a privilege to be under his mentorship.

I am also grateful to all the colleagues I have met over my two years of study. Their dedication and tireless pursuit of knowledge have been a constant source of inspiration for me. In particular, I would like to thank Yuran Wang and Hongrui Liu in Delft, as well as Fransy Fan and Ryan Chen in Rotterdam. I shared many joyful moments with you, and those memories will always be precious to me. It was your warmth and thoughtfulness that carried me through every moment of solitude and discouragement.

Lastly, I wish to thank my parents and families. It is their unwavering support that gave me the opportunity to travel to the other side of the world, to acquire knowledge, and to experience a different life.

Wei Liu
Delft, August 2025

Contents

1 Introduction

Causal inference has roots in ancient philosophy, where thinkers like Aristotle defined different kinds of causes and David Hume questioned how we can justify causal claims beyond mere observation. The statistical era began in the early 20th century, when Ronald A. Fisher formalized randomized controlled trials and Sewall Wright developed path analysis to represent causal relationships mathematically. Mid-century, Jerzy Neyman and Donald Rubin introduced the potential outcomes framework, giving causality a precise probabilistic foundation. In the late 20th century, Judea Pearl's causal diagrams and do-calculus revolutionized the field by unifying graphical models with statistical inference, enabling rigorous causal analysis even from observational data. Today, causal inference is central to fields from epidemiology to artificial intelligence, blending experimental design, econometrics, and machine learning to answer questions about interventions, policies, and complex systems.

As the beginning of the thesis, we introduce the basic counterfactual model of the causal inference, and then its evolution in problems with unmeasured confounders.

1.1. From basic model to proximal model

In the basic counterfactual causal setting, the binary treatment and the outcome are affected by a common confounder. To better illustrate the details, we consider a scenario in which some patients with serious heart diseases made decisions on whether taking heart transplantation surgeries or not (adapted from [17]). After the surgeries, deaths and survivals happened.

In this story, we denote A to be the binary treatment variable taking 1 if the transplantation is received and 0 if not. Y is the binary outcome variable representing a death if taking 1 and a survival if taking 0. The physical condition L of the patients is crucial to the determination of accepting the transplantation as well as the result of the treatment.

The directed acyclic graph for the model representing the relationship of all the three variables is given in Figure 1.1.

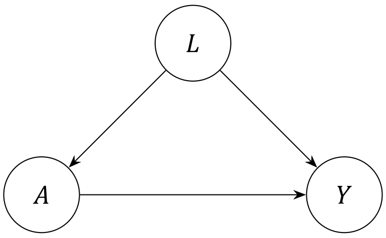


Figure 1.1: A directed acyclic graph (DAG) for basic counterfactual model

Notice that the patient can't be both treated and not treated at the same time, so the outcome variable Y only reflects one possible result based on the decision made by the patient. However, this is not enough to determine the effect of the transplantation surgery. Instead we introduce a set of counterfactual outcome variables $\{Y^a : a \in \{0,1\}\}$. The counterfactual outcome variables follow their own data generating schemes (or to say densities in continuous cases) but which are never observable. The outcome variable Y takes one of the values Y^0 or Y^1 after the patient determined whether to take the treatment. The observable outcome Y can be written as $\mathbb{1}_{A=1}Y^1$ and $\mathbb{1}_{A=0}Y^0$ depending on the treatment for each individual. This is the consistency assumption of the counterfactual model, which states mathematically that

Assumption 1.1 ***(Consistency)*** *On the event $A = a$, we have $Y = Y^a$.*

Now a group of patients are chosen if their physical conditions L are the same. The death rate (or survival rate) of the patients who haven't (have) taken the transplantation surgery would have been the same as the patients who have (haven't) taken the treatment, if they had (hadn't) taken the surgery. Mathematically, it can be represented by

$$Pr\left\{Y^0 = 1|A = 0, L\right\} = Pr\left\{Y^0 = 1|A = 1, L\right\} = Pr\left\{Y^0 = 1|L\right\}$$
$$Pr\left\{Y^1 = 1|A = 0, L\right\} = Pr\left\{Y^1 = 1|A = 1, L\right\} = Pr\left\{Y^1 = 1|L\right\}.$$

The statement is equivalent to the conditional exchangeability assumption:

Assumption 1.2 ***(Conditional exchangeability)*** *$Y^a \perp A|L, \forall a \in \{0,1\}$.*

To emphasize the significance of the conditions of the patients, we assume there must be treated and not treated individuals for any value of this confounder. Mathematically, it is the positivity assumption of the propensity score:

Assumption 1.3 ***(Positivity)*** *$0 < Pr\{A = a|L\} < 1, \forall a \in \{0,1\}$.*

To find the effect of the transplantation surgery on the patients with heart diseases, we need to determine the average outcome EY^a to find the average treatment effect $EY^1 - EY^0$. In fact, we can find it through the standardization formula.

Theorem 1.1 ***(Standardization formula)***
Under the consistency, conditional exchangeability and positivity assumptions, the average outcome under treatment a is identified as

$$EY^a = E_L E(Y|L, A = a).$$

Proof:
By the tower property of conditional expectation, $EY^a = E_L E(Y^a|L)$. By the consistency and conditional exchangeability assumptions, we have $Y|L, A = a \sim Y^a|L$. So, $E_L E(Y^a|L) = E_L E(Y|L, A = a)$. This leads to $EY^a = E_L E(Y|L, A = a)$.

□

However, the conditional exchangeability does not hold all the time. For example when there is another factor called environmental impact that is strongly influential to the condition of patients (L), the decision on taking the treatment (A) and the outcome (Y). This may cover the level of hygiene, the condition of heart donators, the quality of medical devices and so on. To determine the average treatment effect, by the standardization formula (1.1), the average outcome is given by

$$EY^a = E_{U,L} E[Y|U, L, A = a]. \tag{1.1}$$

The standardization formula (1.1) works only when the confounder U is observable. However the environmental compact is difficult to measure because it consists of various factors that are hard to be represented mathematically. An alternative method to compute the average outcome in this case is to identify by some measurable proxies of the unmeasurable confounder.

We can partition the confounder L into three factors (Z, X, W), where Z and W will only affect A and Y respectively, while X will influence all factors except U. We call Z a treatment-inducing confounding

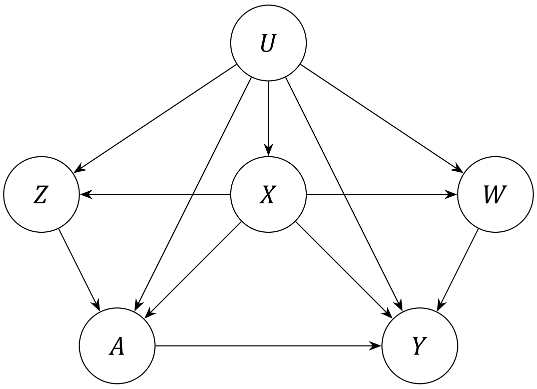


Figure 1.2: A DAG for proximal counterfactual model

proxy and W an outcome-inducing confounding proxy. In other cases, both Z and W can be a group of confounders with the same function [45]. The corresponding directed acyclic graph is shown in Figure 1.2.

Faithful to the random variables (U, Z, X, W, A, Y), the DAG 1.2 implies two conditional independencies for Y and W: $Y \perp Z|U, X, A$ and $W \perp (Z, A)|U, X$. Adaptions of assumptions 1.2 and 1.3 are made to fit the proximal counterfactual model. This gives the assumptions for proximal counterfactual models:

Assumption 1.4 ***(Proximal counterfactual model)***

- *Consistency:* $Y = Y^A$,
- *Conditional independence for* Y*:* $Y \perp Z|U, X, A,$
- *Conditional independence for* W*:* $W \perp (Z, A)|U, X,$
- *Conditional exchangeability:* $Y^a \perp A|U, X, \forall a \in \{0, 1\}$
- *Positivity:* $0 < Pr(A = a|U, X) < 1, \forall a \in \{0, 1\}$.

In this story, the unobservable confounder break the assumption of conditional exchangeability 1.2 and thus makes the classical counterfactual model fail. If people still apply the classical counterfactual model to identify the average treatment effect regardless of the influence from the environmental impact, the average treatment effect calculated from the observed data won't be consistent with the true one. Hence, the use of proxy variables brings possibilities to find methods of standardization through variables that inherit the information from unobservable confounder and directly affect the treatment and outcome. This helps reduce the bias brought up by the confounders.

There are many methods related to proxies (negative control methods) that are proposed to deal with the unobservable confounding. Flander et al.[14] used a so-called indicator variables to detect the unmeasurable confounder under a series of assumptions including linearity standardization formula and conditional expectation of confounder under proxies. Moreover, based on negative control outcomes, Tchetgen [40] proposed a control outcome calibration approach to correct causal effect estimates for bias due to unobserved confounding, while Sofer [34] showed the negative outcome control approach is equivalent to the difference-in-differences approach under certain circumstances. For more negative control methods, Shi et al.[31] made a review of negative control methods in epidemiology, including methods in bias detection, bias reduction and bias correction. Although these approaches are effective under certain assumptions, they are also restricted by the assumptions which narrow down the applicable situations.

One of the most prominent methods based on proxy variables is given by Miao et al.[24]. The author generalized the method of identifying the unknown data generation mechanism of Kuroki et al.[21]. In this way, they put forward a standardization method through the outcome confounding bridge function about proxies such that with at least two independent proxy variables satisfying certain completeness assumptions, the causal effect could be nonparametrically identified without any prior knowledge on the distribution relevant to the unobservable confounder. Therefore, they changed the point of the problem from the uncertainty in the unknown mechanism of confounders to solving the outcome confounding bridge function from integral equations, which provided more connections to operator theory and inverse problems.

After the work of Miao et al., Tchetgen et al.[39] introduced a formal framework for proximal causal inference, where they systematically gave the proximal assumptions and completeness assumption for nonparametric identification. They also described the algorithm for estimating proximal g-formula under a parametric model for outcome confounding bridge function. This spurs the later work on the proximal causal inference for example identifying causality from treatment confounding proxy and finding the g-formula through treatment confounding bridge function by Cui et al.[45].

1.2. Proximal standardization formulae

In this section, we will introduce the proximal standardization formulae through outcome confounding bridge function by Miao et al.[24] and treatment confounding bridge function by Cui et.al [45]. The proximal standardization formulae are based on the existence of bridge functions. The bridge functions connect the average outcome and the proxies without the requirement of the information from the unmeasurable confounders. To build up the proximal standardization formulae, we need the following completeness assumptions 1.5 and 1.6, which are crucial for using the bridge functions to determine the average outcome under certain treatment $E[Y^a]$.

Assumption 1.5 ***(Completeness)***

For any square-integrable function g and for any a, x, $E[g(U)|Z, A = a, X = x] = 0$ almost surely if and only if $g(U) = 0$ almost surely.

Assumption 1.6 ***(Completeness)***

For any square-integrable function g and for any a, x, $E[g(U)|W, A = a, X = x] = 0$ almost surely if and only if $g(U) = 0$ almost surely.

The square-integrable function g must not depend on Z or W if applied to Assumption 1.5 or 1.6.

Example 1.1 ***(Counter example)***

Suppose square-integrable function $g_Z(U) = U - Z$ and $U|(Z, A = a, X = x) \sim \mathcal{N}(Z, 1)$, $\forall a, x$, then completeness assumption 1.5 fails for g_Z. In fact, $E[g_Z(U)|Z, A = a, X = x] = Z - Z = 0$ but $g_Z(U)$ isn't almost surely a zero function.

We use the following Gaussian model to give a straightforward example of distribution families which satisfy the two completeness assumptions.

Example 1.2 ***(Gaussian distribution)***

$U \sim \mathcal{N}(\mu_U, \Sigma_U) \in \mathbb{R}^{d_1}$	$f_A(1|U, Z, X) = \frac{1}{1+\exp\{-\|\mu_A+\gamma_{A|Z}Z+\gamma_{A|X}X+\gamma_{A|U}U\|_2\}}$
$X|U \sim \mathcal{N}(\mu_X + \gamma_{X|U}U, \Sigma_X) \in \mathbb{R}^{d_2}$	$Y^1|U, X, W \sim \mathcal{N}(\mu_1 + \gamma_{1|X}X + \gamma_{1|W}W + \gamma_{1|U}U, \Sigma_1) \in \mathbb{R}^{d_5}$
$Z|U, X \sim \mathcal{N}(\mu_Z + \gamma_{Z|U}U + \gamma_{Z|X}X, \Sigma_Z) \in \mathbb{R}^{d_3}$	$Y^0|U, X, W \sim \mathcal{N}(\mu_0 + \gamma_{0|X}X + \gamma_{0|W}W + \gamma_{0|U}U, \Sigma_0) \in \mathbb{R}^{d_5}$
$W|U, X \sim \mathcal{N}(\mu_W + \gamma_{W|U}U + \gamma_{W|X}X, \Sigma_W) \in \mathbb{R}^{d_4}$	

Table 1.1: Gaussian model in a proximal setting

Consider the two conditional distribution $p(U|Z, A = a, X)$ (A.1) and $p(U|W, A = a, X)$ (A.3) under the Gaussian model in Table 1.1. The distribution families of $T_1(U)$ and $T_2(U)$, which are given by

$$T_1(U) = (\|U\|^2_{\Sigma_U^{-1}}, \|\gamma_{Z|U}U\|^2_{\Sigma_Z^{-1}}, \|\gamma_{Z|U}U\|^2_{\Sigma_Z^{-1}}, \mu(Z, X)^T U, \log f_A(a|U, Z, X))$$

$$T_2(U) = (\|U\|^2_{\Sigma_U^{-1}}, \|\gamma_{X|U}U\|^2_{\Sigma_X^{-1}}, \|\gamma_{W|U}U\|^2_{\Sigma_W^{-1}}, \|\gamma_{Z|U}U\|^2_{\Sigma_Z^{-1}}, \mu(W, X)^T U, \log T(U)),$$

are all complete.

Proof:

The statement is a direct result of Theorem 2.1.

□

From a view in functional analysis, for example in Assumption 1.5, the completeness assumption implies that the conditional expectation operator $E : L_2(P_{U|A=a,X=x}) \mapsto L_2(P_{Z|A=a,X=x})$ is injective, since its null space is just $\{0\}$. By Theorem 2.3, the range of $E^* : L_2(P_{Z|A=a,X=x}) \mapsto L_2(P_{U|A=a,X=x})$ is dense in $L_2(P_{U|A=a,X=x})$. Usually, the completeness can be interpreted as the proxies capturing the variability of the unmeasured confounder.

In the binary treatment cases, if the confounders are all categorical, the completeness assumption is equivalent to assuming the category of the proxies are at least as numerous as the unmeasured confounder to make sure the probability matrix is invertible and the redundant categories of the proxies can be incorporated by some coarsening methods [24]. Suppose $P\,(U|Z, A = a, X = x)$ represents the probability matrix with entries $Pr\left\{u_i|z_j, A = a, X = x\right\}$, for $i, j \in \{1, \cdots, n\}$. And $g(U)$ is the row vector with elements $g(u_i)$, for $i \in \{1, \cdots, n\}$. The left hand side of the assumption 1.5 is equal to $\sum_{i=1}^{n} g(u_i) Pr\left\{u_i|z_j, A = a, X = x\right\} = 0$, a.s. for $j \in \{1, \cdots, n\}$. By the invertibility of the probability matrix, the null space is just $\{0\}$, which means $g(U) = 0$ almost surely.

If the confounders are continuous, in parametric and semiparametric background, the completeness assumption usually requires certain density families to realize. In fact, many commonly used parametric and semiparametric models such as exponential families [27][22] and location-scale families [18] meet the demand. For nonparametric regression models, the results of [11] and [10] based on instrumental variable estimation can be used to justify the completeness conditions.

With the help of the completeness assumptions, we can build the identification formulae of the average treatment effect without any information from the unmeasurable confounders. Two standardization formulae identifying the average treatment effect through bridge functions are elaborated in Theorem 1.2 and 1.3. The existence of outcome-inducing and treatment-inducing bridge functions are introduced in Lemma 1.1 and 1.2, which are applications of Theorem 2.9.

Lemma 1.1 ***(Existence of outcome confounding bridge function)***

Denote the singular system[1] *of the conditional expectation operator $T : L_2(P_{(W,A,X)}) \mapsto L_2(P_{(Z,A,X)})$ by $(\sigma_i, u_i, v_i)_{i\geq 1}$. There exists a solution h for the integral equation:*

$$E[Y|Z, A = a, X] = \int h(w, A, X) dF(w|Z, A = a, X) = E[h(W, a, X)|Z, A = a, X], \tag{1.2}$$

if

$$\sum_{i\geq 1} \frac{|\langle E[Y|Z, A = a, X], v_i\rangle|^2}{\sigma_i^2} < \infty. \tag{1.3}$$

Theorem 1.2 *([24]* ***Outcome confounding standardization formula****)*

Under proximal assumptions for counterfactual models and completeness assumption 1.5, if the solution of equation (1.2) exists, the proximal standardization formula is given by:

$$EY^a = E[h(W, a, X)]. \tag{1.4}$$

The average treatment effect is given by:

$$\chi = E[h(W, 1, X) - h(W, 0, X)].$$

Proof:

By the existence of $h(W, a, X)$, we have

$$E[Y|Z, A = a, X] = E[h(W, A, X)|Z, A = a, X].$$

[1] See Theorem 2.8

By the tower property of conditional expectation, the above equation becomes

$$
\begin{aligned}
&E_U[E[Y|Z, A = a, X, U]|Z, A = a, X] = E_U[E[h(W, A, X)|Z, A = a, X, U]|Z, A = a, X] \\
\Rightarrow &E_U[E[Y - h(W, A, X)|Z, A = a, X, U]|Z, A = a, X] = 0 \\
\Rightarrow &E_U[E[Y - h(W, A, X)|A = a, X, U]|Z, A = a, X] = 0. \qquad (W \perp (Z, A)|U, X)
\end{aligned}
$$

Here $E[Y - h(W, A, X)|A = a, X, U]$ is a function dependent only on U for a fixed X. So, by the completeness assumption 1.5, we have

$$
\begin{aligned}
&E[Y - h(W, A, X)|A = a, X, U] = 0, a.s. \\
\Rightarrow &E[Y|A = a, X, U] = E[h(W, A, X)|A = a, X, U], a.s..
\end{aligned}
$$

The left hand side equals $E[Y^a|X, U]$ by consistency and conditional exchangeability. The right hand side equals $E[h(W, a, X)|X, U]$ since $W \perp (Z, A)|U, X$. So, we have

$$
\begin{aligned}
&E[Y^a|X, U] = E[h(W, a, X)|X, U] \\
\Rightarrow\ &E_{X,U}E[Y^a|X, U] = E_{X,U}E[h(W, a, X)|X, U] \\
\Rightarrow\ &EY^a = E[h(W, a, X)].
\end{aligned}
$$

□

Lemma 1.2 ***(Existence of treatment confounding bridge function)***
Denote the singular system of the conditional expectation operator $T^* : L_2(P_{(Z,A,X)}) \mapsto L_2(P_{(W,A,X)})$ by $(\sigma_i^*, u_i^*, v_i^*)_{i\geq 1}$. There exists a solution q for the integral equation:

$$
\frac{1}{f(a|W, X)} = \int q(z, A, X)dF(z|W, A = a, X) = E[q(Z, a, X)|W, A = a, X], \tag{1.5}
$$

where $f(a|W, X) = Pr\{A = a|W, X\}$, if

$$
\sum_{i\geq 1} \frac{|\left\langle \frac{1}{f(a|W,X)}, v_i^* \right\rangle|^2}{(\sigma_i^*)^2} < \infty. \tag{1.6}
$$

Before giving the treatment confounding standardization formula, we introduce a useful result to be used in the upcoming proof.

Lemma 1.3 *Consider random variables X, Y and Z on a measurable metric space S. Let $f(X|Y, Z)$ and $f(X|Z)$ be the positive conditional density of X relative to a σ-finite measure ν. If the base measure for Y on S is the σ-finite measure μ, then the two conditional densities are connected through the following equation*

$$
\frac{1}{f(X|Z)} = E_Y[\frac{1}{f(X|Y, Z)}|X, Z].
$$

Proof:

$$
\begin{aligned}
1 &= E_Y[\frac{1}{f(X|Y, Z)}f(X|Y, Z)|X, Z] \\
&= \int \frac{1}{f(X|y, Z)}f(X|y, Z)f(y|X, Z)d\mu(y) \\
&= \int \frac{1}{f(X|y, Z)}f(y|X, Z)f(X|Z)d\mu(y) \\
&= f(X|Z)E_Y[\frac{1}{f(X|Y, Z)}|X, Z] \\
\Rightarrow \frac{1}{f(X|Z)} &= E_Y[\frac{1}{f(X|Y, Z)}|X, Z].
\end{aligned}
$$

□

Theorem 1.3 *([45] **Treatment confounding standardization formula**)*
Under proximal assumptions for counterfactual models and completeness assumption 1.6, if the solution of equation (1.5) exists, the proximal standardization formula is given by:

$$EY^a = E[Yq(Z,a,X)\mathbb{1}_{A=a}]. \tag{1.7}$$

The average treatment effect is given by:

$$\chi = E[Y(\mathbb{1}_{A=1}q(Z,1,X) - \mathbb{1}_{A=0}q(Z,0,X))] = E[(-1)^{1-A}Yq(Z,A,X)].$$

Proof:
By the existence of $q(Z,A,X)$, we have

$$\frac{1}{f(a|W,X)} = E[q(Z,a,X)|W,A=a,X]. \tag{1.8}$$

Notice that the left hand side is equivalent to $E[\frac{1}{f(a|U,X)}|W,A=a,X]$ because

$$\begin{aligned} \frac{1}{f(a|W,X)} &= E_U[\frac{1}{f(a|U,W,X)}|W,A=a,X], && \text{(Lemma 1.3)} \\ E_U[\frac{1}{f(a|U,X)}|W,A=a,X] &= E_U[\frac{1}{f(a|U,W,X)}|W,A=a,X]. && (W \perp (Z,A)|U,X) \end{aligned}$$

Hence, Equation (1.8) is equivalent to

$$\begin{aligned} E_U[\frac{1}{f(a|U,X)}|W,A=a,X] &= E[q(Z,a,X)|W,A=a,X] \\ &= E_U[E[q(Z,a,X)|W,A=a,X,U]|W,A=a,X] \\ &= E_U[E[q(Z,a,X)|A=a,X,U]|W,A=a,X]. && (W \perp (Z,A)|U,X) \end{aligned}$$

Here $E[q(Z,a,X)|A=a,X,U]$ is a function dependent only on U for a fixed X. So, by the completeness assumption 1.6, we get

$$\frac{1}{f(a|U,X)} = E[q(Z,a,X)|A=a,X,U].$$

By consistency and conditional exchangeability,

$$\begin{aligned} EY^a =& E_{U,X}E[Y|A=a,U,X] \\ =& E_{U,X}\{E[Y|A=a,U,X]E[q(Z,a,X)|A=a,U,X]f(a|U,X)\} \\ =& E_{U,X}\{E[Yq(Z,a,X)|A=a,U,X]f(a|U,X)\} && (Y \perp Z|U,X,A) \\ =& E_{U,X}\{E[Yq(Z,a,X)\mathbb{1}_{A=a}|U,X]\} && (E[\cdot|U,X] = E_A[E[\cdot|A,U,X]|U,X]) \\ =& E[Yq(Z,a,X)\mathbb{1}_{A=a}]. \end{aligned}$$

The average treatment effect χ is given by $E[Y(\mathbb{1}_{A=1}q(Z,1,X) - \mathbb{1}_{A=0}q(Z,0,X))] = E[(-1)^{1-A}Yq(Z,A,X)]$.
□

The proximal standardization formulae (1.4) and (1.7) provide feasible methods to acquire the average treatment effect regardless the unobservable confounders. In fact, Cui et al.[45] proposed three ATE estimators based on the estimators of the two bridge functions $\widehat{h}$ and $\widehat{q}$, named as proximal outcome regression estimator, proximal inverse probability weighted estimator and proximal double robust estimator. They are given as follows.

$$\hat{\chi}_{POR} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{h}(W_i,1,X_i) - \hat{h}(W_i,0,X_i)\right)$$

$$\hat{\chi}_{PIPW} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i\mathbb{1}_{A_i=1}\hat{q}(Z_i,1,X_i) - Y_i\mathbb{1}_{A_i=0}\hat{q}(Z_i,0,X_i)\right) \tag{1.9}$$

$$\hat{\chi}_{PDR} = \frac{1}{n}\sum_{i=1}^{n}\left((-1)^{1-A_i}\hat{q}(Z_i,A_i,X_i)[Y_i - \hat{h}(W_i,A_i,X_i)] + \hat{h}(W_i,1,X_i) - \hat{h}(W_i,0,X_i)\right). \tag{1.10}$$

Among the three estimators, the double robust estimator (1.10) is able to tolerate situations in which the existence assumptions of bridge functions fail. This means even if consistent estimator for h or q doesn't exist at the same time, only one consistent estimator for any of the bridge functions will preserve the consistency of the ATE estimator.

The estimator for ATE depends on at least one of the consistency bridge function estimators. Ghassami et al.[16] and Kallus et al.[19] used minimax optimization of the two bridge functions restricted to reproducing kernel Hilbert spaces to get the estimators. Ghassami et al. constructed the minimax optimization through the double robustness of the influence function that the expectation of the perturbation of influence function at point (q_{true}, h) or (q, h_{true}) toward (q_{true}, h') or (q', h_{true}) should be zero. While the motivation for Kallus et al. is based on the following equalities

$$E[f(X)|Y] = 0 \Leftrightarrow E[g(Y)f(X)] = 0 \Leftrightarrow \sup_{g\in L_2(P_Y)} (E[g(Y)f(X)])^2,\ \forall g \in L_2(P_Y). \tag{1.11}$$

They also gave the estimators for the two bridge functions after restricting them into reproducing kernel Hilbert spaces. Although their methods are efficient in determining bridge functions in nonparametric models, they also need the assumptions for the existence of the two bridge functions to hold at the same time. However, this is quite challenging in real cases because the Picard's conditions are hard to verify.

There are many other results for estimating the ATE through bridge function estimators. Mastouri et al.[23] used a two-kernel-based approach to estimate outcome confounding bridge function, where in the first stage the conditional covariance operator is learnt from the first group of samples, and in the second stage the estimator for outcome confounding bridge function is derived from a ridge regression problem. Similar to the second stage of the method by Mastouri et al., Singh [32] proposed a family of algorithms based on kernel ridge regression for learning nonparametric treatment effects with negative controls. In addition, Kompa et al.[20] combined (1.11) and reproducing kernels to use neural network to find the estimated outcome confounding bridge function under a loss function transformed from (1.11). Furthermore, beyond the binary treatment, Wu et al.[44] derived a double robust estimator under continuous treatment with a kernel function $K_{h_{bw}}(A-a) := \frac{1}{h_{bw}}K(\frac{A-a}{h_{bw}})$ approximating the indicator function $\mathbb{1}_{A=a}$ as $h_{bw} \longrightarrow 0$. This helps identifying the treatment effect when A is continuous.

Although estimating the ATE only depends on one of the estimated bridge functions and many literatures have already find the right ways to estimating the outcome confounding bridge function, it is still intriguing to find the estimator of the treatment confounding bridge function. Hence, in Chapter 4, we will discuss the approach to estimating the treatment confounding bridge function q.

2 Mathematical preliminary

This chapter aims at introducing all the mathematical preliminaries that are needed in other chapters. Section 2.1 introduces the definition of complete distribution family and gives a theorem about the completeness of exponential family, which is useful for the explanation of the two completeness assumptions determining the standardization formulae. Section 2.3 introduces the Picard's condition for the existence of the solutions to the Fredholm integral of the first kind. For example the existence of the two bridge functions by Lemma 1.1 and 1.2. Section 2.2 focuses on the basic knowledge in functional analysis on Hilbert spaces. The spectral theorems about compact operators are the most important points in this section because they will be widely used through out the whole thesis. Section 2.4 is one of the core parts of the Chapter 2, it includes the Moore-Aronszajn theorem 2.14 which is the key to understanding the structure of reproducing kernel Hilbert spaces. The kernel embeddings are also crucial since they are the basis for the kernel method used for estimating the treatment confounding bridge function. Section 2.7 systematically introduces the results in convex analysis from semi-continuity to the interchange of minimization and integration. These helps explaining the series of transforms given in Chapter 4 combining the ERM theorem 2.18 given by Section 2.6. The rest two Sections 2.5 and 2.8 are designed to illustrate the existence of Fisher information mentioned in Chapter 3 and the derivative of inner product when deriving the closed form expression of extremal value points in Chapter 4.

2.1. Complete distribution family

This section is a brief introduction to complete probability distribution families based on the section 4.3 of [22], which helps explaining the completeness assumptions 1.5 and 1.6.

Definition 2.1 ***(Complete distribution family)***
A probability distribution family $\mathcal{P}$ is complete if $\forall p \in \mathcal{P}$, any measurable function g satisfying $E_P[g(X)] = 0$ implies $P\{g(X) = 0\} = 1$.

An classical example of the complete distribution families is the exponential family, which is given by the following theorem.

Theorem 2.1 ***(Completeness of exponential family)***
Suppose X is a random vector with density

$$p_\theta(x) = C(\theta)\exp\left\{\sum_{i=1}^{n} \theta_i T_i(x)\right\}.$$

Let $\mathcal{P}_T$ be the distribution family of $T = (T_1, \cdots, T_n)$. if $\theta = (\theta_1, \cdots, \theta_n)$ contains a n-dim rectangle, then $\mathcal{P}_T$ is complete.

An application of the completeness of exponential family can be found in Example 1.2, in which the theorem is used to verify the completeness assumptions 1.5 and 1.6 of a Gaussian proximal model.

2.2. Functional analysis on Hilbert spaces

In this section, we set $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, H to be a Hilbert space of functions $f : S \mapsto \mathbb{K}$ and $\mathcal{L}(H)$ to be the set of all bounded linear operators from H to itself. We choose S to be a compact subset of $\mathbb{K}^d$ and denote the linear space of all square integrable functions by $L_2(S)$. Without additional statements, the measure used in the L_2 space is the Lebesgue measure. The following basic definitions and theorems about Hilbert space and operators in Subsection 2.2.1, 2.2.2 and 2.2.3 are introduced based on [26] and [6].

Definition 2.2 ***(Hilbert space)***
A Hilbert space is a linear space equipped with an inner product $\langle\cdot,\cdot\rangle$ that is complete relative to the induced norm $\|\cdot\|$.

Classical Hilbert spaces include $L_2(S, \mu)$ for any measure μ.

2.2.1. Basic definitions and theorems

In this part, we mainly introduce the definitions of adjoint and compact operators through dual spaces and Riesz representation theorem.

Definition 2.3 ***(Linear operator)***
An operator $T : H \mapsto H$ is said to be linear if

$$T(\alpha f + \beta g) = \alpha T(f) + \beta T(g),\ \forall f, g \in H,\ \alpha, \beta \in \mathbb{K}.$$

Definition 2.4 ***(Bounded operator)***
An linear operator $T : H \mapsto H$ is said to be bounded if there exists a positive $C \in \mathbb{K}$ such that

$$\|T(h)\|_H \leq C\|h\|_H,\ \forall h \in H.$$

The norm $\|T\|_{\mathcal{L}(H)}$ of T on $\mathcal{L}(H)$ is defined to be the smallest constant C.

The definition gives the inequality $\|T(h)\|_H \leq \|T\|_{\mathcal{L}(H)}\|h\|_H$.
Furthermore, the definition also implies that for linear operators boundedness is equivalent to continuity.

Example 2.1 ***(Integral operator on $L_2(S)$)***
Let $k(\cdot,\cdot) \in L_2(S^2) : S{\times}S \mapsto \mathbb{K}$. The integral operator $T : L_2(S) \mapsto L_2(S)$, with $T(f)(x) = \int_S k(x,y)f(y)dy$, $\forall f \in L_2(S)$, $\forall x \in S$, is bounded by $\|k\|_2$.

Proof:

$$\begin{aligned}\|T(f)\|_2^2 &= \int_S \left(\int_S k(x,y)f(y)dy\right)^2 dx \\ &\leq \int_S \left(\int_S |k(x,y)|^2 dy\right)\left(\int_S |f(y)|^2 dy\right) dx \qquad \text{(Cauchy-Schwartz inequality B.1)} \\ &= \left(\int_S \int_S |k(x,y)|^2 dy dx\right)\left(\int_S |f(y)|^2 dy\right) \\ &= \|k\|_2^2 \|f\|_2^2.\end{aligned}$$

Hence, we get $\|T(f)\|_2 \leq \|k\|_2\|f\|_2$. For f with nonzero L_2-norm, we have $\frac{\|T(f)\|_2}{\|f\|_2} \leq \|k\|_2$. When taking the supremum over all f with $\|f\|_2 \leq 1$, we finally reach the upper bound of $\|T\|$:

$$\|T\| = \sup_{\|f\|_2 \leq 1} \frac{\|T(f)\|_2}{\|f\|_2} \leq \|k\|_2.$$

□

Theorem 2.2 ***(Riesz representation theorem)***
If $\psi : H \mapsto \mathbb{K}$ is a bounded linear functional, there exists a unique element $\tilde{\psi} \in H$ such that

$$\psi(g) = \langle g, \tilde{\psi} \rangle, \ \forall g \in H.$$

Since every closed subspace of H is also a Hilbert space, the Riesz representation theorem can be applied to any closed subspace of H with the same bounded linear functional if the functional is well-defined everywhere on H. The Riesz representation theorem is not only vital in unveiling the isometrically isomorphic essence in Hilbert space and its dual space, but also quite useful in computationally finding the uniquely existing representative elements in a Hilbert space given a bounded functional. The example of the later statement can be found in finding the efficient influence function of an estimator which is shown in Chapter 3.

Definition 2.5 ***(Dual space)***
The dual space of Hilbert space H is the Hilbert space $H^* := \mathcal{L}(H, \mathbb{K})$.

Since the dual space is always complete, H^* is also a Hilbert space. The Riesz representation theorem implies there exists a bijective map between H and H^*, thus H^* can be canonically identified with H.

Definition 2.6 ***(Adjoint operator)***
Let H_1 and H_2 be two Hilbert spaces. Suppose $T \in \mathcal{L}(H_1, H_2)$, then its adjoint operator T^* is the linear operator belonging to $\mathcal{L}(H_2, H_1)$ and satisfying

$$\langle Th_1, h_2 \rangle = \langle h_1, T^* h_2 \rangle, \ \forall h_1 \in H_1, \ h_2 \in H_2.$$

If $T \in \mathcal{L}(H)$, and $T = T^*$, then T is self-adjoint.

After introducing the adjoint operators on Hilbert spaces, we state a useful decomposition result based on the range and the kernel space of bounded linear operators and their adjoints. The theorem can be used to explain the completeness assumptions 1.5 and 1.6.

Theorem 2.3 ***(Orthogonal decomposition)***
If T is a bounded linear operator in $\mathcal{L}(H_1, H_2)$, then H_1 and H_2 have orthogonal decompositions

$$H_1 = \mathit{Null}(T) \oplus \overline{\mathit{Range}(T^*)}, \ H_2 = \mathit{Null}(T^*) \oplus \overline{\mathit{Range}(T)}.$$

In particular,

- *T (or T^*) is injective if and only if T^* (or T) has dense range;*
- *T (or T^*) is surjective if and only if T^* (or T) is injective and has closed range.*

Definition 2.7 ***(Compact operator)***
A bounded operator T is compact if it maps any bounded sets to relatively compact sets, i.e. sets with compact closures.

Example 2.2 ***(Finite rank operators)***
A bounded operator is said to be finite rank if its image belongs to a finite-dimensional space. We claim that any finite rank operator is compact.

Proof:
Any bounded set in a finite-dimensional space is relatively compact. Since finite rank operators are bounded, they map any bounded set to bounded set. This means finite rank operators are compact.

□

The finite rank operator can be used to approximate a compact linear operator, as shown in the following theorem.

Theorem 2.4 ***(Finite rank operator approximation)***

An linear operator is compact if and only if it is the uniform limit of some finite rank operators.

Example 2.3 ***(Hilbert-Schmidt operator)***

Every Hilbert-Schmidt operator (Definition 2.8) is compact and can be approximated by finite rank operators in Hilbert-Schmidt norm. One can check Proposition 14.5 in [26] for details.

Example 2.4 ***(Hilbert-Schmidt integral operator)***

The Hilbert-Schmidt integral operator $T_k : L_2(S,\mu) \mapsto L_2(S,\mu)$ is given by

$$T_k(f)(x) = \int_S k(x,y)f(y)d\mu(y),\ \forall f \in L_2(S,\mu),$$

where μ is a Borel finite measure on S. The function $k(\cdot,\cdot) \in L_2(S^2, \mu \otimes \mu) : S \times S \mapsto \mathbb{K}$ in the integral is called Hilbert-Schmidt kernel function. Any Hilbert-Schmidt integral operator is bounded and compact. If the Hilbert-Schmidt kernel function is symmetric, the operator is self-adjoint.

Proof:

Since the boundedness has been proved in Example 2.1 and that the Hilbert-Schmidt integral operator is self-adjoint is a direct result of the symmetry of the Hilbert-Schmidt kernel function, we only prove that Hilbert-Schmidt integral operator is compact.

Fix $\epsilon > 0$. Since for a compact S and a finite Borel measure μ, the linear space of all continuous functions on S^2: $\mathcal{C}(S^2)$ is dense in $L_2(S^2)$, we can find a function $\tilde{k} \in \mathcal{C}(S^2)$ such that $\|\tilde{k}-k\|_2 < \epsilon$. By the uniform continuity of $\tilde{k}$, there exists a $\delta > 0$ such that $|\tilde{k}(x,y) - \tilde{k}(x',y')| < \epsilon$, where $|x - x'| + |y - y'| < \delta$. For a large $n > 2$, the compactness of S implies there exists a finite open cover $\{\theta_i : 1 \le i \le n\}$ of S with diameter at most $\frac{1}{2}\delta$. Define $B_1 = \theta_1$, $B_2 = \theta_2 - \theta_1, \cdots, B_i = \theta_i - \sum_{j=1}^{i-1} \theta_j, \cdots$, so that $B_i \cap B_k = \emptyset$, for $i \neq k$. We know B_i, for $1 \le i \le n$, is with diameter at most $\frac{1}{2}\delta$ such that $S = \bigcup_{i=1}^{n} B_i$. Next set

$$k_n(x,y) = \sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{k}(x_i,y_j)\mathbb{1}_{B_i}(x)\mathbb{1}_{B_j}(y),\ x_i \in B_i,\ y_j \in B_j.$$

Then k_n is a kernel function since the symmetry and positive semi-definiteness is guaranteed by the kernel function $\tilde{k}$. We denote the corresponding integral operator by T_n which is given by, for any $f \in L_2(S)$, $x \in S$,

$$\begin{aligned} T_n(f)(x) &= \int_S k_n(x,y)f(y)d\mu(y) \\ &= \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\int_{B_j} \tilde{k}(x_i,y_j)f(y)d\mu(y)\right)\mathbb{1}_{B_i}(x). \end{aligned}$$

The range of T_n is contained in the linear span of $\{\mathbb{1}_{B_1}, \cdots, \mathbb{1}_{B_n}\}$ and hence T_n is finite rank. Now apply the result in Example 2.1, we have

$$\begin{aligned} \|T_k - T_n\| &\le \|k - k_n\|_2 \\ &\le \|k - \tilde{k}\|_2 + \|\tilde{k} - k_n\|_2 \\ &< \epsilon + \left(\int_S\int_S \left(\tilde{k}(x,y) - \sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{k}(x_i,y_j)\mathbb{1}_{B_i}(x)\mathbb{1}_{B_j}(y)\right)^2 dxdy\right)^{\frac{1}{2}} \\ &\le \epsilon + \left(\sum_{i=1}^{n}\sum_{j=1}^{n} \sup_{(x,y)\in B_i\times B_j} |\tilde{k}(x,y) - \tilde{k}(x_i,y_j)|^2 \mu(B_i)\mu(B_j)\right)^{\frac{1}{2}} \\ &< \epsilon(1 + \mu(S)) \end{aligned}$$

This shows that T_k is the uniform limit of some finite rank operators since ϵ is arbitrary. So, the Hilbert-Schmidt integral operator is compact by Theorem 2.4.

□

An application of Example 2.4 is to verify that the conditional expectation operator is compact.

We use $L_2(P_.)$ to represent the Hilbert space consisting of square integrable functions with probability measure $P_.$.

Example 2.5 ***(Compactness of conditional expectation operator)***

Suppose X and Y are two random variables on a measurable space $(\Omega, \mathcal{A})$ with σ-finite measures μ and ν. Their probability distributions are P_X and P_Y respectively. By Radon-Nikodym theorem, their density functions f_X and f_Y exist as $\frac{dP_X}{d\mu}$ and $\frac{dP_Y}{d\nu}$. We claim that the conditional expectation operator $E : L_2(P_X) \mapsto L_2(P_Y)$ is compact, with $E(g)(y) = \int_\Omega g(x) f_{X|Y}(x|y) d\mu(x)$, $\forall g \in L_2(P_X)$, $\forall y \in \Omega$, if $k(x,y) = \frac{f_{XY}(x,y)}{f_X(x) f_Y(y)} \in L_2(P_X \times P_Y)$, i.e.

$$\int_\Omega \int_\Omega (k(x,y))^2 \, dP_X dP_Y < \infty.$$

Proof:

For any $g \in L_2(P_X)$, $y \in \Omega$,

$$\begin{aligned} E(g)(y) &= \int_\Omega g(x) f_{X|Y}(x|y) d\mu(x) \\ &= \int_\Omega \frac{f_{XY}(x,y)}{f_X(x) f_Y(y)} g(x) f_X(x) d\mu(x) \\ &= \int_\Omega \frac{f_{XY}(x,y)}{f_X(x) f_Y(y)} g(x) dP_X \\ &= \int_\Omega k(x,y) g(x) dP_X. \end{aligned}$$

Since $k(x,y) \in L_2(P_X \times P_Y)$, by Example 2.4, the conditional expectation operator is compact.

□

2.2.2. Hilbert-Schmidt operators

In this subsection we introduce the definition and basic properties of the Hilbert-Schmidt operators. Let H and F be two separable Hilbert spaces with orthonormal basis $(e_i)_{i\geq 1}$ and $(l_i)_{i\geq 1}$ respectively.

Definition 2.8 ***(Hilbert-Schmidt operator)***

A bounded linear operator $T \in \mathcal{L}(H,F)$ is Hilbert-Schimdt if its Hilbert-Schmidt norm is finite, i.e.

$$\|T\|_{HS}^2 = \sum_{i\geq 1} \|Te_i\|_F^2 < \infty.$$

Example 2.6 ***(Hilbert-Schmidt integral operator)***

The Hilbert-Schmidt integral operator $T_k : L_2(S,\mu) \mapsto L_2(S,\mu)$ in Example 2.4 is a Hilbert-Schmidt operator.

Proof:

Suppose $(\xi_i)_{i\geq 1}$ is an orthonormal basis of $L_2(S,\mu)$.

$$\begin{aligned}
\|T_k\|_{HS}^2 = \sum_{i\geq 1} \|T_k\xi_i\|_{L_2(S,\mu)}^2 &= \sum_{i\geq 1} \int_S \Big| \int_S k(x,y)\xi_i(y)d\mu(y)\Big|^2 d\mu(x) \\
&= \int_S \sum_{i\geq 1} \langle k(x,\cdot), \xi_i(\cdot)\rangle_{L_2(S,\mu)}\, d\mu(x) \\
&= \int_S \|k(x,\cdot)\|_{L_2(S,\mu)}^2 d\mu(x) \qquad \text{(Parseval's identity B.2)} \\
&= \|k\|_{L_2(S^2,\mu\otimes\mu)}^2 < \infty.
\end{aligned}$$

□

Example 2.7 ***(Rank-one operator)***
For any $h \in H$, $f \in F$, the tensor product (Definition 2.18) operator $f \otimes h$ is a rank-one operator from H to linear expansion of f, which is a 1-dim subspace of F. The operator is given by

$$(f \otimes h)(g) = \langle g, h\rangle_H\, f,\ \forall g \in H. \tag{2.1}$$

The tensor product operator is a Hilbert-Schmidt operator.

Proof:

$$\begin{aligned}
\|f\otimes h\|_{HS}^2 = \sum_{i\geq 1} \|(f\otimes h)e_i\|_F^2 &= \sum_{i\geq 1} \|\, \langle h, e_i\rangle_H\, f\|_F^2 \\
&= \sum_{i\geq 1} \langle h, e_i\rangle_H^2\, \|f\|_F^2 \\
&= \|h\|_H^2 \|f\|_F^2 < \infty. \qquad \text{(Parseval's identity B.2)}
\end{aligned}$$

□

Moreover, all Hilbert-Schmidt operators in $\mathcal{L}(H,F)$ forms a Hilbert space denoted by $HS(H,F)$, equipped with inner product $\langle\cdot,\cdot\rangle_{HS}$. Suppose $T_1, T_2 \in HS(H,F)$. The inner product is given by

$$\langle T_1, T_2\rangle_{HS} = \sum_{i\geq 1} \langle T_1 e_i, T_2 e_i\rangle_F\,, \tag{2.2}$$

which is well-defined by the symmetry, linearity and positive definiteness of the inner product defined on F. From the definitions, it is clear that the Hilbert-Schmidt norm $\|\cdot\|_{HS}$ is induced by the Hilbert-Schmidt inner product $\langle\cdot,\cdot\rangle_{HS}$. Below we give an equivalent expression of the Hilbert-Schmidt inner product.

Proposition 2.1 ***(Equivalent expression of Hilbert-Schmidt inner product)***
The Hilbert-Schmidt inner product (2.2) is equivalent to

$$\langle T_1, T_2\rangle_{HS} = \sum_{i\geq 1}\sum_{j\geq 1} \langle T_1 e_i, l_j\rangle_F \langle T_2 e_i, l_j\rangle_F\,. \tag{2.3}$$

Proof:
Since $T_1 e_i$ and $T_2 e_i$ are elements in F, $\forall i \geq 1$, we suppose their representations are given by

$$T_1 e_i = \sum_{k\geq 1} \lambda_k^{(i)} l_k,\ T_2 e_i = \sum_{k\geq 1} \gamma_k^{(i)} l_k,$$

where $\lambda_k^{(i)}, \gamma_k^{(i)} \in \mathbb{K}$, $\forall i, k \geq 1$. By (2.2),

$$\begin{aligned}
\langle T_1, T_2 \rangle_{HS} = \sum_{i\geq 1} \langle T_1 e_i, T_2 e_i \rangle_F &= \sum_{i\geq 1} \left\langle \sum_{j\geq 1} \lambda_j^{(i)} l_j, \sum_{k\geq 1} \gamma_k^{(i)} l_k \right\rangle_F \\
&= \sum_{i\geq 1} \sum_{j\geq 1} \lambda_j^{(i)} \left\langle l_j, \sum_{k\geq 1} \gamma_k^{(i)} l_k \right\rangle_F \\
&= \sum_{i\geq 1} \sum_{j\geq 1} \lambda_j^{(i)} \gamma_j^{(i)} \\
&= \sum_{i\geq 1} \sum_{j\geq 1} \langle T_1 e_i, l_j \rangle_F \langle T_2 e_i, l_j \rangle_F .
\end{aligned}$$

□

The equivalent expression of the Hilbert-Schmidt inner product (2.3) is crucial for the following property.

Proposition 2.2 ***(Irrelevant orthonormal basis)***
The definition of the Hilbert-Schmidt operator is irrelevant to the choice of the orthonormal basis.

Proof:
Suppose $(\bar{e}_i)_{i\geq 1}$ and $(\bar{l}_i)_{i\geq 1}$ are also the orthonormal basis of H and F respectively. Denote the adjoint operator of T_1 and T_2 by T_1^* and T_2^*.

$$\begin{aligned}
\langle T_1, T_2 \rangle_{HS} &= \sum_{i\geq 1} \sum_{j\geq 1} \left\langle T_1 e_i, \bar{l}_j \right\rangle_F \left\langle T_2 e_i, \bar{l}_j \right\rangle_F \\
&= \sum_{j\geq 1} \sum_{i\geq 1} \left\langle T_1^* \bar{l}_j, e_i \right\rangle_H \left\langle T_2^* \bar{l}_j, e_i \right\rangle_H && \text{(Adjoint operators 2.6)} \\
&= \sum_{j\geq 1} \left\langle T_1^* \bar{l}_j, T_2^* \bar{l}_j \right\rangle_H && \text{(Proposition 2.1)} \\
&= \sum_{j\geq 1} \sum_{i\geq 1} \left\langle T_1^* \bar{l}_j, \bar{e}_i \right\rangle_H \left\langle T_2^* \bar{l}_j, \bar{e}_i \right\rangle_H \\
&= \sum_{i\geq 1} \sum_{j\geq 1} \left\langle T_1 \bar{e}_i, \bar{l}_j \right\rangle_F \left\langle T_2 \bar{e}_i, \bar{l}_j \right\rangle_F \\
&= \sum_{i\geq 1} \langle T_1 \bar{e}_i, T_2 \bar{e}_i \rangle_F , && (2.4)
\end{aligned}$$

where the last three lines use the repeated trick of the first three lines. Comparing (2.2) and (2.4), we find that the choice of orthonormal basis doesn't influence the definition of Hilbert-Schmidt operator.

□

Proposition 2.3 ***(Hilbert-Schmidt inner product between operators)***
Consider the tensor product operator $f \otimes h$ in Example 2.7. For any $L \in HS(H, F)$, the Hilbert-Schmidt inner product between L and $f \otimes h$ satisfies

$$\langle L, f \otimes h \rangle_{HS} = \langle f, Lh \rangle_F . \tag{2.5}$$

Proof:

Given the orthonormal basis of H $(e_i)_{i\geq 1}$, h has the representation

$$h = \sum_{i\geq 1} \langle h, e_i\rangle_H \, e_i. \tag{2.6}$$

The left hand side of (2.5) becomes

$$\begin{aligned}
\langle L, f \otimes h\rangle_{HS} &= \sum_{i\geq 1} \langle Le_i, (f \otimes h)e_i\rangle_F && \text{(By 2.2)}\\
&= \sum_{i\geq 1} \left\langle Le_i, \langle h, e_i\rangle_H f\right\rangle_F \\
&= \sum_{i\geq 1} \langle h, e_i\rangle_H \langle Le_i, f\rangle_F . && (2.7)
\end{aligned}$$

The right hand side of (2.5) becomes

$$\begin{aligned}
\langle f, Lh\rangle_F &= \left\langle f, L(\sum_{i\geq 1} \langle h, e_i\rangle_H \, e_i)\right\rangle_F \\
&= \sum_{i\geq 1} \langle h, e_i\rangle_H \langle Le_i, f\rangle_F . && (2.8)
\end{aligned}$$

Comparing (2.7) and (2.8), we get the identity (2.5).

□

An crucial application of Proposition 2.3 is to find the equivalent expression of Hilbert-Schmidt inner product between two tensor product operators. Substituting L by another tensor product operator $f' \otimes h'$, where $f' \in F$ and $h' \in H$, we have the following equation

$$\langle f' \otimes h', f \otimes h\rangle_{HS} = \langle f, f'\rangle_F \langle h, h'\rangle_H . \tag{2.9}$$

2.2.3. The spectral theorem for compact operators

In this part, the final aim is to introduce the singular value decomposition theorem for compact operators. To achieve this, we start from the definition of spectrum and explain how the set of eigenvalues of a compact operator composes its spectrum. This leads to the spectral theorem for compact self-adjoint operators showing that a self-adjoint compact operator in a Hilbert space can be represented by the combination of its eigenvalues and the outer products between its eigenfunctions. From this important theorem, we are able to introduce the general representation theorem for compact operators.

Definition 2.9 ***(Spectrum)***

The spectrum $\sigma(T)$ of a linear operator $T \in \mathcal{L}(H)$ is the set of all $\lambda \in \mathbb{K}$ such that $\lambda I - T$ is not boundedly invertible, i.e. there is no bounded linear operator $U \in \mathcal{L}(H)$ such that

$$U(\lambda I - T) = (\lambda I - T)U = I.$$

For self-adjoint operator $T \in \mathcal{L}(H)$, its operator norm is chosen from the maximum value between the absolute of infimum and supremum of its spectrum.

Theorem 2.5 ***(Norm of self-adjoint operators)***

If T is self-adjoint on H, then

$$\|T\|_{\mathcal{L}(H)} = \sup_{\|h\|\leq 1} |\langle Th, h\rangle_H| = \max\{|m|, |M|\}$$

and $\{m, M\} \subseteq \sigma(T) \subseteq [m, M]$, where $m = \inf_{\|h\|=1} \langle Th, h\rangle_H$, $M = \sup_{\|h\|=1} \langle Th, h\rangle_H$.

The set $\sigma_p(T)$ of all eigenvalues of T are composed by $\lambda \in \mathbb{K}$ such that $Th = \lambda h$ for some nonzero $h \in H$. Since $\lambda I - T$ is not injective and thus not boundedly invertible if $\lambda \in \sigma_p(T)$, we have $\sigma_p(T) \subseteq \sigma(T)$. In finite-dimensional cases, the eigenvalues of linear operators are equal to their spectrum, like matrices acting on $\mathbb{R}^n$. However for linear operators acting on infinite-dimensional spaces, $\sigma_p(T) \subsetneq \sigma(T)$, which means the points in spectrum don't have to be eigenvalues. When linear operators are compact, the relationship between their sets of eigenvalues and spectra is clearer, as shown below.

Theorem 2.6 ***(Riesz-Schauder theorem)***
Let $T \in \mathcal{L}(H)$ be a compact operator. Then:

1. *Every nonzero $\lambda \in \sigma(T)$ is an eigenvalue of T and the eigenspace $E_\lambda := \{h \in H : Th = \lambda h\}$ is finite-dimensional;*
2. *for every $r > 0$, the number of eigenvalues satisfying $|\lambda| \geq r$ is finite;*
3. *if there exists a sequence of eigenvalues $(\lambda_n)_{n\geq1}$ such that $\lim_{n\to\infty} \lambda_n = \lambda$, then $\lambda = 0$;*
4. *if dim$(H) = \infty$, then $0 \in \sigma(T)$.*

The Riesz-Schauder theorem shows that the nonzero part of the spectrum of a compact operator is discrete and consists of eigenvalues. The only possible accumulation point zero belongs to $\sigma(T)$ only if H is infinite-dimensional. This can be shown by its converse-negative proposition that if $0 \notin \sigma(T)$, then T is boundedly invertible and so does T^{-1}. Since T is compact, that the unit ball $B = T(T^{-1}B)$ is relatively compact implies dim(H) is finite. When dim$(H) = \infty$, $0 \in \sigma_p(T)$ only when the kernel of T is not just $\{0\}$.

Example 2.8 ***(Spectrum of compact and positive semi-definite operators)***
We define a positive semi-definite operator in $\mathcal{L}(H)$ to be any self-adjoint operator C such that

$$\langle h, C(h)\rangle_H \geq 0, \ \forall h \in H - \{0\}.$$

Suppose $T \in \mathcal{L}(H)$ is compact and positive semi-definite. Then every element in its spectrum $\sigma(T)$ is non-negative.

Proof:
Since T is compact, by Riesz-Schauder theorem 2.6, every nonzero element in $\sigma(T)$ is eigenvalue. Denote $(\xi_i)_{i\geq1}$ by its eigenvectors with corresponding eigenvalues $(\lambda_i)_{i\geq1}$. By the positive semi-definiteness of T, we have that $\forall i \geq 1$,

$$\lambda_i \|\xi_i\|_H^2 = \langle \xi_i, T(\xi_i)\rangle_H \geq 0$$
$$\Longrightarrow \lambda_i = \frac{\langle \xi_i, T(\xi_i)\rangle_H}{\|\xi_i\|_H^2} \geq 0.$$

This means the eigenvalues are non-negative. Since $\sigma(T) = \sigma_p(T)$ or $\sigma(T) = \sigma_p(T) \cup \{0\}$, $\sigma(T)$ consists of non-negative elements.

□

The eigenvalues and eigenspaces of compact operator in $\mathcal{L}(H)$ are crucial for building the orthonormal system in H. And the spectral theorem for compact self-adjoint operators explains that the linear span of eigenspaces is dense in H, showing that the eigenfunctions can form the orthonormal basis of the Hilbert space.

Theorem 2.7 ***(The spectral theorem for compact self-adjoint operators)***
Suppose $T \in \mathcal{L}(H)$ is a compact self-joint operator. Let $(\lambda_n)_{n\geq1}$ be the sequence of its distinct eigenvalues, $(\phi_n)_{n\geq1}$ be the corresponding sequence of eigenfunctions. Then

$$T = \sum_{n\geq1} \lambda_n \phi_n \otimes \phi_n,$$

where $\otimes$ denotes the tensor product (Example 2.7). The convergence is in the operator norm in $\mathcal{L}(H)$.

The spectral theorem for compact self-adjoint operator states that any self-adjoint and compact linear operator has a representation built by its eigenvalues and eigenfunctions. The spectral decomposition of T indicates a representation for images of T:

$$T(h) = \sum_{n\geq 1} \lambda_n \langle h, \phi_n \rangle \phi_n.$$

The spectral theorem for compact self-adjoint operator also makes it possible to deduce the general representation theorem for compact operators acting between Hilbert spaces, which is known as the singular value decomposition theorem for compact operators.

Theorem 2.8 ***(Singular value decomposition theorem for compact operators)***
Suppose H_1 and H_2 are two Hilbert spaces. Let $T \in \mathcal{L}(H_1, H_2)$ be a compact operator. Then there exists a sequence of nonzero eigenvalues $(\sigma_n)_{n\geq 1}$ of the compact operator $(T^*T)^{\frac{1}{2}}$ repeated according to multiplicities, a sequence $(\phi_n)_{n\geq 1} \subseteq H_1$ of eigenfunctions of $(T^*T)^{\frac{1}{2}}$ and an orthonormal sequence $(\tilde{\phi}_n)_{n\geq 1} \subseteq H_2$ such that

$$T = \sum_{n\geq 1} \sigma_n \phi_n \otimes \tilde{\phi}_n$$

converges in the operator norm.

The eigenvalues of compact operator $(T^*T)^{\frac{1}{2}}$ are called singular values of T, which make $T\phi_n = \sigma_n\tilde{\phi}_n$ and $T^*\tilde{\phi}_n = \sigma_n\phi_n$. Moreover, the eigenvalues of $(T^*T)^{\frac{1}{2}}$ are non-negative since the operator is positive semi-definite. This can be seem by the definition that $\forall h \in H - \{0\}$, $\left\langle h, (T^*T)^{\frac{1}{2}}(h) \right\rangle_H = ||T^{\frac{1}{2}}(h)||^2 \geq 0$. ϕ is called the left eigenfunction and $\tilde{\phi}$ is called the right eigenfunction. The tuple $(\sigma_n, \phi_n, \tilde{\phi_n})_{n\geq 1}$ is the singular system of the compact operator T.

The singular value decomposition theorem for compact operators is a powerful tool in finding the solutions of Fredholm integral equation of the first kind, which will be briefly introduced in the next subsection.

2.3. Fredholm integral equation of the first kind

Fredholm integral equation of the first kind appears in the causal inference studies with high frequency. The existence of the bridge functions (Lemma 1.1 and 1.2) in proximal inference is an important case. We introduce the definition of this kind of integral equation problem and give Picard's theorem stating the existence of its solution based on [8] and [29].

Definition 2.10 ***(Fredholm integral equation of the first kind on*** $L_2(S, \mu)$***)***
Given a data function $f \in L_2(S, \mu)$ and a kernel function $k \in \mathcal{C}(S^2, \mu \otimes \mu)$, the integral equation

$$f(x) = \int_S k(x,y)g(y)d\mu(y) \tag{2.10}$$

is called Fredholm integral equation of the first kind, where μ is a finite Borel measure.

The problem is to solve the unknown function g in a known Hilbert space H from the integral equation. Since for compact S and finite Borel measure μ, $\mathcal{C}(S^2, \mu \otimes \mu) \subseteq L_2(S^2, \mu \otimes \mu)$, the Fredholm integral equation of the first kind is featured by a Hilbert-Schmidt integral operator $T_k : H \mapsto L_2(S, \mu)$. We represent the Equation (2.10) by

$$T_k(g) = f.$$

The statement below the singular value decomposition theorem for compact operators (Theorem 2.8) claimed that the integral equation can be solved from the singular value decomposition of the compact linear operator A. In fact, for any $g \in L_2(S,\mu)$, there exists a singular system $(\sigma_i, \phi_i, \tilde{\phi}_i)_{i\geq 1}$ with nonzero singular values, such that

$$f = T_k(g) = \sum_{i\geq 1} \sigma_i \langle g, \phi_i \rangle \tilde{\phi}_i. \tag{2.11}$$

Taking inner product with $\tilde{\phi}_i$ on the both side of Equation (2.11), we get

$$\langle f, \tilde{\phi}_i \rangle = \sigma_i \langle g, \phi_i \rangle .$$

This gives the coefficients of g under orthogonal basis $(\phi_i)_{i\geq 1}$

$$\langle g, \phi_i \rangle = \frac{\langle f, \tilde{\phi}_i \rangle}{\sigma_i}.$$

So the unknown function g has a solution

$$g = \sum_{i\geq 1} \frac{\langle f, \tilde{\phi}_i \rangle}{\sigma_i} \phi_i. \tag{2.12}$$

Formally, the solution (2.12) exists only when the coefficients belongs to $l_2(S)$, i.e. the linear space of all square summable vectors. This is given by the Picard's theorem.

Theorem 2.9 ***(Picard's theorem)***
Let $A : H_1 \mapsto H_2$ be a compact linear operator with singular system $(\sigma_i, \phi_i, \tilde{\phi}_i)$. The equation

$$T_k(g) = f, \ g \in H_1$$

is solvable if and only if f belongs to the orthogonal complement $N(T_k^*)^\perp$ and satisfies

$$\sum_{i\geq 1} \frac{|\langle f, \tilde{\phi}_i \rangle|^2}{\sigma_i^2} < \infty.$$

In this case a solution is given by

$$g = \sum_{i\geq 1} \frac{\langle f, \tilde{\phi}_i \rangle}{\sigma_i} \phi_i.$$

2.4. Reproducing kernel Hilbert space

In this part, we set $\mathbb{K}$ to be the set of all real numbers. We denote the inner product on H by $\langle \cdot, \cdot \rangle_H$ and the inner product of Euclidean spaces by $\langle \cdot, \cdot \rangle$. The Subsection 2.4.1 about the basic definitions and theorems of reproducing kernel Hilbert spaces are collected from [38] and [2].

2.4.1. Basic definitions and theorems

This subsection starts from the definition of RKHS and proves that the reproducing kernels are equivalent to kernel functions by introducing the feature maps on H and the Moore-Aronszajn theorem. The properties of kernel functions provide the reproducing kernels with symmetry and positive semi-definiteness, which are useful for the applications of RKHS.

Definition 2.11 ***(Reproducing kernel Hilbert space (RKHS))***
A Hilbert space H of functions $f : S \mapsto \mathbb{K}$ is a reproducing kernel Hilbert space if all linear functionals $T_x : T_x(f) = f(x)$, $\forall f \in H$, $x \in S$, are bounded on H. The functional considered here is called the point evaluation functional.

Example 2.9 *($H \cong l^2$)*

Consider the Hilbert space of functions from S to $\mathbb{K}$ with an orthonormal basis $(e_i)_{i\geq 1}$.

$$H := \left\{ f : S \mapsto \mathbb{K} \middle| f = \sum_{i\geq 1} a_i e_i, \sum_{i\geq 1} |a_i|^2 < \infty \right\}$$

is a RKHS isometrically isomorphic to l^2.

Example 2.10 ***(Counter example** $L_2(S,\mu)$)*

$L_2(S,\mu)$ is the set of the equivalent classes which contain functions agree almost everywhere. This means any square integrable functions f_1, f_2 by measure μ on S, such that $f_1 = f_2$, $a.e.$, are considered to be the same element in $L_2(S,\mu)$, which is denoted by f. So, in $L_2(S,\mu)$, the point evaluation functional is not well-defined because there exists a subset of S with zero measure such that the value of f is ambiguous. This makes $L_2(S,\mu)$ never a RKHS if no point evaluation functional is able to be defined on.

To introduce the function featuring a RKHS, we define the reproducing kernel functions.

Definition 2.12 ***(Reproducing kernel function)***

Any function $k(\cdot,\cdot) : S \times S \mapsto \mathbb{K}$ is called a reproducing kernel of H if $\forall x \in S$, $k(x,\cdot) : S \mapsto \mathbb{K}$ is an element of H satisfying the reproducing property

$$f(x) = \langle k(x,\cdot), f \rangle_H . \tag{2.13}$$

In RKHS, by virtue of reproducing property, norm convergence implies pointwise convergence.

Theorem 2.10 ***(Convergence in norm implies pointwise convergence)***

Suppose a sequence $(h_i)_{i\geq 1}$ converges to $h \in H$ in the RKHS norm, then $\forall x \in S$, we have

$$\lim_{i\longrightarrow\infty} |h_i(x) - h(x)| = 0.$$

Proof:

By reproducing property (2.13) and Cauchy-Schwartz inequality,

$$\lim_{i\longrightarrow\infty} |h_i(x) - h(x)| = \lim_{i\longrightarrow\infty} |\langle h_i - h, k(x,\cdot)\rangle_H| \leq \|h_i - h\|_H \|k(x,\cdot)\|_H.$$

□

Theorem 2.11 *A Hilbert space H of functions $f : S \mapsto \mathbb{K}$ is a reproducing kernel Hilbert space if and only if it has a reproducing kernel.*

Proof:

If H is a RKHS, then by Definition 2.11, $\forall x \in S$, all point evaluation functionals on it are bounded. For a point evaluation functional $T_x : H \mapsto \mathbb{K}$, by the Riesz representation theorem (Theorem 2.2), there exists a unique representative function $k_x : S \mapsto \mathbb{K}$ in H such that

$$T_x(f) = \langle k_x, f \rangle_H . \tag{2.14}$$

We define a bivariate function $k : S \times S \mapsto \mathbb{K}$ by $k(x,y) = \langle k_x, k_y \rangle_H$, $\forall x, y \in S$. Applying the point evaluation functional T_y to k_x, by the symmetry of inner product, we have $k_x(y) = T_y(k_x) = \langle k_x, k_y \rangle_H$. This implies $k(x,\cdot) = k_x \in H$. In addition, the reproducing property of k is directly given by replacing k_x by $k(x,\cdot)$ in (2.14). So, k is a reproducing kernel of H.

If H has a reproducing kernel $k(\cdot,\cdot) : S \times S \mapsto \mathbb{K}$, its reproducing property implies it canonically corresponds to a point evaluation functional $T_x : H \mapsto \mathbb{K}$, $\forall x \in S$. $\forall f \in H$, $|T_x(f)| = |\langle k_x, f \rangle_H| \leq \|k_x\|_H \|f\|_H < \infty$. This means T_x is bounded with norm $\|k_x\|_H$ and hence H is a RKHS.

□

The following theorem shows the relationship between the reproducing kernel function and the orthonormal basis of the RKHS.

Theorem 2.12 *([36]**Reproducing kernel expansion with orthonormal basis**)*
Let H be a RKHS over S with a reproducing kernel function $k(\cdot,\cdot) : S \times S \mapsto \mathbb{K}$. If $(e_i)_{i\geq 1}$ is an orthonormal basis of H, then $\forall x, x' \in S$

$$k(x,x') = \sum_{i\geq 1} e_i(x)e_i(x'),$$

where the convergence is absolute.

From the theorem, we also have that the function $k(x,\cdot) \in H$ has expression

$$k(x,\cdot) = \sum_{i\geq 1} e_i(x)e_i,$$

which means it is a sum of basis functions $e_i : S \mapsto \mathbb{K}$ scaled by $(e_i(x))_{i\geq 1}$.

Example 2.11 *(**Cont. Example 2.9**)*
Recall the orthonormal basis function $(e_i)_{i\geq 1}$ of H satisfies

$$\left\langle e_i(\cdot), e_j(\cdot)\right\rangle_H = \begin{cases} 1, \ i = j \\ 0, \ i \neq j \end{cases}. \tag{2.15}$$

For RKHS isometrically isomorphic to l^2, we can reproduce any function in H by

$$f(x) = \langle k(x,\cdot), f\rangle_H = \left\langle \sum_{i\geq 1} e_i(x)e_i, \sum_{i\geq 1} a_i e_i \right\rangle_H \underbrace{=}_{(2.15)} \sum_{i\geq 1} a_i e_i(x) = \langle \vec{e}(x), \vec{a}\rangle.$$

This means we can simplify the reproducing property of such RKHS by the inner product of Euclidean spaces between $\vec{a} = (a_1, a_2, \cdots, a_n, \cdots)^T \in l^2$ and $\vec{e}(x) = (e_1(x), e_2(x), \cdots, e_n(x), \cdots)^T$.

The reproducing kernel function has a strong relationship with the kernel functions. In fact, the idea of the reproducing kernel function is equivalent to the kernel function under a RKHS background. We will show this step by step starting from the kernel functions and feature maps.

Definition 2.13 *(**Kernel function**)*
The function $k(\cdot,\cdot) : S \times S \mapsto \mathbb{K}$ is a kernel function if

1. *k is symmetric: $k(x,y) = k(y,x)$, $\forall x, y \in S$.*
2. *k is positive semi-definite[1]: The Gram matrix K with entry $K_{ij} = k(x_i, x_j)$, $\forall x_1, \cdots, x_n \in S$, is positive semi-definite.*

Definition 2.14 *(**Feature map**)*
Given a Hilbert space F, a feature map ϕ maps elements from S to F. F is also called a feature space.

The inner product of feature map directly leads to the positive semi-definiteness of the kernel functions.

Proposition 2.4 *The function $k(\cdot,\cdot) : S \times S \mapsto \mathbb{K}$ is a kernel function if there exists a feature map $\phi : S \mapsto F$ such that*

$$k(\cdot,\cdot) = \langle \phi(\cdot), \phi(\cdot)\rangle_F.$$

[1] In some literatures positive semi-definite kernels are called positive definite kernels based on the theory of positive definite functions.

Proof:

Since the symmetry of $k(\cdot,\cdot)$ is a direct result of the symmetry of the inner product, we only show that $k(\cdot,\cdot)$ is positive semi-definite. $\forall \alpha = (\alpha_1, \cdots, \alpha_n)^T \in \mathbb{K}^n$, given the Gram matrix K with entry $K_{ij} = k(x_i, x_j)$, $\forall x_1, \cdots, x_n \in S$,

$$\begin{aligned}
\alpha^T K \alpha &= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \left\langle \phi(x_i), \phi(x_j) \right\rangle_F \\
&= \left\langle \sum_{i=1}^{n} \alpha_i \phi(x_i), \sum_{i=1}^{n} \alpha_i \phi(x_i) \right\rangle_F \\
&= \| \sum_{i=1}^{n} \alpha_i \phi(x_i) \|_F^2 \\
&\geq 0.
\end{aligned}$$

□

So, any kernel functions can be represented by the inner product of feature maps on a Hilbert space. Given the inner product defined on the Hilbert space, one can find the feature map corresponding to the kernel function.

Example 2.12 *(**Common kernel functions and corresponding feature maps**)*

The polynomial kernels and Gaussian radial basis function (RBF) kernel are widely used in machine learning especially in support vector machine (SVM), kernel ridge regression and nonlinear modeling tasks [43, 5], which are given by

- *Polynomial kernel: $k_p(x, y) = (\langle x, y\rangle + c)^d$, where $x, y \in S$, $c \in \mathbb{K}$ and $d \in \mathbb{N}^+$.*
- *Gaussian radial basis function (RBF) kernel: $k_G(x, y) = \exp\{-\gamma \|x - y\|_2^2\}$, where $x, y \in S$ and $\gamma \in \mathbb{K}^+$.*

The feature map of a polynomial kernel depends on the order d and the dimension of S. For example when dim$(S) = d = 2$ and $c = 0$,

$$\begin{aligned}
k_p(x, y) = (x_1 y_1 + x_2 y_2)^2 &= x_1^2 y_1^2 + x_2^2 y_2^2 + 2x_1 x_2 y_1 y_2 \\
&= (x_1^2, x_2^2, \sqrt{2}x_1 x_2)(y_1^2, y_2^2, \sqrt{2}y_1 y_2)^T \\
&= (x_1^2, x_2^2, x_1 x_2, x_1 x_2)(y_1^2, y_2^2, y_1 y_2, y_1 y_2)^T.
\end{aligned}$$

This means the feature map and corresponding feature space are not uniquely determined. In fact, we find $\phi_p : \phi_p(x) = (x_1^2, x_2^2, \sqrt{2}x_1 x_2)$ implying a feature space $F_p \subseteq \mathbb{K}^3$, or $\tilde{\phi}_p : \tilde{\phi}_p(x) = (x_1^2, x_2^2, x_1 x_2, x_1 x_2)$ implying a feature space $\tilde{F}_p \subseteq \mathbb{K}^4$.

The feature map of the Gaussian RBF kernel depends on the basis function and Taylor expansion of exponential function. In fact,

$$\begin{aligned}
k_G(x, y) &= \exp\{-\gamma \|x - y\|_2^2\} \\
&= \exp\{2\gamma \langle x, y\rangle\} \cdot \exp\{-\gamma \|x\|_2^2\} \cdot \exp\{-\gamma \|y\|_2^2\} \\
&= \sum_{n=0}^{\infty} \frac{(2\gamma)^n}{n!} (\left\langle \exp\{-\gamma \|x\|_2^2\} x, \exp\{-\gamma \|y\|_2^2\} y \right\rangle)^n.
\end{aligned}$$

The feature map ϕ_G of k_G doesn't have an explicit form unless the dimension of S is 1. In this case, $\phi_G : \phi_G(x) = \left[1, \sqrt{\frac{4\gamma^2}{2}} x^2 e^{-\gamma x^2}, \cdots, \sqrt{\frac{(2\gamma)^n}{n!}} x^n e^{-\gamma x^2}, \cdots\right]$ implying an infinite dimensional feature space $F_G \subseteq l^2$. For general situation, Steinwart et.al [36] gave the structure of the RKHS of Gaussian RBF kernel where the feature map is subsequently discussed.

If the Hilbert space is a RKHS, then the reproducing kernel is just a kernel function, which means it can also be represented by the inner product of feature maps. This is shown in the following proposition.

Proposition 2.5 *Every reproducing kernel function is a kernel function.*

Proof:
Suppose $k(\cdot,\cdot) : S \times S \mapsto \mathbb{K}$ is the reproducing kernel function of RKHS H. By Definition 2.12, $k(x,\cdot) \in H$. Then there exists a feature map $\phi : S \mapsto H$ such that $\forall x \in S$, $\phi(x) = k(x,\cdot)$. $\forall y \in S$, we have

$$\begin{aligned} k(x,y) &= \langle k(x,\cdot), k(y,\cdot)\rangle_H \qquad &\text{(reproducing property (2.13), } k(x,\cdot) \in H) \\ &= \langle \phi(x), \phi(y)\rangle_H . \end{aligned}$$

□

This proposition is crucial for showing the uniqueness of reproducing kernel given a RKHS.

Theorem 2.13 *(**Uniqueness of reproducing kernel**)*
A RKHS uniquely determines its reproducing kernel function.

Proof:
Suppose $k_1(\cdot,\cdot)$ and $k_2(\cdot,\cdot)$ are the reproducing kernel functions of RKHS H. By Definition 2.12, $\forall x \in S$, $k_1(x,\cdot) \in H$ and $k_2(x,\cdot) \in H$. By Proposition 2.5, $k_1(\cdot,\cdot)$ and $k_2(\cdot,\cdot)$ are symmetric. By reproducing property (2.13), $\forall x, y \in S$, we have

$$k_1(x,y) = \langle k_1(x,\cdot), k_2(y,\cdot)\rangle_H = k_2(x,y).$$

This means $k_1 = k_2$ everywhere on $S \times S$.

□

Recall in Example 2.12 a kernel may not uniquely determines the feature map and the feature space. However, it is able to uniquely determine a RKHS by the following theorem.

Theorem 2.14 *([12]**Moore-Aronszajn theorem**)*
Suppose $k : S \times S \mapsto \mathbb{K}$ is a kernel function. Then there is a unique Hilbert space of functions on S for which k is a reproducing kernel.

Proof:
Given a kernel function k, for $n \in \mathbb{N}^+$, arbitrarily choosing points $x_1, x_2 \cdots, x_n \in S$, we consider the linear span of $k(x_1,\cdot), k(x_2,\cdot) \cdots, k(x_n,\cdot)$.
The elements in the linear span have representation $f = \sum_{i=1}^{n} \alpha_i k(x_i,\cdot)$, $\alpha_i \in \mathbb{K}$. Denote the closure of the linear span by

$$H = \overline{\left\{ f : f = \sum_{i=1}^{n} \alpha_i k(x_i,\cdot),\ x_i \in S,\ \alpha_i \in \mathbb{K},\ n \in \mathbb{N}^+ \right\}}.$$

$\forall f, g \in H$, $\forall m \in \mathbb{N}$, the representations are given by

$$f = \sum_{i=1}^{n} \alpha_i k(x_i,\cdot),\ g = \sum_{i=1}^{m} \beta_i k(y_i,\cdot).$$

The inner product $\langle\cdot,\cdot\rangle_H$ here is defined by $\langle f,g\rangle_H = \sum_i^n \sum_{j=1}^m \alpha_i\beta_j k(x_i,y_j)$, which is verified by its symmetry, linearity and positive definiteness. The symmetry is a direct result of the symmetric kernel function and the linearity is derived from the linear representations of the functions. For the positive definiteness, since the positive semi-definiteness is shown in the proof of Proposition 2.4, we only need to show $\langle f,f\rangle_H = 0 \Rightarrow f(x) = 0$, $\forall x \in S$. In fact, $\forall x \in S$,

$$\begin{aligned}
|f(x)|^2 &= |\sum_{i=1}^n \alpha_i k(x_i,x)|^2 \\
&= \langle f(\cdot), k(x,\cdot)\rangle_H^2 && (2.16) \\
&\leq \underbrace{\langle f,f\rangle_H}_{0} \langle k(x,\cdot),k(x,\cdot)\rangle_H && \text{(Cauchy-Schwartz inequality)} \\
&= 0.
\end{aligned}$$

The Equation (2.16) shows that the kernel function satisfies the reproducing property (2.13). This means k is the reproducing kernel of H.

To verify the uniqueness of H, we suppose there exists a Hilbert space H_0 such that $H \subseteq H_0$ and k is its reproducing kernel. Since H is a closed subspace of H_0, $H_0 = H \oplus H^\perp$. For any $f \in H_0$, $f = f_H + f_{H^\perp}$, where $f_H \in H$ and $f_{H^\perp} \in H^\perp$. From the fact that k is the reproducing kernel in both H and H_0, we have

$$\begin{aligned}
f(x) = \langle k(x,\cdot), f\rangle_{H_0} &= \langle k(x,\cdot), f_H\rangle_{H_0} + \langle k(x,\cdot), f_{H^\perp}\rangle_{H_0} \\
&= \langle k(x,\cdot), f_H\rangle_{H_0} && (k \in H) \\
&= f_H(x).
\end{aligned}$$

This means $\forall f \in H_0$, $f = f_H$ and thus $H = H_0$.

□

Example 2.13 ***(Cont. Example 2.12)***

By Moore-Aronszajn theorem, the RKHSs induced by polynomial kernel $k_p(x,y) = \langle x,y\rangle^2$ and Gaussian RBF kernel $k_G = \exp\{-\gamma|x-y|^2\}$ are given by[2]

$$H_p = \overline{\left\{ f : \sum_{i=1}^n \alpha_i k_p(x_i,\cdot), \alpha_i \in \mathbb{K}, n \in \mathbb{N}^+ \right\}},\ k_p(x,\cdot) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)(\cdot)$$

$$H_G = \overline{\left\{ g : \sum_{i=1}^n \beta_i k_G(x_i,\cdot), \beta_i \in \mathbb{K}, n \in \mathbb{N}^+ \right\}},\ k_G(x,\cdot) = \left(\sqrt{\frac{(2\gamma)^n}{n!}} x^n \exp\{-\gamma x^2\} \right)_{n\geq 0} (\cdot).$$

For any $f \in H_p$, $g \in H_G$, $y \in \mathbb{K}^2$, $z \in \mathbb{K}$,

$$f(y) = \left\langle k_p(y,\cdot), f\right\rangle_{H_p} = \sum_{i=1}^n \alpha_i k_p(x_i,y) = \sum_{i=1}^n \alpha_i \left\langle \phi_p(x_i), \phi_p(y)\right\rangle^2 = \sum_{i=1}^n \alpha_i (x_{i1}^2 y_1^2 + x_{i2}^2 y_2^2 + 2x_{i1}x_{i2}y_1y_2)$$

$$g(z) = \langle k_G(z,\cdot), g\rangle_{H_G} = \sum_{i=1}^n \beta_i k_G(x_i,z) = \sum_{i=1}^n \beta_i \langle \phi_G(x_i), \phi_G(z)\rangle^2 = \sum_{i=1}^n \beta_i \exp\{-\gamma|x_i - z|^2\}.$$

Hence, f matches a unique scalar vector $\sum_{i=1}^n \alpha_i (x_{i1}^2, x_{i2}^2, \sqrt{2}x_{i1}x_{i2})^T \in F_p$ and g matches a unique scalar vector $\sum_{i=1}^n \beta_i \left(\sqrt{\frac{(2\gamma)^n}{n!}} x_i^n \exp\{-\gamma x_i^2\} \right)_{n\geq 0} \in F_G$. This implies that $H_p \cong F_p$ and $H_G \cong F_G$ and $\langle\cdot,\cdot\rangle_{H_p} \equiv \langle\cdot,\cdot\rangle_{H_G} \equiv \langle\cdot,\cdot\rangle$.

[2] Here $x_i \in \mathbb{K}^2$ for k_p and $x_i \in \mathbb{K}$ for k_G.

2.4.2. Continuous functions approximation

In this part, we introduce the idea of universal kernel that is decisive for finding a RKHS to approximate the space of all continuous functions on a compact subset of $\mathbb{R}^n$.

Lemma 2.1 ***(Continuous kernel)***

Given a kernel $k(\cdot,\cdot) : S \times S \mapsto \mathbb{R}$ and its induced RKHS H. k is continuous if its corresponding feature map $\phi : S \mapsto H$ is continuous.

Proof:

The continuity of ϕ implies $\phi(\cdot)$ is a continuous function in H. By the virtue of the compactness of S, $\forall x, y \in S$, $\phi(x)$ and $\phi(y)$ are bounded.

$$\begin{aligned}|k(x,y) - k(x',y')| &\leq |k(x,y) - k(x',y) + k(x',y) - k(x',y')| \\ &\leq |\langle\phi(x) - \phi(x'), \phi(y)\rangle_H| + |\langle\phi(x'), \phi(y) - \phi(y')\rangle_H| \qquad \text{(Proposition 2.4)} \\ &\leq \|\phi(y)\|_H\|\phi(x) - \phi(x')\|_H + \|\phi(x')\|_H\|\phi(y) - \phi(y')\|_H. \quad \text{(Cauchy-Schwartz)}\end{aligned}$$

By the boundedness and continuity of $\phi(\cdot)$, we have the continuity of kernel function $k(\cdot,\cdot)$.

□

Definition 2.15 *([37]**Universal kernel**)*

A continuous kernel $k(\cdot,\cdot) : S \times S \mapsto \mathbb{R}$ is universal if the space of functions H induced by it is dense in the space of all bounded continuous functions on S.

Given a universal kernel acting on $S \times S$, by Theorem 2.14, one can construct a RKHS with the universal kernel as its reproducing kernel. The universality makes sure that for any function $g \in \mathcal{C}(S)$, for any $\epsilon > 0$, there always exists a function f in this RKHS such that

$$\|f - g\|_\infty \leq \epsilon.$$

We denote the space of all continuous functions on a compact subset S of $\mathbb{K}^n$ by $\mathcal{C}(S)$ and the space of all bounded continuous functions on $\mathbb{K}^m$ by $\mathcal{C}_0(\mathbb{K}^m)$.

It is only possible for the RKHS generated by a universal kernel to be dense in spaces whose element functions have infinite infinity norm. The reason is that for any RKHS element f,

$$|f(X)| = |\langle f, k(x,\cdot)\rangle_H| \leq \|f\|_H\sqrt{k(X,X)} < \infty,$$

which implies $\|f\|_\infty < \infty$.

Therefore, since the infinity norm of functions in $\mathcal{C}(S)$ and $\mathcal{C}_0(\mathbb{K})$ are all finite, $\mathcal{C}(S)$ and $\mathcal{C}_0(\mathbb{K})$ can be well approximated by the RKHS generated by an universal kernel acting on the corresponding sets.

Steinwart [37] shows that kernels are universal on any compact subset of certain sets, if they can be expressed by Taylor expandable infinitely differentiable functions or Fourier expandable continuous functions. This further extends to the universality of the Gaussian RBF kernel, infinite polynomial kernel, stronger regularized Fourier kernel and weaker regularized Fourier kernel. One can check Corollary 10 and 11 as well as Example 1 to 4 in [37] for details.

2.4.3. Kernel embedding of mean and cross-covariance

Based on the results from [33] and [35], this part aims at finding methods to compute the expectations and covariances of RKHS functions without distributions of random variables but only with inner product between RKHS functions.

Let D be a compact subset of $\mathbb{K}^n$. Suppose $k(\cdot,\cdot) : S \times S \mapsto \mathbb{K}$ and $l(\cdot,\cdot) : D \times D \mapsto \mathbb{K}$ are kernel functions with induced RKHS H and F respectively. Denote the corresponding feature maps[3] of the two kernel functions by $\phi : S \mapsto H$and $\psi : D \mapsto F$. Let X and Y be random variables defined on S and D separately with marginal distribution P_X and P_Y.

[3] Recall the trick used in the proof of Proposition 2.5 that one can choose the feature map $\phi(x)$ uniquely corresponding to a RKHS function $k(x,\cdot)$.

Definition 2.16 ***(Mean value operator)***
The mean value operator is a linear operator $T_X : H \mapsto \mathbb{K}$ such that $\forall h \in H$

$$T_X(h) := E_X[h(X)] = \int_S h(x) dP_X.$$

The empirical operator is similarly defined.

Definition 2.17 ***(Empirical mean value operator)***
The empirical mean value operator is a linear operator $\widehat{T}_X : H \mapsto \mathbb{K}$ such that $\forall h \in H$

$$\widehat{T}_X(h) := \overline{h(X)} = \frac{1}{n} \sum_{i=1}^{n} h(X_i),$$

where $X_1, \cdots, X_n$ are i.i.d from P_X.

The following lemma shows that the mean value operator uniquely corresponds to a function in H.

Lemma 2.2 ***(Boundedness of the mean value operator)***
If $E_X[\sqrt{k(X,X)}] < \infty$, the mean value operator T_X is a bounded linear functional and thus corresponds to a unique element in H.

Proof:
For any $h \in H$,

$$\begin{aligned} |T_X(h)| = |E_X[h(X)]| &\leq E_X[|h(X)|] && \text{(Jensen's inequality 2.18)} \\ &= E_X[| \langle h, \phi(X) \rangle_H |] && \text{(Reproducing property (2.13))} \\ &\leq E_X[\sqrt{k(X,X)}] \|h\|_H < \infty && \text{(Cauchy-Schwartz \& Reproducing property)} \end{aligned}$$

This implies $\|T_X\|_{\mathcal{L}(H,\mathbb{K})}$ is bounded by $E_X[\sqrt{k(X,X)}]$. By Riesz representation theorem 2.2, T_X canonically corresponds to a unique function in H, which we denote by μ_X. This gives

$$T_X(h) = \langle h, \mu_X \rangle_H . \tag{2.17}$$

□

We can similarly show that the empirical mean value operator $\widehat{T}_X$ uniquely corresponds to a function $\widehat{\mu}_X \in H$ under assumption $\frac{1}{n} \sum_{i=1}^{n} \sqrt{k(x_i, x_i)} < \infty$, with i.i.d. $x_1 = x_1, \cdots, X_N = x_n$ from P_X, which satisfies

$$\widehat{T}_X(h) = \langle h, \widehat{\mu}_X \rangle_H .$$

The lemma implies that if the mean value operator exists, one can find the expectation of a RKHS function by the RKHS inner product between the function and μ_X directly. This shows that μ_X plays the similar role of distribution P_X in computing expectations of RKHS functions. Hence, it is important to find the explicit form of μ_X and $\widehat{\mu}_X$ as elements in H.

Proposition 2.6 ***((Empirical) mean embedding)***
The explicit forms of μ_X and $\widehat{\mu}_X$ as elements in H are given by $\mu_X = E_X[\phi(X)]$ and $\widehat{\mu}_X = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i)$.

Proof:
By reproducing property (2.13), $\forall h \in H$,

$$E_X[h(X)] = E_X[\langle h, \phi(X) \rangle_H] = \big\langle h, E_X[\phi(X)] \big\rangle_H .$$

By (2.17), this further implies

$$\mu_X = E_X[\phi(X)]. \tag{2.18}$$

Similarly, the explicit form of $\widehat{\mu}_X$ is given by

$$\widehat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n}\phi(X_i). \tag{2.19}$$

□

Hence, for any $h \in H$, its expectation or empirical mean relative to random variable X can be calculated by $E_X[h(X)] = \langle h, \mu_X\rangle_H$ and $\overline{h(X)} = \langle h, \widehat{\mu}_X\rangle_H$.

The process (2.18) and (2.19) are called the (empirical) mean embedding of the distribution P_X, where the distribution P_X is mapped to the feature space H as the expectation or empirical mean of the RKHS function $\phi(X) = k(X,\cdot)$.

The mean embedding μ_X and its empirical form $\widehat{\mu}_X$ share good properties. Altun and Smola [1] showed that the convergence speed in RKHS norm is $\mathcal{O}_p(n^{-\frac{1}{2}})$, which is given by the following theorem.

Theorem 2.15 ***($\sqrt{n}$-consistency of mean embedding)***

If $E_X[k(X,X)] < \infty$ and $\frac{1}{n}\sum_{i=1}^{n} k(x_i,x_i) < \infty$, the mean value operator and corresponding empirical mean value operator admit mean embeddings. Moreover,

$$\|\widehat{\mu}_X - \mu_X\|_H = \mathcal{O}_p(n^{-\frac{1}{2}}).$$

Proof:

By Jensen's inequality and concavity of square root,

$$E_X[\sqrt{k(X,X)}] \leq \sqrt{E_X[k(X,X)]} < \infty, \text{ and } \frac{1}{n}\sum_{i=1}^{n}\sqrt{k(x_i,x_i)} \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}k(x_i,x_i)} < \infty.$$

Hence the mean embedding and empirical embedding exist by Lemma 2.2.

$$\begin{aligned}
E_X\left[\|\widehat{\mu}_X - \mu_X\|_H\right] &\leq \left(E_X\left[\|\widehat{\mu}_X - \mu_X\|_H^2\right]\right)^{\frac{1}{2}} && \text{(Jensen's inequality)}\\
&= \frac{1}{n}\left(E_X\left[\left\|\sum_{i=1}^{n}(\phi(X_i) - \mu_X)\right\|_H^2\right]\right)^{\frac{1}{2}}\\
&= \frac{1}{n}\Bigg(\sum_{i=1}^{n}E_{X_i}\left[\left\|\phi(X_i) - \mu_X\right\|_H^2\right]\\
&\qquad + \underbrace{\sum_{i\neq j}^{n}E_{X_i,X_j}\left[\left\langle\phi(X_i) - \mu_X, \phi(X_j) - \mu_X\right\rangle_H\right]}_{\text{Vanishes by } X_1\cdots X_n \text{ i.i.d}}\Bigg)^{\frac{1}{2}} && (2.20)\\
&= \frac{1}{n}\left(nE_X\left[\left\|\phi(X) - \mu_X\right\|_H^2\right]\right)^{\frac{1}{2}} && (X_1\cdots X_n \text{ i.i.d})\\
&= n^{-\frac{1}{2}}\sqrt{E_X[k(X,X)] - \|\mu_X\|_H^2} = \mathcal{O}(n^{-\frac{1}{2}}). && \text{(Reproducing property)}
\end{aligned}$$

For any $h \in H$, by (2.18) we have

$$E_X\left[\langle h, \phi(X) - \mu_X\rangle_H\right] = \left\langle h, E_X[\phi(X)]\right\rangle_H - \langle h, \mu_X\rangle_H = E_X[h(X)] - E_X[h(X)] = 0.$$

For (2.20), since $X_1, \cdots, X_n$ are i.i.d. samples from P_X, we have

$$\sum_{i \neq j}^{n} E_{X_i, X_j}\left[\left\langle \phi(X_i) - \mu_X, \phi(X_j) - \mu_X \right\rangle_H\right] = \sum_{i \neq j}^{n} E_{X_i}\left[\underbrace{E_{X_j}\left[\left\langle \phi(X_i) - \mu_X, \phi(X_j) - \mu_X \right\rangle_H\right]}_{\text{Vanishes by choosing } h=\phi(X_i)-\mu_X}\right] = 0.$$

□

Moreover, when the kernel k is universal, the map from P_X to μ_X is injective. This means if the mean embedding of two random variables are the same then their distribution are identical. The injectivity of the map is derived from a classic result in probability theory that for any two Borel probability measures P and Q defined on metric space (S, d), $P = Q$ if and only if $E_A[f(A)] = E_B[f(B)]$ for any $f \in \mathcal{C}(S)$, where $A \sim P$ and $B \sim Q$[13].

A brief proof is given as follows. Assume $\mu_A = \mu_B$. By the fact that the RKHS induced by k is dense in $\mathcal{C}(S)$, we suppose $\forall \epsilon > 0$, $\exists g \in H$ such that $\|f - g\|_\infty < \epsilon$. Then

$$|E_P[f(A)] - E_Q[f(B)]| < \underbrace{|E_P[f(A)] - E_P[g(A)]|}_{\leq \epsilon} + \underbrace{|E_P[g(A)] - E_Q[g(B)]|}_{\langle g, \mu_A - \mu_B \rangle_H = 0} + \underbrace{|E_Q[g(B)] - E_Q[f(B)]|}_{\leq \epsilon} \leq 2\epsilon.$$

Before introducing the cross-covariance operators, we define the tensor product space.

Definition 2.18 *([26]**Tensor product & tensor product space**)*

Given two linear spaces V_1 and V_2 over $\mathbb{K}$, the tensor product space denoted by $V_1 \otimes V_2$ is the linear span of set $\{v_1 \otimes v_2 : v_1 \in V_1,\ v_2 \in V_2\}$, where $\otimes$ can be any operation satisfying bilinearity rules

1. $c(v_1 \otimes v_2) = (cv_1) \otimes v_2 = v_1 \otimes (cv_2)$, $\forall c \in \mathbb{K}$;
2. $v_1 \otimes (v_2 + v) = v_1 \otimes v_2 + v_1 \otimes v$, $\forall v \in V_2$;
3. $(v' + v_1) \otimes v_2 = v' \otimes v_2 + v_1 \otimes v_2$, $\forall v' \in V_1$.

Example 2.14 *(**Tensor product on finite spaces**)*

If $V_1 = \mathbb{K}^n$ and $V_2 = \mathbb{K}^m$, then $\forall v_1 \in V_1$, $\forall v_2 \in V_2$, we have $v_1 \otimes v_2 = v_1 v_2^T \in \mathbb{K}^{n \times m}$ and $v_2 \otimes v_1 = v_2 v_1^T \in \mathbb{K}^{m \times n}$. The tensor product here is matrix multiplication. When $n = m = 1$, the tensor product degenerates to multiplication between scalars.

Example 2.15 *(**Tensor product feature space**)*

$F \otimes H$ is a tensor product space with a tensor product given by the rank-one operator (2.1). This tensor product space is a Hilbert space of rank-one operators from H to F, which is equipped with Hilbert-Schmidt inner product $\langle \cdot, \cdot \rangle_{HS(H,F)}$ and is complete relative to the induced norm $\| \cdot \|_{HS(H,F)}$. Moreover, $F \otimes H$ is a RKHS with reproducing kernel kl. The corresponding feature map from $D \times S$ to $F \otimes H$ is $\psi \otimes \phi$.

We can generalize the mean embedding to the situation on tensor product feature space $F \otimes H$ where the joint distribution P_{XY} is mapped to this feature space.

Definition 2.19 *(**Cross-covariance operator**)*

The cross-covariance operator is a linear operator $Cov_{XY} : F \otimes H \mapsto \mathbb{K}$ such that $\forall f \otimes h \in F \otimes H$,

$$Cov_{XY}(f \otimes h) := E_{XY}[(f(Y) - E_Y[f(Y)])(h(X) - E_X[h(X)])]. \tag{2.21}$$

The empirical form is similarly defined.

Definition 2.20 *(**Empirical cross-covariance operator**)*

The empirical cross-covariance operator is a linear operator $\widehat{Cov}_{XY} : F \otimes H \mapsto \mathbb{K}$ such that $\forall f \otimes h \in F \otimes H$,

$$\widehat{Cov}_{XY}(f \otimes h) := \frac{1}{n} \sum_{i=1}^{n} [(f(Y_i) - \overline{f(Y)})(h(X_i) - \overline{h(X)})], \tag{2.22}$$

where $(X_1, Y_1), \cdots, (X_n, Y_n)$ are i.i.d from $P_X \times P_Y$.

Like Lemma 2.2, we have the following lemma showing that the cross-covariance operator uniquely corresponds to an element in $F \otimes H$.

Lemma 2.3 ***(Boundedness of cross-covariance operator)***
If $E_{XY}[\sqrt{k(X,X)l(Y,Y)}] < \infty$, $E_X[\sqrt{k(X,X)}] < \infty$ and $E_Y[\sqrt{k(Y,Y)}] < \infty$, the cross-covariance operator Cov_{XY} is a bounded linear functional and thus corresponds to a unique element in $F \otimes H$.

Proof:

$$\begin{aligned}|Cov_{XY}(f \otimes h)| &= |E_{XY}[(f(Y) - E_Y[f(Y)])(h(X) - E_X[h(X)])]| \\ &\leq |E_{XY}[f(Y)h(X)]| + |E_X[h(X)]E_Y[f(Y)]|,\end{aligned}$$

where

$$\begin{aligned}|E_{XY}[f(Y)h(X)]| &= |E_{XY}[\langle f, \psi(Y)\rangle_F \langle h, \phi(X)\rangle_H]| \\ &\leq E_{XY}[|\langle f, \psi(Y)\rangle_F \langle h, \phi(X)\rangle_H|] && \text{(Jensen's ineq.)} \\ &\leq \|f\|_F \|h\|_H E_{XY}[\sqrt{k(X,X)l(Y,Y)}] < \infty. && \text{(Cauchy-Schwartz)}\end{aligned}$$

Since $E_X[\sqrt{k(X,X)}] < \infty$ and $E_Y[\sqrt{k(Y,Y)}] < \infty$ imply the existence of mean embeddings μ_X and μ_Y, by Cauchy-Schwartz inequality and reproducing property,

$$E_X[h(X)]E_Y[f(Y)] \leq \|\mu_X\|_H \|h\|_H \|\mu_Y\|_F \|f\|_F < \infty.$$

Hence $\|Cov_{XY}\|_{\mathcal{L}(F\otimes H, \mathbb{K})}$ is bounded. By Riesz representation theorem, there exists a unique element in $F \otimes H$ denoted by C_{XY} such that

$$Cov_{XY}(f \otimes h) = \langle C_{XY}, f \otimes h\rangle_{HS(H,F)}.$$

□

We can similarly show that the empirical cross-covariance operator uniquely corresponds to an element in $F \otimes H$ under assumptions $\frac{1}{n}\sum_{i=1}^{n}\sqrt{k(x_i,x_i)l(y_i,y_i)} < \infty$, $\frac{1}{n}\sum_{i=1}^{n}\sqrt{k(x_i,x_i)} < \infty$ and $\frac{1}{n}\sum_{i=1}^{n}\sqrt{l(y_i,y_i)} < \infty$, which satisfies

$$\widehat{Cov}_{XY}(f \otimes h) = \langle \widehat{C}_{XY}, f \otimes h\rangle_{HS(H,F)}.$$

Since C_{XY} and $\widehat{C}_{XY}$ are all elements in $HS(H,F)$, they are Hilbert-Schmidt operators. This can be proved by the definition of Hilbert-Schmidt operators.

Proposition 2.7 ***(Hilbert-Schmidt equivalence)***
If $E_{XY}[k(X,X)l(Y,Y)] < \infty$, $E_X[k(X,X)] < \infty$ and $E_Y[l(Y,Y)] < \infty$, then $C_{XY} : H \mapsto F$ is a Hilbert-Schmidt operator in $HS(H,F)$.

Proof:
Suppose $(h_i)_{i\geq 1}$ is an orthonormal basis of H. Then,

$$\begin{aligned}\|C_{XY}\|_{HS}^2 &= \sum_{i=1}^{\infty}\|C_{XY}(h_i)\|_F^2 \\ &= \sum_{i=1}^{\infty}\|E[(h_i(X) - E[h_i(X)])(\psi(Y) - \mu_Y)]\|_F^2 \\ &\leq \underbrace{\sum_{i=1}^{\infty}\|E[h_i(X)\psi(Y)]\|_F^2}_{I_1} + \underbrace{\sum_{i=1}^{\infty}\|E[h_i(X)]\mu_Y\|_F^2}_{I_2} + \underbrace{2\sum_{i=1}^{\infty}|\langle E[h_i(X)\psi(Y)], E[h_i(X)]\mu_Y\rangle_F|}_{I_3}.\end{aligned}$$

For I_2,

$$\begin{aligned} I_2 := \sum_{i=1}^{\infty} E^2[\langle\phi(X), h_i\rangle_H]\|\mu_Y\|_F^2 &\leq E[\sum_{i=1}^{\infty}\langle\phi(X), h_i\rangle_H^2]\|\mu_Y\|_F^2 && \text{(Jensen's inequality)} \\ &= E[\|\phi(X)\|_H^2]\|\mu_Y\|_F^2 && \text{(Parseval's identity)} \\ &= E[k(X,X)]\|\mu_Y\|_F^2 < \infty. \end{aligned}$$

For I_1,

$$\begin{aligned} I_1 := \sum_{i=1}^{\infty}\|E[\langle\phi(X), h_i\rangle_H\,\psi(Y)]\|_F^2 &\leq \sum_{i=1}^{\infty} E[\|\langle\phi(X), h_i\rangle_H\,\psi(Y)\|_F^2] && \text{(Jensen's inequality)} \\ &= E[\sum_{i=1}^{\infty}\langle\phi(X), h_i\rangle_H^2\,\|\psi(Y)\|_F^2] = E[\|\phi(X)\|_H^2\,\langle\psi(Y),\psi(Y)\rangle_F] && \text{(Parseval's identity)} \\ &= E[k(X,X)l(Y,Y)] < \infty. \end{aligned}$$

For I_3,

$$\begin{aligned} I_3 &:= 2\sum_{i=1}^{\infty}|\left\langle E[h_i(X)\psi(Y)], E[h_i(X)]\mu_Y\right\rangle_F| \leq 2\sum_{i=1}^{\infty}\|E[h_i(X)\psi(Y)]\|_F\|E[h_i(X)]\mu_Y\|_F \\ &\leq \sum_{i=1}^{\infty}(E\|h_i(X)\psi(Y)\|_F + E\|h_i(X)\mu_Y\|)^2 && (2ab \leq (a+b)^2) \\ &\leq \sum_{i=1}^{\infty}(E[|\langle h_i,\phi(X)\rangle_H|(\|\psi(Y)\|_F + \|\mu_Y\|_F)])^2 \\ &\leq E[\sum_{i=1}^{\infty}\langle h_i,\phi(X)\rangle_H^2(\|\psi(Y)\|_F + \|\mu_Y\|_F)^2] && \text{(Jensen's inequality)} \\ &= E[\|\phi(X)\|_H^2(\|\psi(Y)\|_F + \|\mu_Y\|_F)^2] && \text{(Parseval's identity)} \\ &= E[\|\phi(X)\|_H^2\|\psi(Y)\|_F^2] + E[\|\phi(X)\|_H^2]\|\mu_Y\|_F^2 + 2E[\phi(X)\|_H^2\|\psi(Y)\|_F]\|\mu_Y\|_F \\ &\leq E[k(X,X)l(Y,Y)] + E[k(X,X)]\|\mu_Y\|_F^2 + \underbrace{2(E[\|\phi(X)\|_H^2])^{\frac{1}{2}}(E\|\phi(X)\|_H^2\|\psi(Y)\|_F^2)^{\frac{1}{2}}\|\mu_Y\|_F}_{\text{By Hölder's inequality}} \\ &= E[k(X,X)l(Y,Y)] + E[k(X,X)]\|\mu_Y\|_F^2 + 2\|\mu_Y\|_F\sqrt{E[k(X,X)]E[k(X,X)l(Y,Y)]} < \infty. \end{aligned}$$

Hence, the operator $C_{XY} : H \mapsto F$ is a Hilbert-Schmidt operator in $HS(H,F)$.

□

The proof for the empirical version is the same as the above, where the assumption is replaced by $\frac{1}{n}\sum_{i=1}^{n} k(x_i,x_i)l(y_i,y_i) < \infty$, $\frac{1}{n}\sum_{i=1}^{n} k(x_i,x_i) < \infty$ and $\frac{1}{n}\sum_{i=1}^{n} l(y_i,y_i) < \infty$. By Example 2.3, the operator C_{XY} and $\widehat{C}_{XY}$ are compact.

So, we can apply useful properties of Hilbert-Schmidt operators to analyze the operators C_{XY} and $\widehat{C}_{XY}$. In fact, by Proposition 2.3, replacing $L \in HS(H,F)$ by $f \otimes h$, we have that they satisfy

$$Cov_{XY}(f \otimes h) = \langle C_{XY}, f \otimes h\rangle_{HS} = \langle f, C_{XY}(h)\rangle_F = E_{XY}[(h(X) - E_X[h(X)])(f(Y) - E_Y[f(Y)])] \quad (2.23)$$

$$\widehat{Cov}_{XY}(f \otimes h) = \left\langle\widehat{C}_{XY}, f \otimes h\right\rangle_{HS} = \left\langle f, \widehat{C}_{XY}(h)\right\rangle_H = \frac{1}{n}\sum_{i=1}^{n}[(f(Y_i) - \overline{f(Y)})(h(X_i) - \overline{h(X)})]. \quad (2.24)$$

This makes it possible to define the operators $C_{XY} : H \mapsto F$ and $\widehat{C}_{XY} : H \mapsto F$ that is in consistent with the definition of (empirical) cross-covariance operators.

Definition 2.21 ***(Riesz representative operator of cross-covariance operator)***
The Riesz representative of cross-covariance operator Cov_{XY} in $F \otimes H$ is the linear operator $C_{XY} : H \mapsto F$ such that $\forall h \in H$,

$$C_{XY}(h) := E_{XY}[(\psi(Y) - \mu_Y)(h(X) - E[h(X)])].$$

Definition 2.22 ***(Riesz representative operator of empirical cross-covariance operator)***
The Riesz representative of empirical cross-covariance operator $\widehat{Cov}_{XY}$ in $F \otimes H$ is the linear operator $\widehat{C}_{XY} : H \mapsto F$ such that $\forall h \in H$,

$$\widehat{C}_{XY}(h) := \frac{1}{n}\sum_{i=1}^{n}[(\psi(Y_i) - \widehat{\mu}_Y)(h(X_i) - \overline{h(X)})].$$

Then by reproducing property and linearity of expectation, the right most two equations of (2.23) and (2.24) hold naturally by the definitions above.

To find the explicit forms of C_{XY} and $\widehat{C}_{XY}$ as elements of $HS(H, F)$, we can use (2.9). This gives the following proposition.

Proposition 2.8 ***((Empirical) cross-covariance embedding)***
The explicit forms of C_{XY} and $\widehat{C}_{XY}$ as elements in $HS(H, F)$ are given by

$$C_{XY} = E_{XY}[(\psi(Y) - \mu_Y) \otimes (\phi(X) - \mu_X)] \text{ and } \widehat{C}_{XY} = \frac{1}{n}\sum_{i=1}^{n}[(\psi(Y_i) - \widehat{\mu}_Y) \otimes (\phi(X_i) - \widehat{\mu}_X)].$$

Proof:
For any $f \otimes h \in HS(H, F)$,

$$\begin{aligned}\langle C_{XY}, f \otimes h\rangle_{HS(H,F)} &= E_{XY}[(f(Y) - E_Y[f(Y)])(h(X) - E_X[h(X)])] \\ &= E_{XY}[\langle f, \psi(Y) - \mu_Y\rangle_F \langle h, \phi(X) - \mu_X\rangle_H] \\ &= E_{XY}[\langle f \otimes h, (\psi(Y) - \mu_Y) \otimes (\phi(X) - \mu_X)\rangle_{HS(H,F)}] \\ &= \left\langle E_{XY}[(\psi(Y) - \mu_Y) \otimes (\phi(X) - \mu_X)], f \otimes h\right\rangle_{HS(H,F)}.\end{aligned}$$

Hence, we have the explicit form of C_{XY} in $HS(H, F)$ as

$$C_{XY} = E_{XY}[(\psi(Y) - \mu_Y) \otimes (\phi(X) - \mu_X)]. \tag{2.25}$$

Similarly, the explicit form of $\widehat{C}_{XY}$ in $HS(H, F)$ is given by

$$\widehat{C}_{XY} = \frac{1}{n}\sum_{i=1}^{n}[(\psi(Y_i) - \widehat{\mu}_Y) \otimes (\phi(X_i) - \widehat{\mu}_X)]. \tag{2.26}$$

□

Like the mean embedding, (2.25) and (2.26) are called (empirical) cross-covariance embedding or (empirical) mean embedding of joint distribution P_{XY}.

Proposition 2.9 ***(Adjoint operators of C_{XY} and $\widehat{C}_{XY}$)***
The adjoint operators of C_{XY} and $\widehat{C}_{XY}$ are C_{YX} and $\widehat{C}_{YX}$ respectively. Furthermore, if $H = F$, when X and Y are random variables on the same measurable space with identical distributions, the operators C_{XY} and $\widehat{C}_{XY}$ are self-adjoint.

Proof:

By definition of cross-covariance and its empirical version, we have

$$Cov_{YX}(h \otimes f) = \langle C_{YX}, h \otimes f \rangle_{HS} = \langle h, C_{YX}(f) \rangle_H = E_{YX}[(f(Y) - E_Y[f(Y)])(h(X) - E_X[h(X)])] \quad (2.27)$$

$$\widehat{Cov}_{YX}(h \otimes f) = \langle \widehat{C}_{YX}, h \otimes f \rangle_{HS} = \langle h, \widehat{C}_{YX}(f) \rangle_H = \frac{1}{n}\sum_{i=1}^{n}[(f(Y) - \overline{f(Y)})(h(X_i) - \overline{h(X)})]. \quad (2.28)$$

Noticing that the right most sides of (2.23) and (2.27) are equivalent, we can immediately have

$$\langle f, C_{XY}(h) \rangle_F = \langle h, C_{YX}(f) \rangle_H .$$

This shows that the adjoint operator of $C_{XY} : H \mapsto F$ is $C_{YX} : F \mapsto H$. When $H = F$, $S = D$ and $P_X = P_Y$, the operator C_{XY} is self-adjoint. Similarly, by (2.24) and (2.28), the adjoint operator of $\widehat{C}_{XY} : H \mapsto F$ is $\widehat{C}_{YX} : F \mapsto H$. When $H = F$, $S = D$ and $P_X = P_Y$, the operator $\widehat{C}_{XY}$ is self-adjoint.

□

Moreover, self-adjoint operator C_{XX} and its empirical version $\widehat{C}_{XX}$ are positive semi-definite.

Proposition 2.10 ***(Positive semi-definiteness)***
The self-adjoint operators C_{XX} and $\widehat{C}_{XX}$ are positive semi-definite operators.

Proof:
For any $h \in H$,

$$\langle h, C_{XX}(h) \rangle_H = E_X[(h(X) - E[h(X)])^2] \geq 0$$

$$\langle h, \widehat{C}_{XX}(h) \rangle_H = \frac{1}{n}\sum_{i=1}^{n}(h(X_i) - \overline{h(X)})^2 \geq 0.$$

□

Following the $\sqrt{n}$-consistency of mean embeddings, we can subsequently find the $\sqrt{n}$-consistency of cross-covariance embeddings.

Theorem 2.16 ***($\sqrt{n}$-consistency of cross-covariance embedding)***
If $E_X[k(X,X)]$, $E_Y[k(Y,Y)]$, $E_{XY}[k(x,x)l(y,y)]$, $\frac{1}{n}\sum_{i=1}^{n} k(x_i,x_i)$, $\frac{1}{n}\sum_{i=1}^{n} l(y_i,y_i)$ and $\frac{1}{n}\sum_{i=1}^{n} k(x_i,x_i)l(y_i,y_i)$ are all finite, the cross-covariance operator and corresponding empirical cross-covariance operator admit cross-covariance embeddings. Moreover,

$$||\widehat{C}_{XY} - C_{XY}||_{HS(H,F)} = O_p(n^{-\frac{1}{2}}).$$

Proof:
By Jensen's inequality and concavity of square root, the finiteness of the second moment of a random variable induces the finiteness of its first moment. Hence, the cross-covariance embeddings exist by Lemma 2.3.

Suppose $(h_i)_{i \geq 1}$ is an orthonormal basis of H. Denote the average of h_i under i.i.d sample $x_1, \cdots, x_n$ by

$\overline{h}_i$.

$$
\begin{aligned}
\|\widehat{C}_{XY} - C_{XY}\|^2_{HS(H,F)} &= \sum_{i=1}^{\infty} \|(\widehat{C}_{XY} - C_{XY})(h_i)\|_F^2 \\
&= \sum_{i=1}^{\infty} \|E[(h_i(X) - E[h_i(X)](\psi(Y) - \mu_Y)] - \frac{1}{n}\sum_{j=1}^{n}(h_i(x_j) - \overline{h}_i)(\psi(y_j) - \widehat{\mu}_Y)]\|_F^2 \\
&= \sum_{i=1}^{\infty} \|E[h_i(X)\psi(Y)] - E[h_i(X)]\mu_Y - \frac{1}{n}\sum_{j=1}^{n} h_i(x_j)\psi(y_j) + \overline{h}_i\widehat{\mu}_Y\|_F^2 \\
&= \sum_{i=1}^{\infty} \|E[h_i(X)\psi(Y)] - E[h_i(X)]\mu_Y - \overline{h}_i\mu_Y + \overline{h}_i\mu_Y - \frac{1}{n}\sum_{j=1}^{n} h_i(x_j)\psi(y_j) + \overline{h}_i\widehat{\mu}_Y\|_F^2 \\
&\le \underbrace{\sum_{i=1}^{\infty} \|\frac{1}{n}\sum_{j=1}^{n} h_i(x_j)\psi(y_j) - E[h_i(X)\psi(Y)]\|_F^2}_{L_1} + \underbrace{\sum_{i=1}^{\infty} \|\overline{h}_i\mu_Y - E[h_i(X)]\mu_Y\|_F^2}_{L_2} \\
&\qquad + \underbrace{\sum_{i=1}^{\infty} \|\overline{h}_i\widehat{\mu}_Y - \overline{h}_i\mu_Y\|_F^2}_{L_3} + L_4,
\end{aligned}
$$

where L_4 is given by

$$
\begin{aligned}
L_4 := 2\sum_{i=1}^{\infty}\left|\left\langle \frac{1}{n}\sum_{j=1}^{n} h_i(x_j)\psi(y_j) - E[h_i(X)\psi(Y)], \overline{h}_i\mu_Y - E[h_i(X)]\mu_Y \right\rangle_F\right| + 2\sum_{i=1}^{\infty}\left|\left\langle \overline{h}_i\mu_Y - E[h_i(X)]\mu_Y, \overline{h}_i\widehat{\mu}_Y - \overline{h}_i\mu_Y \right\rangle_F\right| \\
+ 2\sum_{i=1}^{\infty}\left|\left\langle \frac{1}{n}\sum_{j=1}^{n} h_i(x_j)\psi(y_j) - E[h_i(X)\psi(Y)], \overline{h}_i\widehat{\mu}_Y - \overline{h}_i\mu_Y \right\rangle_F\right|.
\end{aligned}
$$

For L_2,

$$
\begin{aligned}
L_2 := \sum_{i=1}^{\infty}\|\overline{h}_i\mu_Y - E[h_i(X)]\mu_Y\|_F^2 &= \|\mu_Y\|_F^2\sum_{i=1}^{\infty}\left|\frac{1}{n}\sum_{j=1}^{n}\langle\phi(x_j), h_i\rangle_H - E[\langle\phi(X), h_i\rangle_H]\right|^2 \\
&= \|\mu_Y\|_F^2\sum_{i=1}^{\infty}\left|\left\langle\frac{1}{n}\sum_{j=1}^{n}\phi(x_j) - E[\phi(X)], h_i\right\rangle_H\right|^2 = \|\mu_Y\|_F^2\|\widehat{\mu}_X - \mu_X\|_H^2. \qquad \text{(Parseval's identity)}
\end{aligned}
$$

For L_3, also by Parseval's identity,

$$
L_3 := \sum_{i=1}^{\infty}\left|\left\langle\frac{1}{n}\sum_{j=1}^{n}\phi(x_j), h_i\right\rangle_H\right|^2\|\widehat{\mu}_Y - \mu_Y\|_F^2 = \|\widehat{\mu}_X\|_H^2\|\widehat{\mu}_Y - \mu_Y\|_F^2.
$$

By Theorem 2.15, the $\sqrt{n}$-consistency of mean embedding, we know L_2 and L_3 are $O_p(n^{-1})$ as n goes to infinity.

For L_1,

$$
\begin{aligned}
E[L_1] :&= \sum_{i=1}^{\infty} E_{x_1,\cdots,x_n,y_1,\cdots,y_n}\left\|\frac{1}{n}\sum_{j=1}^{n} h_i(x_j)\psi(y_j) - E[h_i(X)\psi(Y)]\right\|_F^2 \\
&= \sum_{i=1}^{\infty} \frac{1}{n^2} E_{x_1,\cdots,x_n,y_1,\cdots,y_n}\left\|\sum_{j=1}^{n} \left[h_i(x_j)\psi(y_j) - E[h_i(X)\psi(Y)]\right]\right\|_F^2 \\
&= \sum_{i=1}^{\infty} \frac{1}{n^2} E_{x_1,\cdots,x_n,y_1,\cdots,y_n}\sum_{j=1}^{n}\left\|h_i(x_j)\psi(y_j) - E[h_i(X)\psi(Y)]\right\|_F^2 \\
&\quad + \sum_{i=1}^{\infty} \frac{1}{n^2} \underbrace{\sum_{j\neq k}^{n} E_{x_1,\cdots,x_n,y_1,\cdots,y_n}\left\langle h_i(x_j)\psi(y_j) - E[h_i(X)\psi(Y)], h_i(x_k)\psi(y_k) - E[h_i(X)\psi(Y)]\right\rangle_F}_{\text{Vanishes by } (x_i,y_i) \text{ i.i.d}} \\
&= \sum_{i=1}^{\infty} \frac{1}{n} E_{XY}\left\|h_i(X)\psi(Y) - E[h_i(X)\psi(Y)]\right\|_F^2 \qquad ((x_i,y_i) \text{ i.i.d}) \\
&= \frac{1}{n} E_{XY}\sum_{i=1}^{\infty}\left\|(\psi(Y) - \mu_Y)\otimes(\phi(X) - \mu_X)(h_i)\right\|_F^2 \\
&= \frac{1}{n} E_{XY}\|(\psi(Y) - \mu_Y)\otimes(\phi(X) - \mu_X)\|_{HS(H,F)}^2 \\
&= \frac{1}{n} E_{XY}\left[\|\psi(Y) - \mu_Y\|_F^2\|\phi(X) - \mu_X\|_H^2\right] \\
&= \frac{1}{n}\underbrace{E_{XY}\left[\left(l(Y,Y) - 2\mu_Y + \|\mu_Y\|_F^2\right)\left(k(X,X) - 2\mu_X + \|\mu_X\|_H^2\right)\right]}_{\text{Finite by assumption}}.
\end{aligned}
$$

This means $L_1 = O_p(n^{-1})$.

At last, for L_4, notice that the three sums of the absolute value of inner products have the same structure. This means we can use the same trick to bound the three terms. In fact, for the first sum,

$$
\begin{aligned}
&2\sum_{i=1}^{\infty}\left|\left\langle\frac{1}{n}\sum_{j=1}^{n} h_i(x_j)\psi(y_j) - E[h_i(X)\psi(Y)], \overline{h}_i\mu_Y - E[h_i(X)]\mu_Y\right\rangle_F\right| \\
\leq& 2\sum_{i=1}^{\infty}\|\frac{1}{n}\sum_{j=1}^{n} h_i(x_j)\psi(y_j) - E[h_i(X)\psi(Y)]\|_F\|\overline{h}_i\mu_Y - E[h_i(X)]\mu_Y\|_F \\
\leq& \sum_{i=1}^{\infty}\|\frac{1}{n}\sum_{j=1}^{n} h_i(x_j)\psi(y_j) - E[h_i(X)\psi(Y)]\|_F^2 + \sum_{i=1}^{\infty}\|\overline{h}_i\mu_Y - E[h_i(X)]\mu_Y\|_F^2, \qquad (2ab \leq a^2 + b^2)
\end{aligned}
$$

which is exactly L_1 + L_2. Hence, following the same trick and combining the convergence rates of L_1, L_2 and L_3, we can have that the convergence rate of L_4 is $O_p(n^{-1})$. Combining the convergence rates of L_1, L_2, L_3 and L_4, we have

$$\|\widehat{C}_{XY} - C_{XY}\|_{HS(H,F)} = O_p(n^{-\frac{1}{2}}).$$

□

Remark about the cross-covariance without centering When applying cross-covariance operators in kernel embeddings in RKHS, people often ignore the centering of the feature maps [35, 15]. We define the cross-covariance operators without centering $\widetilde{Cov}_{XY}$ and its empirical version $\widetilde{\widetilde{Cov}}_{XY}$ by removing the terms $E_Y[f(Y)]$, $E_X[h(X)]$ in (2.21) and $\overline{f(Y)}$, $\overline{h(X)}$ in (2.22). In other words, $\forall f \in F$, $h \in H$, and $(X_i, Y_i)_{1 \le i \le n}$ i.i.d. from P_{XY}, we define the cross-covariance operator and its empirical version by linear operators from $F \otimes H \mapsto \mathbb{K}$ such that

$$\widetilde{Cov}_{XY}(f \otimes h) := E_{XY}[f(Y)h(X)]$$

$$\widetilde{\widehat{Cov}}_{XY} := \frac{1}{n}\sum_{i=1}^{n} f(Y_i)h(X_i).$$

Their Hilbert-Schmidt Riesz representatives in $HS(H, F)$ of cross-covariance operators without centering are given by

$$\tilde{C}_{XY}(h) := E_{XY}[h(X)\psi(Y)] \tag{2.29}$$

$$\tilde{\hat{C}}_{XY}(h) := \frac{1}{n}\sum_{i=1}^{n} h(X_i)\psi(Y_i). \tag{2.30}$$

And their explicit forms in $HS(H, F)$ as well as the cross-covariance embeddings without centering are

$$\tilde{C}_{XY} := E_{XY}[\psi(Y) \otimes \phi(X)]$$

$$\tilde{\hat{C}}_{XY} := \frac{1}{n}\sum_{i=1}^{n} \psi(Y_i) \otimes \phi(X_i).$$

These operators share the same properties as the ones with centering because the proofs given previously for the theorems and propositions about the centered operators have already implied this. For example, in the proof of Theorem 2.16, the rate of convergence of L_1 is just the rate of convergence of $\tilde{\hat{C}}_{XY}$ to $\tilde{C}_{XY}$ in squared Hilbert-Schmidt norm, which is $\mathcal{O}_p(n^{-1})$ as same as the centered case.

2.5. Quadratic mean differentiability

This section aims at introducing the quadratic mean differentiability which is important in the existence for score functions and Fisher information of parametric models. QMD is also used to explain the construction of paths for semiparametric models in Chapter 3. The following definition and theorem are based on the Theorem 7.2 in Chapter 7 of [41].

Definition 2.23 ***(Quadratic mean differentiability (QMD))***
A parametric model $\mathcal{P} = \{P_\theta, \theta \in \Theta \subseteq \mathbb{R}\}$ possessing density p_θ with respect to measure μ is differentiable in quadratic mean at θ_0 if there exists a measurable function $g : \chi \mapsto \mathbb{R}$ such that as $\theta \to \theta_0$,

$$\int \left[\sqrt{p_\theta} - \sqrt{p_{\theta_0}} - \frac{1}{2}(\theta - \theta_0)g\sqrt{p_{\theta_0}}\right]^2 d\mu = o(|\theta - \theta_0|^2). \tag{2.31}$$

For parametric models, quadratic mean differentiability (QMD) is a weaker condition for first order differentiability of p (or $\sqrt{p}$) which is a part of Cramér-Rao regularity condition

$$\sqrt{p_\theta} - \sqrt{p_{\theta_0}} - \frac{1}{2}(\theta - \theta_0)g\sqrt{p_\theta} = o(|\theta - \theta_0|^2), \text{ as } \theta \to \theta_0.$$

However it is enough for the existence of both score functions and Fisher information.

Theorem 2.17 *For a QMD parametric model $\mathcal{P} = \{P_\theta, \theta \in \Theta \subseteq \mathbb{R}\}$, $E[g] = 0$ and $\mathcal{I}(\theta) = E[g^2]$ exists.*

Proof:

We denote $\sqrt{p_\theta}-\sqrt{p_{\theta_0}}-\frac{1}{2}(\theta-\theta_0)g\sqrt{p_{\theta_0}}$ by $f(\theta;\theta_0)$. Then Equation (2.31) is given by $\|f(\theta;\theta_0)\|^2_{L_2(\mu)} = o(|\theta-\theta_0|^2)$.

$$
\begin{aligned}
&\|\sqrt{p_\theta}-\sqrt{p_{\theta_0}}\|^2_{L_2(\mu)}\\
=&\|f(\theta;\theta_0)+\frac{1}{2}(\theta-\theta_0)g\sqrt{p_{\theta_0}}\|^2_{L_2(\mu)}\\
=&\|f(\theta;\theta_0)\|^2_{L_2(\mu)}+\|\frac{1}{2}(\theta-\theta_0)g\sqrt{p_{\theta_0}}\|^2_{L_2(\mu)}+\langle f(\theta;\theta_0),\frac{1}{2}(\theta-\theta_0)g\sqrt{p_{\theta_0}}\rangle_{L_2(\mu)}\\
\leq&\|f(\theta;\theta_0)\|^2_{L_2(\mu)}+\|\frac{1}{2}(\theta-\theta_0)g\sqrt{p_{\theta_0}}\|^2_{L_2(\mu)}+\|f(\theta;\theta_0)\|_{L_2(\mu)}\|\frac{1}{2}(\theta-\theta_0)g\sqrt{p_{\theta_0}}\|_{L_2(\mu)} \quad \text{(Cauchy-Schwartz)}\\
=&o(|\theta-\theta_0|^2)+O(|\theta-\theta_0|^2)+o(|\theta-\theta_0|^2)\\
=&O(|\theta-\theta_0|^2).
\end{aligned}
$$

Hence, by Equation (2.31), as $\theta\to\theta_0$, we have two sequences converging in $L_2(\mu)$ norm, which are $\sqrt{p_\theta}\to\sqrt{p_{\theta_0}}$ and $\frac{\sqrt{p_\theta}-\sqrt{p_{\theta_0}}}{\theta-\theta_0}\to\frac{1}{2}g\sqrt{p_{\theta_0}}$. After rearranging $E[g]$ to an inner product of the two convergent sequences, we can find it's just a zero. In fact,

$$
\begin{aligned}
E[g] &= \int g p_{\theta_0} d\mu\\
&= \int \left(\frac{1}{2}g\sqrt{p_\theta}\right)\left(2\sqrt{p_\theta}\right) d\mu\\
&= \lim_{\theta\to\theta_0}\int \frac{\left(\sqrt{p_\theta}-\sqrt{p_{\theta_0}}\right)}{\theta-\theta_0}\left(\sqrt{p_\theta}+\sqrt{p_{\theta_0}}\right) d\mu && \text{(continuity of inner product)}\\
&= \lim_{\theta\to\theta_0}\frac{1}{\theta-\theta_0}\int p_\theta - p_{\theta_0} d\mu\\
&= \lim_{\theta\to\theta_0}\frac{1}{\theta-\theta_0}(1-1) && (p_{\theta_0}\text{ and }p_\theta\text{ are all densities})\\
&= 0.
\end{aligned}
$$

Define random variable $W_\theta = 2\frac{1}{(\theta-\theta_0)^2}\left(\sqrt{\frac{p_\theta}{p_{\theta_0}}}(X)-1\right)$. Since p_{θ_0} is a density, we know $Pr\left\{p_{\theta_0}=0\right\}=0$. This means W_θ is well defined with probability 1. The expectation of W_θ exists by Cauchy-Schwartz that $\left\langle\sqrt{p_\theta},\sqrt{p_{\theta_0}}\right\rangle \leq \|\sqrt{p_\theta}\|^2_{L_2(\mu)}\|\sqrt{p_{\theta_0}}\|^2_{L_2(\mu)} = 1$.

As $\theta\to\theta_0$,

$$
\begin{aligned}
E[W_\theta] &= \int 2\frac{1}{(\theta-\theta_0)^2}\left(\sqrt{\frac{p_\theta}{p_{\theta_0}}}-1\right)p_{\theta_0}d\mu\\
&= 2\frac{1}{(\theta-\theta_0)^2}\left(\int \sqrt{p_\theta}\sqrt{p_{\theta_0}}d\mu - 1\right)\\
&= -\frac{1}{(\theta-\theta_0)^2}\int\left(\sqrt{p_\theta}-\sqrt{p_{\theta_0}}\right)^2 d\mu\\
&\to -\int\frac{1}{4}g^2 p_{\theta_0}d\mu\\
&= -\frac{1}{4}E[g^2].
\end{aligned}
$$

Hence $E[g^2]$ exists as $-4E[W_\theta]$.

□

2.6. Expected risk minimization

Expected risk minimization (ERM) is a fundamental concept in statistical learning theory. It refers to the process of choosing a predictive model that minimizes the expected loss (or risk) over the distribution of all possible data. In this section, we introduce the ERM problem and derive its minimizer under a square loss function.

Suppose $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is the input-output pair and $P(X, Y)$ is the unknown true data distribution. The hypothesis class $\mathcal{H}$ contains all the model $\mathcal{X} \mapsto \mathcal{Y}$ that can be used in the learning algorithm. After determining the loss function $l(\cdot, \cdot)$, the expected risk under the model $f \in \mathcal{H}$ is the average loss over the true data distribution which is given by

$$R(f) = E_{X,Y}[l(Y, f(X))]. \tag{2.32}$$

The ERM problem is defined by

$$min_{f \in \mathcal{H}} R(f) = E_{X,Y}[l(Y, f(X))]. \tag{2.33}$$

The following theorem determines the minimizer for ERM problem (2.33) with square loss.

Theorem 2.18 ***(Expected risk minimization)***

The solution to the ERM problem (2.33) with square loss $l(y, y') = \frac{1}{2}|y - y'|^2$, $\forall y, y' \in \mathcal{Y}$ is $f^*(X) := E[Y|X]$.

Proof:

Since the goal of ERM is to find the optimal model $f^* \in \mathcal{H}$ that minimizes the expected risk $R(f)$ (2.32) for every input x, noticing that by the tower property of conditional expectation,

$$\begin{aligned} R(f) &= E_{X,Y}[l(Y, f(X))] \\ &= E_X E_{X,Y}[l(Y, f(X))|X], \end{aligned}$$

we know that in order to minimize $R(f)$, we need to find the minimizer for

$$E_{X,Y}[l(Y, f(X))|X]. \tag{2.34}$$

With square loss, Equation (2.34) is

$$\begin{aligned} E_{X,Y}[l(Y, f(X))|X] &= \frac{1}{2} E_{X,Y}[(Y - f(X))^2|X] \\ &= \frac{1}{2} E_{X,Y}[(Y - E[Y|X] + E[Y|X] - f(X))^2|X] \\ &= \frac{1}{2}\left\{E_{X,Y}[(Y - E[Y|X])^2|X] + E_{X,Y}[(Y - E[Y|X])(E[Y|X] - f(X))|X] + E_{X,Y}[(E[Y|X] - f(X))^2|X]\right\} \\ &= \frac{1}{2} E_{X,Y}[(Y - E[Y|X])^2|X] + \frac{1}{2}(E[Y|X] - f(X)) \underbrace{E_{X,Y}[(Y - E[Y|X])|X]}_{0} + \frac{1}{2} E_{X,Y}[(E[Y|X] - f(X))^2|X]. \end{aligned} \tag{2.35}$$

We find the minimizer for $E_{X,Y}[(E[Y|X] - f(X))^2|X]$ because it is the only term relevant to f in Equation (2.35). Since the square function is convex with respect to f, there exists a minimizer. At the minimum point $f = f^*$, by the first-order condition

$$\begin{aligned} \left.\frac{\partial}{\partial f(X)}\right|_{f=f^*} \frac{1}{2} E_{X,Y}[(E[Y|X] - f(X))^2|X] &= E_{X,Y}[f^*(X) - E[Y|X]|X] \\ &= f^*(X) - E[Y|X] \\ &= 0, \end{aligned}$$

we have that $f^*(X) = E[Y|X]$ is a minimizer for the expected risk $R(f)$. Choosing $f(X) := E[Y|X]$ is a necessary condition for the expected risk $R(f)$ (2.32) to be the minimum because applying the first-order necessary condition directly to $R(f)$ with square loss also produce a set of minimizers $\left\{f : f = \text{argmin}_{f \in \mathcal{H}} R(f)\right\}$ such that any model f within the set satisfies $E[f(X)] = E[Y]$. However, the purpose for ERM requires a pointwise minimizer for every input x but not the minimizer on the average scale. Hence $f^*(X) := E[Y|X]$ is the only solution.

□

2.7. Results from convex analysis

This section is introduced to give a short overview of some results from convex analysis. The critical parts include the Fenchel duality and the interchange for minimization and integration, which are used in Chapter 4. The reference includes chapter 11 and 14 of [28] and chapter 7 of [30].

Denote the set of expanded real numbers by $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\} \cup \{+\infty\}$. Let $(\Omega, \mathcal{A})$ be a measurable space and $\langle \cdot, \cdot \rangle$ be the inner product on $\mathbb{R}^n$.

2.7.1. Basic definitions

In this subsection, basic definitions about properness, subdifferential, semicontinuity, convexity and concavity are given.

Definition 2.24 ***(Proper functions)***

A function $f : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ is proper if its domain

$$domf := \{x \in \mathbb{R}^n : f(x) < +\infty\}$$

is nonempty and $f(x) > -\infty$, $\forall x \in \mathbb{R}^n$.

Proper functions can be seen as functions which are always larger than $-\infty$ and with at least one point such that the value is finite. For improper functions, we have $f(x) = +\infty, \forall x \in \mathbb{R}^n$ by its empty domain and $f(x) = -\infty, \forall x \in \mathbb{R}^n$ since there is no point in $\mathbb{R}^n$ to make it larger than $-\infty$.

Definition 2.25 ***(Subdifferential)***

The subdifferential of a function $f : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ at $x_0 \in \mathbb{R}^n$ is the set of all subgradients of f at x_0, denoted by $\partial f(x_0)$, where the subgradients of f at x_0 is a vector $z \in \mathbb{R}^n$ such that

$$f(x) - f(x_0) \geq \langle z, x - x_0 \rangle, \ \forall x \in \mathbb{R}^n. \tag{2.36}$$

The corresponding subdifferential mapping ∂f is set-valued map whose graph $gph\partial f$ is defined by

$$gph\partial f = \{(x, g) \in \mathbb{R}^n \times \mathbb{R}^n : g \in \partial f(x)\}.$$

The subdifferential is a closed convex subset of $\mathbb{R}^n$. If a function f is differentiable at x_0, then its subdifferential at x_0 is just $\{\nabla f(x_0)\}$. If it is nondifferentiable at x_0, then $\partial f(x_0)$ contains multiple subgradients. Moreover, if f is subdifferential at x_0, then f is proper by Equation (2.36).

Example 2.16 ***(Subdifferential of the absolute value)***

The subdifferential $\partial f(x)$ of the function $f(x) = |x|$ defined on $\mathbb{R}$ is given by $\partial f(x) = \begin{cases} \{-1\}, x < 0, \\ \{1\}, x > 0, \\ [-1, 1], x = 0. \end{cases}$

The graph of the subdifferential mapping is

$$gph\partial f = \{(x, -1) : x < 0\} \cup \{(x, 1) : x > 0\} \cup \{(0, g) : g \in [-1, 1]\}.$$

Definition 2.26 ***(Semicontinuity)***

At a point $x_0 \in \mathbb{R}^n$, a function $f : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ is

- *lower semicontinuous, if $f(x_0) \leq \liminf_{x \to x_0} f(x)$;*
- *upper semicontinuous, if $f(x_0) \geq \limsup_{x \to x_0} f(x)$.*

If f is lower (upper) semicontinuous at every point in $\mathbb{R}^n$ then it is a lower (upper) semicontinuous function.

From the definition of semicontinuity, we know any continuous function is both lower semicontinuous and upper semicontinuous everywhere.

Below we give an example for semicontinuous functions to better illustrate the semicontinuity.

Example 2.17 ***(Semicontinuous functions)***

Function $f(x) = \begin{cases} -1, & x \leq 0 \\ 1, & x > 0 \end{cases}$ is lower semicontinuous at $x = 0$ because $\forall \epsilon > 0$, $\exists \delta = \epsilon$ such that $\forall |x| < \delta$, $f(x) - f(0) > -\epsilon$, which means $f(0) \leq \liminf_{x \to 0} f(x)$. Similarly, function $f(x) = \begin{cases} -1, & x < 0 \\ 1, & x \geq 0 \end{cases}$ is upper semicontinuous at $x = 0$ because $\forall \epsilon > 0$, $\exists \delta = \epsilon$ such that $\forall |x| < \delta$, $f(x) - f(0) < \epsilon$.

Definition 2.27 ***(Convexity, concavity)***
A function $f : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ is

- *convex, if $\forall x_1, x_2 \in \mathbb{R}^n$, $\forall \lambda \in (0, 1)$, $\lambda f(x_1) + (1 - \lambda) f(x_2) \geq f(\lambda x_1 + (1 - \lambda) x_2)$;*
- *concave, if $\forall x_1, x_2 \in \mathbb{R}^n$, $\forall \lambda \in (0, 1)$, $\lambda f(x_1) + (1 - \lambda) f(x_2) \leq f(\lambda x_1 + (1 - \lambda) x_2)$.*

When the equality doesn't hold anywhere on $\mathbb{R}^n$, the convexity and concavity are strict.

Example 2.18 ***(Jensen's inequality)***
Suppose X is a random variable. Given a function $\phi : \mathbb{R} \mapsto \mathbb{R}$ differentiable at point $x = E[X]$, we have

- *$\phi(E[X]) \leq E[\phi(X)]$, if ϕ is convex;*
- *$\phi(E[X]) \geq E[\phi(X)]$, if ϕ is concave.*

Proof:
Only the convex case is shown here because the proof of the two cases are similar. The convexity and differentiability at point $x = E[X]$ implies that $\forall \lambda \in [0, 1]$, at $x_0 \in S$,

$$\begin{aligned}
\lambda \phi(x) + (1 - \lambda)\phi(E[X]) &\geq \phi(\lambda x + (1 - \lambda)E[X]) \\
\phi(x) - \phi(E[X]) &\geq \frac{1}{\lambda}(\phi(\lambda x + (1 - \lambda)E[X]) - \phi(E[X])) \\
\phi(x) - \phi(E[X]) &\geq \frac{1}{\lambda} \frac{\phi(\lambda x - \lambda E[X] + E[X]) - \phi(E[X])}{\lambda x - \lambda E[X]}(\lambda x - \lambda E[X]) \\
\phi(x) - \phi(E[X]) &\geq \frac{\phi(\lambda(x - E[X]) + E[X]) - \phi(E[X])}{\lambda(x - E[X])}(x - E[X]).
\end{aligned}$$

As $\lambda \downarrow 0$, we have

$$\begin{aligned}
\phi(x) - \phi(E[X]) &\geq \lim_{\lambda \downarrow 0} \frac{\phi(\lambda(x - E[X]) + E[X]) - \phi(E[X])}{\lambda(x - E[X])}(x - E[X]) \\
\phi(x) - \phi(E[X]) &\geq \phi'(E[X])(x - E[X]) \\
\phi(E[X]) &\leq \phi(x) - \phi'(E[X])(x - E[X]). \qquad (2.37)
\end{aligned}$$

Next is to compute the expectations relative to X on the both sides of (2.37).

$$\begin{aligned}
\phi(E[X]) &\leq E[\phi(X)] - \phi'(E[X])(E[X] - E[X]) \\
\phi(E[X]) &\leq E[\phi(X)].
\end{aligned}$$

□

2.7.2. Fenchel duality

This subsection introduces the Fenchel duality that is derived from the Fenchel-Moreau theorem and the definition of conjugate functions.

Definition 2.28 ***(Conjugate function)***
Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a real valued function. The conjugate function of f is given by

$$f^*(z) := \sup_{x \in \mathbb{R}^n} \{\langle z, x \rangle - f(x)\}. \qquad (2.38)$$

Example 2.19 ***(Conjugate of the square loss)***
The square loss function $l : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ is given by $l(x,y) = c(x-y)^2$, where $c \in (0,1]$. Denote $l(x,\cdot)$ by $l_x(\cdot)$ for a fixed x. The conjugate function $l_x^*(y) = xy + \frac{y^2}{4c}$.

Proof:
By definition 2.28, the conjugate function of l_x is

$$l_x^*(y) := \max_{t\in\mathbb{R}} \{yt - l_x(t)\}.$$

Since $yt - l_x(t) = -ct^2 + (2cx + y)t - cx^2$ is a convex parabola with respect to t, it has a global maximum point $t = x + \frac{y}{2c}$ and a corresponding global maximum $xy + \frac{y^2}{4c}$. This gives the conjugate function l_x^*.

□

The following theorem is important for defining the Fenchel duality.

Theorem 2.19 ***(Fenchel-Moreau)***
Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a proper and convex function, then $f^{**} = lscf$, where $lscf$ is the largest lower semicontinuous function less or equal to f.

For a function $f : \mathbb{R}^n \mapsto \mathbb{R}$, from Equation (2.38), it's clear that

$$f^*(z) \geq \langle z, x\rangle - f(x),$$

which also gives

$$f(x) \geq \langle x, z\rangle - f^*(z).$$

The last inequality implies $f \geq f^{**}$. The following theorem gives the condition for the equality to hold.

Theorem 2.20 ***(Fenchel duality)***
Given a function $f : \mathbb{R}^n \mapsto \mathbb{R}$, $f = f^{**}$ holds if f is proper, lower semicontinuous and convex. (f, f^*) is called the Fenchel duality in which f and f^* are dual to each other.

The proof of the theorem is a direct result of Fenchel-Moreau theorem 2.19 by adding the assumption that f is lower semicontinuous.

Example 2.20 ***(Fenchel duality of the square loss)***
Here we continue the discussion about the square loss function $l_x(\cdot)$ introduced in Example 2.19. Since $l_x(\cdot)$ is continuous, it is a lower semicontinuous function. It is also clear that it is proper and convex. Hence $l_x(y) = l_x^{**}(y)$ everywhere on $\mathbb{R}$ by Theorem 2.20, which gives a Fenchel duality (l_x, l_x^*).

For a Fenchel duality (f, f^*), we have the following result.

Proposition 2.11 ***(Conjugate extreme point)***
Given a Fenchel duality (f, f^*), $z^* \in \partial f(x^*)$ if and only if $x^* \in \partial f^*(z^*)$.

Proof:
By the definition of conjugate function, at the points z^* and x^*,

$$\begin{aligned} f^*(z^*) &= \sup_{x\in\mathbb{R}^n} \{\langle x, z^*\rangle - f(x)\}, \\ f(x^*) &= \sup_{z\in\mathbb{R}^n} \{\langle z, x^*\rangle - f^*(z)\}. \qquad (f = f^{**}) \end{aligned}$$

The functions considered should satisfy the first order condition at extreme points x^* and z^*. This gives

$$\frac{\partial}{\partial x}\Big|_{x=x^*}(\langle x, z^*\rangle - f(x)) = 0 \Leftrightarrow z^* = \nabla f(x^*),$$
$$\frac{\partial}{\partial z}\Big|_{z=z^*}(\langle z, x^*\rangle - f^*(z)) = 0 \Leftrightarrow x^* = \nabla f^*(z^*).$$

From the definition of subdifferential 2.25, it is clear that $x^* \in \partial f^*(z^*)$ and $z^* \in \partial f(x^*)$ at the same time.

□

An application of this proposition is that given a Fenchel duality (f, f^*), if we have some point $x \in \partial f(u)$, then immediately $u \in \partial f^*(x)$.

2.7.3. Interchange of minimization and integration

The interchange of minimization and integral is an useful result in extremal problems related to integrals. In this subsection, important ideas including normal integrands and decomposable spaces are introduced for the final interchangeability theorem. Moreover, Proposition 2.12 (subgradient characterization of convex normality) is given to provide a method to identify the normality of any proper, lower semicontinuous and convex function.

Definition 2.29 ***(Epigraph)***

The epigraph of a function $f : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ is the set of points lying on or above the graph of f:

$$epif := \{(x, \alpha) \in \mathbb{R}^{n+1} : f(x) \le \alpha\}.$$

For a bivariate function $f : \Omega \times \mathbb{R}^n \mapsto \overline{\mathbb{R}}$, its epigraphical mapping S_f is defined by

$$S_f(x) := \text{epi}f(x, \cdot) = \{(\cdot, \alpha) \in \mathbb{R}^{n+1} : f(x, \cdot) \le \alpha\}.$$

The closedness of epigraph is related to the lower semicontinuity. In fact, function $f : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ is lower continuous if and only if its epigraph is a closed subset of $\mathbb{R}^n$.

Definition 2.30 ***(Normal integrand)***

A function $f : \Omega \times \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ is a normal integrand if its epigraphical mapping $S_f : \Omega \mapsto \mathbb{R}^n \times \overline{\mathbb{R}}$ is closed-valued[4] *and measurable.*

Proposition 2.12 ***(Subgradient characterization of convex normality)***

Let $f : \Omega \times \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ be such that $f(x, \cdot)$ is a proper, lower semicontinuous and convex function with respect to $\cdot \in \mathbb{R}^n$ for any $x \in \Omega$. Then f is a normal integrand if and only if the following hold:

1. *the mapping $x \mapsto gph\partial f(x, \cdot)$ is measurable;*
2. *there is a measurable function $\overline{u} : \Omega \mapsto \mathbb{R}^n$ such that $\partial f(x, \overline{u}(x)) \neq \emptyset$ for all $x \in \Omega$ and the function $x \mapsto f(x, \overline{u}(x))$ is measurable.*

Theorem 2.21 ***(Conjugate integrands)***

Given a normal integrand $f : \Omega \times \mathbb{R}^n \mapsto \overline{\mathbb{R}}$, the conjugate $f^*(\cdot) := f^*(\omega, \cdot)$ and biconjugate $f^{**}(\cdot) : f^{**}(\omega, \cdot)$ are normal integrands.

Example 2.21 ***(Normality of the square loss)***

The square loss function $l_x : \mathbb{R} \mapsto \mathbb{R}$ and its conjugate function l_x^* introduced in Example 2.19 and 2.20 are normal integrands.

[4] A set-valued mapping with a closed image.

Proof:
Since $l_x(y)$ is differentiable everywhere on $\mathbb{R}$, the subdifferential is $\partial l_x(y) = \frac{\partial}{\partial y} l_x(y) = 2(y - x)$. The graph of the subdifferential mapping is

$$\text{gph}\partial l_x = \{(y, 2y - 2x) : y \in \mathbb{R}\}.$$

The set-valued mapping $x \mapsto \text{gph}\partial l_x$ corresponds to a continuous function $g_y(x) := 2y - 2x$ which is measurable.
As for the second condition, we choose $\overline{u} : \mathbb{R} \mapsto \mathbb{R}$ by $\overline{u}(x) = x$. Then $\forall x \in \mathbb{R}$, $\partial l_x(\overline{u}(x)) = \{0\} \neq \emptyset$ because $l_x(\overline{u}(x)) = 0$. The function $x \mapsto l_x(\overline{u}(x)) = 0$ is a constant function and thus is measurable.
By Theorem 2.21, the conjugate l_x^* is also a normal integrand.

□

Definition 2.31 ***(Decomposable space)***
Given a measurable space $(\Omega, \mathcal{A})$, a space $\mathcal{U}$ of measurable functions $u : \Omega \mapsto \mathbb{R}^n$ is decomposable relative to a measure μ on $\mathcal{A}$ if any $u_0 \in \mathcal{U}$ and any $A \in \mathcal{A}$ with $\mu(A) < \infty$, any bounded and measurable $u_1 \in \mathcal{U}$, function $u : u(x) = \begin{cases} u_0(x), x \in \Omega - A, \\ u_1(x), x \in A, \end{cases}$ is in $\mathcal{U}$.

The definition means one can "glue" functions in $\mathcal{U}$ together on measurable sets and still stay inside $\mathcal{U}$. The decomposable spaces should always be linear spaces containing all bounded measurable functions that vanish outside some sets with finite measures. Examples of the decomposable spaces include spaces of all measurable functions from Ω to $\mathbb{R}^n$ and spaces of all equivalent classes of functions from Ω to $\mathbb{R}^n$ with finite L_p norm with respect to measure μ.

Theorem 2.22 ***(Interchange of minimization and integration)***
Let X be a random variable on the measurable space $(\Omega, \mathcal{A})$ with a σ-finite measure μ on $\mathcal{A}$. Let $\mathcal{U}$ be a space of measurable functions from Ω to $\mathbb{R}^n$ that is decomposable relative to μ. Let $f : \Omega \times \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ be a normal integrand. Then the minimization of integral of f over $\mathcal{U}$ can be reduced to pointwise minimization if the integral is finite, i.e.

$$\inf_{u(\cdot)\in\mathcal{U}} \int_\Omega f(x, u(x)) d\mu(x) = \int_\Omega \inf_{u\in\mathbb{R}^n} f(x, u) d\mu(x).$$

The interchange of minimization and integration guarantees that if the common infimum are finite and reachable, the optimal function $u^*(\cdot) \in \arg\min_{u(\cdot)\in\mathcal{U}} \int_\Omega f(x, u(x)) d\mu(x)$ satisfies

$$u^*(x) = \arg\min_{u\in\mathbb{R}^n} f(x, u),$$

almost everywhere on Ω by μ. In fact, if $\mu(\{x \in \Omega : f(x, u^*(x)) > \inf_{u\in\mathbb{R}^n} f(x, u)\}) > 0$, we have

$$\begin{aligned} \int_\Omega f(x, u^*(x)) d\mu(x) &> \int_\Omega \inf_{u\in\mathbb{R}^n} f(x, u) d\mu(x) \\ &= \inf_{u(\cdot)\in\mathcal{U}} \int_\Omega f(x, u(x)) d\mu(x). \end{aligned}$$

This contradicts the fact that $u^*(\cdot) \in \arg\min_{u(\cdot)\in\mathcal{U}} \int_\Omega f(x, u(x)) d\mu(x)$.

2.8. Fréchet derivative

The Fréchet derivative is a generalization of the derivative to infinite-dimensional spaces, such as Hilbert or Banach spaces.

Definition 2.32 *([3]**Fréchet derivative**)*
Let $f : A \mapsto \mathbb{R}$ be a functional defined on Banach space A. f is Fréchet differentiable at u, if the Fréchet derivative of f at a point $u \in A$ exists as a bounded linear operator $Df_u : A \mapsto \mathbb{R}$ such that $\forall h \in A$,

$$\lim_{t \to 0} \frac{f(u+th) - f(u)}{t} = Df_u(h).$$

The Fréchet derivative matches the definition of the derivative of simple real functions. In fact, if f is a real function from $\mathbb{R}$ to $\mathbb{R}$, its derivative at $u \in \mathbb{R}$ satisfies

$$\lim_{t \to 0} \frac{f(u+th) - f(u)}{t} = Df_u(h) := f'(u)h,$$

which is just as defined by the Fréchet derivative.

If A is a Hilbert space, then by Riesz representation theorem 2.2, the Fréchet derivative $Df_u : A \mapsto \mathbb{R}$ uniquely corresponds to an element in H. Hence in the Hilbert space background, the Fréchet derivative refers to its Riesz representative but not the operator.

Below we give an example of the Fréchet derivative of a function acting on an infinite dimensional Hilbert space H equipped with a inner product $\langle \cdot, \cdot \rangle_H$.

Example 2.22 *(**Fréchet derivative of inner product**)*
The Fréchet derivative of the inner product $\langle \cdot, \cdot \rangle_H$ at $(g(u), f(u))$, where g and f are Fréchet differentiable operators from H to H, is given by $Dg_u^*(f(u)) + Df_u^*(g(u))$.

Proof:
By definition 2.32, $\forall h \in H$, the Fréchet derivative of the inner product at $(g(u), f(u))$ is given by

$$\begin{aligned} DI_u(h) &= \lim_{t \to 0} \frac{\langle g(u+th), f(u+th) \rangle_H - \langle g(u), f(u) \rangle_H}{t} \\ &= \lim_{t \to 0} \frac{\langle g(u+th), f(u+th) - f(u) \rangle_H}{t} + \lim_{t \to 0} \langle g(u+th) - g(u), f(u) \rangle_H \\ &= \langle g(u), Df_u(h) \rangle_H + \langle Dg_u(h), f(u) \rangle_H . \end{aligned} \tag{2.39}$$

Denote the adjoint operators by Dg_u^* and Df_u^*. Then by Definition of adjoint operators 2.6

$$\begin{aligned} \langle g(u), Df_u(h) \rangle_H &= \langle Df_u^*(g(u)), h \rangle_H \\ \langle Dg_u(h), f(u) \rangle_H &= \langle Dg_u^*(f(u)), h \rangle_H . \end{aligned}$$

Hence (2.39) is equivalent to

$$\langle Dg_u^*(f(u)) + Df_u^*(g(u)), h \rangle_H .$$

By Riesz representation theorem 2.2, $Dg_u^*(f(u)) + Df_u^*(g(u))$ is the unique Hilbert space element that DI_u corresponds to.

□

When f and g are linear operators, from Definition 2.32, it is clear that $Dg_u := g$ and $Df_u := f$. In this situation, the Riesz representative of Fréchet derivative DI_u shown in Example 2.22 is given by

$$g^*(f(u)) + f^*(g(u)). \tag{2.40}$$

3

Semiparametric information theory

Suppose there are some observed samples of a random variable $X_1, \cdots, X_n$ from a distribution P belonging to a set of probability measures $\mathcal{P}$ on the measurable space $(\Omega, \mathcal{A})$. Now the task is to estimate the value $\psi(P)$ where the functional satisfies $\psi : \mathcal{P} \mapsto \mathbb{R}$. If the set $\mathcal{P}$ takes the form $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, i.e. for a finite dimensional parameter θ in the distribution of X, the Fisher information is needed to measure the information that the observed samples carry about the unknown parameter θ, and the Cramér-Rao bound determines the minimal variance for estimating the parameter $\psi(P)$.

Definition 3.1 ***(Fisher information)***
For a one-dimensional parametric model P_θ with density p_θ, the Fisher information of $\theta \in \Theta \subseteq \mathbb{R}$ is

$$\mathcal{I}(\theta) = E_\theta[s(\theta, X)^2],$$

where $s(\theta, X) = \frac{\partial}{\partial\theta} \log p_\theta(X)$ is the score function.

Definition 3.2 ***(Cramér-Rao bound)***
Given the unknown parameter θ in the one-dimensional parametric model P_θ and the one-dimensional parameter $\psi(\theta)$ to be estimated, the Cramér-Rao bound is given by

$$\frac{[\psi'(\theta)]^2}{\mathcal{I}(\theta)}$$

To extend the idea of information like Fisher information and Cramér-Rao bound to semiparametric models, i.e. models that include both unknown finite dimensional parameters and infinite dimensional parameters (or functions), we can restrict the model $\mathcal{P}$ to any of its one-dimensional (sufficient in most situations) smooth parametric submodels with shape $\mathcal{P}_0 = \{P_\theta : \theta \in \Theta\}$ and find the Fisher information. Since the information of the whole model $\mathcal{P}$ is no larger than the infimum of all the Fisher information derived from its parametric submodels, the information of the semiparametric model $\mathcal{P}$ should be defined as this infimum [41].

The idea naturally leads to the construction of submodels. For any one-dimensional parametric model, the QMD (Definition 2.23) suffices for determining the score function and making sure the existence of Fisher information as shown in Theorem 2.17. Under this condition, if there exists a measurable function $g : \Omega \mapsto \mathbb{R}$ which is the derivative in square mean of root density then it is the score function.

This helps constructing the demanded one-dimensional parametric submodels from measurable function g. Now we consider maps $t \mapsto P_t$ from the right half neighborhood of 0: $[0, \epsilon) \subseteq [0, \infty)$ to $\mathcal{P}$ choosing P_0 as the "true" distribution of the observations. The submodels $\{P_t \in \mathcal{P} | 0 \leq t < \epsilon\}$ should be quadratic mean differentiable at $t = 0^+$ with score function g, i.e. it satisfies Equation (3.1).

$$\int \left[\frac{dP_t^{\frac{1}{2}} - dP^{\frac{1}{2}}}{t} - \frac{1}{2} g dP^{\frac{1}{2}} \right]^2 = o(1), \text{ as } t \downarrow 0. \tag{3.1}$$

This means usually the submodels $\{P_t \in \mathcal{P} | 0 \leq t < \epsilon\}$ are constructed in the way such that for any x [41],

$$g(x) = \frac{\partial}{\partial t}|_{t=0} \log dP_t(x).$$

3.1. Tangent sets, influence functions and information

The tangent set of a model $\mathcal{P}$ is defined on the basis of the score functions of its submodels.

Definition 3.3 ***(Tangent set)***

The tangent set $\dot{\mathcal{P}}_P$ of the model $\mathcal{P}$ at P consists of score functions g of all its parametric submodels $\{P_t \in \mathcal{P} | 0 \leq t < \epsilon\}$ passing through P when $t = 0$. The score function g satisfies Equation 3.1.

Since score functions satisfy $E[g^2] = \int g^2 dP < \infty$, the tangent set is a subset of $L_2(P)$.

Before introducing influence functions, we will first restrict the range of parametric submodels to those that are differentiable after being mapped by the functional ψ.

Definition 3.4 ***(Differentiable functional)***

Functional $\psi : \mathcal{P} \mapsto \mathbb{R}$ is said to be differentiable at P relative to a given tangent set $\dot{\mathcal{P}}_P$ if there exists a continuous linear map $\dot{\psi}_P : L_2(P) \mapsto \mathbb{R}$ such that for every $g \in \dot{\mathcal{P}}_P$, there is a submodel P_t with score function g,

$$\lim_{t \downarrow 0} \frac{\psi(P_t) - \psi(P)}{t} = \dot{\psi}_P(g).$$

Denote the closed linear span of the tangent set $\dot{\mathcal{P}}_P$ as $\overline{\text{lin}}\dot{\mathcal{P}}_P$, which is a closed subspace of $L_2(P)$. Although $\dot{\psi}_P$ is defined on the whole space $L_2(P)$, since $g \in \dot{\mathcal{P}}_P \subset \overline{\text{lin}}\dot{\mathcal{P}}_P$, by the Riesz representation theorem for Hilbert space (Theorem 2.2), there exists a Riesz representative function $\tilde{\psi}_P$ uniquely defined in $\overline{\text{lin}}\dot{\mathcal{P}}_P$.

$$\dot{\psi}_P(g) = \langle \tilde{\psi}_P, g \rangle_P = \int \tilde{\psi}_P g dP. \tag{3.2}$$

The uniquely defined function $\tilde{\psi}_P$ is called the efficient influence function of ψ.

Definition 3.5 ***(Efficient influence function)***

Among all the influence function of ψ under model $\mathcal{P}$, there exists a unique one belonging to $\overline{lin}\dot{\mathcal{P}}_P$ which is called the efficient influence function.

Since usually $\overline{\text{lin}}\dot{\mathcal{P}}_P \subsetneqq L_2(P)$, the efficient influence function should be the projection of other functions in $L_2(P)$ onto $\overline{\text{lin}}\dot{\mathcal{P}}_P$. Those functions are called influence functions and are not uniquely defined by the submodel $\mathcal{P}$ and g.

Now we can define the minimal asymptotic variance for estimating $\psi(p)$ in semiparametric models just like the Cramér-Rao bound in parametric models. Notice that for any one-dimensional parametric submodel P_t of the whole model $\mathcal{P}$, its Fisher information is $E[g^2]$. The Cramér-Rao bound for estimating $\psi(P_t)$ at $t = 0$ is given by

$$\frac{[\psi'(P_t)|_{t=0}]^2}{E[g^2]} = \frac{\langle \tilde{\psi}_P, g \rangle_P^2}{\langle g, g \rangle_P} \tag{3.3}$$

When choosing the supremum of Equation 3.3 over all elements in the closed linear span of the tangent set, we get the lower bound of the asymptotic variance for estimating $\psi(P)$. In fact, this lower bound is just the variance of the efficient influence function.

Lemma 3.1 ***(Efficiency bound)***

Suppose that the functional $\psi : \mathcal{P} \mapsto \mathbb{R}$ is differentiable at P relative to the tangent set $\dot{\mathcal{P}}_P$. Then

$$\sup_{g \in \overline{lin}\dot{\mathcal{P}}_P} \frac{\langle \tilde{\psi}_P, g \rangle_P^2}{\langle g, g \rangle_P} = E_P[\tilde{\psi_P^2}].$$

Proof:
The definition of efficient influence function states that $\tilde{\psi}_P \in \overline{\text{lin}}\dot{\mathcal{P}}_P$. By the Cauchy-Schwartz inequality, $E[\tilde{\psi}_P g]^2 \leq E[\tilde{\psi}_P^2]E[g^2]$, with equality if $g = \tilde{\psi}_p$, which directly leads to the equation.

□

3.2. Efficiency bound for ATE

As shown by Lemma 3.1, the efficiency bound for ATE determines the minimal asymptotic variance that the estimator of ATE can achieve. Before finding the efficiency bound, we need the efficient influence function for ATE. Hence, in this section, we introduce the efficient influence function in the basic counterfactual framework and then the attempt to find the efficient influence function in the proximal counterfactual framework.

3.2.1. In basic framework

The average treatment effect in the basic counterfactual model is given by $\chi = E_L E[Y^1 - Y^0|L]$, whose efficient estimators are restricted by efficient influence function and efficieny bound of it. To determine a lower bound of the variance for the efficient estimators, we firstly consider the semiparametric model to be investigated. Notice that the joint density of the observations $(Y = (\mathbb{1}_{A=1}Y^1, \mathbb{1}_{A=0}Y^0), A, L)$ is given by

$$f(Y,A,L) = [f^0(Y|L)]^{1-A}[f^1(Y|L)]^A[1 - f(1|L)]^{1-A}[f(1|L)]^A f^L(L),$$

where f^a is the marginal density for Y^a, $f(a|L)$ is the propensity score under $A = a$ and f^L is the marginal density for confounder L. The nonparametric model should satisfies Assumption 1.2 and 1.3 with unknown segments f^0, f^1, $f(1|L)$ and f^L.

Theorem 3.1 *([42] **Efficient influence function and efficiency bound**)*
Denote the conditional average outcome $E[Y^a|L]$ as μ^a. The efficient influence function $\tilde{\psi}$ of the average treatment effect identification formula $\chi = E_L E[Y^1 - Y^0|L]$ is given by

$$\tilde{\psi} = \frac{\mathbb{1}_{A=1}}{f(1|L)}(Y^1 - \mu^1) - \frac{\mathbb{1}_{A=0}}{f(0|L)}(Y^0 - \mu^0) + \mu^1 - \mu^0 - \chi. \tag{3.4}$$

The corresponding semiparametric local efficiency bound of χ equals $E[\tilde{\psi}^2]$.

Proof:
We consider the parametric submodel passing through true density $f(Y,A,L)$ when $t = 0$.

$$f_t(Y,A,L) = [f_t^0(Y|L)]^{1-A}[f_t^1(Y|L)]^A[1 - f_t(1|L)]^{1-A}[f_t(1|L)]^A f_t^L(L),$$

where the form of paths are set as follows to make sure the submodel is a density.

$$\begin{aligned} f_t(1|L) &= f(1|L) + t\phi(L), && \text{where } -f(1|L) < t\phi(L) < f(0|L), \\ f_t^a(Y|L) &= (1 + t\rho^a(Y|L))f^a(Y|L), a \in \{0,1\}, && \text{where } E[\rho^a(Y|L)|L] = 0, \\ f_t^L(L) &= (1 + t\gamma(L))f^L(L), && \text{where } E[\gamma(L)] = 0. \end{aligned}$$

The score function of the submodel is

$$g(Y,A,L) = \frac{\partial}{\partial t}|_{t=0} log f_t(Y,A,L) = \mathbb{1}_{A=0}\rho^0(Y|L) + \mathbb{1}_{A=1}\rho^1(Y|L) + \frac{\mathbb{1}_{A=1} - f(1|L)}{f(1|L)f(0|L)}\phi(L) + \gamma(L).$$

As defined by equation (3.2), the influence function $\tilde{\psi}$ of χ under the whole model satisfies

$$\begin{aligned}
\frac{\partial \chi_t}{\partial t}|_{t=0} &= \frac{\partial}{\partial t}|_{t=0} \int y(f_t^1(y|l) - f_t^0(y|l))f_t^L(l)d\mu(y,l) \\
&= \int y\left[f^1(y|l)(\rho^1(y|l) + \gamma(l)) - f^0(y|l)(\rho^0(y|l) + \gamma(l))\right] f^L(l)d\mu(y,l) \\
&= \underbrace{\int yf^1(y|l)(\rho^1(y|l) + \gamma(l))f^L(l)d\mu(y,l)}_{I_1} - \underbrace{\int yf^0(y|l)(\rho^0(y|l) + \gamma(l))f^L(l)d\mu(y,l)}_{I_2}.
\end{aligned}$$

Notice that

$$\tilde{\psi} = \left\{\frac{\mathbb{1}_{A=1}}{f(1|L)}(Y^1 - \mu^1) + \mu^1 - E[Y^1]\right\} - \left\{\frac{\mathbb{1}_{A=0}}{f(0|L)}(Y^0 - \mu^0) + \mu^0 - E[Y^0]\right\} := \tilde{\psi}^1 - \tilde{\psi}^0.$$

If we show $I_a = E[\tilde{\psi}^a g]$, for $a \in \{0,1\}$, then $\tilde{\psi}$ is indeed an influence function of χ. In fact,

$$\begin{aligned}
E[\tilde{\psi}^a g] &= E\left\{\left[\frac{\mathbb{1}_{A=a}(Y^a - \mu^a)}{f(A=a|L)} + \mu^a - E[Y^a]\right]\left[\mathbb{1}_{A=0}\rho^0(Y|L) + \mathbb{1}_{A=1}\rho^1(Y|L) + \frac{\mathbb{1}_{A=1} - f(1|L)}{f(1|L)f(0|L)}\phi(L) + \gamma(L)\right]\right\} \\
&= E\left\{\left[\frac{\mathbb{1}_{A=a}(Y^a - \mu^a)}{f(A=a|L)} + \mu^a\right]\left[\mathbb{1}_{A=0}\rho^0(Y|L) + \mathbb{1}_{A=1}\rho^1(Y|L) + \frac{\mathbb{1}_{A=1} - f(1|L)}{f(1|L)f(0|L)}\phi(L) + \gamma(L)\right]\right\} \\
&= \underbrace{E\left[\frac{\mathbb{1}_{A=a}(Y^a - \mu^a)}{f(A=a|L)}\rho^a(Y|L)\right]}_{E_1} + \underbrace{E\left[\frac{\mathbb{1}_{A=a}(Y^a - \mu^a)}{f(A=a|L)}\gamma(L)\right]}_{E_2} + \underbrace{E\left[\frac{\mathbb{1}_{A=a}(Y^a - \mu^a)}{f(A=a|L)}\frac{\mathbb{1}_{A=1} - f(1|L)}{f(1|L)f(0|L)}\phi(L)\right]}_{E_3} \\
&\quad + \underbrace{E\left[\mu^a[\mathbb{1}_{A=0}\rho^0(Y|L) + \mathbb{1}_{A=1}\rho^1(Y|L)]\right]}_{E_4} + \underbrace{E\left[\mu^a\frac{\mathbb{1}_{A=1} - f(1|L)}{f(1|L)f(0|L)}\phi(L)\right]}_{E_5} + \underbrace{E\left[\mu^a\gamma(L)\right]}_{E_6}.
\end{aligned}$$

To find the result of the six expectations, we write them as the repetitive expectation conditional on L: $E_L E[\cdot|L]$. E_2 and E_3 are zeros because they are multiplied by a common term $E[Y^a - \mu^a|L] = 0$. E_4 is a zero because $E[\rho^a(Y|L)|L] = 0$, for $a \in \{0,1\}$. E_5 is a zero because it is multiplied by $E[\mathbb{1}_{A=1} - f(1|L)|L] = 0$. Thus, by conditional exchangeability, the inner product of $\tilde{\psi}^a$ and g is equivalent to

$$\begin{aligned}
& E\left[\frac{\mathbb{1}_{A=a}(Y^a - \mu^a)}{f(A=a|L)}\rho^a(Y|L)\right] + E\left[\mu^a\gamma(L)\right] \\
&= E_L E\left\{\frac{1}{f(A=a|L)}E[\mathbb{1}_{A=a}|L]E[(Y^a - \mu^a)\rho^a(Y|L)|L]\right\} + E\left[\mu^a\gamma(L)\right] \\
&= E_L E[Y^a\rho^a(Y|L)|L] - E_L\left[\mu^a E[\rho^a(Y|L)|L]\right] + E\left[\mu^a\gamma(L)\right] \\
&= E_L[Y^a\rho^a(Y|L)|L] + E[\mu^a\gamma(L)] \\
&= I_a.
\end{aligned}$$

This means $\tilde{\psi}$ is an influence function of χ. Notice that $\tilde{\psi}$ is also a score function of a parametric model passing through the true density at $t = 0$ which chooses

$$\begin{aligned}
\phi(L) &= 0, \\
\rho^a(Y|L) &= (-1)^{1-a}\frac{Y^a - \mu^a}{f(A=a|L)}, \\
\gamma(L) &= \mu^1 - \mu^0 - \chi.
\end{aligned}$$

So, $\tilde{\psi}$ belongs to the closed linear span of the tangent set and thus is the efficient influence function of χ. Its efficiency bound is given by $E[\tilde{\psi}^2]$.

□

3.2.2. In proximal framework

To determine the information, we need the model to be considered in proximal framework.

Definition 3.6 ***(Proximal framework model)***
The whole model $\mathcal{P}_{proximal}$ in proximal framework is composed by all joint densities of (U, Z, X, W, A, Y) complying the factorization rule of random variables based on Assumption 1.4, i.e.

$$p(U, X, Z, W, A, Y) = p(U)p(X|U)p(Z|U, X)p(W|U, X)p(A|Z, U, X)p(Y|U, X, A, W), \tag{3.5}$$

where

$$p(A|Z, U, X) = [f_A(1|Z, U, X)]^A[1 - f_A(1|Z, U, X)]^{1-A}$$
$$p(Y|U, X, A, W) = [f_{Y^1}(y|U, X, W)]^A[f_{Y^0}(y|U, X, W)]^{1-A}.$$

Under the existence of the two bridge functions and the completeness assumptions such that the standardization formulae holds, the efficient influence function and efficiency bound of average treatment effect χ can be determined to test the efficiency of any estimators of χ.

Lemma 3.2 *([45]* ***Influence function)***
If there exists a submodel in $\mathcal{P}_{proximal}$ that supports the existence of bridge function h and q, the average treatment effect identification formula $\chi = E[h(W, A = 1, X) - h(W, A = 0, X)]$ has an influence function

$$\tilde{\psi} = (-1)^{1-A}q(Z, A, X)[Y - h(W, A, X)] + h(W, 1, X) - h(W, 0, X) - \chi. \tag{3.6}$$

Proof:

Denote $\mathcal{O} = (W, Y, Z, A, X)$, we write the joint density of $\mathcal{O}$ to be $f(\mathcal{O})$, and the score function to be S. Suppose f_t is an one-dimensional parametric submodel of the whole model $\mathcal{P}_{proximal}$ defined in Definition 3.6. The model $\mathcal{P}_{proximal}$ has the property in consistent with the Picard's condition (1.3) and (1.6) for the existence of the bridge functions, and satisfies completeness assumption 1.5 and 1.6.

As defined by Equation (3.2), the influence function $\tilde{\psi}$ of χ under the whole model satisfies

$$\frac{\partial \chi_t}{\partial t}|_{t=0} =< \tilde{\psi}, S_t(\mathcal{O}) > |_{t=0} = E[\tilde{\psi}S(\mathcal{O})].$$

Define $h_t(\Delta) := h_t(w, 1, x) - h_t(w, 0, x)$. By chain rule, the derivative equals to

$$\begin{aligned}\frac{\partial \chi_t}{\partial t}|_{t=0} &= \frac{\partial}{\partial t}|_{t=0} \int h_t(\Delta)dF_t(w, x)\\ &= \underbrace{\int h(\Delta)\frac{\partial}{\partial t}|_{t=0}f_t(w, x)d\mu(w, x)}_{I_1} + \underbrace{\int \frac{\partial}{\partial t}|_{t=0}h_t(\Delta)dF(w, x)}_{I_2}.\end{aligned}$$

Part (i): To show $I_1 = E[(h(\Delta) - \chi)S(\mathcal{O})]$.

$$\begin{aligned}I_1 &= \int h(\Delta)\left[\frac{\partial}{\partial t}|_{t=0}logf_t(w, x)\right]f_t(w, x)d\mu(w, x)\\ &= E[h(\Delta)S(W, X)]\\ &= E[h(\Delta)S(\mathcal{O}) - S(Z, Y, A|W, X)]\\ &= E[h(\Delta)S(\mathcal{O})] - E[h(\Delta)S(Z, Y, A|W, X)]\end{aligned} \tag{3.7}$$

The second expectation in the last row is 0 because the expectation of any score function is 0:

$$\begin{aligned}E[h(\Delta)S(Z, Y, A|W, X)] &= \int h(\Delta)S(z, y, a|w, x)f(z, y, a, w, x)d\mu(\mathcal{O})\\ &= \int h(\Delta)\left\{\int S(z, y, a|w, x)f(z, y, a|w, x)d\mu(z, y, a)\right\}f(w, x)d\mu(w, x)\\ &= E_{W,X}[h(\Delta)E_{Z,Y,A}[S(Z, Y, A|W, X)|W, X]]\\ &= E_{W,X}[h(\Delta) \cdot 0].\end{aligned}$$

The desire result is obtained by adding a 0 term $E[-\chi S(O)]$ on the equation (3.7).

Part (ii): To show $I_2 = E[(-1)^{1-A}q(Z,A,X)\epsilon S(O)]$, where $\epsilon = Y - h(W,A,X)$.

We temporarily denote $\frac{\partial}{\partial t}|_{t=0}h_t(w,A=a,x)$ as $dh(a)$ to avoid chaos.

$$\begin{aligned}
I_2 &= \int \frac{\partial}{\partial t}|_{t=0}h_t(\Delta)dF(w,x) \\
&= \int dh(1)f(w,x)d\mu(w,x) - \int dh(0)f(w,x)d\mu(w,x) \\
&= \int \left\{ \sum_{a\in\{0,1\}} (-1)^{1-a}dh(a)\frac{1}{f(A=a|w,x)}f(A=a|w,x) \right\} f(w,x)d\mu(w,x) \\
&= \int E_A[\frac{(-1)^{1-A}}{f(A|w,x)}dh(A)|W=w,X=x]f(w,x)d\mu(w,x) \\
&= E_{W,X}E_A[\frac{(-1)^{1-A}}{f(A|w,x)}dh(A)|W,X] \\
&= E\left[\frac{(-1)^{1-A}}{f(A|W,X)}\frac{\partial}{\partial t}|_{t=0}h_t(W,A=a,X)\right] \\
&= E_{W,A,X}E\left[\frac{(-1)^{1-A}}{f(A|W,X)}\frac{\partial}{\partial t}|_{t=0}h_t(W,A=a,X)|W,A,X\right].
\end{aligned}$$

The bridge function $q(Z,A=a,X)$ given in lemma 1.2 implies the last row equals:

$$\begin{aligned}
&E_{W,A,X}E\left[(-1)^{1-A}E\left[q(Z,A,X)|W,A,X\right]\frac{\partial}{\partial t}|_{t=0}h_t(W,A=a,X)|W,A,X\right] \\
=&E_{W,A,X}E_{W,A,X}\left[E\left[(-1)^{1-A}q(Z,A,X)\frac{\partial}{\partial t}|_{t=0}h_t(W,A=a,X)|W,A,X\right]|W,A,X\right] \\
=&E\left[(-1)^{1-A}q(Z,A,X)\frac{\partial}{\partial t}|_{t=0}h_t(W,A=a,X)\right] \\
=&E_{Z,A,X}\left[(-1)^{1-A}q(Z,A,X)E\left[\frac{\partial}{\partial t}|_{t=0}h_t(W,A=a,X)|Z,A,X\right]\right].
\end{aligned} \tag{3.8}$$

Recall the outcome confounding standardization formula (1.2), on the submodel, we have

$$\frac{\partial}{\partial t}|_{t=0}E_t[Y - h_t(W,A,X)|Z,A,X] = 0$$

$$\int \frac{\partial}{\partial t}|_{t=0}[(y - h_t(w,A,X))f_t(w,y|Z,A,X)]d\mu(w,y) = 0.$$

Let $\epsilon = Y - h(W,A,X)$, by the chain rule, we have

$$E[\epsilon S(W,Y|Z,A,X)|Z,A,X] = E[\frac{\partial}{\partial t}|_{t=0}h_t(W,A,X)|Z,A,X]. \tag{3.9}$$

Applying equation (3.9) to (3.8), we get

$$\begin{aligned}
I_2 &= E\left[(-1)^{1-A}q(Z,A,X)\epsilon S(W,Y|Z,A,X)\right] \\
&= E\left[(-1)^{1-A}q(Z,A,X)\epsilon(S(O - S(Z,A,X)))\right] \\
&= E\left[(-1)^{1-A}q(Z,A,X)\epsilon S(O)\right] - E\left[(-1)^{1-A}q(Z,A,X)\epsilon S(Z,A,X)\right].
\end{aligned}$$

The second expectation in the last row is 0 because

$$\begin{aligned}
&E\left[(-1)^{1-A}q(Z,A,X)\epsilon S(Z,A,X)\right] \\
=&E_{Z,A,X}\left[(-1)^{1-A}q(Z,A,X)S(Z,A,X)\cdot E\left[\epsilon|Z,A,X\right]\right] \\
=&E_{Z,A,X}\left[(-1)^{1-A}q(Z,A,X)S(Z,A,X)\cdot E\left[Y - h(W,A,X)|Z,A,X\right]\right] \\
=&E_{Z,A,X}\left[(-1)^{1-A}q(Z,A,X)S(Z,A,X)\cdot 0\right]. \qquad \text{(formula (1.2))}
\end{aligned}$$

Combining the results from part(i) and (ii), we have

$$E[\tilde{\psi}S(\mathcal{O})] = E\left[((-1)^{1-A}q(Z,A,X)\epsilon + h(\Delta) - \chi)S(\mathcal{O})\right],$$

which means that the influence function $\tilde{\psi}$ is just

$$(-1)^{1-A}q(Z,A,X)[Y - h(W,A,X)] + h(W,1,X) - h(W,0,X) - \chi.$$

□

Although the influence function is not enough for determining the efficiency bound, it still plays an important role in the proximal framework. First of all, its double robustness means it is consistent to zero whenever one of the bridge function exists, which provides a possibility of being a score function of some submodel. Moreover, it inspired ideas of estimating bridge functions for example the result given by Ghassami et al. [16] who also gave a method to estimate the influence function (3.6) and proposed its asymptotic normality.

However, the efficiency of the influence function is not an easy task to determine, since it is difficult to find a submodel whose tangent space includes the influence function (3.6). Cui et al. [45] tried to give The efficiency of the influence function (3.6) under a surjectivity assumption of conditional expectation operator through the following theorem.

Theorem 3.2 *([45]* ***Efficiency bound****)*
The influence function (3.6) is the efficient influence function, if the conditional expectation operators $T : L_2(P_{(W,A,X)}) \mapsto L_2(P_{(Z,A,X)})$ and its adjoint $T^* : L_2(P_{(Z,A,X)}) \mapsto L_2(P_{(W,A,X)})$ are surjective. The corresponding semiparametric local efficiency bound of χ is $E[\tilde{\psi}^2]$.

Proof:
Recall the tangent space of the whole model $\mathcal{P}_{proximal}$ (Definition 3.6) is given by the closed linear span of $\Lambda_1 + \Lambda_2$, where

$$\begin{aligned}\Lambda_1 &:= \left\{S(Z,A,X) : S(Z,A,X) \in L_2(P_{(Z,A,X)}),\ E[S(Z,A,X)] = 0\right\},\\ \Lambda_2 &:= \left\{S(Y,W|Z,A,X) : S(Y,W|Z,A,X) \in L_2(P_{(Z,A,X)})^{\perp},\ E[S(Y,W|Z,A,X)] = 0,\right.\\ &\qquad\qquad \left. E[\epsilon S(Y,W|Z,A,X)|Z,A,X] \in \overline{\text{Range}(T)}\right\}.\end{aligned}$$

The requirement $E[\epsilon S(Y,W|Z,A,X)|Z,A,X] \in \overline{\text{Range}(T)}$ is a direct result of the existence of the outcome confounding bridge function h, given by Equation (3.9).

Notice that Equation (3.6) can be decomposed into two parts with zero mean:

$$\begin{aligned}&(-1)^{1-A}q(Z,A,X)[Y - h(W,A,X)] + h(W,1,X) - h(W,0,X) - \chi\\ &= \underbrace{E[h(\Delta) - \chi|Z,A,X]}_{I_1} + \underbrace{h(\Delta) - \chi - E[h(\Delta) - \chi|Z,A,X] + (-1)^{1-A}q(Z,A,X)\epsilon}_{I_2}.\end{aligned}$$

Since $h(\Delta) - \chi \in L_2(P_{(W,A,X)})$, by the surjectivity of T, it is clear that $I_1 \in \Lambda_1$.
As for I_2, since $E_{Y,W}[I_2|Z,A,X] = 0$, we have $E_{\mathcal{O}}[I_2 \cdot g] = 0$, $\forall g \in L_2(P_{(Z,A,X)})$ and thus $I_2 \in L_2(P_{(Z,A,X)})^{\perp}$. Noticing that $E_Y[\epsilon I_2|W,Z,A,X] \in L_2(P_{(W,A,X)})$ for fixed Z, by the surjectivity of T, we have $E[\epsilon I_2|Z,A,X] = E_W E_Y[\epsilon I_2|W,Z,A,X] = T(E_Y[\epsilon I_2|W,Z,A,X]) \in \overline{\text{Range}(T)}$. This means $I_2 \in \Lambda_2$.
Hence the influence function (3.6) is efficient.

□

Kallus et al.[19] generalized the efficiency bound based on the Theorem 3.2 to their generalized average causal effect although they didn't avoid the surjectivity assumption of the two conditional expectation operators.

The surjectivity assumption in Theorem 3.2 for the conditional expectation operators implies, by Theorem 2.3, T and T^* are also injective and have closed range. This is equivalent to the statement that

T is bijective, which restricts the question to a trivial situation when the existence of bridge functions holds automatically. However it is rarely the case in the real problems.

Considering only the surjectivity of T is needed during the proof, even though dropping the assumption of the surjectivity of T^*, we still find the assumption extremely narrows the opportunity to apply the theorem. Typically, the range of T should be strictly smaller than $L_2(P_{(Z,A,X)})$, i.e.

$$\text{Range}(T) = \{E[f(W,A,X)|Z,A,X] : \forall f \in L_2(W,A,X)\} \subsetneq L_2(P_{(Z,A,X)}).$$

With the surjectivity of T, we have

$$\text{Range}(T) = \{E[f(W,A,X)|Z,A,X] : \forall f \in L_2(W,A,X)\} = L_2(P_{(Z,A,X)}).$$

This happens only when W is rich enough to explain Z, for example when Z and W are identically distributed conditional on U and X, while in real cases it is usually impossible to acquire information of one of the proxies from another.

4

Estimating the treatment confounding bridge function

Identification of the ATE in the proximal framework depends on the existence of the bridge functions while the Picard's conditions determining the existence are difficult to verify. This makes it a though task to calculate the ATE directly from the bridge functions. In this chapter, we focus on estimating the treatment confounding bridge function through the Fredholm integral equation of the first kind (1.5), which is given by

$$\frac{1}{f(a|W,X)} = E[\mathbb{1}_{A=a}q(Z,a,X)|W,A=a,X] \tag{4.1}$$

To achieve this, one requires the true data generating function $f(a|W,X)$ which is usually hard to get when confounder U is unknown. An approach is to find out the transformed integral equation problem that is irrelevant to $f(a|W,X)$ so that a straightforward method for estimating the treatment confounding bridge function is produced. In fact, by the tower property of conditional expectation, integral problem (4.1) is equivalent to

$$\begin{aligned}\frac{1}{f(a|W,X)} &= E[\mathbb{1}_{A=a}q(Z,a,X)|W,A=a,X] \\ 1 &= E[\mathbb{1}_{A=a}q(Z,a,X)|W,A=a,X]f(a|W,X) \\ 1 &= E_A[E_Z[\mathbb{1}_{A=a}q(Z,a,X)|W,A,X]|W,X] \\ 1 &= E_{ZA}[\mathbb{1}_{A=a}q(Z,a,X)|W,X].\end{aligned}$$

Since the treatment confounding bridge function under different treatments are different, we denote $q(Z,a,X)$ by q_a to discriminate between the two functions. If the solution exists, suppose $q_a^\star$ is the true treatment confounding bridge function under treatment $A=a$. Then it solves

$$E_{ZA}[\mathbb{1}_{A=a}q_a(Z,a,X)|W,X] = 1. \tag{4.2}$$

Estimating $q_a^\star$ will produce the estimator of treatment confounding bridge function under treatment $A = a$.

Suppose confounders X, W, Z are random variables over polished subspaces of $\mathbb{R}$ which we denote by $\mathcal{X}$, $\mathcal{W}$ and $\mathcal{Z}$ respectively. Let A be a binary random variable over set $\mathcal{A} = \{0,1\}$.

4.1. Transformed problems and a kernel embedded solution

The dual kernel embedding method was proposed by Dai et al.[9] to solve problems of learning from conditional distributions. These problems can be seen as solving the function connecting the conditional distribution to the target values. The method transforms the equation of conditional expectation to a kernel embedded minimax problem by a series of reformulations, including ERM reformulation, minimax problem reformulation through Fenchel duality and interchange of minimization and integration, and finally the kernel embeddings of means and cross-covariances. The method is based on the definition of universal kernels that any bounded continuous function can be well approximated by a function belonging to the RKHS induced by the universal kernel. Hence the kernel embedded solution approximates the original bounded continuous function arbitrarily well.

After Dai et al., Muandet et al.[25] applied the dual kernel embedding method to the instrumental variable problems and found a closed form solution of the causal function in instrumental variable regression. The instrumental variable regression stands for the inverse problem

$$E[Y|Z] = E[f(X)|Z], \tag{4.3}$$

where X, Y, Z are faithful to graphical model 4.1. Suppose the influence mechanism from X to Y is $Y = f(X) + \epsilon$, where f complies an unknown model and ϵ is the noise from e. If we want to identify the true mechanism f^* when $E[\epsilon] = 0$ and the only known information is $E[Y|X]$, we can't solve the influence from $E[Y|X] = f(X)$. The reason is that from the graphical model, we have $E[e|X] \neq E[e]$ which implies $E[\epsilon|X] \neq E[\epsilon] = 0$. This means the influence we found from $E[Y|X] = f(X)$ will always include the noise from e. To get rid off this noise, by $e \perp Z$, we know $E[e|Z] = E[e]$ and thus $E[\epsilon|Z] = E[\epsilon] = 0$. This makes the influence solved from (4.3) will only reflect the causality from X to Y. Hence, people refer to Z by the instrumental variable.

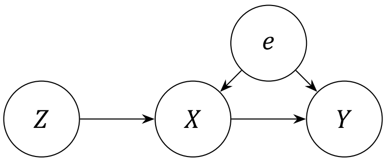


Figure 4.1: A DAG for instrumental variable scenario

By virtue of the similarity of the integral equation determining the existence of the treatment confounding bridge function and the instrumental variable regression, we apply the dual kernel method to find a kernel embedded q_a.

We start with the ERM reformulation of (4.2).

Lemma 4.1 ***(ERM reformulation)***
The true solution $q_a^\star(Z, a, X)$ for integral equation (4.2) solves the expected risk minimization problem with square loss for 1-dim data $l(x, y) = \frac{1}{2}(x - y)^2$

$$\min_{q_a \in L_2(P_{(Z,A,X|W,X)})} E_{W,X}[l(1, E_{ZA}[\mathbb{1}_{A=a} q_a(Z, a, X)|W, X])]. \tag{4.4}$$

Proof:

First we consider the solution $h^\star \in L_2(P_{(W,X)})$ for minimizing the mean square error

$$E_{W,X}[\frac{1}{2}(1 - h(W, X))^2] \tag{4.5}$$

By Theorem 2.18, the expected risk (4.5) has the minimizer $h^\star(W, X) := 1$. It is clear that $h^\star(W, X) = E_Z[\mathbb{1}_{A=a} q_a^\star(Z, a, X)|W, X]$ by Equation (4.2). Hence the true solution $q_a^\star(Z, a, X)$ for integral equation (4.2) solves the ERM problem.

□

Next comes the minimax problem reformulation of ERM reformulation (4.4).

Lemma 4.2 ***(Minimax problem reformulation)***
The ERM form (4.4) of integral equation (4.2) is equivalent to a minimax problem

$$\min_{q_a \in L_2(P_{(Z,A,X|W,X)})} \max_{u(\cdot,\cdot) \in \mathcal{M}(W,X)} E_{Z,W,A,X}[(\mathbb{1}_{A=a} q_a(Z, a, X) - 1)u(W, X)] - \frac{1}{2} E_{W,X}[u^2(W, X)], \tag{4.6}$$

where $\mathcal{M}(W, X)$ is the space of all measurable functions from $\mathcal{W} \times \mathcal{X} \mapsto \mathbb{R}$.

Proof:
For simplicity, we denote the square loss function $l(x,y)$ by $l_x(y)$.

$$
\begin{aligned}
&\min_{q_a \in L_2(P_{(Z,A,X|W,X)})} E_{W,X}[l_1(E_{ZA}[\mathbb{1}_{A=a} q_a(Z,a,X)|W,X])] \\
&= \min_{q_a \in L_2(P_{(Z,A,X|W,X)})} E_{W,X}[\max_{u\in\mathbb{R}} \{uE_{ZA}[\mathbb{1}_{A=a} q_a(Z,a,X)|W,X] - l_1^*(u)\}] \quad (4.7) \\
&= \min_{q_a \in L_2(P_{(Z,A,X|W,X)})} E_{W,X}[\max_{u\in\mathbb{R}} \left\{uE_{ZA}[\mathbb{1}_{A=a} q_a(Z,a,X)|W,X] - u - \frac{1}{2}u^2\right\}] \quad (4.8) \\
&= \min_{q_a \in L_2(P_{(Z,A,X|W,X)})} \max_{u(\cdot,\cdot)\in\mathcal{M}(W,X)} E_{W,X}[u(W,X)E_{ZA}[\mathbb{1}_{A=a} q_a(Z,a,X)|W,X] - u(W,X) - \frac{1}{2}u^2(W,X)] \\
&= \min_{q_a \in L_2(P_{(Z,A,X|W,X)})} \max_{u(\cdot,\cdot)\in\mathcal{M}(W,X)} E_{Z,W,A,X}[(\mathbb{1}_{A=a} q_a(Z,a,X) - 1)u(W,X)] - \frac{1}{2}E_{W,X}[u^2(W,X)].
\end{aligned}
$$

In Example 2.20, it is shown that (l_1^*, l_1) are dual to each other. Since the supremum over $\mathbb{R}$ is achievable by maximum, (4.4) can be transferred to (4.7), a minimum problem with a maximum inside, by $l_1 = l_1^{**}$. Substituting l_1^* by the result with $c = \frac{1}{2}$ in Example 2.19, one can get the equivalent expression (4.8). By Example 2.21, l_1^* is a normal integrands. And it is easy to show that $l_1^*(u)$ plus a linear term of u is still normal. As defined that $\mathcal{M}(W,X)$ contains all measurable function from $\mathcal{W} \times \mathcal{X} \mapsto \mathbb{R}$, $\mathcal{M}(W,X)$ is a decomposable space by Definition 2.31. Hence after applying Theorem 2.22 to the negative function of the normal integrand $\frac{1}{2}u^2 - u(E_{ZA}[\mathbb{1}_{A=a} q_a(Z,a,X)|W,X] - 1)$, we interchange the maximum and the expectation from finding the maximum over R to the function space $\mathcal{M}(W,X)$. This gives the equivalent problem (4.6), which is a minimax problem.

□

To remain consistent with the terminology used by Dai et al. [9], we will continue to refer to $u(W,X)$ as the dual function. Below we give two important properties of the optimal dual function $u^*(W,X)$.

Proposition 4.1 ***(Uniqueness)***
The optimal dual function u^* is uniquely defined by the derivative of the square loss function $l_1(\cdot)$ at $E_{ZA}[\mathbb{1}_{A=a} q_a(Z,a,X)|W,X]$, which is $u^*(W,X) = E_{ZA}[\mathbb{1}_{A=a} q_a(Z,a,X)|W,X] - 1$.

Proof:
By Theorem 2.22, since the maximum is finite, the optimal function u^* satisfies

$$u^*(W,X) = \arg\max_{u\in\mathbb{R}} \{uE_{ZA}[\mathbb{1}_{A=a} q_a(Z,a,X)|W,X] - l_1^*(u)\},$$

everywhere on $\mathcal{W} \times \mathcal{X}$. Combining $E_{ZA}[\mathbb{1}_{A=a} q_a(Z,a,X)|W,X] \in \partial l_1^*(u^*(W,X))$ implied by (4.7) and Proposition 2.11, we have that the optimal dual function u^* satisfies

$$u^*(W,X) \in \partial l_1(E_{ZA}[\mathbb{1}_{A=a} q_a(Z,a,X)|W,X]).$$

Since the square loss $l_1(\cdot)$ is differentiable with derivative $x - 1$ at point x, the optimal dual function is uniquely defined by the derivative of $l_1(\cdot)$ at $E_{ZA}[\mathbb{1}_{A=a} q_a(Z,a,X)|W,X]$, which is $E_{ZA}[\mathbb{1}_{A=a} q_a(Z,a,X)|W,X] - 1$.

□

Proposition 4.2 ***(Bounded continuity)***
The optimal dual function $u^*(W,X)$ is bounded continuous on $\mathcal{W} \times \mathcal{X}$ if

- *for $a \in \mathcal{A}$, $\mathbb{1}_{A=a} q_a(Z,a,X)$ is continuous in both Z and X or at least continuous in Z for any X;*
- *the conditional density $p(Z|W,a,X)$ and conditional probability $f(a|W,X)$ are continuous at any $(W,X) \in \mathcal{W} \times \mathcal{X}$.*

Proof:

The two continuity assumptions make sure $E_{ZA}[\mathbb{1}_{A=a}q_a(Z,a,X)|W,X]$ is a bounded continuous function of (W,X). Since the derivative of square loss $l_1(\cdot)$ is continuous, combining the result in Proposition 4.1, we have the bounded continuity of the optimal dual function.

□

Proposition 4.1 and 4.2 show that the optimal dual function $u^*(W,X)$ is uniquely defined on the space of all bounded continuous functions $\mathcal{W}\times\mathcal{X}\mapsto\mathbb{R}$, which we denote by $\mathcal{C}_0(W,X)$.

To commence further transforms about the minimax problem (4.6), we need the following assumption.

Assumption 4.1 ***(Continuous treatment confounding bridge function)***

The Picard's condition for the existence of the treatment confounding bridge function (1.6) holds in a way such that there exists at least one bounded continuous treatment confounding bridge function for each treatment.

Since $\mathcal{C}_0(W,X)\subseteq\mathcal{M}(W,X)$ and $\mathcal{C}_0(Z,A,X)\subseteq L_2(P_{(Z,A,X|W,X)})$ over any polished subspaces of $\mathbb{R}$, under Assumption 4.1, we can restrict the minimax problem (4.6) to

$$\min_{q_a\in\mathcal{C}_0(Z,A,X)}\max_{u\in\mathcal{C}_0(W,X)} E_{Z,W,A,X}[(\mathbb{1}_{A=a}q_a(Z,a,X)-1)u(W,X)]-\frac{1}{2}E_{W,X}[u^2(W,X)]. \tag{4.9}$$

By the definition of universal kernels, the RKHS induced by a universal kernel is dense in the space of bounded continuous functions. We can further restrict (4.9) to a minimax problem finding solutions in RKHSs. This gives

$$\min_{q_a\in F}\max_{u\in H} E_{Z,W,A,X}[(\mathbb{1}_{A=a}q_a(Z,a,X)-1)u(W,X)]-\frac{1}{2}E_{W,X}[u^2(W,X)]. \tag{4.10}$$

Lemma 4.3 ***(Kernel embedded reformulation)***

For any two universal kernels $k:(\mathcal{W}\times\mathcal{X})\times(\mathcal{W}\times\mathcal{X})\mapsto\mathbb{R}$ and $\tilde{l}:(\mathcal{Z}\times\{a\}\times\mathcal{X})\times(\mathcal{Z}\times\{a\}\times\mathcal{X})\mapsto\mathbb{R}$, we denote H, $\tilde{F}$ and F by RKHSs induces by k, $\tilde{l}$ and l, where $l:(\mathcal{Z}\times\mathcal{A}\times\mathcal{X})\times(\mathcal{Z}\times\mathcal{A}\times\mathcal{X})\mapsto\mathbb{R}$ is defined by $l((Z,A,X),(Z',A',X')):=\mathbb{1}_{A=a}\mathbb{1}_{A'=a}\tilde{l}((Z,a,X),(Z',a,X'))$. The corresponding feature maps of k, $\tilde{l}$ and l are given by $\phi:\mathcal{W}\times\mathcal{X}\mapsto H$, $\tilde{\psi}:\mathcal{Z}\times\{a\}\times\mathcal{X}\mapsto\tilde{F}$ and $\psi:\mathcal{Z}\times\mathcal{A}\times\mathcal{X}\mapsto F$. Denote the Hilbert-Schmidt Riesz representatives of (cross)-covariance operators by the operator $C_{(ZAX)(WX)}$ from F to H and the operator $C_{(WX)}$ from H to H respectively. Denote the mean embedding on H by $\mu_{(WX)}$. The kernel embedded form of (4.10) is given by

$$\min_{q_a\in F}\max_{u\in H}\Gamma(q_a,u),\ \textit{for}\ \Gamma(q_a,u):=\left\langle C_{(ZAX)(WX)}q_a-\mu_{(WX)}-\frac{1}{2}C_{(WX)}u,u\right\rangle_H, \tag{4.11}$$

where the operators $C_{(ZAX)(WX)}$ and $C_{(WX)}$ are not centered[1].

Proof:

After restrict to the RKHSs, the expectation of dual function u and the cross-covariance of $\mathbb{1}_{A=a}q_a$ and dual function u can be represented by mean embedding (2.18) and cross-covariance embedding (2.25), where the embeddings considered here are not centered.

First notice that $q_a\in\tilde{F}\cap F$, since by Moore-Aronszajn theorem 2.14, q_a has a representation in F which is given by

$$q_a=\sum_{i=1}^n\alpha_i\tilde{\psi}(z_i,a,x_i)=\sum_{i=1}^n\alpha_i\mathbb{1}_{a_i=a}\tilde{\psi}(z_i,a,x_i)=\sum_{i=1}^n\alpha_i\psi(z_i,a_i,x_i),\ \text{where}\ a_1=\cdots=a_n=a.$$

[1] Recall the remark about the cross-covariance without centering in Subsection 2.4.3, where the Hilbert-Schmidt Riesz representatives of not centered cross-covariance and its empirical version are defined by (2.29) and (2.30).

This gives

$$\begin{aligned}
E_{WX}[u(W,X)] &= \langle \mu_{(WX)}, u\rangle_H ,\\
E_{Z,W,A,X}[\mathbb{1}_{A=a} q_a(Z,a,X)u(W,X)] &= E_{Z,W,A,X}\left[\mathbb{1}_{A=a}\left\langle q_a, \tilde{\psi}(Z,a,X)\right\rangle_{\tilde{F}} \langle u, \phi(W,X)\rangle_H\right]\\
&= E_{Z,W,A,X}\left[\langle q_a, \underbrace{\mathbb{1}_{A=a}\tilde{\psi}(Z,a,X)}_{\psi(Z,A,X)}\rangle_F \langle u, \phi(W,X)\rangle_H\right]\\
&= \left\langle q_a \otimes u, E_{Z,W,A,X}[\psi(Z,A,X)\otimes\phi(W,X)]\right\rangle_{HS(H,F)}\\
&= \left\langle q_a \otimes u, C_{(WX)(ZAX)}\right\rangle_{HS(H,F)}\\
&= \left\langle q_a, C_{(WX)(ZAX)}u\right\rangle_F\\
&= \left\langle C_{(ZAX)(WX)}q_a, u\right\rangle_H .
\end{aligned}$$

Hence the kernel embedded form of (4.10) is given by (4.11).

□

Notice that although the kernel embedded form (4.11) is concave relative to u and convex about q_a, the uniqueness of the saddle point does not hold since the convexity and concavity are not strict. If the saddle point (q_a^*, u^*) exists, then by first order condition, it must satisfy

$$\begin{aligned}
\nabla_{q_a}\Gamma\Big|_{(q_a=q_a^*,u)} &= C_{(WX)(ZAX)}u = 0\\
\nabla_u\Gamma\Big|_{(q_a,u=u^*)} &= C_{(ZAX)(WX)}q_a - \mu_{(WX)} - C_{(WX)}u^* = 0.
\end{aligned}$$

Hence the sufficient condition for the existence of saddle points is that

1. $u \in \mathrm{Null}(C_{(WX)(ZAX)})$;
2. $C_{(WX)}u^* + \mu_{(WX)} \in \mathrm{Range}(C_{(ZAX)(WX)})$.

Furthermore, to determine the uniqueness of the saddle point, it must be assumed that operators

$$C_{WX} : H \mapsto H \text{ and } C_{(WX)(ZAX)}C_{(WX)}^{-1}C_{(ZAX)(WX)} : F \mapsto F$$

are invertible. However, by the Riesz-Schauder theorem 2.6, if the Hilbert space is infinite dimensional, then 0 belongs to the spectrum of the compact operator. This means the compact operator like C_{WX} (Proposition 2.7 and Example 2.3) acting on an infinite dimensional Hilbert space can't be boundedly invertible. Hence the invertibility assumption restricts H and F to be finite dimensional, which isn't true in general for RKHSs induced by universal kernels. So, the series of transforms from the original inverse problem (4.2) to kernel embedded minimax problem (4.11) have not eliminated the ill-posedness of the original problem but shifted it to an ill-posed problem that can be handled by regularization.

Hence, we seek to find the solution of the regularized form of kernel embedded minimax problem. It is given by

$$\min_{q_a\in F}\max_{u\in H}\Gamma_{\lambda_1,\lambda_2}(q_a,u), \text{ for } \Gamma_{\lambda_1,\lambda_2}(q_a,u) := \left\langle C_{(ZAX)(WX)}q_a - \mu_{(WX)} - \frac{1}{2}C_{(WX)}u, u\right\rangle_H - \frac{1}{2}\lambda_1\|u\|_H^2 + \frac{1}{2}\lambda_2\|q_a\|_F^2, \tag{4.12}$$

where λ_1 and λ_2 guarantee the bounded invertibility of $C_{(WX)}+\lambda_1 I$ and $C_{(WX)(ZAX)}(C_{(WX)}+\lambda_1 I)^{-1}C_{(ZAX)(WX)}+\lambda_2 I$ with the identity operator I.

The regularization terms make the kernel embedded minimax problem strictly convex in q_a and strictly concave in u. This means there exists a unique saddle point with a closed-form expression.

Theorem 4.1 ***(Regularized kernel embedded reformulation)***

The regularized kernel embedded minimax problem $\min_{q_a \in F} \max_{u \in H} \Gamma_{\lambda_1,\lambda_2}(q_a, u)$ *(4.12) has a unique saddle point* (q_a^*, u^*) *given by*

$$u^* = (C_{(WX)} + \lambda_1 I)^{-1}(C_{(ZAX)(WX)} q_a - \mu_{(WX)})$$
$$q_a^* = (C_{(WX)(ZAX)}(C_{(WX)} + \lambda_1 I)^{-1} C_{(ZAX)(WX)} + \lambda_2 I)^{-1} C_{(WX)(ZAX)}(C_{(WX)} + \lambda_1 I)^{-1} \mu_{(WX)}.$$

Proof:

By Example 3.4.1 of [4], the saddle point of a minimax problem is unique only if the second derivatives on the saddle point is negative definite relative to the maximizer and positive definite about the minimizer. In (4.12), the uniqueness of the saddle point derives from the strict concavity relative to u and strict convexity relative to q_a, which means the negative definiteness and positive definiteness of second Fréchet derivative of u and q_a.

The unique optimal kernel embedded dual function $u^* \in H$ satisfies the first order condition of the extreme value point. Combining the result (2.40) given by Example 2.22 and setting Fréchet derivative relative to u about $\Gamma_{\lambda_1,\lambda_2}(q_a, u)$ at $(q_a, u = u^*)$ to zero, by the self-adjointness of $C_{(WX)}$, we have

$$-\frac{1}{2}(C_{(WX)} + \lambda_1 I)u^* + C_{(ZAX)(WX)} q_a - \mu_{(WX)} - \frac{1}{2}(C_{(WX)} + \lambda_1 I)u^* = 0$$
$$C_{(ZAX)(WX)} q_a - \mu_{(WX)} = (C_{(WX)} + \lambda_1 I)u^*.$$

This yields the optimal kernel embedded dual function $u^* \in H$

$$u^* = (C_{(WX)} + \lambda_1 I)^{-1}(C_{(ZAX)(WX)} q_a - \mu_{(WX)}).$$

Putting back the u^* to the original function (4.12), we have

$$\begin{aligned}\min_{q_a \in F} \Gamma_{\lambda_1,\lambda_2}(q_a, u^*) : = &\left\langle C_{(ZAX)(WX)} q_a - \mu_{(WX)}, (C_{(WX)} + \lambda_1 I)^{-1}(C_{(ZAX)(WX)} q_a - \mu_{(WX)})\right\rangle_H \\ &-\left\langle \frac{1}{2} C_{(WX)}(C_{(WX)} + \lambda_1 I)^{-1}(C_{(ZAX)(WX)} q_a - \mu_{(WX)}), (C_{(WX)} + \lambda_1 I)^{-1}(C_{(ZAX)(WX)} q_a - \mu_{(WX)})\right\rangle_H \\ &-\left\langle \frac{1}{2}\lambda_1 (C_{(WX)} + \lambda_1 I)^{-1}(C_{(ZAX)(WX)} q_a - \mu_{(WX)}), (C_{(WX)} + \lambda_1 I)^{-1}(C_{(ZAX)(WX)} q_a - \mu_{(WX)})\right\rangle_H \\ &+\frac{1}{2}\lambda_2 \|q_a\|_F^2.\end{aligned}$$

Notice that

$$I - \frac{1}{2} C_{(WX)}(C_{(WX)} + \lambda_1 I)^{-1} - \frac{1}{2}\lambda_1 (C_{(WX)} + \lambda_1 I)^{-1} = I - \frac{1}{2}(C_{(WX)} + \lambda_1 I)(C_{(WX)} + \lambda_1 I)^{-1} = \frac{1}{2} I.$$

Hence, the minimization problem is equivalent to

$$\min_{q_a \in F} \Gamma_{\lambda_1,\lambda_2}(q_a, u^*) := \left\langle \frac{1}{2}(C_{(ZAX)(WX)} q_a - \mu_{(WX)}), (C_{(WX)} + \lambda_1 I)^{-1}(C_{(ZAX)(WX)} q_a - \mu_{(WX)})\right\rangle_H + \frac{1}{2}\lambda_2 \|q_a\|_F^2. \tag{4.13}$$

By the uniqueness of minimizer q_a^*, also combining the result (2.40) given by Example 2.22, we set the Fréchet derivative of the $\Gamma_{\lambda_1,\lambda_2}(q_a, u^*)$ at point $(q_a = q_a^*, u)$ to be zero and get

$$\begin{aligned}&\frac{1}{2} C_{(WX)(ZAX)}(C_{(WX)} + \lambda_1 I)^{-1}(C_{(ZAX)(WX)} q_a^* - \mu_{(WX)}) \\ &\qquad + C_{(WX)(ZAX)}(C_{(WX)} + \lambda_1 I)^{-1}\frac{1}{2}(C_{(ZAX)(WX)} q_a^* - \mu_{(WX)}) + \lambda_2 q_a^* = 0 \\ &\Longrightarrow (C_{(WX)(ZAX)}(C_{(WX)} + \lambda_1 I)^{-1} C_{(ZAX)(WX)} + \lambda_2 I) q_a^* = C_{(ZAX)(WX)}(C_{(WX)} + \lambda_1 I)^{-1} \mu_{(WX)}.\end{aligned}$$

This produces the optimal regularized kernel embedded solution $q_a^* \in F$

$$q_a^* = (C_{(WX)(ZAX)}(C_{(WX)} + \lambda_1 I)^{-1} C_{(ZAX)(WX)} + \lambda_2 I)^{-1} C_{(WX)(ZAX)}(C_{(WX)} + \lambda_1 I)^{-1} \mu_{(WX)}. \tag{4.14}$$

□

By virtue of the definition of universal kernels, the RKHS induced by the universal kernel is dense in the space of bounded continuous functions over a polished space. This means q_a^* (4.14) approximates to a bounded continuous function in $\mathcal{C}_0(Z, A, X)$ arbitrarily well.

4.2. Convergence of the kernel embedded solution

If the original integral equation (4.2) has a bounded continuous solution, we claim that q_a^* asymptotically converges to a kernel embedded solution in F as λ_2 goes to 0, and then, by the definition of universal kernels, approximates the true bounded continuous solution. This true bounded continuous solution should has the smallest L_2 norm by the nature of Tikhonov regularization. We will prove the statement by showing that finding the regularized kernel embedded minimizer $q_a^* \in F$ is equivalent to solving the ill-posed problem $C_{(ZAX)(WX)}q_a = \mu_{(WX)}$ by Tikhonov regularization. We start by transforming the original integral equation (4.2) into a kernel embedded ill-posed problem.

Lemma 4.4 ***(Kernel embedded form of the original problem)***
If Assumption 4.1 holds, then there exists a RKHS solution of

$$C_{(ZAX)(WX)}q_a = \mu_{(WX)}$$

that approximates the bounded continuous solution of (4.2) arbitrarily well.

Proof:
For any solution q_a solving the original integral equation (4.2), for all $g \in H$, we have

$$\begin{aligned} E_{W,X,A,Z}[\mathbb{1}_{A=a}q_a(Z,a,X)g(W,X)] &= E_{W,X}E_{A,Z}[\mathbb{1}_{A=a}q_a(Z,a,X)g(W,X)|W,X] \\ &= E_{W,X}[g(W,X)\underbrace{E_{A,Z}[\mathbb{1}_{A=a}q_a(Z,a,X)|W,X]}_{=1 \text{ by } (4.2)}] \\ &= E_{W,X}[g(W,X)]. \end{aligned} \tag{4.15}$$

Hence, any solution of (4.2) must be a solution of (4.15). Moreover, if Assumption 4.1 holds, we can restrict $q_a \in F$ so as to find a RKHS function in F that approximates a bounded continuous solution in $\mathcal{C}_0(Z, A, X)$ arbitrarily well.
By kernel embeddings, the transformed equation is given by

$$\langle C_{(ZAX)(WX)}q_a, g\rangle_H = \langle \mu_{(WX)}, g\rangle_H,$$

which implies that

$$C_{(ZAX)(WX)}q_a = \mu_{(WX)}. \tag{4.16}$$

□

Lemma 4.5 ***(Equivalence of two ill-posed problems)***
Assume λ_1 is a fixed positive number so that the operator $C_{(WX)} + \lambda_1 I$ is positive definite. The ill-posed problem (4.16) is equivalent to the preconditioned ill-posed problem

$$A_{\lambda_1}q_a = \mu_{\lambda_1}, \tag{4.17}$$

where $A_{\lambda_1} := (C_{(WX)} + \lambda_1 I)^{-\frac{1}{2}}C_{(ZAX)(WX)}$, and $\mu_{\lambda_1} := (C_{(WX)} + \lambda_1 I)^{-\frac{1}{2}}\mu_{(WX)}$.

Proof:
Since the operator $C_{(WX)} + \lambda_1 I$ is positive definite, it has a decomposition

$$C_{(WX)} + \lambda_1 I = (C_{(WX)} + \lambda_1 I)^{\frac{1}{2}}(C_{(WX)} + \lambda_1 I)^{\frac{1}{2}}, \tag{4.18}$$

where $(C_{(WX)}+\lambda_1 I)^{\frac{1}{2}}$ is also positive definite. Applying the invertible operator $(C_{(WX)}+\lambda_1 I)^{\frac{1}{2}}$ on the left of the both sides of (4.16), we get the preconditioned ill-posed problem (4.17). If exist, the solutions to the two problems (4.16) and (4.17) are equivalent by the invertible operator.

□

Lemma 4.6 ***(Tikhonov regularization equivalence of minimization problem)***

Assume λ_2 is a positive regularization parameter. The minimization problem (4.13) is the least square problem finding the solution to the Tikhonov regularized problem of (4.17), which is given by

$$\min_{q_a \in F} \frac{1}{2}\|A_{\lambda_1} q_a - \mu_{\lambda_1}\|_H^2 + \frac{1}{2}\lambda_2 \|q_a\|_F^2.$$

Furthermore, the regularized kernel embedded solution q_a^* (4.14) is just the Tikhonov regularized solution of (4.17).

Proof:

Based on the operator decomposition (4.18), the minimization relative to q_a in the proof of Theorem 4.1 is given by

$$\begin{aligned}\min_{q_a \in F} \Gamma_{\lambda_1,\lambda_2}(q_a, u^*) :&= \left\langle \frac{1}{2}(C_{(ZAX)(WX)} q_a - \mu_{(WX)}), (C_{(WX)} + \lambda_1 I)^{-1}(C_{(ZAX)(WX)} q_a - \mu_{(WX)}) \right\rangle_H + \frac{1}{2}\lambda_2 \|q_a\|_F^2 \\ &= \frac{1}{2}\|(C_{(WX)} + \lambda_1 I)^{-\frac{1}{2}}(C_{(ZAX)(WX)} q_a - \mu_{(WX)})\|_H^2 + \frac{1}{2}\lambda_2 \|q_a\|_F^2 \\ &= \frac{1}{2}\|(C_{(WX)} + \lambda_1 I)^{-\frac{1}{2}} C_{(ZAX)(WX)} q_a - (C_{(WX)} + \lambda_1 I)^{-\frac{1}{2}}\mu_{(WX)}\|_H^2 + \frac{1}{2}\lambda_2 \|q_a\|_F^2 \\ &= \frac{1}{2}\|A_{\lambda_1} q_a - \mu_{\lambda_1}\|_H^2 + \frac{1}{2}\lambda_2 \|q_a\|_F^2.\end{aligned}$$

This the least square problem for finding the solution to the Tikhonov regularized problem of the preconditioned ill-posed problem (4.17). We can verify the equivalence of q_a^* and the least square minimizer. In fact, by first order condition of the extreme values, we set the Fréchet derivative of $\Gamma_{\lambda_1,\lambda_2}$ at the minimizer to be 0 and get

$$\begin{aligned}\frac{\partial \Gamma_{\lambda_1,\lambda_2}}{\partial q_a} &= A_{\lambda_1}^*(A_{\lambda_1} q_a - \mu_{\lambda_1}) + \lambda_2 q_a \\ &= (A_{\lambda_1}^* A_{\lambda_1} + \lambda_2 I) q_a - A_{\lambda_1}^* \mu_{\lambda_1} = 0.\end{aligned}$$

This implies that the unique least square minimizer is given by

$$\begin{aligned}&(A_{\lambda_1}^* A_{\lambda_1} + \lambda_2 I)^{-1} A_{\lambda_1}^* \mu_{\lambda_1} \\ =&(C_{(WX)(ZAX)}(C_{(WX)} + \lambda_1 I)^{-1} C_{(ZAX)(WX)} + \lambda_2 I)^{-1} C_{(WX)(ZAX)}(C_{(WX)} + \lambda_1 I)^{-1}\mu_{(WX)},\end{aligned}$$

which is exactly the regularized kernel embedded solution q_a^* (4.14).

□

Many literatures have shown that under some source conditions, the Tikhonov regularized solution, i.e. the minimizer of the least square problem converges to the minimum-norm solution of the ill-posed problem with a convergence rate relevant to the perturbation of the noises from samples [7]. Since we don't have any prior knowledge about the sample noise, we will show the convergence by spectral decompositions of the operators under the source condition that the true solution lies in the RKHS F.

Theorem 4.2 ***(Convergence of the regularized kernel embedded solution)***

Given the positive regularization parameter λ_2, then as $\lambda_2 \longrightarrow 0$, the regularized kernel embedded solution q_a^* (4.14) converges to a solution of preconditioned ill-posed problem (4.17) contained in F, which we denote by $q_a^\dagger$.

Proof:

By spectral theorem (2.7) and (2.8), the compact operators $A^*_{\lambda_1}A_{\lambda_1} + \lambda_2 I$ and A_{λ_1} have spectral decompositions

$$A^*_{\lambda_1}A_{\lambda_1} + \lambda_2 I = \sum_{i\geq 1}(\sigma_i^2 + \lambda_2) f_i \otimes f_i \tag{4.19}$$

$$A_{\lambda_1} = \sum_{i\geq 1} \sigma_i h_i \otimes f_i, \tag{4.20}$$

where the singular values $(\sigma_i)_{i\geq 1}$ of A_{λ_1} with corresponding left and right eigenvectors $(f_i)_{i\geq 1}$, $(h_i)_{i\geq 1}$ forming the singular system of A_{λ_1}, which is denoted by $(\sigma_i, h_i, f_i)_{i\geq 1}$. Moreover, $(f_i)_{i\geq 1}$ and $(h_i)_{i\geq 1}$ form orthonormal bases in F and H respectively.

In the proofs of previous lemmas and theorems, we have that the solution $q_a^\dagger$ and kernel embedded solution q_a^* satisfy

$$A_{\lambda_1} q_a^\dagger = \mu_{\lambda_1} \tag{4.21}$$

$$(A^*_{\lambda_1}A_{\lambda_1} + \lambda_2 I) q_a^* = A^*_{\lambda_1}\mu_{\lambda_1}, \tag{4.22}$$

Replacing the operator A_{λ_1} in (4.21) by its spectral decomposition (4.20), we have

$$\begin{aligned}
\mu_{\lambda_1} &= \sum_{i\geq 1} \sigma_i (h_i \otimes f_i)(q_a^\dagger) \\
&= \sum_{i:\sigma_i\neq 0}^{\infty} \sigma_i \left\langle q_a^\dagger, f_i \right\rangle_F h_i \\
\Longrightarrow \left\langle h_i, \mu_{\lambda_1}\right\rangle_H &= \sigma_i \left\langle q_a^\dagger, f_i \right\rangle_F \\
\frac{\left\langle h_i, \mu_{\lambda_1}\right\rangle_H}{\sigma_i} &= \left\langle q_a^\dagger, f_i \right\rangle_F .
\end{aligned}$$

Since $q_a^\dagger \in F$, this implies that

$$q_a^\dagger = \sum_{i:\sigma_i\neq 0}^{\infty} \frac{\left\langle h_i, \mu_{\lambda_1}\right\rangle_H}{\sigma_i} f_i. \tag{4.23}$$

Replacing the operator $A^*_{\lambda_1}A_{\lambda_1} + \lambda_2 I$ in (4.22) by its spectral decomposition (4.19), we have

$$\begin{aligned}
A^*_{\lambda_1}\mu_{\lambda_1} &= \sum_{i\geq 1}(\sigma_i^2 + \lambda_2)(f_i \otimes f_i)(q_a^*) \\
&= \sum_{i\geq 1}(\sigma_i^2 + \lambda_2)\left\langle f_i, q_a^*\right\rangle_F f_i \\
\Longrightarrow \left\langle f_i, A^*_{\lambda_1}\mu_{\lambda_1}\right\rangle_F &= (\sigma_i^2 + \lambda_2)\left\langle f_i, q_a^*\right\rangle_F \\
\left\langle A_{\lambda_1} f_i, \mu_{\lambda_1}\right\rangle_H &= (\sigma_i^2 + \lambda_2)\left\langle f_i, q_a^*\right\rangle_F \\
\left\langle \sum_{j\geq 1} \sigma_j (h_j \otimes f_j)(f_i), \mu_{\lambda_1}\right\rangle_H &= (\sigma_i^2 + \lambda_2)\left\langle f_i, q_a^*\right\rangle_F \\
\left\langle \sigma_i h_i, \mu_{\lambda_1}\right\rangle_H &= (\sigma_i^2 + \lambda_2)\left\langle f_i, q_a^*\right\rangle_F \\
\frac{\sigma_i \left\langle h_i, \mu_{\lambda_1}\right\rangle_H}{\sigma_i^2 + \lambda_2} &= \left\langle f_i, q_a^*\right\rangle_F .
\end{aligned}$$

By the closed-form of q_a^* (4.14), it is an element in F. So, its decomposition in F is given by

$$q_a^* = \sum_{i\geq 1} \frac{\sigma_i \left\langle h_i, \mu_{\lambda_1} \right\rangle_H}{\sigma_i^2 + \lambda_2} f_i = \sum_{i:\sigma_i \neq 0}^{\infty} \frac{\sigma_i \left\langle h_i, \mu_{\lambda_1} \right\rangle_H}{\sigma_i^2 + \lambda_2} f_i. \tag{4.24}$$

Hence, as $\lambda_2 \longrightarrow 0$, the difference of the two solutions in RKHS norm is given by

$$\begin{aligned} \|q_a^* - q_a^\dagger\|_F &= \left\| \sum_{i:\sigma_i\neq 0}^{\infty} \frac{\sigma_i \left\langle h_i, \mu_{\lambda_1}\right\rangle_H}{\sigma_i^2 + \lambda_2} f_i - \sum_{i:\sigma_i\neq 0}^{\infty} \frac{\left\langle h_i, \mu_{\lambda_1}\right\rangle_H}{\sigma_i} f_i \right\|_F \\ &= \left\| \sum_{i:\sigma_i\neq 0}^{\infty} \left\langle h_i, \mu_{\lambda_1}\right\rangle_H f_i \left(\frac{\sigma_i}{\sigma_i^2 + \lambda_2} - \frac{1}{\sigma_i}\right) \right\|_F \\ &= \left\| \sum_{i:\sigma_i\neq 0}^{\infty} \left\langle h_i, \mu_{\lambda_1}\right\rangle_H f_i \left(\frac{\lambda_2}{\sigma_i^3 + \lambda_2\sigma_i}\right) \right\|_F \\ &= \left\| \sum_{i:\sigma_i\neq 0}^{\infty} \left\langle h_i, \mu_{\lambda_1}\right\rangle_H f_i \left(\frac{1}{\lambda_2^{-1}\sigma_i^3 + \sigma_i}\right) \right\|_F \\ &= \sqrt{\sum_{i:\sigma_i\neq 0}^{\infty} \frac{\left\langle h_i, \mu_{\lambda_1}\right\rangle_H^2}{(\lambda_2^{-1}\sigma_i^3 + \sigma_i)^2}}. \end{aligned}$$

Denote the spectral expansion of μ_{λ_1} in H by $\mu_{\lambda_1} = \sum_{i\geq 1} c_i h_i$, where $\sum_{i\geq 1} c_i^2 < \infty$ by its finite square RKHS norm. Then the difference of the two solutions in RKHS norm is upper bounded by

$$\|q_a^* - q_a^\dagger\|_F^2 = \sum_{i:\sigma_i\neq 0}^{\infty} \frac{\left\langle h_i, \mu_{\lambda_1}\right\rangle_H^2}{(\lambda_2^{-1}\sigma_i^3 + \sigma_i)^2} = \sum_{i:\sigma_i\neq 0}^{\infty} \frac{c_i^2}{\sigma_i^2} \frac{1}{(\lambda_2^{-1}\sigma_i + 1)^2}. \tag{4.25}$$

Since $q_a^\dagger \in F$ implies the finite square RKHS norm of $q_a^\dagger$, its decomposition in F gives $\sum_{i:\sigma_i\neq 0}^{\infty} \frac{c_i^2}{\sigma_i^2} < \infty$. Combining $\lim_{\lambda_2\longrightarrow 0} \frac{1}{(\lambda_2^{-1}\sigma_i+1)^2} = 0$ and $0 < \frac{c_i^2}{\sigma_i^2}\frac{1}{(\lambda_2^{-1}\sigma_i+1)^2} < \frac{c_i^2}{\sigma_i^2}$ for any $i : \sigma_i \neq 0$, we have (4.25) asymptotically converges to 0 as $\lambda_2 \longrightarrow 0$.

□

To get the convergence rate of q_a^* to $q_a^\dagger$ in RKHS norm, we need more assumptions on the decay of the singular values of A_{λ_1}. In fact, if singular values have polynomial decay $\sigma_i \asymp i^{-u}$ such that $\sum_{i:\sigma_i\neq 0}^{\infty} \frac{c_i^2}{\sigma_i^2} \leq 1$ for some positive u, then $\sum_{i:\sigma_i\neq 0}^{\infty} \frac{c_i^2}{\sigma_i^2}\frac{1}{(\lambda_2^{-1}\sigma_i+1)^2} \asymp \lambda_2^{((\frac{2}{u}-2)\wedge 2)}$.

The following corollary is an immediate result of the last step in the proof of Theorem 4.2.

Corollary 4.1 ***(Second moment convergence)***

Given the positive regularization parameter λ_2, then as $\lambda_2 \longrightarrow 0$, the square of the difference between regularized kernel embedded solution q_a^* and the true solution $q_a^\dagger$ in RKHS norm converges to 0.

4.3. Empirical estimator and consistency

The empirical version of the regularized kernel embedded solution $q_a^* \in F$ is given by

$$\widehat{q}_a^* = (\widehat{C}_{(WX)(ZAX)}(\widehat{C}_{(WX)} + \lambda_1 I)^{-1}\widehat{C}_{(ZAX)(WX)} + \lambda_2 I)^{-1}\widehat{C}_{(WX)(ZAX)}(\widehat{C}_{(WX)} + \lambda_1 I)^{-1}\widehat{\mu}_{(WX)}. \tag{4.26}$$

Let $\Phi = (\phi(w_1, x_1), \cdots, \phi(w_n, x_n))^T$, $\Psi = (\psi(z_1, a_1, x_1), \cdots, \psi(z_n, a_n, x_n))^T = (\mathbb{1}_{a_1=a}\tilde{\psi}(z_1, a, x_1), \cdots, \mathbb{1}_{a_n=a}\tilde{\psi}(z_n, a, x_n))^T$, and $\Psi_a = (\psi(z_1, a, x_1), \cdots, \psi(z_n, a, x_n))$. where $(w_i, x_i, a_i, z_i)_{1\leq i\leq n}$ are i.i.d samples from $P_{(W,X,A,Z)}$. Then without centering, the empirical mean embedding and empirical (cross)-covariance embeddings under the samples are given by

$$\widehat{C}_{(WX)} = \frac{1}{n}\sum_{i=1}^{n}\phi(w_i,x_i)\otimes\phi(w_i,x_i) = \frac{1}{n}\Phi^T\Phi, \qquad \widehat{\mu}_{(WX)} = \frac{1}{n}\sum_{i=1}^{n}\phi(w_i,x_i) = \frac{1}{n}\mathbb{1}_n^T\Phi,$$

$$\widehat{C}_{(ZAX)(WX)} = \frac{1}{n}\sum_{i=1}^{n}(\mathbb{1}_{a_i=a}\tilde{\psi}(z_i,a,x_i))\otimes\phi(w_i,x_i) = \frac{1}{n}\sum_{i=1}^{n}\psi(z_i,a_i,x_i)\otimes\phi(w_i,x_i) = \frac{1}{n}\Psi^T\Phi,$$

$$\widehat{C}_{(WX)(ZAX)} = \frac{1}{n}\sum_{i=1}^{n}\phi(w_i,x_i)\otimes(\mathbb{1}_{a_i=a}\tilde{\psi}(z_i,a,x_i)) = \frac{1}{n}\sum_{i=1}^{n}\phi(w_i,x_i)\otimes\psi(z_i,a_i,x_i) = \frac{1}{n}\Phi^T\Psi.$$

Recall that in Section 2.4 we have shown the equivalence of kernels and reproducing kernels as well as the uniqueness between a RKHS and its reproducing kernel. This means the kernels $k(\cdot,\cdot) = \langle\phi(\cdot),\phi(\cdot)\rangle_H$, $l(\cdot,\cdot) = \langle\psi(\cdot),\psi(\cdot)\rangle_F$ and the RKHSs H, F uniquely determine each other. Since the Moore-Aronszajn theorem 2.14 gives the form of elements consisting of the RKHSs induced by kernels, we can subsequently acquire the expressions of $q_a^* \in F$ and $u^* \in H$, where q_a^* is given by

$$q_a^* = \sum_{i=1}^{n}\alpha_i\psi(z_i,a,x_i) = \Lambda^T\Psi_a,\ \Lambda = (\alpha_1,\cdots,\alpha_n)^T. \tag{4.27}$$

In the empirical version, the coefficients should be represented by empirical embeddings. To get the explicit expressions of $\widehat{q}_a^* = \widehat{\Lambda}^T\Psi_a$, we have the following proposition.

Proposition 4.3 ***(Kernel representor of*** $\widehat{q}_a^*$***)***
Denote the Gram matrices and mixed Gram matrix of kernels $k(\cdot,\cdot)$ and $l(\cdot,\cdot)$ by $K = \Phi\Phi^T$, $L = \Psi\Psi^T$, $M = \Phi\Psi^T$ and $M_a = \Phi\Psi_a^T$. Then the regularized kernel embedded solution $q_a^* \in F$ has an kernel representor $\widehat{q}_a^* = \widehat{\Lambda}^T\Psi_a$, where the coefficient $\widehat{\Lambda}$ is given by

$$\widehat{\Lambda} = (KL(K+n\lambda_1 I_n)^{-1}M + n\lambda_2 M_a)^{-1}K(K+n\lambda_1 I_n)^{-1}M\mathbb{1}_n.$$

Proof:
Substituting the kernel embeddings by their matrices versions, the empirical version (4.26) satisfies

$$\begin{aligned}\widehat{q}_a^* &= (\frac{1}{n}\Phi^T\Psi(\frac{1}{n}\Phi^T\Phi+\lambda_1)^{-1}\frac{1}{n}\Psi^T\Phi+\lambda_2)^{-1}\frac{1}{n}\Phi^T\Psi(\frac{1}{n}\Phi^T\Phi+\lambda_1)^{-1}\frac{1}{n}\mathbb{1}_n^T\Phi\\ &= (\Phi^T\Psi(\Phi^T\Phi+n\lambda_1)^{-1}\Psi^T\Phi+n\lambda_2)^{-1}\Phi^T\Psi(\Phi^T\Phi+n\lambda_1)^{-1}\mathbb{1}_n^T\Phi.\end{aligned}$$

Multiply the scalar term $\Phi^T\Psi(\Phi^T\Phi+n\lambda_1)^{-1}\Psi^T\Phi+n\lambda_2$ on both sides and get[2]

$$\begin{aligned}(\Phi^T\Psi(\Phi^T\Phi+n\lambda_1)^{-1}\underline{\Psi^T\Phi}+n\lambda_2)\widehat{q}_a^* &= \Phi^T\Psi(\Phi^T\Phi+n\lambda_1)^{-1}\underline{\mathbb{1}_n^T\Phi}\\ (\Phi^T\Psi\Psi^T\Phi(\Phi^T\Phi+n\lambda_1)^{-1}+n\lambda_2)\Psi_a^T\widehat{\Lambda} &= \Phi^T\underline{\Psi\mathbb{1}_n^T}\Phi(\Phi^T\Phi+n\lambda_1)^{-1} && (\widehat{q}_a^* = \widehat{\Lambda}^T\Psi_a)\\ \Phi^T\Psi\Psi^T(\Phi\Phi^T+n\lambda_1 I_n)^{-1}\Phi\Psi^T\widehat{\Lambda}+n\lambda_2\Psi_a^T\widehat{\Lambda} &= \Phi^T(\Phi\Phi^T+n\lambda_1 I_n)^{-1}\Phi\underline{\Psi\mathbb{1}_n^T}. && \text{(identity (B.1))}\end{aligned}$$

After left multiplying Φ on the both sides, we get

$$\begin{aligned}\Phi\Phi^T\Psi\Psi^T(\Phi\Phi^T+n\lambda_1 I_n)^{-1}\Phi\Psi^T\widehat{\Lambda}+n\lambda_2\Phi\Psi_a^T\widehat{\Lambda} &= \Phi\Phi^T(\Phi\Phi^T+n\lambda_1 I_n)^{-1}\Phi\Psi^T\mathbb{1}_n\\ (KL(K+n\lambda_1 I_n)^{-1}M+n\lambda_2 M_a)\widehat{\Lambda} &= K(K+n\lambda_1 I_n)^{-1}M\mathbb{1}_n.\end{aligned}$$

If $KL(K+n\lambda_1 I_n)^{-1}M+n\lambda_2 M_a$ is invertible, this gives the kernel embedded coefficient

$$\widehat{\Lambda} = (KL(K+n\lambda_1 I_n)^{-1}M+n\lambda_2 M_a)^{-1}K(K+n\lambda_1 I_n)^{-1}M\mathbb{1}_n.$$

[2] The underlined terms are scalars about to shift positions in the coming step.

□

Since $KL(K+n\lambda_1 I_n)^{-1}M+n\lambda_2 M_a$ is not guaranteed to be invertible or well-posed by the sparsity of M, we can use the Tikhonov regularized solution instead, which is given by

$$\widehat{\Lambda}_\lambda := (C^*C+\lambda I)^{-1}C^*b, \tag{4.28}$$

where $C := KL(K+n\lambda_1 I_n)^{-1}M+n\lambda_2 M_a$, $b := K(K+n\lambda_1 I_n)^{-1}M\mathbb{1}_n$. (4.28) converges to the minimum-norm solution as $\lambda \longrightarrow 0$.

Theorem 4.3 *(**Asymptotic convergence of the empirical kernel embedded solution**) Suppose the universal kernel functions $k(\cdot,\cdot)$ and $l(\cdot,\cdot)$ are bounded. Given regularization parameter λ_1 which is a fixed positive number and $\lambda_2 = n^{-\beta}$ with $0 < \beta < \frac{1}{4}$, the empirical regularized kernel embedded solution $\widehat{q}_a^*$ converges to the regularized kernel embedded solution q_a^* of the regularized task (4.12) in RKHS norm with a convergence rate of $\mathcal{O}_p(n^{2\beta-\frac{1}{2}})$, i.e.*

$$E\|\widehat{q}_a^* - q_a^*\|_F = \mathcal{O}(n^{2\beta-\frac{1}{2}}).$$

Proof:

For simplicity, we continue to use the notations given by Muandet et al. [25],

$$\begin{aligned}
C_{\lambda_1} &= C_{(WX)} + \lambda_1 I, & R_{\lambda_1} &= C_{(WX)(ZAX)}C_{\lambda_1}^{-1}C_{(ZAX)(WX)},\\
\widehat{C}_{\lambda_1} &= \widehat{C}_{(WX)} + \lambda_1 I, & \widehat{R}_{\lambda_1} &= \widehat{C}_{(WX)(ZAX)}\widehat{C}_{\lambda_1}^{-1}\widehat{C}_{(ZAX)(WX)}.
\end{aligned}$$

Then the difference of $\widehat{q}_a^*$ and q_a^* in RKHS norm is given by

$$\begin{aligned}
E\|\widehat{q}_a^* - q_a^*\|_F &= E\|(\widehat{R}_{\lambda_1}+\lambda_2 I)^{-1}\widehat{C}_{(WX)(ZAX)}\widehat{C}_{\lambda_1}^{-1}\widehat{\mu}_{(WX)} - (R_{\lambda_1}+\lambda_2 I)^{-1}C_{(WX)(ZAX)}C_{\lambda_1}^{-1}\mu_{(WX)}\|_F\\
&\le E\|(\widehat{R}_{\lambda_1}+\lambda_2 I)^{-1}\widehat{C}_{(WX)(ZAX)}\widehat{C}_{\lambda_1}^{-1}\widehat{\mu}_{(WX)} - (R_{\lambda_1}+\lambda_2 I)^{-1}\widehat{C}_{(WX)(ZAX)}\widehat{C}_{\lambda_1}^{-1}\widehat{\mu}_{(WX)}\|_F\\
&\quad + E\|(R_{\lambda_1}+\lambda_2 I)^{-1}\widehat{C}_{(WX)(ZAX)}\widehat{C}_{\lambda_1}^{-1}\widehat{\mu}_{(WX)} - (R_{\lambda_1}+\lambda_2 I)^{-1}C_{(WX)(ZAX)}C_{\lambda_1}^{-1}\mu_{(WX)}\|_F\\
&=: E[T_1] + E[T_2].
\end{aligned}$$

Part i) Bounding $E[T_1]$:

$$\begin{aligned}
E[T_1] =&E\left[\|(\widehat{R}_{\lambda_1}+\lambda_2 I)^{-1}\widehat{C}_{(WX)(ZAX)}\widehat{C}_{\lambda_1}^{-1}\widehat{\mu}_{(WX)} - (R_{\lambda_1}+\lambda_2 I)^{-1}\widehat{C}_{(WX)(ZAX)}\widehat{C}_{\lambda_1}^{-1}\widehat{\mu}_{(WX)}\|_F\right]\\
\le&E\left[\|\widehat{C}_{(WX)(ZAX)}\widehat{C}_{\lambda_1}^{-1}\widehat{\mu}_{(WX)}\|_F\|(\widehat{R}_{\lambda_1}+\lambda_2 I)^{-1} - (R_{\lambda_1}+\lambda_2 I)^{-1}\|_{\mathcal{L}(F)}\right]\\
=&E\left[\|\widehat{C}_{(WX)(ZAX)}\widehat{C}_{\lambda_1}^{-1}\widehat{\mu}_{(WX)}\|_F\|(\widehat{R}_{\lambda_1}+\lambda_2 I)^{-1}(\widehat{R}_{\lambda_1}+\lambda_2 I)((\widehat{R}_{\lambda_1}+\lambda_2 I)^{-1} - (R_{\lambda_1}+\lambda_2 I)^{-1})\|_{\mathcal{L}(F)}\right]\\
=&E\left[\|\widehat{C}_{(WX)(ZAX)}\widehat{C}_{\lambda_1}^{-1}\widehat{\mu}_{(WX)}\|_F\|(\widehat{R}_{\lambda_1}+\lambda_2 I)^{-1}(R_{\lambda_1}-\widehat{R}_{\lambda_1})(R_{\lambda_1}+\lambda_2 I)^{-1}\|_{\mathcal{L}(F)}\right] \qquad \text{(Identity (B.2))}\\
\le&E\Bigg[\|\widehat{C}_{(WX)(ZAX)}\|_{\mathcal{L}(H,F)}\|\widehat{\mu}_{(WX)}\|_F\|\widehat{C}_{\lambda_1}^{-1}\|_{\mathcal{L}(H)}\|(\widehat{R}_{\lambda_1}+\lambda_2 I)^{-1}\|_{\mathcal{L}(F)}\\
&\qquad\cdot\|(R_{\lambda_1}+\lambda_2 I)^{-1}\|_{\mathcal{L}(F)}\|R_{\lambda_1}-\widehat{R}_{\lambda_1}\|_{\mathcal{L}(F)}\Bigg].
\end{aligned} \tag{4.29}$$

By the assumption that kernels are bounded, we know the reproducing property implies the feature maps are bounded in RKHS norms. We denote the upper bound for $k(\cdot,\cdot)$ over $(\mathcal{W}\times\mathcal{X})\times(\mathcal{W}\times\mathcal{X})$ by t_ϕ^2 and the upper bound for $l(\cdot,\cdot)$ over $(\mathcal{Z}\times\mathcal{A}\times\mathcal{X})\times(\mathcal{Z}\times\mathcal{A}\times\mathcal{X})$ by t_ψ^2. Then, the upper bounds for the feature maps in RKHS norms are given by

$$\begin{aligned}
\|\phi(W,X)\|_H &= \sqrt{k((W,X),(W,X))} \le t_\phi,\ \forall(W,X)\in(\mathcal{W}\times\mathcal{X}),\\
\|\psi(Z,A,X)\|_F &= \sqrt{l((Z,A,X),(Z,A,X))} \le t_\psi,\ \forall(Z,A,X)\in(\mathcal{Z}\times\mathcal{A}\times\mathcal{X}).
\end{aligned}$$

Hence

$$\begin{aligned}
||\widehat{C}_{(WX)(ZAX)}||_{\mathcal{L}(H,F)} &= ||\widehat{C}_{(WX)(ZAX)}||_{HS(H,F)} \leq \frac{1}{n}\sum_{i=1}^{n} ||\phi(w_i,x_i)||_H ||\psi(z_i,a_i,x_i)||_F \leq t_\phi t_\psi, \\
||C_{(WX)(ZAX)}||_{\mathcal{L}(H,F)} &= ||E[\phi(W,X)\otimes\psi(Z,A,X)]||_{HS(H,F)} \leq E||\phi(W,X)\otimes\psi(Z,A,X)||_{HS(H,F)} \\
&\qquad = E\left[||\phi(W,X)||_H ||\psi(Z,A,X)||_F\right] \leq t_\phi t_\psi, \\
||\widehat{\mu}_{(WX)}||_H &\leq \frac{1}{n}\sum_{i=1}^{n} ||\phi(w_i,x_i)||_H \leq t_\phi.
\end{aligned}$$

Next is to bound the operator norm of the inverse regularized operators. Recall by Theorem 2.5 the norm of self-adjoint operator is chosen as the maximum between the absolute values of the infimum of spectrum and maximum of spectrum. Since all the inverse operators considered here are positive semi-definite by Proposition 2.10, their spectra are non-negative by Example 2.8. So, their operator norms are just the the supremum of the spectrum.

$$\begin{aligned}
||\widehat{C}_{\lambda_1}^{-1}||_{\mathcal{L}(H)} &\leq \frac{1}{\inf_\lambda\{\lambda\in\sigma(\widehat{C}_{(WX)}+\lambda_1 I)\}} = \frac{1}{\inf_\lambda\{\lambda\in\sigma(\widehat{C}_{WX})\}+\lambda_1} \leq \frac{1}{\lambda_1}, \\
||(\widehat{R}_{\lambda_1}+\lambda_2 I)^{-1}||_{\mathcal{L}(F)} &\leq \frac{1}{\inf_\lambda\{\lambda\in\sigma(\widehat{R}_{\lambda_1}+\lambda_2 I)\}} = \frac{1}{\inf_\lambda\{\lambda\in\sigma(\widehat{R}_{\lambda_1})\}+\lambda_2} \leq \frac{1}{\lambda_2}, \\
||(R_{\lambda_1}+\lambda_2 I)^{-1}||_{\mathcal{L}(F)} &\leq \frac{1}{\inf_\lambda\{\lambda\in\sigma(R_{\lambda_1}+\lambda_2 I)\}} = \frac{1}{\inf_\lambda\{\lambda\in\sigma(R_{\lambda_1})\}+\lambda_2} \leq \frac{1}{\lambda_2}.
\end{aligned} \tag{4.30}$$

Last, we have

$$\begin{aligned}
||R_{\lambda_1}-\widehat{R}_{\lambda_1}||_{\mathcal{L}(F)} &= ||C_{(WX)(ZAX)}C_{\lambda_1}^{-1}C_{(ZAX)(WX)} - \widehat{C}_{(WX)(ZAX)}\widehat{C}_{\lambda_1}^{-1}\widehat{C}_{(ZAX)(WX)}||_{\mathcal{L}(F)} \\
&\leq ||C_{(WX)(ZAX)}C_{\lambda_1}^{-1}C_{(ZAX)(WX)} - \widehat{C}_{(WX)(ZAX)}C_{\lambda_1}^{-1}\widehat{C}_{(ZAX)(WX)}||_{\mathcal{L}(F)} \\
&\quad + ||\widehat{C}_{(WX)(ZAX)}C_{\lambda_1}^{-1}\widehat{C}_{(ZAX)(WX)} - \widehat{C}_{(WX)(ZAX)}\widehat{C}_{\lambda_1}^{-1}\widehat{C}_{(ZAX)(WX)}||_{\mathcal{L}(F)} \\
&\leq ||C_{(WX)(ZAX)}C_{\lambda_1}^{-1}C_{(ZAX)(WX)} - C_{(WX)(ZAX)}C_{\lambda_1}^{-1}\widehat{C}_{(ZAX)(WX)}||_{\mathcal{L}(F)} \\
&\quad + ||C_{(WX)(ZAX)}C_{\lambda_1}^{-1}\widehat{C}_{(ZAX)(WX)} - \widehat{C}_{(WX)(ZAX)}C_{\lambda_1}^{-1}\widehat{C}_{(ZAX)(WX)}||_{\mathcal{L}(F)} \\
&\quad + ||\widehat{C}_{(WX)(ZAX)}(C_{\lambda_1}^{-1}-\widehat{C}_{\lambda_1}^{-1})\widehat{C}_{(ZAX)(WX)}||_{\mathcal{L}(F)}.
\end{aligned} \tag{4.31}$$

Notice that by matrix identity (B.2), we have

$$C_{\lambda_1}^{-1}-\widehat{C}_{\lambda_1}^{-1} = C_{\lambda_1}^{-1}(\widehat{C}_{\lambda_1}-C_{\lambda_1})\widehat{C}_{\lambda_1}^{-1} = C_{\lambda_1}^{-1}(\widehat{C}_{WX}-C_{WX})\widehat{C}_{\lambda_1}^{-1}. \tag{4.32}$$

Applying (4.32) to (4.31), we get

$$\begin{aligned}
E||R_{\lambda_1}-\widehat{R}_{\lambda_1}||_{\mathcal{L}(F)} \leq E\Bigg[& ||C_{(WX)(ZAX)}||_{\mathcal{L}(H,F)}||C_{\lambda_1}^{-1}||_{\mathcal{L}(H)}||\widehat{C}_{(ZAX)(WX)}-C_{(ZAX)(WX)}||_{HS(F,H)} \\
& + ||C_{\lambda_1}^{-1}||_{\mathcal{L}(H)}||\widehat{C}_{(ZAX)(WX)}||_{\mathcal{L}(F,H)}||\widehat{C}_{(WX)(ZAX)}-C_{(WX)(ZAX)}||_{HS(H,F)} \\
& + ||\widehat{C}_{(WX)(ZAX)}||_{\mathcal{L}(H,F)}||C_{\lambda_1}^{-1}||_{\mathcal{L}(H)}||\widehat{C}_{\lambda_1}^{-1}||_{\mathcal{L}(H)}||\widehat{C}_{(ZAX)(WX)}||_{\mathcal{L}(F,H)}||\widehat{C}_{WX}-C_{WX}||_{HS(H)}\Bigg] \\
\leq E\Bigg[& \frac{t_\phi t_\psi}{\lambda_1}||\widehat{C}_{(ZAX)(WX)}-C_{(ZAX)(WX)}||_{HS(F,H)} + \frac{t_\phi t_\psi}{\lambda_1}||\widehat{C}_{(WX)(ZAX)}-C_{(WX)(ZAX)}||_{HS(H,F)} \\
& + \frac{t_\phi^2 t_\psi^2}{\lambda_1^2}||\widehat{C}_{WX}-C_{WX}||_{HS(H)}\Bigg].
\end{aligned} \tag{4.33}$$

By Theorem 2.16, the $\sqrt{n}$-consistency of cross-covariance embedding, (4.33) is $\mathcal{O}(\frac{1}{\sqrt{n}})$. Hence, combined with the upper bounds of other terms in (4.29), the convergence rate of $E[T_1]$ is $\mathcal{O}(\frac{1}{\lambda_2^2\sqrt{n}})$.

Part ii) Bounding T_2:

$$
\begin{aligned}
E[T_2] &= E\|(R_{\lambda_1}+\lambda_2 I)^{-1}\widehat{C}_{(WX)(ZAX)}\widehat{C}_{\lambda_1}^{-1}\widehat{\mu}_{(WX)} - (R_{\lambda_1}+\lambda_2 I)^{-1}C_{(WX)(ZAX)}C_{\lambda_1}^{-1}\mu_{(WX)}\|_F \\
&\leq E\left[\frac{1}{\lambda_2}\|\widehat{C}_{(WX)(ZAX)}\widehat{C}_{\lambda_1}^{-1}\widehat{\mu}_{(WX)} - C_{(WX)(ZAX)}C_{\lambda_1}^{-1}\mu_{(WX)}\|_F\right] && \text{(By 4.30)} \\
&\leq E\left[\frac{1}{\lambda_2}\|\widehat{C}_{(WX)(ZAX)}\widehat{C}_{\lambda_1}^{-1}\widehat{\mu}_{(WX)} - C_{(WX)(ZAX)}\widehat{C}_{\lambda_1}^{-1}\widehat{\mu}_{(WX)}\|_F\right. \\
&\qquad + \frac{1}{\lambda_2}\|C_{(WX)(ZAX)}\widehat{C}_{\lambda_1}^{-1}\widehat{\mu}_{(WX)} - C_{(WX)(ZAX)}C_{\lambda_1}^{-1}\widehat{\mu}_{(WX)}\|_F \\
&\qquad\qquad\qquad \left. + \frac{1}{\lambda_2}\|C_{(WX)(ZAX)}C_{\lambda_1}^{-1}\widehat{\mu}_{(WX)} - C_{(WX)(ZAX)}C_{\lambda_1}^{-1}\mu_{(WX)}\|_F\right] \\
&\leq E\left[\frac{1}{\lambda_2}\|\widehat{C}_{(WX)(ZAX)} - C_{(WX)(ZAX)}\|_{HS(H,F)}\|\widehat{C}_{\lambda_1}^{-1}\|_{\mathcal{L}(H)}\|\widehat{\mu}_{(WX)}\|_F\right. \\
&\qquad + \frac{1}{\lambda_2}\|C_{(WX)(ZAX)}\|_{\mathcal{L}(H,F)}\|C_{\lambda_1}^{-1}\|_{\mathcal{L}(H)}\|(\widehat{C}_{(WX)} - C_{(WX)})\|_{HS(H,F)}\|\widehat{C}_{\lambda_1}^{-1}\|_{\mathcal{L}(H)}\|\widehat{\mu}_{(WX)}\|_F && \text{(By 4.32)} \\
&\qquad\qquad\qquad \left. + \frac{1}{\lambda_2}\|C_{(WX)(ZAX)}\|_{\mathcal{L}(H,F)}\|C_{\lambda_1}^{-1}\|_{\mathcal{L}(H)}\|\widehat{\mu}_{(WX)} - \mu_{(WX)}\|_F\right] \\
&\leq E\left[\frac{t_\phi}{\lambda_1\lambda_2}\|\widehat{C}_{(WX)(ZAX)} - C_{(WX)(ZAX)}\|_{HS(H,F)} + \frac{t_\phi^2 t_\psi}{\lambda_1^2\lambda_2}\|(\widehat{C}_{(WX)} - C_{(WX)})\|_{HS(H,F)}\right. \\
&\qquad \left. + \frac{t_\phi t_\psi}{\lambda_1\lambda_2}\|\widehat{\mu}_{(WX)} - \mu_{(WX)}\|_F\right]. && (4.34)
\end{aligned}
$$

By Theorem 2.15 and 2.16, the $\sqrt{n}$-consistency of mean embedding and cross-covariance embedding, the convergence rate of $E[T_2]$ is $\mathcal{O}(\frac{1}{\lambda_2\sqrt{n}})$.

Part iii) Asymptotic convergence behavior of $\|\widehat{q}_a^* - q_a^*\|_F$:

The convergence rates of $E[T_1]$ and $E[T_2]$ produce the convergence rate of $\|\widehat{q}_a^* - q_a^*\|_F$, which is given by

$$\mathcal{O}_p(\frac{1}{\sqrt{n}}(\frac{1}{\lambda_2} + \frac{1}{\lambda_2^2})) = \mathcal{O}_p(\frac{1}{\lambda_2^2\sqrt{n}}) = \mathcal{O}_p(n^{2\beta-\frac{1}{2}}).$$

The asymptotic convergence behavior depends on the growth rate of $\lambda_2 = n^{-\beta}$. To make the expectation of the difference between $\widehat{q}_a^*$ and q_a^* in RKHS norm converges to 0 as n goes to infinity, we need

$$2\beta - \frac{1}{2} < 0 \Longrightarrow 0 < \beta < \frac{1}{4}.$$

□

Corollary 4.2 ***(Second moment convergence)***

Suppose kernel functions $k(\cdot,\cdot)$ and $l(\cdot,\cdot)$ are bounded. Given regularization parameter λ_1 which is a fixed positive number and $\lambda_2 = n^{-\beta}$ for $\beta \in (0,1)$ such that $\beta < \frac{1}{4}$, the second moment of the difference between empirical regularized kernel embedded solution $\widehat{q}_a^*$ and regularized kernel embedded solution q_a^* in RKHS norm converges to 0 as n goes to ∞ with a convergence rate of $\mathcal{O}(n^{4\beta-1})$, i.e.

$$E\|\widehat{q}_a^* - q_a^*\|_F^2 = \mathcal{O}(n^{4\beta-1}).$$

Proof:

We continue to use the notations in the proof of Theorem 4.3. Then, by Hölder's inequality, the second moment of the difference in RKHS norm can be upper bounded by

$$E\|\widehat{q}_a^* - q_a^*\|_F^2 \leq E[T_1^2] + E[T_2^2] + E[T_1 T_2] \leq E[T_1^2] + E[T_2^2] + \sqrt{E[T_1^2]E[T_2^2]}. \tag{4.35}$$

where

$$T_1 := \|(\widehat{R}_{\lambda_1} + \lambda_2 I)^{-1}\widehat{C}_{(WX)(ZAX)}\widehat{C}_{\lambda_1}^{-1}\widehat{\mu}_{(WX)} - (R_{\lambda_1} + \lambda_2 I)^{-1}\widehat{C}_{(WX)(ZAX)}\widehat{C}_{\lambda_1}^{-1}\widehat{\mu}_{(WX)}\|_F$$
$$T_2 := \|(R_{\lambda_1} + \lambda_2 I)^{-1}\widehat{C}_{(WX)(ZAX)}\widehat{C}_{\lambda_1}^{-1}\widehat{\mu}_{(WX)} - (R_{\lambda_1} + \lambda_2 I)^{-1}C_{(WX)(ZAX)}C_{\lambda_1}^{-1}\mu_{(WX)}\|_F.$$

By (4.29), combined with the upper bounds of operator norms and empirical embedding norms, $E[T_1^2]$ can be upper bounded by

$$\begin{aligned} E[T_1^2] \leq & E\Bigg[\|\widehat{C}_{(WX)(ZAX)}\|^2_{\mathcal{L}(H,F)}\|\widehat{\mu}_{(WX)}\|_F^2\|\widehat{C}_{\lambda_1}^{-1}\|^2_{\mathcal{L}(H)}\|(\widehat{R}_{\lambda_1} + \lambda_2 I)^{-1}\|^2_{\mathcal{L}(F)} \\ & \qquad \cdot \|(R_{\lambda_1} + \lambda_2 I)^{-1}\|^2_{\mathcal{L}(F)}\|R_{\lambda_1} - \widehat{R}_{\lambda_1}\|^2_{\mathcal{L}(F)}\Bigg] \\ \leq & \frac{t_\phi^4 t_\psi^2}{\lambda_1^2\lambda_2^4} E\|R_{\lambda_1} - \widehat{R}_{\lambda_1}\|^2_{\mathcal{L}(F)}. \end{aligned}$$

By (4.33) and Theorem 2.16, $E\|R_{\lambda_1} - \widehat{R}_{\lambda_1}\|^2_{\mathcal{L}(F)}$ has an upper bound

$$\begin{aligned} E\|R_{\lambda_1} - \widehat{R}_{\lambda_1}\|^2_{\mathcal{L}(F)} \leq E\Bigg[\Bigg(& \frac{t_\phi t_\psi}{\lambda_1}\|\widehat{C}_{(ZAX)(WX)} - C_{(ZAX)(WX)}\|_{HS(F,H)} + \frac{t_\phi t_\psi}{\lambda_1}\|\widehat{C}_{(WX)(ZAX)} - C_{(WX)(ZAX)}\|_{HS(H,F)} \\ & + \frac{t_\phi^2 t_\psi^2}{\lambda_1^2}\|\widehat{C}_{WX} - C_{WX}\|_{HS(H)}\Bigg)^2\Bigg], \end{aligned}$$

which has a convergence rate of $\mathcal{O}(\frac{1}{n})$ after expanded by Hölder's inequality. Hence, we have that the convergence rate of $E[T_1^2]$ is $\mathcal{O}(\frac{1}{\lambda_2^4 n})$.

By (4.34), Theorem 2.15 and 2.16, we have the upper bound of $E[T_2^2]$, which is given by

$$\begin{aligned} E[T_2^2] & \leq E\left[\left(\frac{t_\phi}{\lambda_1\lambda_2}\|\widehat{C}_{(WX)(ZAX)} - C_{(WX)(ZAX)}\|_{HS(H,F)} + \frac{t_\phi^2 t_\psi}{\lambda_1^2\lambda_2}\|(\widehat{C}_{(WX)} - C_{(WX)})\|_{HS(H,F)} + \frac{t_\phi t_\psi}{\lambda_1\lambda_2}\|\widehat{\mu}_{(WX)} - \mu_{(WX)}\|_F\right)^2\right] \\ & = \mathcal{O}(\frac{1}{\lambda_2^2 n}). \end{aligned}$$

combining the convergence rates of $E[T_1^2]$ and $E[T_2^2]$, by (4.35), we have that the second moment of the difference in RKHS norm has a convergence rate of

$$\mathcal{O}(\frac{1}{\lambda_2^4 n} + \frac{1}{\lambda_2^2 n} + \frac{1}{\lambda_2^3 n}) = \mathcal{O}(\frac{1}{\lambda_2^4 n}) = \mathcal{O}(n^{4\beta-1}).$$

When $\beta < \frac{1}{4}$, the second moment of the difference in RKHS norm converges to 0 as n goes to ∞.

□

4.4. Kernel embedded ATE estimator

In this section, we put forward an kernel embedded estimator for the ATE based on the treatment confounding standardization formula (1.7) and PIPW estimator (1.9). which is given by

$$\chi = E[Y(\mathbb{1}_{A=1}q_1^\star(Z,1,X) - \mathbb{1}_{A=0}q_0^\star(Z,0,X))], \tag{4.36}$$

where $q_1^\star$ and $q_0^\star$ are the true treatment confounding bridge functions under treatment $a = 1$ and $a = 0$ respectively. In the previous sections, we have derived a consistent regularized kernel embedded solution q_a^* for

$$E[\mathbb{1}_{A=a}q_a(Z,a,X)|W,X] = 1,$$

which is able to approximate $q_a^\star$ if it is bounded continuous and has the smallest L_2 norm among other solutions.

Definition 4.1 ***(Kernel embedded ATE estimator)***
Under the i.i.d. observations $(y_i, z_i, a_i, x_i)_{1\le i\le n}$, the kernel embedded ATE estimator is given by

$$\widehat{\chi} = \frac{1}{n}\sum_{i=1}^{n} y_i(\mathbb{1}_{a_i=1}\widehat{q}_1^*(z_i,1,x_i) - \mathbb{1}_{a_i=0}\widehat{q}_0^*(z_i,0,x_i)). \tag{4.37}$$

Theorem 4.4 ***(Consistency of the kernel embedded ATE estimator)***
Assume $\lambda_2 = n^{-\beta}$ with $0 < \beta < \frac{1}{4}$ and the following conditions are true.

1. *Var*$(Y(\mathbb{1}_{A=1}q_1^\dagger(Z,1,X) - \mathbb{1}_{A=0}q_0^\dagger(Z,0,X))) < \infty$;
2. $E_{ZAX}[(E_Y[|Y||Z,A,X])^2] < \infty$;
3. $|l(\cdot,\cdot)| \le t_\psi^2$ *everywhere on* $(\mathcal{Z}\times\mathcal{A}\times\mathcal{X})\times(\mathcal{Z}\times\mathcal{A}\times\mathcal{X})$.

Then if the assumption 4.1 holds, under the distribution family satisfying the completeness assumption 1.5, the kernel embedded ATE estimator (4.37) is a consistent estimator of the ATE (4.36), i.e.

$$\lim_{n\to\infty} E|\widehat{\chi} - \chi| = 0.$$

Proof:
In the proof we will use the triangle inequality, Cauchy-Schwartz inequality, Hölder's inequality, Jensen's inequality and the reproducing property for reproducing kernel functions without mentioning.
We consider the following expected difference.

$$\begin{aligned}
&E\left|\left(\frac{1}{n}\sum_{i=1}^{n} y_i(\mathbb{1}_{a_i=1}\widehat{q}_1^*(z_i,1,x_i) - \mathbb{1}_{a_i=0}\widehat{q}_0^*(z_i,0,x_i))\right) - \left(E[Y(\mathbb{1}_{A=1}q_1^\dagger(Z,1,X) - \mathbb{1}_{A=0}q_0^\dagger(Z,0,X))]\right)\right|\\
\le& E\left|\left(\frac{1}{n}\sum_{i=1}^{n} y_i(\mathbb{1}_{a_i=1}\widehat{q}_1^*(z_i,1,x_i) - \mathbb{1}_{a_i=0}\widehat{q}_0^*(z_i,0,x_i))\right) - \left(\frac{1}{n}\sum_{i=1}^{n} y_i(\mathbb{1}_{a_i=1}q_1^\dagger(z_i,1,x_i) - \mathbb{1}_{a_i=0}q_0^\dagger(z_i,0,x_i))\right)\right|\\
&+ E\left|\left(\frac{1}{n}\sum_{i=1}^{n} y_i(\mathbb{1}_{a_i=1}q_1^\dagger(z_i,1,x_i) - \mathbb{1}_{a_i=0}q_0^\dagger(z_i,0,x_i))\right) - \left(E[Y(\mathbb{1}_{A=1}q_1^\dagger(Z,1,X) - \mathbb{1}_{A=0}q_0^\dagger(Z,0,X))]\right)\right|\\
:=& l_1 + l_2.
\end{aligned} \tag{4.38}$$

Part i) Bounding l_1:

$$\begin{aligned}
l_1 :=& E\left|\left(\frac{1}{n}\sum_{i=1}^{n} y_i(\mathbb{1}_{a_i=1}\widehat{q}_1^*(z_i,1,x_i) - \mathbb{1}_{a_i=0}\widehat{q}_0^*(z_i,0,x_i))\right) - \left(\frac{1}{n}\sum_{i=1}^{n} y_i(\mathbb{1}_{a_i=1}q_1^\dagger(z_i,1,x_i) - \mathbb{1}_{a_i=0}q_0^\dagger(z_i,0,x_i))\right)\right|\\
\le& \frac{1}{n}\sum_{i=1}^{n} E\left|\left(y_i\mathbb{1}_{a_i=1}(\widehat{q}_1^*(z_i,1,x_i) - q_1^\dagger(z_i,1,x_i))\right) - \left(y_i\mathbb{1}_{a_i=0}(\widehat{q}_0^*(z_i,0,x_i) - q_0^\dagger(z_i,0,x_i))\right)\right|\\
=& E\left|\left(Y\mathbb{1}_{A=1}(\widehat{q}_1^*(Z,1,X) - q_1^\dagger(Z,1,X))\right) - \left(Y\mathbb{1}_{A=0}(\widehat{q}_0^*(Z,0,X) - q_0^\dagger(Z,0,X))\right)\right|\\
\le& E\left|Y\mathbb{1}_{A=1}(\widehat{q}_1^*(Z,1,X) - q_1^\dagger(Z,1,X))\right| + E\left|Y\mathbb{1}_{A=0}(\widehat{q}_0^*(Z,0,X) - q_0^\dagger(Z,0,X))\right|.
\end{aligned}$$

Notice that

$$\begin{aligned}
E\left|Y\mathbb{1}_{A=a}(\widehat{q}_a^*(Z,a,X) - q_a^\dagger(Z,a,X))\right| &= E_{ZAX}\left[\left|\mathbb{1}_{A=a}(\widehat{q}_a^*(Z,a,X) - q_a^\dagger(Z,a,X))\right| E_Y[|Y||Z,A,X]\right] \\
&= E_{ZAX}\left[\left| \left\langle \widehat{q}_a^* - q_a^\dagger, \psi(Z,A,X) \right\rangle_F \right| E_Y[|Y||Z,A,X]\right] \\
&\leq E_{ZAX}\left[\|\widehat{q}_a^* - q_a^\dagger\|_F \ \underbrace{\|\psi(Z,A,X)\|_F}_{\text{upper bounded by } t_\psi} \ E_Y[|Y||Z,A,X] \right] \\
&\leq \left(E\|\widehat{q}_a^* - q_a^\dagger\|_F^2\right)^{\frac{1}{2}} \ \underbrace{(E_{ZAX}[(E_Y[|Y||Z,A,X])^2])^{\frac{1}{2}} \, t_\psi}_{<\infty}.
\end{aligned} \tag{4.39}$$

To bound (4.39), we need the upper bound of $E\|\widehat{q}_a^* - q_a^\dagger\|_F^2$.

$$\begin{aligned}
E\|\widehat{q}_a^* - q_a^\dagger\|_F^2 &= E\|\widehat{q}_a^* - q_a^* + q_a^* - q_a^\dagger\|_F^2 \\
&= E\|\widehat{q}_a^* - q_a^*\|_F^2 + E\|q_a^* - q_a^\dagger\|_F^2 + 2E\left[\left\langle \widehat{q}_a^* - q_a^*, q_a^* - q_a^\dagger \right\rangle_F\right] \\
&\leq E\|\widehat{q}_a^* - q_a^*\|_F^2 + E\|q_a^* - q_a^\dagger\|_F^2 + 2E\left[\|\widehat{q}_a^* - q_a^*\|_F \|q_a^* - q_a^\dagger\|_F\right] \\
&\leq E\|\widehat{q}_a^* - q_a^*\|_F^2 + E\|q_a^* - q_a^\dagger\|_F^2 + 2\left(E\|\widehat{q}_a^* - q_a^*\|_F^2\right)^{\frac{1}{2}} \left(E\|q_a^* - q_a^\dagger\|_F^2\right)^{\frac{1}{2}}.
\end{aligned}$$

In fact, by Corollary 4.1 and 4.2, as $n \longrightarrow \infty$, the three parts in the upper bound converges to 0, if $0 < \beta < \frac{1}{4}$. Hence, by (4.39), l_1 asymptotically converges to 0 as $n \longrightarrow \infty$.

Part ii) Bounding l_2:

Since we have assumed the variance of $Y(\mathbb{1}_{A=1}q_1^\dagger(Z,1,X) - \mathbb{1}_{A=0}q_0^\dagger(Z,0,X))$ is finite, l_2 can be upper bounded by

$$\begin{aligned}
l_2 := &E\left|\left(\frac{1}{n}\sum_{i=1}^{n} y_i(\mathbb{1}_{a_i=1}q_1^\dagger(z_i,1,x_i) - \mathbb{1}_{a_i=0}q_0^\dagger(z_i,0,x_i))\right) - \left(E[Y(\mathbb{1}_{A=1}q_1^\dagger(Z,1,X) - \mathbb{1}_{A=0}q_0^\dagger(Z,0,X))]\right)\right| \\
\leq &\sqrt{\text{Var}(\frac{1}{n}\sum_{i=1}^{n} y_i(\mathbb{1}_{a_i=1}q_1^\dagger(z_i,1,x_i) - \mathbb{1}_{a_i=0}q_0^\dagger(z_i,0,x_i)))} \\
= &\frac{1}{\sqrt{n}}\sqrt{\text{Var}(Y(\mathbb{1}_{A=1}q_1^\dagger(Z,1,X) - \mathbb{1}_{A=0}q_0^\dagger(Z,0,X)))} \\
= &\mathcal{O}(n^{-\frac{1}{2}}).
\end{aligned}$$

Combining the asymptotic behaviors of l_1 and l_2, we have (4.38) converges to 0 as $n \longrightarrow \infty$. By the definition of universal kernel, $\widehat{\chi}$ should approximate the true ATE χ arbitrarily well. Hence, when $0 < \beta < \frac{1}{4}$, the kernel embedded ATE estimator $\widehat{\chi}$ is a consistent estimator of χ.

□

4.5. A simple numerical test

In this section, we explain the numerical test of the kernel embedded ATE estimator (4.37) and analyze its convergence behaviors in a simple trial.

Data generation The data generation scheme of the numerical test is based on the graphical model 1.2. We assume the data is sequentially generated as follows.

$$\begin{aligned}
U &= \epsilon_1 \\
X &= 0.5U + \epsilon_2 \\
Z &= 0.3U + 0.4X + \epsilon_3 \\
W &= 0.6U + 0.3X + \epsilon_4 \\
A &\sim \text{Bernoulli}(p),\ p = E[\mathbb{1}_{0.3U+0.5X+0.7Z+\epsilon_5>0}] \\
Y^1 &= 2 + 0.7U + 0.2X + 0.8W + \epsilon_6 \\
Y^0 &= 1 + 0.3U + 0.4X + 0.3W + \epsilon_7 \\
Y &= \mathbb{1}_{A=1}Y^1 + \mathbb{1}_{A=0}Y^0,
\end{aligned}$$

where $\epsilon_1, \cdots, \epsilon_7$ are i.i.d. from $\mathcal{N}(0,1)$.

According to the model, since $E[U] = E[X] = E[W] = 0$, the true ATE is given by

$$\chi = E[Y^1 - Y^0] = 1.$$

Kernel function To compute the empirical regularized kernel embedded solution $\widehat{q}_a^*$ (4.27) by Proposition 4.3, we choose the Gaussian RBF kernel to construct the RKHSs. The kernel is given by

$$\begin{aligned}
k((W,X),(W',X')) &:= \exp\{-\gamma\|[WX] - [W'X']\|^2\} \\
\tilde{l}((Z,a,X),(Z',a,X')) &:= \exp\{-\gamma\|[ZX] - [Z'X']\|^2\},\ \forall a \in \{0,1\} \\
l(((Z,A,X),(Z',A',X'))) &:= \mathbb{1}_{A=a}\mathbb{1}_{A'=a}\exp\{-\gamma\|[ZX] - [Z'X']\|^2\},\ \forall a \in \{0,1\},
\end{aligned}$$

where $[\cdot\cdot]$ is a concatenated vector, i.e. $\forall A, B \in \mathbb{R}^n$, then $[AB] = A \oplus B \in \mathbb{R}^{2n}$.

For any $a \in \{0,1\}$, we define the mixed Gram matrices M_a of k and $\tilde{l}$, and M of k and l by matrices whose entries are

$$\begin{aligned}
M_a(i,j) &:= \exp\{-\gamma\|[w_i x_i] - [z_j x_j]\|^2\} \\
M(i,j) &:= \mathbb{1}_{a_j=a}\exp\{-\gamma\|[w_i x_i] - [z_j x_j]\|^2\}.
\end{aligned}$$

The parameter γ represents the flexibility of the kernel, since low γ gives wide Gaussian curves implying smoother basis functions, while large γ gives narrow Gaussian curves and captures fine detail easily but require more care to avoid oscillations or instability.

Hyperparameter selection During the first few attempts, we find that the matrix $KL(K + n\lambda_1 I_n)^{-1}M + n\lambda_2 M_a$ is usually ill-posed, which means its inverse is unstable. So, we turn to the Tikhonov regularized solution of the kernel embedded coefficient $\widehat{\Lambda}$ (4.28). This leads to selection of 4 parameters: $\lambda_1 \in (0,+\infty)$, $\beta \in (0,\frac{1}{4})$, $\lambda \in (0,+\infty)$, $\gamma \in (0,+\infty)$. Since the spaces where the parameters lie in are extremely large to find the optimal values, we choose some potentially optimal points and restrict the searching area of the parameters within

$$\begin{aligned}
\lambda_1 &\in \{0.0001, 0.005, 0.001, 0.05, 0.01, 0.5, 0.4, 0.3, 0.2, 0.1\} \\
\beta &\in \{0.00001, 0.0001, 0.001, 0.01, 0.1, 0.125, 0.15, 0.175, 0.2, 0.225\} \\
\lambda &\in \{0.0001, 0.005, 0.001, 0.05, 0.01, 0.5, 0.4, 0.3, 0.2, 0.1\} \\
\gamma &\in \{0.001, 0.05, 0.01, 0.5, 0.1, 1, 5, 50, 500, 1000\}.
\end{aligned}$$

This still provides 10000 combinations. To decrease the training cost, we randomly choose 100 potential combinations from the 10000 options to select the optimal parameters. Furthermore, here R, the number of repetitions in a single iteration, is set to 100, which is a relatively small value compared to 500 or above in general training tasks. Therefore, we believe when choosing the optimal parameters from the whole parameter space and increasing R, the performance of the estimator will improve.

In a test with a fixed sample size $n = 2000$, we find the optimal parameters in the randomly chosen subset are

$$(\lambda_1, \beta, \lambda, \gamma)_{\chi=1} = (0.1, 0.01, 0.1, 0.001). \tag{4.40}$$

The mean absolute error (MAE) of the difference between $\hat{\chi}$ and χ is given by 0.088 ± 0.008, which is around 8.8% of the true ATE. However, when applying the optimal parameters (4.40) to a different data generation scheme for example changing $Y^1 = 2 + 0.7U + 0.2X + 0.8W + \epsilon_6$ to $Y^1 = 20 + 0.7U + 0.2X + 0.8W + \epsilon_6$ while remaining the others unchanged, we find the convergence behavior is bad. The MAE for the new data generation scheme under the same setting is 10.447 ± 0.061, which is more than a half of the true ATE $\chi = 19$. When the parameter is replaced by

$$(\lambda_1, \beta, \lambda, \gamma)_{\chi=19} = (0.2, 0.225, 0.01, 0.001), \tag{4.41}$$

the MAE decreases to 1.514 ± 0.153 which is around 8.0% to the true ATE. Hence, the behavior of the estimator greatly depends on the choice of parameters. In addition, even a slight change in the data generation scheme will lead to the change on the optimal parameters.

Test on convergence rate We conduct two simple experiments on the convergence of the ATE estimator as sample size grows. The data generation processes are mentioned in the hyperparameter selection with corresponding true ATE 1 and 19. We set the sample size to increase from 100 to 2000 with the interval of 50 between two adjacent points. And we fix the parameters by (4.40) for $\chi = 1$ and (4.41) for $\chi = 19$.

The results are shown in Figure 4.2 and 4.3.

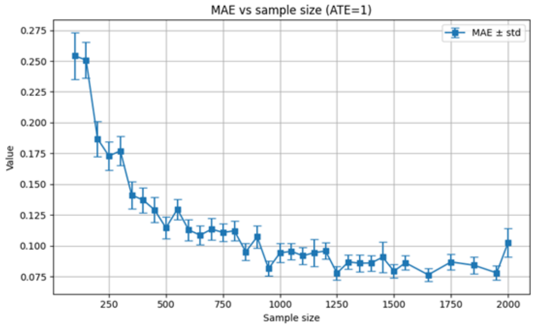


Figure 4.2: Curves of MAE when true ATE is 1 with parameters (4.40).

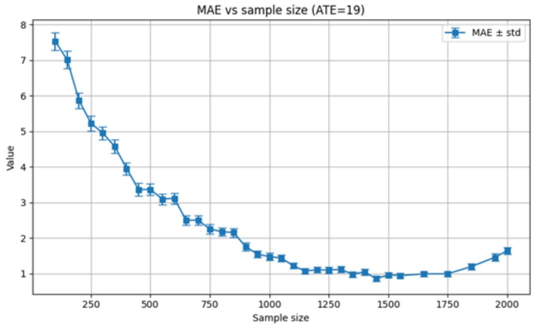


Figure 4.3: Curves of MAE when true ATE is 19 with parameters (4.41).

If we ignore the influence from the discrete sample sizes, the oscillations of the curves in both figures are caused by the relatively low R, which result in large uncertainty in the estimation of the MAE. We believe as R goes larger, the curve will be smoother. That the lowest MAEs for the two cases keep above 0 also matches the bias brought by regularization parameters. Moreover, the result for $\chi = 19$ shows that the estimator maintains the lowest MAE when the sample size lies in the range from 1500 to 1750, and loses its convergence when sample size continues to grow. The potential reason is that the fixed regularization parameter λ can't control the well-posedness of $KL(K + n\lambda_1 I_n)^{-1}M + n\lambda_2 M_a$ anymore when the size of matrix is larger than a certain scale. Hence, we can conclude that the experiments witness the consistency of our ATE estimator, although the regularization parameters bring uncertainty as sample size keeps growing.

5

Discussion

Conclusion This thesis proposes a method to estimate the q-bridge function under nonparametric model in the proximal causal inference framework. The method depends on a series of problem transformations and the existence of a bounded continuous solution. It starts from transforming the original integral equation determining the existence of the q-bridge function to a new form without prior knowledge of propensity score. Then it combines ERM reformulation, minimax problem reformulation through the Fenchel duality and the interchange of minimization and integration, and finally the kernel embedding of means and cross-covariances to acquire a kernel embedded minimax problem about the dual function and q-bridge function. To get the unique kernel embedded solution, we applied Tikhonov regularization to get a regularized kernel embedded solution q_a^* belonging to the RKHS induced by a universal kernel about Z, A, X, which converges to the true kernel embedded solution $q_a^\dagger$ also lying in the same RKHS. We find the convergence of q_a^* to $q_a^\dagger$ in RKHS norm only depends on λ_2, the regularization parameter of the RKHS norm of q-bridge function, but irrelevant to the penalty of the dual function. To prove this convergence, we first transform the original equation problem with an idea similar to the transformation (1.11), then give the kernel embedded version of the new equation problem to derive an equality that $q_a^\dagger$ satisfies. And in the end we use spectral decompositions of compact operators to decompose both q_a^* and $q_a^\dagger$. We also prove the consistency of the empirical regularized kernel embedded solution $\hat{q}_a^*$ to q_a^* in RKHS norm. The corresponding convergence rate is $\mathcal{O}_p(\frac{1}{\lambda_2^2\sqrt{n}})$. The proof mainly depends on the $\sqrt{n}$-consistency of (cross)-covariance embeddings and mean embeddings which are shown in Chapter 2. At last, we propose an ATE estimator $\hat{\chi}$ based on the PIPW estimator and show the convergence of $\hat{\chi}$ to χ in absolute difference when $\lambda_2 = n^{-\beta}$ with $0 < \beta < \frac{1}{4}$. The simple numerical test also helps illustrate the consistency of the estimator $\hat{\chi}$.

Future work Unlike the existing estimators of treatment confounding bridge function [16, 19], which require both bridge functions to exist, our estimator only depends on the existence of the bounded continuous treatment confounding bridge function and thus has potential to be applied to more real-life scenarios. However, although the definition of the universal kernels guarantees the RKHS induced by the universal kernel to be dense in the space of all bounded continuous functions, it is still a problem to understand how well can the kernel embedded solution approximate the true bounded continuous solution. This relies on more investigations on the literature about universal kernel in the future. Moreover, in the numerical test part, we only conducted an experiment about the new proposed estimator on a toy case, which can’t truly reflect the power and deficiency of the estimator. Hence, we expect the opportunity to apply the estimator to some real cases in the future.

A

Gaussian model

In this part, we derive the distribution of U conditional on Z, A, X and the distribution of U conditional on W, A, X under Gaussian model in Example 1.2. For simplicity, we denote $x^T A x$ by $\|x\|_A^2$ and $x^T A y$ by $\langle x, y\rangle_A$.

Part (i): $p(U|Z, A = a, X)$**.**

$$\begin{aligned}
p(U|Z, A = a, X) &= \frac{p(U, Z, a, X)}{p(Z, a, X)} \\
&\propto p(U, Z, a, X) \\
&= \mathcal{N}(\mu_U, \Sigma_U)\mathcal{N}(\mu_X + \gamma_{X|U}U, \Sigma_X)\mathcal{N}(\mu_Z + \gamma_{Z|U}U + \gamma_{Z|X}X, \Sigma_Z) f_A(a|U, Z, X) \\
&\propto \exp\left\{-\frac{1}{2}\|U\|_{\Sigma_U^{-1}}^2 - \frac{1}{2}\|\gamma_{X|U}U\|_{\Sigma_X^{-1}}^2 - \frac{1}{2}\|\gamma_{Z|U}U\|_{\Sigma_Z^{-1}}^2 + \mu(Z, X)^T U + \log f_A(a|U, Z, X)\right\}.
\end{aligned}$$

Hence, $p(U|Z, A = a, X)$ is given by

$$p(U|Z, A = a, X) = C(\theta(Z, a, X)) \exp\left\{-\frac{1}{2}\|U\|_{\Sigma_U^{-1}}^2 - \frac{1}{2}\|\gamma_{Z|U}U\|_{\Sigma_Z^{-1}}^2 - \frac{1}{2}\|\gamma_{Z|U}U\|_{\Sigma_Z^{-1}}^2 + \mu(Z, X)^T U + \log f_A(a|U, Z, X)\right\}, \tag{A.1}$$

where $C(\theta(Z, a, X))$ is the normalizing term and

$$\mu(Z, X) = \Sigma_U^{-1}\mu_U + (X - \mu_X)^T \Sigma_X^{-1}\gamma_{X|U} + (Z - \mu_Z - \gamma_{Z|X})^T \Sigma_Z^{-1}\gamma_{Z|U}.$$

Part (ii): $p(U|W, A = a, X)$**.**

$$\begin{aligned}
p(U|W, A = a, X) &= \frac{p(U, W, a, X)}{p(W, a, X)} \\
&\propto p(U, W, a, X) \\
&= p(W|U, X)p(U, X, a) \\
&= p(W|U, X)\int_{\mathbb{R}^{d_3}} p(U)p(X|U)p(z|U, X) f_A(a|U, z, X)dz \\
&= p(W, U, X)\int_{\mathbb{R}^{d_3}} p(z|U, X) f_A(a|U, z, X)dz.
\end{aligned}$$

$$\begin{aligned}
p(W, U, X) &= \mathcal{N}(\mu_U, \Sigma_U)\mathcal{N}(\mu_X + \gamma_{X|U}U, \Sigma_X)\mathcal{N}(\mu_W + \gamma_{W|U}U + \gamma_{W|X}X, \Sigma_W) \\
&\propto \exp\left\{-\frac{1}{2}\|U\|_{\Sigma_U^{-1}}^2 - \frac{1}{2}\|\gamma_{X|U}U\|_{\Sigma_X^{-1}}^2 - \frac{1}{2}\|\gamma_{W|U}U\|_{\Sigma_W^{-1}}^2 + \mu(W, X)^T U\right\},
\end{aligned}$$

where

$$\mu(W, X) = \Sigma_U^{-1}\mu_U + (X - \mu_X)^T \Sigma_X^{-1}\gamma_{X|U} + (W - \mu_W - \gamma_{W|X})^T \Sigma_W^{-1}\gamma_{W|U}.$$

As for the integral part,

$$\int_{\mathbb{R}^{d_3}} p(z|U,X) f_A(a|z,X) dz \propto \exp\left\{-\frac{1}{2}\|\gamma_{Z|U}U\|^2_{\Sigma_Z^{-1}}\right\} \exp\{\log T(U)\},$$

where

$$T(U) = \int_{\mathbb{R}^{d_3}} \exp\left\{\langle z-\mu_Z-\gamma_{Z|X}X, \gamma_{Z|U}U\rangle_{\Sigma_Z^{-1}} - \frac{1}{2}\|z-\mu_Z-\gamma_{Z|X}X\|^2_{\Sigma_Z^{-1}}\right\} f_A(a|U,z,X) dz. \tag{A.2}$$

The exponent in (A.2) is dominated by the quadratic term. As $\|z\|_2 \to \infty$, when $a = 1$, logistic function $f_A(1|U,z,X)$ is close to 1. The integrand is the exponential function with a quadratic decreasing speed that makes the integral convergent. When $a = 0$, logistic function $f_A(0|U,z,X) \leq \exp\{-\|\mu_A + \gamma_{A|Z}Z + \gamma_{A|X}X + \gamma_{A|U}U\|_2\}$, whose exponent is still linear in z. In this way, the integrand is controlled by an exponential function decaying quadraticall, implying a convergence. Hence $T(U) < \infty$.

From the above deduction, $p(U|W,a,X)$ is given by

$$p(U|W,a,X) = C(\theta(W,a,X)) \exp\left\{-\frac{1}{2}\|U\|^2_{\Sigma_U^{-1}} - \frac{1}{2}\|\gamma_{X|U}U\|^2_{\Sigma_X^{-1}} - \frac{1}{2}\|\gamma_{W|U}U\|^2_{\Sigma_W^{-1}} - \frac{1}{2}\|\gamma_{Z|U}U\|^2_{\Sigma_Z^{-1}} + \mu(W,X)^T U + \log T(U)\right\}, \tag{A.3}$$

where $C(\theta(W,a,X))$ is the normalizing term.

B

Inequality and identities

In this part, we introduce the Cauchy-Schwartz inequality, the Parseval's identity on Hilbert spaces and the Hölder's inequality. Some matrix identities are also included.

Let H be a separable Hilbert space with countable orthonormal basis $(e_i)_{i\geq 1}$. Denote the inner product on H by $\langle \cdot, \cdot \rangle_H$.

Theorem B.1 *(**Cauchy-Schwartz inequality**)*
For any $x, y \in H$, we have $\langle x, y \rangle_H \leq \sqrt{\langle x, x \rangle_H \langle y, y \rangle_H}$. The equality holds if and only if $x = \lambda y$, $\forall \lambda \in \mathbb{R}$.

Proof:
Assume $y \neq 0$ and fix a $k = \frac{\langle x,y \rangle_H}{\langle y,y \rangle_H}$.

$$\begin{aligned} 0 \leq \langle x - ky, x - ky \rangle_H &= \langle x, x \rangle_H - 2k \langle x, y \rangle_H + k^2 \langle y, y \rangle_H \\ &= \langle x, x \rangle_H - 2\frac{\langle x, y \rangle_H}{\langle y, y \rangle_H} \langle x, y \rangle_H + \frac{\langle x, y \rangle_H^2}{\langle y, y \rangle_H}. \end{aligned}$$

Multiplying $\langle y, y \rangle_H$ at both sides, we have

$$\begin{aligned} 0 &\leq \langle x, x \rangle_H \langle y, y \rangle_H - \langle x, y \rangle_H^2 \\ \Longrightarrow \langle x, y \rangle_H &\leq \sqrt{\langle x, x \rangle_H \langle y, y \rangle_H}. \end{aligned}$$

It is clear that the equality holds if and only if x and y are linearly dependent.

□

The Cauchy-Schwarz inequality is a fundamental result in linear algebra and functional analysis. It plays a key role in the theory of Hilbert spaces because it's used to prove the triangle inequality for norms and ensures the continuity of the inner product. One can check section 3.1 of [26] for more details.

Theorem B.2 *([26]**Parseval's identity**)*
For any $x \in H$, the Parseval's identity is given by

$$\|x\|_H^2 = \sum_{i\geq 1} \langle x, e_i \rangle_H^2 .$$

Proof:

The orthonormality of the basis means that $\langle e_i, e_j \rangle_H = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases}$. Under the orthonormal basis $(e_i)_{i\geq 1}$, element $x \in H$ has the representation

$$x = \sum_{i\geq 1} \langle x, e_i \rangle_H \, e_i.$$

Hence,

$$\begin{aligned} \|x\|_H^2 = \langle x, x \rangle_H &= \left\langle \sum_{i\geq 1} \langle x, e_i \rangle_H \, e_i, \sum_{j\geq 1} \langle x, e_j \rangle_H \, e_j \right\rangle_H \\ &= \sum_{i\geq 1} \langle x, e_i \rangle_H \left\langle e_i, \sum_{j\geq 1} \langle x, e_j \rangle_H \, e_j \right\rangle_H \\ &= \sum_{i\geq 1} \langle x, e_i \rangle_H \left\langle e_i, \langle x, e_i \rangle_H \, e_i \right\rangle_H \\ &= \sum_{i\geq 1} \langle x, e_i \rangle_H^2 . \end{aligned}$$

□

The Parseval's identity can be seen as the analog of the Pythagorean theorem on infinite dimensional spaces. This means in infinite dimensions the norm squared is the sum of squared projections onto the basis vectors.

Next we give the Hölder's inequality without proof.

Theorem B.3 ***(Hölder's inequality)***

Let $(\Omega, \mathcal{A}, \mu)$ be a measurable space and let $p, q \in [1, \infty]$ with $\frac{1}{p} + \frac{1}{q} = 1$. Then for all measurable functions $f \in L_p(\Omega, \mu)$ and $g \in L_q(\Omega, \mu)$,

$$\|fg\|_1 \leq \|f\|_p \|g\|_q.$$

The equality holds when $|f|^p$ and $|g|^q$ are linearly independent almost everywhere on Ω, i.e. there exists $\alpha, \beta \in \mathbb{R}$ such that $\alpha|f|^p = \beta|g|^q$.

When μ generalizes to probability measure, the Hölder's inequality is given by

$$E[|fg|] \leq (E[|f|^p])^{\frac{1}{p}} (E[|g|^q])^{\frac{1}{q}}.$$

At last we show two useful matrix identities.

Theorem B.4 ***(Matrix identities)***

Assume $A \in \mathbb{R}^{m\times n}$, $B, C \in \mathbb{R}^{n\times n}$ and $t \in \mathbb{R}$. The following identities hold

$$A(A^TA + tI_n)^{-1} = (AA^T + tI_m)^{-1}A, \tag{B.1}$$

$$B(B^{-1} - C^{-1}) = (C - B)C^{-1}. \tag{B.2}$$

Proof:

i) Identity $A(A^TA + tI_n)^{-1} = (AA^T + tI_m)^{-1}A$:

Multiplying $A^TA + tI_n$ on the right on the both sides, we notice the LHS is A and get the RHS

$$\begin{aligned} (AA^T + tI_m)^{-1}A(A^TA + tI_n) &= (AA^T + tI_m)^{-1}(AA^TA + tA) \\ &= (AA^T + tI_m)^{-1}(AA^T + tI_m)A = A, \end{aligned}$$

which means the two sides are equivalent.

ii) Identity $B(B^{-1} - C^{-1}) = (C - B)C^{-1}$:

Since both sides can be simplified to $I_n - BC^{-1}$, the identity holds.

□

Bibliography

[1] Yasemin Altun and Alex Smola. "Unifying Divergence Minimization and Statistical Inference Via Convex Duality". In: *Learning Theory*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 139–153. URL: https://home.ttic.edu/~altun/pubs/AltSmo-COLT06.pdf.

[2] N. Aronszajn and Nachman. "Theory of reproducing kernels". In: *Transactions of the American mathematical society* 68.3 (1950), pp. 337–404. URL: https://www.ams.org/tran/1950-068-03/S0002-9947-1950-0051437-7/S0002-9947-1950-0051437-7.pdf.

[3] Daryoush Behmardi and Encyeh Dehghan Nayeri. "Introduction of Fréchet and Gâteaux derivative". In: *Applied Mathematical Sciences* 2.20 (2008), pp. 975–980. URL: https://www.m-hikari.com/ams/ams-password-2008/ams-password17-20-2008/behmardiAMS17-20-2008.pdf.

[4] Dimitri Bertsekas. *Convex optimization theory*. Vol. 1. Athena Scientific, 2009.

[5] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006.

[6] Alberto Bressan. *Lecture Notes on Functional Analysis: With Applications to Linear Partial Differential Equations*. American Mathematical Society, 2010. ISBN: 978-0-8218-9406-4. URL: https://www.researchgate.net/profile/Alberto_Bressan/publication/268054345_Lecture_Notes_on_Functional_Analysis_and_Linear_Partial_Differential_Equations/links/555a74ee08ae6fd2d82822c4/Lecture-Notes-on-Functional-Analysis-and-Linear-Partial-Differential-Equations.pdf.

[7] Albrecht Böttcher et al. "Convergence rates for Tikhonov regularization from different kinds of smoothness conditions". In: *Applicable Analysis* 85.5 (2006), pp. 555–578. DOI: 10.1080/00036810500474838. URL: https://doi.org/10.1080/00036810500474838.

[8] Marine Carrasco, Jean-Pierre Florens, and Eric Renault. "Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization". In: *Handbook of econometrics* 6 (2007), pp. 5633–5751. URL: https://publications.ut-capitole.fr/15955/1/handbook20.pdf.

[9] Bo Dai et al. "Learning from conditional distributions via dual embeddings". In: *Artificial Intelligence and Statistics*. PMLR. 2017, pp. 1458–1467. URL: http://proceedings.mlr.press/v54/dai17a/dai17a.pdf.

[10] S. Darolles et al. "Nonparametric Instrumental Regression". In: *Econometrica* 79.5 (2011), pp. 1541–1565. DOI: https://doi.org/10.3982/ECTA6539. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA6539. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA6539.

[11] Xavier D'Haultfoeuille. "On the completeness condition in nonparametric instrumental problems". In: *Econometric Theory* 27.3 (2011), pp. 460–471. DOI: 10.1017/S0266466610000368.

[12] Monika Drewnik, Tomasz Miller, and Zbigniew Pasternak-Winiarski. *Reproducing Kernel Hilbert Space Associated with a Unitary Representation of a Groupoid*. 2021. arXiv: 2102.09585 [math.FA]. URL: https://arxiv.org/abs/2102.09585.

[13] R. M. Dudley. *Real Analysis and Probability*. 2nd ed. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2002.

[14] W. Dana Flanders, Matthew J. Strickland, and Mitchel Klein. *A New Method for Partial Correction of Residual Confounding in Time-Series and other Observational Studies*. 2015. arXiv: 1510.06905 [stat.ME]. URL: https://arxiv.org/abs/1510.06905.

[15] Kenji Fukumizu, Francis R Bach, and Michael I Jordan. “Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces”. In: *Journal of Machine Learning Research* 5.Jan (2004), pp. 73–99. URL: https://www.jmlr.org/papers/volume5/fukumizu04a/fukumizu04a.pdf.

[16] 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 [stat.ML]. URL: https://arxiv.org/abs/2104.02929.

[17] Miguel A. Hernan and James M. Robin. *Causal inference: What if*. Florida: CRC Press, 2024. ISBN: 1420076167.

[18] Yingyao Hu and Ji-Liang Shiu. “Nonparametric identification using instrumental variables: Sufficient conditions for completeness”. In: *Econometric Theory* 34.3 (2018), pp. 659–693. DOI: 10.1017/S0266466617000251.

[19] 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.

[20] Benjamin Kompa et al. *Deep Learning Methods for Proximal Inference via Maximum Moment Restriction*. 2022. arXiv: 2205.09824 [stat.ML]. URL: https://arxiv.org/abs/2205.09824.

[21] Manabu Kuroki and Judea Pearl. “Measurement bias and effect restoration in causal inference”. In: *Biometrika* 101.2 (Mar. 2014), pp. 423–437. ISSN: 0006-3444. URL: https://doi.org/10.1093/biomet/ast066.

[22] E. L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Third. Springer Texts in Statistics. Springer, 2005. ISBN: 0-387-98864-5.

[23] Afsaneh Mastouri et al. *Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restriction*. 2023. arXiv: 2105.04544 [cs.LG]. URL: https://arxiv.org/abs/2105.04544.

[24] Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. “Identifying causal effects with proxy variables of an unmeasured confounder”. In: *Biometrika* 105.4 (Aug. 2018), pp. 987–993. ISSN: 0006-3444. DOI: 10.1093/biomet/asy038. URL: https://doi.org/10.1093/biomet/asy038.

[25] Krikamol Muandet et al. *Dual Instrumental Variable Regression*. 2020. arXiv: 1910.12358 [stat.ML]. URL: https://arxiv.org/abs/1910.12358.

[26] Jan van Neerven. *Functional analysis*. Cambridge Studies in Advanced Mathematics. Cambridge: Cambridge University Press, 2022. ISBN: 9781009232487.

[27] Whitney K. Newey and James L. Powell. “Instrumental Variable Estimation of Nonparametric Models”. In: *Econometrica* 71.5 (2003), pp. 1565–1578. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1555512 (visited on 03/01/2025).

[28] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*. Vol. 317. Springer Science & Business Media, 2009.

[29] Eberhard Schock. “Nonlinear ill-posed equations: Singular value decomposition and the Picard criterion”. In: *Journal of Mathematical Analysis and Applications* 116.1 (1986), pp. 200–208. ISSN: 0022-247X. DOI: https://doi.org/10.1016/0022-247X(86)90052-1. URL: https://www.sciencedirect.com/science/article/pii/0022247X86900521.

[30] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

[31] Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. “A selective review of negative control methods in epidemiology”. In: *Current epidemiology reports* 7.4 (2020), pp. 190–202. URL: https://pubmed-ncbi-nlm-nih-gov.tudelft.idm.oclc.org/33996381/.

[32] Rahul Singh. *Kernel Methods for Unobserved Confounding: Negative Controls, Proxies, and Instruments*. 2023. arXiv: 2012.10315 [stat.ML]. URL: https://arxiv.org/abs/2012.10315.

[33] Alex Smola et al. “A Hilbert space embedding for distributions”. In: *International conference on algorithmic learning theory*. Springer. 2007, pp. 13–31. URL: `https://www.gatsby.ucl.ac.uk/~gretton/papers/SmoGreSonSch07.pdf`.

[34] Tamar Sofer et al. “On negative outcome control of unobserved confounding as a generalization of difference-in-differences”. In: *Statistical science: a review journal of the Institute of Mathematical Statistics* 31.3 (2016), p. 348. URL: `https://pubmed-ncbi-nlm-nih-gov.tudelft.idm.oclc.org/28239233/`.

[35] Le Song, Kenji Fukumizu, and Arthur Gretton. “Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models”. In: *IEEE Signal Processing Magazine* 30.4 (2013), pp. 98–111. URL: `http://www.gatsby.ucl.ac.uk/~gretton/papers/SonFukGre13.pdf`.

[36] I. Steinwart, D. Hush, and C. Scovel. “An Explicit Description of the Reproducing Kernel Hilbert Spaces of Gaussian RBF Kernels”. In: *IEEE Transactions on Information Theory* 52.10 (2006), pp. 4635–4643. DOI: `10.1109/TIT.2006.881713`.

[37] Ingo Steinwart. “On the influence of the kernel on the consistency of support vector machines”. In: *Journal of machine learning research* 2.Nov (2001), pp. 67–93. URL: `https://www.jmlr.org/papers/volume2/steinwart01a/steinwart01a.pdf`.

[38] Xinlu Tan. “Notes on reproducing kernel Hilbert space”. STAT-926. URL: `http://stat.wharton.upenn.edu/~buja/STAT-926/Notes-on-Kernelizing-by-Xin-Lu.pdf`.

[39] Eric J Tchetgen Tchetgen et al. *An Introduction to Proximal Causal Learning*. 2020. arXiv: `2009.10982 [stat.ME]`. URL: `https://arxiv.org/abs/2009.10982`.

[40] Eric Tchetgen Tchetgen. “The control outcome calibration approach for causal inference with unobserved confounding”. In: *American journal of epidemiology* 179.5 (2014), pp. 633–640. URL: `https://pubmed-ncbi-nlm-nih-gov.tudelft.idm.oclc.org/24363326/`.

[41] A. W. van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 1998. ISBN: 9780511802256.

[42] A. W. van der Vaart. “Causality and Graphical Models”. Lecture notes, Advance topics in statistics, EEMCS, TU Delft. 2024.

[43] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[44] 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`.

[45] Cui Yifan et al. “Semiparametric Proximal Causal Inference”. In: *Journal of the American Statistical Association* 119.546 (Mar. 2024), pp. 1348–1359. DOI: `10.1080/01621459.2023.2191817`. URL: `https://doi.org/10.1080/01621459.2023.2191817`.