FUNDAÇÃO GETULIO VARGAS
SCHOOL OF APPLIED MATHEMATICS

CAIO F. LINS PEIXOTO

NONPARAMETRIC INSTRUMENTAL VARIABLE REGRESSION THROUGH STOCHASTIC GRADIENTS AND KERNEL METHODS

Rio de Janeiro
2023

CAIO F. LINS PEIXOTO

NONPARAMETRIC INSTRUMENTAL VARIABLE REGRESSION THROUGH STOCHASTIC GRADIENTS AND KERNEL METHODS

Bachelor's dissertation presented to the School of Applied Mathematics (FGV/EMAp) to obtain the Bachelor's degree in Applied Mathematics.

Area of Study: Nonparametric Regression, Instrumental Variables, Stochastic Optimization, Kernel Methods.

Advisor: Yuri F. Saporito

Rio de Janeiro

2023

Ficha catalográfica elaborada pela BMHS/FGV

Lins, Caio

Nonparametric Instrumental Variable Regression Through Stochastic Gradients and Kernel Methods/ Caio F. Lins Peixoto. – 2023.

49f.

Bachelor's Dissertation (Undergraduate) – School of Applied Mathematics.

Advisor: Yuri F. Saporito.

Includes bibliography.

1. Nonparametric Regression 2. Instrumental Variables 2. Stochastic Optimization I. Saporito, Yuri Fahham II. School of Applied Mathematics. III. Nonparametric Instrumental Variable Regression Through Stochastic Gradients and Kernel Methods

CAIO F. LINS PEIXOTO

NONPARAMETRIC INSTRUMENTAL VARIABLE REGRESSION THROUGH STOCHASTIC GRADIENTS AND KERNEL METHODS

Bachelor’s dissertation presented to the School of Applied Mathematics (FGV/EMAp) to obtain the Bachelor’s degree in Applied Mathematics.

Area of Study: Nonparametric Regression, Instrumental Variables, Stochastic Optimization, Kernel Methods.

Approved on December 15, 2023
By the organizing committee

Yuri F. Saporito
School of Applied Mathematics

Luiz Max Carvalho
School of Applied Mathematics

Eduardo Mendes
FGV/EESP

I dedicate this thesis to ...

Acknowledgements

Thanks, ...

“ Bipedi bopedi bum! ”
Albert Einstein

Abstract

Keywords:

Resumo

Palavras-chave:

List of Figures

List of Tables

List of symbols

$\ker \mathcal{P}$ Kernel of the operator $\mathcal{P}$, that is, the set of vectors h such that $\mathcal{P}[h] = 0$.

Contents

1 Introduction

Remember to cite every person (NEWEY; POWELL, 2003).

2 Instrumental Variable Regression

This chapter provides an introduction to both parametric and nonparametric instrumental variable regression. It is goal is twofold. Firstly, we want to introduce the subject to readers unfamiliar with it. To make the exposition more fluid, we chose to delay the precise definition of all mathematical objects involved until Section 2.4, which deals with the nonparametric approach. The second goal is to precisely state the nonparametric regression problem which will be addressed in the remainder of this thesis. Along the exposition, we will cover the basics of Two Stages Least Squares, the IV regression method most widely employed in practice.

2.1 Endogeneity

We start by introducing the problem of endogenous covariates. The structural equation we consider is the following:

$$Y = h^\star(X) + \varepsilon, \tag{2.1}$$

where X is a vector of explanatory variables, Y is the scalar response, ε is a zero mean noise and the function $h^\star$ is the structural parameter we would like to estimate. The simplest estimation method for this model specification — and, therefore, one we would like to be able to use — is ordinary least squares (OLS), which works by finding, within a given class of functions $\mathcal{H}$, the element which minimizes the mean squared error:

$$\widehat{h} = \underset{h \in \mathcal{H}}{\arg\min}\ \mathbb{E}[(Y - h(X))^2]. \tag{2.2}$$

A reasonable and ample choice for $\mathcal{H}$ is the set of all square-integrable functions of X, that is, such that $\mathbb{E}[h(X)^2] < \infty$. Under this choice, we recover the conditional expectation of Y given X, i.e., $\widehat{h}(X) = \mathbb{E}[Y \mid X]$. Expanding Y through (2.1), we find that $\widehat{h}(X) = h^\star(X) + \mathbb{E}[\varepsilon \mid X]$. Hence, if $\mathbb{E}[\varepsilon \mid X]$ is not identically null, we have introduced bias in our estimation.

This is one of the problems which appear when $\mathbb{E}[\varepsilon \mid X] \neq 0$, or, more generally, when X and ε are correlated in some way. When this happens, we say that X is *endogenous*. There are several causes for endogenous covariates, the most common of which are (WOOLDRIDGE, 2001):

Omitted Variables This means ε can be decomposed as $g^\star(W) + \eta$, where $\mathbb{E}[\eta \mid X, W] = 0$ and X and W are correlated. Hence, when we do not observe W and leave it to the error term, we end up estimating

$$\begin{aligned}\mathbb{E}[Y \mid X] &= h^\star(X) + \mathbb{E}[\varepsilon \mid X] \\ &= h^\star(X) + \mathbb{E}[g^\star(W) + \eta \mid X] \\ &= h^\star(X) + \mathbb{E}[g^\star(W) \mid X]\end{aligned}$$

which is likely different from $h^\star(X)$, if W is correlated with X. For example, if we want to regress a person's wage solely on her number of schooling years (this is X), there are other variables, unaccounted for, which influence both wages and schooling, such as natural ability (this is W). Innately skilled people may tend to be successful in school — and, therefore, pursue higher levels of education — as well as show higher performance in their future jobs, resulting in better wages. Thus, we fail to estimate $h^\star$.

Measurement Error If we are unable to exactly measure one of the covariates, X_k, and instead measure X_k' subject to some stochastic error, by using X_k' in our regression instead of X_k we are delegating to ε some measure of the difference between X_k and X_k'. Depending on how these two variables are related, we may introduce endogeneity. For example, X_k may be a marginal tax rate, but we may only have access to an average tax rate X_k'.

Simultaneity Simultaneity arises when one covariate X_k is determined simultaneously with Y. For example, if we are regressing neighborhood murder rates using the size of the local task force as a covariate, there is a simultaneity problem, since larger murder rates in a place cause a larger task force to be allocated there.

As we have said, bias in the estimation procedure is only one of the problems which arise when there are endogenous covariates. It is well known that the OLS estimate for linear regression fails to be consistent if any one of the covariates is endogenous (WOOLDRIDGE, 2001). To overcome endogeneity a few approaches exist, but by far the one most used by empirical economic research is instrumental variable estimation (WOOLDRIDGE, 2001).

2.2 Instrumental Variables

2.1 Definition An *instrumental variable* for regression problem (2.1) is a random variable Z such that

(i) There is some influence of Z upon X, that is, the marginal distribution of X is not the same as the distribution of X conditioned on Z;

(ii) The conditional mean of ε given Z is almost surely null, i.e., $\mathbb{E}[\varepsilon \mid Z] = 0$.

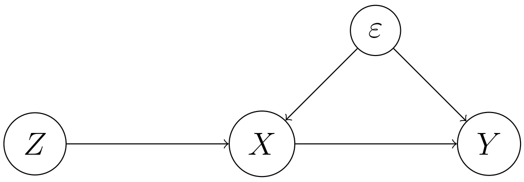

Figure 1 – Causal diagram for equation (2.1), where X is endogenous and Z is an IV.

The idea behind an instrumental variable is that it is exogenous (ii) while still influencing Y through X (i). An exogenous covariate, in contrast to an endogenous one, as a variable that is determined outside of the system described by (2.1).

Condition (ii) is only one of the possible meanings for the statement that Z is exogenous. Two possible alternatives are requiring that Z be (1) independent from, or (2) uncorrelated with ε. Of course, (1) is a much more strict requirement which implies (ii), while (2) is a softer condition, implied by (ii). Independence is almost always impossible to verify in real scenarios, so (1) is not a good option. In contrast, there are situations where condition (2) is enough for ensuring good properties of IV estimators, including one we will present shortly, the linear model (WOOLDRIDGE, 2001). However, in order to prepare grounds for the nonparametric methods that will come later, we chose to use the definition which serves both.

Instrumental variables are also studied in the context of causal inference, where the conditions above are presented differently, in terms of causal diagrams. In this field, instrumental variables are also required that to satisfy a third condition, phrased in terms of the causal diagram describing the relations between variables of interest (HERNÁN; ROBINS, 2020):

(iii) All paths from Z to Y must pass through X, that is, Z *only* influences Y through X.

In this sense, a typical causal diagram for an IV problem is the one in Figure 1.

2.3 Two Stages Least Squares (2SLS)

In this section, we restrict the structural function $h^\star$ in (2.1) to be affine:

$$h^\star(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{d_X} x_{d_X}, \tag{2.3}$$

and assume to have access to a random variable Z, taking values in $\mathbf{R}^{d_Z}$, satisfying conditions 2.1 (i) and (ii), so that Z is a valid instrumental variable. To ease the notation, we augment the variables X and Z to have a 1 as a first coordinate, so we may write $h^\star(X) = \beta^\top X$, where $\beta = (\beta_0, \ldots, \beta_{d_X})$. Their new dimensions are $d'_X = d_X + 1$ and $d'_Z = d_Z + 1$. Our data is then composed of n independent joint samples $\{(X_i, Z_i, Y_i)\}_{i=1}^n$.

Let $\boldsymbol{X} \in \mathbf{R}^{n\times d'_X}$ and $\boldsymbol{Z} \in \mathbf{R}^{n\times d'_Z}$ be the experiment design matrices with 1's in the first column, and let $\boldsymbol{Y} \in \mathbf{R}^n$ be the vector with all observations of Y. Each line of $\boldsymbol{X}$ and $\boldsymbol{Z}$ corresponds to one sample of the vectors X and Z, respectively. The idea of 2SLS is to first perform a regression of X on Z (the *first stage*) and then regress Y on the fitted values $\widehat{X}$ (the *second stage*). In what follows, we will derive this method and give some numerical examples to show its applicability. To avoid misunderstandings during computations, we explicitly state that all vectors are regarded as *column* vectors.

2.3.1 Constructing the estimator

Since we have access to an exogenous covariate, a possible idea is to use this covariate to extract from X a component which is uncorrelated with ε. The simplest way to do this is to perform the *linear orthogonal projection* of X onto Z, that is, to find the matrix P which minimizes the MSE:

$$P = \underset{M\in\mathbf{R}^{d'_X\times d'_Z}}{\arg\min}\ \mathbb{E}[\|X - MZ\|^2] = \mathcal{L}(M).$$

A straightforward computation shows that $\nabla\mathcal{L}(M) = M\mathbb{E}[ZZ^\top] - \mathbb{E}[XZ^\top]$. Since $\mathcal{L}$ is clearly convex, we may find the optimal value by setting the gradient to 0:

$$\nabla\mathcal{L}(P) = 0 \iff P\mathbb{E}[ZZ^\top] = \mathbb{E}[XZ^\top].$$

We now make the hypothesis that $\mathbb{E}[ZZ^\top]$ is invertible. This means that the coordinates of Z are almost surely linearly independent, which is easy to guarantee in practice. With that assumption, we have

$$P = \mathbb{E}[XZ^\top]\mathbb{E}[ZZ^\top]^{-1}. \tag{2.4}$$

Therefore, if we denote the fitted values PZ by $\widehat{X}$, we may write

$$X = \widehat{X} + \eta, \tag{2.5}$$

where $\mathbb{E}[\widehat{X}\eta^\top] = 0$, that is, the residual is orthogonal to the projection.

Now we go back to the structural equation, which, in the linear setting, is the following:

$$Y = X^\top\beta + \varepsilon. \tag{2.6}$$

If we substitute X using equation (2.5), we get

$$Y = \widehat{X}^\top\beta + \eta^\top\beta + \varepsilon.$$

Multiply on the left by $\widehat{X}$ and take expectations to obtain

$$\mathbb{E}[\widehat{X}Y] = \mathbb{E}[\widehat{X}\widehat{X}^\top]\beta + \mathbb{E}[\widehat{X}\eta^\top]\beta + \mathbb{E}[\widehat{X}\varepsilon]. \tag{2.7}$$

We have already established that $\mathbb{E}[\widehat{X}\eta^\top] = 0$. Notice also that

$$\mathbb{E}[\widehat{X}\varepsilon] = \mathbb{E}[PZ\varepsilon] = \mathbb{E}[\mathbb{E}[PZ\varepsilon \mid Z]] = \mathbb{E}[PZ\mathbb{E}[\varepsilon \mid Z]] = 0,$$

by our definition of instrumental variable. Equation (2.7) then reduces to

$$\mathbb{E}[\widehat{X}Y] = \mathbb{E}[\widehat{X}\widehat{X}^\top]\beta.$$

We would like to multiply both sides on the left by $\mathbb{E}[\widehat{X}\widehat{X}^\top]^{-1}$, but we must first check if this matrix is invertible. Expanding we have:

$$\mathbb{E}[\widehat{X}\widehat{X}^\top] = P\mathbb{E}[ZZ^\top]P^\top = \mathbb{E}[XZ^\top]\mathbb{E}[ZZ^\top]^{-1}\mathbb{E}[ZX^\top].$$

Therefore, if we require the rank of the matrix $\mathbb{E}[ZX^\top]$ to be d'_X, we have invertibility of $\mathbb{E}[\widehat{X}\widehat{X}^\top]$. Thus, we need to make two more assumptions: $d'_Z \geq d'_X$, which is equivalent to $d_Z \geq d_X$, and $\operatorname{rk}\mathbb{E}[ZX^\top] = d'_X$. The first assumption is a requirement of the second, and means that we need at least as many exogenous covariates as endogenous covariates in order to identify β. As for the second assumption, it is satisfied if X is sufficiently linearly related to Z (WOOLDRIDGE, 2001). Under these conditions, we have

$$\begin{aligned}\beta &= \mathbb{E}[\widehat{X}\widehat{X}^\top]^{-1}\mathbb{E}[\widehat{X}Y] && (2.8)\\ &= \left[\mathbb{E}[XZ^\top]\mathbb{E}[ZZ^\top]^{-1}\mathbb{E}[ZX^\top]\right]^{-1}\mathbb{E}[XZ^\top]\mathbb{E}[ZZ^\top]^{-1}\mathbb{E}[ZY]. && (2.9)\end{aligned}$$

The 2SLS estimator is then obtained by substituting the expectations by empirical versions, using the data in $\boldsymbol{X}, \boldsymbol{Z}$ and $\boldsymbol{Y}$. The analogue of expression (2.9) would be

$$\begin{aligned}\widehat{\beta} = &\left[\left(\frac{1}{n}\sum_{i=1}^n X_i Z_i^\top\right)\left(\frac{1}{n}\sum_{i=1}^n Z_i Z_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n Z_i X_i^\top\right)\right]^{-1}\\ &\cdot\left(\frac{1}{n}\sum_{i=1}^n X_i Z_i^\top\right)\left(\frac{1}{n}\sum_{i=1}^n Z_i Z_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n Z_i Y_i\right).\end{aligned} \tag{2.10}$$

Notice that all n^{-1} factors cancel out, so this may be equivalently written as [1]

$$\begin{aligned}\widehat{\beta} &= \left[\left(\sum_{i=1}^n X_i Z_i^\top\right)\left(\sum_{i=1}^n Z_i Z_i^\top\right)^{-1}\left(\sum_{i=1}^n Z_i X_i^\top\right)\right]^{-1}\\ &\qquad\cdot\left(\sum_{i=1}^n X_i Z_i^\top\right)\left(\sum_{i=1}^n Z_i Z_i^\top\right)^{-1}\left(\sum_{i=1}^n Z_i Y_i\right)\\ &= \left[\boldsymbol{X}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{X}\right]^{-1}\boldsymbol{X}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{Y}.\end{aligned}$$

Letting $\widehat{\boldsymbol{X}}$ denote $\boldsymbol{X}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$, we have

$$\widehat{\beta} = \left(\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top\right)^{-1}\widehat{\boldsymbol{X}}\boldsymbol{Y},$$

[1] We remind the reader that the transpose sign changes place when writing the 2SLS estimator using data matrices, since each observation of X and Z is a *line* in the corresponding matrix, not a column.

which is the empirical analogue of equation (2.8). This final form makes it clear that the estimator $\widehat{\beta}$ is obtained by first performing one linear regression of $\boldsymbol{X}$ onto $\boldsymbol{Z}$, and then taking the fitted values $\widehat{\boldsymbol{X}}$ and linearly regressing $\boldsymbol{Y}$ on them.

Using equation (2.10), together with the Law of Large Numbers and Slutsky's theorem, one may prove the consistency of $\widehat{\beta}$. Similar inspection allows one to establish asymptotic normality. For further theoretical properties of the 2SLS estimator, we refer the reader to (WOOLDRIDGE, 2001, Chapter 5), this section's main reference.

2.3.2 Numerical examples

We now provide numerical examples to strengthen our intuition about the differences between OLS and 2SLS. We present two scenarios with the same joint distribution for X and Z, in which both are one dimensional. The scenarios also share the same structural function, which we set to be $h^\star(x) = \beta_0 + \beta_1 x$, with $\beta_0 = 0.3$ and $\beta_1 = 0.7$. The difference between both experiments is the distribution of ε, which is mildly codependent with that of X in the first experiment, and strongly codependent in the second.

The data generating process we use for X and Z is the following:

$$\begin{aligned}
&\delta_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \quad i = 1, 2;\\
&Z = \Phi(\delta_1);\\
&X = \Phi(\rho\delta_1 + \sqrt{1-\rho^2}\delta_2),
\end{aligned}$$

where Φ is the cumulative distribution function of the standard normal and $\rho \in (0,1)$ is fixed at 0.8. For the first scenario, we generate ε as follows:

$$\begin{aligned}
&\delta_3 \sim \mathcal{N}(0,1), \quad \delta_3 \perp\!\!\!\perp (\delta_1, \delta_2);\\
&\varepsilon = \sigma \cdot (\eta\delta_2 + \sqrt{1-\eta^2}\delta_3),
\end{aligned}$$

where $\sigma > 0$ and $\eta \in (0,1)$ are set to be 0.1 and 0.6, respectively. For the second scenario, we introduce more interdependence between X and ε:

$$\varepsilon = \sigma \cdot \left(\left[\eta\delta_2 + \sqrt{1-\eta^2}\delta_3 \right] + C_b(\delta_2 - b)^+ - C_a(\delta_2 - a)^- \right).$$

Here, σ and η have the same values as before. The additional terms are $C_b = 6, b = 0.7, C_a = 2$ and $a = 0.3$. Finally, in both formulations we have $Y = h^\star(x) + \varepsilon$.

The results of the experiments are in Figure 2. We can see that in both of them the 2SLS estimate is closer to the true values of β_0 and β_1 than the OLS estimate. As expected, the OLS estimate does not take endogeneity into account and, therefore, incorporates some bias into the final estimates. This effect worsens as the endogeneity becomes stronger. An important observation, which is not visible in the figure, is that, as the number of observations grows, the OLS estimate drifts further from the true values, while the 2SLS estimate becomes closer to them.

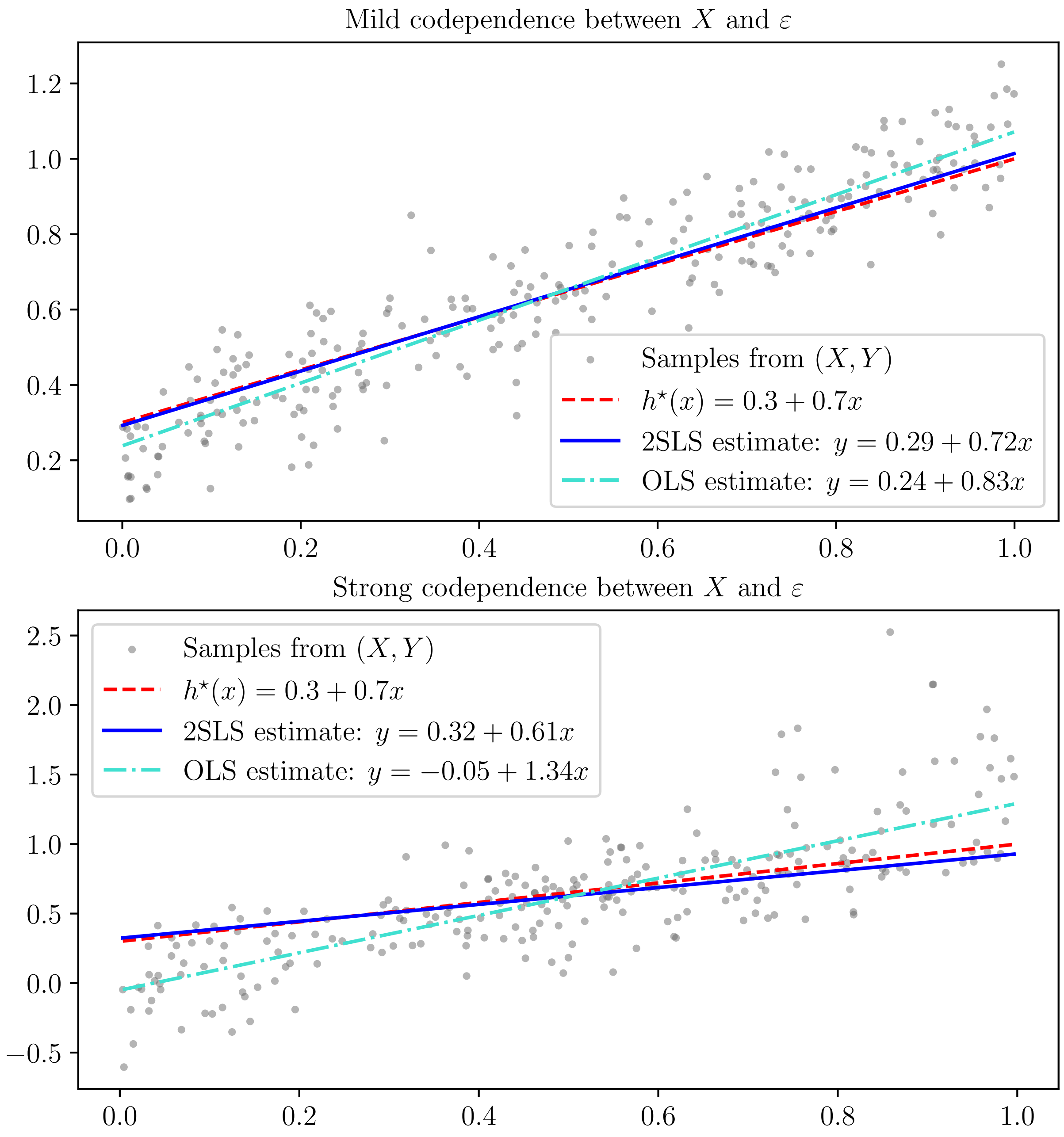

Figure 2 – Comparison between OLS and 2SLS under two different levels of endogeneity.

2.4 Nonparametric Instrumental Variable Regression

In nonparametric regression, we do not specify *a priori* a finite dimensional parametric form for the structural function (such as restricting it to be affine), and so we allow our search space to potentially be infinite dimensional. However, in doing this, we must still precisely define the infinite dimensional space where the solution will be searched for. Hence, we start by precisely defining the nonparametric regression problem given by (2.1)

2.4.1 Problem specification

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be the underlying probability space. Assume that $X : (\Omega, \mathcal{A}) \to (\mathbf{R}^{d_X}, \mathcal{B}(\mathbf{R}^{d_X}))$ and $\varepsilon : (\Omega, \mathcal{A}) \to (\mathbf{R}, \mathcal{B}(\mathbf{R}))$ are measurable[2] and, furthermore, that $\varepsilon \in L^1(\Omega, \mathcal{A}, \mathbb{P})$ with $\mathbb{E}[\varepsilon] = 0$. We also assume that $\mathbb{E}[\varepsilon \mid X]$ is *not* almost surely null and, hence, X is endogenous. Denote by $\mathbb{P}_X$ the distribution of the random variable[3] X, that is, the pushforward measure $\mathbb{P} \circ X^{-1}$ defined on $\mathcal{B}(\mathbf{R}^{d_X})$. We write $L^2(X)$ as a shorthand for the space $L^2(\mathbf{R}^{d_X}, \mathcal{B}(\mathbf{R}^{d_X}), \mathbb{P}_X)$ of real and square integrable (equivalence classes of) measurable functions defined on the measure space $(\mathbf{R}^{d_X}, \mathcal{B}(\mathbf{R}^{d_X}), \mathbb{P}_X)$. It is important to recall that the inner product and norm in $L^2(X)$ are given by $\langle h, g\rangle_{L^2(X)} = \mathbb{E}[h(X)g(X)]$ and $\|h\|_{L^2(X)}^2 = \langle h, h\rangle_{L^2(X)} = \mathbb{E}[h(X)^2]$.

We assume there exists $h^\star \in L^2(X)$ such that (2.1) holds, that is, $Y = h^\star(X) + \varepsilon$. Finally, we assume there exists a random variable $Z : (\Omega, \mathcal{A}) \to (\mathbf{R}^{d_Z}, \mathcal{B}(\mathbf{R}^{d_Z}))$ such that Z qualifies as an instrumental variable, i.e., Z satisfies conditions 2.1 (i) and (ii). We define $\mathbb{P}_Z$ and $L^2(Z)$ in an manner analogous to $\mathbb{P}_X$ and $L^2(X)$. Our goal is to estimate $h^\star$ based on i.i.d. samples from the joint distribution of X, Z and Y.

2.4.2 Identification

An important question to ask after specifying the problem is whether the function $h^\star$ is identified. The answer is negative without further assumptions, which will be presented in this subsection. This discussion was inspired on (NEWEY; POWELL, 2003, Section 2).

Suppose there exists $\delta \in L^2(X)$ such that $\delta \neq 0$, but $\mathbb{E}[\delta(X) \mid Z] = 0$. Without loss of generality, we can assume[4] $\delta(X) \neq \mathbb{E}[\varepsilon \mid X]$. Defining $g \triangleq h^\star + \delta$ and $\eta \triangleq \varepsilon - \delta(X)$, we have

$$Y = g(X) + \eta,$$

where $\mathbb{E}[\eta \mid Z] = 0$ and $\mathbb{E}[\eta \mid X] \neq 0$. Hence, $g \neq h^\star$ but they are indistinguishable from the data generating process' perspective. Reciprocally, suppose that the only member of $L^2(X)$ which has null mean conditioned on Z is the null function. Then, given $g \in L^2(X)$ such that

$$Y = g(X) + \eta$$

with $\mathbb{E}[\eta \mid Z] = 0$, we have

$$0 = (g - h^\star)(X) + \eta - \varepsilon.$$

[2] We denote by $\mathcal{B}(\mathbf{R}^k)$ the Borel σ-algebra in $\mathbf{R}^k$.

[3] We use the term "random variable" when referring to scalar or vector valued measurable functions defined on $(\Omega, \mathcal{A})$.

[4] If it happens to be the case that $\mathbb{E}\left[\mathbb{E}[\varepsilon \mid X] \mid Z\right] = 0$ (which is *not* implied by our assumptions so far), we can simply take $\delta(X) = \lambda \mathbb{E}[\varepsilon \mid X]$, for some $\lambda \in \mathbf{R} \setminus \{0, 1\}$. Since, by hypothesis, $\mathbb{E}[\varepsilon \mid X] \neq 0$, this satisfies our requirements and is different from $\mathbb{E}[\varepsilon \mid X]$.

Conditioning on Z, we get

$$\mathbb{E}[(g - h^\star)(X) \mid Z] = 0,$$

which, by assumption, implies $h^\star = g$.

Therefore, a necessary and sufficient condition for identification of our regression problem is the following:

Assumption (Identification) If $\delta \in L^2(X)$ satisfies $\mathbb{E}[\delta(X) \mid Z] = 0$, then $\delta = 0$.

This condition has an interpretation in terms of the conditional expectation operator, which will be a key object in the construction of our estimator for $h^\star$. Let $h \in L^2(X)$. Notice that, by Jensen's inequality,

$$\mathbb{E}[(\mathbb{E}[h(X) \mid Z])^2] \leq \mathbb{E}[\mathbb{E}[h(X)^2 \mid Z]] = \mathbb{E}[h(X)^2] < +\infty. \tag{2.11}$$

Furthermore, since $\mathbb{E}[h(X) \mid Z]$ is a $\sigma(Z)$-measurable[5] random variable, by the Doob-Dynkin Lemma there exists a measurable function $f_h : (\mathbf{R}^{d_Z}, \mathcal{B}(\mathbf{R}^{d_Z})) \to (\mathbf{R}, \mathcal{B}(\mathbf{R}))$ such that

$$\mathbb{E}[h(X) \mid Z] = f_h(Z).$$

Under these conditions, we write $\mathbb{E}[h(X) \mid Z = z]$ for $f_h(z)$. The computation in (2.11) shows that $f_h \in L^2(Z)$ for every $h \in L^2(X)$. Therefore, we can define the operator $\mathcal{P} : L^2(X) \to L^2(Z)$ given by $\mathcal{P}[h] = f_h = \mathbb{E}[h(X) \mid Z = \cdot]$. This operator, called the *conditional expectation operator*, is clearly linear and, again by (2.11), also bounded, satisfying $\|\mathcal{P}\|_{\text{op}} \leq 1$. The identification assumption thus amounts to saying that the kernel of $\mathcal{P}$ is trivial, i.e., $\mathcal{P}$ is injective.

It is hard to quantify how restrictive this condition is for arbitrary X and Z, so we analyze it in a more familiar setting, the exponential family. We will use a classic completeness result for statistics in this family of distributions to reformulate the identification assumption in terms of more familiar objects. We first define completeness:

2.2 Definition (LEHMANN, 1959) We say that a family $\mathscr{P}$ of probability distributions on a measurable space $(E, \mathcal{E})$ is *complete* if

$$\int_E f(x)\ P(\mathrm{d}x) = 0 \quad \text{for all} \quad P \in \mathscr{P}$$

implies $f(x) = 0$ $\mathscr{P}$-a.e[6].

Then, a remarkable fact about the exponential family is the completeness of the natural statistics under a mild condition on the set of parameters:

[5] We denote by $\sigma(Z)$ the smallest σ-algebra in Ω with respect to which Z is measurable.

[6] We say that a statement $Q(x)$ is true $\mathscr{P}$-a.e. if there exists a set $N \in \mathcal{E}$ such that $Q(x)$ is true for $x \in E \setminus N$ and $P(N) = 0$ for all $P \in \mathscr{P}$.

2.3 Theorem (LEHMANN, 1959) Let Ξ be a subset of an Euclidian space with nonempty interior. Let X be a random vector with distribution P^θ parametrized by $\theta \in \Xi$ in the following manner:

$$P^\theta(\mathrm{d}x) = C(\theta) \exp \left\{ \sum_{i=1}^{s} \theta_i T_i(x) \right\} \mu(\mathrm{d}x),$$

where μ is the underlying measure. Then, the family $\mathscr{P}_T$, formed by the distributions of the random vector $T(X) = (T_1(X), \dots, T_s(X))$ as θ ranges through Ξ, is complete.

Using this result and under suitable hypotheses, we can reformulate the identification condition.

2.4 Theorem For $z \in \mathbf{R}^{d_z}$, let $\mathbb{Q}_z : \mathcal{B}(\mathbf{R}^{d_X}) \to [0,1]$ denote the conditional distribution of X given $Z = z$. Assume there exists $U \in \mathcal{B}(R^{d_Z})$ such that $\mathbb{P}_Z(U) = 1$ and for all $z \in U$ we have

$$\mathbb{Q}_z(\mathrm{d}x) = C(z) \exp(\alpha(z)^\top T(x))\ \mu(\mathrm{d}x)$$

for an underlying measure μ on $\mathcal{B}(\mathbf{R}^{d_X})$ and some functions $\alpha : \mathbf{R}^{d_z} \to \mathbf{R}^s$ and $T : \mathbf{R}^{d_x} \to \mathbf{R}^s$. Assume that T is injective and that the image of α restricted to U contains an open set. Then, $h^\star$ is identified.

Proof. Taking $\Xi = \alpha(U)$ and $\theta = \alpha(z)$, we see that the hypotheses of theorem 2.3 are satisfied and, hence,

$$\mathscr{P} \triangleq \left\{ \mathbb{Q}_z \circ T^{-1} : z \in U \right\}$$

is a complete family of probability distributions. Let $h \in L^2(X)$ be such that $\mathcal{P}[h] = 0$, i.e., $\mathbb{E}[h(X) \mid Z] = 0$ $\mathbb{P}_Z$-a.s. This means that the function

$$z \longmapsto \int_{\mathbf{R}^{d_x}} h(x)\ \mathbb{Q}_z(\mathrm{d}x)$$

is null $\mathbb{P}_Z$-a.s. Without loss of generality, we may assume that its null on U. But notice that, since T is injective, we can rewrite this integral as

$$\begin{aligned} 0 = \int_{\mathbf{R}^{d_x}} h(x)\ \mathbb{Q}_z(\mathrm{d}x) &= \int_{\mathbf{R}^{d_x}} (h \circ T^{-1})(T(x))\ \mathbb{Q}_z(\mathrm{d}x) \\ &= \int_{\mathbf{R}^s} (h \circ T^{-1})(t)\ (\mathbb{Q}_z \circ T^{-1})(\mathrm{d}t) \end{aligned}$$

for all $z \in U$ and some left inverse T^{-1} of T. By completeness of $\mathscr{P}$, this implies $h \circ\ T^{-1}(t) = 0$ $\mathscr{P}$-a.s. which, in turn, means that for all $z \in U$ we have

$$\begin{aligned} 1 = (\mathbb{Q}_z \circ T^{-1})[(h \circ T^{-1})(t) = 0] &= \mathbb{Q}_z[(h \circ T^{-1})(T(x)) = 0] \\ &= \mathbb{Q}_z[h(x) = 0]. \end{aligned}$$

Now, by the definition of conditional probability we have

$$\begin{aligned}\mathbb{P}_X[h(x) = 0] &= \int_{\mathbf{R}^{d_z}} \mathbb{Q}_z[h(x) = 0]\ \mathbb{P}_Z(\mathrm{d}z) \\ &= \int_U \mathbb{Q}_z[h(x) = 0]\ \mathbb{P}_Z(\mathrm{d}z) \\ &= 0.\end{aligned}$$

Therefore, h is the null function, which means $h^\star$ is identified. □

Maybe say something more about identification and conclude the chapter a bit better?

3 NPIV Regression and Linear Inverse Problems

Having defined what we mean by nonparametric regression instrumental variable (NPIV) regression, we now present two well established approaches, namely the ones in (DAROLLES et al., 2011) and (NEWEY; POWELL, 2003), the first of which will serve as a starting point for our method. We start by connecting nonparametric instrumental variable regression with inverse problems.

3.1 NPIV regression as an ill-posed linear inverse problem

Recalling the notation presented in section 2.4, we want to find $h^\star \in L^2(X)$ which satisfies

$$Y = h^\star(X) + \varepsilon, \tag{3.1}$$

where $\mathbb{E}[\varepsilon \mid Z] = 0$. Letting $r(Z) \triangleq \mathbb{E}[Y \mid Z]$ and assuming that Y has a finite second moment, we have $r \in L^2(Z)$, so that (3.1) is equivalent to

$$r = \mathcal{P}[h^\star], \tag{3.2}$$

where $\mathcal{P} : L^2(X) \to L^2(Z)$ is the conditional expectation operator. Assume that the joint distribution of (X, Z) is absolutely continuous with respect to Lebesgue measure in $\mathbf{R}^{d_X + d_Z}$, so that we may rewrite (3.2) as

$$r(z) = \mathbb{E}[h^\star(X) \mid Z = z] = \int_{\mathbf{R}^{d_X}} h^\star(x) p_{X|Z}(x \mid z) \, \mathrm{d}x, \tag{3.3}$$

where $p_{X|Z}(x \mid z)$ is the conditional density of X given Z. This is a Fredholm integral equation of the first kind (KRESS, 1989) which, due to the nature of our problem, will most likely be ill-posed. We dedicate some space to make this statement precise.

Equation (3.2) would describe a well-posed problem if the operator $\mathcal{P}$ were invertible and the inverse $\mathcal{P}^{-1}$, continuous. There are three ways in which these conditions may be violated. One of them was already discussed in subsection 2.4.2, the identification problem. It corresponds to non-injectivity of $\mathcal{P}$, that is, to $\ker \mathcal{P} \neq \{0\}$. A possible correction for this problem is to look for the least norm solution.

Another violation happens if $\mathcal{P}$ is not surjective, which leaves the possibility of $r \notin \operatorname{Im}(\mathcal{P})$. A way to bypass this difficulty is to project r onto the subspace $\operatorname{Im}(\mathcal{P})$ of $L^2(Z)$ and find the inverse image of this projection. However, the orthogonal projection onto a subspace is only well defined for $\overline{\operatorname{Im}(\mathcal{P})}$, since we need the subspace to be closed, which may not be the case for $\operatorname{Im}(\mathcal{P})$. Hence, we would need to assume that the projection of r onto $\overline{\operatorname{Im}(\mathcal{P})}$ belongs to $\operatorname{Im}(\mathcal{P})$. The methods we will analyze assume that the model (3.1) is correctly specified and, therefore, that $r \in \operatorname{Im}(\mathcal{P})$, so this will not be a concern for us. The interested reader may consult (CARRASCO; FLORENS; RENAULT, 2007) for ways to proceed without this assumption.

The last way for a problem such as (3.1) to be ill-posed is to have $\mathcal{P}$ injective and $r \in \operatorname{Im}(\mathcal{P})$, but $\mathcal{P}^{-1} : \operatorname{Im}(\mathcal{P}) \to L^2(X)$ discontinuous. By the open mapping theorem (RUDIN, 1991), if $\operatorname{Im}(\mathcal{P})$ is closed, then $\mathcal{P} : L^2(X) \to \operatorname{Im}(\mathcal{P})$ is an open map and the inverse is automatically continuous. Hence, we must require that $\operatorname{Im}(\mathcal{P})$ is not closed. A prototypical example for this situation is when $\mathcal{P}$ is a compact operator with infinite dimensional range. If $\operatorname{Im}(\mathcal{P})$ were closed, since $\mathrm{id}_{\operatorname{Im}(\mathcal{P})} = \mathcal{P}^{-1} \circ \mathcal{P}$ it would be a compact operator, which would imply compactness of the closed unit ball of $\operatorname{Im}(\mathcal{P})$ and, hence, $\dim \operatorname{Im}(\mathcal{P}) < \infty$ (CARRASCO; FLORENS; RENAULT, 2007).

In light of this discussion, we now take the opportunity to study additional properties of the operator $\mathcal{P}$, in particular, with the goal of providing a sufficient condition for it to be compact in terms of the distribution of X and Z. Notice that

$$\begin{aligned}
\mathcal{P}[h](z) = \mathbb{E}[h(X) \mid Z = z] &= \int_{\mathbf{R}^{d_X}} h(x) p_{X|Z}(x \mid z) \,\mathrm{d}x \\
&= \int_{\mathbf{R}^{d_X}} h(x) \frac{p_{X,Z}(x,z)}{p_X(x)p_Z(z)} p_X(x) \,\mathrm{d}x \\
&= \mathbb{E}\left[h(X) \frac{p_{X,Z}(X,z)}{p_X(X)p_Z(z)} \right] \\
&= \mathbb{E}\left[h(X)\Phi(X,z) \right]. \qquad (3.4)
\end{aligned}$$

Hence, $\mathcal{P}$ is an integral operator with kernel[1]

$$\Phi(x,z) = \frac{p_{X,Z}(x,z)}{p_X(x)p_Z(z)}.$$

Furthermore, since

$$\begin{aligned}
\langle \mathcal{P}[h], g \rangle_{L^2(Z)} &= \mathbb{E}\left[\mathbb{E}[h(X) \mid Z] g(Z) \right] \\
&= \mathbb{E}\left[h(X) g(Z) \right] \\
&= \mathbb{E}\left[h(X) \mathbb{E}[g(Z) \mid X] \right] \\
&= \langle h, \mathbb{E}[g(Z) \mid X = \cdot] \rangle_{L^2(X)},
\end{aligned}$$

by the defining inequality of the adjoint operator $\mathcal{P}^* : L^2(Z) \to L^2(X)$ we have

$$\mathcal{P}^*[g](X) = \mathbb{E}[g(Z) \mid X].$$

Proceeding in a manner analogous to the one which got us (3.4), we may find that $\mathcal{P}^*$ is also an integral operator with the same kernel Φ of $\mathcal{P}$.

[1] Not to be confused with $\ker \mathcal{P}$, a subspace of $L^2(X)$.

This observation is useful because, from (CARRASCO; FLORENS; RENAULT, 2007, theorem 2.34) we know that if the kernel Φ satisfies

$$\int_{\mathbf{R}^{d_X}} \int_{\mathbf{R}^{d_Z}} \Phi(x,z)^2 p_X(x) p_Z(z) \, \mathrm{d}z\mathrm{d}x < \infty, \tag{3.5}$$

then $\mathcal{P}$ (and, consequently, $\mathcal{P}^*$) is a Hilbert-Schmidt operator and, in particular, compact.

It is also appropriate to note that in the case where X and Z share any coordinates, the operator $\mathcal{P}$ *cannot* be compact. To see this, suppose that $X = (X_1, W)$ and $Z = (Z_1, W)$ for some random vector W. Denote by $L^2(W)$ the subspace of $L^2(X)$ (and $L^2(Z)$) of functions which depend only on W. Notice that, if $f \in L^2(W)$, then

$$\mathcal{P}[f](Z) = \mathcal{P}[f](Z_1, W) = \mathbb{E}[f(W) \mid W, Z_1] = f(W) \quad \text{i.e.} \quad \mathcal{P}[f] = f.$$

Hence, the image of the unit ball in $L^2(X)$ by $\mathcal{P}$ contains the unit ball of $L^2(W)$. Since the norm $\|\cdot\|_{L^2(Z)}$ coincides with $\|\cdot\|_{L^2(W)}$ in $L^2(W)$, and since $L^2(W)$ is closed as a subspace of $L^2(Z)$, then the unit ball in $L^2(W)$ is compact for $\|\cdot\|_{L^2(W)}$ if and only if it is compact for $\|\cdot\|_{L^2(Z)}$. Then, since $L^2(W)$ is infinite dimensional, its unit ball is not compact, which means that the closure of image of the unit ball in $L^2(X)$ by $\mathcal{P}$ is not compact and, hence, $\mathcal{P}$ is not a compact operator.

We remark that, aside from requiring $r \in \operatorname{Im}(\mathcal{P})$, we leave the possibility of any other form of ill-posedness to be present in our problem until stated otherwise.

3.2 NPIV regression through Tikhonov regularization

In (DAROLLES et al., 2011), the authors assume identification and also that the joint density $p_{X,Z}$ is dominated by the product of marginals $p_x \cdot p_Z$. This is the same as demanding the kernel Φ to be bounded, which implies (3.5). Hence, the operators $\mathcal{P}$ and $\mathcal{P}^*$ are assumed to be Hilbert-Schmidt. Ill-posedness can then be characterized in terms of the singular values of $\mathcal{P}$, which we will show nextly.

Since $\mathcal{P}$ is compact, it has a *singular value decomposition (SVD)* — see (KRESS, 1989, section 15.4) — that is, there exists orthonormal sequences $(\varphi_n)_{n\in\mathbf{N}}$ and $(\psi_n)_{n\in\mathbf{N}}$, in $L^2(X)$ and $L^2(Z)$ respectively, as well as a decreasing sequence $\lambda_0 \geq \lambda_1 \geq \cdots \geq 0$ of nonnegative real numbers (the singular values), such that

(a) The eigenvalues of $\mathcal{P}^*\mathcal{P}$ are precisely $(\lambda_i^2)_{i\geq 0}$;

(b) $\mathcal{P}[\varphi_i] = \lambda_i \psi_i$ and $\mathcal{P}^*[\psi_i] = \lambda_i \varphi_i$ for all $i \geq 0$;

(c) For all $h \in L^2(X)$ we have $h = \sum_{i=0}^{\infty} \langle h, \varphi_i \rangle \varphi_i + \bar{h}$, where $\bar{h} \in \ker \mathcal{P}$;

(d) For all $g \in L^2(Z)$ we have $g = \sum_{i=0}^{\infty} \langle g, \psi_i \rangle \psi_i + \bar{g}$, where $\bar{g} \in \ker \mathcal{P}^*$;

Therefore, we have

$$\mathcal{P}[h] = \sum_{i=0}^{\infty} \lambda_i \langle h, \varphi_i \rangle \psi_i$$

as well as

$$\mathcal{P}^*[g] = \sum_{i=0}^{\infty} \lambda_i \langle g, \psi_i \rangle \varphi_i.$$

We can also see that, because of (a), identification is equivalent to all the λ_i being strictly greater than 0.

Since $\mathcal{P}^*\mathcal{P}$ is compact, its sequence of eigenvalues $(\lambda_i^2)_{i\geq 0}$ and, hence, the sequence of singular values $(\lambda_i)_{i\geq 0}$, must converge to 0. This fact allows the NPIV regression problem to be ill posed when combined with the following proposition (KRESS, 1989, theorem 15.18), sometimes known as the Picard theorem:

3.1 Proposition We have $r \in \operatorname{Im}(\mathcal{P})$ if, and only if, $r \in (\ker \mathcal{P}^*)^{\perp}$ and

$$\sum_{i=0}^{\infty} \frac{1}{\lambda_i^2} \langle r, \psi_i \rangle^2 < \infty.$$

Then, the solution to $\mathcal{P}[h] = r$ is given by

$$h = \sum_{i=0}^{\infty} \frac{1}{\lambda_i} \langle r, \psi_i \rangle \varphi_i. \tag{3.6}$$

As we are assuming $r \in \operatorname{Im}(\mathcal{P})$, this proposition tells us that if instead of r we measure $r + \delta\psi_i$, the solution will be perturbed by $\frac{\delta}{\lambda_i}\varphi_i$, which may be very large in norm since λ_i can be arbitrarily small. To overcome this issue of noncontinuity, a classical technique in inverse problems is to apply Tikhonov regularization.

Observe that solving $\mathcal{P}[h] = r$ is equivalent to solving

$$\operatorname*{arg\,min}_{h \in L^2(X)} \|\mathcal{P}[h] - r\|_{L^2(Z)}^2.$$

Tikhonov regularization modifies this objective function by adding a penalization for large norm solutions. Given a regularization parameter $\alpha > 0$, the new optimization problem is

$$\operatorname*{arg\,min}_{h \in L^2(X)} \; L_\alpha(h) = \|\mathcal{P}[h] - r\|_{L^2(Z)}^2 + \alpha \|h\|_{L^2(X)}^2.$$

The function L_α is clearly convex and a straightforward application of the chain rule gives us

$$\nabla L_\alpha(h) = 2\left(\mathcal{P}^*[\mathcal{P}[h] - r] + \alpha h\right). \tag{3.7}$$

Hence, we can minimize L_α by setting the gradient to be null. Denoting the solution by h_α, we have:

$$\begin{aligned} 2(\mathcal{P}^*[\mathcal{P}[h_\alpha] - r] + \alpha h_\alpha) = 0 \iff & \mathcal{P}^*\mathcal{P}[h_\alpha] + \alpha h_\alpha = \mathcal{P}^*[r] \\ \iff & h_\alpha = (\mathcal{P}^*\mathcal{P} + \alpha I)^{-1}\mathcal{P}^*[r]. \end{aligned} \tag{3.8}$$

Since the (eigenvector, eigenvalue) pairs of $(\mathcal{P}^*\mathcal{P}+\alpha I)^{-1}$ are precisely $(1/(\lambda_i^2+\alpha),\varphi_i)$, a computation shows that

$$h_\alpha = \sum_{i=0}^{\infty} \frac{\lambda_i}{\lambda_i^2+\alpha}\langle r,\psi_i\rangle\varphi_i.$$

This clearly controls the decay of λ_i by replacing $1/\lambda_i$ in (3.6) with $\lambda_i/(\lambda_i^2+\alpha)$.

To be able to better control the convergence of h_α to $h^\star$ as $\alpha \to 0$, the authors of (DAROLLES et al., 2011) impose a so-called *source condition*, which states that there exists $\beta > 0$ such that

$$\sum_{i=0}^{\infty} \frac{\langle h^\star,\varphi_i\rangle^2}{\lambda_i^{2\beta}} < \infty \tag{3.9}$$

or, equivalently, that $h^\star \in \mathrm{Im}[(\mathcal{P}^*\mathcal{P})^{\beta/2}]$. This condition, although somewhat common in the inverse problems literature, is considerably restrictive, even more so as the degree of ill-posedness of the problem increases. For example, if the λ_i decay exponentially, then (3.9) requires the Fourier coefficients $\langle h^\star,\varphi_i\rangle$ of $h^\star$ to decay even more rapidly, which means that $h^\star$ must have a highly precise finite-dimensional approximation.

Another technique employed in (DAROLLES et al., 2011) to obtain better consistency results is, instead of performing a standard Tikhonov regularization with (3.8), to build a sequence of estimators iteratively:

$$\begin{aligned} h_\alpha^{(1)} &= (\mathcal{P}^*\mathcal{P}+\alpha I)^{-1}\mathcal{P}^*[r],\\ h_\alpha^{(k+1)} &= (\mathcal{P}^*\mathcal{P}+\alpha I)^{-1}[\mathcal{P}^*r + h_\alpha^{(k)}]. \end{aligned} \tag{3.10}$$

While for the estimator (3.8) one has $\|h^\star - h_\alpha\| = \mathcal{O}(\alpha^{(2\wedge\beta)})$, for the iterated estimator it is possible to show $\left\|h^\star - h_\alpha^{(k)}\right\| = \mathcal{O}(\alpha^{(2k\wedge\beta)})$ (DAROLLES et al., 2011).

To be able to apply the iteration scheme (3.10), one must first obtain estimates for $\mathcal{P},\mathcal{P}^*$ and r, since these are unknown. The choice was made to use Nadaraya-Watson kernel estimators — see (NADARAYA, 1964) and (WATSON, 1964).

The important aspects of this method that the reader should keep in mind are the following:

- Ill-posedness is tackled by assuming identification and performing (iterated) Tikhonov regularization;
- A source condition is necessary for providing convergence bounds;
- The estimation of conditional expectations is performed using Nadaraya-Watson kernels.

3.3 Nonparametric 2SLS

In (NEWEY; POWELL, 2003), the authors overcome the difficulties presented in section 3.1 by assuming identification and restricting the solution $h^\star$ to belong to a compact set. With this assumption, the problem of continuous inverse is automatically satisfied, since the inverse of a continuous function defined on a compact set and taking values in a Hausdorff space is automatically continuous (MUNKRES, 2000).

However, we must advise the reader that the formulation of the nonparametric regression problem presented in (NEWEY; POWELL, 2003) is a somewhat different then ours. More specifically, the authors do not use the notation presented in subsection 2.4.1, neither the spaces $L^2(X)$ and $L^2(Z)$ for domain and codomain of the conditional expectation operator. Therefore, the compact set where the solution is restricted to live is not necessarily a compact subset of $L^2(X)$, and it becomes difficult to compare both approaches from a theoretical perspective, a task which is left for future work. Nonetheless, we decided to include a section on this paper because of its importance to the research in nonparametric methods for instrumental variables. In the following, we will provide a high level description of their method, adapting the notation and focusing on the aspects we feel are the most important. We refer the reader to the original paper for details.

Suppose that the structural function can be approximated as follows:

$$h^\star(x) \approx \sum_{j=1}^{J} \gamma_j p_j(x), \tag{3.11}$$

where $p_1, p_2, \ldots$ is a sequence of “basis” functions and γ is the corresponding vector of coefficients. Substituting this into (3.3), we get

$$\mathbb{E}[Y \mid Z = z] = r(z) = \sum_{j=1}^{J} \gamma_j \mathbb{E}[p_j(X) \mid Z = z]. \tag{3.12}$$

This suggests a two stage procedure analogous to 2SLS. In the first stage, a nonparametric estimate of $\mathbb{E}[p_j(X) \mid Z = z]$ is obtained for $j = 1, \ldots, J$. Then, in the second stage, the Y samples on these estimates to obtain the vector of coefficients γ. It is worth emphasizing that the second stage regression is substantially sensitive to the number J of approximating functions and the precision of the first stage estimators (NEWEY; POWELL, 2003).

More specifically, their search space consists of structural functions which take the following form:

$$h(x) = \beta^\top a(x) + h_1(x), \tag{3.13}$$

where β is a vector of unknown coefficients, a is a vector of known functions and h_1 is an unknown function. The compactness property is obtained by placing bounds on β and restricting h_1 and its derivatives to be small in the tails, which is done by demanding a certain Sobolev norm of h_1 to be finite. The details can be found in (NEWEY; POWELL, 2003). It is important to notice that (3.13) is a semiparametric model which allows h to be nonparametric in the center of the support of X, but restricts it to be parametric in the tails. Let us denote by $\mathcal{G}$ the set of functions which are of the form (3.13) and satisfy the mentioned regularity conditions. It is assumed that $h^\star = (\beta^\star)^\top a(x) + h_1^\star(x) \in \mathcal{G}$.

The nonparametric estimate for $h_1^\star$ is based on an expansion such as (3.11), where the p_j are chosen to be Hermite polynomials and the coefficients are restricted in a way that the final estimate belongs to $\mathcal{G}$.

With these choices, it is not yet possible to then use equation (3.12) directly, since the conditional distribution of X given Z is not known. Hence, a separate step is necessary, in which estimates for $\mathbb{E}[Y \mid Z = z]$, $\mathbb{E}[a(X) \mid Z = z]$ and $\mathbb{E}[p_j(X) \mid Z = z]$ are obtained. Assuming that these estimates are available and denoting them by $\widehat{\mathbb{E}}[Y \mid Z = z], \widehat{\mathbb{E}}[a(X) \mid Z = z]$ and $\widehat{\mathbb{E}}[p_j(X) \mid Z = z]$, equation (3.3) indicates that one should optimize the parameters in order to have

$$\widehat{\mathbb{E}}[Y \mid Z = z] \approx \beta^\top \widehat{\mathbb{E}}[a(X) \mid Z = z] + \sum_{j=1}^{J} \gamma_j \widehat{\mathbb{E}}[p_j(X) \mid Z = z].$$

Therefore, the optimization objective chosen in (NEWEY; POWELL, 2003) is

$$\widetilde{Q}(\beta, \gamma) = \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i - \beta^\top \widehat{\mathbb{E}}[a(X) \mid Z = z_i] - \sum_{j=1}^{J} \gamma_j \widehat{\mathbb{E}}[p_j(X) \mid Z = z_i] \right\}^2 . \tag{3.14}$$

In the paper, it is stated that it is not necessary to use $\widehat{\mathbb{E}}[Y_i \mid Z = z]$ inside the objective instead of Y_i because of the choice made for first stage $\widehat{\mathbb{E}}$-estimators, which we will shortly clarify. With this objective function, the nonparametric 2SLS estimator is obtained minimizing $\widetilde{Q}$ over (β, γ), that is,

$$\widehat{h}(x) = \widehat{\beta}^\top a(x) + \sum_{j=1}^{J} \widehat{\gamma}_j p_j(x) \tag{3.15}$$

where $(\widehat{\beta}, \widehat{\gamma}) = \arg\min\ \widetilde{Q}(\beta, \gamma)$ subject to $h = \beta^\top a + \sum_{j=1}^{J} \gamma_j p_j \in \mathcal{G}$. This restriction is shown to be equivalent to a quadratic inequality restriction on β and γ and, therefore, there is a closed form solution for $\widehat{\gamma}$ and $\widehat{\beta}$.

What is left to do is to specify how the first stage estimators are computed. For this task, a series estimator is employed, using splines or power series as approximating functions. Then, the estimated values of $\widehat{\mathbb{E}}[a(X) \mid Z = z_i]$ and $\widehat{\mathbb{E}}[p_j(X) \mid Z = z_i]$ are employed in the computation of $\widetilde{Q}(\beta, \gamma)$. The specific form taken by these estimators may be consulted in (NEWEY; POWELL, 2003).

The key aspects of the nonparametric 2SLS (NP2SLS) estimator given by (3.15) which we would like the reader to have in mind are the following:

- The ill-posedness of the inverse problem is solved by requiring identification and that the structural function $h^\star$ belong to a compact set;
- There is a clear two stage procedure analogous to 2SLS, where the first stage estimates the conditional expectation using splines/power series and the second stage employs a truncated basis expansion with Hermite polynomials.

To end this section, we remark that although (NEWEY; POWELL, 2003) provides consistency results of the form $\left\|\widehat{h} - h^\star\right\| \to 0$ (in probability for a norm $\|\cdot\|$ related to the space $\mathcal{G}$), these results rest on non-trivial assumptions about the quality of the first stage regression, the denseness of the chosen basis functions and the already mentioned compactness of the parameter set.

4 NPIV Regression through Stochastic Approximate Gradients

In this chapter, we present a novel approach to NPIV regression, which leverages stochastic approximate gradients to avoid computation of the adjoint operator $\mathcal{P}^*$. The main idea is to search for $h \in L^2(X)$ which minimizes a certain risk measure $\mathcal{R} : L^2(X) \to \mathbf{R}$. The minimization is performed through an approximate SGD procedure, where $h_{m+1} = h_m - \alpha_m u_m$ and u_m is an approximate stochastic gradient for $\mathcal{R}$ at h_m.

4.1 Risk measure

Motivated by the fact that $h^\star$ satisfies[1]

$$\mathcal{P}[h^\star] = r,$$

we introduce a pointwise loss function $\ell : \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ and define the *populational risk measure* $\mathcal{R} : L^2(X) \to \mathbf{R}$ associated with it to be

$$\mathcal{R}(h) = \mathbb{E}[\ell(r(Z), \mathcal{P}[h](Z))].$$

Throughout the rest of this chapter, the example the reader should keep in mind is the quadratic[2] loss function $\ell(y, y') = \frac{1}{2}(y - y')^2$. Our goal is to solve the NPIV regression problem stated in subsection 2.4.1 by solving

$$\inf_{h \in \mathcal{H}} \mathcal{R}(h),$$

where $\mathcal{H}$ is a closed, convex, bounded subset of $L^2(X)$ such that $(h^\star + \ker \mathcal{P}) \cap \mathcal{H} \neq \emptyset$. This is a weaker condition than $h^\star \in \mathcal{H}$, but which is sufficient for our theoretical results. We also require $0 \in \mathcal{H}$. For future reference, we state these conditions in

4.1 Assumption (Regularity of $\mathcal{H}$) The set $\mathcal{H}$ is a closed, convex, bounded subset of $L^2(X)$, which contains the origin and satisfies $(h^\star + \ker \mathcal{P}) \cap \mathcal{H} \neq \emptyset$.

The only part of this assumption which concerns the data generating process is $(h^\star + \ker \mathcal{P}) \cap \mathcal{H} \neq \emptyset$. This essentially means that the set $\mathcal{H}$ is large enough so that there exists $h \in \mathcal{H}$ to that $\mathcal{R}(h) = \mathcal{R}(h^\star)$. For $\mathcal{H}$ satisfying assumption 4.1, we let

[1] Recall the notation introduced in the beginning of chapter 3.

[2] Although our results apply to more general loss functions, detailed experiments with setups in which other loss functions are more suitable are left for future work.

$D \triangleq \operatorname{diam}\mathcal{H} < \infty$, so that $\|h\| < D$ for every $h \in \mathcal{H}$, since $\|h\| = \|h - 0\| < \operatorname{diam}\mathcal{H}$, as $0 \in \mathcal{H}$. One possible choice for the set $\mathcal{H}$ is the $L^\infty(X)$ ball contained in $L^2(X)$, that is

$$\mathcal{H} = \left\{h \in L^2(X) : \|h\|_\infty < A\right\}, \tag{4.1}$$

where $A > 0$ is a constant. This set is obviously convex and bounded in the $L^2(X)$ norm. It can be shown that it is also closed, but not necessarily compact, as can be seen by taking a $\|\cdot\|_\infty$–bounded orthonormal basis for $L^2(X)$, if one exists. We denote by $\pi_{\mathcal{H}}$ the orthogonal projection onto $\mathcal{H}$. In case $\mathcal{H}$ is given by (4.1), we have the explicit formula:

$$\pi_{\mathcal{H}}[h] = (h^+ \wedge A) - (h^- \wedge A).$$

We now state all the assumptions needed on the pointwise loss ℓ. We denote by ∂_2 a partial derivative with respect to the second argument.

4.2 Assumption (Regularity of ℓ)

(i) The function $\ell : \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ is convex and C^2 with respect to its second argument;

(ii) The function ℓ has Lipschitz first derivative with respect to the second argument, i.e., there exists $L \geq 0$ such that, for all $y, y', u, u' \in \mathbf{R}$ we have

$$|\partial_2\ell(y, y') - \partial_2\ell(u, u')| \leq L(|y - u| + |y' - u'|).$$

Some useful facts which follow immediately from these assumptions are:

4.3 Proposition Assume that ℓ satisfies Assumption 4.2. Then:

(i) Setting $C_0 = |\partial_2\ell(0, 0)|$ we have

$$|\partial_2\ell(y, y')| \leq C_0 + L(|y| + |y'|)$$

for all $y, y' \in \mathbf{R}$;

(ii) The map $f \mapsto \partial_2\ell(r_0(\cdot), f(\cdot))$ from $L^2(Z)$ to $L^2(Z)$ is well-defined and L-Lipschitz.

(iii) The second derivative with respect to the second argument is bounded: $|\partial_2^2\ell(y, y')| \leq L$ for all $y, y' \in \mathbf{R}$;

Proof.

(i) Write $\partial_2\ell(y, y') = \partial_2\ell(y, y') - \partial_2\ell(0, 0) + \partial_2\ell(0, 0)$ and apply the triangle inequality as well as Assumption 4.2.(ii).

(ii) From the previous item we know this map is well-defined. If f and g belong to $L^2(Z)$, we have

$$\begin{aligned}\|\partial_2\ell(r_0(\cdot), f(\cdot)) - \partial_2\ell(r_0(\cdot), g(\cdot))\|^2_{L^2(Z)} &= \mathbb{E}\left[|\partial_2(r_0(Z), f(Z)) - \partial_2(r_0(Z), g(Z))|^2\right] \\ &\leq L^2\mathbb{E}\left[|f(Z) - g(Z)|^2\right] \\ &= L^2\|f - g\|^2_{L^2(Z)}.\end{aligned}$$

(iii) Follows from the definition of derivative and Assumption 4.2 (ii). □

4.2 Gradient computation

As our strategy is based on minimizing the risk measure $\mathcal{R}$, we would like to compute an analytical formula for $\nabla\mathcal{R}(h)$, where $h \in L^2(X)$. We start by computing the directional derivative of $\mathcal{R}$ at h in the direction f, denoted by $D\mathcal{R}[h](f)$:

$$\begin{aligned}D\mathcal{R}[h](f) &= \lim_{\delta\to 0}\frac{1}{\delta}\left[\mathcal{R}(h + \delta f) - \mathcal{R}(f)\right] \\ &= \lim_{\delta\to 0}\frac{1}{\delta}\mathbb{E}\left[\ell(r(Z), \mathcal{P}[h + \delta f](Z)) - \ell(r(Z), \mathcal{P}[h](Z))\right] \\ &= \lim_{\delta\to 0}\frac{1}{\delta}\mathbb{E}\left[\ell(r(Z), \mathcal{P}[h](Z) + \delta\mathcal{P}[f](Z)) - \ell(r(Z), \mathcal{P}[h](Z))\right] \\ &= \lim_{\delta\to 0}\frac{1}{\delta}\mathbb{E}\Bigg[\delta\partial_2\ell(r(Z), \mathcal{P}[h](Z))\cdot\mathcal{P}[f](Z) \\ &\qquad + \frac{\delta^2}{2}\partial_2^2\ell(r(Z), \mathcal{P}[h + \theta f](Z))\cdot\mathcal{P}[f](Z)^2\Bigg] \\ &= \mathbb{E}\left[\partial_2\ell(r(Z), \mathcal{P}[h](Z))\cdot\mathcal{P}[f](Z)\right] \\ &\qquad + \lim_{\delta\to 0}\mathbb{E}\left[\frac{\delta}{2}\partial_2^2\ell(r(Z), \mathcal{P}[h + \theta f](Z))\cdot\mathcal{P}[f](Z)^2\right] \\ &= \mathbb{E}\left[\partial_2\ell(r(Z), \mathcal{P}[h](Z))\cdot\mathcal{P}[f](Z)\right],\end{aligned}$$

where $\theta \in (0, \delta)$ is due to Taylor's formula. The last step is then due to Proposition 4.3 (iii). We can in fact expand the calculation a bit more, as follows:

$$\begin{aligned}D\mathcal{R}[h](f) &= \mathbb{E}\left[\partial_2\ell(r(Z), \mathcal{P}[h](Z))\cdot\mathcal{P}[f](Z)\right] \\ &= \langle\partial_2\ell(r(\cdot), \mathcal{P}[h](\cdot)), \mathcal{P}[f]\rangle_{L^2(Z)} \\ &= \langle\mathcal{P}^*[\partial_2\ell(r(Z), \mathcal{P}[h](\cdot))], f\rangle_{L^2(X)}.\end{aligned}$$

This shows that $\mathcal{R}$ is Gateux-differentiable, with Gateux derivative at h given by

$$D\mathcal{R}[h] = \mathcal{P}^*[\partial_2\ell(r(\cdot), \mathcal{P}[h](\cdot))].$$

By Proposition 4.3 (ii) we have that $h \mapsto D\mathcal{R}[h]$ is a continuous mapping from $L^2(X)$ to $L^2(X)$, which implies that $\mathcal{R}$ is also Fréchet-differentiable, and both derivatives coincide

(PATHAK, 2018, Theorem 3.3). Therefore,

$$\nabla\mathcal{R}(h) = \mathcal{P}^*[\partial_2\ell(r(\cdot), \mathcal{P}[h](\cdot))]. \tag{4.2}$$

Some previous approaches to NPIV, such as the one discussed in section 3.2, involved approximating $\mathcal{P}^*$ directly using kernel methods. We, in contrast, perform one more step before plugging in estimators. From section 3.1, we know that $\mathcal{P}^*$ is an integral operator with kernel $\Phi(x,z) = p_{X,Z}(x,z)/(p_X(x)p_Z(z))$, that is,

$$\mathcal{P}^*[g](x) = \mathbb{E}[\Phi(x,Z)g(Z)] \quad \text{for all } g \in L^2(Z).$$

Therefore, taking $g(z) = \Phi(x,z)\partial_2\ell(r(z), \mathcal{P}[h](z))$, we have that the random variable $\Phi(x,Z)\partial_2\ell(r(Z), \mathcal{P}[h](Z))$ is an unbiased stochastic estimate of $\nabla\mathcal{R}(h)(x)$. In fact, a stronger result is true: we can consider $\Phi(\cdot,Z)\partial_2\ell(r(Z), \mathcal{P}[h](Z))$ as a random element of the Hilbert space $L^2(X)$, with mean vector — in the Bochner integral sense, see (HSING; EUBANK, 2015, Chapter 7) — equal to $\nabla\mathcal{R}(h)$.

A hypothesis which is needed for this proof and which is necessary for the following theoretical analysis is a finite $L^2(\mathbb{P}_X \otimes \mathbb{P}_Z)$–norm of Φ, which, as we have seen in section 3.1, amounts to saying $\mathcal{P}$ is Hilbert-Schmidt.

4.4 Assumption The kernel Φ satisfies

$$\|\Phi\|_{L^2(\mathbb{P}_X\otimes\mathbb{P}_Z)} \triangleq \left(\int_{\mathbf{R}^{d_X}\times\mathbf{R}^{d_Z}} \Phi(x,z)^2 p_X(x)p_Z(z)\ \mathrm{d}x\mathrm{d}z\right)^{\frac{1}{2}} < \infty.$$

This is the only restrictive assumption on the data generating process we have made so far, and, as we have discussed in section 3.1, it still allows the problem to be ill-posed. This assumption is also present in the paper analyzed in section 3.2.

4.3 Stochastic Approximate Gradient Descent IV

Our approximate stochastic gradient is then built using estimators $\widehat{\Phi}, \widehat{r}$ and $\widehat{\mathcal{P}}$ of Φ, r and $\mathcal{P}$ respectively. With this notation, given a sample Z we have

$$\nabla\mathcal{R}(h)(x) \approx \widehat{\Phi}(x,Z)\partial_2\ell(\widehat{r}(Z), \widehat{\mathcal{P}}[h](Z)). \tag{4.3}$$

We remain agnostic to the specific ways in which Φ, r and $\mathcal{P}$ are being estimated. All we require of these estimators is the following:

4.5 Assumption (Properties of $\widehat{\Phi}, \widehat{r}$ and $\widehat{\mathcal{P}}$)

(i) $\widehat{\Phi}, \widehat{r}$ and $\widehat{\mathcal{P}}$ are computed using a dataset $\mathcal{D}$ independent from the Z samples used in Algorithm 1.

(ii) $\widehat{r} \in L^2(Z)$ a.s.;

(iii) $\widehat{\mathcal{P}} : L^2(X) \to L^2(Z)$ is a bounded linear operator a.s., that is

$$\left\|\widehat{\mathcal{P}}\right\|_{\text{op}} = \sup_{\|h\|_{L^2(X)}=1} \left\|\widehat{\mathcal{P}}[h]\right\|_{L^2(Z)} < \infty \quad \text{a.s.;}$$

(iv) $\left\|\widehat{\Phi}\right\|_{\infty} \triangleq \sup_{\substack{x\in\mathbf{R}^{d_X}\\ z\in\mathbf{R}^{d_Z}}} \left|\widehat{\Phi}(x,z)\right| < \infty$. This implies, in particular, that $\left\|\widehat{\Phi}\right\|_{L^2(\mathbb{P}_X\otimes\mathbb{P}_Z)} < \infty$.

We now present Stochastic Approximate Gradient Descent IV (SAGD–IV), an algorithm for estimating $h^\star$ using the approximation given by (4.3).

Algorithm 1: SAGD–IV

```
input  : Samples $\{(\boldsymbol{z}_m)_{m=1}^M\}$. Estimators $\widehat{\Phi}, \widehat{r}$ and $\widehat{\mathcal{P}}$. Sequence of learning rates
         $(\alpha_m)_{m=1}^M$. Initial guess $\widehat{h}_0 \in L^2(X)$.
output: $\widehat{h}$
for $1 \le m \le M$ do
    Set $u_m = \widehat{\Phi}(\cdot, \boldsymbol{z}_m)\partial_2\ell\left(\widehat{r}(\boldsymbol{z}_m), \widehat{\mathcal{P}}[\widehat{h}_{m-1}](\boldsymbol{z}_m)\right)$ ;
    Set $\widehat{h}_m = \pi_{\mathcal{H}}\left[\widehat{h}_{m-1} - \alpha_m u_m\right]$ ;
end
Set $\widehat{h} = \frac{1}{M}\sum_{m=1}^M \widehat{h}_m$ ;
```

We note the fact that the internal loop only needs samples from the instrumental variable Z to unfold. This is due to the fact that our risk measure $\mathcal{R}$ directly compares $r(Z) = \mathbb{E}[Y \mid Z]$ and $\mathcal{P}[h](Z)$, instead of comparing $\mathcal{P}[h](Z)$ with Y. The drawback is that we have to estimate r.

4.4 Theoretical results

4.6 Remark (Comment on notation) Let X and Y be arbitrary independent random variables. Then, given a function of two arguments f, we have

$$\mathbb{E}[f(X,Y) \mid Y] = g(Y),$$

where $g(y) = \mathbb{E}[f(X,y)]$. That is, to compute $\mathbb{E}[f(X,Y) \mid Y]$ we simply treat Y as constant and integrate in X. We will denote this by writing $\mathbb{E}_X[f(X,Y)]$.

Since we are directly optimizing the risk measure, we are able to provide guarantees for $\mathcal{R}(\widehat{h})$ in mean with respect to the training data $\boldsymbol{z}_{1:M} = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_m\}$. Our main result is the following, whose proof we present in the Appendix:

4.7 Theorem Let $\widehat{h}_0, \dots, \widehat{h}_{M-1}$ be generated according to Algorithm 1. Assume that ℓ satisfies Assumption 4.2, $\mathcal{H}$ satisfies Assumption 4.1, Φ satisfies Assumption 4.4 and

$\widehat{\Phi}, \widehat{r}, \widehat{\mathcal{P}}$ satisfy Assumption 4.5. Then, if we let $\widehat{h} = \sum_{m=1}^{M} \widehat{h}_{m-1}$, the following bound holds:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{z}_{1:M}}\left[\mathcal{R}(\widehat{h}) - \mathcal{R}(h^{\star})\right] \leq \frac{D^2}{2M\alpha_M} + \frac{\xi}{M}\sum_{m=1}^{M}\alpha_m \\ + \tau \cdot \left(\left\|\Phi - \widehat{\Phi}\right\|^2_{L^2(\mathbb{P}_X \otimes \mathbb{P}_Z)} + \|r - \widehat{r}\|^2_{L^2(Z)} + \left\|\mathcal{P} - \widehat{\mathcal{P}}\right\|^2_{\mathrm{op}}\right)^{\frac{1}{2}},\end{aligned} \tag{4.4}$$

where

$$\begin{aligned}\xi = \xi\left(\widehat{\Phi}, \widehat{r}, \widehat{\mathcal{P}}\right) &= \frac{3}{2}\left\|\widehat{\Phi}\right\|^2_{\infty}\left(C_0^2 + L^2\|\widehat{r}\|^2_{L^2(Z)} + L^2D^2\left\|\widehat{\mathcal{P}}\right\|^2_{\mathrm{op}}\right), \\ \tau = \tau\left(\widehat{\Phi}\right) &= 2D\max\left\{3(C_0^2 + L^2\mathbb{E}[Y^2] + L^2D^2), 2L^2\left\|\widehat{\Phi}\right\|^2_{\infty}, 2L^2D^2\left\|\widehat{\Phi}\right\|^2_{\infty}\right\}.\end{aligned}$$

It is productive to analyze the RHS of the bound in (4.4) more carefully. If we choose the sequence (α_m) to satisfy

$$M\alpha_M \to \infty \quad \text{and} \quad \frac{1}{M}\sum_{m=1}^{M}\alpha_m \to 0$$

as $M \to \infty$, then the first two terms in the sum vanish as M grows. The last term is due to the fact that we do not know Φ, r nor $\mathcal{P}$, but it explicitly quantifies how the estimation errors of $\widehat{\Phi}, \widehat{r}$ and $\widehat{\mathcal{P}}$ come together to determine the quality of the final estimator.

4.5 Computing the necessary estimators with kernel methods

In this section, we give references for how we chose to compute $\widehat{\Phi}, \widehat{r}$ and $\widehat{\mathcal{P}}$ in our practical implementation. All of these estimators are based on kernel Ridge regression, where the unknown function is approximated using members of a Reproducing Kernel Hilbert Space (RKHS) — see (STEINART; CHRISTMANN, 2008, Chapter 4) for an introduction to these spaces. Kernel methods have the advantage of providing closed form solutions to regression problems due to the Representer Theorem (SCHÖLKOPF; HERBRICH; SMOLA, 2001, Theorem 1).

Starting with the estimator of

$$\Phi(x, z) = \frac{p(x, z)}{p(x)p(z)},$$

we chose to use the Unconstrained Least Squares Importance Fitting (uLSIF) framework described in (MASASHI SUGIYAMA TAIJI SUZUKI, 2012, Chapter 6). Using a gaussian kernel for the estimator, wee can see that it adheres to Assumption 4.5.

The method of estimation of $\widehat{\mathcal{P}}$ was taken to be the first stage of Kernel Instrumental Variable (KIV) (SINGH; SAHANI; GRETTON, 2019), a method for NPIV regression which provides two stages based on kernel Ridge regression. The first stage consists of computing a conditional mean embedding (SONG et al., 2009) of the conditional distribution of X given $Z = z$ in a RKHS of functions $\mathcal{H}_{\mathcal{X}}$, which acts as a proxy for $L^2(X)$. This conditional mean embedding encodes the same information as $\mathcal{P}$. To estimate $r(Z) = \mathbb{E}[Y \mid Z]$, we simply used this same method to obtain an approximation to the operator $L^2(Y) : f \mapsto \mathbb{E}[f(Y) \mid Z = \cdot\ ] \in L^2(Z)$, which we then apply to the identity function.

4.6 Numerical experiment

We tested SAGD-IV using the following data generating process, obtained from (BENNET; KALLUS; SCHNABEL, 2019):

$$\begin{aligned} &Y = h^\star(X) + \varepsilon + \delta && X = Z_1 + \varepsilon + \gamma \\ &Z = (Z_1, Z_2) \sim \text{Uniform}([-3,3]^2) && \varepsilon \sim \mathcal{N}(0,1),\ \ \gamma, \delta \sim \mathcal{N}(0, 0.1) \end{aligned}$$

For the response function $h^\star$, we tested two cases:

$$\textbf{sin} : h^\star(x) = \sin(x) \quad \text{and} \quad \textbf{abs} : h^\star(x) = |x|.$$

We chose to compare our model to a direct competitor: Kernel Instrumental Variable (KIV) (SINGH; SAHANI; GRETTON, 2019), due to the relationship between both methods discussed in section 4.5. All regularization parameters necessary for the computation of $\widehat{\Phi}, \widehat{r}, \widehat{\mathcal{P}}$, as well as the ones in the two stages of KIV, were chosen using cross-validation on held-out training samples. All kernels employed were gaussian, with lenghtscale parameter set to be equal to the inverse of the median distance between samples. With the notation of Algorithm 1, for SAGD-IV we set the learning rate to be $\alpha_m = \frac{1}{\sqrt{M}}$ for $1 \leq m \leq M$. For SAGD-IV, the dataset was comprised of 600 samples from the triple (X, Z, Y) and an additional 1200 samples of Z alone, to conduct the loop in Algorithm 1. The dataset for KIV was comprised of the same 600 samples of the triple (X, Z, Y), plus additional 600 samples of the pair (Z, Y) (these are necessary for the algorithm, see (SINGH; SAHANI; GRETTON, 2019)). In this way, both algorithms have access to the same amount of samples and share as many as possible.

The results obtained are in Figure 3. We can see that, while both methods performed similarly in the **sin** case, KIV was slightly better. This might be due to the smooth nature of the KIV estimator, since both stages are conducted using Kernel Ridge Regression. However, the relative performance changes drastically when looking at the **abs** case, where SAGD-IV greatly outperformed KIV. This indicates that SAGD-IV may perform better with non-smooth response functions. Of course, this isn't a thorough investigation of the empirical performance of SAGD-IV, nor irrefutable proof that it is better then KIV, and a more detailed analysis, involving other recent methods which have appeared in the NPIV literature, is left for future work.

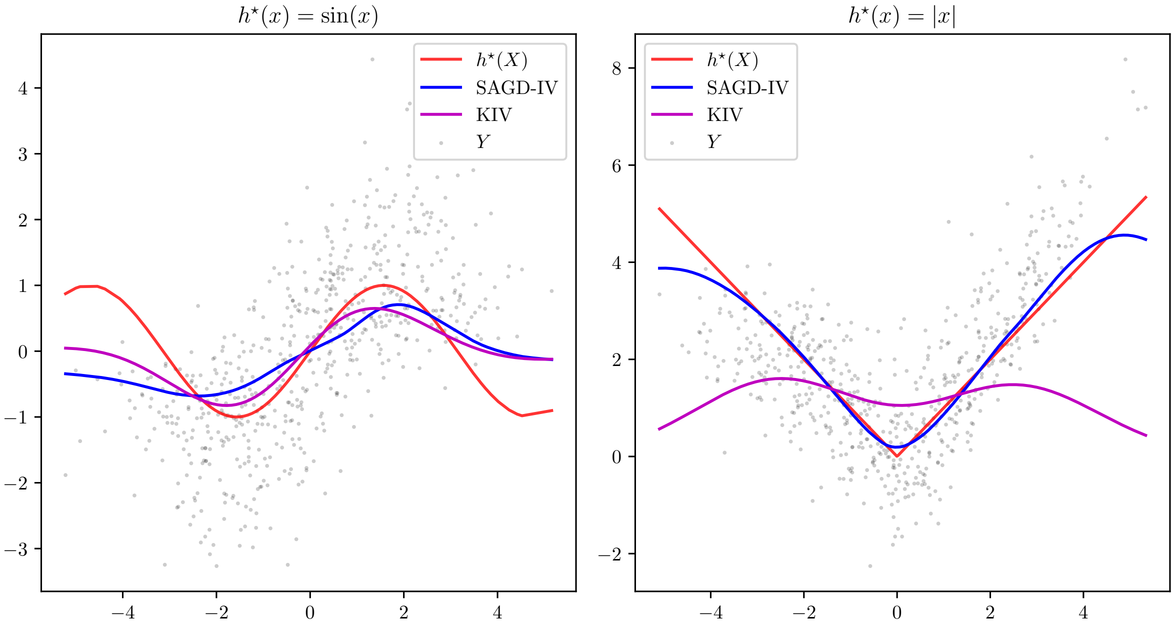

Figure 3 – SAGD-IV and KIV for two different response functions.

5 Conclusion

References

BENNET, Andrew; KALLUS, Nathan; SCHNABEL, Tobias. Deep Generalized Method of Moments for Instrumental Variable Analysis. **NeurIPS**, 2019.

CARRASCO, Marine; FLORENS, Jean-Pierre; RENAULT, Eric. Handbook of Econometrics. In: [s.l.]: Elsevier, 2007. 6B Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization.

DAROLLES, S. et al. Nonparametric Instrumental Regression. **Econometrica**, v. 79, n. 5, p. 1541–5165, 2011.

HERNÁN, Miguel A.; ROBINS, James M. **Causal Inference: What If**. [S.l.]: Chapman & Hall/CRC, 2020.

HSING, Tailen; EUBANK, Randall. **Theoretical Foundations of Functional Data Analysis, with and Introduction to Linear Operators**. [S.l.]: John Wiley & Sons, 2015. (Wiley Series in Probability and Statistics).

KRESS, Rainer. **Linear Integral Equations**. [S.l.]: Springer-Verlag, 1989. (Applied Mathematical Sciences).

LEHMANN, E. L. **Testing Statistical Hypotheses**. [S.l.]: John Wiley & Sons, 1959.

MASASHI SUGIYAMA TAIJI SUZUKI, Takafumi Kanamori. **Density Ration Estimation in Machine Learning**. [S.l.]: Cambridge University Press, 2012.

MUNKRES, James R. **Topology**. [S.l.]: Prentice Hall, Inc, 2000.

NADARAYA, E. A. On Estimating Regression. **Theory of Probability & its Applications**, v. 9, n. 1, p. 141–142, 1964.

NEWEY, Whitney K.; POWELL, James L. Instrumental Variable Estimation of Nonparametric Models. **Econometrica**, v. 71, n. 5, p. 1565–1578, 2003. DOI: http://dx.doi.org/10.1111/1468-0262.00459.

PATHAK, Hemant Kumar. **An Introduction to Nonlinear Analysis and Fixed Point Theory**. [S.l.]: Springer, 2018.

RUDIN, Walter. **Functional Analysis**. [S.l.]: McGraw-Hill, 1991.

SCHÖLKOPF, Bernhard; HERBRICH, Ralf; SMOLA, Alex J. A Generalized Representer Theorem. In________. **Computational Learning Theory**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001. P. 416–426. ISBN 978-3-540-44581-4.

SINGH, Rahul; SAHANI, Maneesh; GRETTON, Arthur. Kernel Instrumental Variable Regression. In: WALLACH, H. et al. (Eds.). **Advances in Neural Information Processing Systems**. [S.l.]: Curran Associates, Inc., 2019. v. 32. Available from: <https://proceedings.neurips.cc/paper_files/paper/2019/file/17b3c7061788dbe82de5abe9f6fe22b3-Paper.pdf>.

SONG, Le et al. Hilbert Space Embeddings of Conditional Distributions with Applications to Dynamical Systems. In: PROCEEDINGS of the 26th Annual International Conference on Machine Learning. Montreal, Quebec, Canada: Association for Computing Machinery, 2009. (ICML '09), p. 961–968. ISBN 9781605585161. DOI: 10.1145/1553374.1553497. Available from: <https://doi.org/10.1145/1553374.1553497>.

STEINART, Ingo; CHRISTMANN, Andreas. **Support Vector Machines**. [S.l.]: Springer, 2008. (Information Science and Statistics).

WATSON, Geoffrey S. Smooth Regression Analysis. **Sankhyā: The Indian Journal of Statistics, Series A**, v. 26, n. 4, p. 359–372, 1964.

WOOLDRIDGE, Jeffrey M. **Econometric Analysis of Cross Section and Panel Data**. [S.l.]: The MIT Press, 2001. ISBN 9780262232197.

Appendix

APPENDIX A – Proofs of theoretical results

A.1 Proof of Theorem 4.7

To lighten the notation, the symbols $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$, when written without a subscript to specify which space they refer to, will act as the norm and inner product, respectively, of $L^2(X)$. Before presenting the proof of Theorem 4.7, we need to prove two auxiliary lemmas:

A.1 Lemma In the procedure of Algorithm 1 we have $u_m \in L^2(X)$ for all $1 \leq m \leq M$ and, furthermore,

$$\mathbb{E}_{\boldsymbol{z}_{1:M}}[\|u_m\|^2] \leq \rho\left(\widehat{\Phi}, \widehat{r}, \widehat{\mathcal{P}}\right),$$

where

$$\rho\left(\widehat{\Phi}, \widehat{r}, \widehat{\mathcal{P}}\right) = 3\left\|\widehat{\Phi}\right\|_\infty^2 \left(C_0^2 + L^2\|\widehat{r}\|_{L^2(Z)}^2 + L^2 D^2 \left\|\widehat{\mathcal{P}}\right\|_{\text{op}}^2\right).$$

Proof. By Assumption 4.5 we have:

$$\begin{aligned}
\|u_m\|_{L^2(X)}^2 &= \left\|\widehat{\Phi}(\cdot, \boldsymbol{z}_m)\partial_2\ell\left(\widehat{r}(\boldsymbol{z}_m), \widehat{\mathcal{P}}[\widehat{h}_{m-1}](\boldsymbol{z}_m)\right)\right\|_{L^2(X)}^2 \\
&= \mathbb{E}_X\left[\left|\widehat{\Phi}(X, \boldsymbol{z}_m)\partial_2\ell\left(\widehat{r}(\boldsymbol{z}_m), \widehat{\mathcal{P}}[\widehat{h}_{m-1}](\boldsymbol{z}_m)\right)\right|^2\right] \\
&\leq \partial_2\ell\left(\widehat{r}(\boldsymbol{z}_m), \widehat{\mathcal{P}}[\widehat{h}_{m-1}](\boldsymbol{z}_m)\right)^2 \left\|\widehat{\Phi}\right\|_\infty^2 \\
&< \infty.
\end{aligned}$$

Hence, $u_m \in L^2(X)$ for all m. This computation and Proposition 4.3 (i) then imply

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{z}_{1:M}}\left[\|u_m\|^2\right] &\leq 3\left\|\widehat{\Phi}\right\|_\infty^2 \left(C_0^2 + L^2\left(\|\widehat{r}\|_{L^2(Z)}^2 + \left\|\widehat{\mathcal{P}}[\widehat{h}_{m-1}]\right\|_{L^2(Z)}^2\right)\right) \\
&\leq 3\left\|\widehat{\Phi}\right\|_\infty^2 \left(C_0^2 + L^2\left(\|\widehat{r}\|_{L^2(Z)}^2 + \left\|\widehat{\mathcal{P}}\right\|_{\text{op}}^2 \left\|\widehat{h}_{m-1}\right\|^2\right)\right) \\
&\leq 3\left\|\widehat{\Phi}\right\|_\infty^2 \left(C_0^2 + L^2\left(\|\widehat{r}\|_{L^2(Z)}^2 + D^2\left\|\widehat{\mathcal{P}}\right\|_{\text{op}}^2\right)\right) \\
&= 3\left\|\widehat{\Phi}\right\|_\infty^2 \left(C_0^2 + L^2\|\widehat{r}\|_{L^2(Z)}^2 + L^2D^2\left\|\widehat{\mathcal{P}}\right\|_{\text{op}}^2\right) \triangleq \rho\left(\widehat{\Phi}, \widehat{r}, \widehat{\mathcal{P}}\right).
\end{aligned}$$

□

A.2 Lemma In the procedure of Algorithm 1 we have

$$\left\|\mathbb{E}_{\boldsymbol{z}_m}\left[\nabla\mathcal{R}(\widehat{h}_{m-1}) - u_m\right]\right\| \leq \kappa\left(\widehat{\Phi}\right)\left(\left\|\Phi - \widehat{\Phi}\right\|_{L^2(\nu_X \otimes \nu_Z)}^2 + \|r - \widehat{r}\|_{L^2(Z)}^2 + \left\|\mathcal{P} - \widehat{\mathcal{P}}\right\|_{\text{op}}^2\right)^{\frac{1}{2}},$$

where

$$\kappa^2\left(\widehat{\Phi}\right) \triangleq 2\max\left\{3(C_0^2 + L^2\mathbb{E}[Y^2] + L^2D^2), 2L^2\left\|\widehat{\Phi}\right\|_\infty^2, 2L^2D^2\left\|\widehat{\Phi}\right\|_\infty^2\right\}.$$

Proof. To ease the notation, we define

$$\begin{aligned}\Psi_m(Z) &\triangleq \partial_2\ell(r(Z), \mathcal{P}[\widehat{h}_{m-1}](Z)),\\ \widehat{\Psi}_m(Z) &\triangleq \partial_2\ell(\widehat{r}(Z), \widehat{\mathcal{P}}[\widehat{h}_{m-1}](Z)).\end{aligned}$$

Let's expand the definition of $\|\cdot\|$:

$$\begin{aligned}\left\|\mathbb{E}_{\boldsymbol{z}_m}\left[\nabla\mathcal{R}(\widehat{h}_{m-1}) - u_m\right]\right\| &= \mathbb{E}_X\left[\mathbb{E}_{\boldsymbol{z}_m}\left[\nabla\mathcal{R}(\widehat{h}_{m-1})(X) - u_m(X)\right]^2\right]^{\frac{1}{2}}\\ &= \mathbb{E}_X\left[\left(\nabla\mathcal{R}(\widehat{h}_{m-1})(X) - \mathbb{E}_{\boldsymbol{z}_m}\left[u_m(X)\right]\right)^2\right]^{\frac{1}{2}}\\ &= \mathbb{E}_X\left[\left(\mathbb{E}_Z\left[\Phi(X,Z)\Psi_m(Z)\right] - \mathbb{E}_{\boldsymbol{z}_m}\left[\widehat{\Phi}(X,\boldsymbol{z}_m)\widehat{\Psi}_m(\boldsymbol{z}_m)\right]\right)^2\right]^{\frac{1}{2}}\\ &= \mathbb{E}_X\left[\left(\mathbb{E}_Z\left[\Phi(X,Z)\Psi_m(Z) - \widehat{\Phi}(X,Z)\widehat{\Psi}_m(Z)\right]\right)^2\right]^{\frac{1}{2}},\end{aligned}$$

Now we add and subtract $\widehat{\Phi}(X,Z)\Psi_m(Z)$, so that

$$\begin{aligned}&\mathbb{E}_X\left[\left(\mathbb{E}_Z\left[\Phi(X,Z)\Psi_m(Z) - \widehat{\Phi}(X,Z)\widehat{\Psi}_m(Z)\right]\right)^2\right]^{\frac{1}{2}}\\ &\quad= \mathbb{E}_X\left[\left(\mathbb{E}_Z\left[\Psi_m(Z)\left(\Phi(X,Z) - \widehat{\Phi}(X,Z)\right) + \widehat{\Phi}(X,Z)\left(\Psi_m(Z) - \widehat{\Psi}_m(Z)\right)\right]\right)^2\right]^{\frac{1}{2}}\\ &\quad\le \mathbb{E}_X\left[\left(\|\Psi_m\|_{L^2(Z)}\left\|\Phi(X,\cdot) - \widehat{\Phi}(X,\cdot)\right\|_{L^2(Z)} + \left\|\widehat{\Phi}(X,\cdot)\right\|_{L^2(Z)}\left\|\Psi_m - \widehat{\Psi}_m\right\|_{L^2(Z)}\right)^2\right]^{\frac{1}{2}}\\ &\quad\le \sqrt{2}\mathbb{E}_X\left[\|\Psi_m\|_{L^2(Z)}^2\left\|\Phi(X,\cdot) - \widehat{\Phi}(X,\cdot)\right\|_{L^2(Z)}^2 + \left\|\widehat{\Phi}(X,\cdot)\right\|_{L^2(Z)}^2\left\|\Psi_m - \widehat{\Psi}_m\right\|_{L^2(Z)}^2\right]^{\frac{1}{2}}\\ &\quad= \sqrt{2}\left(\|\Psi_m\|_{L^2(Z)}^2\left\|\Phi - \widehat{\Phi}\right\|_{L^2(\nu_X\otimes\nu_Z)}^2 + \left\|\widehat{\Phi}\right\|_{L^2(\nu_X\otimes\nu_Z)}^2\left\|\Psi_m - \widehat{\Psi}_m\right\|_{L^2(Z)}^2\right)^{\frac{1}{2}},\end{aligned}$$

where

$$\|\Phi\|_{L^2(\nu_X\otimes\nu_Z)}^2 = \int_{\mathcal{X}\times\mathcal{Z}} \Phi(x,z)^2 p(x)p(z)\,\mathrm{d}x\mathrm{d}z$$

is the norm with respect to the independent coupling of the distributions of X and Z. By Proposition 4.3.(i) we have

$$\begin{aligned}\|\Psi_m\|_{L^2(Z)}^2 &= \mathbb{E}_Z\left[\partial_2\ell(r(Z), \mathcal{P}[\widehat{h}_{m-1}](Z))^2\right]\\ &\le \mathbb{E}_Z\left[\left(C_0 + L\left(|r(Z)| + \left|\mathcal{P}[\widehat{h}_{m-1}](Z)\right|\right)\right)^2\right]\\ &\le 3\left(C_0^2 + L^2\|r\|_{L^2(Z)}^2 + L^2\left\|\mathcal{P}[\widehat{h}_{m-1}]\right\|_{L^2(Z)}^2\right)\\ &\le 3\left(C_0^2 + L^2\mathbb{E}[Y^2] + L^2D^2\right).\end{aligned}$$

It is also clear that, by Assumption 4.5,

$$\left\|\widehat{\Phi}\right\|^2_{L^2(\nu_X\otimes\nu_Z)} \le \left\|\widehat{\Phi}\right\|^2_{\infty}.$$

Finally, by Assumption 4.2.(ii) we also have

$$\begin{aligned}\left\|\Psi_m - \widehat{\Psi}_m\right\|^2_{L^2(Z)} &= \mathbb{E}_Z\left[\left(\partial_2\ell(r(Z),\mathcal{P}[\widehat{h}_{m-1}](Z)) - \partial_2\ell(\widehat{r}(Z),\widehat{\mathcal{P}}[\widehat{h}_{m-1}](Z))\right)^2\right]\\ &\le 2L^2\left(\|r-\widehat{r}\|^2_{L^2(Z)} + \left\|(\mathcal{P}-\widehat{\mathcal{P}})[\widehat{h}_{m-1}]\right\|^2_{L^2(Z)}\right)\\ &\le 2L^2\left(\|r-\widehat{r}\|^2_{L^2(Z)} + D^2\left\|\mathcal{P}-\widehat{\mathcal{P}}\right\|^2_{\mathrm{op}}\right).\end{aligned}$$

To combine all terms, we first define

$$\kappa^2\left(\widehat{\Phi}\right) \triangleq 2\max\left\{3(C_0^2 + L^2\mathbb{E}[Y^2] + L^2D^2), 2L^2\left\|\widehat{\Phi}\right\|^2_{\infty}, 2L^2D^2\left\|\widehat{\Phi}\right\|^2_{\infty}\right\}.$$

Then, it's easy to see that

$$\left\|\mathbb{E}_{\boldsymbol{z}_m}\left[\nabla\mathcal{R}(\widehat{h}_{m-1}) - u_m\right]\right\| \le \kappa\left(\widehat{\Phi}\right)\left(\left\|\Phi-\widehat{\Phi}\right\|^2_{L^2(\nu_X\otimes\nu_Z)} + \|r-\widehat{r}\|^2_{L^2(Z)} + \left\|\mathcal{P}-\widehat{\mathcal{P}}\right\|^2_{\mathrm{op}}\right)^{\frac{1}{2}},$$

as we wanted to show. □

We now have everything needed to show the

Proof of Theorem 4.7. We start by checking that $\mathcal{R}$ is convex in $\mathcal{H}$: if $h, g \in \mathcal{H}$ and $\lambda \in [0,1]$, then

$$\begin{aligned}\mathcal{R}(\lambda h + (1-\lambda)g) &= \mathbb{E}[\ell(r(Z),\mathcal{P}[\lambda h + (1-\lambda)g](Z))]\\ &= \mathbb{E}[\ell(r(Z),\lambda\mathcal{P}[h](Z) + (1-\lambda)\mathcal{P}[g](Z))]\\ &\le \lambda\mathbb{E}[\ell(r(Z),\mathcal{P}[h](Z))] + (1-\lambda)\mathbb{E}[\ell(r(Z),\mathcal{P}[g](Z))]\\ &= \lambda\mathcal{R}(h) + (1-\lambda)\mathcal{R}(g).\end{aligned}$$

By Assumption 4.1, there exists $\bar{h} \in (h^\star + \ker\mathcal{P}) \cap \mathcal{H}$. By the Algorithm 1 procedure, we have

$$\begin{aligned}\frac{1}{2}\left\|\widehat{h}_m - \bar{h}\right\|^2 &= \frac{1}{2}\left\|\pi_{\mathcal{H}}\left[\widehat{h}_{m-1} - \alpha_m u_m\right] - \bar{h}\right\|^2\\ &\le \frac{1}{2}\left\|\widehat{h}_{m-1} - \alpha_m u_m - \bar{h}\right\|^2\\ &= \frac{1}{2}\left\|\widehat{h}_{m-1} - \bar{h}\right\|^2 - \alpha_m\langle u_m, \widehat{h}_{m-1} - \bar{h}\rangle + \frac{\alpha_m^2}{2}\|u_m\|^2.\end{aligned}$$

After adding and subtracting $\alpha_m\langle\nabla\mathcal{R}(\widehat{h}_{m-1}), \widehat{h}_{m-1} - \bar{h}\rangle$, we are left with

$$\frac{1}{2}\left\|\widehat{h}_{m-1} - \bar{h}\right\|^2 - \alpha_m\langle u_m - \nabla\mathcal{R}(\widehat{h}_{m-1}), \widehat{h}_{m-1} - \bar{h}\rangle + \frac{\alpha_m^2}{2}\|u_m\|^2 - \alpha_m\langle\nabla\mathcal{R}(\widehat{h}_{m-1}), \widehat{h}_{m-1} - \bar{h}\rangle.$$

Applying the first order convexity inequality on the last term give us, in total,

$$\begin{aligned}\frac{1}{2}\left\|\widehat{h}_m-\bar{h}\right\|^2 \le \frac{1}{2}\left\|\widehat{h}_{m-1}-\bar{h}\right\|^2 &- \alpha_m\langle u_m-\nabla\mathcal{R}(\widehat{h}_{m-1}), \widehat{h}_{m-1}-\bar{h}\rangle\\ &+\frac{\alpha_m^2}{2}\|u_m\|^2-\alpha_m(\mathcal{R}(\widehat{h}_{m-1})-\mathcal{R}(\bar{h})).\end{aligned}$$

Notice that, by the definition of $\bar{h}$, we have $\mathcal{R}(\bar{h})=\mathcal{R}(h^\star)$. Hence, making this substitution and rearranging terms, we get

$$\begin{aligned}\mathcal{R}(\widehat{h}_{m-1})-\mathcal{R}(h^\star) \le \frac{1}{2\alpha_m}&\left(\left\|\widehat{h}_{m-1}-\bar{h}\right\|^2-\left\|\widehat{h}_m-\bar{h}\right\|^2\right)\\ &+\frac{\alpha_m}{2}\|u_m\|^2-\langle u_m-\nabla\mathcal{R}(\widehat{h}_{m-1}), \widehat{h}_{m-1}-\bar{h}\rangle.\end{aligned}$$

Finally, summing over $1\le m\le M$ leads to

$$\begin{aligned}\sum_{n=1}^{M}\left[\mathcal{R}(\widehat{h}_{m-1})-\mathcal{R}(h^\star)\right] \le \sum_{m=1}^{M}\frac{1}{2\alpha_m}&\left(\left\|\widehat{h}_{m-1}-\bar{h}\right\|^2-\left\|\widehat{h}_m-\bar{h}\right\|^2\right)\\ &+\sum_{m=1}^{M}\frac{\alpha_m}{2}\|u_m\|^2\\ &+\sum_{m=1}^{M}\langle\nabla\mathcal{R}(\widehat{h}_{m-1})-u_m, \widehat{h}_{m-1}-\bar{h}\rangle.\end{aligned}\tag{A.1}$$

The next step is to take the average of both sides with respect to $\boldsymbol{z}_{1:M}$, taking advantage of the independence between $\boldsymbol{z}_{1:M}$ and $\mathcal{D}$, the data used to compute $\widehat{\Phi}, \widehat{r}$ and $\widehat{\mathcal{P}}$. Each summation in the RHS is then bounded separately.

The first summation admits a deterministic bound. By assumption, we the diameter D of $\mathcal{H}$ is finite. Hence

$$\begin{aligned}\sum_{m=1}^{M}\frac{1}{2\alpha_m}\left(\left\|\widehat{h}_{m-1}-\bar{h}\right\|^2-\left\|\widehat{h}_m-\bar{h}\right\|^2\right) &= \sum_{m=2}^{M}\left(\frac{1}{2\alpha_m}-\frac{1}{2\alpha_{m-1}}\right)\left\|\widehat{h}_{m-1}-\bar{h}\right\|^2\\ &\quad+\frac{1}{2\alpha_1}\left\|\widehat{h}_0-\bar{h}\right\|^2-\frac{1}{2\alpha_M}\left\|\widehat{h}_M-\bar{h}\right\|^2\\ &\le\sum_{m=2}^{M}\left(\frac{1}{2\alpha_m}-\frac{1}{2\alpha_{m-1}}\right)D^2+\frac{1}{2\alpha_1}D^2\\ &=\frac{D^2}{2\alpha_M}.\end{aligned}\tag{A.2}$$

The second summation can be bounded with the aid of Lemma A.1:

$$\mathbb{E}_{\boldsymbol{z}_{1:M}}\left[\sum_{m=1}^{M}\frac{\alpha_m}{2}\|u_m\|^2\right]=\frac{\mathbb{E}_{\boldsymbol{z}_{1:M}}\left[\|u_m\|^2\right]}{2}\sum_{m=1}^{M}\alpha_m\le\frac{\rho\left(\widehat{\Phi},\widehat{r},\widehat{\mathcal{P}}\right)}{2}\sum_{m=1}^{M}\alpha_m.\tag{A.3}$$

Finally, the third summation can be bounded using Lemma A.2. Let $\mathbb{E}_{\boldsymbol{z}_{-m}}$ denote the expectation with respect to $\boldsymbol{z}_1, \dots, \boldsymbol{z}_{m-1}, \boldsymbol{z}_{m+1}, \dots, \boldsymbol{z}_M$ and notice that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{z}_{1:M}}\left[\langle \nabla\mathcal{R}(\widehat{h}_{m-1}) - u_m, \widehat{h}_{m-1} - \bar{h}\rangle\right] &= \mathbb{E}_{\boldsymbol{z}_{-m}}\left[\mathbb{E}_{\boldsymbol{z}_m}\left[\langle \nabla\mathcal{R}(\widehat{h}_{m-1}) - u_m, \widehat{h}_{m-1} - \bar{h}\rangle\right]\right] \\
&= \mathbb{E}_{\boldsymbol{z}_{-m}}\left[\langle \mathbb{E}_{\boldsymbol{z}_m}\left[\nabla\mathcal{R}(\widehat{h}_{m-1}) - u_m\right], \widehat{h}_{m-1} - \bar{h}\rangle\right] \\
&= \mathbb{E}_{\boldsymbol{z}_{-m}}\left[\left\|\mathbb{E}_{\boldsymbol{z}_m}\left[\nabla\mathcal{R}(\widehat{h}_{m-1}) - u_m\right]\right\|\left\|\widehat{h}_{m-1} - \bar{h}\right\|\right] \\
&\leq D\mathbb{E}_{\boldsymbol{z}_{-m}}\left[\left\|\mathbb{E}_{\boldsymbol{z}_m}\left[\nabla\mathcal{R}(\widehat{h}_{m-1}) - u_m\right]\right\|\right].
\end{aligned}$$

Then, applying Lemma A.2 and setting $\tau \triangleq D\kappa$ we get

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}_{1:M}}\left[\langle \nabla\mathcal{R}(\widehat{h}_{m-1}) - u_m, \widehat{h}_{m-1} - \bar{h}\rangle\right] \\
&\qquad \leq \tau\left(\widehat{\Phi}\right)\left(\left\|\Phi - \widehat{\Phi}\right\|^2_{L^2(\nu_X \otimes \nu_Z)} + \|r - \widehat{r}\|^2_{L^2(Z)} + \left\|\mathcal{P} - \widehat{\mathcal{P}}\right\|^2_{\text{op}}\right)^{\frac{1}{2}}.
\end{aligned} \tag{A.4}$$

All that is left to do is to apply equations (A.1), (A.2), (A.3) and (A.4) along with the inequality which defines convexity. Let $\widehat{h} \triangleq \frac{1}{M}\sum_{m=1}^{M} \widehat{h}_{m-1}$ and $\xi \triangleq \rho/2$. Then:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}_{1:M}}\left[\mathcal{R}(\widehat{h}) - \mathcal{R}(h^\star)\right] \\
&\quad \leq \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\boldsymbol{z}_{1:M}}\left[\mathcal{R}(\widehat{h}_m) - \mathcal{R}(h^\star)\right] \\
&\quad \leq \frac{D^2}{2M\alpha_M} + \xi\left(\widehat{\Phi}, \widehat{r}, \widehat{\mathcal{P}}\right)\frac{1}{M}\sum_{m=1}^{M}\alpha_m \\
&\qquad\quad + \tau\left(\widehat{\Phi}\right)\left(\left\|\Phi - \widehat{\Phi}\right\|^2_{L^2(\nu_X \otimes \nu_Z)} + \|r - \widehat{r}\|^2_{L^2(Z)} + \left\|\mathcal{P} - \widehat{\mathcal{P}}\right\|^2_{\text{op}}\right)^{\frac{1}{2}}. \qquad \square
\end{aligned}$$