

Nonparametric Instrumental Variable Regression with Observed Covariates

Zikai Shen^{*1}, Zonghao Chen^{*2}, Dimitri Meunier³, Ingo Steinwart⁴, Arthur Gretton^{†3}, and Zhu Li^{†3}

¹Department of Statistical Science, University College London

²Department of Computer Science, University College London

³Gatsby Computational Neuroscience Unit, University College London

⁴Department of Mathematics, University of Stuttgart

Abstract

We study the problem of nonparametric instrumental variable regression with observed covariates, which we refer to as NPIV-O. Compared with standard nonparametric instrumental variable regression (NPIV), the additional observed covariates facilitate causal identification and enables heterogeneous causal effect estimation. However, the presence of observed covariates introduces two challenges for its theoretical analysis. First, it induces a partial identity structure, which renders previous NPIV analyses—based on measures of ill-posedness, stability conditions, or link conditions—inapplicable. Second, it imposes anisotropic smoothness on the structural function. To address the first challenge, we introduce a novel *Fourier measure of partial smoothing*; for the second challenge, we extend the existing kernel 2SLS instrumental variable algorithm with observed covariates, termed KIV-O, to incorporate Gaussian kernel lengthscales adaptive to the anisotropic smoothness. We prove upper L^2 -learning rates for KIV-O and the first L^2 -minimax lower learning rates for NPIV-O. Both rates interpolate between known optimal rates of NPIV and nonparametric regression (NPR). Interestingly, we identify a gap between our upper and lower bounds, which arises from the choice of kernel lengthscales tuned to minimize a projected risk. Our theoretical analysis also applies to proximal causal inference, an emerging framework for causal effect estimation that shares the same conditional moment restriction as NPIV-O.

1 Introduction

We consider the problem of identifying and estimating the *causal effect* of a treatment variable X on an outcome variable Y , where their relationship is confounded by an unobserved confounder ϵ . Despite the existence of unobserved confounding, it is nonetheless possible to identify the causal effect by leveraging an *instrumental variable* Z . For instance, if one aims to identify the causal effect of smoking (X) on the risk of lung disease (Y), which may be potentially confounded by an individual’s occupation and early childhood environment (ϵ), the cigarette cost (Z) would be a valid instrument as it only affects the risk of lung disease via smoking [Leigh and Schembri, 2004]. For cases where X and Y are continuous (and possibly multivariate) random variables, *Nonparametric Instrumental Variables Regression* (NPIV) has received significant attention. NPIV is particularly

^{*}, [†] Equal contribution in random order.

valuable as it avoids imposing potentially misspecified parametric or semiparametric assumptions when such structure is not warranted [Newey and Powell, 2003, Horowitz, 2011]. Prior literature has explored various algorithms for NPIV. These include methods based on: kernel density estimation [Hall and Horowitz, 2005, Darolles et al., 2011], sieve minimum distance estimators [Chen and Pouzo, 2012, Newey and Powell, 2003, Chen and Christensen, 2018] and more recently Reproducing Kernel Hilbert Spaces (RKHSs) with the Kernel Instrumental Variables (KIV) algorithm [Singh et al., 2019, Meunier et al., 2024a], which is a nonparametric generalization of the two-stage least squares (2SLS) algorithm. Another family of nonparametric algorithms is based on min-max optimization [Bennett et al., 2019, Dikkala et al., 2020, Bennett et al., 2023]. We defer a full discussion of these approaches to [Section 3](#).

Practitioners often have access to *observed covariates* O . These observed covariates encode individual-level characteristics, which allow for the estimation of *heterogenous causal effects*. Returning to the smoking example, confounders such as an individual’s occupation fall into this subset as it is readily observable. Such extra information enables the estimation of heterogeneous treatment effects—for instance, the causal effect of smoking on lung disease specifically for manual workers. In this work, we refer to the NPIV framework that incorporates observed covariates as NPIV-O. Formally, we introduce the following NPIV-O model,

$$Y = f_*(X, O) + \epsilon, \quad \mathbb{E}[\epsilon | Z, O] = 0. \quad (1)$$

We refer to f_* as the *heterogeneous dose response curve*, and it is our target of interest. Here, ϵ is an unobserved confounder that affects Y additively. By Eq. (1), we implicitly assume that Z can only possibly affect Y through X , a condition known as *exclusion restriction*. The mean independence $\mathbb{E}[\epsilon | Z, O] = 0$ is a relaxation of the stronger *unconfoundedness assumption*, that requires (Z, O) to be independent of ϵ . We also require that Z and X are not independent, a condition known as *instrumental relevance*. A random variable Z that satisfies these requirements is referred to as a valid instrumental variable. For a detailed discussion of our assumptions, we refer to [Section 4](#). We also note that the observed covariates O aid identification by capturing non-linear confounding effects, thereby relaxing the strict additivity assumption required for *all* confounders in classical NPIV.

Incorporating observed covariates in NPIV estimation poses both an algorithmic and a theoretical challenge. To understand this, define the conditional expectation operator

$$T : L^2(P_{XO}) \rightarrow L^2(P_{ZO}), \quad f \mapsto \mathbb{E}[f(X, O) | Z, O].$$

Eq. (1) is equivalent to the following *conditional moment restriction* for f_* :

$$\mathbb{E}[Y | Z, O] = (Tf_*)(Z, O). \quad (2)$$

We refer the reader to [Assumption 4.1](#) for the technical assumption ensuring unique identification of f_* from this equation. With the presence of O , T acts as an identity operator on the infinite dimensional function space

$$\mathcal{F}_1 = \{f : f(X, O) = f_1(O), \quad f_1 \in L^2(P_O)\} \subset L^2(P_{XO}), \quad (3)$$

rendering the operator non-compact. A naive application of kernel instrumental variables without observed covariates [Singh et al., 2019] augments both X and Z to (X, O) and (Z, O) , without exploiting the fact that T is partially an identity operator. As a result, this algorithm is not consistent, a point we elaborate on in [Section 2.1](#). From a theoretical standpoint, T being partially

an identity operator fundamentally alters the statistical properties of NPIV-O estimation compared with classical NPIV, posing significant challenges as detailed below.

The first challenge is to characterize the degree of ill-posedness of the inverse problem, Eq. (2), while accounting for the fact that T is partially an identity operator. We consider another function space $\mathcal{F}_2 = \{f : f(X, O) = f_2(X), f_2 \in L^2(P_X)\} \subset L^2(P_{XO})$. In this case, under mild conditions on the conditional distributions [Darolles et al., 2011, Assumption A.1], T when treated as a mapping from \mathcal{F}_2 to $L^2(P_{ZO})$ is a compact operator whose smoothing effect can be quantified through the rate of decay of its singular values. *This mixed behaviour of T reveals the nature of NPIV-O as a hybrid between NPIV and Nonparametric Regression (NPR).* Existing theoretical analyses of NPIV estimator mostly preclude the case where T acts as a partial identity operator [Blundell et al., 2007, Chen and Reiss, 2011, Chen and Pouzo, 2012, Chen and Christensen, 2018, Meunier et al., 2024a, Kim et al., 2025, Chen et al., 2024], making them unsuitable for the NPIV-O setting in Eq. (1). We refer the reader to Section 3 and Remark 4.2 for a more in-depth discussion.

Another challenge is the *anisotropic smoothness* of the heterogenous dose response curve f_* across the treatment X and observed covariates O . In real-world applications, the treatment X (e.g. smoking) is often one dimensional, while the observed covariates O (e.g. occupation, age, gene) are of higher dimensionality. This is because practitioners tend to adjust for as many observed covariates as possible, in an effort to overcome unobserved confounding. Therefore, a desirable algorithm would adapt to the *intrinsic smoothness* of f_* , thereby adapts appropriately to the high intrinsic smoothness when the directional smoothness is highly anisotropic [Hoffman and Lepski, 2002]. To achieve this desirable property, we modify the KIV-O algorithm to select kernel lengthscales adaptively to the varying directional smoothness of f_* . In contrast, existing NPIV algorithms are typically analyzed under an isotropic smoothness assumption, meaning their established convergence rates are dictated by the worst smoothness across all dimensions [Singh et al., 2019, 2024]. This limitation of these theoretical guarantees becomes increasingly severe in high dimensions.

In this paper, we tackle the above two challenges and make the following contributions.

1. *Fourier measure of partial smoothing effect of T :* In Section 4.1, we introduce a new framework based on Fourier spectra which quantifies the *partial* smoothing effect of T for functions $f : \mathcal{X} \times \mathcal{O} \rightarrow \mathbb{R}$ with only high frequencies on X ; while conversely quantifies the *partial* anti-smoothing effect of T for functions $f : \mathcal{X} \times \mathcal{O} \rightarrow \mathbb{R}$ with only low frequencies on X . Our Fourier measure of partial smoothing effect resembles existing ones based on sieves [Blundell et al., 2007, Chen et al., 2024, Kim et al., 2025], but features two key distinctions: 1) Our framework takes into account the partial identity structure of T caused by the existence of observed covariates O . 2) Our framework aligns, through Bochner's theorem, with the RKHS of a continuous translational invariant kernel. This alignment will be useful for our next contribution.
2. *Upper and minimax lower $L^2(P_{XO})$ -learning rate:* Under the above framework that quantifies the *partial* smoothing effect of T , we prove an upper learning rate (Theorem 4.1) for a kernel based algorithm for instrumental variable regression with observed covariates proposed in Singh et al. [2024], termed KIV-O. Furthermore, we prove the first minimax lower learning rate (Theorem 4.2) for NPIV-O defined in Eq. (1). All our bounds hold in the strong $L^2(P_{XO})$ norm rather than the pseudo-metric $\|T(\cdot)\|_{L^2(P_{ZO})}$ considered in Singh et al. [2019]. For the following two edge cases: 1) No observed covariates: NPIV-O reduces to classical NPIV; 2) No hidden confounding: instrument variables Z are unnecessary and NPIV-O reduces to NPR, our upper bound of KIV-O matches the minimax lower bound, recovering earlier results on minimax optimality of classical kernel instrumental variable regression [Meunier et al., 2024a] and kernel ridge regression [Hang and Steinwart, 2021, Fischer and Steinwart, 2020], respectively established under analogous

assumptions to our work. In the general intermediate case where T exhibits a partial identity structure, both our upper and lower learning rates interpolate accordingly, however, there exists a gap between the upper and minimax lower bound. We posit that this gap is fundamental and we provide insights on why this gap emerges in [Section 5](#).

3. *Adaptivity to model intrinsic smoothness:* We modify the existing KIV-O algorithm to select kernel lengthscales separately for each dimension, adaptive to the varying directional smoothness of f_* . We prove that its learning rate takes into account the anisotropic smoothness of f_* across the treatment X and observed covariates O . Compared with existing KIV algorithms and their associated analyses that assume isotropic smoothness [Fischer and Steinwart, 2020, Meunier et al., 2024a, Singh, 2020, Singh et al., 2024], our learning rate is adaptive to the target function's intrinsic smoothness, and alleviates the slow rate caused by the need to account for the worst-case smoothness, when the anisotropic smoothness is highly imbalanced.
4. *Interpretable anisotropic smoothness assumption:* Another key feature of our upper and lower bounds is that they highlight the separate contribution of the partial smoothing effect of T and the anisotropic smoothness of f_* , characterized by an anisotropic Besov space. In contrast, much work in the NPIV literature employ a generalized source condition with respect to the unknown conditional expectation operator T [Engl et al., 1996, Singh et al., 2019, Mastouri et al., 2021, Singh, 2020, Bozkurt et al., 2025a, Hall and Horowitz, 2005], which is less interpretable because T is unknown a priori and cannot reveal the separate contribution of the intrinsic (anisotropic) smoothness of f_* and the smoothness of T .

1.1 Organization of the paper

An outline of the paper is as follows. In [Section 2](#), we introduce the RKHS-based 2SLS algorithm for instrumental variable regression with observed covariates, referred to as KIV-O. In [Section 3](#), we discuss related work on NPIV in the literature. In [Section 4](#), we present the main assumptions and theoretical results, and discuss the interpretation of our findings. In [Section 5](#), we highlight the fundamental challenges towards obtaining minimax optimal rates of KIV-O.

2 Setup

Consider P the joint data-generating probability measure over (Z, O, X, Y) , where $Z \in \mathcal{Z} := [0, 1]^{d_Z}$ denotes the instrument, $O \in \mathcal{O} := [0, 1]^{d_O}$ denotes the observed covariates, $Y \in \mathbb{R}$ denotes the outcome variable, and $X \in \mathcal{X} := [0, 1]^{d_X}$ denotes the treatment variable. We use p to denote the probability density functions; for example, $p(\mathbf{x} | \mathbf{z}, \mathbf{o})$ denotes the density of the conditional distribution $P_{X|Z=\mathbf{z}, O=\mathbf{o}}$. As stated in [Section 1](#), we define the conditional expectation operator T :

$$T : L^2(P_{XO}) \rightarrow L^2(P_{ZO}), \quad f \mapsto ((\mathbf{z}, \mathbf{o}) \mapsto \int_{\mathcal{X}} f(\mathbf{x}, \mathbf{o}) p(\mathbf{x} | \mathbf{z}, \mathbf{o}) d\mathbf{x}).$$

Notations: Let \mathbb{N}_+ denote the set of positive integers and $\mathbb{N} = \mathbb{N}_+ \cup \{0\}$ denote the set of non-negative integers. We use boldfaced letters, such as \mathbf{x} , to denote a vector in \mathbb{R}^d for $d \geq 1$. Specifically, $\mathbf{x} = [x_1, \dots, x_d]^\top \in \mathcal{X} \subset \mathbb{R}^d$. For a distribution P defined on a measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and $0 < p < \infty$, $L^p(P)$ is the space of functions $h : \mathcal{X} \rightarrow \mathbb{R}$ such that $\|h\|_{L^p(P)} := \mathbb{E}_{X \sim P} [|h(X)|^p]^{\frac{1}{p}} < \infty$ and $L^\infty(P)$ is the space of functions that are bounded P -almost everywhere. When P is the Lebesgue measure $\mathcal{L}_{\mathcal{X}}$ over \mathcal{X} , we write $L^p(\mathcal{X}) := L^p(\mathcal{L}_{\mathcal{X}})$. For H a separable Hilbert space, we let $L^p(\mathcal{X}; H)$ denote the space of Bochner 2-integrable functions from \mathcal{X} to H with norm $\|F\|_{L^2(\mathcal{X}; H)}^2 = \int_{\mathcal{X}} \|F(\mathbf{x})\|_H^2 d\mathbf{x}$. Two Banach spaces E_1, E_2 are said to be isometrically isomorphic,

denoted $E_1 \cong E_2$, if there exists an isometric isomorphism S , such that $\|Sh\|_{E_2} = \|h\|_{E_1}$ for all $h \in E_1$. Two Banach spaces E_1, E_2 are said to be norm equivalent, denoted $E_1 \simeq E_2$, if E_1, E_2 coincide as sets and there are constants $c_1, c_2 > 0$ such that $c_1\|h\|_{E_1} \leq \|h\|_{E_2} \leq c_2\|h\|_{E_1}$ holds for all $h \in E_1$. For an operator $T : E_1 \rightarrow E_2$, $\|T\|$ denotes its operator norm and T^* denotes its adjoint. For two Hilbert spaces H_1, H_2 , $S_2(H_1, H_2)$ is the Hilbert space of Hilbert-Schmidt operators from H_1 to H_2 . For two numbers α and β , we let $\alpha \wedge \beta = \min(\alpha, \beta)$ and $\alpha \vee \beta = \max(\alpha, \beta)$. \lesssim (resp. \gtrsim) means \leq (resp. \geq) up to positive multiplicative constants.

2.1 Algorithm

In this section, we introduce a kernel two-stage least-squares approach for instrumental variable regression with observed covariates, which we term the KIV-O algorithm. KIV-O algorithm adopts a sample splitting strategy. In Stage I, we learn the conditional expectation operator T with dataset $\mathcal{D}_1 := \{(\tilde{\mathbf{z}}_i, \tilde{\mathbf{o}}_i, \tilde{\mathbf{x}}_i)\}_{i=1}^n$ (see Eq. (6)); in Stage II, we perform regression of the outcome Y on the features learned in Stage I with dataset $\mathcal{D}_2 := \{(\mathbf{z}_i, \mathbf{o}_i, y_i)\}_{i=1}^n$ (see Eq. (7)). The KIV-O algorithm is a generalization of the KIV algorithm proposed in Singh et al. [2019]. We note that many existing NPIV learning methods employ a two-stage estimation procedure, see for instance Hartford et al. [2017], Singh et al. [2019], Xu et al. [2021a], Li et al. [2024b], Petrushev et al. [2024], Khouri et al. [2025].

We now briefly review the relevant reproducing kernel Hilbert space (RKHS) theory, following Berlinet and Thomas-Agnan [2004]. For a domain $\mathcal{X} \subseteq \mathbb{R}^d$, a Hilbert space \mathcal{H} of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ is called a *Reproducing Kernel Hilbert Space* (RKHS) if the evaluation functional $\delta_{\mathbf{x}} : \mathcal{H} \rightarrow \mathbb{R}$ defined by $f \mapsto f(\mathbf{x})$ is continuous for every $\mathbf{x} \in \mathcal{X}$. Every RKHS \mathcal{H} has a unique symmetric, positive definite *reproducing kernel* $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, which satisfies $k(\mathbf{x}, \cdot) \in \mathcal{H}$ for all $\mathbf{x} \in \mathcal{X}$ and $\langle f, k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}} = f(\mathbf{x})$ for all $f \in \mathcal{H}$ and $\mathbf{x} \in \mathcal{X}$ (the reproducing property). To describe the KIV-O algorithm, we introduce RKHSs \mathcal{H}_X on \mathcal{X} , $\mathcal{H}_{O,1}$ and $\mathcal{H}_{O,2}$ on \mathcal{O} and \mathcal{H}_Z on \mathcal{Z} . The reasoning for defining two distinct RKHSs on \mathcal{O} will be clear later in the algorithm. We denote the associated unique reproducing kernels via $k_X : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, $k_{O,1} : \mathcal{O} \times \mathcal{O} \rightarrow \mathbb{R}$, $k_{O,2} : \mathcal{O} \times \mathcal{O} \rightarrow \mathbb{R}$, $k_Z : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$. We denote the canonical feature map of \mathcal{H}_X as $\phi_X(\mathbf{x}) := k_X(\mathbf{x}, \cdot)$, and similarly for feature maps $\phi_{O,1}, \phi_{O,2}, \phi_Z$.

Assumption 2.1. All kernels ($k_X, k_{O,1}, k_{O,2}$ and k_Z) are measurable and bounded.

An immediate consequence of Assumption 2.1 is that the embedding $I_{P_X} : \mathcal{H}_X \rightarrow L^2(P_X)$, which maps a function $f \in \mathcal{H}_X$ to its P_X -equivalence class $[f]_{P_X}$ is well-defined, compact and Hilbert-Schmidt [Steinwart and Scovel, 2012, Lemma 2.3]. We define $[\mathcal{H}_X]_{P_X} \subseteq L^2(P_X)$ as the image of I_{P_X} . For $\beta > 0$, we denote by $[\mathcal{H}_X]_{P_X}^{\beta}$ the β -th *power space*, as introduced in Steinwart and Scovel [2012, Theorem 4.6]. For $0 \leq \beta \leq 1$, this space is shown to be isomorphic to the β -interpolation space $[L^2(P_X), [\mathcal{H}_X]_{P_X}]_{\beta,2}$ [Steinwart and Scovel, 2012, Theorem 4.6]. It is known that the 1-interpolation space $[\mathcal{H}_X]_{P_X}^1$ is isometrically isomorphic to the closed subspace $(\ker I_{P_X})^{\perp}$ of \mathcal{H}_X via I_{P_X} [Steinwart and Scovel, 2012, Lemma 2.12]. For $\beta \geq 1$, the space contains functions that are smoother than those in \mathcal{H}_X . The same definitions and properties hold for \mathcal{H}_Z , $\mathcal{H}_{O,1}$ and $\mathcal{H}_{O,2}$ as well.

For two Hilbert spaces H, H' , we let $H \otimes H'$ denote their tensor product Hilbert space, defined as $H \otimes H' := \overline{\text{span}\{u \otimes u' : u \in H, u' \in H'\}}$, where $u \otimes u'$ is the linear rank-one operator $H' \rightarrow H$ defined by $(u \otimes u')v' = \langle u', v' \rangle_{H'} u$ [Aubin, 2011, Section 12]. In the case of RKHSs, the tensor product $\mathcal{H}_{ZO,1} := \mathcal{H}_Z \otimes \mathcal{H}_{O,1}$ and $\mathcal{H}_{XO,2} := \mathcal{H}_X \otimes \mathcal{H}_{O,2}$ are the unique RKHSs associated with the product kernels $k_{ZO,1}((\mathbf{z}, \mathbf{o}), (\mathbf{z}', \mathbf{o}')) = k_Z(\mathbf{z}, \mathbf{z}') \cdot k_{O,1}(\mathbf{o}, \mathbf{o}')$ and $k_{XO,2}((\mathbf{x}, \mathbf{o}), (\mathbf{x}', \mathbf{o}')) = k_X(\mathbf{x}, \mathbf{x}') \cdot k_{O,2}(\mathbf{o}, \mathbf{o}')$, respectively [Berlinet and Thomas-Agnan, 2004]. We define the embedding $I_{P_{XO,2}} : \mathcal{H}_{XO,2} \rightarrow L^2(P_{XO})$ which maps a function $f \in \mathcal{H}_{XO,2}$ to its P_{XO} -equivalence class $[f]_{P_{XO}}$,

and define the β -th power spaces as $[\mathcal{H}_{XO,2}]_{P_{XO}}^\beta$. An analogous construction applies to $\mathcal{H}_{ZO,1}$, yielding the spaces $[\mathcal{H}_{ZO,1}]_{P_{ZO}}^\beta$. In the rest of the paper, we omit the subscript and use the notation $[\cdot]$ to denote equivalence classes in L^2 .

We are now ready to present the KIV-O algorithm.

Stage I. The action of the operator T on the RKHS $\mathcal{H}_{XO,2}$ can be represented with the aid of the *conditional mean embedding* (CME) [Song et al., 2009, Park and Muandet, 2020, Klebanov et al., 2020, Li et al., 2022]. We define the CME F_* as the mapping from $\mathcal{Z} \times \mathcal{O}$ to \mathcal{H}_X , given by $(\mathbf{z}, \mathbf{o}) \mapsto \mathbb{E}[\phi_X(X) | Z = \mathbf{z}, O = \mathbf{o}]$. Equipped with the CME, we note that the image of T acting on a function $f \in \mathcal{H}_{XO,2}$ admits the following representation: for any $(\mathbf{z}, \mathbf{o}) \in \mathcal{Z} \times \mathcal{O}$,

$$\begin{aligned} (Tf)(\mathbf{z}, \mathbf{o}) &= \mathbb{E}[f(X, O) | Z = \mathbf{z}, O = \mathbf{o}] = \mathbb{E}[\langle f, \phi_X(X) \otimes \phi_{O,2}(O) \rangle_{\mathcal{H}_{XO,2}} | Z = \mathbf{z}, O = \mathbf{o}] \\ &= \langle f, \mathbb{E}[\phi_X(X) | Z = \mathbf{z}, O = \mathbf{o}] \otimes \phi_{O,2}(\mathbf{o}) \rangle_{\mathcal{H}_{XO,2}} = \langle f, F_*(\mathbf{z}, \mathbf{o}) \otimes \phi_{O,2}(\mathbf{o}) \rangle_{\mathcal{H}_{XO,2}}, \end{aligned}$$

where the second equality follows from the *reproducing property* and the third equality requires a Bochner integrable feature map ϕ_X (true for bounded kernels) from Assumption 2.1 [Steinwart and Christmann, 2008, Definition A.5.20]. Note that the feature map ϕ_X is projected by the conditional expectation of the conditional distribution $P_{X|Z,O}$, while the feature map $\phi_{O,2}$ remains unprojected. This is the key distinction from classical KIV. In Stage I, our goal is to estimate the CME, F_* , by performing a regularized least squares regression in a vector-valued RKHS \mathcal{G} induced by the operator-valued kernel [Grünewälder et al., 2012, Li et al., 2022]

$$K := k_{ZO,1} \text{Id}_{\mathcal{H}_X} : (\mathcal{Z} \times \mathcal{O}) \times (\mathcal{Z} \times \mathcal{O}) \rightarrow \mathcal{L}(\mathcal{H}_X), \quad (4)$$

where $\mathcal{L}(\mathcal{H}_X)$ denotes the space of bounded linear operators $\mathcal{H}_X \rightarrow \mathcal{H}_X$, and $\text{Id}_{\mathcal{H}_X} \in \mathcal{L}(\mathcal{H}_X)$ denotes the identity operator on \mathcal{H}_X . An important property of \mathcal{G} is that it is isometrically isomorphic to the space $S_2(\mathcal{H}_{ZO,1}, \mathcal{H}_X)$ of Hilbert-Schmidt operators from $\mathcal{H}_{ZO,1}$ to \mathcal{H}_X . On the other hand, by Aubin [2011, Theorem 12.6.1], $S_2(L^2(P_{ZO}), \mathcal{H}_X)$ is isometrically isomorphic to the Bochner space $L^2(P_{ZO}, \mathcal{H}_X)$, and we denote this isomorphism as Ψ . We can define vector-valued β -th power spaces [Li et al., 2022, Definition 4]:

$$[\mathcal{G}]^\beta := \Psi(S_2([\mathcal{H}_{ZO,1}]^\beta, \mathcal{H}_X)) = \{F | F = \Psi(C), C \in S_2([\mathcal{H}_{ZO,1}]^\beta, \mathcal{H}_X)\}. \quad (5)$$

The space $[\mathcal{G}]^\beta$ generalizes the definition of scalar-valued *power space* to vector-valued RKHSs, quantifying the smoothness of F_* relative to the RKHS \mathcal{G} (see Eq. (79)). We refer the reader to Carmeli et al. [2006, 2010] for definitions and properties of more general vector-valued RKHSs.

Given $\mathcal{D}_1 = \{(\tilde{\mathbf{z}}_i, \tilde{\mathbf{o}}_i, \tilde{\mathbf{x}}_i)\}_{i=1}^{\tilde{n}}$ sampled i.i.d from the joint distribution P_{XZO} , a regularized estimator of F_* is obtained as the solution to the following optimization problem:

$$\hat{F}_\xi := \arg \min_{F \in \mathcal{G}} \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \|\phi_X(\tilde{\mathbf{x}}_i) - F(\tilde{\mathbf{z}}_i, \tilde{\mathbf{o}}_i)\|_{\mathcal{H}_X}^2 + \xi \|F\|_{\mathcal{G}}^2, \quad (6)$$

where $\xi > 0$ denotes the Stage I regularization parameter.

Stage II. In Stage II, we perform regularized least squares regression in the RKHS $\mathcal{H}_{XO,2}$, using features derived from the estimated conditional mean embedding \hat{F}_ξ . Specifically, the features are $\hat{F}_\xi(Z, O) \otimes \phi_{O,2}(O)$. Given $\mathcal{D}_2 = \{(\mathbf{z}_i, \mathbf{o}_i, y_i)\}_{i=1}^n$ i.i.d sampled from the joint distribution P_{ZOY} and independent of \mathcal{D}_1 , the regularized estimator \hat{f}_λ is defined as:

$$\hat{f}_\lambda := \arg \min_{f \in \mathcal{H}_{XO,2}} \frac{1}{n} \sum_{i=1}^n \left(y_i - \langle f, \hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) \otimes \phi_{O,2}(\mathbf{o}_i) \rangle_{\mathcal{H}_{XO,2}} \right)^2 + \lambda \|f\|_{\mathcal{H}_X \otimes \mathcal{H}_{O,2}}^2, \quad (7)$$

where $\lambda > 0$ denotes the Stage II regularization parameter.* Owing to the favourable properties of kernel ridge regression, \hat{f}_λ admits a closed-form expression, given in [Section B](#) in the Supplement. Upon learning \hat{f}_λ , the quantity $\hat{f}_\lambda(\mathbf{x}^*, \mathbf{o}^*)$ represents the estimated heterogenous dose response of a new treatment \mathbf{x}^* on a new individual with observed covariates \mathbf{o}^* . The estimated dose response curve evaluated at \mathbf{x}^* can then be obtained as the expectation of $\hat{f}_\lambda(\mathbf{x}^*, O)$ with respect to the marginal distribution of the observed covariates.

Our primary goal is to study the $L^2(P_{XO})$ -risk:

$$\|\hat{f}_\lambda - f_*\|_{L^2(P_{XO})}. \quad (8)$$

To this end, we need to impose regularity conditions on the regression targets in both stages. Specifically, we characterize the regularity of the conditional mean embedding $F_* : \mathcal{Z} \times \mathcal{O} \rightarrow \mathcal{H}_X$ through a *dominating mixed-smoothness Sobolev space*, as discussed in [Section 2.2](#); and we characterize the regularity of the function $f_* : \mathcal{X} \times \mathcal{O} \rightarrow \mathbb{R}$ through an *anisotropic Besov space*, as introduced in [Section 2.3](#). It is thus natural to use two different kernels $k_{O,1}$ and $k_{O,2}$, because the regularity of F_* and f_* with respect to \mathcal{O} might not be the same. Since the choice of kernel in both stages is dependent on the regularity of their respective regression targets F_* and f_* , we provide a more in-depth description and justification of the kernels we use in stages I and II in Remarks [2.1](#) and [2.2](#) respectively.

2.2 Mixed-smoothness Sobolev spaces

In this section, we introduce vector-valued mixed-smoothness Sobolev spaces to characterize the smoothness of the conditional mean embedding (CME) $F_* : (\mathbf{z}, \mathbf{o}) \mapsto \mathbb{E}[\phi_X(X) \mid Z = \mathbf{z}, O = \mathbf{o}]$ in Stage I. In fact, the smoothness of F_* can be identified via the differentiability of the conditional density.

Let $(\mathbb{N}^+)^d$ be the set of all multi-indices $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d)$ with $\alpha_i \in \mathbb{N}$ and $|\boldsymbol{\alpha}| = \sum_{i=1}^d \alpha_i$. For $\boldsymbol{\alpha} \in \mathbb{N}^d$ and $f : \mathbb{R}^d \rightarrow \mathbb{R}$, $\partial^{\boldsymbol{\alpha}}$ denotes the classical (pointwise) partial derivative, and $D^{\boldsymbol{\alpha}}f$ denotes the corresponding weak (distributional) partial derivative.

Assumption 2.2. *Let $m_o, m_z \in \mathbb{N}^+$. For any $\mathbf{x} \in \mathcal{X}$, the map $(\mathbf{z}, \mathbf{o}) \mapsto p(\mathbf{x} \mid \mathbf{z}, \mathbf{o})$ has bounded, continuous derivatives of order m_o with respect to \mathbf{o} and order m_z with respect to \mathbf{z} on the interior of $\mathcal{Z} \times \mathcal{O}$.*

$$\rho := \max_{|\boldsymbol{\alpha}| \leq m_z} \max_{|\boldsymbol{\beta}| \leq m_o} \sup_{\mathbf{x} \in \mathcal{X}, \mathbf{z} \in \mathcal{Z}, \mathbf{o} \in \mathcal{O}} \left| \partial_{\mathbf{z}}^{\boldsymbol{\alpha}} \partial_{\mathbf{o}}^{\boldsymbol{\beta}} p(\mathbf{x} \mid \mathbf{z}, \mathbf{o}) \right| < \infty.$$

The differentiability conditions on the conditional density imposed in [Assumption 2.2](#) imply that F_* belong to a certain vector-valued dominating mixed-smoothness Sobolev space, as defined below.

Definition 1 (Vector-valued dominating mixed-smoothness Sobolev space). *Let H be a Hilbert space. Let $m_z, m_o \in \mathbb{N}^+$. We define*

$$MW_2^{m_z, m_o}(\mathcal{Z} \times \mathcal{O}; H) := \left\{ F \mid F \in L^2(\mathcal{Z} \times \mathcal{O}; H), \|F\|_{MW_2^{m_z, m_o}(\mathcal{Z} \times \mathcal{O}; H)} < \infty \right\}.$$

where $\|F\|_{MW_2^{m_z, m_o}(\mathcal{Z} \times \mathcal{O}; H)} := \sum_{|\boldsymbol{\alpha}| \leq m_z} \sum_{|\boldsymbol{\beta}| \leq m_o} \|D_{\mathbf{z}}^{\boldsymbol{\alpha}} D_{\mathbf{o}}^{\boldsymbol{\beta}} F\|_{L^2(\mathcal{Z} \times \mathcal{O}; H)}$.

*The naive extension to observed covariates in [KIV Singh et al. \[2019\]](#) considers augmenting X, Z to $(X, O), (Z, O)$. This approach is not consistent because Stage I would then require estimating the conditional mean embedding $(\mathbf{z}, \mathbf{o}) \mapsto \mathbb{E}[\phi_X(X) \otimes \phi_O(O) \mid Z = \mathbf{z}, O = \mathbf{o}]$, which is *not* Hilbert-Schmidt and for which vector-valued kernel ridge regression is not consistent, see also [\[Mastouri et al., 2021, Appendix B.9\]](#) for an illustration.

The real-valued dominating mixed-smoothness Sobolev space [Schmeisser, 1987, 2007, Sickel and Ullrich, 2009] $MW_2^{m_z, m_o}(\mathcal{Z} \times \mathcal{O}; \mathbb{R})$ is a special case of $MW_2^{\tilde{m}_z, \tilde{m}_o}(\mathcal{Z} \times \mathcal{O}; H)$ when $H = \mathbb{R}$. When $d_z = 0$ (or $d_o = 0$), we recover the vector-valued Sobolev spaces $W_2^{m_z}(\mathcal{Z}; H)$ (or $W_2^{m_o}(\mathcal{O}; H)$) as defined in Aubin [2011, Section 12.7]. Assumption 2.2 implies that $\|D_{\mathbf{z}}^{\boldsymbol{\alpha}} D_{\mathbf{o}}^{\boldsymbol{\beta}} F_*\|_{L^2(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_X)}$ is bounded for any multi-indices $|\boldsymbol{\alpha}| \leq m_z$ and $|\boldsymbol{\beta}| \leq m_o$. Hence, $F_* \in MW_2^{m_z, m_o}(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_X)$.

Now we are ready to state our choice of kernels $k_Z, k_{O,1}$ in Stage I.

Remark 2.1 (Choice of Stage I kernels $k_Z, k_{O,1}$). *We let k_Z and $k_{O,1}$ be any positive definite kernels such that their associated RKHSs $\mathcal{H}_Z, \mathcal{H}_{O,1}$ are respectively norm equivalent to real-valued Sobolev spaces $W_2^{t_z}(\mathcal{Z})$ and $W_2^{t_o}(\mathcal{O})$, where $t_z > \frac{d_z}{2}, t_o > \frac{d_o}{2}$. Following Chen et al. [2025], we say that $k_Z, k_{O,1}$ are Sobolev reproducing kernels of smoothness t_z, t_o . An important example of Sobolev reproducing kernel is the Matérn- ν kernel whose RKHS is norm equivalent to a Sobolev space W_2^t of smoothness $t = \nu + d/2$ [Wendland, 2004, Corollary 10.48]. Since all Sobolev reproducing kernels are bounded and measurable, $k_Z, k_{O,1}$ satisfy Assumption 2.1.*

With the above choice of k_Z and $k_{O,1}$, it follows from Lemma C.8 that the vector-valued RKHS \mathcal{G} associated with the operator-valued kernel in Eq. (4) is norm equivalent to the mixed-smoothness Sobolev space $MW^{t_z, t_o}(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_X)$. Since $F_* \in MW_2^{m_z, m_o}(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_X)$ as established above, F_* lies in the appropriate power space of \mathcal{G} for suitably chosen (t_z, t_o) . Consequently, Li et al. [2022], Meunier et al. [2024b] show that estimating the CME via Eq. (6) achieves the minimax-optimal rate in both the $L^2(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_X)$ and \mathcal{G} norms, provided that the regularization parameter ξ is selected adaptively with respect to the sample size \tilde{n} .

2.3 Anisotropic Besov spaces

In this section, we introduce the definition of anisotropic Besov spaces [Leisner, 2003], which is used to characterize the smoothness of f_* .

Definition 2 (Modulus of smoothness). *Let $\mathcal{X} = \prod_{i=1}^d \mathcal{X}^{(i)} \subseteq \mathbb{R}^d$ be a subset with non-empty interior, ν be a product measure on \mathcal{X} with $\nu = \otimes_{i=1}^d \nu_i$, and $f : \mathcal{X} \rightarrow \mathbb{R}$ be a function in $L^p(\nu)$ for some $p \in (0, \infty]$. The r -th modulus of smoothness of f is defined by*

$$\omega_{r,p}(f, \mathbf{t}, \mathcal{X}) = \sup_{0 < |h_i| \leq t_i} \|\Delta_{\mathbf{h}}^r f\|_{L^p(\nu)}, \quad (9)$$

where the r -th difference of f in the direction \mathbf{h} at point \mathbf{x} , denoted as $\Delta_{\mathbf{h}}^r f(\mathbf{x})$, is defined through recursion: $\Delta_{\mathbf{h}}^0 f(\mathbf{x}) := f(\mathbf{x})$ and $\Delta_{\mathbf{h}}^r f(\mathbf{x}) := \Delta_{\mathbf{h}}^{r-1} f(\mathbf{x} + \mathbf{h}) - \Delta_{\mathbf{h}}^{r-1} f(\mathbf{x})$ if $\mathbf{x}, \mathbf{x} + \mathbf{h}, \dots, \mathbf{x} + r\mathbf{h} \in \mathcal{X}$ and 0 otherwise.

Definition 3 (Anisotropic Besov space $B_{p,q}^{\mathbf{s}}(\nu)$). *For $p \in [1, \infty)$, $q \in [1, \infty]$ and $\mathbf{s} = (s_1, \dots, s_d) \in \mathbb{R}_+^d$, the anisotropic Besov space $B_{p,q}^{\mathbf{s}}(\nu)$ is defined by*

$$B_{p,q}^{\mathbf{s}}(\nu) := \left\{ f \in L^p(\nu) : \|f\|_{B_{p,q}^{\mathbf{s}}(\nu)} := \|f\|_{L^p(\nu)} + |f|_{B_{p,q}^{\mathbf{s}}(\nu)} < \infty \right\}, \quad (10)$$

where the Besov semi-norm $|f|_{B_{p,q}^{\mathbf{s}}(\nu)}$ is defined as,

$$|f|_{B_{p,q}^{\mathbf{s}}(\nu)} := \left[\int_0^1 \left[t^{-1} \omega_{r,p}(f, t^{1/s_1}, \dots, t^{1/s_d}, \mathcal{X}) \right]^q \frac{dt}{t} \right]^{1/q}, \quad (11)$$

for $r = \max\{\lfloor s_1 \rfloor, \dots, \lfloor s_d \rfloor\} + 1$. When $q = \infty$, we replace the integral by a supremum in Eq. (11). When ν is the Lebesgue measure over \mathcal{X} , we use the notation $B_{p,q}^{\mathbf{s}}(\mathcal{X}) := B_{p,q}^{\mathbf{s}}(\nu)$.

If $s_1 = \dots = s_d = s$, then the anisotropic Besov space recovers the standard isotropic Besov space [DeVore and Popov, 1988, DeVore and Sharpley, 1993]. Since f_* takes as input both the treatment X and observed covariate O , it naturally exhibits different smoothness with respect to X and O . Hence, as opposed to an isotropic Besov space which imposes uniform smoothness along all directions, an anisotropic Besov space captures such heterogeneous regularity. To simplify the exposition, we focus on anisotropic smoothness across X and O , while assuming isotropic smoothness within X and within O . In other words, we only consider $\mathbf{s} = (s_x, \dots, s_x, s_o, \dots, s_o) \in \mathbb{R}_{\geq 0}^{d_x + d_o}$ and denote $B_{2,\infty}^{\mathbf{s}}(\mathcal{X} \times \mathcal{O})$ as $B_{2,\infty}^{s_x, s_o}(\mathcal{X} \times \mathcal{O})$. Let $\mathbb{U}(B_{2,\infty}^{s_x, s_o}(\mathbb{R}^{d_x + d_o}))$ denote the unit ball of $B_{2,\infty}^{s_x, s_o}(\mathbb{R}^{d_x + d_o})$ with respect to the Besov norm. Let $C^0(\mathbb{R}^{d_x + d_o})$ denote the space of continuous functions $\mathbb{R}^{d_x + d_o} \rightarrow \mathbb{R}$.

Assumption 2.3. $f_* \in \mathfrak{S}$ where we define $\mathfrak{S} := \mathbb{U}(B_{2,\infty}^{s_x, s_o}(\mathbb{R}^{d_x + d_o})) \cap L^\infty(\mathbb{R}^{d_x + d_o}) \cap L^1(\mathbb{R}^{d_x + d_o}) \cap C^0(\mathbb{R}^{d_x + d_o})$.

In particular, under **Assumption 2.3**, $\|f_*\|_{L^2(\mathbb{R}^{d_x + d_o})} \leq \|f_*\|_{B_{2,\infty}^{s_x, s_o}(\mathbb{R}^{d_x + d_o})} \leq 1$. Moreover, by continuity, for all $\mathbf{o} \in \mathbb{R}^{d_o}$ the slice function $f_*(\cdot, \mathbf{o})$ is well-defined. We assume $p = 2$ in accordance with our $L^2(P_{X,O})$ -norm learning risk in Eq. (8). We assume $q = \infty$ because $B_{2,\infty}^{s_x, s_o}(\mathcal{X} \times \mathcal{O})$ is the largest anisotropic Besov space among all $B_{2,q}^{s_x, s_o}(\mathcal{X} \times \mathcal{O})$ spaces [Triebel, 2011]. To the best of our knowledge, we are the first to consider anisotropic smoothness in the NPIV literature.

We are now ready to state our choice of kernels $k_X, k_{O,2}$ in Stage II.

Remark 2.2 (Choice of Stage II kernels $k_X, k_{O,2}$). *We choose k_X and $k_{O,2}$ to be Gaussian kernels k_{γ_x} and k_{γ_o} with bandwidths $\gamma_x \in (0, 1), \gamma_o \in (0, 1)$. Denote \mathcal{H}_{γ_x} and \mathcal{H}_{γ_o} as the associated Gaussian RKHSs; ϕ_{γ_x} and ϕ_{γ_o} as the associated feature maps. The tensor product RKHS $\mathcal{H}_{\gamma_x, \gamma_o} := \mathcal{H}_{\gamma_x} \otimes \mathcal{H}_{\gamma_o}$ is the unique RKHS associated with the product kernel*

$$k_{\gamma_x}(\mathbf{x}, \mathbf{x}') \cdot k_{\gamma_o}(\mathbf{o}, \mathbf{o}') = \exp \left(- \sum_{j=1}^{d_x} \frac{(x_j - x'_j)^2}{\gamma_x^2} - \sum_{j=1}^{d_o} \frac{(o_j - o'_j)^2}{\gamma_o^2} \right). \quad (12)$$

This kernel is called an anisotropic Gaussian kernel and its associated RKHS $\mathcal{H}_{\gamma_x, \gamma_o}$ is the corresponding anisotropic Gaussian RKHS. Hang and Steinwart [2021] proves that kernel ridge regression with anisotropic Gaussian kernel in the form of Eq. (12) is minimax optimal for anisotropic Besov space target functions, provided that both the regularization parameter and the kernel lengthscale γ_x, γ_o are adaptive to the number of samples n . Such adaptivity will also be evident in our setting (see **Theorem 4.1**). Singh et al. [2019] adopt the median heuristic for selecting the kernel lengthscale in KIV algorithm, a widely used practical choice. Unlike ours, however, the theoretical relationship between their heuristic and the underlying smoothness of the target function remains unclear.

Remark 2.3 (Why use different kernels in Stage I and Stage II). *We briefly explain the rationale for selecting different types of kernels for Stage I and Stage II. For Stage I, the regression target is the conditional mean embedding F_* . By **Assumption 2.2**, we have shown that F_* belongs to the mixed Sobolev space $MW^{m_z, m_o}(\mathcal{Z} \times \mathcal{O}; \mathcal{H})$, which can be learned at the minimax optimal rate using a tensor-product Sobolev RKHS (see **Proposition D.4**). On the other hand, the Stage II regression target f_* belongs to an anisotropic Besov space (**Assumption 2.3**). Hang and Steinwart [2021] has proved that learning an anisotropic function in a nonparametric regression setting is minimax optimal via an anisotropic Gaussian RKHS. We have followed their approach with additional refinements to our setting, that reveals the interplay between the effect of T and the anisotropic smoothness of f_* . See **Section 4** for the details.*

3 Related work

Early NPIV literature focuses on series estimators [Newey and Powell, 2003, Blundell et al., 2007, Chen, 2007, Horowitz, 2011] and methods based on kernel density estimation [Hall and Horowitz, 2005, Darolles et al., 2011, Florens et al., 2011]. These works established minimax optimal convergence rates under various ill-posedness and smoothness conditions [Hall and Horowitz, 2005, Chen and Reiss, 2011, Chen and Christensen, 2018]. Recent NPIV algorithms leverage modern machine learning techniques, including RKHSs [Singh et al., 2019, Zhang et al., 2023b, Meunier et al., 2024a] and neural networks [Hartford et al., 2017, Bennett et al., 2019, Xu et al., 2021a, Petrushev et al., 2024, Kim et al., 2025, Sun et al., 2025, Meunier et al., 2025]. These modern methods mainly fall into two categories: two-stage estimation and min-max optimization. Min-max approaches [Bennett et al., 2019, Dikkala et al., 2020, Liao et al., 2020, Bennett et al., 2023, Zhang et al., 2023b, Wang et al., 2022] formulate NPIV as a saddle point optimization problem, which can be unstable and may fail to converge, especially when deep neural networks are used as function classes. In contrast, two-stage methods—such as the KIV-O algorithm studied in this manuscript (Section 2.1)—first estimate the conditional expectation operator T , and then perform a second-stage regression using the learned operator [Hartford et al., 2017, Singh et al., 2019, Xu et al., 2021a, Li et al., 2024b, Kim et al., 2025]. One recent paper [Kankanala, 2025] employs a sieve estimator in the first stage and a Gaussian process (a Bayesian analogue of an RKHS) estimator in the second stage.

In the introduction, we outlined two challenges for the theoretical analysis of NPIV-O. The first challenge concerns the fact that T is an identity operator restricted to the infinite dimensional function space \mathcal{F}_1 (defined in Eq. (3)). This has the following consequences. The *L^2 -stability condition* imposed in the NPIV literature (cf. Blundell et al. [2007, Assumption 6], Chen and Pouzo [2012, Assumption 5.2(ii)] and Chen and Christensen [2018, Assumption 4.2]) fails to hold except for the degenerate case where the sieve measure of ill-posedness is 1 (i.e. $Z = X$, see also Remark 4.2). The *link condition* imposed in the optimal rate literature for NPIV (cf. Hall and Horowitz [2005], Chen and Reiss [2011, Assumption 2.2], Chen and Christensen [2018, Condition LB]) implies that $\|Tf\|_{L^2(P_{ZO})} \leq \|B^r f\|_{L^2(P_{XO})}$ for some known compact operator B , where a larger r corresponds to a more ill-posed model. However, for any $f \in \mathcal{F}_1$ defined in Eq. (3), we have $\|Tf\|_{L^2(P_{ZO})} = \|f\|_{L^2(P_{XO})}$ so the link condition only holds with $r = 0$.

The second challenge lies in deriving a unified analysis where f_* lies in an anisotropic Besov space. Several prior works on NPIV-O and nonparametric proxy methods (see Section 3.1) Singh et al. [2019], Mastouri et al. [2021], Singh [2020], Bozkurt et al. [2025a], Hall and Horowitz [2005], Bozkurt et al. [2025b] choose instead to assume the generalized source condition $f_* \in \mathcal{R}((T^*T)^\beta)$ for some $\beta \geq 0$. However, as we have critiqued in the introduction, such an approach does not shed light on the separate contribution of the intrinsic smoothness of f_* and the smoothing effect of T . It also suffers from a lack of interpretability since T is a priori unknown. We also mention that Hall and Horowitz [2005] derived optimal rates for a kernel density based estimator for NPIV-O, where the smoothness of $f_*(\cdot, \mathbf{o})$ is characterized via a generalized source condition with respect to the partial conditional expectation operator $T_{\mathbf{o}}$ [Hall and Horowitz, 2005, Section 4.3] for $\mathbf{o} \in \mathcal{O}$. Such an assumption suffers from similar drawbacks to those outlined above, and it is moreover unclear how f_* 's smoothness in the direction of O impacts learning rates.

To the best of our knowledge, our paper is the first theoretical analysis that simultaneously addresses the anisotropic smoothness of both f_* and the operator T in the X and O directions. We address both challenges by (i) introducing a novel Fourier-based measure of partial smoothing of T , and (ii) employing Gaussian kernel lengthscales that adapt to the anisotropic smoothness of f_* .

3.1 Proximal causal learning (PCL)

The two challenges mentioned above also arise in a recent popular framework called proximal causal learning (PCL), which has gained considerable interest as a framework to identify and estimate causal effects from observational data, where the analyst only has access to imperfect proxies of the true underlying confounding mechanism without being able to observe the confounders directly [Miao et al., 2018, Tchetgen Tchetgen et al., 2024]. Our contributions to NPIV-O can directly be extended to this context. In the context of PCL, the (heterogeneous) dose response curve f^* can be identified either via the *outcome bridge function* [Miao et al., 2018, Deaner, 2018, Mastouri et al., 2021, Xu et al., 2021b, Kallus et al., 2021, Singh, 2020], which generalizes outcome regression, or via the *treatment bridge function* [Cui et al., 2024, Kallus et al., 2021, Bozkurt et al., 2025a,b], which generalizes inverse propensity weighting estimators. Analogous to the modern NPIV literature, the nonparametric estimators for bridge functions fall under the 2SLS approach [Deaner, 2018, Mastouri et al., 2021, Singh, 2020, Bozkurt et al., 2025a,b], min-max optimization approach with either RKHS or deep neural networks as function classes [Mastouri et al., 2021, Ghassami et al., 2022, Kallus et al., 2021], or via spectral methods [Sun et al., 2025]. Notably, both outcome bridge function and treatment bridge function are identified via conditional moment constraints of the same form as NPIV-O (see Eq. (2)), thus our theory in NPIV-O could be extended to the estimation of bridge functions in PCL.

3.2 Kernel ridge regression (KRR)

Our theoretical analysis of KIV-O builds on and extends existing theory in kernel ridge regression (KRR). The literature on KRR primarily follows two methodological lines: one based on *empirical process* [Steinwart and Christmann, 2008, Steinwart et al., 2009, Eberts and Steinwart, 2013, Hang and Steinwart, 2021, Hamm and Steinwart, 2021] and one based on *integral operator* techniques [De Vito et al., 2005, Smale and Zhou, 2005, 2007, Blanchard and Mücke, 2018, Lin and Cevher, 2020, Fischer and Steinwart, 2020, Zhang et al., 2023a, 2024]. In the context of learning an anisotropic Besov space function f_* using KRR, the only available convergence rate is provided by Hang and Steinwart [2021], which builds on an oracle inequality derived using empirical process techniques and hence necessitates a clipping operation on the KRR estimator. In our work, we are the first to remove the clipping step by leveraging integral operator techniques to directly control the finite sample estimation error. Moreover, our analysis leverages state of the art results on optimal rates for vector-valued kernel ridge regression [Li et al., 2024a, Meunier et al., 2024b] to bound the statistical error arising from estimating a conditional mean embedding.

4 Theory

This section presents our main theoretical results on the non-asymptotic convergence rate of the learning risk defined in Eq. (8). Section 4.1 presents our assumptions on the conditional expectation operator T . Section 4.2 presents an upper bound. Section 4.3 presents a minimax lower bound.

4.1 Partial smoothing effect of T

The challenge of estimating f_* via the inverse problem $\mathbb{E}[Y | Z, O] = (Tf_*)(Z, O)$ arises from its ill-posed nature: a small error in estimating $\mathbb{E}[Y | Z, O]$ may lead to a large error in estimating f_* . To address this challenge, we make the following assumptions. The first assumption enables unique identification of f_* .

Assumption 4.1 (L^∞ -completeness). *The conditional distribution $P_{X|Z,O}$ satisfies that, for every bounded measurable function $f : \mathcal{X} \times \mathcal{O} \rightarrow \mathbb{R}$, if $\mathbb{E}[f(X, O) | Z, O] = 0$ holds $P_{Z,O}$ -almost surely, then $f(X, O) = 0$ holds $P_{X,O}$ -almost surely.*

Assumption 4.1, known as bounded completeness or L^∞ -completeness [D'Haultfoeuille, 2011, Blundell et al., 2007], is weaker than the L^2 -completeness condition, which assumes that $T : L^2(P_{X,O}) \rightarrow L^2(P_{Z,O})$ is injective. The latter is standard in the NPIV literature [Newey and Powell, 2003, Hall and Horowitz, 2005, Carrasco et al., 2007, Darolles et al., 2011, Andrews, 2017, Chen et al., 2014, Chen and Christensen, 2018, Chen et al., 2024]. Although we do not assume that the outcome Y is bounded (see Assumption 4.5), the target heterogenous response curve f_* is bounded (assumed in Assumption 2.3), hence it suffices to impose the weaker L^∞ -completeness identification. We refer the reader to Andrews [2017], D'Haultfoeuille [2011] for sufficient conditions on the conditional distribution $P_{X|Z,O}$ such that bounded completeness holds.

Beyond identification, to establish a non-asymptotic rate of convergence for f_* , existing work on NPIV imposes additional assumptions on the smoothing properties of T which are not compatible with the partial identity structure of T imposed by the common variable O [Blundell et al., 2007, Chen and Christensen, 2018, Chen and Reiss, 2011]. In contrast, as highlighted in Section 1, with the existence of observed covariates O , our T exhibits characteristics of a compact operator in the X direction and acts as an identity operator in the O direction. We thus propose a novel framework to characterize the *partial* smoothing properties of T .

We describe this partial smoothing in terms of the Fourier representation of a function f on which T acts. For $f \in L^1(\mathbb{R}^{d_x})$, its Fourier transform is defined as a Lebesgue integral [Rudin, 1987, 9.1]: $\hat{f}(\cdot) = \int_{\mathbb{R}^{d_x}} f(\mathbf{x}) \exp(-i\langle \mathbf{x}, \cdot \rangle) d\mathbf{x}$. One can extend the Fourier transform to $L^2(\mathbb{R}^{d_x})$ by defining it as a *unitary* operator on $L^2(\mathbb{R}^{d_x})$ [Rudin, 1987, Theorem 9.13]. We use \mathcal{F} to denote this operator, and let \mathcal{F}^{-1} denote its inverse. In particular, $\mathcal{F}^{-1}[\mathbb{1}[A]]$ is well-defined, where $\mathbb{1}[A]$ denotes the indicator function of a compact set $A \subset \mathbb{R}^{d_x}$. For any scalar $\gamma \in (0, 1)$, we define the following two sets of functions:

$$\begin{aligned} \text{LF}(\gamma) &:= \{f : \mathbb{R}^{d_x+d_o} \rightarrow \mathbb{R} \mid \forall \mathbf{o} \in \mathcal{O}, [f(\cdot, \mathbf{o})] \in L^2(\mathbb{R}^{d_x}), \\ &\quad \text{supp}(\mathcal{F}[f(\cdot, \mathbf{o})]) \subseteq \{\boldsymbol{\omega}_x \in \mathbb{R}^{d_x} : \|\boldsymbol{\omega}_x\|_2 \leq \gamma^{-1}\}\}, \\ \text{HF}(\gamma) &:= \{f : \mathbb{R}^{d_x+d_o} \rightarrow \mathbb{R} \mid \forall \mathbf{o} \in \mathcal{O}, [f(\cdot, \mathbf{o})] \in L^2(\mathbb{R}^{d_x}), \\ &\quad \text{supp}(\mathcal{F}[f(\cdot, \mathbf{o})]) \subseteq \{\boldsymbol{\omega}_x \in \mathbb{R}^{d_x} : \|\boldsymbol{\omega}_x\|_2 \geq \gamma^{-1}\}\}. \end{aligned} \tag{13}$$

where supp for an element of $L^2(\mathbb{R}^{d_x})$ is defined in Definition 5 and Definition 6 in the Supplementary. The set $\text{LF}(\gamma)$ (respectively, $\text{HF}(\gamma)$) consists of functions such that for every $\mathbf{o} \in \mathcal{O}$, the slice function $f(\cdot, \mathbf{o})$ belongs to $L^1(\mathbb{R}^{d_x})$ and its Fourier transform is supported inside (respectively, outside) the centered ball of radius γ^{-1} in the Fourier domain. See Figure 1 for an illustration.

Assumption 4.2 (Fourier measure of partial ill-posedness of T). *There exists a constant $c_0 > 0$ and a parameter $\eta_0 \in [0, \infty)$ depending only on T , such that for all $\gamma \in (0, 1)$ and all functions $f \in \text{LF}(\gamma) \cap L^\infty(P_{X,O})$, the following inequality is satisfied:*

$$\|Tf\|_{L^2(P_{Z,O})} \geq c_0 \gamma^{d_x \eta_0} \|f\|_{L^2(P_{X,O})}.$$

In particular, c_0 does not depend on γ .

Assumption 4.3 (Fourier measure of partial contractivity of T). *There exists a constant $c_1 > 0$ and a parameter $\eta_1 \in [0, \infty)$ depending only on T , such that for all $\gamma \in (0, 1)$ and all functions*

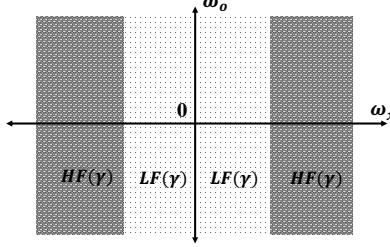


Figure 1: Illustration for $\text{LF}(\gamma)$ and $\text{HF}(\gamma)$.

$f \in \text{HF}(\gamma) \cap L^\infty(P_{XO})$, the following inequality is satisfied:

$$\|Tf\|_{L^2(P_{Z_O})} \leq c_1 \gamma^{d_x \eta_1} \|f\|_{L^2(P_{XO})}.$$

In particular, c_1 does not depend on γ .

Assumptions 4.2 and 4.3 are assumptions about the conditional distribution $P(X | Z, O)$. In Section A in the Supplement, for any positive integer $k \geq 1$, we construct a distribution $P_k(X, Z, O)$ such that, for the conditional expectation operator T defined by $P_k(X | Z, O)$, a weaker version of Assumptions 4.3 and 4.2 is satisfied with $\eta_0 = \eta_1 = k$ (where we restrict to considering functions $f(\mathbf{x}, \mathbf{o}) = g(\mathbf{x})h(\mathbf{o})$, and impose a further technical restriction for Assumption 4.2). If Assumption 4.2 and 4.3 hold simultaneously, and P_{XO} is absolutely continuous with respect to the Lebesgue measure on $\mathcal{X} \times \mathcal{O}$, then $\eta_0 \geq \eta_1$ and $c_0 \leq c_1$ (Lemma A.1 in Section A). In the remainder of the manuscript, we assume that the distribution P_{ZXOY} is fixed and we set the constants $c_0 = c_1 = 1$ for notational simplicity. Assumption 4.2 and Assumption 4.3 characterize the *mildly ill-posed* regime in the NPIV literature.

Assumption 4.3 quantifies the *partial smoothing* effect of T on the high-frequency components of a function f with respect to X ; while Assumption 4.2 captures the *partial anti-smoothing* behaviour of T on the low-frequency components of a function f with respect to X . When the treatment X is exogenous and we take $X = Z$, then T is an identity mapping so $\eta_0 = \eta_1 = 0$ and NPIV-O reduces to non-parametric regression from (X, O) to Y .

To motivate the partial smoothing effect of T , notice that the bounded self-adjoint operator $T^*T : L^2(P_{XO}) \rightarrow L^2(P_{XO})$ acts on $f \in L^2(P_{XO})$ as follows:

$$((T^*T)f)(\mathbf{x}', \mathbf{o}) = \int_{\mathcal{X}} f(\mathbf{x}, \mathbf{o}) L(\mathbf{x}, \mathbf{x}'; \mathbf{o}) d\mathbf{x}, \quad (14)$$

where $L(\mathbf{x}, \mathbf{x}'; \mathbf{o}) := \int_{\mathcal{Z}} p(\mathbf{x} | \mathbf{z}, \mathbf{o}) p(\mathbf{z} | \mathbf{x}', \mathbf{o}) d\mathbf{z}$. Consider two subsets of $L^2(P_{XO})$.

$$\begin{aligned} \mathfrak{G}_X &= \left\{ g \in L^2(P_{XO}) \mid \exists \tilde{g} \in L^2(P_X) \text{ such that } \forall \mathbf{x} \in \mathcal{X}, \mathbf{o} \in \mathcal{O}, g(\mathbf{x}, \mathbf{o}) = \tilde{g}(\mathbf{x}) \right\}, \\ \mathfrak{G}_O &= \left\{ g \in L^2(P_{XO}) \mid \exists \tilde{g} \in L^2(P_O) \text{ such that } \forall \mathbf{x} \in \mathcal{X}, \mathbf{o} \in \mathcal{O}, g(\mathbf{x}, \mathbf{o}) = \tilde{g}(\mathbf{o}) \right\}. \end{aligned}$$

Under mild conditions on the conditional distribution $p(\mathbf{x} | \mathbf{z}, \mathbf{o})$ [Darolles et al., 2011, Assumption A.1], $T^*T|_{\mathfrak{G}_X}$ (T^*T restricted to \mathfrak{G}_X) is compact and its smoothing effect can be quantified through its eigenvalue decay; while in contrast, $T^*T|_{\mathfrak{G}_O}$ is an identity operator: for $g \in \mathfrak{G}_O$

$$((T^*T)g)(\mathbf{x}, \mathbf{o}) = \int_{\mathcal{X}} g(\mathbf{x}, \mathbf{o}) L(\mathbf{x}, \mathbf{x}'; \mathbf{o}) d\mathbf{x} = \tilde{g}(\mathbf{o}) \int_{\mathcal{X}} L(\mathbf{x}, \mathbf{x}'; \mathbf{o}) d\mathbf{x} = g(\mathbf{x}, \mathbf{o}). \quad (15)$$

Therefore, when we incorporate observed covariates O , the conditional expectation operator T acts as a compact operator in the X direction and as an identity operator in the O direction. As a result, we propose to characterize its *partial* smoothing properties through measure of *partial* contractivity (Assumption 4.3) and measure of *partial* ill-posedness (Assumption 4.2).

In the literature on NPIV, conditions similar to [Assumption 4.3](#) and [Assumption 4.2](#) have been employed to quantify the smoothing effect of T . For instance, [Chen and Reiss \[2011\]](#), [Meunier et al. \[2024a\]](#) use the so-called link condition and reverse link condition which relate the smoothness of the hypothesis space with that of $\mathcal{R}(T^*T)$; [Chen and Christensen \[2018\]](#), [Blundell et al. \[2007\]](#), [Chen et al. \[2024\]](#) employ the sieve measure of ill-posedness and stability conditions, which quantify the smoothing effect of T on functions in the hypothesis space spanned by a sieve basis. Our [Assumption 4.3](#) and [Assumption 4.2](#) share strong resemblance with the latter. To see why, recall the definition of Gaussian RKHS \mathcal{H}_{γ_x} with length-scale γ_x through Fourier transforms [[Wendland, 2004](#), Theorem 10.12]:

$$\mathcal{H}_{\gamma_x} = \left\{ f : \mathbb{R}^{d_x} \rightarrow \mathbb{R} \mid \int_{\mathbb{R}^{d_x}} |\mathcal{F}[f](\omega_x)|^2 \exp\left(\frac{1}{4}\gamma_x^2\|\omega_x\|_2^2\right) d\omega_x < \infty \right\},$$

where we can see that for $f \in \mathcal{H}_{\gamma_x}$, the bulk of its Fourier spectrum would belong to the ball $\{\omega_x : \|\omega_x\|_2 \leq \gamma_x^{-1}\}$ with the remaining spectrum decaying exponentially as $\|\omega_x\|_2 \rightarrow \infty$. We formulate our [Assumption 4.3](#) and [Assumption 4.2](#) with Fourier transforms rather than Gaussian RKHSs for potential applications beyond Gaussian RKHSs. A closely related work is [Kankanala \[2025\]](#), which employs a *local* sieve measure of ill-posedness for functions in the RKHS. Unfortunately, these conditions, including ours, are hard to verify in practice as T is unknown.

Remark 4.1 (Connection with sieve measure of ill-posedness). *In this remark, we connect [Assumption 4.2](#) to the sieve measure of ill-posedness condition employed in the analysis of sieve 2SLS [[Blundell et al., 2007](#), [Chen and Christensen, 2018](#), [Chen et al., 2024](#), [Kim et al., 2025](#)]. For this remark, we omit observed covariates O , and take $T : L^2(P_X) \rightarrow L^2(P_Z)$. The sieve measure of ill-posedness is defined as*

$$\tau_J^{\text{sieve}} := \sup_{0 \neq f \in \Psi_J} \frac{\|f\|_{L^2(P_X)}}{\|Tf\|_{L^2(P_Z)}}, \quad (16)$$

where Ψ_J denotes the J th sieve space for the treatment variable [[Chen and Christensen, 2018](#), Section 3]. For this remark, we let Ψ_J be the linear span of cardinal B-splines of order \mathfrak{m} up to resolution \mathfrak{K} with $J = (2^{\mathfrak{K}} + \mathfrak{m} + 1)^{d_x} \asymp 2^{\mathfrak{K}d_x}$ [[DeVore and Lorentz, 1993](#), Section 5]. We note that the parameter γ^{-1} , where γ occurs in the definition of the function space $\text{LF}(\gamma)$ in Eq. (13), plays a role analogous to the resolution level \mathfrak{K} for cardinal B-splines. Indeed, $\text{LF}(\gamma)$ for smaller values of γ (or Ψ_J for larger values of J) correspond to a class of less smooth functions. The above observation and the form of Eq. (16) thus suggests an analogous definition:

$$\tau_\gamma^{\text{Fourier}} := \sup_{0 \neq f \in \text{LF}(\gamma) \cap L^\infty(P_{XO})} \frac{\|f\|_{L^2(P_X)}}{\|Tf\|_{L^2(P_Z)}}. \quad (17)$$

We can thus restate [Assumption 4.2](#) as: there exists a constant $c_0 > 0$ and a parameter $\eta_0 \in [0, \infty)$ depending only on T , such that for all $\gamma \in (0, 1)$, the following inequality is satisfied: $\tau_\gamma^{\text{Fourier}} \leq c_0^{-1}\gamma^{-d_x\eta_0}$. In the sieve NPIV literature [[Blundell et al., 2007](#), [Chen et al., 2024](#), [Chen and Christensen, 2018](#)], an NPIV model is said to be mildly ill-posed if $\tau_J^{\text{sieve}} = O(J^\eta) = O(2^{\eta \mathfrak{K}d_x})$. By the analogy between $\tau_\gamma^{\text{Fourier}}$ and τ_J^{sieve} , we see that our [Assumption 4.2](#) characterizes a mildly ill-posed regime.

Remark 4.2 (Existing treatment of partial smoothing of T in NPIV-O). *Existing work that concerns NPIV-O in the literature circumvents the partial identity structure of T either by imposing additional structural assumptions [[Blundell et al., 2007](#), [Syrgkanis et al., 2019](#)] or by stratifying the*

problem on O [Horowitz, 2011] thereby reducing NPIV- O to NPIV, which is statistically inefficient and scales poorly with the dimension of O . Instead of T , Chen and Christensen [2018, Section 3.3] consider the compactness and the smoothing effect of the partial conditional expectation operator $T_{\mathbf{o}} : L^2(P_{X|O=\mathbf{o}}) \rightarrow L^2(P_{Z|O=\mathbf{o}})$ for each $\mathbf{o} \in \mathcal{O}$. Chen and Christensen [2018, Section 3.3] proves that, if Ψ_J (defined in the above remark) equals the span of the first J eigenvectors of $T_{\mathbf{o}}^* T_{\mathbf{o}}$ for any $\mathbf{o} \in \mathcal{O}$, then

$$\tau_J^{\text{sieve}} \asymp \mathbb{E}_{\mathbf{o} \sim O} [\mu_{J,\mathbf{o}}^2]^{-\frac{1}{2}}, \quad (18)$$

where $\mu_{J,\mathbf{o}}$ is the J th singular value of $T_{\mathbf{o}}$ arranged in non-increasing order. Chen and Christensen [2018] then claims that the convergence rates derived for the standard NPIV can be extended to NPIV- O . However, we identify an essential oversight: Chen and Christensen [2018, Assumption 4 (ii)], which is required for their L^2 -norm upper bound, cannot hold in the NPIV- O setting. This assumption states that for Ψ_J , f_* satisfies the following inequality:

$$\tau_J^{\text{sieve}} \|T(f_* - \Pi_J f_*)\|_{L^2(P_{ZO})} \leq \|f_* - \Pi_J f_*\|_{L^2(P_{XO})},$$

where $\Pi_J : L^2(P_{XO}) \rightarrow \Psi_J$ denotes the L^2 -orthogonal projection from $L^2(P_{XO})$ onto Ψ_J . In the NPIV- O setting, if f_* depends only on \mathbf{o} and P_{XO} is the Lebesgue measure, then the projection $\Pi_J f_*$ also depends only on \mathbf{o} . This implies $(T^* T)(f_* - \Pi_J f_*) = f_* - \Pi_J f_*$ as per Eq. (15), forcing $\tau_J^{\text{sieve}} = O(1)$, causing a contradiction with the measure of ill-posedness condition that $\tau_J^{\text{sieve}} = \mathcal{O}(J^{\eta_0})$ unless $\eta_0 = 0$. Hall and Horowitz [2005] addresses NPIV- O by assuming smoothness of $f_*(\cdot, \mathbf{o})$ relative to $\mathcal{R}(T_{\mathbf{o}}^* T_{\mathbf{o}})$ for each $\mathbf{o} \in \mathcal{O}$, which is hard to interpret because $T_{\mathbf{o}}$ is unknown in NPIV- O .

4.2 Upper Bound for KIV-O

To obtain the upper learning rate for KIV-O in Section 2.1, we need to impose the following assumptions about the data-generating distribution.

Assumption 4.4. 1. The joint probability measures P_{ZO} and P_{XO} admit probability density functions p_{ZO} and p_{XO} . There exists a universal constant $a > 0$ such that $a^{-1} \geq p_{ZO}(\mathbf{z}, \mathbf{o}) \geq a$ for all $(\mathbf{z}, \mathbf{o}) \in [0, 1]^{d_z+d_o}$ and $a^{-1} \geq p_{XO}(\mathbf{x}, \mathbf{o})$ for all $(\mathbf{x}, \mathbf{o}) \in [0, 1]^{d_x+d_o}$ and $p_{XO}(\mathbf{x}, \mathbf{o}) \geq a$ for all $(\mathbf{x}, \mathbf{o}) \in [1/4, 3/4]^{d_x+d_o}$. 2. Assumption 2.2 holds with the following $m_z, m_o \in \mathbb{N}^+$:

$$m_o := \left\lceil \frac{\frac{d_o}{2} \frac{1+2\left(\frac{s_x}{d_x} + \eta_1\right) + \frac{d_o}{s_o}\left(\frac{s_x}{d_x} + \eta_1\right)}{1+2\left(\frac{s_x}{d_x} + \eta_1\right)}}{d_o} \right\rceil + 1, \quad m_z := \left\lceil \frac{d_z}{d_o} m_o \right\rceil. \quad (19)$$

The requirement that $p_{XO}(\mathbf{x}, \mathbf{o}) \geq a$ for all $(\mathbf{x}, \mathbf{o}) \in [1/4, 3/4]^{d_x+d_o}$ is a mild assumption to ensure that the $\mathcal{H}_{\gamma_x, \gamma_o}$ -norm of the kernel mean embedding of P_{XO} is bounded away from zero. This plays a role in the control of the estimation error for Stage II regression (see Proposition D.8 in the Supplement). The choice of $1/4, 3/4$ is arbitrary and can be replaced by any fixed unequal values in $(0, 1)$. The constraint on m_o depends on (s_x, s_o) because the embedding norm of the RKHS \mathcal{H}_{FO} (defined in Section C) into the mixed-smoothness Sobolev space scales as $\gamma_o^{-m_o}$, where γ_o itself depends on both s_x and s_o in Theorem 4.1; this embedding norm must be controlled, a requirement referred to as the (EMB) condition in Fischer and Steinwart [2020]. See Eq. (89) for details.

Assumption 4.5. For all $(\mathbf{z}, \mathbf{o}) \in \mathcal{Z} \times \mathcal{O}$, the residual $v := Y - (Tf_*)(Z, O)$ is subgaussian conditioned on $Z = \mathbf{z}, O = \mathbf{o}$ with subgaussian norm at most σ .

In the NPIV literature, existing work assumes a moment condition on the residual v [Blundell et al., 2007, Hall and Horowitz, 2005, Chen and Christensen, 2018, Chen and Reiss, 2011], which is weaker than our Assumption 4.5. However, their corresponding high probability upper bounds only guarantee polynomially decaying tails, as a result of Chebyshev's inequality (see, for example, the proof of Lemma F.9 in Chen and Christensen [2018], p.40). In contrast, our upper bound holds in high probability with *subexponential* tails. Our sharper guarantee is a consequence of our applying Bernstein concentration inequality and the more advanced techniques in the analysis of kernel ridge regression [Fischer and Steinwart, 2020, Eberts and Steinwart, 2013, Hang and Steinwart, 2021].

Now we are ready to state the upper learning rate of $\|\hat{f}_\lambda - f_*\|_{L^2(P_{XO})}$. We remind the reader that \tilde{n}, n denote respectively the number of Stage 1 and Stage 2 samples. We let Stage I kernels $k_{O,1}, k_Z$ be Matérn kernels whose associated RKHSs $\mathcal{H}_{O,1}$ and \mathcal{H}_Z are respectively norm equivalent to $W_2^{t_o}(\mathcal{O})$ and $W_2^{t_z}(\mathcal{Z})$, for t_o, t_z to be specified. We let Stage II kernels $k_X, k_{O,2}$ be Gaussian kernels with respective lengthscales γ_x, γ_o .

Theorem 4.1 (Upper learning rate for KIV-O). *Suppose Assumptions 2.2, 4.4 hold with parameters $m_z, m_o \in \mathbb{N}^+$, Assumption 2.3 holds with parameters $s_x, s_o > 0$, Assumptions 4.2, 4.3 hold with parameters η_0, η_1 . We further suppose Assumptions 4.1 and 4.5 hold. We assume t_o, t_z satisfy $2t_o \geq m_o > t_o > d_o/2$, $2t_z \geq m_z > t_z > d_z/2$. We let*

$$\gamma_x = n^{-\frac{\frac{1}{d_x}}{1+2(\frac{s_x}{d_x}+\eta_1)+\frac{d_o}{s_o}(\frac{s_x}{d_x}+\eta_1)}}, \quad \gamma_o = n^{-\frac{\frac{1}{s_o}(\frac{s_x}{d_x}+\eta_1)}{1+2(\frac{s_x}{d_x}+\eta_1)+\frac{d_o}{s_o}(\frac{s_x}{d_x}+\eta_1)}}. \quad (20)$$

Define $m^\dagger := (m_z t_z^{-1}) \wedge (m_o t_o^{-1})$ and $d^\dagger := (d_z t_z^{-1}) \vee (d_o t_o^{-1})$. Let Stage I regularization parameter ξ be given by $\tilde{n}^{-\frac{1}{m^\dagger+d^\dagger+\zeta}}$ for any $\zeta > 0$; and Stage II regularization parameter λ be given by n^{-1} . Suppose that $n \geq 1$ is sufficiently large, and $\tilde{n} \geq 1$ satisfies

$$\tilde{n} \gtrsim n^{\frac{m^\dagger+d^\dagger/2+\zeta}{m^\dagger-1}} \vee n^{2\frac{m^\dagger+d^\dagger/2+\zeta}{m^\dagger}}. \quad (21)$$

Then with $P^{n+\tilde{n}}$ -probability at least $1 - 40e^{-\tau}$, we have

$$\left\| \left[\hat{f}_\lambda \right] - f_* \right\|_{L^2(P_{XO})} \lesssim \tau n^{-\frac{\frac{s_x}{d_x}+\eta_1-\eta_0}{1+2(\frac{s_x}{d_x}+\eta_1)+\frac{d_o}{s_o}(\frac{s_x}{d_x}+\eta_1)}} \cdot (\log n)^{\frac{d_x+d_o+1+d_x\eta_0}{2}}. \quad (22)$$

Remark 4.3. In Theorem 4.1, we present the regime where the Stage I sample size \tilde{n} is sufficiently large relative to the Stage II sample size (see Eq. (21)). This is the appropriate regime where we can study rate-optimality because we present a minimax lower bound in Theorem 4.2 with respect to the class of estimators that only utilize Stage II samples. It is nevertheless possible to derive upper bounds with respect to both n and \tilde{n} without restrictions on the relative size of n, \tilde{n} . It remains a challenging open problem, however, to establish rate optimality for estimators utilizing a split dataset of the form $\{(\tilde{\mathbf{z}}_i, \tilde{\mathbf{o}}_i, \tilde{\mathbf{x}}_i)\}_{i=1}^{\tilde{n}}$ and $\{(\mathbf{z}_i, \mathbf{o}_i, y_i)\}_{i=1}^n$, even for the standard NPIV setting [Chen and Reiss, 2011, Chen and Christensen, 2018, Meunier et al., 2024a].

Remark 4.4 (Interpolation between NPIV and non-parametric regression). *Our derived upper rate interpolates between the known optimal L^2 -rates for NPIV without observed covariates and anisotropic kernel ridge regression.*

- When $\eta_0 = \eta_1 = 0$ and $X = Z$, i.e. T is the identity mapping, our setting reduces to nonparametric regression where the target function lies in the anisotropic Besov space $B_{2,\infty}^{s_x, s_o}(\mathcal{X} \times \mathcal{O})$. The upper rate simplifies to $\tilde{\mathcal{O}}_P(n^{-\frac{1}{2\tilde{s}+1}})$ with $\tilde{s} = (d_o/s_o + d_x/s_x)^{-1}$ being the intrinsic smoothness,

which matches the known optimal learning rate of regression with an anisotropic Besov target function [Hoffman and Lepski, 2002, Hang and Steinwart, 2021].

2. When $d_o = 0$, our setting reduces to NPIV without observed covariates where the target function f_* lies in an isotropic Besov space $B_{2,\infty}^{s_x}(\mathcal{X})$. We take $\eta_1 = \eta_0 = \eta$ following Chen and Christensen [2018], Chen and Reiss [2011] which employs a single parameter to characterize both the ill-posedness and contractivity, then our upper learning rate simplifies to $\tilde{\mathcal{O}}_P(n^{-\frac{s_x}{d_x+2(s_x+\eta d_x)}})$, which matches the known optimal rate in NPIV regression [Chen and Christensen, 2018, Corollary 3.1].

4.2.1 Proof sketch

The proof of [Theorem 4.1](#) is given in [Section D](#) in the Supplement. Here we give an outline of our proof to facilitate a deeper understanding of both the assumptions and the results. Define \bar{f}_λ as the oracle estimator for Stage II with access to the true conditional mean embedding F_* and recall \hat{f}_λ for comparison:

$$\bar{f}_\lambda := \arg \min_{f \in \mathcal{H}_{\gamma_x, \gamma_o}} \lambda \|f\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 + \frac{1}{n} \sum_{i=1}^n (y_i - \langle f, F_*(\mathbf{z}_i, \mathbf{o}_i) \otimes \phi_{O, \gamma_o}(\mathbf{o}_i) \rangle_{\mathcal{H}_{\gamma_x, \gamma_o}})^2. \quad (23)$$

$$\hat{f}_\lambda := \arg \min_{f \in \mathcal{H}_{\gamma_x, \gamma_o}} \lambda \|f\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 + \frac{1}{n} \sum_{i=1}^n (y_i - \langle f, \hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) \otimes \phi_{O, \gamma_o}(\mathbf{o}_i) \rangle_{\mathcal{H}_{\gamma_x, \gamma_o}})^2. \quad (24)$$

The proof can be summarized in 3 steps. We upper bound $\|T[\hat{f}_\lambda] - T[\bar{f}_\lambda]\|_{L^2(P_{ZO})}$ in *Step 1* and upper bound $\|T[\bar{f}_\lambda] - Tf_*\|_{L^2(P_{ZO})}$ in *Step 2*, which induce an upper bound on $\|T[\hat{f}_\lambda] - Tf_*\|_{L^2(P_{ZO})}$ via a triangular inequality. In *Step 3*, we apply the partial measure of ill-posedness to obtain an upper bound on $\|\hat{f}_\lambda - f_*\|_{L^2(P_{ZO})}$. We highlight our technical contributions in each step with [Remark 4.5](#) and [Remark 4.6](#).

Step 1 We upper bound $\|T[\hat{f}_\lambda] - T[\bar{f}_\lambda]\|_{L^2(P_{ZO})}$. By their definition in Eq. (24) and Eq. (23), the discrepancy between \hat{f}_λ and \bar{f}_λ arises solely from the difference between \hat{F}_ξ and F_* ; hence it corresponds to Stage I error. First, we prove in [Proposition D.3](#) that $\|T[\hat{f}_\lambda] - T[\bar{f}_\lambda]\|_{L^2(P_{ZO})}$ can be upper bounded by an expression involving $\|\hat{F}_\xi - F_*\|_{\mathcal{G}}$ and $\|F_* - \hat{F}_\xi\|_{L^2(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_{X, \gamma_x})}$, where we recall that \mathcal{G} denotes the unique vector-valued RKHS induced by the operator-valued kernel K defined in Eq. (4). To obtain high-probability upper bounds on both of these quantities, we adapt the existing optimal learning rates on CME from [Li et al. \[2022\]](#) to our setting with a tensor product RKHS. The fact that we require upper learning rate for $\|F_* - \hat{F}_\xi\|_{\mathcal{G}}$ imposes the conditions $m_o > t_o$ and $m_z > t_z$, so the CME F_* lies in a smoother space than RKHS \mathcal{G} . Specifically, F_* belongs to the *power space* $[\mathcal{G}]^{m^\dagger}$ (See Eq. (79)). In addition, the constraints $2t_o > m_o$ and $2t_z > m_z$ reflect the saturation effect inherent in Tikhonov regularization [[Bauer et al., 2007](#), [Lu et al., 2024](#), [Meunier et al., 2024b](#)]. These constraints can be removed to allow for greater smoothness of F_* by employing spectral regularization [[Meunier et al., 2024b](#)]. The appearance of an arbitrarily small $\zeta > 0$ in Stage I regularization parameter ξ reflects the fact that, after reordering, the eigenvalues of the tensor product operator exhibit slower decay than those of the individual components, owing to an extra logarithmic term [[Krieg, 2018](#)].

Remark 4.5 (Tensor product kernel ridge regression). *In the above step, we generalize the upper learning rate for vector-valued kernel ridge regression to the setting of tensor product kernels (See [Proposition D.4](#)). Although tensor product kernels have been widely used in kernel-based hypothesis tests [[Gretton et al., 2007](#), [Gretton and Gyorfi, 2010](#), [Sejdinovic et al., 2013](#), [Gretton, 2015](#), [Zhang et al., 2018](#), [Albert et al., 2022](#), [Szabó and Sriperumbudur, 2018](#)], kernel independent component*

analysis [Bach and Jordan, 2002, Shen et al., 2009], and feature selection [Song et al., 2012, Li et al., 2021], they have been less well studied in kernel ridge regression, with the exception of Hang and Steinwart [2021] for Gaussian kernels. Our analysis is also applicable to real-valued kernel ridge regression with tensor product kernels.

Step 2 We upper bound $\|T[\bar{f}_\lambda] - Tf_*\|_{L^2(P_{ZO})}$. We follow the approach in Blanchard and Mücke [2018], Meunier et al. [2024a], where it is observed (in the standard NPIV case) that this term corresponds to the learning risk of a kernel ridge regression problem with an appropriately defined RKHS $\mathcal{H}_{FO} \subseteq \{\mathcal{Z} \times \mathcal{O} \rightarrow \mathbb{R}\}$, namely the RKHS induced by the feature map $(\mathbf{z}, \mathbf{o}) \mapsto F_*(\mathbf{z}, \mathbf{o}) \otimes \phi_{O,2}(\mathbf{o})$. We refer the reader to Section C in the Supplement for the definition of \mathcal{H}_{FO} . Our construction is adapted from Meunier et al. [2024a, Appendix E.1.2]. However, unlike Blanchard and Mücke [2018], Meunier et al. [2024a] who only consider fixed RKHSs, we employ tensor product Gaussian RKHS $\mathcal{H}_{\gamma_x, \gamma_o}$ with adaptive length-scales γ_x, γ_o (as in Eq. (20)) to capture the anisotropic smoothness of $f_* \in B_{2,\infty}^{s_x, s_o}(\mathcal{X} \times \mathcal{O})$ [Hang and Steinwart, 2021]. When $\eta_1 = 0$, our choice of length-scales γ_x, γ_o coincides with that of kernel ridge regression in Hang and Steinwart [2021]. The logarithmic factor $(\log n)^{\frac{d_x+d_o+1}{2}}$ in Eq. (22) arises from the entropy numbers of the Gaussian RKHS in this step (see Section C.1 in the Supplement and Hang and Steinwart [2021, Proposition 1]).

Remark 4.6 (Gaussian kernel ridge regression with Besov space target functions). *To establish learning rates for kernel ridge regression, there are two main techniques in the literature: the empirical process technique [Steinwart and Christmann, 2008, Steinwart et al., 2009] and the integral operator technique [Fischer and Steinwart, 2020, Lin et al., 2020, Smale and Zhou, 2007, Caponnetto and De Vito, 2007, Blanchard and Mücke, 2018]. Previous works on Gaussian kernel ridge regression with Besov space targets [Eberts and Steinwart, 2013, Hang and Steinwart, 2021, Hamm and Steinwart, 2021] rely on an oracle inequality proved via empirical process techniques [Steinwart and Christmann, 2008, Theorem 7.23], which necessitates a clipping operation on the estimator. On the other hand, the integral operator technique avoids the clipping operation, but it requires the target f_* in a power space of the RKHS—a condition known as the source condition [Fischer and Steinwart, 2020, SRC]—which does not hold for Besov space targets and Gaussian RKHSs.*

In our proof in step 2, we prove an upper learning rate of Gaussian kernel ridge regression with Besov space target functions without clipping the estimator. Specifically, we combine the two techniques above, in that we bound the approximation error with Hang and Steinwart [2021], while we bound the estimation error with the integral operator technique [Fischer and Steinwart, 2020]. To be precise, define

$$f_\lambda := \arg \min_{f \in \mathcal{H}_{\gamma_x, \gamma_o}} \lambda \|f\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 + \|T([f] - f_*)\|_{L^2(P_{ZO})}^2. \quad (25)$$

The learning risk $\|T([\bar{f}_\lambda] - f_)\|_{L^2(P_{ZO})}$ can be decomposed into an estimation error term $\|T([\bar{f}_\lambda] - [f_\lambda])\|_{L^2(P_{ZO})}$ and an approximation error term $\|T([f_\lambda] - f_*)\|_{L^2(P_{ZO})}$. We upper bound the estimation error with Fischer and Steinwart [2020, Theorem 16], once we prove that the RKHS \mathcal{H}_{FO} satisfies an embedding property (see Fischer and Steinwart [2020, EMB]), which avoids the clipping operation. We upper bound the approximation error with Hang and Steinwart [2021, Theorem 4], which avoids the source condition.*

Step 3 We combine the above two terms $\|T[\hat{f}_\lambda] - T[\bar{f}_\lambda]\|_{L^2(P_{ZO})}$ and $\|T[\bar{f}_\lambda] - Tf_*\|_{L^2(P_{ZO})}$ through a triangle inequality, which gives an upper bound on the projected risk $\|T[\hat{f}_\lambda] - Tf_*\|_{L^2(P_{ZO})}$.

$$\|T[\hat{f}_\lambda] - Tf_*\|_{L^2(P_{ZO})} \lesssim (\log n)^{\frac{d_x+d_o+1}{2}} n^{-\frac{\frac{s_x}{d_x}+\eta_1}{1+2(\frac{s_x}{d_x}+\eta_1)+\frac{d_o}{s_o}(\frac{s_x}{d_x}+\eta_1)}}. \quad (26)$$

To bound the unprojected risk $\|[\hat{f}_\lambda] - f_*\|_{L^2(P_{ZO})}$, it seems all that remains is to apply the Fourier measure of partial ill-posedness in [Assumption 4.2](#) to remove T . Unfortunately, however, [Assumption 4.2](#) only holds for functions in $\text{LF}(\gamma)$ (Eq. (13)) with low partial Fourier frequency in \mathcal{X} . Notice that for a function $f \in \mathcal{H}_{\gamma_x, \gamma_o}$, its partial Fourier spectrum $|\mathcal{F}[f(\cdot, \mathbf{o})](\boldsymbol{\omega}_x)|$ decays exponentially fast as $\|\boldsymbol{\omega}_x\|_2 \rightarrow \infty$ [[Wendland, 2004](#), Theorem 10.12]

$$(\forall \mathbf{o} \in \mathcal{O}), \quad \int_{\mathbb{R}^{d_x}} |\mathcal{F}[f(\cdot, \mathbf{o})](\boldsymbol{\omega}_x)|^2 \exp\left(\frac{1}{4}\gamma_x^2\|\boldsymbol{\omega}_x\|_2^2\right) d\boldsymbol{\omega}_x < \infty. \quad (27)$$

This exponential decay implies that $\hat{f}_\lambda(\cdot, \mathbf{o})$ satisfies the conditions of [Assumption 4.2](#) up to some logarithmic factors as reflected by $(\log n)^{\frac{d_x \eta_0}{2}}$ in Eq. (22). For f_* , we find an auxiliary function $f_{\text{aux}} \in \mathcal{H}_{\gamma_x, \gamma_o}$ that is close to f_* in $L^2(P_{XO})$ -norm and agrees with f_* at low frequencies. To be precise, we require that $\text{supp}(\mathcal{F}[(f_{\text{aux}} - f_*)(\cdot, \mathbf{o})]) \subseteq \{\boldsymbol{\omega}_x : \|\boldsymbol{\omega}_x\|_2 \geq \gamma_x^{-1}\}$ for any $\mathbf{o} \in \mathcal{O}$, and we refer the reader to Eq. (40) for the exact definition of f_{aux} . Hence,

$$\|[\hat{f}_\lambda] - f_*\|_{L^2(P_{XO})} \lesssim \|[\hat{f}_\lambda] - f_{\text{aux}}\|_{L^2(P_{XO})} \lesssim \gamma_x^{-\eta_0 d_x} (\log n)^{\frac{d_x \eta_0}{2}} \|T[\hat{f}_\lambda] - T f_{\text{aux}}\|_{L^2(P_{ZO})}.$$

In the last step, we utilize the Fourier measure of partial ill-posedness in [Assumption 4.2](#), and the fact that $\hat{f}_\lambda, f_{\text{aux}} \in \mathcal{H}_{\gamma_x, \gamma_o}$. The above equation is a sketch where formal derivations can be found at the beginning of [Section D](#), particularly Eq. (74) and Eq. (75). Combining the above relation and the choice of γ_x in Eq. (20) concludes the proof of [Theorem 4.1](#).

Remark 4.7 (Extension to more anisotropy). *For simplicity of presentation, we focus on the case where anisotropic smoothness exists across (X, O) but we assume no anisotropy within X and O . The KIV-O algorithm with adaptive length-scales and its associated learning rate can both be easily extended to the fully anisotropic setting, where $f_* \in B_{2,\infty}^s(\mathcal{X} \times \mathcal{O})$ with $s = [s_1, \dots, s_{d_x}, s_{d_x+1}, \dots, s_{d_x+d_o}] \in \mathbb{R}^{d_x+d_o}$, a generalization of previous results from anisotropic non-parametric regression [[Hang and Steinwart, 2021](#), [Suzuki and Nitanda, 2021](#), [Hoffman and Lepski, 2002](#)] to anisotropic NPIV-O. This is particularly relevant in applied work, where the observed covariates O are often of high dimensionality, because practitioners tend to adjust for as many observed covariates as possible to mitigate unobserved confounding. In such cases, our KIV-O algorithm with adaptive length-scales adapts to the intrinsic smoothness with respect to O . This mitigates the slow rates typically caused by a high ambient dimension when the intrinsic smoothness is high, as our learning rates avoid being limited by the worst-case smoothness across all dimensions.*

4.3 Minimax lower bound for NPIV-O

In [Theorem 4.2](#), we prove a minimax lower bound for the NPIV-O problem. We call *admissible* a distribution P_{ZXOY} over (Z, X, O, Y) satisfying [Assumption 2.2](#), [Assumption 4.1](#), [Assumption 4.4](#) and [Assumption 4.5](#), inducing a model of the form

$$Y = f_*(X, O) + \epsilon, \quad \mathbb{E}[\epsilon | Z, O] = 0,$$

where f_* satisfies [Assumption 2.3](#) and the conditional expectation operator $T : L^2(P_{XO}) \rightarrow L^2(P_{ZO})$ satisfies [Assumption 4.3](#). For an *admissible* distribution P_{ZXOY} , consider $\mathcal{D} = (\mathbf{z}_i, \mathbf{x}_i, \mathbf{o}_i, y_i)_{i=1}^N, N \geq 1$ sampled i.i.d from P_{ZXOY} and consider $\mathcal{D}_1 = (\tilde{\mathbf{z}}_i, \tilde{\mathbf{o}}_i, \tilde{\mathbf{x}}_i)_{i=1}^{\tilde{n}}$ and $\mathcal{D}_2 = (\mathbf{z}_i, \mathbf{o}_i, y_i)_{i=1}^n$ with $\tilde{n}, n \leq N$.

Theorem 4.2 (Minimax lower bound). *There exists an admissible distribution P_{ZXOY} such that for all learning methods $(\mathcal{D}_1, \mathcal{D}_2) \mapsto \hat{f}_{(\mathcal{D}_1, \mathcal{D}_2)}$, for all $\tau > 0$, and sufficiently large $n \geq 1$, the following minimax lower bound holds with P^n -probability at least $1 - C_1 \tau^2$ and $P^{\tilde{n}}$ -almost surely ,*

$$\left\| \hat{f}_{(\mathcal{D}_1, \mathcal{D}_2)} - f_* \right\|_{L^2(P_{XO})} \geq C_0 \tau^2 n^{-\frac{s_x}{1+2(\frac{s_x}{d_x}+\eta_1)+\frac{d_o}{s_o}\frac{s_x}{d_x}}} (\log n)^{-2s_x-d_x}. \quad (28)$$

We remind the reader here of our upper bound in Eq. (22): with $P^{n+\tilde{n}}$ -probability at least $1 - 40e^{-\tau}$, we have

$$\left\| \left[\hat{f}_\lambda \right] - f_* \right\|_{L^2(P_{XO})} \lesssim \tau n^{-\frac{\frac{s_x}{d_x} + \eta_1 - \eta_0}{1+2(\frac{s_x}{d_x} + \eta_1) + \frac{d_o}{s_o}(\frac{s_x}{d_x} + \eta_1)}} \cdot (\log n)^{\frac{d_x + d_o + 1 + d_x \eta_0}{2}}.$$

There remains a gap between the upper and minimax lower bounds even if we take $\eta_1 = \eta_0$ and ignore the logarithmic terms. We dedicate [Section 5](#) to its discussion.

Remark 4.8 (Interpolation between NPIV and nonparametric regression). *Similar to the Remark 4.4 of the upper learning rate, our minimax lower learning rate also interpolates between the known optimal L^2 -rates for NPIV without observed covariates ($d_o = 0$) and anisotropic nonparametric regression ($\eta_1 = \eta_0 = 0$).*

Remark 4.9 (Comparison to existing L^2 -minimax lower bounds for NPIV regression). *The work Chen and Reiss [2011] established minimax rate-optimality in L^2 -norm for NPIV under an approximation condition and link condition. For a conveniently chosen compact operator B , the approximation condition characterizes the smoothness of the structural function by the optimal L^2 -rate approximation by the eigenvectors of B , and the link condition relates the mapping properties of the conditional expectation operator T to the Hilbert scale generated by B . This framework subsumes an earlier minimax convergence rate result in Hall and Horowitz [2005], and was subsequently instantiated in Chen and Christensen [2018] for a B constructed via a wavelet basis, and generalized to kernel instrumental variables in Meunier et al. [2024a] with a Hilbert scale given by the covariance operator of the RKHS. A crucial limitation of this framework is the link condition can only hold if the singular values of the conditional expectation operator are decaying to zero. Thus the minimax framework given by the link condition is applicable only when T is compact, and is not suitable for our observed covariates setting.*

We give an outline of our proof for [Theorem 4.2](#) to facilitate a deeper understanding of both the assumptions and the results. The full proof is in [Section E.2](#) in the Supplementary. The primary strategy in deriving minimax lower bounds is to construct a family of distributions that are similar enough so that they are statistically indistinguishable, but for which the target function of interest is maximally separated. This implies no estimator can have error uniformly smaller than this separation.

Following prior work in the literature [Chen and Reiss, 2011, Meunier et al., 2024a], we observe that NPIV-O in Eq. (1) is statistically more challenging than the reduced form non-parametric indirect regression with observed covariates (NPIR-O) with a *known* operator $T : L^2(P_{XO}) \rightarrow L^2(P_{ZO})$. The NPIR-O model is defined below

$$Y = (Tf_*)(Z, O) + v, \quad (29)$$

where v is a random variable such that $\mathbb{E}[v | Z, O] = 0$, and it satisfies [Assumption 4.5](#), and where T satisfies [Assumption 4.1](#), [Assumption 4.3](#) and the associated conditional distribution $P_{X|Z,O}$ satisfies [Assumption 2.2](#). We refer the reader to [Section E.1](#) in the Supplementary Material for a detailed definition of the NPIR-O model class. We formally prove in [Lemma E.1](#) in the Supplementary that it suffices to construct a minimax lower bound for the NPIR-O model. To this end, we adapt Theorem 20 of Fischer and Steinwart [2020], which is itself an adaptation of Proposition 2.3 of Tsybakov [2008]. Recall that the *Kullback-Leibler divergence* of two probability measures P_1, P_2 on some measurable space (Ω, \mathcal{A}) is given by $\text{KL}(P_1, P_2) := \int_{\Omega} \log(\frac{dP_1}{dP_2}) dP_1$ if P_1 is absolutely continuous with respect to P_2 , and $+\infty$ otherwise.

To apply Theorem 20 of Fischer and Steinwart [2020] on the measurable space $\Omega = (\mathcal{Z} \times \mathcal{O} \times \mathbb{R})^n$, we construct a family of probability measures P_0, P_1, \dots, P_M over $(\mathcal{Z} \times \mathcal{O} \times \mathbb{R})$ that share the same marginal distribution over (Z, O) but different conditional distributions $P_{Y|Z,O}$. Our strategy follows the construction in Chen and Reiss [2011] and can be explained as follows. We fix a marginal probability measure P_{ZO} over $\mathcal{Z} \times \mathcal{O}$. We then propose a family of conditional probability measures $P_{i;Y|Z,O}$ indexed by $f_i \in \mathfrak{F}$ for $0 \leq i \leq M$. Since f_i 's are functions on $\mathcal{X} \times \mathcal{O}$, we fix a marginal probability measure P_X on \mathcal{X} under the constraint that X is independent of O . Then, we define a smooth copula to parametrize the dependence between P_X and P_Z , which fully specifies $P_{X|Z,O}$. This induces a fixed conditional expectation operator $T : L^2(P_{XO}) \rightarrow L^2(P_{ZO})$. We then specify $P_{i;Y|Z,O}$ for $0 \leq i \leq M$ via the following equation:

$$Y | Z = \mathbf{z}, O = \mathbf{o} \sim \mathcal{N}(Tf_i)(\mathbf{z}, \mathbf{o}), \sigma^2,$$

for a fixed $\sigma > 0$, and for any $(\mathbf{z}, \mathbf{o}) \in \mathcal{Z} \times \mathcal{O}$. P_i denotes the probability measure on $\mathcal{Z} \times \mathcal{O} \times \mathbb{R}$ by coupling $P_{i;Y|Z,O}$ and P_{ZO} . In the above construction, we require $P_{X|Z,O}$ to satisfy [Assumption 2.2](#), and all f_i 's to satisfy [Assumption 2.3](#), namely $f_i \in \mathfrak{S}$.

Concretely, we fix P_{ZO} to be the Lebesgue measure over $\mathcal{Z} \times \mathcal{O}$. Without loss of generality, we take $\mathcal{X} = [-0.5, 0.5]^{d_x}$ rather than $[0, 1]^{d_x}$. By working with a symmetric domain, we exploit the fact that even functions have real-valued Fourier transforms, which simplify our subsequent calculations. We fix P_X via the density function

$$p_X(\mathbf{x}) \propto \prod_{i=1}^{d_x} g_i(x_i), \quad g_i(x_i) := \exp\left(-\frac{2}{1-4x_i^2}\right) \mathbb{1}_{x_i \in [-0.5, 0.5]}, \quad (30)$$

where each g_i is a smooth, compactly supported *bump function*. We fix an arbitrary smooth copula [Nelsen, 2006], such that $P_{X|Z,O}$ satisfies [Assumption 2.2](#) and T satisfies [Assumption 4.3](#) with parameter $\eta_1 > 0$.

The main technical challenge of the minimax lower bound is the construction of the function class \mathfrak{F} whose each element f_i induces a conditional distribution $P_{X|Z,O}$. Our goal is to design \mathfrak{F} so that it satisfies three desirable properties: 1) we want $\mathfrak{F} \in B_{2,\infty}^{s_x, s_o}(\mathbb{R}^{d_x + d_o})$ as per [Assumption 2.3](#). 2) we want the size of \mathfrak{F} to be large to enable the application of Fischer and Steinwart [2020, Theorem 20]. 3) we want all elements of \mathfrak{F} to exhibit high partial Fourier frequency such that after applying T , the resulting functions become difficult to distinguish. This increases the intrinsic difficulty of the problem and yields a tighter lower bound. The construction of such a function class satisfying the first two properties is standard and can be achieved using anisotropic B-splines [Ibragimov and Khas' minskii, 1984, Suzuki and Nitanda, 2021, Schmidt-Hieber, 2020]. However, the third property poses a challenge: the Fourier transform of B-splines has full support. Consequently, conventional B-splines cannot enforce the desired partial high-frequency restriction. To address this limitation, we convolve the B-splines with a high frequency bandpass filter in the X -direction. Since convolution acts as a smoothing operator, the function class constructed with such modified B-splines remains in the Besov space.

To start with, we introduce the definition of anisotropic B-splines [Leisner, 2003].

Definition 4 (Anisotropic B-spline [Leisner, 2003]). *The cardinal B-spline of order $\mathfrak{m} \in \mathbb{N}$ is defined as the repeated \mathfrak{m} times convolution $\iota_{\mathfrak{m}} = \iota_0 * \dots * \iota_0$ where ι_0 is the indicator function $\mathbb{1}_{[0,1]}$. Let d denote the ambient dimension, and let $\mathbf{s} = (s_1, \dots, s_d)$ denote a smoothness vector. Define the notation $\underline{s} = \min(s_1, \dots, s_d)$ and $s'_i := \frac{\underline{s}}{s_i}$ for $1 \leq i \leq d$. The (anisotropic) B-spline of order \mathfrak{m} with resolution $\mathfrak{K} \in \mathbb{N}_+$ and location vector $\ell \in \prod_{i=1}^d \{-\mathfrak{m}, -\mathfrak{m} + 1, \dots, 2^{\lfloor \mathfrak{K}s'_i \rfloor}\}$ is defined as $M_{\mathfrak{K}, \ell}(\mathbf{x}) = \prod_{i=1}^d \iota_{\mathfrak{m}}(2^{\lfloor \mathfrak{K}s'_i \rfloor} x_i - \ell_i)$.*

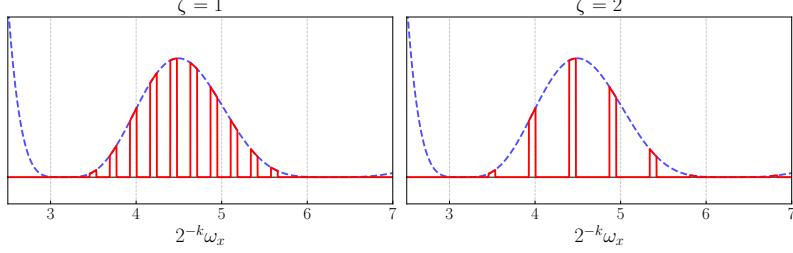


Figure 2: This figure illustrates the partial Fourier transform of f_v in Eq. (35), where we take $d_x = 1$, $\kappa = 2$, $m = 4$, $s_x = \underline{s} \leq s_o$. The dashed blue line represents the Fourier transform of the B-spline $M_{\kappa, -2}$. The red line shows the result of applying frequency masks to this B-spline: $\sum_{\ell_x} \beta_{v(\ell_x, \ell_o)} \mathcal{F}[M_{\kappa, -2}](\omega_x) \cdot \mathbb{1}_{\ell_x}(2^{-\kappa} \omega_x)$. The Fourier transform of $f_v(\cdot, \mathbf{o})$ is equal to this result up to scaling factors dependent only on \mathbf{o} .

Let $m > s_x \vee s_o$ be a non-negative even integer. The resolution $\kappa = \kappa(n)$ is defined later in Eq. (131), such that $\kappa \rightarrow \infty$ as $n \rightarrow \infty$. The B-spline basis in the O -direction follows the standard construction. Define the set of location vectors \mathcal{L}_O and the associated B-splines

$$\mathcal{L}_O := \left\{ 2m \mathbb{N} \cap \left\{ 0, \dots, 2^{\lfloor \frac{\kappa \frac{s}{s_o}}{2} \rfloor} \right\} \right\}^{d_o}, \quad M_{\kappa, \ell_o}(\mathbf{o}) = \prod_{j=1}^{d_o} \iota_m \left(2^{\lfloor \frac{\kappa \frac{s}{s_o}}{2} \rfloor} o_j - \ell_{o,j} \right). \quad (31)$$

Note that here we enforce all location vectors to be a multiple of $2m$ such that the $\{M_{\kappa, \ell_o}\}_{\ell_o \in \mathcal{L}_O}$ have disjoint supports, which will simplify the calculations later on. In contrast, the basis functions in the X -direction are constructed by convolving a standard B-spline $M_{0, -\frac{m}{2}}$ on X with the inverse Fourier transform of the indicator function $\mathbb{1}_{\ell_x}$, for $\ell_x \in \mathcal{L}_X$:

$$\mathcal{L}_X := \left\{ 0, 1, \dots, \left\lfloor \frac{0.8\pi}{\zeta} 2^{\frac{\kappa \frac{s}{s_x}}{2}} \right\rfloor \right\}^{d_x}, \quad \Omega_{\kappa \ell_x}(\mathbf{x}) := \left(M_{0, -\frac{m}{2}} * \mathcal{F}^{-1}[\mathbb{1}_{\ell_x}] \right) \left(2^{\frac{\kappa \frac{s}{s_x}}{2}} \mathbf{x} \right). \quad (32)$$

where $\mathbb{1}_{\ell_x}$ is the indicator function over the following hyper-rectangle:

$$I_{\ell_x} := \bigtimes_{j=1}^{d_x} \left[1.1\pi + \zeta \ell_{x,j} 2^{-\frac{\kappa \frac{s}{s_x}}{2}}, 1.1\pi + (\zeta \ell_{x,j} + 1) 2^{-\frac{\kappa \frac{s}{s_x}}{2}} \right]. \quad (33)$$

$\zeta > 0$ is a width hyperparameter that determines the spacing of different hyper-rectangles; its value will be specified later. For sufficiently large $n \geq 1$, since $\ell_{x,j} \leq \frac{0.8\pi}{\zeta} 2^{\frac{\kappa \frac{s}{s_x}}{2}}$, we know that $1.1\pi + (\zeta \ell_{x,j} + 1) 2^{-\frac{\kappa \frac{s}{s_x}}{2}} \leq 1.9\pi + 2^{-\frac{\kappa \frac{s}{s_x}}{2}} \leq 1.95\pi$. Thus we have $I_{\ell_x} \subseteq [1.1\pi, 1.95\pi]^{d_x}$.

The main consequence of this construction is revealed via the convolution theorem [Rudin, 1987, Theorem 9.2], which gives

$$\mathcal{F}[\Omega_{\kappa \ell_x}](\omega_x) = \mathcal{F}[M_{\kappa, -\frac{m}{2}}](\omega_x) \cdot \mathbb{1}_{\ell_x}(2^{-\frac{\kappa \frac{s}{s_x}}{2}} \omega_x). \quad (34)$$

As a result, we have that $\text{supp}(\mathcal{F}[\Omega_{\kappa \ell_x}]) = 2^{\frac{\kappa \frac{s}{s_x}}{2}} \cdot I_{\ell_x} \subseteq [\pi 2^{\frac{\kappa \frac{s}{s_x}}{2}}, 2\pi 2^{\frac{\kappa \frac{s}{s_x}}{2}}]^{d_x}$. Since $\kappa \rightarrow \infty$ as $n \rightarrow \infty$, the construction guarantees that $\Omega_{\kappa \ell_x}$ only has high-frequency spectrum, thereby fulfilling the third desirable property for \mathfrak{F} .

Define the basis $\Omega_{\kappa(\ell_x, \ell_o)}(\mathbf{x}, \mathbf{o}) := \Omega_{\kappa \ell_x}(\mathbf{x}) \cdot M_{\kappa, \ell_o}(\mathbf{o})$ and the set of all location vectors $\mathcal{L} = \mathcal{L}_X \times \mathcal{L}_O$. We have $|\mathcal{L}| \asymp \zeta^{-d_x} 2^{\kappa \frac{s}{s_x} (\frac{d_x}{s_x} + \frac{d_o}{s_o})}$. We construct a function class \mathfrak{F} as follows:

$$\mathfrak{F} := \left\{ f_{\mathbf{v}} : f_{\mathbf{v}}(\mathbf{x}, \mathbf{o}) = 2^{-\kappa \frac{s}{s_x} \left(1 - \frac{d_x}{2s_x} \right)} \sum_{(\ell_x, \ell_o) \in \mathcal{L}} \beta_{\mathbf{v}(\ell_x, \ell_o)} \Omega_{\kappa(\ell_x, \ell_o)}(\mathbf{x}, \mathbf{o}) \mid \mathbf{v} \in \{0, 1\}^{|\mathcal{L}|} \right\}, \quad (35)$$

where $\beta_{\mathbf{v}(\ell_x, \ell_o)} \in \{0, 1\}$ is the value assigned by $\mathbf{v} \in \{0, 1\}^{|\mathcal{L}|}$ to the location vector $(\ell_x, \ell_o) \in \mathcal{L}$. Specifically, for each $f_{\mathbf{v}} \in \mathfrak{F}$ and for a fixed $\mathbf{o} \in \mathcal{O}$, we can see that $f_{\mathbf{v}}(\cdot, \mathbf{o})$ is constructed by applying a sum of *frequency masks* indexed by $\ell_x \in \mathcal{L}_X$ to the original B-spline $M_{\mathfrak{K}, -\frac{m}{2}}$,

$$\mathcal{F}[f_{\mathbf{v}}(\cdot, \mathbf{o})](\omega_x) = \mathcal{F}\left[M_{\mathfrak{K}, -\frac{m}{2}}\right](\omega_x) \cdot \underbrace{\left(\sum_{\ell_o \in \mathcal{L}_O} \sum_{\ell_x \in \mathcal{L}_X} \beta_{\mathbf{v}(\ell_x, \ell_o)} \mathbb{1}_{\ell_x} \left(2^{-\frac{\mathfrak{K}s}{s_x}} \omega_x\right) M_{\mathfrak{K}, \ell_o}(\mathbf{o})\right)}_{\text{frequency masks}}.$$

Figure 2 is provided to illustrate this effect of frequency masking.

We prove in Eq. (108) in the Supplementary that $\mathfrak{F} \subseteq \mathbb{U}(B_{2,\infty}^{s_x, s_o}(\mathbb{R}^{d_x+d_o})) \cap L^\infty(\mathbb{R}^{d_x+d_o}) \cap L^1(\mathbb{R}^{d_x+d_o}) \cap C^0(\mathbb{R}^{d_x+d_o}) \subseteq \mathfrak{G}$, satisfying [Assumption 2.3](#). For each $f_{\mathbf{v}} \in \mathfrak{F}$, we construct the conditional distribution $P_{\mathbf{v}}(\cdot | \mathbf{z}, \mathbf{o}) := \mathcal{N}((T f_{\mathbf{v}})(\mathbf{z}, \mathbf{o}), \sigma^2)$ as the normal distribution on \mathbb{R} with mean $(T f_{\mathbf{v}})(\mathbf{z}, \mathbf{o})$ and variance σ^2 , such that $P_{\mathbf{v}}$ satisfies [Assumption 4.5](#). Together with the marginal distribution P_{ZO} over $Z \times O$, we have constructed a family of distributions $\{P_{\mathbf{v}}, \mathbf{v} \in \{0, 1\}^{|\mathcal{L}|}\}$ over $Z \times O \times \mathbb{R}$ with each $P_{\mathbf{v}}$ associated to a $f_{\mathbf{v}} \in \mathfrak{F}$. Next, by the Gilbert-Varshamov Bound [[Tsybakov, 2008, Lemma 2.9](#)], we prove in Eq. (123) and Eq. (127) in the Supplementary that there exists a subset $V_{\mathfrak{K}} \subseteq \{0, 1\}^{|\mathcal{L}|}$ with $|V_{\mathfrak{K}}| \geq 2^{\frac{L}{8}}$ such that for any $f_{\mathbf{v}}, f_{\mathbf{v}'} \in \mathfrak{F}_{\text{pruned}} = \{f_{\mathbf{v}} | \mathbf{v} \in V_{\mathfrak{K}}\}$, there is

$$\|f_{\mathbf{v}} - f_{\mathbf{v}'}\|_{L^2(P_{ZO})}^2 \geq 2^{-2\mathfrak{K}s} \zeta^{-d_x}, \quad \text{KL}(P_{\mathbf{v}}^{\otimes n}, P_{\mathbf{0}}^{\otimes n}) \leq n 2^{-2\mathfrak{K}s (\frac{d_x}{s_x} \eta_1 + 1)}.$$

Therefore, we have established that the function class $\mathfrak{F}_{\text{pruned}}$ satisfies all the three properties as desired, namely $\mathfrak{F}_{\text{pruned}} \subset \mathfrak{G}$, functions $f_{\mathbf{v}} \in \mathfrak{F}_{\text{pruned}}$ are well-separated in $L^2(P_{ZO})$ yet statistically indistinguishable. Finally, we take $\zeta = (\log n)^2$, we apply [Fischer and Steinwart \[2020, Theorem 20\]](#) and the general reduction scheme in [Tsybakov \[2008, Section 2.2\]](#) to obtain the desired minimax lower bound.

5 On the gap between the upper bound and minimax lower bound

As explained in [Section 4.3](#), in the general setting where $\frac{d_x}{s_x} > 0$ and $\frac{d_o}{s_o} > 0$, a gap arises between the upper bound given in [Theorem 4.1](#) and the minimax lower bound given in [Theorem 4.2](#). By setting $\eta_0 = \eta_1 = \eta$ in [Assumption 4.3](#) and [Assumption 4.2](#), which gives a precise characterization of the *partial* smoothing effect of T , we obtain the following upper and lower bounds (ignoring the logarithmic terms for simplicity):

$$\text{Lower Bound: } n^{-\frac{\frac{s_x}{d_x}}{1+2(\frac{s_x}{d_x}+\eta)+\frac{d_o}{s_o}\frac{s_x}{d_x}}}, \quad \text{Upper Bound: } n^{-\frac{\frac{s_x}{d_x}}{1+2(\frac{s_x}{d_x}+\eta)+\frac{d_o}{s_o}\frac{s_x}{d_x}+\frac{d_o}{s_o}\eta}}$$

Note that the denominator in the exponent of the upper bound contains an extra $\frac{d_o}{s_o}\eta$ term, which vanishes when T has no *partial* smoothing effect ($\eta = 0$) or when the ratio $\frac{d_o}{s_o}$ is zero.

The gap between the upper bound and the minimax lower bound arises due to the existence of observed covariates O in the analysis of the approximation error term in the upper bound,

$$\|[f_{\lambda}] - f_*\|_{L^2(P_{ZO})},$$

where f_{λ} is defined in Eq. (25). To provide an in-depth discussion, we begin with a warm-up in [Section 5.1](#) by considering the standard KIV setting without observed covariates, where our upper and minimax lower bounds match. We then move on to KIV-O in [Section 5.2](#) and demonstrate how the presence of observed covariates O causes a gap to emerge.

5.1 Analysis of the approximation error in KIV

In NPIV, the conditional expectation operator is $T : L^2(P_X) \rightarrow L^2(P_Z)$, $f \mapsto \mathbb{E}[f(X) | Z]$. Hence f_λ defined in Eq. (25) reduces to the following

$$f_\lambda := \arg \min_{f \in \mathcal{H}_{X,\gamma_x}} \lambda \|f\|_{\mathcal{H}_{X,\gamma_x}}^2 + \|T([f] - f_*)\|_{L^2(P_Z)}^2.$$

We first employ the fact that f_λ is the minimizer, hence for some $f_{\text{aux}} \in \mathcal{H}_{X,\gamma_x}$,

$$\|T([f_\lambda] - f_*)\|_{L^2(P_Z)}^2 \leq \lambda \|f_{\text{aux}}\|_{\mathcal{H}_{X,\gamma_x}}^2 + \|T([f_{\text{aux}}] - f_*)\|_{L^2(P_Z)}^2.$$

Here, f_{aux} is defined as $f_{\text{aux}} := f_{*,\text{low}} + K_{\gamma_x} * f_{*,\text{high}}$, where K_{γ_x} is defined, following [Hang and Steinwart \[2021\]](#), [Eberts and Steinwart \[2013\]](#), as

$$K_{\gamma_x}(\mathbf{x}) := \sum_{j=1}^r \binom{r}{j} (-1)^{1-j} \frac{1}{(j\gamma_x)^d} \left(\frac{2}{\pi}\right)^{\frac{d}{2}} \exp\left(-2 \sum_{i=1}^d \frac{x_i^2}{(j\gamma_x)^2}\right), \quad (36)$$

and $f_{*,\text{low}}, f_{*,\text{high}}$ are defined such that $f_{*,\text{low}}$ (resp. $f_{*,\text{high}}$) corresponds to the low-frequency (resp. high-frequency) component of $f_* = f_{*,\text{low}} + f_{*,\text{high}}$:

$$\begin{aligned} \mathcal{F}[f_{*,\text{low}}](\boldsymbol{\omega}_x) &= \mathcal{F}[f_*](\boldsymbol{\omega}_x) \cdot \mathbb{1}[\|\boldsymbol{\omega}_x\|_2 \leq \gamma_x^{-1}] \\ \mathcal{F}[f_{*,\text{high}}](\boldsymbol{\omega}_x) &= \mathcal{F}[f_*](\boldsymbol{\omega}_x) \cdot \mathbb{1}[\|\boldsymbol{\omega}_x\|_2 \geq \gamma_x^{-1}]. \end{aligned}$$

To see why $f_{\text{aux}} \in \mathcal{H}_{X,\gamma_x}$, note that $\|f_{*,\text{low}}\|_{\mathcal{H}_{X,\gamma_x}}^2 \lesssim \int_{\mathbb{R}^{d_x}} |\mathcal{F}[f_*](\boldsymbol{\omega}_x)|^2 d\boldsymbol{\omega}_x = \|f_*\|_{L^2(\mathbb{R}^{d_x})}^2 < \infty$ and $K_{\gamma_x} * f_{*,\text{high}} \in \mathcal{H}_{X,\gamma_x}$ proved by [Eberts and Steinwart \[2013\]](#). Then we have

$$\begin{aligned} \|T([f_\lambda] - f_*)\|_{L^2(P_Z)}^2 &\leq \lambda \|f_{\text{aux}}\|_{\mathcal{H}_{X,\gamma_x}}^2 + \|T(f_{*,\text{high}} - K_{\gamma_x} * f_{*,\text{high}})\|_{L^2(P_Z)}^2 \\ &\stackrel{(*)}{\leq} \lambda \|f_{\text{aux}}\|_{\mathcal{H}_{X,\gamma_x}}^2 + \gamma_x^{2d_x\eta} \|f_{*,\text{high}} - K_{\gamma_x} * f_{*,\text{high}}\|_{L^2(P_X)}^2 \\ &\stackrel{(**)}{\lesssim} \lambda \gamma_x^{-d_x} + \gamma_x^{2d_x\eta+2s_x}. \end{aligned} \quad (37)$$

where (*) follows by [Assumption 4.3](#), and (**) follows by the same derivations in [Eberts and Steinwart \[2013\]](#), Theorem 2.2, Theorem 2.3]. Eq. (37) is the key step which would require significant modifications in the setting of NPIV-O due to the existence of observed covariates. Finally, we have

$$\begin{aligned} &\|[f_\lambda] - f_*\|_{L^2(P_X)}^2 \\ &\leq \|[f_\lambda] - [f_{\text{aux}}]\|_{L^2(P_X)}^2 + \|[f_{\text{aux}}] - f_*\|_{L^2(P_X)}^2 \\ &\stackrel{(a)}{\leq} \gamma_x^{-2d_x\eta} \|T[f_\lambda] - T[f_{\text{aux}}]\|_{L^2(P_Z)}^2 + \|[f_{\text{aux}}] - f_*\|_{L^2(P_X)}^2 \\ &\leq \gamma_x^{-2d_x\eta} \left(\|T([f_\lambda] - f_*)\|_{L^2(P_Z)}^2 + \|T([f_{\text{aux}}] - f_*)\|_{L^2(P_Z)}^2 \right) + \|[f_{\text{aux}}] - f_*\|_{L^2(P_X)}^2 \\ &\stackrel{(b)}{\leq} \gamma_x^{-2d_x\eta} \left(\|T([f_\lambda] - f_*)\|_{L^2(P_Z)}^2 + \gamma_x^{2d_x\eta} \|[f_{\text{aux}}] - f_*\|_{L^2(P_X)}^2 \right) + \|[f_{\text{aux}}] - f_*\|_{L^2(P_X)}^2 \\ &\lesssim \gamma_x^{-2d_x\eta} \left(\lambda \gamma_x^{-d_x} + \gamma_x^{2d_x\eta+2s_x} \right) + 2 \|[f_{\text{aux}}] - f_*\|_{L^2(P_X)}^2 \stackrel{(c)}{\lesssim} n^{-\frac{2s_x}{d_x+2s_x+2\eta d_x}}. \end{aligned}$$

In the above derivations, (a) follows by an application of [Assumption 4.2](#) and [Lemma D.9](#) since $\|f_\lambda\|^2 \leq \lambda^{-1} \|Tf_*\|_{L^2(P_Z)}^2 \lesssim n$ by the optimality of f_λ and $\|f_{\text{aux}}\| \leq n$ by the same derivations as in Eq. (67) in the Supplement. The second term of Eq. (96) in the Supplement is subsumed by the following choice of γ_x . Furthermore, (b) follows from [Assumption 4.3](#), and (c) follows from using Eq. (84) in the Supplement and choosing $\lambda = n^{-1}$, $\gamma_x = n^{-\frac{1}{d_x+2s_x+2\eta d_x}}$. The upper bound on the approximation error above is *optimal* in the sense that it matches the minimax lower bound of NPIV with Besov targets (see e.g [Chen and Christensen, 2018](#) and our Theorem 4.2 with $d_o = 0$).

5.2 Analysis of the approximation error in KIV-O

We now proceed to our NPIV-O setting, and demonstrate how the existence of observed covariates O fundamentally changes the problem. As above, to upper bound the approximation error $\|[f_\lambda] - f_*\|_{L^2(P_{ZO})}$, we first need to upper bound the projected approximation error $\|Tf_\lambda - Tf_*\|_{L^2(P_{ZO})}$. To this end, we employ the fact that f_λ is the minimizer in Eq. (25), hence for some $f_{\text{aux}} \in \mathcal{H}_{\gamma_x, \gamma_o}$,

$$\|T([f_\lambda] - f_*)\|_{L^2(P_{ZO})}^2 \leq \lambda \|f_{\text{aux}}\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 + \|T([f_{\text{aux}}] - f_*)\|_{L^2(P_{ZO})}^2.$$

We construct such an auxiliary function f_{aux} . For any $\mathbf{o} \in \mathcal{O}$, we construct $f_{*,\text{low}}(\cdot, \mathbf{o})$ and $f_{*,\text{high}}(\cdot, \mathbf{o})$ such that

$$\mathcal{F}[f_{*,\text{low}}(\cdot, \mathbf{o})](\boldsymbol{\omega}_x) = \mathcal{F}[f_*(\cdot, \mathbf{o})](\boldsymbol{\omega}_x) \cdot \mathbb{1}[\boldsymbol{\omega}_x : \|\boldsymbol{\omega}_x\|_2 \leq \gamma_x^{-1}] \quad (38)$$

$$\mathcal{F}[f_{*,\text{high}}(\cdot, \mathbf{o})](\boldsymbol{\omega}_x) = \mathcal{F}[f_*(\cdot, \mathbf{o})](\boldsymbol{\omega}_x) \cdot \mathbb{1}[\boldsymbol{\omega}_x : \|\boldsymbol{\omega}_x\|_2 \geq \gamma_x^{-1}]. \quad (39)$$

Note that $f_{*,\text{high}} = f_* - f_{*,\text{low}}$. Despite $f_{*,\text{low}}(\cdot, \mathbf{o}) \in \mathcal{H}_{X, \gamma_x}$ for any $\mathbf{o} \in \mathcal{O}$, a crucial difference with NPIV is that, $f_{*,\text{low}} \notin \mathcal{H}_{\gamma_x, \gamma_o}$ since $f_{*,\text{low}}$ is only constructed by a cut-off with respect to the partial Fourier spectrum of on X . Hence, to construct f_{aux} we convolve $f_{*,\text{low}}$ again with K_{γ_o} :

$$f_{\text{aux}} = f_{*,\text{low}} * K_{\gamma_o} + f_{*,\text{high}} * K_{\gamma_x, \gamma_o} \in \mathcal{H}_{\gamma_x, \gamma_o}. \quad (40)$$

Here, $K_{\gamma_x, \gamma_o}(\mathbf{x}, \mathbf{o}) := K_{\gamma_x}(\mathbf{x}) \cdot K_{\gamma_o}(\mathbf{o})$. Proceeding from the above, we have

$$\begin{aligned} \|T([f_\lambda] - f_*)\|_{L^2(P_{ZO})}^2 &\leq \lambda \|f_{\text{aux}}\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 \\ &\quad + \|T(f_{*,\text{low}} - f_{*,\text{low}} * K_{\gamma_o})\|_{L^2(P_{ZO})}^2 + \|T(f_{*,\text{high}} - f_{*,\text{high}} * K_{\gamma_x, \gamma_o})\|_{L^2(P_{ZO})}^2 \\ &\leq \lambda \|f_{\text{aux}}\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 + \|f_{*,\text{low}} - f_{*,\text{low}} * K_{\gamma_o}\|_{L^2(P_{ZO})}^2 \\ &\quad + \gamma_x^{2d_x \eta} \|f_{*,\text{high}} - f_{*,\text{high}} * K_{\gamma_x, \gamma_o}\|_{L^2(P_{ZO})}^2. \end{aligned} \quad (41)$$

The last inequality follows by Jensen's inequality for the second term and [Assumption 4.3](#) for the third term. We would like to highlight here that the second term in Eq. (41) is absent in Eq. (37) for the NPIV setting when there are no observed covariates O . Unlike the high frequency term in Eq. (41), we cannot employ the Fourier measure of *partial* contractivity in [Assumption 4.3](#) because $f_{*,\text{low}} - f_{*,\text{low}} * K_{\gamma_o}$ only contains low frequency spectrum on X by construction. From the above derivations, we can also deduce

$$\begin{aligned} \|T([f_{\text{aux}}] - f_*)\|_{L^2(P_{ZO})}^2 &\lesssim \|f_{*,\text{low}} - f_{*,\text{low}} * K_{\gamma_o}\|_{L^2(P_{XO})}^2 \\ &\quad + \gamma_x^{2d_x \eta} \|f_{*,\text{high}} - f_{*,\text{high}} * K_{\gamma_x, \gamma_o}\|_{L^2(P_{XO})}^2. \end{aligned}$$

To upper bound the approximation error $\|[f_\lambda] - f_*\|_{L^2(P_{XO})}$, we notice that

$$\begin{aligned} &\|[f_\lambda] - f_*\|_{L^2(P_{XO})}^2 \\ &\leq \|[f_\lambda] - [f_{\text{aux}}]\|_{L^2(P_{XO})}^2 + \|[f_{\text{aux}}] - f_*\|_{L^2(P_{XO})}^2 \\ &\leq \gamma_x^{-2d_x \eta} \|T([f_\lambda] - [f_{\text{aux}}])\|_{L^2(P_{ZO})}^2 + \|[f_{\text{aux}}] - f_*\|_{L^2(P_{XO})}^2 \\ &\lesssim \gamma_x^{-2d_x \eta} \|T([f_\lambda] - f_*)\|_{L^2(P_{ZO})}^2 + \gamma_x^{-2d_x \eta} \|T[f_{\text{aux}}] - Tf_*\|_{L^2(P_{ZO})}^2 + \|[f_{\text{aux}}] - f_*\|_{L^2(P_{XO})}^2 \\ &\lesssim \gamma_x^{-2d_x \eta} \left(\lambda \|f_{\text{aux}}\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 + \|f_{*,\text{low}} - f_{*,\text{low}} * K_{\gamma_o}\|_{L^2(P_{XO})}^2 \right. \\ &\quad \left. + \gamma_x^{2d_x \eta} \|f_{*,\text{high}} - f_{*,\text{high}} * K_{\gamma_x, \gamma_o}\|_{L^2(P_{XO})}^2 \right). \end{aligned}$$

The last inequality holds by plugging into the upper bound on $\|T([f_\lambda] - f_{\text{aux}})\|_{L^2(P_{ZO})}^2$ and $\|T[f_{\text{aux}}] - Tf_*\|_{L^2(P_{ZO})}^2$ derived above. Notice that the term $\|f_{*,\text{low}} - f_{*,\text{low}} * K_{\gamma_o}\|_{L^2(P_{XO})}$ is unnecessarily inflated by the Fourier measure of *partial* ill-posedness $\gamma_x^{-d_x\eta} \gg 1$ ([Assumption 4.2](#)) even if there is no smoothing effect of T acting on O . We obtain

$$\begin{aligned} \| [f_\lambda] - f_* \|_{L^2(P_{XO})}^2 &\lesssim \gamma_x^{-2d_x\eta} \left(\lambda \gamma_x^{-d_x} \gamma_o^{-d_o} + \gamma_o^{2s_o} + \gamma_x^{2s_x+2d_x\eta} \right) + \gamma_o^{2s_o} + \gamma_x^{2s_x} \\ &\lesssim n^{-\frac{\frac{s_x}{d_x}}{1+2(\frac{s_x}{d_x}+\eta)+\frac{d_o}{s_o}(\frac{s_x}{d_x}+\eta)}}. \end{aligned}$$

The last step holds by choosing $\lambda, \gamma_x, \gamma_o$ as in [Theorem 4.1](#). The upper bound on the approximation error does not match the minimax lower bound in [Theorem 4.2](#).

We conclude this section by offering a more practical perspective on why the gap emerges. Hyperparameters including the kernel lengthscales are selected via cross-validation in practice. For the KIV-O estimator \hat{f}_λ computed from Stage II samples $\mathcal{D}_2 = \{(\mathbf{z}_i, \mathbf{o}_i, y_i)\}_{i=1}^n$, the kernel lengthscales γ_x, γ_o are selected by minimizing the following cross-validation criterion.

$$\text{CV}(\gamma_x, \gamma_o, \lambda) = \frac{1}{n} \sum_{i=1}^n (y_i - \langle \hat{f}_{-i, \gamma_x, \gamma_o, \lambda}, \hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) \rangle_{\mathcal{H}_{\gamma_x, \gamma_o}})^2, \quad (42)$$

where $\hat{f}_{-i, \gamma_x, \gamma_o, \lambda}$ denotes the version of $\hat{f}_{\gamma_x, \gamma_o, \lambda}$ computed from $n-1$ samples from \mathcal{D}_2 excluding its i -th sample. This cross-validation criterion $\text{CV}(\gamma_x, \gamma_o, \lambda)$ is a consistent estimator for the *projected* risk $\|T([\hat{f}_\lambda] - f_*)\|_{L^2(P_{ZO})}$ and has been used in many 2SLS approaches [[Hartford et al., 2017](#), [Xu et al., 2021b](#), [Mastouri et al., 2021](#), [Xu and Gretton, 2025](#)].

Similarly, the choice of lengthscales $\gamma_x, \gamma_o, \lambda$ in our upper bound analysis is also obtained by minimizing the projected risk $\|T([\hat{f}_\lambda] - f_*)\|_{L^2(P_{XO})}$ which results $\gamma_x^{s_x+d_x\eta} = \gamma_o^{s_o}$. This can be contrasted with the choice of kernel lengthscales in anisotropic kernel ridge regression which imposes a different balance condition $\gamma_x^{s_x} = \gamma_o^{s_o}$ [[Hang and Steinwart, 2021](#)]. The extra $d_x\eta$ in the exponent of γ_x arises due to the partial smoothing effect of T , which maps a function that is (s_x, s_o) -smooth on $\mathcal{X} \times \mathcal{O}$ to a function that is $(s_x + d_x\eta, s_o)$ -smooth on $\mathcal{Z} \times \mathcal{O}$. On the other hand, our minimax lower bound requires constructing B-splines in Eq. (98) in the Supplement with resolution \mathfrak{K} and $J_x := 2^{\lfloor \frac{\mathfrak{K}s}{s_x} \rfloor}$ and $J_o := 2^{\lfloor \frac{\mathfrak{K}s}{s_o} \rfloor}$. J_x, J_o play a role analogous to γ_x, γ_o (see [Remark 4.1](#)), but they satisfy the balance condition $J_x^{s_x} \asymp J_o^{s_o}$. This discrepancy between the kernel lengthscales balance condition and B-spline lengthscales balance condition gives rise to the gap between our upper and lower bound. Simply enforcing $\gamma_x^{s_x} = \gamma_o^{s_o}$ in our upper bound analysis would result in a slower rate.

6 Conclusion

We study nonparametric instrumental variable regression with observed covariates (NPIV-O), a setting that generalizes NPIV by incorporating observed covariates to enable heterogeneous treatment effect estimation. The conditional expectation operator T behaves as a partial identity operator, which makes NPIV-O a hybrid of NPIV and NPR. We prove an upper bound for kernel 2SLS and the first minimax lower bound. Our upper and lower bounds interpolate between the known optimal rates for NPIV and NPR, and adapt to the anisotropic smoothness of f_* . Our analysis reveals a gap between the upper and lower bounds in the general setting, and closing this gap remains an open direction for NPIV-O.

References

- Mélisande Albert, Bé'atrice Laurent, Amandine Marrel, and Anouar Meynaoui. Adaptive test of independence based on hsic measures. *The Annals of Statistics*, 50(2):858–879, 2022.
- Donald WK Andrews. Examples of l^2 -complete and boundedly-complete distributions. *Journal of Econometrics*, 199(2):213–220, 2017.
- Jean-Pierre Aubin. *Applied functional analysis*. John Wiley & Sons, 2011.
- Francis R Bach and Michael I Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.
- Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007.
- Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- Andrew Bennett, Nathan Kallus, Xiaojie Mao, Whitney Newey, Vasilis Syrgkanis, and Masatoshi Uehara. Minimax instrumental variable regression and l^2 convergence guarantees without identification or closedness. In *The 36th Annual Conference on Learning Theory*, pages 2291–2318. PMLR, 2023.
- Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, 2004. ISBN 978-1-4419-9096-9.
- Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013, 2018.
- Richard Blundell, Xiaohong Chen, and Dennis Kristensen. Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669, 2007.
- Bariscan Bozkurt, Ben Deaner, Dimitri Meunier, Liyuan Xu, and Arthur Gretton. Density ratio-based proxy causal learning without density ratios. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtilay Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 5095–5103. PMLR, 03–05 May 2025a. URL <https://proceedings.mlr.press/v258/bozkurt25a.html>.
- Bariscan Bozkurt, Houssam Zenati, Dimitri Meunier, Liyuan Xu, and Arthur Gretton. Density ratio-free doubly robust proxy causal learning. *arXiv preprint arXiv:2505.19807*, 2025b.
- Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.
- Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 4(04):377–408, 2006. URL https://www.worldscientific.com/doi/10.1142/S0219530506000838?srslid=AfmB0opECjt_9-RIZIsquke2XT40kUwY1kvYRzJUfbAL5-e9JDTAXR6A.

- Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanitá. Vector-valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.
- Marine Carrasco, Jean-Pierre Florens, and Eric Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of Econometrics*, 6:5633–5751, 2007.
- Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632, 2007.
- Xiaohong Chen and Timothy M. Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression: nonlinear functionals of nonparametric iv. *Quantitative Economics*, 9(1):39–84, March 2018. ISSN 1759-7323.
- Xiaohong Chen and Demian Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321, January 2012.
- Xiaohong Chen and Markus Reiss. On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory*, 27(3):497–521, 2011. ISSN 0266-4666, 1469-4360.
- Xiaohong Chen, Victor Chernozhukov, Sokbae Lee, and Whitney K Newey. Local identification of nonparametric and semiparametric models. *Econometrica*, 82(2):785–809, 2014.
- Xiaohong Chen, Timothy Christensen, and Sid Kankanala. Adaptive estimation and uniform confidence bands for nonparametric structural functions and elasticities, 2024. URL <https://arxiv.org/abs/2107.11869>.
- Zonghao Chen, Masha Naslidnyk, and Francois-Xavier Briol. Nested expectations with kernel quadrature. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 8760–8793. PMLR, 13–19 Jul 2025.
- Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359, 2024.
- Serge Darolles, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.
- Ernesto De Vito, Lorenzo Rosasco, Andrea Caponnetto, Umberto De Giovannini, Francesca Odone, and Peter Bartlett. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6(5), 2005.
- Ben Deaner. Proxy controls and panel data. *arXiv preprint arXiv:1810.00283*, 2018.
- Andreas Defant and Carsten Michels. A complex interpolation formula for tensor products of vector-valued banach function spaces. *Archiv der Mathematik*, 74:441–451, 2000.
- Ronald A DeVore and George G Lorentz. *Constructive approximation*. Springer Science & Business Media, 1993.
- Ronald A. DeVore and Vasil A. Popov. Interpolation of besov spaces. *Transactions of the American Mathematical Society*, 305(1):397–414, January 1988.

- Ronald A DeVore and Robert C Sharpley. Besov spaces on domains in \mathbb{R}^d . *Transactions of the American Mathematical Society*, 335(2):843–864, 1993.
- Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12248–12262. Curran Associates, Inc., 2020.
- Xavier D’Haultfoeuille. On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, 27(3):460–471, 2011.
- Mona Eberts and Ingo Steinwart. Optimal regression rates for svms using gaussian kernels. *Electronic Journal of Statistics*, 7(none):1 – 42, 2013.
- David Eric Edmunds and Hans Triebel. *Function spaces, entropy numbers, differential operators*. Cambridge University Press, 1996.
- Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*. Springer Science & Business Media, 1996.
- Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21(205):1–38, 2020.
- Jean-Pierre Florens, Jan Johannes, and Sébastien Van Bellegem. Identification and estimation by penalization in nonparametric instrumental regression. *Econometric Theory*, 27(3):472–496, 2011.
- Friedrich Gerard Friedlander. *Introduction to the theory of distributions*. Cambridge University Press, 1998.
- Amiremad Ghassami, Andrew Ying, Ilya Shpitser, and Eric Tchetgen Tchetgen. Minimax kernel machine learning for a class of doubly robust functionals with application to proximal causal inference. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7210–7239. PMLR, 28–30 Mar 2022. URL <https://proceedings.mlr.press/v151/ghassami22a.html>.
- Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press, 2021.
- A. Gretton. A simpler condition for consistency of a kernel independence test. *arXiv preprint arXiv:1501.06103*, 2015.
- A. Gretton and L. Gyorfi. Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11:1391–1423, 2010.
- Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita, editors, *Algorithmic Learning Theory*, pages 63–77, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.
- Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

- Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimiliano Pontil. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1803—1810. PMLR, 2012.
- Peter Hall and Joel L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33(6):2904–2929, 2005.
- Thomas Hamm and Ingo Steinwart. Adaptive learning rates for support vector machines working on data with low intrinsic dimension. *The Annals of Statistics*, 49(6):3153–3180, 2021.
- Hanyuan Hang and Ingo Steinwart. Optimal learning with anisotropic gaussian svms. *Applied and Computational Harmonic Analysis*, 55:337–367, 2021.
- Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, pages 1414–1423. PMLR, 2017.
- M Hoffman and Oleg Lepski. Random rates in anisotropic regression. *The Annals of Statistics*, 30(2):325–396, 2002.
- Joel L. Horowitz. Applied nonparametric instrumental variables estimation. *Econometrica*, 79(2):347–394, Mar 2011.
- Tuomas Hytönen, Jan Van Neerven, Mark Veraar, and Lutz Weis. *Analysis in Banach spaces*, volume 12. Springer, 2016.
- IA Ibragimov and RZ Khas' minskii. Asymptotic bounds on the quality of the nonparametric regression estimation in. *Journal of Soviet Mathematics*, 24(5):540–550, 1984.
- Steven G. Johnson. Saddle-point integration of c_∞ “bump” functions. *arXiv preprint arXiv:1508.04376*, 2015.
- Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Causal inference under unmeasured confounding with negative controls: a minimax learning approach. *arXiv preprint arXiv:2103.14029*, 2021.
- Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.
- Sid Kankanala. Generalized bayes in conditional moment restriction models. *arXiv preprint arXiv:2510.01036*, 2025.
- Yitzhak Katznelson. *An introduction to harmonic analysis*. Cambridge University Press, 2004.
- Fares El Khoury, Edouard Pauwels, Samuel Vaïter, and Michael Arbel. Learning theory for kernel bilevel optimization. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025.
- Juno Kim, Dimitri Meunier, Arthur Gretton, Taiji Suzuki, and Zhu Li. Optimality and adaptivity of deep neural features for instrumental variable regression. In *International Conference on Learning Representations*, 2025.

- Ilja Klebanov, Ingmar Schuster, and Timothy John Sullivan. A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, 2(3):583–606, 2020.
- Achim Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013.
- Yurii Kolomoitsev and Sergey Tikhonov. Properties of moduli of smoothness in $l_p(\mathbb{R}^d)$. *Journal of Approximation Theory*, 257:105423, 2020. ISSN 0021-9045.
- David Krieg. Tensor power sequences and the approximation of tensor product operators. *Journal of Complexity*, 44:30–51, 2018.
- J Paul Leigh and Michael Schembri. Instrumental variables technique: cigarette price provided better estimate of effects of smoking on sf-12. *Journal of Clinical Epidemiology*, 57(3):284–293, 2004.
- Christopher Leisner. Nonlinear wavelet approximation in anisotropic besov spaces. *Indiana University Mathematics Journal*, 52(2):437–455, 2003. ISSN 00222518, 19435258. URL <https://www.math.purdue.edu/~lucier/692/caarticle.pdf>.
- Yujia Li, Roman Pogodin, Dougal Sutherland, and Arthur Gretton. Self-supervised learning with kernel dependence maximization. In Marc’Aurelio Ranzato, Alina Beygelzimer, Yann Dauphin, Percy S. Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, Inc., 2021.
- Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Optimal rates for regularized conditional mean embedding learning. In Sanmi Koyejo, Shakir Mohamed, Alekh Agarwal, Danielle Belgrave, Kyunghyun Cho, and Alice Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 4433–4445. Curran Associates, Inc., 2022.
- Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Towards optimal sobolev norm rates for the vector-valued regularized least-squares algorithm. *Journal of Machine Learning Research*, 25(181):1–51, 2024a.
- Zihao Li, Hui Lan, Vasilis Syrgkanis, Mengdi Wang, and Masatoshi Uehara. Regularized deep IV with model selection. *arXiv preprint arXiv:2403.04236*, 2024b.
- Luofeng Liao, You-Lin Chen, Zhuoran Yang, Bo Dai, Mladen Kolar, and Zhaoran Wang. Provably efficient neural estimation of structural equation models: An adversarial approach. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8947–8958. Curran Associates, Inc., 2020.
- Junhong Lin and Volkan Cevher. Optimal convergence for distributed learning with stochastic gradient methods and spectral algorithms. *Journal of Machine Learning Research*, 21(147):1–63, 2020.
- Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020.

Weihao Lu, Yicheng Li, Qian Lin, et al. On the saturation effects of spectral algorithms in large dimensions. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 7011–7059. Curran Associates, Inc., 2024.

Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt Kusner, Arthur Gretton, and Krikamol Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, pages 7512–7523. PMLR, 2021.

Dimitri Meunier, Zhu Li, Tim Christensen, and Arthur Gretton. Nonparametric instrumental regression via kernel methods is minimax optimal. *arXiv preprint arXiv:2411.19653*, 2024a.

Dimitri Meunier, Zikai Shen, Mattes Mollenhauer, Arthur Gretton, and Zhu Li. Optimal rates for vector-valued spectral regularization learning algorithms. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2024b. URL <https://arxiv.org/abs/2405.14778>.

Dimitri Meunier, Antoine Moulin, Jakub Wornbard, Vladimir R Kostic, and Arthur Gretton. Demystifying spectral feature learning for instrumental variable regression. *arXiv preprint arXiv:2506.10899*, 2025.

Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Roger B Nelsen. *An introduction to copulas*. Springer, 2006.

Whitney K. Newey and James L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21247–21259. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f340f1b1f65b6df5b5e3f94d95b11daf-Paper.pdf.

Ieva Petruļionytė, Julien Mairal, and Michael Arbel. Functional bilevel optimization for machine learning. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 14016–14065. Curran Associates, Inc., 2024.

Michael Reed and Barry Simon. *Methods of modern mathematical physics: Functional analysis*, volume 1. Gulf Professional Publishing, 1980.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In Corinna Cortes, Neil Lawrence, Daniel Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Walter Rudin. *Real and complex analysis*. McGraw-Hill, Inc., 1987.

- H J Schmeisser. An unconditional basis in periodic spaces with dominating mixed smoothness properties. *Analysis Mathematica*, 13(2):153–168, 1987.
- Hans-Jürgen Schmeisser. Recent developments in the theory of function spaces with dominating mixed smoothness. *Nonlinear Analysis, Function Spaces and Applications*, pages 145–204, 2007.
- Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.
- Dino Sejdinovic, Arthur Gretton, and Wicher Bergsma. A kernel test for three-variable interactions. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/076a0c97d09cf1a0ec3e19c7f2529f2b-Paper.pdf.
- H. Shen, S. Jegelka, and A. Gretton. Fast kernel-based independent component analysis. *IEEE Transactions on Signal Processing*, 57:3498–3511, 2009.
- Winfried Sickel and Tino Ullrich. Tensor products of sobolev–besov spaces and applications to approximation from the hyperbolic cross. *Journal of Approximation Theory*, 161(2):748–786, 2009. ISSN 0021-9045.
- Rahul Singh. Kernel methods for unobserved confounding: negative controls, proxies, and instruments. *arXiv preprint arXiv:2012.10315*, 2020.
- Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. In Hanna Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché Buc, Emily Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.
- Rahul Singh, Liyuan Xu, and Arthur Gretton. Kernel methods for causal functions: dose, heterogeneous and incremental response curves. *Biometrika*, 111(2):497–516, 2024.
- Steve Smale and Ding-Xuan Zhou. Shannon sampling ii: connections to learning theory. *Applied and Computational Harmonic Analysis*, 19(3):285–302, 2005.
- Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.
- L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13:1393–1434, 2012.
- Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968. ACM, 2009.
- Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.
- Ingo Steinwart and Clint Scovel. Mercer’s theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35:363–417, 2012.
- Ingo Steinwart, Don R Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *The 22nd Annual Conference on Learning Theory*, pages 79–93, 2009.

- Haotian Sun, Antoine Moulin, Tongzheng Ren, Arthur Gretton, and Bo Dai. Spectral representation for causal estimation with hidden confounders. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258. PMLR, 2025.
- Danica J Sutherland. Fixing an error in caponnetto and de vito (2007). *arXiv preprint arXiv:1702.02982*, 2017.
- Taiji Suzuki and Atsushi Nitanda. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic besov space. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann Dauphin, Percy S. Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3609–3621, 2021.
- Vasilis Syrgkanis, Victor Lei, Miruna Oprescu, Maggie Hei, Keith Battocchi, and Greg Lewis. Machine learning estimation of heterogeneous treatment effects with instruments. In Hanna Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.
- Zoltán Szabó and Bharath K Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18(233):1–29, 2018.
- Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal inference. *Statistical Science*, 39(3):375–390, 2024.
- Hans Triebel. Entropy numbers in function spaces with mixed integrability. *Revista matemática complutense*, 24:169–188, 2011.
- Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519.
- Roman Vershynin. *High-dimensional probability: an introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- Ziyu Wang, Yuchen Luo, Yueru Li, Jun Zhu, and Bernhard Schölkopf. Spectral representation learning for conditional moment models. *arXiv preprint arXiv:2210.16525*, 2022.
- Holger Wendland. *Scattered Data Approximation*, volume 17. Cambridge university press, 2004.
- Liyuan Xu and Arthur Gretton. Kernel single proxy control for deterministic confounding. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 3736–3744. PMLR, 2025.
- Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. In *International Conference on Learning Representations*, 2021a.
- Liyuan Xu, Heishiro Kanagawa, and Arthur Gretton. Deep proxy causal learning and its application to confounded bandit policy evaluation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26264–26275. Curran Associates, Inc., 2021b. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/dcf3219715a7c9cd9286f19db46f2384-Paper.pdf.

Haobo Zhang, Yicheng Li, Weihao Lu, and Qian Lin. On the optimality of misspecified kernel ridge regression. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 41th International Conference on Machine Learning*, pages 41331–41353. PMLR, 2023a.

Haobo Zhang, Yicheng Li, and Qian Lin. On the optimality of misspecified spectral algorithms. *Journal of Machine Learning Research*, 25(188):1–50, 2024.

Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, 2018.

Rui Zhang, Masaaki Imaizumi, Bernhard Schölkopf, and Krikamol Muandet. Instrumental variable regression via kernel maximum moment loss. *Journal of Causal Inference*, 11(1):20220073, 2023b.

A Examples for the partial smoothing effect of \mathbf{T}

In this section, we extend [Assumption 4.2](#) and [Assumption 4.3](#) from the main text—on the Fourier measure of partial ill-posedness and on the Fourier measure of partial contractivity of the operator T , respectively—to the space of distributions. This allows us to apply these assumptions to periodic functions and to construct an explicit example verifying [Assumption 4.3](#). First, we prove the following Lemma relating [Assumption 4.2](#) and [Assumption 4.3](#).

Lemma A.1. *If Assumption 4.2 and 4.3 hold simultaneously, and P_{XO} is absolutely continuous with respect to the Lebesgue measure on $\mathcal{X} \times \mathcal{O}$, then $\eta_0 \geq \eta_1$.*

Proof. Let $\gamma \in (0, 1)$ be arbitrary. Let $\epsilon > 0$ be arbitrary. We define

$$f_{\gamma,\epsilon}(\mathbf{x}, \mathbf{o}) := \mathcal{F}^{-1} [\mathbb{1}[\|\boldsymbol{\omega}\|_2 \in [\gamma^{-1}, \gamma^{-1} + \epsilon]]] (\mathbf{x}).$$

We have $f_{\gamma,\epsilon} \in \text{LF}((\gamma^{-1} + \epsilon)^{-1}) \cap \text{HF}(\gamma) \cap L^\infty(P_{XO})$. Since P_{XO} admits a density function on $\mathcal{X} \times \mathcal{O}$, and $f_{\gamma,\epsilon}$ only vanishes on sets of null Lebesgue measure, we have $\|f_{\gamma,\epsilon}\|_{L^2(P_{XO})} \neq 0$. As imposed by [Assumption 4.3](#) and [Assumption 4.2](#), we thus have

$$c_0^{-1} (\gamma^{-1} + \epsilon)^{-d_x \eta_0} \|f_{\gamma,\epsilon}\|_{L^2(P_{XO})} \leq \|T f_{\gamma,\epsilon}\|_{L^2(P_{Z_O})} \leq c_1 \gamma^{d_x \eta_1} \|f_{\gamma,\epsilon}\|_{L^2(P_{XO})}.$$

Since $\|f_{\gamma,\epsilon}\|_{L^2(P_{XO})} \neq 0$, we have $(\forall \gamma \in (0, 1)) (\forall \epsilon > 0) c_0^{-1} (\gamma^{-1} + \epsilon)^{-d_x \eta_0} \leq c_1 \gamma^{d_x \eta_1}$. For a fixed γ , taking the limit $\epsilon \rightarrow 0$, we have by continuity

$$(\forall \gamma \in (0, 1)) \frac{1}{c_0 c_1} \leq \gamma^{d_x(\eta_1 - \eta_0)}. \quad (43)$$

Taking the limit $\gamma \rightarrow 1$ in Eq. (43), we find $c_0 c_1 \geq 1$. Then since Eq. (43) holds for all $\gamma \in (0, 1)$, we find that $\eta_1 \leq \eta_0$. \square

Definition 5 (Distribution and distribution of a function). *Let $\Omega \subseteq \mathbb{R}^d$ be a domain. A distribution $u \in \mathcal{D}'(\Omega)$ is a continuous linear functional on the space of test functions $\mathcal{D}(\Omega)$, where $\mathcal{D}(\Omega)$ is the set $C_c^\infty(\Omega)$ endowed with the canonical limit of Fréchet topology. For a locally integrable function $f \in L^1_{\text{loc}}(\mathbb{R}^d)$, for all $\phi \in \mathcal{D}(\Omega)$, we define the distribution T_f by*

$$T_f \phi := \int_{\Omega} f(\mathbf{x}) \phi(\mathbf{x}) \, d\mathbf{x}.$$

Definition 6 (Support of a distribution). A distribution $u \in \mathcal{D}'(\Omega)$ is supported in the closed set $K \subset \Omega$ if $u[\phi] = 0 \forall \phi \in C_c^\infty(\Omega \setminus K)$. The support of u , $\text{supp } u$ is the set

$$\text{supp } u = \cap\{K : u \text{ is supported in } K\}.$$

Definition 7 (Tempered distribution). We define the Schwartz space \mathcal{S} via

$$\mathcal{S} = \left\{ \phi \in C^\infty(\mathbb{R}^d) \mid \forall \alpha, \forall N \in \mathbb{N}, \sup_{\mathbf{x} \in \mathbb{R}^d} |(1 + |\mathbf{x}|)^N D^\alpha \phi(\mathbf{x})| < \infty \right\}.$$

We say that a sequence $\{\phi_j\}_{j=1}^\infty \subset \mathcal{S}$ tends to zero iff

$$\sup_{\mathbf{x} \in \mathbb{R}^d} |(1 + |\mathbf{x}|)^N D^\alpha \phi_j(\mathbf{x})| \rightarrow 0$$

for all $N \in \mathbb{N}$ and all multi-indices α . This endows \mathcal{S} with a topology. We define the space of tempered distributions to be the continuous dual space of \mathcal{S} , denoted as \mathcal{S}' .

Definition 8 (Periodic distribution). We define the translation of a distribution $u \in \mathcal{D}'(\mathbb{R}^d)$ via $\tau_{\mathbf{z}} u[\phi] = u[\tau_{-\mathbf{z}} \phi]$ for all $\phi \in \mathcal{D}(\mathbb{R}^d)$, where we use $\tau_{\mathbf{z}} \phi(\mathbf{x}) := \phi(\mathbf{x} - \mathbf{z})$. We say that a distribution $u \in \mathcal{D}'(\mathbb{R}^d)$ is periodic if for each $\mathbf{g} \in \mathbb{Z}^d$ we have $\tau_{\mathbf{g}} u = u$. Clearly, if $f \in L^1_{\text{loc}}(\mathbb{R}^d)$ is periodic, then T_f is a periodic distribution.

Definition 9 (Fourier transform of L^1 functions). For $f \in L^1(\mathbb{R}^d)$, we define the Fourier transform $\mathcal{F}[f] = \hat{f} : \mathbb{R}^d \rightarrow \mathbb{C}$ by

$$\mathcal{F}[f](\xi) = \hat{f}(\xi) := \int_{\mathbb{R}^d} f(\mathbf{x}) \exp(-i\langle \mathbf{x}, \xi \rangle) d\mathbf{x}.$$

Definition 10 (Fourier transform of tempered distributions). For a distribution $u \in \mathcal{S}'$, we define the Fourier transform of u , written $\hat{u} \in \mathcal{S}'$, to be the distribution satisfying:

$$\hat{u}[\phi] = u[\hat{\phi}], \quad \forall \phi \in \mathcal{S},$$

which is well defined since the Fourier transform maps \mathcal{S} to \mathcal{S} continuously.

We note that, for all $\phi \in \mathcal{S}$, by the Fourier inversion theorem [Rudin, 1987, 9.11],

$$T_1[\hat{\phi}] = \int_{\mathbb{R}^d} \hat{\phi}(\mathbf{x}) d\mathbf{x} = (2\pi)^d \phi(\mathbf{0}) = (2\pi)^d \delta_{\mathbf{0}}[\phi].$$

Hence, $\forall \mathbf{x} \in \mathbb{R}^d$, we define $e_\xi(\mathbf{x}) = \exp(i2\pi\langle \mathbf{x}, \xi \rangle)$. So $\forall \phi \in \mathcal{S}$, we have

$$T_{e_\xi}[\hat{\phi}] = \int_{\mathbb{R}^d} \exp(i2\pi\langle \mathbf{x}, \xi \rangle) \hat{\phi}(\mathbf{x}) d\mathbf{x} = (2\pi)^d \phi(2\pi\xi). \quad (44)$$

i.e. $\mathcal{F}[T_{e_\xi}] = (2\pi)^d \delta_{2\pi\xi}$, the Dirac delta distribution at $2\pi\xi$. We can now make sense of the Fourier transform of a periodic function. The following Lemma is from Friedlander [1998, 8.5]:

Lemma A.2 (Periodic distributions are tempered). Let $u \in \mathcal{D}'(\mathbb{R}^d)$ be a periodic distribution. Then u is in fact a tempered distribution, i.e. $u \in \mathcal{S}'$.

Proposition A.1. Suppose $u \in \mathcal{D}'(\mathbb{R}^d)$ is a periodic distribution. Then there exist constants $c_\xi \in \mathbb{C}$ such that u can be represented as a (generalized) Fourier series,

$$u = \sum_{\xi \in \mathbb{Z}^d} c_\xi T_{e_\xi},$$

with c_ξ satisfying the bound $|c_\xi| \leq K(1 + |\xi|)^N$ for some $K > 0$ and $N \in \mathbb{Z}$.

We can now make sense of the following definitions: for any scalar $\gamma \in (0, 1)$, we define the following two sets of functions which are generalization of the $\text{LF}(\gamma)$ and $\text{HF}(\gamma)$ defined in the main text to *distributions*:

$$\begin{aligned} \text{LF}(\gamma) &:= \{f \in L^1_{\text{loc}}(\mathbb{R}^{d_x+d_o}) \mid \forall \mathbf{o} \in \mathcal{O}, T_{f(\cdot, \mathbf{o})} \in \mathcal{S}'(\mathbb{R}^{d_x}), \\ &\quad \text{supp}(\mathcal{F}[f(\cdot, \mathbf{o})]) \subseteq \{\boldsymbol{\omega}_x \in \mathbb{R}^{d_x} : \|\boldsymbol{\omega}_x\|_2 \leq \gamma^{-1}\}\}. \\ \text{HF}(\gamma) &:= \{f \in L^1_{\text{loc}}(\mathbb{R}^{d_x+d_o}) \mid \forall \mathbf{o} \in \mathcal{O}, T_{f(\cdot, \mathbf{o})} \in \mathcal{S}'(\mathbb{R}^{d_x}), \\ &\quad \text{supp}(\mathcal{F}[f(\cdot, \mathbf{o})]) \subseteq \{\boldsymbol{\omega}_x \in \mathbb{R}^{d_x} : \|\boldsymbol{\omega}_x\|_2 \geq \gamma^{-1}\}\}. \end{aligned} \tag{45}$$

We now recall the statements of Assumption 4.2 and 4.3 from the main text with the above generalization of $\text{LF}(\gamma)$ and $\text{HF}(\gamma)$.

Assumption A.1 (Fourier measure of partial ill-posedness of T). *There exists a constant $c_0 > 0$ and a parameter $\eta_0 \in [0, \infty)$ only depending on T , such that for all $\gamma \in (0, 1)$ and all functions $f \in \text{LF}(\gamma) \cap L^\infty(P_{XO})$, the following inequality is satisfied:*

$$\|f\|_{L^2(P_{XO})} \leq c_0 \gamma^{-d_x \eta_0} \|Tf\|_{L^2(P_{ZO})}.$$

In particular, c_0 does not depend on γ .

Assumption A.2 (Fourier measure of partial contractivity of T). *There exists a constant $c_1 > 0$ and a parameter $\eta_1 \in [0, \infty)$ only depending on T , such that for all $\gamma \in (0, 1)$ and all functions $f \in \text{HF}(\gamma) \cap L^\infty(P_{XO})$, the following inequality is satisfied:*

$$\|Tf\|_{L^2(P_{ZO})} \leq c_1 \gamma^{d_x \eta_1} \|f\|_{L^2(P_{XO})}.$$

In particular, c_1 does not depend on γ .

Let $(S^1, +)$ denote the unit circle (equipped with a group structure via addition), and let dx denote the Haar measure on S^1 , which coincides with the pushforward of the Lebesgue measure under the quotient map $S^1 \cong \frac{\mathbb{R}}{\mathbb{Z}}$. We make use of the obvious identification between functions on $(S^1)^d$ and 1-periodic functions on \mathbb{R}^d , for any $d \geq 1$.

Definition 11 (Fourier series [Katznelson, 2004]). *Let $f \in L^1(S^1)$. We define the n th Fourier coefficient of f by $\mathcal{F}[f][n] = \int_{S^1} f(x) e^{-i2\pi n x} dx$. The Fourier series of $f \in L^1(S^1)$ is the trigonometric series $f(x) = \sum_{n=-\infty}^{\infty} \mathcal{F}[f][n] e^{i2\pi n x}$.*

We make use of the group structure of S^1 and the *translation invariance* of the measure dx on S^1 to define the convolution operation in $L^1(S^1)$, following [Katznelson, 2004, Section 1.7].

Proposition A.2. Let $f, g \in L^1(S^1)$. For almost all $t \in S^1$, the function $f(t - \tau)g(\tau)$ is $L^1(S^1)$ -integrable as a function of τ , and if we define the convolution

$$h(t) = \int_{S^1} f(t - \tau)g(\tau) d\tau$$

then $h \in L^1(S^1)$ with $\|h\|_{L^1(S^1)} \leq \|f\|_{L^1(S^1)}\|g\|_{L^1(S^1)}$. Moreover, ($\forall n \in \mathbb{Z}$), we have

$$\mathcal{F}[h][n] = \mathcal{F}[f][n]\mathcal{F}[g][n].$$

For any $\gamma > 0$, we now exhibit a class of functions $f \in \text{HF}(\gamma)$ and a distribution $p(x, z, o)$ satisfying the statement in [Assumption A.2](#), with the help of [Proposition A.2](#). For simplicity, we assume that $d_x = d_z = d_o = 1$. Fix a scalar $\gamma \in (0, 1)$. Let $g \in L^1(S^1)$ be a function whose Fourier coefficients vanish on low frequencies, in the sense that

$$\mathcal{F}[g][n] = 0 \quad \text{for all } n \in \mathbb{Z} \text{ such that } |n| \leq (2\pi\gamma)^{-1}. \quad (46)$$

An example of such a function is $g(x) = \exp(i2\pi mx)$, where $m \in \mathbb{Z}$ and $m > (2\pi\gamma)^{-1}$. Let $h \in L^1_{\text{loc}}(\mathbb{R}^{d_o})$ be such that it does not vanish identically. We then define $f : \mathbb{R}^2 \rightarrow \mathbb{C}$:

$$f(x + t, o) := g(x)h(o) \quad \text{for all } x \in [0, 1), t \in \mathbb{Z}, o \in \mathbb{R}.$$

Then it follows that $f \in L^1_{\text{loc}}(\mathbb{R}^{d_x+d_o})$ since $g \in L^1(S^1)$ and $h \in L^1_{\text{loc}}(\mathbb{R})$. It follows from [Lemma A.2](#) that ($\forall o \in \mathbb{R}$), $T_{f(\cdot, o)} \in \mathcal{S}'(\mathbb{R})$ for any $o \in \mathbb{R}$. Moreover, for every $o \in \mathbb{R}$, we observe that: i) $h(o) = 0$, then $\text{supp}(\mathcal{F}_x[T_{f(\cdot, o)}]) = \emptyset$, ii) $h(o) \neq 0$, then

$$\begin{aligned} \text{supp}(\mathcal{F}[T_{f(\cdot, o)}]) &= \text{supp}(\mathcal{F}[T_g]) \\ &= \text{supp}\left(\mathcal{F}\left[\sum_{n=-\infty}^{\infty} \mathcal{F}[g][n] T_{e^{i2\pi n \cdot}}\right]\right) \\ &\stackrel{(a)}{=} \text{supp}\left(\sum_{n=-\infty}^{\infty} \mathcal{F}[g][n](2\pi)\delta_{2\pi n}\right) \\ &\stackrel{(b)}{=} \text{supp}\left(\sum_{|n|>(2\pi\gamma)^{-1}} \mathcal{F}[g][n](2\pi)\delta_{2\pi n}\right) \\ &\subseteq \left\{\omega_x \in \mathbb{R} : |\omega_x| > \frac{1}{\gamma}\right\}. \end{aligned}$$

Here, step (a) follows from (44) and the use of distributional support as defined in [Definition 6](#), and step (b) follows from Eq. (46). Thus, from both cases, we conclude that $f \in \text{HF}(\gamma)$ defined in Eq. (45).

Fix an integer $k \geq 1$. We consider a probability space $\Omega = S^1 \times S^1 \times \underbrace{S^1 \times \cdots \times S^1}_k$, where each copy of S^1 is equipped with its Borel σ -algebra and the normalized Haar measure, and Ω is equipped with the product Borel σ -algebra and the product measure. We define the following mappings, where $1 \leq i \leq k$:

$$\begin{aligned} \pi_Z : \Omega &\rightarrow S^1, \quad \pi_Z(z, o, u_1, \dots, u_k) = z \\ \pi_O : \Omega &\rightarrow S^1, \quad \pi_O(z, o, u_1, \dots, u_k) = o \end{aligned}$$

$$\pi_{U_i} : \Omega \rightarrow S^1, \quad \pi_{U_i}(z, o, u_1, \dots, u_k) = u_i,$$

Since π_Z, π_O, π_{U_i} 's are measurable, they are valid random variables, and we also denote them by Z, O, U_1, \dots, U_k . We then write $W = 0.1(U_1 + \dots + U_k)$ and $X = Z + W$, where addition is understood as a group operation on S^1 . By construction, X is independent of O . The random tuple (X, Z, O) is as considered in the NPIV-O set-up in the main text.

We are now going to show that there exists some constant $c_1 > 0$ (which does not depend on γ), such that

$$\|Tf\|_{L^2(P_{ZO})} \leq c_1 \gamma^k \|f\|_{L^2(P_{XO})}.$$

We observe that

$$\begin{aligned} \|f\|_{L^2(P_{XO})}^2 &= \int_{\mathbb{R}^2} |f(x, o)|^2 p(x, o) dx do \\ &= \int_{S^1 \times S^1} |f(x, o)|^2 p(x, o) dx do = \|g\|_{L^2(P_X)}^2 \|h\|_{L^2(P_O)}^2. \end{aligned} \quad (47)$$

We also observe that

$$\begin{aligned} \|Tf\|_{L^2(P_{ZO})}^2 &= \langle f, T^*Tf \rangle_{L^2(P_{XO})} \\ &= \int_{S^1 \times S^1} \overline{f(x, o)} (T^*Tf)(x, o) p(x, o) dx do. \end{aligned} \quad (48)$$

We also observe that for all $x \in S^1$ and $o \in S^1$,

$$\begin{aligned} (T^*Tf)(x, o) &= \int_{S^1} \int_{S^1} f(x', o) p(x' | z, o) dx' p(z | x, o) dz \\ &\stackrel{(*)}{=} \int_{S^1} f(x', o) \left(\int_{S^1} p(x' | z, o) p(z | x, o) dz \right) dx' \\ &= \int_{S^1} f(x', o) L(x, x', o) dx', \end{aligned}$$

where $(*)$ follows by Fubini's theorem. L is defined as, for all $x, x' \in S^1$ and $o \in S^1$,

$$L(x, x', o) := \int_{S^1} p(x' | z, o) p(z | x, o) dz = \int_{S^1} p(x' | z) p(z | x) dz.$$

The last step holds because X is independent of O . As L is not dependent on o , we write $L(x, x', o) = L(x, x')$. Hence $(\forall x \in S^1, o \in S^1)$

$$(T^*Tf)(x, o) = \left(\int_{S^1} g(x') L(x, x') dx' \right) h(o).$$

Hence continuing from Eq. (48), we find

$$\begin{aligned} \|Tf\|_{L^2(P_{ZO})}^2 &= \int_{S^1 \times S^1} \overline{g(x) h(o)} (T^*Tf)(x, o) p(x, o) dx do \\ &= \left(\int_{S^1} \overline{g(x)} \left(\int_{S^1} g(x') L(x, x') p(x) dx' \right) dx \right) \|h\|_{L^2(P_O)}^2. \end{aligned} \quad (49)$$

We also notice that ($\forall x \in S^1, x' \in S^1$), the following hold via Bayes' rule and the fact that $p(z)$ is the Haar measure on S^1 :

$$\begin{aligned}
L(x, x')p(x) &= \int_{z \in S^1} p(x' | z)p(z | x)p(x) dz \\
&= \int_{z \in S^1} p(x' | z)p(x | z)p(z) dz \\
&= \int_{z \in S^1} p(x' | z)p(x | z) dz \\
&= \int_{z \in S^1} p_W(x' - z)p_W(x - z) dz \\
&= \int_{z \in S^1} p_W(z)p_W(z + (x - x')) dz \\
&=: \mathcal{L}(x - x').
\end{aligned}$$

where the second last step follows from the change of variable $z \leftarrow x' - z$, and in the last step we use the fact that $L(x, x')p(x)$ is translation-invariant. As throughout the calculations, the difference $x - x'$ denotes the group operation on S^1 rather than the usual difference on \mathbb{R} . We calculate the Fourier coefficients of \mathcal{L} as follows:

$$\begin{aligned}
\mathcal{F}[\mathcal{L}][n] &= \int_{w \in S^1} \mathcal{L}(w)e^{-i2\pi nw} dw = \int_{w \in S^1} \int_{z \in S^1} p_W(z)p_W(z + w) dz e^{-i2\pi nw} dw \\
&= \left| \int_{z \in S^1} p_W(z)e^{2\pi inz} dz \right|^2 = |\mathcal{F}[p_W][n]|^2.
\end{aligned}$$

Since $W = 0.1(U_1 + \dots + U_k)$, we have p_W is the k -times convolution of the probability density function of $0.1U_1$. By the convolution Theorem, we find

$$\mathcal{F}[\mathcal{L}][n] = \left| 10 \int_0^{0.1} e^{-2\pi inz} dz \right|^{2k} = \left| \frac{10}{2\pi in} (1 - e^{-0.2\pi in}) \right|^{2k} = \left(\frac{10}{\pi n} \right)^{2k} \sin^{2k}(0.1\pi n). \quad (50)$$

Hence

$$\mathcal{F}[\mathcal{L}][n] \leq \left(\frac{10}{\pi n} \right)^{2k}. \quad (51)$$

We have

$$\begin{aligned}
\frac{\|Tf\|_{L^2(P_{ZO})}^2}{\|h\|_{L^2(P_O)}^2} &= \int_{S^1} \overline{g(x)} \left(\int_{S^1} g(x') \mathcal{L}(x - x') dx' \right) dx \\
&\stackrel{(a)}{=} \sum_{n=-\infty}^{\infty} \overline{\mathcal{F}[g][n]} \mathcal{F} \left[\int_{S^1} g(x') \mathcal{L}(\cdot - x') dx' \right] [n] \\
&\stackrel{(b)}{=} \sum_{n=-\infty}^{\infty} \overline{\mathcal{F}[g][n]} \mathcal{F}[g][n] \mathcal{F}[\mathcal{L}][n] = \sum_{n=-\infty}^{\infty} |\mathcal{F}[g][n]|^2 \cdot \mathcal{F}[\mathcal{L}][n] \\
&\stackrel{(c)}{=} \sum_{n \in \mathbb{Z}, |n| > (2\pi\gamma)^{-1}} |\mathcal{F}[g][n]|^2 \cdot \mathcal{F}[\mathcal{L}][n] \stackrel{(d)}{\leq} (20\gamma)^{2k} \sum_{n \in \mathbb{Z}, |n| > (2\pi\gamma)^{-1}} |\mathcal{F}[g][n]|^2 \\
&= (20\gamma)^{2k} \sum_{n=-\infty}^{\infty} |\mathcal{F}[g][n]|^2 \stackrel{(e)}{=} (20\gamma)^{2k} \|g\|_{L^2(S^1)}^2
\end{aligned} \quad (52)$$

$$\stackrel{(f)}{=} (20\gamma)^{2k} \|g\|_{L^2(P_X)}^2.$$

In the above derivations, step (a) follows from Parseval's theorem [Katznelson, 2004, 5.4] and the fact that $\{e^{i2\pi nx}\}_{n \in \mathbb{Z}}$ forms an orthonormal system in $L^2(S^1)$, step (b) follows from Proposition A.2, step (c) follows Eq. (46), step (d) follows from Eq. (51), step (e) follows again from Parseval's Theorem, and finally step (f) follows from the fact that P_X is the Haar measure on S^1 . Continuing from Eq. (49), we thus have

$$\|Tf\|_{L^2(P_{ZO})}^2 \leq (20\gamma)^{2k} \|g\|_{L^2(P_X)}^2 \|h\|_{L^2(P_O)}^2 = (20\gamma)^{2k} \|f\|_{L^2(P_{XO})}^2,$$

where the last equality follows from Eq. (47). Therefore, we have proved that Assumption A.2 is satisfied with $\eta_1 = k$ and $c_1 = 20^k$. For any $\gamma > 0$, we now exhibit a class of functions $f' \in \text{LF}(\gamma)$ and a distribution $p(x, z, o)$ satisfying the statement in Assumption A.1. We let $p(x, z, o)$ be the probability distribution constructed above. Let $g' \in L^1(S^1)$ be a function whose Fourier coefficients vanish on high frequencies, in the sense that

$$\mathcal{F}[g'][n] = 0 \quad \text{for all } n \in \mathbb{Z} \text{ such that } |n| \geq (2\pi\gamma)^{-1}.$$

We further assume that $\mathcal{F}[g'][n] \neq 0$ only if $\sin^{2k}(0.1\pi n) \geq c$ for a fixed positive constant $c > 0$. Let $h \in L^1_{\text{loc}}(\mathbb{R}^{d_o})$ be such that it does not vanish identically. We then define $f' : \mathbb{R}^2 \rightarrow \mathbb{C}$:

$$f'(x + t, o) := g'(x)h(o) \quad \text{for all } x \in [0, 1], t \in \mathbb{Z}, o \in \mathbb{R}.$$

We can show that $f' \in \text{LF}(\gamma)$ defined in Eq. (45) via a similar argument as before. By Eq. (52), we have

$$\begin{aligned} \frac{\|Tf'\|_{L^2(P_{ZO})}^2}{\|h\|_{L^2(P_O)}^2} &= \sum_{n=-\infty}^{\infty} |\mathcal{F}[g'][n]|^2 \cdot \mathcal{F}[\mathcal{L}][n] \\ &= \sum_{n \in \mathbb{Z}, |n| < (2\pi\gamma)^{-1}} |\mathcal{F}[g'][n]|^2 \cdot \mathcal{F}[\mathcal{L}][n] \\ &\stackrel{(a)}{=} \sum_{n \in \mathbb{Z}, |n| < (2\pi\gamma)^{-1}} |\mathcal{F}[g'][n]|^2 \left(\frac{10}{\pi n} \right)^{2k} \sin^{2k}(0.1\pi n) \\ &\stackrel{(b)}{\geq} \sum_{n \in \mathbb{Z}, |n| < (2\pi\gamma)^{-1}} |\mathcal{F}[g'][n]|^2 \left(\frac{10}{\pi n} \right)^{2k} c \\ &\geq c(20\gamma)^{2k} \sum_{n \in \mathbb{Z}, |n| < (2\pi\gamma)^{-1}} |\mathcal{F}[g'][n]|^2 = c(20\gamma)^{2k} \|g'\|_{L^2(S^1)}^2, \end{aligned}$$

where step (a) follows from Eq. (50), and step (b) follows from the assumption on the Fourier coefficient of g . We thus have

$$\|Tf'\|_{L^2(P_{ZO})}^2 \geq c(20\gamma)^{2k} \|f'\|_{L^2(P_{XO})}^2.$$

B Explicit Solutions of KIV-O

The following derivation is adapted from Meunier et al. [2024a, Section D], which only covers the case of no observed covariates. We refer the reader to Singh et al. [2019, Section A.5.1] for the original derivation of closed-form solution for KIV with no observed covariates, and to Mastouri

et al. [2021], Singh [2020], Xu and Gretton [2025] for derivation of closed-form solutions for the RKHS two-stage proximal causal learning framework, which is mathematically equivalent to KIV with observed covariates, as discussed in [Section 3.1](#). Whenever an operator or Gram matrix require distinguishing between Stage I or Stage II kernel on O , we denote this via ;1 or ;2 in the subscript. \odot denotes Hadamard product. For a matrix $J \in \mathbb{R}^{m \times n}$, $J_{:,j}$ denotes its j th column.

Stage 1 We follow the closed-form solution given in [Li et al. \[2022\]](#). We define

$$\begin{aligned}\Phi_{\tilde{Z}\tilde{O};1} : \mathcal{H}_{ZO} &\rightarrow \mathbb{R}^{\tilde{n}}, \quad \Phi_{\tilde{Z}\tilde{O};1} = [\phi_Z(\tilde{\mathbf{z}}_1) \otimes \phi_{O,1}(\tilde{\mathbf{o}}_1), \dots, \phi_Z(\tilde{\mathbf{z}}_{\tilde{n}}) \otimes \phi_{O,1}(\tilde{\mathbf{o}}_{\tilde{n}})]^*, \\ \Phi_{\tilde{X}} : \mathcal{H}_X &\rightarrow \mathbb{R}^{\tilde{n}}, \quad \Phi_{\tilde{X}} = [\phi_X(\tilde{\mathbf{x}}_1), \dots, \phi_X(\tilde{\mathbf{x}}_{\tilde{n}})]^*\end{aligned}$$

We obtain the following estimator

$$\hat{F}_\xi(\cdot, \cdot) = \hat{C}_{X|Z,O;\xi} \phi_Z(\cdot) \otimes \phi_{O,1}(\cdot), \quad \hat{C}_{X|Z,O;\xi} = \Phi_{\tilde{X}}^* \left(K_{\tilde{Z}\tilde{O};1} + \tilde{n}\xi \text{Id} \right)^{-1} \Phi_{\tilde{Z}\tilde{O};1}, \quad (53)$$

where we introduce the Gram matrix

$$K_{\tilde{Z}\tilde{O};1} = \Phi_{\tilde{Z}\tilde{O};1} \Phi_{\tilde{Z}\tilde{O};1}^*, \quad [K_{\tilde{Z}\tilde{O};1}]_{ij} = k_Z(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j) k_{O,1}(\tilde{\mathbf{o}}_i, \tilde{\mathbf{o}}_j).$$

Stage 2 The Stage 2 solution can be written as

$$\hat{f}_\lambda = \left(\frac{1}{n} \Phi_{\hat{F}O;2}^* \Phi_{\hat{F}O;2} + \lambda \text{Id} \right)^{-1} \frac{1}{n} \Phi_{\hat{F}O;2}^* \mathbf{Y}$$

where we define

$$\Phi_{\hat{F}O;2} : \mathcal{H}_{XO} \rightarrow \mathbb{R}^n \quad \Phi_{\hat{F}O;2} = \left[\hat{F}_\xi(\mathbf{z}_1, \mathbf{o}_1) \otimes \phi_{O,2}(\mathbf{o}_1), \dots, \hat{F}_\xi(\mathbf{z}_n, \mathbf{o}_n) \otimes \phi_{O,2}(\mathbf{o}_n) \right]^*.$$

We then write this in a dual form

$$\hat{f}_\lambda = \Phi_{\hat{F}O;2}^* \left(K_{\hat{F}O;2} + n\lambda \text{Id} \right)^{-1} \mathbf{Y}.$$

where we introduce the Gram matrix

$$K_{\hat{F}O;2} = \Phi_{\hat{F}O;2} \Phi_{\hat{F}O;2}^*, \quad [K_{\hat{F}O;2}]_{ij} = \langle \hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i), \hat{F}_\xi(\mathbf{z}_j, \mathbf{o}_j) \rangle_{\mathcal{H}_X} k_{O,2}(\mathbf{o}_i, \mathbf{o}_j).$$

By Eq. (53), we obtain, for $1 \leq j \leq n$,

$$\hat{F}_\xi(\mathbf{z}_j, \mathbf{o}_j) = \Phi_{\tilde{X}}^* \underbrace{\left(K_{\tilde{Z}\tilde{O},1} + \tilde{n}\xi \text{Id} \right)^{-1} \left(K_{\tilde{Z}Z} \odot K_{\tilde{O}O;1} \right)_{:,j}}_{=:J_{:,j}} = \sum_{i=1}^{\tilde{n}} J_{ij} \phi_X(\tilde{\mathbf{x}}_i), \quad (54)$$

where J as defined column-wise is a $\tilde{n} \times n$ matrix, and we define the (cross) Gram matrices

$$[K_{\tilde{Z}Z}]_{ij} = k_Z(\tilde{\mathbf{z}}_i, \mathbf{z}_j), \quad [K_{\tilde{O}O;1}]_{ij} = k_{O,1}(\tilde{\mathbf{o}}_i, \mathbf{o}_j),$$

for $1 \leq i \leq \tilde{n}$, $1 \leq j \leq n$. Consequently, for $1 \leq i, j \leq n$, we have

$$K_{\hat{F}O;2} = (J^T K_{\tilde{X}\tilde{X}} J) \odot K_{OO;2},$$

where we define the Gram matrices

$$[K_{\tilde{X}\tilde{X}}]_{ij} = k_X(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j), \quad [K_{OO;2}]_{lm} = k_{O,2}(\mathbf{o}_l, \mathbf{o}_m),$$

for $1 \leq i, j \leq \tilde{n}$, $1 \leq l, m \leq n$. For a new test point $(\mathbf{x}, \mathbf{o}) \in \mathcal{X} \times \mathcal{O}$, we have, for $1 \leq j \leq n$,

$$\begin{aligned} \left\langle \phi_X(\mathbf{x}) \otimes \phi_{O,2}(\mathbf{o}), \hat{F}_\xi(\mathbf{z}_j, \mathbf{o}_j) \otimes \phi_{O,2}(\mathbf{o}_j) \right\rangle_{\mathcal{H}_{X,O}} &= k_{O,2}(\mathbf{o}, \mathbf{o}_j) \left(\sum_{i=1}^{\tilde{n}} J_{ij} k_X(\tilde{\mathbf{x}}_i, \mathbf{x}) \right) \\ &= (K_{O,2} \odot (J^T K_{\tilde{X}\mathbf{x}}))_j, \end{aligned}$$

where we define $K_{\tilde{X}\mathbf{x}} \in \mathbb{R}^{\tilde{n} \times 1}$ and $K_{O,2} \in \mathbb{R}^{n \times 1}$ respectively as follows:

$$[K_{\tilde{X}\mathbf{x}}]_i = k_X(\tilde{\mathbf{x}}_i, \mathbf{x}), \quad [K_{O,2}]_i = k_{O,2}(\mathbf{o}_i, \mathbf{o}).$$

Thus we have

$$\begin{aligned} \hat{f}_\lambda(\mathbf{x}, \mathbf{o}) &= \left\langle \phi_X(\mathbf{x}) \otimes \phi_{O,2}(\mathbf{o}), \Phi_{FO;2}^* \left(K_{FO;2} + n\lambda \text{Id} \right)^{-1} \mathbf{Y} \right\rangle_{\mathcal{H}_{X,O}} \\ &= \left(K_{O,2}^T \odot (K_{\tilde{X}\mathbf{x}}^T J) \right) \left((J^T K_{\tilde{X}\mathbf{x}} J) \odot K_{OO;2} + n\lambda \text{Id} \right)^{-1} \mathbf{Y}, \end{aligned}$$

where the last line follows by Eq. (54), Thus the derivation is concluded.

C RKHS \mathcal{H}_{FO}

We recall that the NPIV-O problem can be written as

$$Y = (T f_*)(Z, O) + v, \quad \mathbb{E}[v | Z, O] = 0, \quad (55)$$

where $v := f_*(X, O) - (T f_*)(Z, O) + \epsilon$. As introduced in Meunier et al. [2024a, Appendix E.1.2], we define an RKHS \mathcal{H}_{FO} induced by the statistical inverse problem Eq. (55), following Steinwart and Christmann [2008, Theorem 4.21]. Our construction can be obtained from that of Meunier et al. [2024a] with an appropriate feature map construction, as follows. Recall the definition of $F_* : \mathcal{Z} \times \mathcal{O} \rightarrow \mathcal{H}_{X,\gamma_x}$ that for any $\mathbf{z} \in \mathcal{Z}, \mathbf{o} \in \mathcal{O}$, $F_*(\mathbf{z}, \mathbf{o}) = \mathbb{E}[\phi_{X,\gamma_x}(X) | Z = \mathbf{z}, O = \mathbf{o}]$.

Definition 12. We define a reproducing kernel Hilbert space \mathcal{H}_{FO} as

$$\mathcal{H}_{FO} = \left\{ f : \mathcal{Z} \times \mathcal{O} \rightarrow \mathbb{R} \mid \exists w \in \mathcal{H}_{\gamma_x, \gamma_o}, f(\mathbf{z}, \mathbf{o}) \equiv \langle w, F_*(\mathbf{z}, \mathbf{o}) \otimes \phi_{\gamma_o}(\mathbf{o}) \rangle_{\mathcal{H}_{\gamma_x, \gamma_o}} \right\},$$

equipped with the norm

$$\|f\|_{\mathcal{H}_{FO}} := \inf \left\{ \|w\|_{\mathcal{H}_{\gamma_x, \gamma_o}} \mid f(\mathbf{z}, \mathbf{o}) \equiv \langle w, F_*(\mathbf{z}, \mathbf{o}) \otimes \phi_{\gamma_o}(\mathbf{o}) \rangle_{\mathcal{H}_{\gamma_x, \gamma_o}} \right\}.$$

We observe that $(\mathbf{z}, \mathbf{o}) \mapsto F_*(\mathbf{z}, \mathbf{o}) \otimes \phi_{\gamma_o}(\mathbf{o})$ is not the canonical feature map of \mathcal{H}_{FO} . To construct its canonical feature map, we define $V : \mathcal{H}_{\gamma_x, \gamma_o} \rightarrow \mathcal{H}_{FO}$ such that

$$(Vw)(\mathbf{z}, \mathbf{o}) \equiv \langle w, F_*(\mathbf{z}, \mathbf{o}) \otimes \phi_{\gamma_o}(\mathbf{o}) \rangle_{\mathcal{H}_{\gamma_x, \gamma_o}}, \quad w \in \mathcal{H}_{\gamma_x, \gamma_o}.$$

From Theorem 4.21 of Steinwart and Christmann [2008], V is a metric surjection. By definition, for any $f \in \mathcal{H}_{\gamma_x, \gamma_o}$, we have

$$(Vf)(\mathbf{z}, \mathbf{o}) = \langle f, F_*(\mathbf{z}, \mathbf{o}) \otimes \phi_{\gamma_o}(\mathbf{o}) \rangle_{\mathcal{H}_{\gamma_x, \gamma_o}} = \mathbb{E}[f(X, O) | Z = \mathbf{z}, O = \mathbf{o}] = (T[f])(\mathbf{z}, \mathbf{o}).$$

Furthermore, we know that V is also a surjective partial isometry, i.e. V is surjective and satisfies

$$(\forall f \in \ker(V)^\perp), \quad \|f\|_{\mathcal{H}_{\gamma_x, \gamma_o}} = \|Vf\|_{\mathcal{H}_{FO}}.$$

Equivalently, ($\forall r \in \mathcal{H}_{FO}$),

$$\|r\|_{\mathcal{H}_{FO}} = \inf\{\|h\|_{\mathcal{H}_{\gamma_x, \gamma_o}} : h \in \mathcal{H}_{\gamma_x, \gamma_o}, r = Vh\}, \quad (56)$$

Thus, the canonical feature map of \mathcal{H}_{FO} is $(\mathbf{z}, \mathbf{o}) \mapsto V(F_*(\mathbf{z}, \mathbf{o}) \otimes \phi_{\gamma_o}(\mathbf{o}))$, a fact which was also observed on Meunier et al. [2024a, Page 28].

Recall the definition of f_λ in Eq. (25),

$$f_\lambda := \arg \min_{f \in \mathcal{H}_{\gamma_x, \gamma_o}} \lambda \|f\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 + \|T([f] - f_*)\|_{L^2(P_{ZO})}^2. \quad (57)$$

and \bar{f}_λ in Eq. (23),

$$\bar{f}_\lambda := \arg \min_{f \in \mathcal{H}_{\gamma_x, \gamma_o}} \lambda \|f\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 + \frac{1}{n} \sum_{i=1}^n (y_i - \langle f, F_*(\mathbf{z}_i, \mathbf{o}_i) \otimes \phi_{\gamma_o}(\mathbf{o}_i) \rangle_{\mathcal{H}_{\gamma_x, \gamma_o}})^2. \quad (58)$$

We define the images of f_λ and \bar{f}_λ respectively under the metric surjection V as follows:

$$h_\lambda := Vf_\lambda, \quad \bar{h}_\lambda := V\bar{f}_\lambda \in \mathcal{H}_{FO}, \quad h_* := Tf_*$$

We observe by combining Eq. (56) and Eq. (58) that \bar{h}_λ is the solution to a standard kernel ridge regression problem with the RKHS \mathcal{H}_{FO} :

$$\bar{h}_\lambda = \arg \min_{h \in \mathcal{H}_{FO}} \lambda \|h\|_{\mathcal{H}_{FO}}^2 + \frac{1}{n} \sum_{i=1}^n (h(\mathbf{z}_i, \mathbf{o}_i) - y_i)^2. \quad (59)$$

Similarly, for h_λ we have

$$h_\lambda = \arg \min_{h \in \mathcal{H}_{FO}} \lambda \|h\|_{\mathcal{H}_{FO}}^2 + \|h_* - h\|_{L^2(P_{ZO})}^2. \quad (60)$$

Finally, we have

$$\|T([\bar{f}_\lambda] - [f_\lambda])\|_{L^2(P_{ZO})} = \|[V\bar{f}_\lambda] - [Vf_\lambda]\|_{L^2(P_{ZO})} = \|[\bar{h}_\lambda] - [h_\lambda]\|_{L^2(P_{ZO})}, \quad (61)$$

which means that the projected error $\|T([\bar{f}_\lambda] - [f_\lambda])\|_{L^2(P_{ZO})}$ has been translated to the generalization error $\|[\bar{h}_\lambda] - [h_\lambda]\|_{L^2(P_{ZO})}$ of a standard kernel ridge regression (KRR) with this new hypothesis space \mathcal{H}_{FO} , which allows us to apply techniques from the analysis of kernel ridge regression. To this end, we will analyse the capacity (Section C.1) along with the embedding property (Section C.2) of \mathcal{H}_{FO} , both of which are crucial properties for characterizing the generalization error of KRR [Fischer and Steinwart, 2020].

C.1 Capacity of \mathcal{H}_{FO}

We have

$$\begin{aligned} & \sup_{\mathbf{o}, \mathbf{o}'} \sup_{\mathbf{z}, \mathbf{z}'} \langle V(F_*(\mathbf{z}, \mathbf{o}) \otimes \phi_{\gamma_o}(\mathbf{o})), V(F_*(\mathbf{z}', \mathbf{o}') \otimes \phi_{\gamma_o}(\mathbf{o}')) \rangle_{\mathcal{H}_{FO}} \\ &= \sup_{\mathbf{o}, \mathbf{o}'} \sup_{\mathbf{z}, \mathbf{z}'} \langle F_*(\mathbf{z}, \mathbf{o}) \otimes \phi_{\gamma_o}(\mathbf{o}), F_*(\mathbf{z}', \mathbf{o}') \otimes \phi_{\gamma_o}(\mathbf{o}') \rangle_{\mathcal{H}_{\gamma_x, \gamma_o}} \\ &= \sup_{\mathbf{o}, \mathbf{o}'} k_{O, \gamma_o}(\mathbf{o}, \mathbf{o}') \sup_{\mathbf{z}, \mathbf{z}'} \iint_{\mathcal{X} \times \mathcal{X}} k_{X, \gamma_x}(\mathbf{x}, \mathbf{x}') p(\mathbf{x} \mid \mathbf{z}, \mathbf{o}) p(\mathbf{x}' \mid \mathbf{z}', \mathbf{o}') d\mathbf{x} d\mathbf{x}' \end{aligned}$$

$$\begin{aligned}
&\leq \sup_{\mathbf{o}, \mathbf{o}'} k_{O, \gamma_o}(\mathbf{o}, \mathbf{o}') \sup_{\mathbf{x}, \mathbf{x}'} k_{X, \gamma_x}(\mathbf{x}, \mathbf{x}') \sup_{\mathbf{z}, \mathbf{z}'} \iint_{\mathcal{X} \times \mathcal{X}} p(\mathbf{x} \mid \mathbf{z}, \mathbf{o}) p(\mathbf{x}' \mid \mathbf{z}', \mathbf{o}') d\mathbf{x} d\mathbf{x}' \\
&= \sup_{\mathbf{o}, \mathbf{o}'} k_{O, \gamma_o}(\mathbf{o}, \mathbf{o}') \sup_{\mathbf{x}, \mathbf{x}'} k_{X, \gamma_x}(\mathbf{x}, \mathbf{x}').
\end{aligned}$$

Hence the RKHS \mathcal{H}_{FO} has a bounded kernel, therefore by [Steinwart and Scovel \[2012, Lemma 2.3\]](#), \mathcal{H}_{FO} is compactly embedded into $L^2(P_{ZO})$. We define the covariance operator $C_{FO} : \mathcal{H}_{FO} \rightarrow \mathcal{H}_{FO}$

$$C_{FO} := V \mathbb{E}_{ZO} [(F_*(Z, O) \otimes \phi_{\gamma_o}(O)) \otimes (F_*(Z, O) \otimes \phi_{\gamma_o}(O))] V^*$$

It is a self-adjoint compact operator by [Steinwart and Scovel \[2012, Lemma 2.2\]](#). The spectral theorem for self-adjoint compact operators [[Reed and Simon, 1980, Theorems VI.16, VI.17](#)] yields, there exists countable $\mu_{FO,1} \geq \mu_{FO,2} \geq \dots \geq 0$, $([e_{FO,i}])_{i \geq 1}$ an orthonormal system of $L^2(P_{ZO})$ and $(\sqrt{\mu_{FO,i}} e_{FO,i}) \subseteq \mathcal{H}_{FO}$ an orthonormal system in \mathcal{H}_{FO} , such that

$$C_{FO} = \sum_{i \geq 1} \mu_{FO,i} \langle \cdot, \sqrt{\mu_{FO,i}} e_{FO,i} \rangle_{\mathcal{H}_{FO}} \sqrt{\mu_{FO,i}} e_{FO,i}. \quad (62)$$

Definition 13 (Effective dimension). *The effective dimension of \mathcal{H}_{FO} , denoted as $\mathcal{N}_{FO} : [0, \infty) \rightarrow [0, \infty)$ is defined as*

$$\mathcal{N}_{FO}(\lambda) = \text{tr}((C_{FO} + \lambda)^{-1} C_{FO}) = \sum_{i \geq 1} \frac{\mu_{FO,i}}{\mu_{FO,i} + \lambda}.$$

Proposition C.1. *Let $n \geq 10$ and $\lambda = n^{-1}$. Let $\gamma_x, \gamma_o \in (0, 1]$ be the lengthscales for the RKHS $\mathcal{H}_{\gamma_x, \gamma_o}$. Suppose that the distribution P_{XO} satisfies Assumption 2.2. Then we have,*

$$\mathcal{N}_{FO}(\lambda) \leq C' (\log n)^{d_x + d_o + 1} \left(\gamma_x^{d_x} \gamma_o^{d_o} \right)^{-1}$$

for some constant C' independent of n, γ_x, γ_o .

Proof. From [Corollary C.1](#) and [Lemma C.1](#) and setting $p = \frac{1}{\log n}$, ($\forall i \geq 1$), we have

$$\mu_{FO,i} \leq (C \log n)^{2(d_x + d_o + 1) \log n} (\gamma_x^{d_x} \gamma_o^{d_o})^{-2 \log n} i^{-2 \log n}.$$

Next, we use [[Caponnetto and De Vito, 2007, Proposition 3](#)] (with error corrected in [Sutherland \[2017\]](#)) to obtain

$$\begin{aligned}
\mathcal{N}_{FO}(\lambda) &\leq \frac{\pi / (2 \log n)}{\sin(\pi / (2 \log n))} \left((C \log n)^{2(d_x + d_o + 1) \log n} (\gamma_x^{d_x} \gamma_o^{d_o})^{-2 \log n} \right)^{\frac{1}{2 \log n}} \lambda^{-(2 \log n)^{-1}} \\
&= \frac{\pi / (2 \log n)}{\sin(\pi / (2 \log n))} (C \log n)^{(d_x + d_o + 1)} \left(\gamma_x^{d_x} \gamma_o^{d_o} \right)^{-1} \lambda^{-(2 \log n)^{-1}} \\
&\stackrel{(i)}{\leq} 3 C^{(d_x + d_o + 1)} (\log n)^{(d_x + d_o + 1)} \left(\gamma_x^{d_x} \gamma_o^{d_o} \right)^{-1} n^{(2 \log n)^{-1}} \\
&\stackrel{(ii)}{=} 3 \sqrt{e} C^{(d_x + d_o + 1)} (\log n)^{(d_x + d_o + 1)} \left(\gamma_x^{d_x} \gamma_o^{d_o} \right)^{-1} \\
&= C' (\log n)^{d_x + d_o + 1} \left(\gamma_x^{d_x} \gamma_o^{d_o} \right)^{-1},
\end{aligned}$$

where $C' = \sqrt{e} 3 C^{(d_x + d_o + 1)}$. In the above chain of derivations, (i) holds because $\frac{t}{\sin t} \leq 3$ for $t \leq \pi/2$ and $\pi / (2 \log n) \leq \pi/2$ and (ii) holds because $n^{\frac{1}{2 \log n}} = \sqrt{e}$ for $n > 1$. \square

In the proof of Proposition C.1, to bound the effective dimension, we need to bound the eigendecay of the compact self-adjoint operator C_{FO} . We first control the decay of the entropy numbers of the RKHS \mathcal{H}_{FO} , which translates into a bound on the eigendecay of C_{FO} as shown in Corollary C.1. In Lemma C.2, we show that the i th entropy number of \mathcal{H}_{FO} is bounded above by the i th entropy number of $\mathcal{H}_{\gamma_x, \gamma_o}$. The entropy numbers of $\mathcal{H}_{\gamma_x, \gamma_o}$ are well-understood by the results of Hang and Steinwart [2021] (restated in Lemma C.1), which completes the derivation.

In this section, for real-valued Hilbert spaces E, F and a bounded, linear, compact operator $S : E \rightarrow F$, $s_i(S)$ denotes the i th singular value of S , as defined in Steinwart and Christmann [2008, Eq. (A.25) Page 505]; $e_i(S)$ denotes the i th entropy number of S , as defined in Steinwart and Christmann [2008, Definition A.5.26 Page 516]; $a_i(S)$ denotes the i th approximation number of S , as defined in Steinwart and Christmann [2008, Eq. (A.29) Page 506].

Lemma C.1. *Suppose that P_{XO} satisfies Assumption 2.2 in the main text. Then, $(\forall i \geq 1)$, $(\gamma_x, \gamma_o \in (0, 1])$, there exists a constant $C > 0$ such that for any $p > 0$,*

$$e_i(\text{id} : \mathcal{H}_{\gamma_x, \gamma_o} \hookrightarrow L^2(P_{XO})) \leq (3C)^{\frac{1}{p}} \left(\frac{d_x + d_o + 1}{ep} \right)^{\frac{d_x + d_o + 1}{p}} \left(\gamma_x^{d_x} \gamma_o^{d_o} \right)^{-\frac{1}{p}} i^{-\frac{1}{p}}.$$

Proof. The corollary follows immediately from Hang and Steinwart [2021, Proposition 1] by setting $\gamma = [\underbrace{\gamma_x, \dots, \gamma_x}_{d_x}, \underbrace{\gamma_o, \dots, \gamma_o}_{d_o}]^\top \in (0, 1]^{d_x + d_o}$, and noting that $L^\infty(\mathcal{X} \times \mathcal{O})$ continuously embeds into $L^2(P_{XO})$ with $\|L^\infty(\mathcal{X} \times \mathcal{O}) \hookrightarrow L^2(P_{XO})\| \leq 1$ and Steinwart and Christmann [2008, Eq. (A.38)]. \square

Lemma C.2. *We have, $(\forall i \geq 1)$,*

$$e_i(\text{id} : \mathcal{H}_{FO} \hookrightarrow L^2(P_{ZO})) \leq e_i(\text{id} : \mathcal{H}_{\gamma_x, \gamma_o} \hookrightarrow L^2(P_{XO})).$$

Proof. Fix $i \geq 1$. For a Hilbert space \mathcal{H} , $B_\mathcal{H}$ denotes the unit ball in \mathcal{H} , and $B(x, r, \|\cdot\|_\mathcal{H})$ denotes the ball in \mathcal{H} centred at $x \in \mathcal{H}$ with radius r . Fix $\epsilon > 0$ such that $\exists g_1, \dots, g_{2^{i-1}} \in B_{\mathcal{H}_{\gamma_x, \gamma_o}}$ such that

$$B_{\mathcal{H}_{\gamma_x, \gamma_o}} \subseteq \bigcup_{i=1}^{2^{i-1}} B(g_i, \epsilon, \|\cdot\|_{L^2(P_{XO})}).$$

Fix $f \in B_{\mathcal{H}_{FO}}$ and an arbitrary $\tilde{\epsilon} > 0$. By Steinwart and Christmann [2008, Eq. (4.11)], there exists $g \in \mathcal{H}_{\gamma_x, \gamma_o}$ such that

$$f = \langle g, F_*(\cdot, \cdot) \otimes \phi_{\gamma_o}(\cdot) \rangle_{\mathcal{H}_{\gamma_x, \gamma_o}} = Vg,$$

and $\|g\|_{\mathcal{H}_{\gamma_x, \gamma_o}} \leq 1 + \tilde{\epsilon}$. By the preceding statement, $(\exists i \in \{1, \dots, 2^{i-1}\})$ such that

$$\|g - (1 + \tilde{\epsilon})g_i\|_{L^2(P_{XO})} \leq \epsilon(1 + \tilde{\epsilon}).$$

For $i \in \{1, \dots, 2^{i-1}\}$, define $f_i = (1 + \tilde{\epsilon})Vg_i = (1 + \tilde{\epsilon})\langle g_i, F_*(\cdot, \cdot) \otimes \phi_{\gamma_o}(\cdot) \rangle_{\mathcal{H}_{\gamma_x, \gamma_o}}$. Then we have

$$\begin{aligned} \|f - f_i\|_{L^2(P_{ZO})}^2 &= \int_{\mathcal{Z} \times \mathcal{O}} \langle g - (1 + \tilde{\epsilon})g_i, F_*(\mathbf{z}, \mathbf{o}) \otimes \phi_{\gamma_o}(\mathbf{o}) \rangle_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 p(\mathbf{z}, \mathbf{o}) d\mathbf{z} d\mathbf{o} \\ &= \int_{\mathcal{Z} \times \mathcal{O}} \mathbb{E}[(g - (1 + \tilde{\epsilon})g_i)(X, O) | Z = \mathbf{z}, O = \mathbf{o}]^2 p(\mathbf{z}, \mathbf{o}) d\mathbf{z} d\mathbf{o} \end{aligned}$$

$$\leq \|g - (1 + \tilde{\epsilon})g_i\|_{L^2(P_{XO})}^2,$$

where the inequality is deduced by Jensen's inequality. We find thus that

$$\mathcal{H}_{FO} \subseteq \bigcup_{i=1}^{2^{i-1}} B(f_i, \epsilon(1 + \tilde{\epsilon}), \|\cdot\|_{L^2(P_{ZO})}).$$

Hence $e_i(\text{id} : \mathcal{H}_{FO} \hookrightarrow L^2(P_{ZO})) \leq \epsilon(1 + \tilde{\epsilon})$. Since $\epsilon > e_i(\text{id} : \mathcal{H}_{\gamma_x, \gamma_o} \hookrightarrow L^2(P_{XO}))$ and $\tilde{\epsilon} > 0$ are arbitrary, it follows that $e_i(\text{id} : \mathcal{H}_{FO} \hookrightarrow L^2(P_{ZO})) \leq e_i(\text{id} : \mathcal{H}_{\gamma_x, \gamma_o} \hookrightarrow L^2(P_{XO}))$. \square

Corollary C.1. Suppose that P_{XO} satisfies Assumption 2.2 in the main text. We have that

$$\mu_{FO,i} \leq 4e_i(\text{id} : \mathcal{H}_{\gamma_x, \gamma_o} \hookrightarrow L^2(P_{XO}))^2$$

Proof. Let $\iota_{FO} : \mathcal{H}_{FO} \rightarrow L^2(P_{ZO})$ denote the embedding $\mathcal{H}_{FO} \hookrightarrow L^2(P_{ZO})$. We have, ($\forall i \geq 1$), the following chain of derivations

$$\begin{aligned} \mu_{FO,i} &= \lambda_i(\iota_{FO}^* \iota_{FO}) \stackrel{(a)}{=} s_i(\iota_{FO})^2 \stackrel{(b)}{=} a_i(\iota_{FO})^2 \\ &\stackrel{(c)}{\leq} 4e_i(\iota_{FO})^2 \stackrel{(d)}{\leq} 4e_i(\text{id} : \mathcal{H}_{\gamma_x, \gamma_o} \hookrightarrow L^2(P_{XO}))^2. \end{aligned}$$

In the above derivations, (a) follows from Steinwart and Christmann [2008, Eq. (A.25)], (b) follows from the paragraph after Steinwart and Christmann [2008, Eq. (A.29)], (c) follows from Steinwart and Christmann [2008, Eq. (A.44)], and (d) follows from Lemma C.2. \square

C.2 Embedding property of \mathcal{H}_{FO}

Definition 14 (Continuous embedding). A Hilbert space $(X, \|\cdot\|)$ is said to continuously embed into Hilbert space $(Y, \|\cdot\|)$ if $X \subset Y$ and there exists a constant C such that $\|x\|_Y \leq C\|x\|_X$ for all $x \in X$. We denote this as $X \hookrightarrow Y$. The embedding norm $\|X \hookrightarrow Y\|$ is defined as the smallest constant C for which the above inequality holds.

Definition 15 (Interpolation space). Assume that X_0 and X_1 are Hilbert spaces and $X_1 \subseteq X_0$. For $\theta \in (0, 1)$ and $x \in X_0$, we define the K -functional $K(t, x; X_0, X_1)$ as follows

$$K(t, x; X_0, X_1) := \inf_{y \in X_1} \{\|x - y\|_{X_0} + t\|y\|_{X_1}\}. \quad (63)$$

For $\theta \in (0, 1)$, we define interpolation space Hytönen et al. [2016]/[Definition C.3.1]

$$(X_0, X_1)_{\theta, 2} := \{x \in X_0 \mid \|x\|_{\theta, 2} < \infty\}, \quad \|x\|_{\theta, 2} := \left(\int_0^\infty \left(t^{-\theta} K(t, x; X_0, X_1) \right)^2 \frac{dt}{t} \right)^{\frac{1}{2}}.$$

We have proved in Section C.1 that \mathcal{H}_{FO} can be compactly embedded into $L^2(P_{ZO})$. Since P_{ZO} is equivalent to the Lebesgue measure over $\mathcal{Z} \times \mathcal{O}$ as per Assumption 4.4, \mathcal{H}_{FO} can also be compactly embedded into $L^2(\mathcal{Z} \times \mathcal{O})$. So we can define the θ -power space of \mathcal{H}_{FO} following [Steinwart and Scovel, 2012, Eq. (36)]

$$\left([\mathcal{H}_{FO}]_{P_{ZO}}^\theta, \|\cdot\|_{[\mathcal{H}_{FO}]_{P_{ZO}}^\theta} \right) := \left\{ \sum_{i \geq 1} a_i \mu_{FO,i}^{\theta/2} [e_{FO,i}] : (a_i) \in \ell_2(\mathbb{N}) \right\} \subseteq L^2(P_{ZO}).$$

For $0 < \theta < 1$, the θ -power space coincides with the interpolation space $[\mathcal{H}_{FO}]_{L^2(P_{ZO})}^\theta \cong [L_2(P_{ZO}), [\mathcal{H}_{FO}]_{P_{ZO}}]_{\theta, 2}$ [Steinwart and Scovel, 2012, Theorem 4.6]. We write $[\mathcal{H}_{FO}]_{L^2(P_{ZO})} := [\mathcal{H}_{FO}]_{L^2(P_{ZO})}^1$. Similarly, we write $[\mathcal{H}_{FO}]_{L^2(\mathcal{Z} \times \mathcal{O})} := [\mathcal{H}_{FO}]_{L^2(\mathcal{Z} \times \mathcal{O})}^1$.

Proposition C.2. Suppose Assumption 2.2 holds, and let m_o, m_z, ρ be as defined in the statement of Assumption 2.2. Suppose $\theta \in (0, 1)$ satisfies $\frac{d_o}{2m_o\theta} < 1$ and $\frac{d_z}{2m_z\theta} < 1$. Then there exists a constant $C_\theta > 0$ independent of n such that

$$\left\| k_{FO}^\theta \right\|_\infty := \left\| [\mathcal{H}_{FO}]_{P_{ZO}}^\theta \hookrightarrow L^\infty(P_{ZO}) \right\| \lesssim C_\theta \rho^\theta \gamma_o^{-\theta m_o}.$$

Proof. By Lemma C.3, we have $[\mathcal{H}_{FO}]_{L^2(\mathcal{Z} \times \mathcal{O})} \hookrightarrow MW^{m_z, m_o}(\mathcal{Z} \times \mathcal{O})$ and

$$\left\| [\mathcal{H}_{FO}]_{L^2(\mathcal{Z} \times \mathcal{O})} \hookrightarrow MW^{m_z, m_o}(\mathcal{Z} \times \mathcal{O}) \right\| \lesssim \rho \gamma_o^{-m_o}.$$

By Lemma C.6, since $\frac{d_o}{2s_o} < 1$ and $\frac{d_z}{2s_z} < 1$, we have $MW^{\theta m_z, \theta m_o}(\mathcal{Z} \times \mathcal{O}) \hookrightarrow L^\infty(\mathcal{Z} \times \mathcal{O})$, with embedding norm a universal constant independent of n . By Lemma C.5, we thus have $(L^2(\mathcal{Z} \times \mathcal{O}), [\mathcal{H}_{FO}]_{L^2(\mathcal{Z} \times \mathcal{O})})_{\theta, 2} \hookrightarrow L^\infty(\mathcal{Z} \times \mathcal{O})$ and

$$\left\| (L^2(\mathcal{Z} \times \mathcal{O}), [\mathcal{H}_{FO}]_{L^2(\mathcal{Z} \times \mathcal{O})})_{\theta, 2} \hookrightarrow L^\infty(\mathcal{Z} \times \mathcal{O}) \right\| \lesssim \rho^\theta \gamma_o^{-\theta m_o},$$

Next, since we know by Assumption 2.2 that P_{ZO} is equivalent to Lebesgue measure on $\mathcal{Z} \times \mathcal{O}$, Lemma C.4 shows that we have

$$(L^2(P_{ZO}), [\mathcal{H}_{FO}]_{L^2(P_{ZO})})_{\theta, 2} \hookrightarrow (L^2(\mathcal{Z} \times \mathcal{O}), [\mathcal{H}_{FO}]_{L^2(\mathcal{Z} \times \mathcal{O})})_{\theta, 2} \hookrightarrow L^\infty(P_{ZO}),$$

and we have

$$\left\| (L^2(P_{ZO}), [\mathcal{H}_{FO}]_{L^2(P_{ZO})})_{\theta, 2} \hookrightarrow L^\infty(P_{ZO}) \right\| \lesssim \rho^\theta \gamma_o^{-\theta m_o} \quad (64)$$

Finally, by Steinwart and Scovel [2012][Theorem 4.6], $(L^2(P_{ZO}), [\mathcal{H}_{FO}]_{L^2(P_{ZO})})_{\theta, 2} \cong [\mathcal{H}_{FO}]_{L^2(P_{ZO})}^\theta$ with the constant of equivalence depends only on θ . Putting it back to Eq. (64) completes the proof of the proposition. \square

Lemma C.3. Suppose Assumption 2.2 holds, and let m_z, m_o, ρ be as defined in the statement of Assumption 2.2. We have $[\mathcal{H}_{FO}]_{L^2(\mathcal{Z} \times \mathcal{O})} \hookrightarrow MW_2^{m_z, m_o}(\mathcal{Z} \times \mathcal{O})$ with

$$\left\| [\mathcal{H}_{FO}]_{L^2(\mathcal{Z} \times \mathcal{O})} \hookrightarrow MW_2^{m_z, m_o}(\mathcal{Z} \times \mathcal{O}) \right\| \lesssim \rho \gamma_o^{-m_o}.$$

Proof. Let $\text{id} : \mathcal{H}_{FO} \hookrightarrow L^2(\mathcal{Z} \times \mathcal{O})$ denote the canonical inclusion map. Since we have $(\ker \text{id})^\perp \cong [\mathcal{H}_{FO}]_{L^2(\mathcal{Z} \times \mathcal{O})}$, we may represent an arbitrary element of $[\mathcal{H}_{FO}]_{L^2(\mathcal{Z} \times \mathcal{O})}$ as $[f]$, where $f \in (\ker \text{id})^\perp$. Moreover, we have $\|[f]\|_{L^2(\mathcal{Z} \times \mathcal{O})} = \|f\|_{\mathcal{H}_{FO}}$. We fix f for the remainder of the proof.

Let $\boldsymbol{\alpha} \in \mathbb{N}^{d_z}$ with $|\boldsymbol{\alpha}| \leq m_z$ and $\boldsymbol{\beta} \in \mathbb{N}^{d_o}$ with $|\boldsymbol{\beta}| \leq m_o$, we have for any $f \in \mathcal{H}_{FO}$,

$$\begin{aligned} \left\| \partial_{\mathbf{z}}^{\boldsymbol{\alpha}} \partial_{\mathbf{o}}^{\boldsymbol{\beta}} f \right\|_{L^2(\mathcal{Z} \times \mathcal{O}; \mathbb{R})}^2 &= \int_{\mathcal{Z}} \int_{\mathcal{O}} \left(\partial_{\mathbf{z}}^{\boldsymbol{\alpha}} \partial_{\mathbf{o}}^{\boldsymbol{\beta}} f(\mathbf{z}, \mathbf{o}) \right)^2 d\mathbf{z} d\mathbf{o} \\ &\stackrel{(a)}{=} \int_{\mathcal{Z}} \int_{\mathcal{O}} \left(\partial_{\mathbf{z}}^{\boldsymbol{\alpha}} \partial_{\mathbf{o}}^{\boldsymbol{\beta}} \int_{\mathcal{X}} w(\mathbf{x}, \mathbf{o}) p(\mathbf{x} \mid \mathbf{z}, \mathbf{o}) d\mathbf{x} \right)^2 d\mathbf{z} d\mathbf{o} \\ &\stackrel{(b)}{=} \int_{\mathcal{Z}} \int_{\mathcal{O}} \left(\sum_{\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 = \boldsymbol{\beta}} \binom{\boldsymbol{\beta}}{\boldsymbol{\beta}_1} \int_{\mathcal{X}} \partial_{\mathbf{o}}^{\boldsymbol{\beta}_1} w(\mathbf{x}, \mathbf{o}) \partial_{\mathbf{z}}^{\boldsymbol{\alpha}} \partial_{\mathbf{o}}^{\boldsymbol{\beta}_2} p(\mathbf{x} \mid \mathbf{z}, \mathbf{o}) d\mathbf{x} \right)^2 d\mathbf{z} d\mathbf{o} \\ &\stackrel{(c)}{\leq} \int_{\mathcal{Z}} \int_{\mathcal{O}} \left(\sum_{\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 = \boldsymbol{\beta}} \binom{\boldsymbol{\beta}}{\boldsymbol{\beta}_1} \left(\int_{\mathcal{X}} (\partial_{\mathbf{o}}^{\boldsymbol{\beta}_1} w(\mathbf{x}, \mathbf{o}))^2 d\mathbf{x} \right)^{\frac{1}{2}} \left(\int_{\mathcal{X}} (\partial_{\mathbf{z}}^{\boldsymbol{\alpha}} \partial_{\mathbf{o}}^{\boldsymbol{\beta}_2} p(\mathbf{x} \mid \mathbf{z}, \mathbf{o}))^2 d\mathbf{x} \right)^{\frac{1}{2}} \right)^2 d\mathbf{z} d\mathbf{o} \end{aligned}$$

$$\begin{aligned}
&\leq \left(\sum_{\beta_1+\beta_2=\beta} \binom{\beta}{\beta_1} \right)^2 \left(\max_{\beta_2 \leq \beta} \sup_{\mathbf{x}, \mathbf{z}, \mathbf{o}} |\partial_{\mathbf{z}}^{\alpha} \partial_{\mathbf{o}}^{\beta_2} p(\mathbf{x} | \mathbf{z}, \mathbf{o})|^2 \right) \int_{\mathcal{Z}} \int_{\mathcal{O}} \max_{\beta_1 \leq \beta} \int_{\mathcal{X}} (\partial_{\mathbf{o}}^{\beta_1} w(\mathbf{x}, \mathbf{o}))^2 d\mathbf{x} d\mathbf{z} d\mathbf{o} \\
&= \left(\sum_{\beta_1+\beta_2=\beta} \binom{\beta}{\beta_1} \right)^2 \left(\max_{\beta_2 \leq \beta} \sup_{\mathbf{x}, \mathbf{z}, \mathbf{o}} |\partial_{\mathbf{z}}^{\alpha} \partial_{\mathbf{o}}^{\beta_2} p(\mathbf{x} | \mathbf{z}, \mathbf{o})|^2 \right) \max_{\beta_1 \leq \beta} \int_{\mathcal{O}} \int_{\mathcal{X}} (\partial_{\mathbf{o}}^{\beta_1} w(\mathbf{x}, \mathbf{o}))^2 d\mathbf{x} d\mathbf{o} \\
&\stackrel{(d)}{=} c_{m_o, d_o} \left(\max_{\beta_2 \leq \beta} \sup_{\mathbf{x}, \mathbf{z}, \mathbf{o}} |\partial_{\mathbf{z}}^{\alpha} \partial_{\mathbf{o}}^{\beta_2} p(\mathbf{x} | \mathbf{z}, \mathbf{o})|^2 \right) \gamma_o^{-2|\beta|} \|w\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 \\
&\leq c_{m_o, d_o} \left(\max_{\alpha \leq m_z} \max_{\beta \leq m_o} \sup_{\mathbf{x}, \mathbf{z}, \mathbf{o}} |\partial_{\mathbf{z}}^{\alpha} \partial_{\mathbf{o}}^{\beta} p(\mathbf{x} | \mathbf{z}, \mathbf{o})|^2 \right) \gamma_o^{-2m_o} \|w\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2,
\end{aligned}$$

where (a) follows from the fact that, for any $f \in \mathcal{H}_{FO}$, there exists $w \in \mathcal{H}_{\gamma_x, \gamma_o}$ such that $f(\mathbf{z}, \mathbf{o}) = \langle w, F_*(\mathbf{z}, \mathbf{o}) \otimes \phi_{\gamma_o}(\mathbf{o}) \rangle_{\mathcal{H}_{\gamma_x, \gamma_o}} = \int_{\mathcal{X}} w(\mathbf{x}, \mathbf{o}) p(\mathbf{x} | \mathbf{z}, \mathbf{o}) d\mathbf{x}$ and $\|f\|_{\mathcal{H}_{FO}} = \|w\|_{\mathcal{H}_{\gamma_x, \gamma_o}}$. (b) follows from generalized Leibniz's rule and differentiation under the integral sign Klenke [2013][Theorem 6.28], (c) follows from a Cauchy-Schwarz inequality. (d) follows from the following arguments.

From Steinwart and Christmann [2008, Theorem 4.21], we know that for any $w \in \mathcal{H}_{\gamma_x, \gamma_o}$, there exists $g \in L^2(\mathbb{R}^{d_x+d_o})$ such that

$$w(\mathbf{x}, \mathbf{o}) = \langle g, \Phi_{\gamma_x}(\mathbf{x}) \Phi_{\gamma_o}(\mathbf{o}) \rangle_{L^2(\mathbb{R}^{d_x+d_o})}, \quad \|w\|_{\mathcal{H}_{\gamma_x, \gamma_o}} = \|g\|_{L^2(\mathbb{R}^{d_x+d_o})},$$

with $\Phi_{\gamma_x} : \mathcal{X} \rightarrow L^2(\mathbb{R}^{d_x})$, $\Phi_{\gamma_o} : \mathcal{O} \rightarrow L^2(\mathbb{R}^{d_o})$ defined in Steinwart and Christmann [2008, Lemma 4.45]. We have

$$\begin{aligned}
&\int_{\mathcal{O}} \int_{\mathcal{X}} (\partial_{\mathbf{o}}^{\beta_1} w(\mathbf{x}, \mathbf{o}))^2 d\mathbf{x} d\mathbf{o} \\
&= \int_{\mathcal{O}} \int_{\mathcal{X}} \left(\partial_{\mathbf{o}}^{\beta_1} \int_{\mathbb{R}^{d_x+d_o}} g(\mathbf{x}', \mathbf{o}') \Phi_{\gamma_x}(\mathbf{x})(\mathbf{x}') \Phi_{\gamma_o}(\mathbf{o})(\mathbf{o}') d\mathbf{x}' d\mathbf{o}' \right)^2 d\mathbf{x} d\mathbf{o} \\
&\stackrel{(i)}{=} \int_{\mathcal{O}} \int_{\mathcal{X}} \left(\int_{\mathbb{R}^{d_x+d_o}} g(\mathbf{x}', \mathbf{o}') \Phi_{\gamma_x}(\mathbf{x})(\mathbf{x}') \partial_{\mathbf{o}}^{\beta_1} (\Phi_{\gamma_o}(\mathbf{o})(\mathbf{o}')) d\mathbf{x}' d\mathbf{o}' \right)^2 d\mathbf{x} d\mathbf{o} \\
&\leq \|g\|_{L^2(\mathbb{R}^{d_x+d_o})}^2 \int_{\mathcal{O}} \int_{\mathcal{X}} \int_{\mathbb{R}^{d_x+d_o}} \left(\Phi_{\gamma_x}(\mathbf{x})(\mathbf{x}') \partial_{\mathbf{o}}^{\beta_1} (\Phi_{\gamma_o}(\mathbf{o})(\mathbf{o}')) \right)^2 d\mathbf{x}' d\mathbf{o}' d\mathbf{x} d\mathbf{o} \\
&= \|g\|_{L^2(\mathbb{R}^{d_x+d_o})}^2 \int_{\mathcal{O}} \int_{\mathcal{X}} \left(\int_{\mathbb{R}^{d_x}} (\Phi_{\gamma_x}(\mathbf{x})(\mathbf{x}'))^2 d\mathbf{x}' \right) \left(\int_{\mathbb{R}^{d_o}} \left(\partial_{\mathbf{o}}^{\beta_1} (\Phi_{\gamma_o}(\mathbf{o})(\mathbf{o}')) \right)^2 d\mathbf{o}' \right) d\mathbf{x} d\mathbf{o} \\
&\stackrel{(ii)}{=} \|g\|_{L^2(\mathbb{R}^{d_x+d_o})}^2 \int_{\mathcal{O}} \int_{\mathbb{R}^{d_o}} \left(\partial_{\mathbf{o}}^{\beta_1} (\Phi_{\gamma_o}(\mathbf{o})(\mathbf{o}')) \right)^2 d\mathbf{o}' d\mathbf{o} \\
&\stackrel{(iii)}{\leq} \|g\|_{L^2(\mathbb{R}^{d_x+d_o})}^2 c_{\beta_1, d_o} \gamma_o^{-2|\beta_1|} \\
&= \|w\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 c_{\beta_1, d_o} \gamma_o^{-2|\beta_1|},
\end{aligned}$$

where we're allowed to exchange differentiation and integration in (i) using the differentiation lemma Klenke [2013][Theorem 6.28], (ii) follows from the fact that

$$\int_{\mathbb{R}^{d_x}} (\Phi_{\gamma_x}(\mathbf{x})(\mathbf{x}'))^2 d\mathbf{x}' = \int_{\mathbb{R}^{d_x}} \left(\frac{2^{d_x/2}}{\pi^{d_x/4} \gamma_x^{d_x/2}} \exp \left(-2 \frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{\gamma_x^2} \right) \right)^2 d\mathbf{x}' = 1,$$

and (iii) follows from the proof of Steinwart and Christmann [2008][Theorem 4.48]. Here c_{β_1, d_o} is a constant only depending on β_1, d_o . Hence we've shown

$$\|f\|_{MW_2^{m_z, m_o}(\mathcal{Z} \times \mathcal{O}; \mathbb{R})}^2$$

$$\begin{aligned}
&= \sum_{\alpha \leq m_z, \beta \leq m_o} \left\| \partial_{\mathbf{z}}^{\alpha} \partial_{\mathbf{o}}^{\beta} f \right\|_{L^2(\mathcal{Z} \times \mathcal{O}; \mathbb{R})}^2 \\
&\lesssim \sum_{\alpha \leq m_z, \beta \leq m_o} \left(\max_{\alpha \leq m_z} \max_{\beta \leq m_o} \sup_{\mathbf{x}, \mathbf{z}, \mathbf{o}} \left| \partial_{\mathbf{z}}^{\alpha} \partial_{\mathbf{o}}^{\beta} p(\mathbf{x} \mid \mathbf{z}, \mathbf{o}) \right|^2 \right) \gamma_o^{-2m_o} \|w\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 \\
&= \left(\sum_{\alpha \leq m_z, \beta \leq m_o} 1 \right) \left(\max_{\alpha \leq m_z} \max_{\beta \leq m_o} \sup_{\mathbf{x}, \mathbf{z}, \mathbf{o}} \left| \partial_{\mathbf{z}}^{\alpha} \partial_{\mathbf{o}}^{\beta} p(\mathbf{x} \mid \mathbf{z}, \mathbf{o}) \right|^2 \right) \gamma_o^{-2m_o} \|f\|_{\mathcal{H}_{FO}}^2 \\
&\lesssim \rho^2 \gamma_o^{-2m_o} \|f\|_{\mathcal{H}_{FO}}^2 \\
&= \rho^2 \gamma_o^{-2m_o} \|f\|_{[\mathcal{H}_{FO}]_{L^2(\mathcal{Z} \times \mathcal{O})}}^2.
\end{aligned}$$

The second last inequality holds by Assumption 2.2. This concludes the proof. \square

C.3 Auxiliary results for Section C

Lemma C.4. Let ν_1, ν_2 be two measures on \mathcal{Z} which are equivalent, i.e $0 < c' \leq \frac{d\nu_2}{d\nu_1} \leq c < \infty$. Let $\mathcal{H} \subseteq L^2(\nu_1)$ be a Hilbert space. Then $(L^2(\nu_1), \mathcal{H})_{\theta,2} \hookrightarrow (L^2(\nu_2), \mathcal{H})_{\theta,2}$ with embedding norm $\|(L^2(\nu_1), \mathcal{H})_{\theta,2} \hookrightarrow (L^2(\nu_2), \mathcal{H})_{\theta,2}\| \leq c^{1-\theta}$.

Proof. For any $f \in (L^2(\nu_1), \mathcal{H})_{\theta,2}$, we have

$$\begin{aligned}
\|f\|_{(L^2(\nu_2), \mathcal{H})_{\theta,2}} &= \left(\int_0^\infty \left(t^{-\theta} K(t, x; L^2(\nu_2), \mathcal{H}) \right)^2 \frac{dt}{t} \right)^{\frac{1}{2}} \\
&= \left(\int_0^\infty \left(t^{-\theta} \inf_{y \in \mathcal{H}} \{ \|x - y\|_{L^2(\nu_2)} + t\|y\|_{\mathcal{H}} \} \right)^2 \frac{dt}{t} \right)^{\frac{1}{2}} \\
&\stackrel{(a)}{\leq} \left(\int_0^\infty \left(ct^{-\theta} \inf_{y \in \mathcal{H}} \left\{ \|x - y\|_{L^2(\nu_1)} + \frac{t}{c}\|y\|_{\mathcal{H}} \right\} \right)^2 \frac{dt}{t} \right)^{\frac{1}{2}} \\
&\stackrel{(b)}{=} c^{1-\theta} \left(\int_0^\infty \left(t^{-\theta} \inf_{y \in \mathcal{H}} \{ \|x - y\|_{L^2(\nu_1)} + t\|y\|_{\mathcal{H}} \} \right)^2 \frac{dt}{t} \right)^{\frac{1}{2}} \\
&= c^{1-\theta} \|f\|_{(L^2(\nu_1), \mathcal{H})_{\theta,2}}
\end{aligned}$$

In the above derivations, we use $\|x - y\|_{L^2(\nu_2)} \leq c\|x - y\|_{L^2(\nu_1)}$ in (a), and the change of variables $t \mapsto ct$ in (b). \square

Lemma C.5. Let $\mathcal{Z} = [0, 1]^d$. Let $\mathcal{H}, W \subseteq L^2(\mathcal{Z})$ be two Hilbert spaces. Suppose that $\mathcal{H} \hookrightarrow W$ with embedding norm $\|\mathcal{H} \hookrightarrow W\| \leq c$. Suppose that $\theta \in (0, 1)$ is chosen so that $(L^2(\mathcal{Z}), W)_{\theta,2} \hookrightarrow L^\infty(\mathcal{Z})$ holds. Then we have $(L^2(\mathcal{Z}), \mathcal{H})_{\theta,2} \hookrightarrow L^\infty(\mathcal{Z})$ with embedding norm $\leq c^\theta c_0$, where c_0 only depends on θ, m and d_z .

Proof. We adapt the proof of Kanagawa et al. [2018][Lemma A.2, Corollary 4.13]. We have, for any $x \in (L^2(\mathcal{Z}), W)_{\theta,2}$,

$$\|x\|_{(L^2(\mathcal{Z}), W)_{\theta,2}} = \left(\int_0^\infty \left(t^{-\theta} K(t, x; L^2(\mathcal{Z}), W) \right)^2 \frac{dt}{t} \right)^{\frac{1}{2}}$$

$$\begin{aligned}
&= \left(\int_0^\infty \left(t^{-\theta} \inf_{y \in W} \{ \|x - y\|_{L^2(\mathcal{Z})} + t\|y\|_W \} \right)^2 \frac{dt}{t} \right)^{\frac{1}{2}} \\
&\leq \left(\int_0^\infty \left(t^{-\theta} \inf_{y \in \mathcal{H}} \{ \|x - y\|_{L^2(\mathcal{Z})} + t\|y\|_W \} \right)^2 \frac{dt}{t} \right)^{\frac{1}{2}} \\
&\leq \left(\int_0^\infty \left(t^{-\theta} \inf_{y \in \mathcal{H}} \{ \|x - y\|_{L^2(\mathcal{Z})} + tc\|y\|_{\mathcal{H}} \} \right)^2 \frac{dt}{t} \right)^{\frac{1}{2}} \\
&= c^\theta \left(\int_0^\infty \left((tc)^{-\theta} \inf_{y \in \mathcal{H}} \{ \|x - y\|_{L^2(\mathcal{Z})} + tc\|y\|_{\mathcal{H}} \} \right)^2 \frac{dt}{t} \right)^{\frac{1}{2}} \\
&= c^\theta \left(\int_0^\infty \left(t^{-\theta} K(t, x; L^2(\mathcal{Z}), \mathcal{H}) \right)^2 \frac{dt}{t} \right)^{\frac{1}{2}} \\
&= c^\theta \|x\|_{(L^2(\mathcal{Z}), \mathcal{H})_{\theta,2}}.
\end{aligned}$$

Notice that we correct an error in [Kanagawa et al. \[2018\]](#) which obtained the exponent $-\theta$ instead of θ due to an error in the change of variables argument in (a). Now we have proved that $(L^2(\mathcal{Z}), \mathcal{H})_{\theta,2} \hookrightarrow (L^2(\mathcal{Z}), W)_{\theta,2}$ and $\|(L^2(\mathcal{Z}), \mathcal{H})_{\theta,2} \hookrightarrow (L^2(\mathcal{Z}), W)_{\theta,2}\| \leq c^\theta$. On the other hand, by assumption, we have $(L^2(\mathcal{Z}), W)_{\theta,2} \hookrightarrow L^\infty(\mathcal{Z})$. Consequently, we have $(L^2(\mathcal{Z}), \mathcal{H})_{\theta,2} \hookrightarrow W(L^2(\mathcal{Z}), W)_{\theta,2} \hookrightarrow L^\infty(\mathcal{Z})$. The embedding norm is bounded by $c^\theta c_0$, where c_0 depends only on θ, m , and d_z . \square

Lemma C.6. *If $\frac{d_o}{2s_o} < 1$ and $\frac{d_z}{2s_z} < 1$, then the dominating mixed smoothness Sobolev space $MW_2^{s_o, s_z}([0, 1]^{d_o+d_z})$ continuously embeds into $L^\infty([0, 1]^{d_o+d_z})$.*

Proof. See [\[Schmeisser, 2007, Equation 1.13\]](#) for the result where the domain is $\mathbb{R}^{d_z+d_o}$. By [DeVore and Sharpley \[1993\]](#), there exists a continuous extension operator $\mathcal{E} : MW_2^{s_o, s_z}([0, 1]^{d_o+d_z}) \rightarrow MW_2^{s_o, s_z}(\mathbb{R}^{d_o+d_z})$ such that $\|\mathcal{E}[f]\|_{MW_2^{s_o, s_z}(\mathbb{R}^{d_o+d_z})} \leq C' \|f\|_{MW_2^{s_o, s_z}([0, 1]^{d_o+d_z})}$ holds for some universal constant C' . Hence, for $f \in MW_2^{s_o, s_z}([0, 1]^{d_o+d_z})$, we thus have $\|f\|_{L^\infty([0, 1]^{d_o+d_z})} \leq \|\mathcal{E}[f]\|_{L^\infty(\mathbb{R}^{d_o+d_z})} \leq C \|\mathcal{E}[f]\|_{MW_2^{s_o, s_z}(\mathbb{R}^{d_o+d_z})} \leq C' \|f\|_{MW_2^{s_o, s_z}([0, 1]^{d_o+d_z})}$, for some universal constants $C, C' > 0$. \square

Lemma C.7. *Let $\mathcal{Z} = [0, 1]^{d_z}$ and $\mathcal{O} = [0, 1]^{d_o}$. The interpolation space $[L^2(\mathcal{Z} \times \mathcal{O}), W_2^{s_z}(\mathcal{Z}) \otimes W_2^{s_o}(\mathcal{O})]_{\theta,2} \cong W_2^{s_z\theta}(\mathcal{Z}) \otimes W_2^{s_o\theta}(\mathcal{O})$.*

Proof. The proof follows immediately from [Defant and Michels \[2000\]](#). \square

Lemma C.8. *Let \mathcal{G} be the vector-valued RKHS of the operator-valued kernel defined in Eq. (4) with $k_Z, k_{O,1}$ being Sobolev reproducing kernels of smoothness t_z, t_o . Then,*

$$\begin{aligned}
\mathcal{G} &\stackrel{(a)}{\cong} \mathcal{H}_X \otimes \mathcal{H}_Z \otimes \mathcal{H}_{O,1} \stackrel{(b)}{\cong} \mathcal{H}_X \otimes W_2^{t_z}(\mathcal{Z}; \mathbb{R}) \otimes W_2^{t_o}(\mathcal{O}; \mathbb{R}) \\
&\stackrel{(c)}{\cong} \mathcal{H}_X \otimes MW_2^{t_z, t_o}(\mathcal{Z} \times \mathcal{O}; \mathbb{R}) \stackrel{(d)}{\cong} MW_2^{t_z, t_o}(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_X),
\end{aligned} \tag{65}$$

Proof. (a) is proved in [Li et al. \[2024a, Theorem 1\]](#), (b) holds by definition of Sobolev reproducing kernels, (c) is proved in [Aubin \[2011, Theorem 12.7.2\]](#) and (d) is an extension of [Aubin \[2011, Theorem 12.7.1, Theorem 12.4.1\]](#). \square

Roadmap for Proof of Theorem 4.1

The following figures illustrate how [Theorem 4.1](#) follows from supporting propositions and lemmas (with arrows indicate that one result is used in the proof of another).

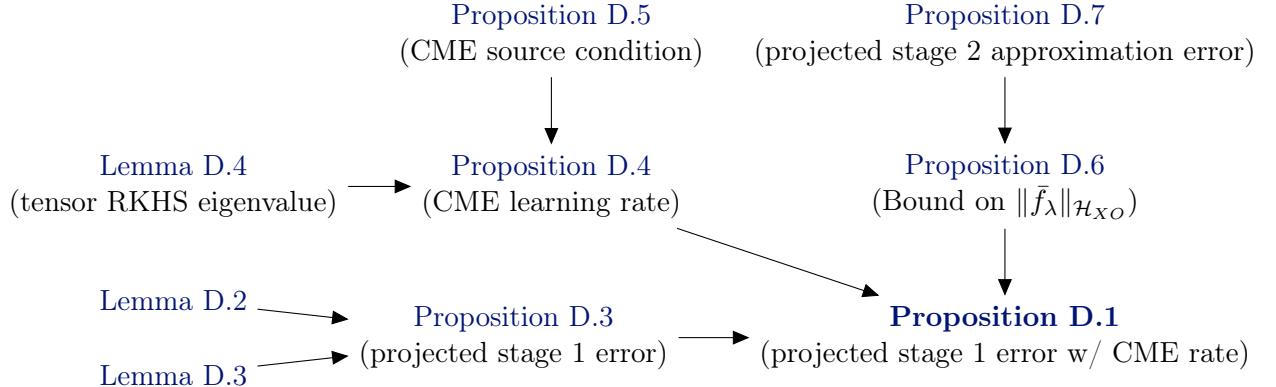


Figure 3: Roadmap for proof of Proposition D.1.

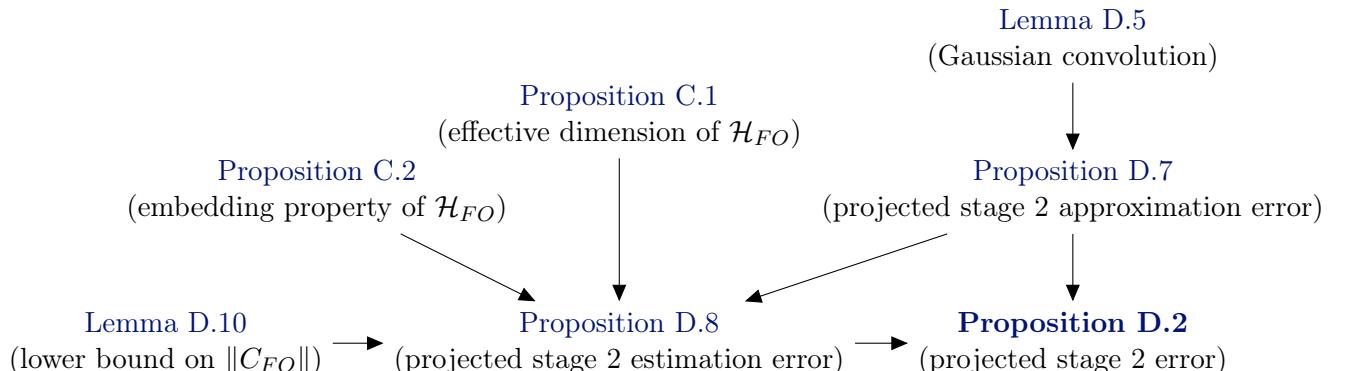


Figure 4: Roadmap for proof of Proposition D.2.

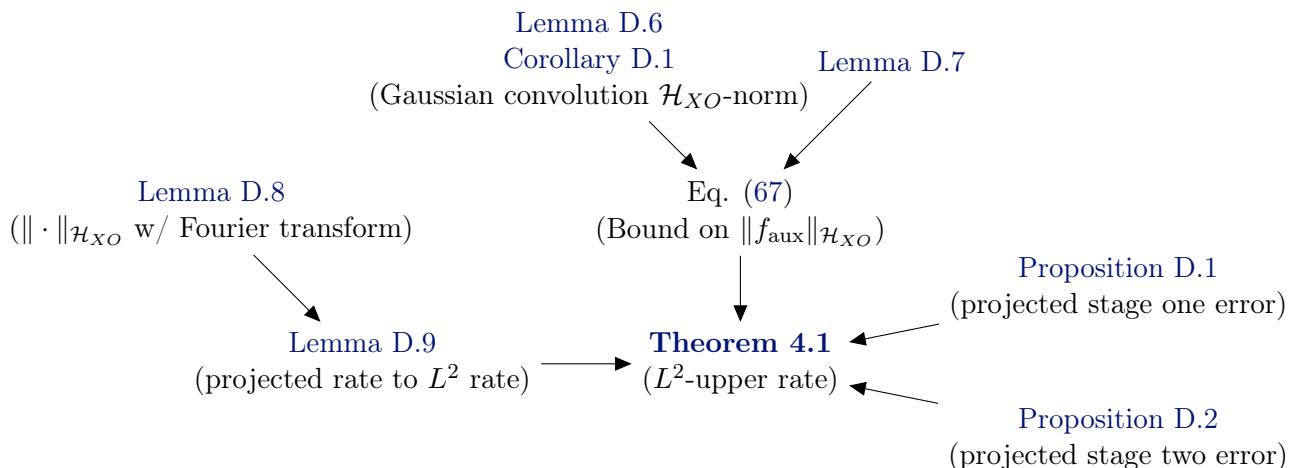


Figure 5: Roadmap for proof of Theorem 4.1.

D Proof of Theorem 4.1 in the main text

We first construct the auxiliary RKHS function f_{aux} defined in Eq. (40) in the main text, and give an upper bound on $\|f_{\text{aux}}\|_{\mathcal{H}_{\gamma_x, \gamma_o}}$. Let $r = \max\{\lfloor s_x \rfloor, \lfloor s_o \rfloor\} + 1$. For $\gamma = (\gamma_1, \dots, \gamma_d)$, we define an *approximate identity* [Giné and Nickl, 2021, Section 4.1.2] $K_\gamma : \mathbb{R}^d \rightarrow \mathbb{R}$ via

$$K_1(\mathbf{x}) := \sum_{j=1}^r \binom{r}{j} (-1)^{1-j} \frac{1}{j^d} \left(\frac{2}{\pi}\right)^{\frac{d}{2}} \exp\left(-2 \sum_{i=1}^d \frac{x_i^2}{j^2}\right), \quad K_\gamma(\mathbf{x}) := K_1\left(\frac{\mathbf{x}}{\gamma}\right) \prod_{i=1}^d \frac{1}{\gamma_i} \quad (66)$$

Note that Eq. (66) reduces to Eberts and Steinwart [2013][Eq. (8)] when $\gamma_1 = \dots = \gamma_d, s_1 = \dots, s_d$. We define $K_{\gamma_x} : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$ to be K_{γ_x} with $\gamma_x = [\gamma_x, \dots, \gamma_x] \in \mathbb{R}^{d_x}$ (note this agrees with Eq. (36) in the main) and $K_{\gamma_o} : \mathbb{R}^{d_o} \rightarrow \mathbb{R}$ to be K_{γ_o} with $\gamma_o = [\gamma_o, \dots, \gamma_o] \in \mathbb{R}^{d_o}$. Define $\iota_{x, \gamma_x^{-1}} : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$ as an indicator function of $\{\omega_x : \|\omega_x\| \leq \gamma_x^{-1}\}$.

By the Young's Convolution Inequality and the fact that \mathcal{F} is unitary, we have

$$\left\| f_*(\cdot, \mathbf{o}) * \mathcal{F}^{-1}[\iota_{x, \gamma_x^{-1}}] \right\|_{L^2(\mathbb{R}^{d_x})} \leq \|f_*(\cdot, \mathbf{o})\|_{L^1(\mathbb{R}^{d_x})} \|\iota_{x, \gamma_x^{-1}}\|_{L^2(\mathbb{R}^{d_x})}$$

The right hand side is finite by [Assumption 2.3](#) in the main text. Hence we apply the Fourier operator \mathcal{F} to $f_*(\cdot, \mathbf{o}) * \mathcal{F}^{-1}[\iota_{x, \gamma_x^{-1}}]$ to find that, ($\forall \mathbf{o} \in \mathcal{O}$),

$$\mathcal{F}[f_*(\cdot, \mathbf{o}) * \mathcal{F}^{-1}[\iota_{x, \gamma_x^{-1}}]] = \mathcal{F}[f_*(\cdot, \mathbf{o})] \cdot \iota_{x, \gamma_x^{-1}}.$$

Therefore, $f_* * \mathcal{F}^{-1}[\iota_{x, \gamma_x^{-1}}] = f_{*, \text{low}}$ where $f_{*, \text{low}}$ is defined in Eq. (38). We define $f_{*, \text{high}} = f_* - f_{*, \text{low}}$. Then we have, ($\forall \mathbf{o} \in \mathcal{O}$),

$$\mathcal{F}[f_{*, \text{high}}(\cdot, \mathbf{o})] = \mathcal{F}[f_*(\cdot, \mathbf{o})] - \mathcal{F}[f_{*, \text{low}}(\cdot, \mathbf{o})]$$

satisfies Eq. (39) in the main text. We let f_{aux} be as defined in Eq. (40) in the main text. We have

$$\|f_{\text{aux}}\|_{\mathcal{H}_{\gamma_x, \gamma_o}} \quad (67)$$

$$\leq \left\| f_* * K_{\gamma_o} * \mathcal{F}^{-1}[\iota_{x, \gamma_x^{-1}}] \right\|_{\mathcal{H}_{\gamma_x, \gamma_o}} + \|K_{\gamma_x} * K_{\gamma_o} * (f_* - f_{*, \text{low}})\|_{\mathcal{H}_{\gamma_x, \gamma_o}}$$

$$\stackrel{(a)}{\leq} n^{\frac{1}{2} \frac{1+\frac{d_o}{s_o}(\frac{s_x}{d_x}+\eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}} \|f_*\|_{L^2(\mathbb{R}^{d_x+d_o})} + n^{\frac{1}{2} \frac{1+\frac{d_o}{s_o}(\frac{s_x}{d_x}+\eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}} \|f_* - f_{*, \text{low}}\|_{L^2(\mathbb{R}^{d_x+d_o})}$$

$$\stackrel{(b)}{\leq} 3n^{\frac{1}{2} \frac{1+\frac{d_o}{s_o}(\frac{s_x}{d_x}+\eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}}. \quad (68)$$

In the above chain of derivations, (a) holds by applying [Lemma D.7](#) to the first term and applying [Corollary D.1](#) to the second term, (b) holds by $\|f_{*, \text{low}}\|_{L^2(\mathbb{R}^{d_x+d_o})} \leq \|f_*\|_{L^2(\mathbb{R}^{d_x+d_o})} \leq 1$ by Plancherel's Theorem and [Assumption 2.3](#). Notice that,

$$(*) := \left\| \hat{f}_\lambda - f_* \right\|_{L^2(P_{XO})} \leq \left\| \hat{f}_\lambda - f_{\text{aux}} \right\|_{L^2(P_{XO})} + \|f_* - f_{\text{aux}}\|_{L^2(P_{XO})}. \quad (69)$$

By definition of \hat{f}_λ in Eq. (7) in the main text, it satisfies

$$\lambda \|\hat{f}_\lambda\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 + \frac{1}{n} \sum_{i=1}^n \left(y_i - \left\langle f, \hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) \otimes \phi_{\gamma_o}(o_i) \right\rangle_{\mathcal{H}_{\gamma_x, \gamma_o}} \right)^2 \leq \frac{1}{n} \sum_{i=1}^n y_i^2. \quad (70)$$

Notice that

$$\frac{1}{n} \sum_{i=1}^n y_i^2 = \frac{1}{n} \sum_{i=1}^n ((Tf_*)(\mathbf{z}_i, \mathbf{o}_i) + v_i)^2 \lesssim 2 + \frac{2}{n} \sum_{i=1}^n v_i^2,$$

where the inequality follows from

$$\|Tf_*\|_{L^\infty(\mathcal{Z} \times \mathcal{O})} \leq \|f_*\|_{L^\infty(\mathcal{X} \times \mathcal{O})} \leq \|f_*\|_{L^\infty(\mathbb{R}^{d_x+d_o})} \lesssim 1,$$

by [Assumption 2.3](#). By [Assumption 4.5](#), v_1, \dots, v_n are n i.i.d. mean zero σ -sub-Gaussian random variables so v_1^2, \dots, v_n^2 are n i.i.d. σ^2 -sub-exponential random variables [[Vershynin, 2018](#), Lemma 2.7.6] with mean $\mathbb{E}[v_1^2] < \infty$. By Exercise 2.7.10 (Centering) of [Vershynin \[2018\]](#) we know that $v_1^2 - \mathbb{E}[v_1^2], \dots, v_n^2 - \mathbb{E}[v_n^2]$ are n i.i.d. $C\sigma^2$ -sub-exponential random variables for some universal constant C . By [Vershynin \[2018\]](#), Theorem 2.8.1 (Bernstein's inequality)], we have

$$(\forall t \geq 0), P\left(\left|\sum_{i=1}^n v_i^2\right| \geq n(1 + \mathbb{E}[v_i^2])\right) \leq 2 \exp\left(-c \min\left(\frac{n}{C^2\sigma^4}, \frac{n}{C\sigma^2}\right)\right),$$

where c is a universal constant. For a fixed $\tau \geq 1$, for sufficiently large $n \geq 1$, the above right hand side $\leq 2 \exp(-\tau)$. Thus for a fixed $\tau \geq 1$, with P^n -probability $\geq 1 - 2e^{-\tau}$, for sufficiently large $n \geq 1$, we have

$$\frac{1}{n} \sum_{i=1}^n y_i^2 \lesssim 2 + \frac{2}{n} \sum_{i=1}^n v_i^2 \leq 4 + 2\mathbb{E}[v_i^2].$$

Under the same high probability event, from Eq. (70), using $\lambda = \frac{1}{n}$, we have

$$\|\hat{f}_\lambda\|_{\mathcal{H}_{\gamma_x, \gamma_o}} \lesssim \sqrt{4 + 2\mathbb{E}[v_i^2]} \sqrt{n}. \quad (71)$$

Also, we have

$$\|f_{\text{aux}}\|_{\mathcal{H}_{\gamma_x, \gamma_o}} \leq 2\sqrt{n} \quad (72)$$

proved above in Eq. (67). Continuing from Eq. (69), we apply [Lemma D.9](#) to $\|\hat{f}_\lambda - f_{\text{aux}}\|_{L^2(P_{X_O})}$, where we notice that $\hat{f}_\lambda, f_{\text{aux}} \in \mathcal{H}_{\gamma_x, \gamma_o}$, and we use Eq. (71) and Eq. (72) to verify the assumption of that lemma. We find, with P^n -probability $\geq 1 - 2e^{-\tau}$,

$$(*) \leq \gamma_x^{-d_x \eta_0} (\log n)^{\frac{d_x \eta_0}{2}} \left\| T[\hat{f}_\lambda] - Tf_{\text{aux}} \right\|_{L^2(P_{Z_O})} + n^{-\frac{\frac{s_x}{d_x}}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}} + \|f_* - f_{\text{aux}}\|_{L^2(P_{X_O})}.$$

From Eq. (84), $\|f_* - f_{\text{aux}}\|_{L^2(P_{X_O})}$ can be upper bounded (up to a constant) by the second last term above and hence subsumed. Through a triangular inequality, we have with P^n -probability $\geq 1 - 2e^{-\tau}$,

$$\begin{aligned} (*) &\leq \gamma_x^{-d_x \eta_0} (\log n)^{\frac{d_x \eta_0}{2}} \left(\left\| T[\hat{f}_\lambda] - Tf_* \right\|_{L^2(P_{Z_O})} + \|Tf_* - Tf_{\text{aux}}\|_{L^2(P_{Z_O})} \right) \\ &\quad + n^{-\frac{\frac{s_x}{d_x}}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}}. \end{aligned} \quad (73)$$

We prove in Eq. (83) in Section D.2 that $\|Tf_* - Tf_{\text{aux}}\|_{L^2(P_{ZO})} \leq n^{-\frac{\frac{s_x}{d_x} + \eta_1}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}}$, so we continue from above to obtain, with P^n -probability $\geq 1 - 2e^{-\tau}$,

$$(*) \lesssim \gamma_x^{-d_x \eta_0} (\log n)^{\frac{d_x \eta_0}{2}} \left\| T \left[\hat{f}_\lambda \right] - Tf_* \right\|_{L^2(P_{ZO})} + n^{-\frac{\frac{s_x}{d_x} + \eta_1 - \eta_0}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} (\log n)^{\frac{d_x \eta_0}{2}}. \quad (74)$$

Note the last term of Eq. (73) is subsumed by the second term above since $\eta_0 \geq \eta_1$. Recall the definition of \bar{f}_λ in Eq. (23) in the main text. Then, we have, through a triangular inequality, with P^n -probability $\geq 1 - 2e^{-\tau}$, the following holds

$$\begin{aligned} (*) &\leq \gamma_x^{-d_x \eta_0} (\log n)^{\frac{d_x \eta_0}{2}} \left(\underbrace{\left\| T \left[\hat{f}_\lambda \right] - T \left[\bar{f}_\lambda \right] \right\|_{L^2(P_{ZO})}}_{\text{Projected Stage I Error}} + \underbrace{\left\| T \left[\bar{f}_\lambda \right] - Tf_* \right\|_{L^2(P_{ZO})}}_{\text{Projected Stage II Error}} \right) \\ &\quad + n^{-\frac{\frac{s_x}{d_x} + \eta_1 - \eta_0}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} (\log n)^{\frac{d_x \eta_0}{2}}. \end{aligned}$$

The projected stage I error can be upper bounded by Proposition D.1 as $n > A_{\lambda, \tau}$ is satisfied for sufficiently large $n \geq 1$ proved in Eq. (89).

$$\text{Projected Stage I Error} = \left\| T \left[\hat{f}_\lambda \right] - T \left[\bar{f}_\lambda \right] \right\|_{L^2(P_{ZO})} \leq \tau n^{-\frac{\frac{s_x}{d_x} + \eta_1}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}}$$

holds with $P^{n+\tilde{n}}$ -probability $\geq 1 - 34e^{-\tau}$. The projected stage II error can be upper bounded by Proposition D.2.

$$\text{Projected Stage II Error} = \left\| T \left[\bar{f}_\lambda \right] - Tf_* \right\|_{L^2(P_{ZO})} \leq \tau n^{-\frac{\frac{s_x}{d_x} + \eta_1}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} \cdot (\log n)^{\frac{d_x + d_o + 1}{2}}$$

holds with P^n -probability $\geq 1 - 4e^{-\tau}$. Combine the above two upper bounds, and we have

$$(*) = \left\| \left[\hat{f}_\lambda \right] - f_* \right\|_{L^2(P_{ZO})} \leq \tau n^{-\frac{\frac{s_x}{d_x} + \eta_1 - \eta_0}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} (\log n)^{\frac{d_x + d_o + 1 + d_x \eta_0}{2}} \quad (75)$$

holds with $P^{n+\tilde{n}}$ -probability $1 - 40e^{-\tau}$. So the proof is concluded. \square

Proposition D.1 (Projected stage-I error). *Suppose that the assumptions of Proposition D.4 hold. Suppose that $n > A_{\lambda, \tau}$ with $A_{\lambda, \tau}$ defined in Eq. (87). Suppose Assumptions 2.2 hold. Suppose that $\tilde{n} \geq 1$ satisfies Eq. (21). Let $\lambda \asymp n^{-1}$ and*

$$\gamma_x = n^{-\frac{\frac{1}{d_x}}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}}, \quad \gamma_o = n^{-\frac{\frac{1}{s_o}(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}}.$$

Then, with $P^{n+\tilde{n}}$ -probability $\geq 1 - 34e^{-\tau}$, we have

$$\left\| T \left(\left[\bar{f}_\lambda \right] - \left[\hat{f}_\lambda \right] \right) \right\|_{L^2(P_{ZO})} \lesssim \tau n^{-\frac{\frac{s_x}{d_x} + \eta_1}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}}.$$

Proof. From Proposition D.3, with $P^{n+\tilde{n}}$ -probability $\geq 1 - 28e^{-\tau}$, we have

$$\begin{aligned} & \left\| T \left([\bar{f}_\lambda] - [\hat{f}_\lambda] \right) \right\|_{L^2(P_{ZO})} \\ & \lesssim \underbrace{\tau \lambda^{-\frac{1}{2}} \left(\frac{\|\hat{F}_\xi - F_*\|_{\mathcal{G}}}{\sqrt{n}} + \|F_* - [\hat{F}_\xi]\|_{L^2(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_{X, \gamma_x})} \right)}_{(\ddagger)} \left(1 + \|\bar{f}_\lambda\|_{\mathcal{H}_{X, \gamma_x, \gamma_0}} \right). \end{aligned}$$

By Proposition D.6, we have with P^n -probability $\geq 1 - 2e^{-\tau}$

$$\|\bar{f}_\lambda\|_{\mathcal{H}_{\gamma_x, \gamma_o}} \lesssim \tau n^{\frac{1}{2} \frac{1 + \frac{d_o}{s_o} (\frac{s_x}{d_x} + \eta_1)}{1 + (2 + \frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}}.$$

Under the probabilistic event that Proposition D.3 holds, the bounds in Proposition D.4 also hold, so we have

$$\left\| [\hat{F}_\xi] - F_* \right\|_{L^2(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_{X, \gamma_x})} \leq J \tau \tilde{n}^{-\frac{1}{2} \frac{m^\dagger}{m^\dagger + d^\dagger / 2 + \zeta}}, \quad \left\| \hat{F}_\xi - F_* \right\|_{\mathcal{G}} \leq J \tau \tilde{n}^{-\frac{1}{2} \frac{m^\dagger - 1}{m^\dagger + d^\dagger / 2 + \zeta}},$$

hold for any $\zeta > 0$. J is some constant independent of \tilde{n}, n . Thus under this probabilistic event, a sufficient condition so that $(\ddagger) \lesssim \frac{1}{n}$ is given by

$$\tilde{n} \geq (J \tau n)^{2 \frac{m^\dagger + d^\dagger / 2 + \zeta}{m^\dagger}} \vee (J^2 \tau^2 n)^{\frac{m^\dagger + d^\dagger / 2 + \zeta}{m^\dagger - 1}}.$$

This is satisfied since \tilde{n} satisfies Eq. (21). Hence,

$$\left\| [\hat{F}_\xi] - F_* \right\|_{L^2(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_{X, \gamma_x})} \leq n^{-1}, \quad \left\| \hat{F}_\xi - F_* \right\|_{\mathcal{G}} \leq n^{-\frac{1}{2}}. \quad (76)$$

By the union bound, we have that with $P^{n+\tilde{n}}$ -probability $\geq 1 - 30e^{-\tau}$, the following bound holds

$$\left\| T \left([\bar{f}_\lambda] - [\hat{f}_\lambda] \right) \right\|_{L^2(P_{ZO})} \lesssim \tau n^{-\frac{\frac{s_x}{d_x} + \eta_1}{1 + (2 + \frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}}.$$

Hence the proof is concluded. \square

Proposition D.2 (Projected stage-II error). *Suppose Assumptions 4.3, 4.2, 4.1, 2.3, 2.2 hold. Let $\lambda \asymp n^{-1}$ and*

$$\gamma_x = n^{-\frac{\frac{1}{d_x}}{1 + (2 + \frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}}, \quad \gamma_o = n^{-\frac{\frac{1}{s_o}(\frac{s_x}{d_x} + \eta_1)}{1 + (2 + \frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}}.$$

Then, with P^n -probability $\geq 1 - 4e^{-\tau}$, for sufficiently large $n \geq 1$, we have

$$\left\| T \left(f_* - [\bar{f}_\lambda] \right) \right\|_{L^2(P_{ZO})} \leq 2C\tau(\log n)^{\frac{d_x + d_o + 1}{2}} n^{-\frac{\frac{s_x}{d_x} + \eta_1}{1 + (2 + \frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}}$$

for some constant $C > 0$ independent of n .

Proof. The proposition is proved through the following triangular inequality

$$\begin{aligned} \|T(f_* - [\bar{f}_\lambda])\|_{L^2(P_{ZO})} &\leq \|T(f_* - [f_\lambda])\|_{L^2(P_{ZO})} + \|T([\bar{f}_\lambda] - [f_\lambda])\|_{L^2(P_{ZO})} \\ &\leq n^{-\frac{\frac{s_x}{d_x} + \eta_1}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} + n^{-\frac{\frac{s_x}{d_x} + \eta_1}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} (\log n)^{\frac{d_x+d_o+1}{2}} \\ &\leq 2n^{-\frac{\frac{s_x}{d_x} + \eta_1}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} (\log n)^{\frac{d_x+d_o+1}{2}}. \end{aligned}$$

The second last inequality holds by using Proposition D.7 for the first term and Proposition D.8 for the second term, which holds with P^n -probability $\geq 1 - 4e^{-\tau}$. \square

D.1 Projected Stage I Error

With the introduction of \mathcal{H}_{FO} and the partial isometry $V : \mathcal{H}_{XO} \rightarrow \mathcal{H}_{FO}$ in Section C, we define

$$\begin{aligned} \hat{h}_\lambda &:= \arg \min_{h \in \mathcal{H}_{FO}} \frac{1}{n} \sum_{i=1}^n \left(y_i - \left\langle h, V \left(\hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) \otimes \phi_{\gamma_o}(\mathbf{o}_i) \right) \right\rangle_{\mathcal{H}_{FO}} \right)^2 + \lambda \|h\|_{\mathcal{H}_{FO}}^2 \\ \bar{h}_\lambda &= \arg \min_{h \in \mathcal{H}_{FO}} \frac{1}{n} \sum_{i=1}^n \left(y_i - \left\langle h, V \left(F_*(\mathbf{z}_i, \mathbf{o}_i) \otimes \phi_{\gamma_o}(\mathbf{o}_i) \right) \right\rangle_{\mathcal{H}_{FO}} \right)^2 + \lambda \|h\|_{\mathcal{H}_{FO}}^2. \end{aligned}$$

As shown in Section C, we have $\hat{h}_\lambda = V \hat{f}_\lambda$ and $\bar{h}_\lambda = V \bar{f}_\lambda$, where $V : \mathcal{H}_{XO} \rightarrow \mathcal{H}_{FO}$ is the metric surjection map introduced in Section C. $\hat{f}_\lambda \in \mathcal{H}_{\gamma_x, \gamma_o}$ is defined in Eq. (7) and $\bar{f}_\lambda \in \mathcal{H}_{\gamma_x, \gamma_o}$ is defined in Eq. (23) in the main text.

We further define

$$\begin{aligned} \Phi_{\hat{F}O} : \mathcal{H}_{FO} &\rightarrow \mathbb{R}^n = \left[V \left(\hat{F}_\xi(\mathbf{z}_1, \mathbf{o}_1) \otimes \phi_{\gamma_o}(\mathbf{o}_1) \right), \dots, V \left(\hat{F}_\xi(\mathbf{z}_n, \mathbf{o}_n) \otimes \phi_{\gamma_o}(\mathbf{o}_n) \right) \right]^* \\ \Phi_{FO} : \mathcal{H}_{FO} &\rightarrow \mathbb{R}^n = [V(F_*(\mathbf{z}_1, \mathbf{o}_1) \otimes \phi_{\gamma_o}(\mathbf{o}_1)), \dots, V(F_*(\mathbf{z}_n, \mathbf{o}_n) \otimes \phi_{\gamma_o}(\mathbf{o}_n))]^* \\ \mathbf{Y} &\in \mathbb{R}^n, \quad \mathbf{Y} = [y_1, \dots, y_n]^\top, \end{aligned} \tag{77}$$

and the following operators on \mathcal{H}_{FO}

$$\begin{aligned} C_{\hat{F}O} &:= \mathbb{E} \left[V \left(\hat{F}_\xi(Z, O) \otimes \phi_{\gamma_o}(O) \right) \otimes V \left(\hat{F}_\xi(Z, O) \otimes \phi_{\gamma_o}(O) \right) \right], \quad \hat{C}_{\hat{F}O} := \frac{1}{n} \Phi_{\hat{F}O}^* \Phi_{\hat{F}O} \\ C_{FO} &:= \mathbb{E} [V(F_*(Z, O) \otimes \phi_{\gamma_o}(O)) \otimes V(F_*(Z, O) \otimes \phi_{\gamma_o}(O))], \quad \hat{C}_{FO} := \frac{1}{n} \Phi_{FO}^* \Phi_{FO}. \end{aligned}$$

Hence, we have closed form expression for

$$\hat{h}_\lambda = \frac{1}{n} \left(\hat{C}_{\hat{F}O} + \lambda \right)^{-1} \Phi_{\hat{F}O}^* \mathbf{Y}, \quad \bar{h}_\lambda = \frac{1}{n} \left(\hat{C}_{FO} + \lambda \right)^{-1} \Phi_{FO}^* \mathbf{Y}. \tag{78}$$

Proposition D.3. Fix $\tau \geq 1$. Suppose that the assumptions of Proposition D.4 hold. Suppose that $n > A_{\lambda, \tau}$ with $A_{\lambda, \tau}$ defined in Eq. (87), and $\tilde{n} \geq 1$ satisfies Eq. (21) in the main text. We have with $P^{n+\tilde{n}}$ -probability $\geq 1 - 28e^{-\tau}$, the following inequality holds

$$\begin{aligned} &\|T[\hat{f}_\lambda] - T[\bar{f}_\lambda]\|_{L^2(P_{ZO})} \\ &\lesssim \tau \lambda^{-\frac{1}{2}} \left(\frac{\|\hat{F}_\xi - F_*\|_{\mathcal{G}}}{\sqrt{n}} + \left\| [F_* - \hat{F}_\xi] \right\|_{L^2(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_{X, \gamma_x})} \right) (1 + \|\bar{f}_\lambda\|_{\mathcal{H}_{\gamma_x, \gamma_o}}). \end{aligned}$$

Under the same probabilistic event, the bounds in Proposition D.4 also hold.

Proof. We find

$$\left\| T[\hat{f}_\lambda] - T[\bar{f}_\lambda] \right\|_{L^2(P_{ZO})} = \left\| [\hat{h}_\lambda] - [\bar{h}_\lambda] \right\|_{L^2(P_{ZO})} \leq \left\| C_{FO}^{\frac{1}{2}} (\hat{h}_\lambda - \bar{h}_\lambda) \right\|_{\mathcal{H}_{FO}}.$$

The last step follows from Lemma 12 of Fischer and Steinwart [2020]. We can proceed by plugging in the closed form expressions from Eq. (78) to have

$$\begin{aligned} & \left\| C_{FO}^{\frac{1}{2}} (\hat{h}_\lambda - \bar{h}_\lambda) \right\|_{\mathcal{H}_{FO}} \\ &= \left\| C_{FO}^{\frac{1}{2}} \left((\hat{C}_{FO} + \lambda)^{-1} \frac{1}{n} \Phi_{FO}^* \mathbf{Y} - (\hat{C}_{FO} + \lambda)^{-1} \frac{1}{n} \Phi_{FO}^* \mathbf{Y} \right) \right\|_{\mathcal{H}_{FO}} \\ &\leq \left\| C_{FO}^{\frac{1}{2}} (C_{FO} + \lambda)^{-\frac{1}{2}} \right\| \cdot \left\| (C_{FO} + \lambda)^{\frac{1}{2}} \left((\hat{C}_{FO} + \lambda)^{-1} \frac{1}{n} \Phi_{FO}^* \mathbf{Y} - (\hat{C}_{FO} + \lambda)^{-1} \frac{1}{n} \Phi_{FO}^* \mathbf{Y} \right) \right\|_{\mathcal{H}_{FO}} \\ &\leq \left\| (C_{FO} + \lambda)^{\frac{1}{2}} \left((\hat{C}_{FO} + \lambda)^{-1} \frac{1}{n} \Phi_{FO}^* \mathbf{Y} - (\hat{C}_{FO} + \lambda)^{-1} \frac{1}{n} \Phi_{FO}^* \mathbf{Y} \right) \right\|_{\mathcal{H}_{FO}} \\ &\leq S_{-1} + S_0, \end{aligned}$$

where we define

$$\begin{aligned} S_{-1} &:= \left\| (C_{FO} + \lambda)^{\frac{1}{2}} \left((\hat{C}_{FO} + \lambda)^{-1} \frac{1}{n} (\Phi_{FO} - \Phi_{FO})^* \mathbf{Y} \right) \right\|_{\mathcal{H}_{FO}} \\ S_0 &:= \left\| (C_{FO} + \lambda)^{\frac{1}{2}} \left((\hat{C}_{FO} + \lambda)^{-1} (\hat{C}_{FO} - \hat{C}_{FO}) (\hat{C}_{FO} + \lambda)^{-1} \frac{1}{n} \Phi_{FO}^* \mathbf{Y} \right) \right\|_{\mathcal{H}_{FO}} \end{aligned}$$

We bound S_{-1} with $P^{n+\tilde{n}}$ -high probability by Lemma D.3 and S_0 in $P^{n+\tilde{n}}$ -high probability by Lemma D.2. We thus have, with $P^{n+\tilde{n}}$ -probability $\geq 1 - 28e^{-\tau}$, the following bound

$$\begin{aligned} & \left\| C_{FO}^{\frac{1}{2}} (\hat{h}_\lambda - \bar{h}_\lambda) \right\|_{\mathcal{H}_{FO}} \\ &\leq c\tau\sqrt{n} \left(\frac{\|\hat{F}_\xi - F_*\|_{\mathcal{G}}}{\sqrt{n}} + \left\| [\hat{F}_\xi - F_*] \right\|_{L^2(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_{X, \gamma_x})} \right) (\|\bar{h}_\lambda\|_{\mathcal{H}_{FO}} + 1). \end{aligned}$$

$\|\bar{h}_\lambda\|_{\mathcal{H}_{FO}} = \|\bar{f}_\lambda\|_{\mathcal{H}_{\gamma_x, \gamma_o}}$ since V is a metric surjection. \square

Proposition D.4 (CME rate). *Suppose Assumption 2.2 in the main text holds. Suppose $k_{\mathcal{O}} : \mathcal{O} \times \mathcal{O} \rightarrow \mathbb{R}$ and $k_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ are Matérn reproducing kernels whose RKHSs $\mathcal{H}_{\mathcal{O}}$ and $\mathcal{H}_{\mathcal{Z}}$ are norm equivalent to $W_2^{t_o}(\mathcal{O})$ and $W_2^{t_z}(\mathcal{Z})$ with $m_o > t_o > d_o/2$, $m_z > t_z > d_z/2$ and $\frac{m_z}{t_z} \wedge \frac{m_o}{t_o} \leq 2$. Define $m^\dagger = (m_z t_z^{-1}) \wedge (m_o t_o^{-1})$ and $d^\dagger = (d_z t_z^{-1}) \vee (d_o t_o^{-1})$. For $\zeta > 0$ arbitrarily small, we take $\xi = \tilde{n}^{-\frac{1}{m^\dagger + d^\dagger / 2 + \zeta}}$, then with $P^{\tilde{n}}$ -probability at least $1 - 4e^{-\tau}$, we have*

$$\left\| [\hat{F}_\xi - F_*] \right\|_{L^2(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_{X, \gamma_x})} \leq J\tau\tilde{n}^{-\frac{1}{2} \frac{m^\dagger}{m^\dagger + d^\dagger / 2 + \zeta}}, \quad \left\| \hat{F}_\xi - F_* \right\|_{\mathcal{G}} \leq J\tau\tilde{n}^{-\frac{1}{2} \frac{m^\dagger - 1}{m^\dagger + d^\dagger / 2 + \zeta}}.$$

J is a constant independent of \tilde{n} .

Proof. We are going to apply Li et al. [2024a, Theorem 3] to obtain the desired result. To this end, we need to verify the assumptions (EVD) and (SRC) made in Li et al. [2024a]. Note that we let the assumption (EMB) of Li et al. [2024a] be trivially verified with $\alpha = 1$, as we prove in

[Proposition D.5](#) that $F_* \in [\mathcal{G}]^\beta$ with $\beta > 1$, therefore $\beta + p > 1 + p > 1 = \alpha$ falls in Case 2 of [Theorem 3](#) of [Li et al. \[2024a\]](#).

Verification of (EVD) Let ϕ_Z (resp. ϕ_O) denote the feature map of \mathcal{H}_Z (resp. \mathcal{H}_O). Define $\bar{C}_{ZO} : \mathcal{H}_Z \otimes \mathcal{H}_O \rightarrow \mathcal{H}_Z \otimes \mathcal{H}_O$ as the covariance operator.

$$C_{ZO} = \iint_{\mathcal{Z} \times \mathcal{O}} (\phi_Z(\mathbf{z}) \otimes \phi_O(\mathbf{o})) \otimes (\phi_Z(\mathbf{z}) \otimes \phi_O(\mathbf{o})) p_{ZO}(\mathbf{z}, \mathbf{o}) d\mathbf{z} d\mathbf{o}.$$

We also define another two auxiliary covariance operators $\bar{C}_Z : \mathcal{H}_Z \rightarrow \mathcal{H}_Z$ and $\bar{C}_O : \mathcal{H}_O \rightarrow \mathcal{H}_O$.

$$\bar{C}_Z = \int_{\mathcal{Z}} \phi_Z(\mathbf{z}) \otimes \phi_Z(\mathbf{z}) p_Z(\mathbf{z}) d\mathbf{z}, \quad \bar{C}_O = \int_{\mathcal{O}} \phi_O(\mathbf{o}) \otimes \phi_O(\mathbf{o}) p_O(\mathbf{o}) d\mathbf{o}.$$

Since k_Z and k_O are bounded, $C_{ZO}, \bar{C}_Z, \bar{C}_O$ are all self-adjoint compact operators. From [Edmunds and Triebel \[1996\]](#) and [Fischer and Steinwart \[2020\]\[Section 4\]](#), we have $\lambda_i(\bar{C}_Z) \asymp i^{-2t_z/d_z}$ and $\lambda_i(\bar{C}_O) \asymp i^{-2t_o/d_o}$. We know from [Assumption 2.2](#) that P_{ZO} is equivalent to the Lebesgue measure, so by Jensen's inequality, we have $C_{ZO} \leq \bar{C}_Z \otimes \bar{C}_O$, where \otimes here denotes an operator tensor product. Hence, from [Lemma 17](#) of [Meunier et al. \[2024a\]](#) we have $\lambda_i(C_{ZO}) \leq \lambda_i(\bar{C}_Z \otimes \bar{C}_O)$. Finally, from [Lemma D.4](#) we have

$$\lambda_i(C_{ZO}) \leq \lambda_i(\bar{C}_Z \otimes \bar{C}_O) \lesssim i^{-2t_z/d_z \wedge 2t_o/d_o + \zeta}$$

for any $\zeta > 0$. Therefore, we have proved that (EVD) hold with $1/p = 2t_z/d_z \wedge 2t_o/d_o - \zeta$. Hence, $p = d^\dagger/2 + \zeta$ for any $\zeta > 0$.

Verification of (SRC) Let $\beta = m^\dagger$. By the definition of vector-valued interpolation space in [Li et al. \[2024a\]\[Definition 2\]](#) and the Assumption that P_{ZO} is equivalent to the Lebesgue measure, we have that

$$\begin{aligned} [\mathcal{G}]^\beta &\cong \mathcal{H}_{X,\gamma_x} \otimes [L^2(\mathcal{Z} \times \mathcal{O}), \mathcal{H}_Z \otimes \mathcal{H}_O]_{\beta,2} \\ &\stackrel{(a)}{\cong} \mathcal{H}_{X,\gamma_x} \otimes [L^2(\mathcal{Z} \times \mathcal{O}), W_2^{t_z}(\mathcal{Z}) \otimes W_2^{t_o}(\mathcal{O})]_{\beta,2} \\ &\stackrel{(b)}{\cong} \mathcal{H}_{X,\gamma_x} \otimes W_2^{t_z\beta}(\mathcal{Z}) \otimes W_2^{t_o\beta}(\mathcal{O}) \\ &\stackrel{(*)}{\supseteq} \mathcal{H}_{X,\gamma_x} \otimes W_2^{m_z}(\mathcal{Z}) \otimes W_2^{m_o}(\mathcal{O}) \\ &\stackrel{(c)}{\cong} \mathcal{H}_{X,\gamma_x} \otimes MW_2^{m_z, m_o}(\mathcal{Z} \times \mathcal{O}; \mathbb{R}) \\ &\stackrel{(d)}{\cong} MW_2^{m_z, m_o}(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_{X,\gamma_x}). \end{aligned}$$

where in (a) we use Corollary 10.13 and Theorem 10.46 of [Wendland \[2004\]](#), in (b) we use [Lemma C.7](#), in (c) we use proposition 3.1 of [Sickel and Ullrich \[2009\]](#), which shows $MW_2^{m_z, m_o}(\mathcal{Z} \times \mathcal{O}) \cong W_2^{m_z}(\mathcal{Z}) \otimes W_2^{m_o}(\mathcal{O})$, and in (d) we use [Lemma D.1](#). Furthermore, the inclusion in (*) is a continuous embedding. We show in [Proposition D.5](#) that under [Assumption 2.2](#), $\|F_*\|_{MW_2^{m_z, m_o}(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_{X,\gamma_x})}$ is bounded by a constant. Hence we deduce from the above embedding that

$$F_* \in [\mathcal{G}]^\beta, \quad \beta = m^\dagger = \frac{m_z}{t_z} \wedge \frac{m_o}{t_o}. \quad (79)$$

Hence $\|F_*\|_{[\mathcal{G}]^\beta} \leq C$ for some constant $C > 0$. Hence F_* satisfies (SRC) in [Li et al. \[2024a\]](#) with $\beta = m^\dagger$.

Now we use Case 2 in Theorem 3 of Li et al. [2024a] to obtain the following result: if we take $\xi = \tilde{n}^{-\frac{1}{m^\dagger + d^\dagger / 2 + \zeta}}$, then for any $0 \leq \gamma < m^\dagger$, the following bound on the γ -norm

$$\left\| [\hat{F}_\xi - F_*] \right\|_\gamma^2 \leq \tau^2 \tilde{n}^{-\frac{m^\dagger - \gamma}{m^\dagger + d^\dagger / 2 + \zeta}}.$$

holds with $P^{\tilde{n}}$ -probability at least $1 - 4e^{-\tau}$. Therefore, noting that $m^\dagger > 1$, by taking $\gamma = 0$ and $\gamma = 1$, we obtain with $P^{\tilde{n}}$ -probability at least $1 - 4e^{-\tau}$,

$$\left\| [\hat{F}_\xi - F_*] \right\|_{L^2(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_{X, \gamma_x})} \leq \tau \tilde{n}^{-\frac{1}{2} \frac{m^\dagger}{m^\dagger + d^\dagger / 2 + \zeta}}, \quad \left\| \hat{F}_\xi - F_* \right\|_{\mathcal{G}} \leq \tau \tilde{n}^{-\frac{1}{2} \frac{m^\dagger - 1}{m^\dagger + d^\dagger / 2 + \zeta}}.$$

□

Proposition D.5. Suppose that Assumption 2.2 in the main text is satisfied. Then,

$$\|F_*\|_{MW_2^{m_z, m_o}(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_{X, \gamma_x})} \leq \sqrt{\binom{m_z + d_z - 1}{d_z - 1}} \sqrt{\binom{m_o + d_o - 1}{d_o - 1}} \rho.$$

Proof. Notice that, by definition of F_* ,

$$\begin{aligned} \|F_*\|_{MW_2^{m_z, m_o}(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_{X, \gamma_x})} &= \left\| \int_{\mathcal{X}} \phi_{\gamma_x}(\mathbf{x}) p(\mathbf{x} | \cdot, \cdot) \, d\mathbf{x} \right\|_{MW_2^{m_z, m_o}(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_{X, \gamma_x})} \\ &\leq \int_{\mathcal{X}} \|\phi_{\gamma_x}(\mathbf{x}) p(\mathbf{x} | \cdot, \cdot)\|_{MW_2^{m_z, m_o}(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_{X, \gamma_x})} \, d\mathbf{x}. \end{aligned}$$

In the last inequality above, we use Hytönen et al. [2016][Proposition 1.2.11]. Next, notice that

$$\begin{aligned} &\|\phi_{\gamma_x}(\mathbf{x}) p(\mathbf{x} | \cdot, \cdot)\|_{MW_2^{m_z, m_o}(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_{X, \gamma_x})}^2 \\ &= \sum_{|\alpha| \leq m_z} \sum_{|\beta| \leq m_o} \int_{\mathcal{Z} \times \mathcal{O}} \left\| \partial_{\mathbf{z}}^\alpha \partial_{\mathbf{o}}^\beta (\phi_{\gamma_x}(\mathbf{x}) p(\mathbf{x} | \mathbf{z}, \mathbf{o})) \right\|_{\mathcal{H}_{X, \gamma_x}}^2 \, d\mathbf{z} \, d\mathbf{o} \\ &= \sum_{|\alpha| \leq m_z} \sum_{|\beta| \leq m_o} \int_{\mathcal{Z} \times \mathcal{O}} \left\| \phi_{\gamma_x}(\mathbf{x}) \partial_{\mathbf{z}}^\alpha \partial_{\mathbf{o}}^\beta p(\mathbf{x} | \mathbf{z}, \mathbf{o}) \right\|_{\mathcal{H}_{X, \gamma_x}}^2 \, d\mathbf{z} \, d\mathbf{o} \\ &\leq \sum_{|\alpha| \leq m_z} \sum_{|\beta| \leq m_o} \int_{\mathcal{Z} \times \mathcal{O}} \left| \partial_{\mathbf{z}}^\alpha \partial_{\mathbf{o}}^\beta p(\mathbf{x} | \mathbf{z}, \mathbf{o}) \right|^2 \, d\mathbf{z} \, d\mathbf{o} \\ &\leq \binom{m_z + d_z - 1}{d_z - 1} \binom{m_o + d_o - 1}{d_o - 1} \rho^2. \end{aligned}$$

In the last step, $\binom{m_z + d_z - 1}{d_z - 1}$ shows up as the evaluation of $\sum_{|\alpha| \leq m_z} 1$. □

D.1.1 RKHS norm of \bar{f}_λ

We now provide a refined control of the RKHS norm of \bar{f}_λ by invoking an oracle inequality for the RKHS \mathcal{H}_{FO} . We remind the reader that $V\bar{f}_\lambda \in \mathcal{H}_{FO}$, i.e. the image of \bar{f}_λ under the metric surjection V , is the solution to a KRR problem with respect to the RKHS \mathcal{H}_{FO} (see Eq. (59)). In Section C.1 we characterized the rate of decay of the entropy numbers of \mathcal{H}_{FO} . The only remaining technical hurdle is that the noise $v = Y - (Tf_*)(Z, O)$ is unbounded; however, under Assumption 4.5 it is σ -subgaussian. Concretely, we introduce a logarithmically growing sequence

of clipping values M_n , which facilitates a key step in the derivation of the oracle inequality and ensures that the conclusions of [Steinwart and Christmann \[2008, Theorem 7.23\]](#) are applicable.

Firstly, by [Vershynin \[2018, Eq. \(2.14\)\]](#), $P(|v| \geq t) \leq 2 \exp(-\frac{ct^2}{\sigma^2})$ for some universal constant c . Define $v_i = y_i - (Tf_*)(\mathbf{z}_i, \mathbf{o}_i)$ for stage II samples $\{\mathbf{z}_i, \mathbf{o}_i, y_i\}_{i=1}^n$. Hence

$$\begin{aligned} & P^n \left(\{\mathbf{z}_i, \mathbf{o}_i, y_i\}_{i=1}^n \in (\mathcal{Z} \times \mathcal{O} \times Y)^n : \max_i |v_i| \leq t \right) \\ & \geq 1 - \sum_{i=1}^n P(|v_i| \geq t) \\ & \geq 1 - 2 \exp \left(\ln n - \frac{ct^2}{\sigma^2} \right). \end{aligned}$$

Hence, with P^n -probability $\geq 1 - 2 \exp(-\hat{\rho})$, we have

$$\max_{1 \leq i \leq n} |v_i| \leq \sigma \sqrt{\frac{\ln n + \hat{\rho}}{c}}.$$

By [Assumption 2.3](#), we have $\|Tf_*\|_\infty \leq \|f_*\|_\infty \leq 1$. For $n \geq 1$, we define

$$M_n = 1 + \sigma \sqrt{\frac{\ln n + \hat{\rho}}{c}}.$$

Hence $y_i \in [-M_n, M_n]$ for all $1 \leq i \leq n$ with P^n -probability $\geq 1 - 2 \exp(-\hat{\rho})$.

Secondly, we verify the assumptions of [Steinwart and Christmann \[2008, Theorem 7.23\]](#). By [Steinwart and Christmann \[2008, Example 7.3\]](#), the *supremum bound* is satisfied for $B = 4M_n^2$ and the *variance bound* is satisfied for $V = 16M_n^2$ and $\nu = 1$. Define $D_{ZO} = \{\mathbf{z}_i, \mathbf{o}_i\}_{i=1}^n$ and $L^2(D_{ZO})$ as the L^2 space with respect to the empirical data measure D_{ZO} . We are about to show that $(\forall n \geq 1)(\exists p \in (0, 1))(\exists a \geq B = 4M_n^2)$ such that

$$(\forall i \geq 1) \mathbb{E}_{D_{ZO} \sim P_{ZO}^n} e_i(\text{id} : \mathcal{H}_{FO} \rightarrow L^2(D_{ZO})) \leq ai^{-\frac{1}{2p}}. \quad (80)$$

To this end, we invoke [Steinwart and Christmann \[2008, Corollary 7.31\]](#), [Lemma C.1](#) and [Lemma C.2](#). We conclude that, there exists a constant $c_p > 0$ only depending on p , such that $(\forall i \geq 1)(\forall n \geq 1)$, we have

$$\begin{aligned} & \mathbb{E}_{D_{ZO} \sim P_{ZO}^n} e_i(\text{id} : \mathcal{H}_{FO} \hookrightarrow L^2(D_{ZO})) \\ & \leq c_p (3C)^{\frac{1}{2p}} \left(\frac{d_x + d_o + 1}{2ep} \right)^{\frac{d_x + d_o + 1}{2p}} \left(\gamma_x^{d_x} \gamma_o^{d_o} \right)^{-\frac{1}{2p}} (\min\{i, n\})^{\frac{1}{2p}} i^{-\frac{1}{p}} \\ & \leq c_p (3C)^{\frac{1}{2p}} \left(\frac{d_x + d_o + 1}{2ep} \right)^{\frac{d_x + d_o + 1}{2p}} \left(\gamma_x^{d_x} \gamma_o^{d_o} \right)^{-\frac{1}{2p}} i^{-\frac{1}{2p}}. \end{aligned}$$

We define a new constant $\tilde{c}_p := c_p (3C)^{\frac{1}{2p}} \left(\frac{d_x + d_o + 1}{2ep} \right)^{\frac{d_x + d_o + 1}{2p}}$. Since for all $n \geq 1$, $(\gamma_x^{d_x} \gamma_o^{d_o})^{-\frac{1}{2p}} \geq 1$, we set

$$a = \max\{4M_n^2, \tilde{c}_p\} \left(\gamma_x^{d_x} \gamma_o^{d_o} \right)^{-\frac{1}{2p}}$$

in Eq. (80), which satisfies $a \geq B = 4M_n^2$ for all $n \geq 1$.

We thus apply Steinwart and Christmann [2008, Theorem 7.23] restricted to the probabilistic event where $y_i \in [-2M_n, 2M_n]$ for all $1 \leq i \leq n$. We find, with P^n -probability $\geq 1 - 5 \exp(-\hat{\rho})$,

$$\begin{aligned} \lambda \|\bar{f}_\lambda\|_{\mathcal{H}_{XO}}^2 &\leq 9 \left(\lambda \|f_0\|_{\mathcal{H}_{XO}}^2 + \|T(f_0 - f_*)\|_{L^2(P_{ZO})}^2 \right) \\ &+ K \left(\frac{\max\{4M_n^2, \tilde{c}_p\}^{2p} \gamma_x^{-d_x} \gamma_o^{-d_o}}{\lambda^p n} \right) + \frac{216M_n^2 \hat{\rho}}{n} + \frac{15B_0 \hat{\rho}}{n} \end{aligned}$$

where $K \geq 1$ is a constant depending on $p, M_n, f_0 \in \mathcal{H}_{\gamma_x, \gamma_o}$. $B_0 \geq B$ is a constant that satisfies

$$\|L \circ (Tf_0)\|_\infty := \sup_{(\mathbf{z}, \mathbf{o}, y) \in \mathcal{Z} \times \mathcal{O} \times [-2M_n, 2M_n]} (y - (Tf_0)(\mathbf{z}, \mathbf{o}))^2 \leq B_0,$$

By checking the dependence of K on p, M_n in the proof of Steinwart and Christmann [2008, Theorem 7.23], we find that $K \leq c(p)M_n^2$ for some constant $c(p)$ depending only on p . We choose $f_0 = f_\lambda$, where $f_\lambda \in \mathcal{H}_{\gamma_x, \gamma_o}$ is defined in Eq. (57). We have, since $Y = [-2M_n, 2M_n]$,

$$\begin{aligned} \|L \circ (Tf_\lambda)\|_\infty &= \sup_{(\mathbf{z}, \mathbf{o}, y) \in \mathcal{Z} \times \mathcal{O} \times [-2M_n, 2M_n]} (y - (Tf_\lambda)(\mathbf{z}, \mathbf{o}))^2 \leq (2M_n + \|Tf_\lambda\|_\infty)^2 \\ &\leq (2M_n + \|f_\lambda\|_\infty)^2 \leq (2M_n + \|f_\lambda\|_{\mathcal{H}_{\gamma_x, \gamma_o}})^2. \end{aligned}$$

We choose thus

$$B_0 = (2M_n + \|f_\lambda\|_{\mathcal{H}_{\gamma_x, \gamma_o}})^2 \geq 4M_n^2 = B.$$

Hence we deduce from the above high probability inequality that

$$\begin{aligned} \lambda \|\bar{f}_\lambda\|_{\mathcal{H}_{XO}}^2 &\leq C \left(\lambda \|f_\lambda\|_{\mathcal{H}_{XO}}^2 + \|T(f_\lambda - f_*)\|_{L^2(P_{ZO})}^2 \right) \\ &+ C \left(M_n^{2+2p} \left(\frac{\max\{4, \tilde{c}_p\}^{2p} \gamma_x^{-d_x} \gamma_o^{-d_o}}{\lambda^p n} \right) + \frac{(M_n + \|f_\lambda\|_{\mathcal{H}_{XO}})^2 \hat{\rho}}{n} \right). \end{aligned}$$

for some constant C that is independent of $\lambda, M_n, n, \gamma_x, \gamma_o, B, V, \hat{\rho}$. We now state the main Proposition in this subsection.

Proposition D.6. *Suppose Assumptions 2.2, 2.3, 4.1, 4.2 and 4.3 in the main text hold. Let $\lambda = n^{-1}$ and*

$$\gamma_x = n^{-\frac{1}{1+(2+\frac{d_O}{s_O})(\frac{s_x}{d_x}+\eta_1)}}, \quad \gamma_o = n^{-\frac{\frac{1}{s_O}(\frac{s_x}{d_x}+\eta_1)}{1+(2+\frac{d_O}{s_O})(\frac{s_x}{d_x}+\eta_1)}}.$$

Fix an arbitrary $p > 0$. Then with probability $\geq 1 - 5e^{-\tau}$, for sufficiently large $n \geq 1$, the following bound holds

$$\|\bar{f}_\lambda\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 \lesssim \tau n^{\frac{1+\frac{d_O}{s_O}(\frac{s_x}{d_x}+\eta_1)}{1+(2+\frac{d_O}{s_O})(\frac{s_x}{d_x}+\eta_1)}+p}.$$

Proof. Fix $p > 0$. We combine the results of Proposition D.7 with the inequality immediately preceding this proposition to find, for $\hat{\rho} \geq 1$, there exists a constant C_1, C_2, C_3 independent of $\lambda, M_n, n, \gamma_x, \gamma_o, B, V, \hat{\rho}$ such that, with probability $\geq 1 - 5e^{-\hat{\rho}}$ the following bound holds

$$\lambda \|\bar{f}_\lambda\|_{\mathcal{H}_{XO}}^2$$

$$\begin{aligned}
&\stackrel{(a)}{\leq} C_1 \hat{\rho} \left(\lambda \|f_\lambda\|_{\mathcal{H}_{XO}}^2 + \|T(f_\lambda - f_*)\|_{L^2(P_{ZO})}^2 + M_n^{2+2p} \left(\frac{\max\{4, \tilde{c}_p\}^{2p} \gamma_x^{-d_x} \gamma_o^{-d_o}}{\lambda^p n} \right) + \frac{M_n^2}{n} \right) \\
&\stackrel{(b)}{\leq} C_2 \hat{\rho} \left(n^{-\frac{2(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} + M_n^{2+2p} \left(\frac{\max\{4, \tilde{c}_p\}^{2p} \gamma_x^{-d_x} \gamma_o^{-d_o}}{\lambda^p n} \right) + \frac{M_n^2}{n} \right) \\
&\stackrel{(c)}{=} C_2 \hat{\rho} \left(n^{-\frac{2(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} \left(1 + M_n^{2+2p} \frac{\max\{4, \tilde{c}_p\}^{2p}}{\lambda^p} \right) + \frac{M_n^2}{n} \right) \\
&\stackrel{(d)}{\leq} C_3 \hat{\rho} n^{-\frac{2(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} n^{2p}.
\end{aligned}$$

In the above derivations, (a) holds by $\lambda = n^{-1}$, (b) holds by the conclusion of [Proposition D.7](#), (c) holds by the choice of γ_x, γ_o as functions of n , (d) holds for sufficiently large $n \geq 1$, since p is fixed and $n^p > M_n^{2+2p} = \left(1 + \sigma \sqrt{\frac{\ln n + \hat{\rho}}{c}}\right)^{2+2p}$ for sufficiently large $n \geq 1$, and $n^{-\frac{2(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} \geq \frac{M_n^2}{n}$ for sufficiently large $n \geq 1$. In particular, the constant C_3 depends on p . \square

D.2 Projected approximation error in Stage II

Proposition D.7. Suppose Assumptions 2.2, 2.3, 4.1, 4.2 and 4.3 in the main text hold. Let $\lambda = n^{-1}$ and

$$\gamma_x = n^{-\frac{\frac{1}{d_x}}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}}, \quad \gamma_o = n^{-\frac{\frac{1}{s_o}(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}}.$$

Then we have

$$\lambda \|f_\lambda\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 + \|T(f_* - [f_\lambda])\|_{L^2(P_{ZO})}^2 \leq C n^{-\frac{2(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}},$$

for some constant $C > 0$ independent of n .

Proof. Recall $f_{\text{aux}} \in \mathcal{H}_{\gamma_x, \gamma_o}$ defined in Eq. (40). We write

$$f_{\text{aux}} = f_* * K_{\gamma_o} - (f_* - f_{*, \text{low}}) * K_{\gamma_o} + (f_* - f_{*, \text{low}}) * K_{\gamma_o} * K_{\gamma_x}.$$

By definition of f_λ , we have

$$\begin{aligned}
&\lambda \|f_\lambda\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 + \|T(f_* - [f_\lambda])\|_{L^2(P_{ZO})}^2 \\
&= \inf_{f \in \mathcal{H}_{\gamma_x, \gamma_o}} \lambda \|f\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 + \|Tf_* - T[f]\|_{L^2(P_{ZO})}^2 \\
&\leq \lambda \|f_{\text{aux}}\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 + \|Tf_* - T[f_{\text{aux}}]\|_{L^2(P_{ZO})}^2 \\
&\leq \lambda \|f_{\text{aux}}\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 + \|Tf_* - T(f_* * K_{\gamma_o})\|_{L^2(P_{ZO})}^2 \\
&\quad + \|T((f_* - f_{*, \text{low}}) * K_{\gamma_o} - (f_* - f_{*, \text{low}}) * K_{\gamma_o} * K_{\gamma_x})\|_{L^2(P_{ZO})}^2. \tag{81}
\end{aligned}$$

For the first term, we know from Eq. (67) and [Assumption 2.3](#) in the main text that

$$\lambda \|f_{\text{aux}}\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 \leq n^{-1} \cdot 2n^{\frac{1+\frac{d_o}{s_o}(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} \|f_*\|_{L^2(\mathbb{R}^{d_x+d_o})} \leq 2n^{-\frac{2(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}}.$$

For the second term, we have

$$\begin{aligned}
\|Tf_* - T(f_* * K_{\gamma_o})\|_{L^2(P_{ZO})}^2 &\leq \|f_* - f_* * K_{\gamma_o}\|_{L^2(P_{XO})}^2 \\
&\lesssim \|f_* - f_* * K_{\gamma_o}\|_{L^2(\mathcal{X} \times \mathcal{O})}^2 \\
&\lesssim |f_*|_{B_{2,q}^{s_x,s_o}(\mathbb{R}^{d_x+d_o})}^2 \cdot \max\{0, \gamma_o^{2s_o}\} \\
&= |f_*|_{B_{2,q}^{s_x,s_o}(\mathbb{R}^{d_x+d_o})}^2 n^{-\frac{2(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}}.
\end{aligned}$$

The first inequality above holds because T is a bounded operator; the second inequality holds by Assumption 2.2 that P_{XO} admits a bounded density. The second last inequality above holds by using Lemma D.5 for $\gamma = (\underbrace{0, \dots, 0}_{d_x}, \underbrace{\gamma_o, \dots, \gamma_o}_{d_o})$.

For the third term, we note that ($\forall \mathbf{o} \in \mathbb{R}^{d_o}$),

$$\mathcal{F}[((f_* - f_{*,\text{low}}) * K_{\gamma_o} - (f_* - f_{*,\text{low}}) * K_{\gamma_o} * K_{\gamma_x})(\cdot, \mathbf{o})]$$

is supported on the complement of $\{\mathbf{x} : \|\mathbf{x}\| \leq \gamma_x^{-1}\}$ (see Eq. (38)). Thus it follows from Assumption 4.3 that

$$\begin{aligned}
&\|T((f_* - f_{*,\text{low}}) * K_{\gamma_o} - (f_* - f_{*,\text{low}}) * K_{\gamma_o} * K_{\gamma_x})\|_{L^2(P_{ZO})}^2 \\
&\leq \gamma_x^{2d_x\eta_1} \|(f_* - f_{*,\text{low}}) * K_{\gamma_o} - (f_* - f_{*,\text{low}}) * K_{\gamma_o} * K_{\gamma_x}\|_{L^2(P_{XO})}^2 \\
&\stackrel{(i)}{\lesssim} \gamma_x^{2d_x\eta_1+2s_x} \cdot |(f_* - f_{*,\text{low}}) * K_{\gamma_o}|_{B_{2,q}^{s_x,s_o}(\mathbb{R}^{d_x+d_o})}^2 \\
&\stackrel{(ii)}{\leq} \gamma_x^{2d_x\eta_1+2s_x} \cdot |f_* * K_{\gamma_o}|_{B_{2,q}^{s_x,s_o}(\mathbb{R}^{d_x+d_o})}^2 \\
&\stackrel{(iii)}{\leq} \gamma_x^{2d_x\eta_1+2s_x} \cdot |f_*|_{B_{2,q}^{s_x,s_o}(\mathbb{R}^{d_x+d_o})}^2 \|K_{\gamma_o}\|_{L^1(\mathbb{R}^{d_o})}^2 \\
&\stackrel{(iv)}{\lesssim} n^{-\frac{2(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}}.
\end{aligned}$$

In the above derivations, (i) follows by using Lemma D.5 for $\gamma = (\underbrace{\gamma_x, \dots, \gamma_x}_{d_x}, \underbrace{0, \dots, 0}_{d_o})$ and the

Assumption that P_{XO} admits a bounded density, (ii) follows from the proof of Lemma E.3, (iii) follows from Lemma E.2, and (iv) follows by the fact that $\|K_{\gamma_o}\|_{L^1(\mathbb{R})} = 1$ [Giné and Nickl, 2021, Section 4.1.2] and Assumption 2.3 in the main text.

Combine the upper bound on the three terms in Eq. (81) and we obtain

$$\lambda \|f_\lambda\|_{\mathcal{H}_{\gamma_x,\gamma_o}}^2 + \|T(f_* - [f_\lambda])\|_{L^2(P_{ZO})}^2 \leq n^{-\frac{2(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}}. \quad (82)$$

The proof concludes here. In addition to that, from Eq. (81), we also have

$$\|Tf_* - T[f_{\text{aux}}]\|_{L^2(P_{ZO})} \leq Cn^{-\frac{\frac{s_x}{d_x} + \eta_1}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}}. \quad (83)$$

Also, if we follow the same derivations as above, we obtain

$$\|f_* - [f_{\text{aux}}]\|_{L^2(P_{XO})} \leq Cn^{-\frac{\frac{s_x}{d_x}}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}}. \quad (84)$$

□

D.3 Projected estimation error in Stage II

Proposition D.8. Suppose Assumptions 2.2, 2.3, 4.1, 4.2 and 4.3 in the main text hold. Let $\lambda \asymp n^{-1}$ and

$$\gamma_x = n^{-\frac{1}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}}, \quad \gamma_o = n^{-\frac{\frac{1}{s_o}(\frac{s_x}{d_x}+\eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}}.$$

With P^n -probability $\geq 1 - 4e^{-\tau}$, for sufficiently large $n \geq 1$, we have

$$\|T([\bar{f}_\lambda] - [f_\lambda])\|_{L^2(P_{ZO})}^2 \leq C\tau^2 n^{-\frac{2(\frac{s_x}{d_x}+\eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}} (\log n)^{d_x+d_o+1},$$

for some constant C independent of n .

Proof. Recall the definition of \bar{h}_λ defined in Eq. (60) and h_λ defined in Eq. (59).

$$\begin{aligned} \bar{h}_\lambda &= \arg \min_{h \in \mathcal{H}_{FO}} \lambda \|h\|_{\mathcal{H}_{FO}}^2 + \frac{1}{n} \sum_{i=1}^n (h(\mathbf{z}_i, \mathbf{o}_i) - y_i)^2. \\ h_\lambda &= \arg \min_{h \in \mathcal{H}_{FO}} \lambda \|h\|_{\mathcal{H}_{FO}}^2 + \|h_* - h\|_{L^2(P_{ZO})}^2. \end{aligned}$$

Recall that we have proved that $\|T([\bar{f}_\lambda] - [f_\lambda])\|_{L^2(P_{ZO})} = \|[\bar{h}_\lambda] - [h_\lambda]\|_{L^2(P_{ZO})}$ in Eq. (61), so the proof of Proposition D.8 is translated to the estimation error of a standard kernel ridge regression with hypothesis space \mathcal{H}_{FO} and the target function $h_* := Tf_* \in L^2(P_{ZO})$. Next, we are going to apply existing results, mainly Theorem 16 of Fischer and Steinwart [2020], to our setting.

From Eq. (61) and Fischer and Steinwart [2020][Lemma 12], we have

$$\|[\bar{h}_\lambda] - [h_\lambda]\|_{L^2(P_{ZO})} \leq \left\| C_{FO}^{\frac{1}{2}} (\bar{h}_\lambda - h_\lambda) \right\|_{\mathcal{H}_{FO}} =: (*) \quad (85)$$

Next, we will upper bound $(*)$ using Theorem 16 from Fischer and Steinwart [2020]. To apply this result, we must verify the underlying assumptions and control the auxiliary quantities specified in Theorem 16 from Fischer and Steinwart [2020].

First, we are going to verify the (MOM) condition. By Assumption 4.5, we know that

$$\begin{aligned} \int_{\mathbb{R}} |y - (Tf_*)(\mathbf{z}, \mathbf{o})|^m p(y \mid \mathbf{z}, \mathbf{o}) dy &= \mathbb{E}[|v|^m \mid Z = \mathbf{z}, O = \mathbf{o}] \\ &\leq (\sqrt{2}C\sigma)^m (m/2)^{\frac{m}{2}} \leq \frac{1}{2}(2C\sigma)^m m! \end{aligned}$$

by Vershynin [2018, Eq. (2.15)], for some universal constant $C > 0$. Hence the (MOM) condition of Fischer and Steinwart [2020] is satisfied.

Next, we control the auxiliary quantities in Fischer and Steinwart [2020][Theorem 16]. Recall m_z, m_o as defined in Eq. (19). By definition, they satisfy

$$\frac{d_o}{2m_o} < \frac{1 + 2\left(\frac{s_x}{d_x} + \eta_1\right)}{1 + 2\left(\frac{s_x}{d_x} + \eta_1\right) + \frac{d_o}{s_o}\left(\frac{s_x}{d_x} + \eta_1\right)} < 1, \quad \frac{d_z}{2m_z} \leq \frac{d_o}{2m_o} < 1.$$

Rearranging, we deduce that

$$\frac{d_o}{2m_o} < \left(2m_o \frac{\frac{1}{s_o}(\frac{s_x}{d_x} + \eta_1)}{1 + (2 + \frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)} + 1 \right)^{-1}$$

$$\frac{d_o}{2m_o} < \frac{1 + (2 + \frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}{1 + 2(\frac{s_x}{d_x} + \eta_1) + \frac{2m_o + d_o}{s_o}(\frac{s_x}{d_x} + \eta_1)}.$$

Hence there exists $\theta \in (0, 1)$ such that the following inequalities hold simultaneously

$$\begin{aligned} \frac{d_o}{2m_o\theta} &< 1, \quad \frac{d_z}{2m_z\theta} < 1 \\ \theta &\leq \frac{1 + (2 + \frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}{1 + 2(\frac{s_x}{d_x} + \eta_1) + \frac{2m_o + d_o}{s_o}(\frac{s_x}{d_x} + \eta_1)} \\ \theta \left(2m_o \frac{\frac{1}{s_o}(\frac{s_x}{d_x} + \eta_1)}{1 + (2 + \frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)} + 1 \right) &< 1 \end{aligned} \tag{86}$$

Next, we define

$$\begin{aligned} g_\lambda &:= \log \left(2e\mathcal{N}_{FO}(\lambda) \frac{\|C_{FO}\| + \lambda}{\|C_{FO}\|} \right) \\ A_{\lambda,\tau} &:= 8 \left\| k_{FO}^\theta \right\|_\infty^2 \tau g_\lambda \lambda^{-\theta} \\ L_\lambda &:= \max \left\{ L, \|h_* - [h_\lambda]\|_{L^\infty(P_{ZO})} \right\} \end{aligned} \tag{87}$$

Controlling g_λ and $A_{\lambda,\tau}$: From Proposition C.1, we know that

$$\mathcal{N}_{FO}(\lambda) \lesssim (\log n)^{d_x + d_o + 1} \left(\gamma_x^{d_x} \gamma_o^{d_o} \right)^{-1} = (\log n)^{d_x + d_o + 1} n^{\frac{1 + \frac{d_o}{s_o}(\frac{s_x}{d_x} + \eta_1)}{1 + (2 + \frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}}.$$

From Lemma D.10, we know that

$$\|C_{FO}\| \geq a_f^{-1} \left(\frac{\sqrt{\pi}}{4} \right)^{\frac{d_x + d_o}{2}} n^{-\frac{1 + \frac{d_o}{s_o}(\frac{s_x}{d_x} + \eta_1)}{1 + (2 + \frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}}.$$

Therefore, since $\lambda \asymp n^{-1}$, we have

$$\begin{aligned} g_\lambda &= \log (2e\mathcal{N}_{FO}(\lambda) (1 + \lambda\|C_{FO}\|^{-1})) \\ &\lesssim \log \left((\log n)^{d_x + d_o + 1} n^{\frac{1 + \frac{d_o}{s_o}(\frac{s_x}{d_x} + \eta_1)}{1 + (2 + \frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} \left(1 + n^{-1} \cdot n^{\frac{1 + \frac{d_o}{s_o}(\frac{s_x}{d_x} + \eta_1)}{1 + (2 + \frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} \right) \right) \\ &\leq \log \left(2(\log n)^{d_x + d_o + 1} n^{\frac{1 + \frac{d_o}{s_o}(\frac{s_x}{d_x} + \eta_1)}{1 + (2 + \frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} \right) \\ &= (d_x + d_o + 1) \log(\log n) + \frac{1 + \frac{d_o}{s_o}(\frac{s_x}{d_x} + \eta_1)}{1 + (2 + \frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)} \log(n) \\ &\leq 2 \frac{1 + \frac{d_o}{s_o}(\frac{s_x}{d_x} + \eta_1)}{1 + (2 + \frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)} \log(n). \end{aligned}$$

The last step holds because $\log(n)$ dominates a constant term and $\log(\log n)$ for sufficiently large n . Since $\frac{d_o}{2m_o\theta} < 1$, $\frac{d_z}{2m_z\theta} < 1$, by Proposition C.2, we have

$$\left\| k_{FO}^\theta \right\|_\infty \lesssim \rho^\theta \gamma_o^{-\theta m_o}. \quad (88)$$

Therefore,

$$A_{\lambda,\tau} \lesssim \left\| k_{FO}^\theta \right\|_\infty^2 \tau \log(n) n^\theta \lesssim \gamma_o^{-2m_o\theta} \tau \log(n) n^\theta = \tau \log(n) n^{\theta \left(2m_o \frac{\frac{1}{s_o}(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)} + 1 \right)}. \quad (89)$$

Since $\theta \left(2m_o \frac{\frac{1}{s_o}(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)} + 1 \right) < 1$ from Eq. (86), $n > A_{\lambda,\tau}$ holds for sufficiently large $n \geq 1$.

Controlling L_λ : By Eq. (82), we have

$$\lambda \|f_\lambda\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 \leq n^{-\frac{2(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} \|f_*\|_{B_{2,q}^{s_x, s_o}(\mathbb{R}^{d_x+d_o})}^2.$$

Thus

$$\|f_\lambda\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 \leq n^{\frac{1+\frac{d_o}{s_o}(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} \|f_*\|_{B_{2,q}^{s_x, s_o}(\mathbb{R}^{d_x+d_o})}^2. \quad (90)$$

Thus we have

$$\begin{aligned} L_\lambda &= \max \left\{ 2C\sigma, \|T(f_* - [f_\lambda])\|_{L^\infty(\mathcal{Z} \times \mathcal{O})} \right\} \\ &\leq \max \left\{ 2C\sigma, \|f_* - [f_\lambda]\|_{L^\infty(\mathcal{X} \times \mathcal{O})} \right\} \\ &\leq \max \left\{ 2C\sigma, \|f_*\|_{L^\infty(\mathcal{X} \times \mathcal{O})} + \| [f_\lambda] \|_{L^\infty(\mathcal{X} \times \mathcal{O})} \right\} \\ &\stackrel{(i)}{\leq} \max \left\{ 2C\sigma, \|f_*\|_{L^\infty(\mathcal{X} \times \mathcal{O})} + \|f_\lambda\|_{\mathcal{H}_{\gamma_x, \gamma_o}} \right\} \\ &\stackrel{(ii)}{\leq} \max \left\{ 2C\sigma, \|f_*\|_{L^\infty(\mathcal{X} \times \mathcal{O})} + n^{\frac{1+\frac{d_o}{s_o}(\frac{s_x}{d_x} + \eta_1)}{2(1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1))}} \|f_*\|_{B_{2,q}^{s_x, s_o}(\mathbb{R}^{d_x+d_o})} \right\} \\ &\stackrel{(iii)}{\leq} 2n^{\frac{1+\frac{d_o}{s_o}(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} \|f_*\|_{B_{2,q}^{s_x, s_o}(\mathbb{R}^{d_x+d_o})}, \end{aligned} \quad (91)$$

where (i) holds by the reproducing property, (ii) holds by Eq. (90), and (iii) holds for sufficiently large n .

Bounding $()$:* Now we are ready to apply Fischer and Steinwart [2020][Theorem 16] to upper bound $(*)$. By Fischer and Steinwart [2020][Theorem 16], with P^n -probability $\geq 1 - 4e^{-\tau}$, for $n \geq A_{\lambda,\tau}$,

$$\begin{aligned} (*) &= \left\| C_{FO}^{\frac{1}{2}} (\bar{h}_\lambda - h_\lambda) \right\|_{\mathcal{H}_{FO}}^2 \\ &\leq \frac{576\tau^2}{n} \left(\sigma^2 \mathcal{N}_{FO}(\lambda) + \|k_{FO}^\theta\|_\infty^2 \frac{\|h_* - [h_\lambda]\|_{L^2(P_{ZO})}^2}{\lambda^\theta} + 2 \|k_{FO}^\theta\|_\infty^2 \frac{L_\lambda^2}{n\lambda^\theta} \right) \\ &\stackrel{(a)}{\lesssim} \frac{\tau^2}{n} \left((\log n)^{d_x+d_o+1} n^{\frac{1+\frac{d_o}{s_o}(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} + \|k_{FO}^\theta\|_\infty^2 n^\theta \left(\|h_* - [h_\lambda]\|_{L^2(P_{ZO})}^2 + \frac{L_\lambda^2}{n} \right) \right) \end{aligned}$$

$$\begin{aligned}
&\stackrel{(b)}{\lesssim} \tau^2 \left(n^{-\frac{2(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} (\log n)^{d_x + d_o + 1} + \gamma_o^{-2m_o\theta} n^{\theta-1} \cdot n^{-\frac{2(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} \right) \\
&\stackrel{(c)}{\lesssim} \tau^2 \left(n^{-\frac{2(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} (\log n)^{d_x + d_o + 1} + n^{-\frac{2(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} \right) \\
&\lesssim \tau^2 n^{-\frac{2(\frac{s_x}{d_x} + \eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}} (\log n)^{d_x + d_o + 1}.
\end{aligned}$$

We use Proposition C.1 to upper bound $\mathcal{N}_{FO}(\lambda)$ in (a), we use Eq. (91) to upper bound L_λ , Eq. (88) to upper bound $\|k_{FO}^\theta\|_\infty$ and Proposition D.7 to upper bound $\|h_* - [h_\lambda]\|_{L^2(P_{ZO})} = \|Tf_* - T[f_\lambda]\|_{L^2(P_{ZO})}$ in (b), and we use the following fact in (c): by Eq. (86), we have

$$\begin{aligned}
\theta &\leq \frac{1 + (2 + \frac{d_o}{s_o})(\frac{s_x}{d_x} + \eta_1)}{1 + 2(\frac{s_x}{d_x} + \eta_1) + \frac{2m_o+d_o}{s_o}(\frac{s_x}{d_x} + \eta_1)} \\
&\iff n^{\frac{2m_o\theta}{s_o}(\frac{s_x}{d_x} + \eta_1)} n^{\theta-1} \leq 1 \\
&\iff \gamma_o^{-2m_o\theta} n^{\theta-1} \leq 1.
\end{aligned}$$

The proof is concluded. \square

D.4 Auxiliary Lemmas for Section D

Lemma D.1. *Let Ω_1, Ω_2 be two open sets in \mathbb{R}^{n_1} and \mathbb{R}^{n_2} respectively. Let \mathcal{H} be a Hilbert space. Then we have*

$$MW^{m,p}(\Omega_1 \times \Omega_2, \mathcal{H}) \cong W^m(\Omega_1, W^p(\Omega_2, \mathcal{H})) \cong W^m(\Omega_1) \otimes W^p(\Omega_2) \otimes \mathcal{H}$$

Proof. We adapt the proof of Aubin [2011][Theorem 12.7.2]. Let $f \in W^m(\Omega_1, W^p(\Omega_2, \mathcal{H}))$. Then $f \in L^2(\Omega_1, W^p(\Omega_2, \mathcal{H}))$ such that $\partial_1^\alpha f \in L^2(\Omega_1, W^p(\Omega_2, \mathcal{H}))$ for $|\alpha| \leq m$. As explained in Hytönen et al. [2016][Definition 2.5.1], to say that $\partial_1^\alpha f \in L^2(\Omega_1, W^p(\Omega_2, \mathcal{H}))$ is equivalent to saying that for every $\phi \in C_c^\infty(\Omega_1)$, we have

$$\int_{\Omega_1} f(x, \cdot) \partial_1^\alpha \phi(x) \, dx = (-1)^{|\alpha|} \int_{\Omega_1} (\partial_1^\alpha f)(x, \cdot) \phi(x) \, dx \quad (92)$$

belongs to $W^p(\Omega_2, \mathcal{H})$. But this exactly says that $f \in MW^{m,p}(\Omega_1 \times \Omega_2, \mathcal{H})$. Thus the first isometric isomorphism is proved. The second isomorphism is proved by two successive applications of Aubin [2011, Theorem 12.7.1], namely the fact that the Hilbert-space valued Sobolev space $W^m(\Omega, \mathcal{H})$ is isometrically isomorphic to the Hilbertian tensor product $W^m(\Omega) \otimes \mathcal{H}$. \square

Lemma D.2. *Fix $\tau \geq 1$. Suppose that the assumptions of Proposition D.4 hold. Let $\lambda = \frac{1}{n}$. We assume that $n \geq 1$ satisfies $n > A_{\lambda, \tau}$ with $A_{\lambda, \tau}$ defined in Eq. (87) and $\tilde{n} \geq 1$ satisfies Eq. (21). Then, with $P^{n+\tilde{n}}$ -probability $\geq 1 - 10e^{-\tau}$*

$$S_0 \leq c' \tau \sqrt{n} \left(\frac{\|\hat{F}_\xi - F_*\|_{\mathcal{G}}}{\sqrt{n}} + \left\| [\hat{F}_\xi] - F_* \right\|_{L^2(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_{X, \gamma_x})} \right) \|\bar{h}_\lambda\|_{\mathcal{H}_{FO}},$$

where c' is a constant independent of n, \tilde{n} . Under the same probabilistic event, the bounds in Proposition D.4 hold.

Proof. We adapt the proof of Meunier et al. [2024a][Theorem 12].

$$\begin{aligned}
S_0 &= \left\| (C_{FO} + \lambda)^{1/2} \left(\hat{C}_{\hat{F}O} + \lambda \right)^{-1} \left(\hat{C}_{\hat{F}O} - \hat{C}_{FO} \right) \left(\hat{C}_{FO} + \lambda \right)^{-1} \frac{1}{n} \Phi_{FO}^* \mathbf{Y} \right\|_{\mathcal{H}_{FO}} \\
&\leq \lambda^{-1/2} \underbrace{\left\| (C_{FO} + \lambda)^{1/2} \left(\hat{C}_{FO} + \lambda \right)^{-1/2} \right\|}_{(A)} \\
&\quad \cdot \underbrace{\left\| \left(\hat{C}_{FO} + \lambda \right)^{1/2} \left(\hat{C}_{\hat{F}O} + \lambda \right)^{-1/2} \right\|}_{(B)} \cdot \underbrace{\left\| \hat{C}_{\hat{F}O} - \hat{C}_{FO} \right\|}_{(C)} \cdot \|\bar{h}_\lambda\|_{\mathcal{H}_{FO}}
\end{aligned}$$

Upper bound for term (A). Notice that all assumptions from Fischer and Steinwart [2020, Lemma 17] have been checked already in the analysis of projected estimation error in Section D.3. Since $n \geq A_{\lambda, \tau}$ with $A_{\lambda, \tau}$ defined in Eq. (87), with P^n -probability $\geq 1 - 2e^{-\tau}$,

$$\begin{aligned}
\left\| (C_{FO} + \lambda)^{-1/2} \left(C_{FO} - \hat{C}_{FO} \right) (C_{FO} + \lambda)^{-1/2} \right\| &\leq \frac{4 \|k_{FO}^\theta\|_\infty^2 \tau g_\lambda}{3n\lambda^\theta} + \sqrt{\frac{2 \|k_{FO}^\theta\|_\infty^2 \tau g_\lambda}{n\lambda^\theta}} \\
&\leq \frac{2}{3}.
\end{aligned}$$

By Rudi et al. [2015, Proposition 7], we have

$$(A) \leq \left(1 - \left\| (C_{FO} + \lambda)^{-1/2} \left(C_{FO} - \hat{C}_{FO} \right) (C_{FO} + \lambda)^{-1/2} \right\| \right)^{-\frac{1}{2}} \leq (1/3)^{-\frac{1}{2}} \leq 2.$$

Upper bound for term (C). Define

$$\begin{aligned}
\tilde{\Phi}_{\hat{F}O} : \mathcal{H}_{\gamma_x, \gamma_o} &\rightarrow \mathbb{R}^n = \left[\left(\hat{F}_\xi(\mathbf{z}_1, \mathbf{o}_1) \otimes \phi_{\gamma_o}(\mathbf{o}_1) \right), \dots, \left(\hat{F}_\xi(\mathbf{z}_n, \mathbf{o}_n) \otimes \phi_{\gamma_o}(\mathbf{o}_n) \right) \right]^* \\
\tilde{\Phi}_{FO} : \mathcal{H}_{\gamma_x, \gamma_o} &\rightarrow \mathbb{R}^n = [(F_*(\mathbf{z}_1, \mathbf{o}_1) \otimes \phi_{\gamma_o}(\mathbf{o}_1)), \dots, (F_*(\mathbf{z}_n, \mathbf{o}_n) \otimes \phi_{\gamma_o}(\mathbf{o}_1))]^*.
\end{aligned}$$

An immediate consequence is that $V \tilde{\Phi}_{\hat{F}O}^* = \Phi_{\hat{F}O}^*$ and $V \tilde{\Phi}_{FO}^* = \Phi_{FO}^*$ for V defined following Definition 12 and $\Phi_{\hat{F}O}, \Phi_{FO}$ defined in Eq. (77). Hence, we have

$$\begin{aligned}
(C) &= \left\| \hat{C}_{\hat{F}O} - \hat{C}_{FO} \right\| \\
&= \left\| \frac{1}{n} \Phi_{FO}^* \Phi_{FO} - \frac{1}{n} \Phi_{\hat{F}O}^* \Phi_{\hat{F}O} \right\| \\
&= \left\| \frac{1}{n} V \tilde{\Phi}_{FO}^* \tilde{\Phi}_{FO} V^* - \frac{1}{n} V \tilde{\Phi}_{\hat{F}O}^* \tilde{\Phi}_{\hat{F}O} V^* \right\| \\
&= \left\| \frac{1}{n} \tilde{\Phi}_{FO}^* \tilde{\Phi}_{FO} - \frac{1}{n} \tilde{\Phi}_{\hat{F}O}^* \tilde{\Phi}_{\hat{F}O} \right\| \\
&= \left\| \frac{1}{n} \sum_{i=1}^n \left(\hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) \otimes \phi_{\gamma_o}(\mathbf{o}_i) \right) \otimes \left(\hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) \otimes \phi_{\gamma_o}(\mathbf{o}_i) \right) \right. \\
&\quad \left. - \frac{1}{n} \sum_{i=1}^n (F_*(\mathbf{z}_i, \mathbf{o}_i) \otimes \phi_{\gamma_o}(\mathbf{o}_i)) \otimes (F_*(\mathbf{z}_i, \mathbf{o}_i) \otimes \phi_{\gamma_o}(\mathbf{o}_i)) \right\|.
\end{aligned}$$

The second last equality holds because $V : \mathcal{H}_{XO} \rightarrow \mathcal{H}_{FO}$ is an isometry. We start with the following decomposition

$$\begin{aligned}
& \left(\hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) \otimes \phi_{\gamma_o}(\mathbf{o}_i) \right) \otimes \left(\hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) \otimes \phi_{\gamma_o}(\mathbf{o}_i) \right) \\
& \quad - (F_*(\mathbf{z}_i, \mathbf{o}_i) \otimes \phi_{\gamma_o}(\mathbf{o}_i)) \otimes (F_*(\mathbf{z}_i, \mathbf{o}_i) \otimes \phi_{\gamma_o}(\mathbf{o}_i)) \\
& = \left(\left(\hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) - F_*(\mathbf{z}_i, \mathbf{o}_i) \right) \otimes \phi_{\gamma_o}(\mathbf{o}_i) \right) \otimes \left(\left(\hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) - F_*(\mathbf{z}_i, \mathbf{o}_i) \right) \otimes \phi_{\gamma_o}(\mathbf{o}_i) \right) \\
& \quad + \left(\left(\hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) - F_*(\mathbf{z}_i, \mathbf{o}_i) \right) \otimes \phi_{\gamma_o}(\mathbf{o}_i) \right) \otimes (F_*(\mathbf{z}_i, \mathbf{o}_i) \otimes \phi_{\gamma_o}(\mathbf{o}_i)) \\
& \quad + (F_*(\mathbf{z}_i, \mathbf{o}_i) \otimes \phi_{\gamma_o}(\mathbf{o}_i)) \otimes \left(\left(\hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) - F_*(\mathbf{z}_i, \mathbf{o}_i) \right) \otimes \phi_{\gamma_o}(\mathbf{o}_i) \right).
\end{aligned}$$

Thus we have

$$\begin{aligned}
(C) & \leq \left\| \frac{1}{n} \sum_{i=1}^n \left(\left(\hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) - F_*(\mathbf{z}_i, \mathbf{o}_i) \right) \otimes \phi_{\gamma_o}(\mathbf{o}_i) \right) \otimes \left(\left(\hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) - F_*(\mathbf{z}_i, \mathbf{o}_i) \right) \otimes \phi_{\gamma_o}(\mathbf{o}_i) \right) \right\| \\
& \quad + 2 \left\| \frac{1}{n} \sum_{i=1}^n \left(\left(\hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) - F_*(\mathbf{z}_i, \mathbf{o}_i) \right) \otimes \phi_{\gamma_o}(\mathbf{o}_i) \right) \otimes (F_*(\mathbf{z}_i, \mathbf{o}_i) \otimes \phi_{\gamma_o}(\mathbf{o}_i)) \right\| \\
& \leq \frac{1}{n} \sum_{i=1}^n \left\| \left(\left(\hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) - F_*(\mathbf{z}_i, \mathbf{o}_i) \right) \otimes \phi_{\gamma_o}(\mathbf{o}_i) \right) \otimes \left(\left(\hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) - F_*(\mathbf{z}_i, \mathbf{o}_i) \right) \otimes \phi_{\gamma_o}(\mathbf{o}_i) \right) \right\| \\
& \quad + 2 \frac{1}{n} \sum_{i=1}^n \left\| \left(\left(\hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) - F_*(\mathbf{z}_i, \mathbf{o}_i) \right) \otimes \phi_{\gamma_o}(\mathbf{o}_i) \right) \otimes (F_*(\mathbf{z}_i, \mathbf{o}_i) \otimes \phi_{\gamma_o}(\mathbf{o}_i)) \right\| \\
& \leq \frac{1}{n} \sum_{i=1}^n \left\| \hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) - F_*(\mathbf{z}_i, \mathbf{o}_i) \right\|_{\mathcal{H}_{X,\gamma_x}}^2 + \frac{2}{n} \sum_{i=1}^n \left\| \hat{F}_\xi(\mathbf{z}_i, \mathbf{o}_i) - F_*(\mathbf{z}_i, \mathbf{o}_i) \right\|_{\mathcal{H}_{X,\gamma_x}},
\end{aligned}$$

where in the last step we use $\|a \otimes b\|_{H_1 \otimes H_2} = \|a\|_{H_1} \|b\|_{H_2}$ for $a \in H_1, b \in H_2$ for Hilbert spaces H_1, H_2 [Gretton et al., 2005, Eq. (3)], $\|\phi_{\gamma_o}(\mathbf{o}_i) \otimes \phi_{\gamma_o}(\mathbf{o}_i)\| \leq 1$ and $\|F_*(\mathbf{z}_i, \mathbf{o}_i)\|_{\mathcal{H}_{X,\gamma_x}} \leq 1$. Note that the last line is exactly analyzed in the proof of Meunier et al. [2024a, Lemma 7]. Their analysis employs a Hoeffding's concentration bounds with respect to Stage 2 samples \mathcal{D}_2 , and *conditioned* on Stage 1 samples \mathcal{D}_1 to show that, conditioned on \mathcal{D}_1 ,

$$(C) \leq J_0 \left(\sqrt{\frac{\tau}{n}} \left\| F_* - \hat{F}_\xi \right\|_{\mathcal{G}} + \left\| F_* - [\hat{F}_\xi] \right\|_{L^2(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_{X,\gamma_x})} \right) \quad (93)$$

with P^n -probability $\geq 1 - 4e^{-\tau}$, under the assumptions that

$$\left\| F_* - [\hat{F}_\xi] \right\|_{L^2(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_{X,\gamma_x})} \vee \left\| F_* - \hat{F}_\xi \right\|_{\mathcal{G}} \leq 1. \quad (94)$$

We note that the independence of \mathcal{D}_1 and \mathcal{D}_2 is implicitly used in this step to ensure that \mathcal{D}_2 remains i.i.d. after conditioning on \mathcal{D}_1 . J_0 is a constant independent of n, \tilde{n} . Denote as \mathfrak{D} the $P^{\tilde{n}}$ -probabilistic event that the bounds in Proposition D.4 hold (which has $P^{\tilde{n}}$ -probability $\geq 1 - 4e^{-\tau}$). Under this event \mathfrak{D} along with the fact that \tilde{n} satisfies Eq. (21), Eq. (94) is satisfied so Eq. (93) holds. Additionally, from Eq. (93), a sufficient condition for $(C) \leq \frac{1}{6n}$ is given by $\tilde{n} \geq (6JJ_0\tau^{\frac{3}{2}}n)^{2^{\frac{m^\dagger+d^\dagger/2+\zeta}{m^\dagger}}} \vee (36J^2J_0^2\tau^2n)^{\frac{m^\dagger+d^\dagger/2+\zeta}{m^\dagger-1}}$, which is satisfied since \tilde{n} satisfies Eq. (21).

Upper bound for term (B). By Rudi et al. [2015, Proposition 7], we have $(B) \leq (1-t)^{-\frac{1}{2}}$, where

$$t := \left\| \left(\hat{C}_{FO} + \lambda \right)^{-\frac{1}{2}} \left(\hat{C}_{FO} - \hat{C}_{\hat{F}O} \right) \left(\hat{C}_{FO} + \lambda \right)^{-\frac{1}{2}} \right\| \leq \lambda^{-1} \left\| \hat{C}_{FO} - \hat{C}_{\hat{F}O} \right\| = n \cdot (C) \leq \frac{1}{6},$$

from where we have $(B) \leq \frac{6}{5}$. Under the event \mathfrak{D} , the upper bounds $(A) \leq 2$, $(B) \leq 6/5$, $(C) \leq \frac{1}{6}$ hold simultaneously with P^n -probability $\geq 1 - 6e^{-\tau}$. Since \mathfrak{D} holds with $P^{\tilde{n}}$ -probability $\geq 1 - 4e^{-\tau}$, by independence of Stage 1 and Stage 2 samples (which is a consequence of the sample splitting strategy), we have that the above upper bounds hold simultaneously with $P^{n+\tilde{n}}$ -probability $\geq (1 - 6e^{-\tau})(1 - 4e^{-\tau}) \geq 1 - 10e^{-\tau}$. \square

Lemma D.3. Fix $\tau \geq 1$. Suppose that the assumptions of Proposition D.4 hold. Let $\lambda = \frac{1}{n}$. We assume that $n \geq 1$ satisfies $n \geq A_{\lambda, \tau}$ with $A_{\lambda, \tau}$ defined in Eq. (87) and $\tilde{n} \geq 1$ satisfies Eq. (21). Then, with $P^{n+\tilde{n}}$ -probability $\geq 1 - 18e^{-\tau}$

$$S_{-1} \leq c\tau\sqrt{n} \left(\frac{\|\hat{F}_\xi - F_*\|_{\mathcal{G}}}{\sqrt{n}} + \left\| [\hat{F}_\xi] - F_* \right\|_{L^2(\mathcal{Z} \times \mathcal{O}; \mathcal{H}_{X, \gamma_x})} \right) \|\bar{h}_\lambda\|_{\mathcal{H}_{FO}}.$$

where c is a constant independent of \tilde{n}, n . Under the same probabilistic event, the bounds in Proposition D.4 hold.

Proof. We omit the proof since it is similar to Meunier et al. [2024a, Theorem 11], with adaptations similar to those in the proof of Lemma D.2. \square

Lemma D.4. Suppose the eigenvalues of the operator Σ_1 satisfy $\mu_{1,i} \asymp i^{-1/p_1}$ and the eigenvalues of the operator Σ_2 satisfy $\mu_{2,i} \asymp i^{-1/p_2}$. Then the eigenvalues of their tensor product $\Sigma_1 \otimes \Sigma_2$ satisfy, for any $\zeta > 0$,

$$\lambda_i(\Sigma_1 \otimes \Sigma_2) \leq i^{-1/p_1 \wedge 1/p_2 + \zeta}.$$

Proof. Suppose $(\mu_{1,i})_{i \in \mathbb{N}^+}$ (resp. $(\mu_{2,j})_{j \in \mathbb{N}^+}$) are the eigenvalues of Σ_1 (resp. Σ_2) with corresponding eigenfunctions $(q_{1,i})_{i \in \mathbb{N}^+}$ (resp. $(q_{2,j})_{j \in \mathbb{N}^+}$). By definition of operator tensor product, we have

$$(\Sigma_1 \otimes \Sigma_2)(q_{1,i} \otimes q_{2,j}) = (\Sigma_1 q_{1,i}) \otimes (\Sigma_2 q_{2,j}) = \mu_{1,i} \mu_{2,j} (q_{1,i} \otimes q_{2,j}).$$

Therefore, for any $i \in \mathbb{N}^+, j \in \mathbb{N}^+$ we obtain that $\mu_{1,i} \cdot \mu_{2,j}$ is the eigenvalue of $\Sigma_1 \otimes_{op} \Sigma_2$ with corresponding eigenfunction $q_{1,i} \otimes q_{2,j}$. Therefore the eigenvalues of $\Sigma_1 \otimes_{op} \Sigma_2$ equal the tensor product of two sequences $(\mu_{1,i})_{i \in \mathbb{N}^+}$ and $(\mu_{2,j})_{j \in \mathbb{N}^+}$.

Without loss of generality, we assume $p_1 \leq p_2$ so that $\mu_{1,i} \leq \mu_{2,i}$ for i large enough. Denote $(\sigma_n)_{n \in \mathbb{N}^+}$ as the tensor product of two sequences $(\mu_{2,i})_{i \in \mathbb{N}^+}$ and $(\mu_{2,j})_{j \in \mathbb{N}^+}$ rearranged in a non-increasing order. Hence, we have $\lambda_n(\Sigma_1 \otimes \Sigma_2) \leq \sigma_n$. From Krieg [2018], we have

$$\sigma_n \leq n^{-1/p_2} (\log n)^{1/p_2} \leq n^{-1/p_2 + \zeta},$$

for any $\zeta > 0$. It concludes the proof. \square

Lemma D.5. Let $\mathcal{X} = [0, 1]^d$. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ and $f \in B_{2,q}^s(\mathcal{X}) \cap L^1(\mathbb{R}^d)$ with $s = [s_1, \dots, s_d]^\top$. Recall K_γ defined in Eq. (66) with $\gamma = [\gamma_1, \dots, \gamma_d]^\top \in (0, 1]^d$. Then, we have

$$\|[K_\gamma * f] - f\|_{L^2(\mathcal{X})} \leq C_0 |f|_{B_{2,q}^s(\mathcal{X})} \cdot \max\{\gamma_1^{s_1}, \dots, \gamma_d^{s_d}\}$$

for some constant C_0 that only depends on d, s .

Proof. Notice that

$$\begin{aligned}
K_{\gamma} * f(\mathbf{x}) &= \int_{\mathbb{R}^d} \sum_{j=1}^r \binom{r}{j} (-1)^{1-j} \frac{1}{j^d} \left(\frac{2}{\pi} \right)^{\frac{d}{2}} \left(\prod_{i=1}^d \frac{1}{\gamma_i} \right) \exp \left(-2 \sum_{i=1}^d \frac{(x_i - t_i)^2}{j^2 \gamma_i^2} \right) f(\mathbf{t}) d\mathbf{t} \\
&= \sum_{j=1}^r \binom{r}{j} (-1)^{1-j} \frac{1}{j^d} \left(\frac{2}{\pi} \right)^{\frac{d}{2}} \left(\prod_{i=1}^d \frac{1}{\gamma_i} \right) \int_{\mathbb{R}^d} \exp \left(-2 \sum_{i=1}^d \frac{(x_i - t_i)^2}{j^2 \gamma_i^2} \right) f(\mathbf{t}) d\mathbf{t} \\
&= \sum_{j=1}^r \binom{r}{j} (-1)^{1-j} \frac{1}{j^d} \left(\frac{2}{\pi} \right)^{\frac{d}{2}} \left(\prod_{i=1}^d \frac{1}{\gamma_i} \right) \int_{\mathbb{R}^d} \exp \left(-2 \sum_{i=1}^d \frac{h_i^2}{\gamma_i^2} \right) f(\mathbf{x} + j\mathbf{h}) j^d d\mathbf{h} \\
&= \sum_{j=1}^r \binom{r}{j} (-1)^{1-j} \left(\frac{2}{\pi} \right)^{\frac{d}{2}} \int_{\mathbb{R}^d} \exp(-2\|\mathbf{h}\|^2) f(\mathbf{x} + j\mathbf{h} \odot \boldsymbol{\gamma}) d\mathbf{h} \\
&= \int_{\mathbb{R}^d} \left(\frac{2}{\pi} \right)^{\frac{d}{2}} \exp(-2\|\mathbf{h}\|^2) \left(\sum_{j=1}^r \binom{r}{j} (-1)^{1-j} f(\mathbf{x} + j\mathbf{h} \odot \boldsymbol{\gamma}) \right) d\mathbf{h}.
\end{aligned}$$

Next, since $\int_{\mathbb{R}^d} \left(\frac{2}{\pi} \right)^{\frac{d}{2}} \exp(-2\|\mathbf{h}\|^2) d\mathbf{h} = 1$, we have that

$$\begin{aligned}
&\| [K_{\gamma} * f] - f \|_{L^2(\mathcal{X})}^2 \\
&= \int_{\mathcal{X}} \left(\int_{\mathbb{R}^d} \left(\frac{2}{\pi} \right)^{\frac{d}{2}} \exp(-2\|\mathbf{h}\|^2) \left(\sum_{j=0}^r \binom{r}{j} (-1)^{1-j} f(\mathbf{x} + j\mathbf{h} \odot \boldsymbol{\gamma}) \right) d\mathbf{h} \right)^2 d\mathbf{x} \\
&= \int_{\mathcal{X}} \left(\int_{\mathbb{R}^d} \left(\frac{2}{\pi} \right)^{\frac{d}{2}} \exp(-2\|\mathbf{h}\|^2) \left(\sum_{j=0}^r \binom{r}{j} (-1)^{2r+1-j} f(\mathbf{x} + j\mathbf{h} \odot \boldsymbol{\gamma}) \right) d\mathbf{h} \right)^2 d\mathbf{x} \\
&= \int_{\mathcal{X}} \left(\int_{\mathbb{R}^d} (-1)^{r+1} \left(\frac{2}{\pi} \right)^{\frac{d}{2}} \exp(-2\|\mathbf{h}\|^2) |\Delta_{\mathbf{h} \odot \boldsymbol{\gamma}}^r f(\mathbf{x})| d\mathbf{h} \right)^2 d\mathbf{x}.
\end{aligned}$$

Where the last step follows from the definition of modulus of smoothness in Eq. (9). Then, from Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\leq \int_{\mathcal{X}} \left(\int_{\mathbb{R}^d} \left(\frac{2}{\pi} \right)^{\frac{d}{2}} \exp(-2\|\mathbf{h}\|^2) d\mathbf{h} \right) \left(\int_{\mathbb{R}^d} \left(\frac{2}{\pi} \right)^{\frac{d}{2}} \exp(-2\|\mathbf{h}\|^2) |\Delta_{\mathbf{h} \odot \boldsymbol{\gamma}}^r f(\mathbf{x})|^2 d\mathbf{h} \right) d\mathbf{x} \\
&= \int_{\mathcal{X}} \int_{\mathbb{R}^d} \left(\frac{2}{\pi} \right)^{\frac{d}{2}} \exp(-2\|\mathbf{h}\|^2) |\Delta_{\mathbf{h} \odot \boldsymbol{\gamma}}^r f(\mathbf{x})|^2 d\mathbf{h} d\mathbf{x} \\
&= \int_{\mathbb{R}^d} \left(\frac{2}{\pi} \right)^{\frac{d}{2}} \exp(-2\|\mathbf{h}\|^2) \|\Delta_{\mathbf{h} \odot \boldsymbol{\gamma}}^r f\|_{L^2(\mathcal{X})}^2 d\mathbf{h}. \tag{95}
\end{aligned}$$

Define $N := \min\{\gamma_1^{-s_1}, \dots, \gamma_d^{-s_d}\}$. Since $f \in B_{2,q}^{\mathbf{s}}(\mathcal{X})$, we have

$$\begin{aligned}
|f|_{B_{2,q}^{\mathbf{s}}(\mathcal{X})} &\geq |f|_{B_{2,\infty}^{\mathbf{s}}(\mathcal{X})} := \sup_t \left(t^{-1} \omega_{r,2} \left(f, t^{\frac{1}{s_1}}, \dots, t^{\frac{1}{s_d}}, \mathcal{X} \right) \right) \\
&\geq \left(\left(\sum_{i=1}^d h_i^{s_i} \right) N^{-1} \right)^{-1} \omega_{r,2} \left(f, \left(\sum_{i=1}^d h_i^{s_i} \right)^{\frac{1}{s_1}} N^{-\frac{1}{s_1}}, \dots, \left(\sum_{i=1}^d h_i^{s_i} \right)^{\frac{1}{s_d}} N^{-\frac{1}{s_d}}, \mathcal{X} \right)
\end{aligned}$$

$$\begin{aligned}
&\geq N \left(\sum_{i=1}^d h_i^{s_i} \right)^{-1} \omega_{r,2}(f, h_1 \gamma_1, \dots, h_d \gamma_d, \mathcal{X}) \\
&\geq N \left(\sum_{i=1}^d h_i^{s_i} \right)^{-1} \|\Delta_{\mathbf{h} \odot \boldsymbol{\gamma}}^r f\|_{L^2(\mathcal{X})}.
\end{aligned}$$

As a result, we have $\|\Delta_{\mathbf{h} \odot \boldsymbol{\gamma}}^r f\|_{L^2(\mathcal{X})} \leq |f|_{B_{2,q}^s(\mathcal{X})} N^{-1} (\sum_{i=1}^d h_i^{s_i})$. Plugging the above back to Eq. (95), we obtain

$$\begin{aligned}
\|K_{\boldsymbol{\gamma}} * f - f\|_{L^2(\mathcal{X})}^2 &\leq \int_{\mathbb{R}^d} \left(\frac{2}{\pi} \right)^{\frac{d}{2}} \exp(-2\|\mathbf{h}\|^2) \left(|f|_{B_{2,q}^s(\mathcal{X})} N^{-1} \left(\sum_{i=1}^d h_i^{s_i} \right) \right)^2 d\mathbf{h} \\
&\leq |f|_{B_{2,q}^s(\mathcal{X})}^2 N^{-2} d \int_{\mathbb{R}^d} \left(\frac{2}{\pi} \right)^{\frac{d}{2}} \exp(-2\|\mathbf{h}\|^2) \left(\sum_{i=1}^d h_i^{2s_i} \right) d\mathbf{h}.
\end{aligned}$$

Notice that

$$C_0^2 := d \int_{\mathbb{R}^d} \left(\frac{2}{\pi} \right)^{\frac{d}{2}} \exp(-2\|\mathbf{h}\|^2) \left(\sum_{i=1}^d h_i^{2s_i} \right) d\mathbf{h} = d \sum_{i=1}^d \int_{\mathbb{R}} \left(\frac{2}{\pi} \right)^{\frac{1}{2}} \exp(-2h_i^2) h_i^{2s_i} dh_i,$$

which is the sum of s_i -th moment of a one dimensional Gaussian distribution $\mathcal{N}(0, 1/4)$ from $i = 1$ to $i = d$. Finally, we obtain

$$\|K_{\boldsymbol{\gamma}} * f - f\|_{L^2(\mathcal{X})}^2 \leq C_0^2 |f|_{B_{2,q}^s(\mathcal{X})}^2 N^{-2} = C_0^2 |f|_{B_{2,q}^s(\mathcal{X})}^2 (\max\{\gamma_1^{s_1}, \dots, \gamma_d^{s_d}\})^2.$$

□

Lemma D.6. *Given a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, we have*

$$\|K_{\boldsymbol{\gamma}} * f\|_{\mathcal{H}_{X,\boldsymbol{\gamma}}} \leq \left(\prod_{i=1}^d \gamma_i^{-\frac{1}{2}} \right) \pi^{-\frac{d}{4}} (2^r - 1) \|f\|_{L^2(\mathbb{R}^d)}.$$

Proof. Following the same arguments in the proof of proposition 4 in Hang and Steinwart [2021], we have $K_{\boldsymbol{\gamma}} * f \in \mathcal{H}_{X,\boldsymbol{\gamma}}$. For a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, we define $\tau_{\boldsymbol{\gamma}} f(\mathbf{x}) = f(\boldsymbol{\gamma} \odot \mathbf{x})$. We have, for all $\mathbf{x} \in \mathbb{R}^d$,

$$\begin{aligned}
\tau_{\boldsymbol{\gamma}} (K_{\boldsymbol{\gamma}} * f)(\mathbf{x}) &= \int_{\mathbb{R}^d} K_{\boldsymbol{\gamma}}(\boldsymbol{\gamma} \odot \mathbf{x} - \mathbf{t}) f(\mathbf{t}) d\mathbf{t} \\
&= \int_{\mathbb{R}^d} K_{\mathbf{1}} \left(\frac{\boldsymbol{\gamma} \odot \mathbf{x} - \mathbf{t}}{\boldsymbol{\gamma}} \right) \left(\prod_{i=1}^d \frac{1}{\gamma_i} \right) f(\mathbf{t}) d\mathbf{t} \\
&= \int_{\mathbb{R}^d} K_{\mathbf{1}} \left(\mathbf{x} - \frac{\mathbf{t}}{\boldsymbol{\gamma}} \right) \left(\prod_{i=1}^d \frac{1}{\gamma_i} \right) f(\mathbf{t}) d\mathbf{t} \\
&= \int_{\mathbb{R}^d} K_{\mathbf{1}}(\mathbf{x} - \mathbf{t}) f(\mathbf{t} \odot \boldsymbol{\gamma}) d\mathbf{t} \\
&= (K_{\mathbf{1}} * (\tau_{\boldsymbol{\gamma}} f))(\mathbf{x}).
\end{aligned}$$

Proposition 4.37 of Steinwart and Christmann [2008] can be generalized to anisotropic case which shows that for any $f \in \mathcal{H}_{X,\gamma}$, $\tau_\gamma f \in \mathcal{H}_{X,1}$ and $\tau_\gamma : \mathcal{H}_{X,\gamma} \rightarrow \mathcal{H}_{X,1}$ is an isometric isomorphism. Hence we have

$$\begin{aligned}\|K_\gamma * f\|_{\mathcal{H}_{X,\gamma}} &= \|\tau_\gamma(K_\gamma * f)\|_{\mathcal{H}_{X,1}} \\ &= \|K_1 * (\tau_\gamma f)\|_{\mathcal{H}_{X,1}} \\ &\leq \pi^{-\frac{d}{4}}(2^r - 1) \|\tau_\gamma f\|_{L^2(\mathbb{R}^d)} \\ &= \left(\prod_{i=1}^d \gamma_i^{-\frac{1}{2}} \right) \pi^{-\frac{d}{4}}(2^r - 1) \|f\|_{L^2(\mathbb{R}^d)},\end{aligned}$$

where the second last inequality follows from Eberts and Steinwart [2013][Theorem 2.3] setting $\gamma = 1$. \square

Corollary D.1. *Let $f : \mathbb{R}^{d_x+d_o} \rightarrow \mathbb{R}$ and*

$$\gamma_x = n^{-\frac{\frac{1}{d_x}}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}}, \quad \gamma_o = n^{-\frac{\frac{1}{s_o}(\frac{s_x}{d_x}+\eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}}.$$

Recall $K_{\gamma_x} : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$ and $K_{\gamma_o} : \mathbb{R}^{d_o} \rightarrow \mathbb{R}$ defined in Eq. (66), then we have

$$\|K_{\gamma_x} * K_{\gamma_o} * f\|_{\mathcal{H}_{\gamma_x, \gamma_o}} \leq C_1 n^{\frac{1}{2} \frac{1+\frac{d_o}{s_o}(\frac{s_x}{d_x}+\eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}} \|f\|_{L^2(\mathbb{R}^{d_x+d_o})},$$

for some constant C_1 that only depends on d_x, d_o, s_x, s_o .

Proof. The proof is a direct application of Lemma D.6. \square

Lemma D.7. *Let $\gamma_x = n^{-\frac{\frac{1}{d_x}}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}}$, $\gamma_o = n^{-\frac{\frac{1}{s_o}(\frac{s_x}{d_x}+\eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}}$. Let $\iota_{x, \gamma_x^{-1}} : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$ be an indicator function over $\{\omega_x : \|\omega_x\|_2 \leq \gamma_x^{-1}\}$ and $K_{\gamma_o} : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$ be as defined in Eq. (66). Then we have*

$$\|\mathcal{F}^{-1}[\iota_{x, \gamma_x^{-1}}] * K_{\gamma_o} * f\|_{\mathcal{H}_{\gamma_x, \gamma_o}} \lesssim 2^r n^{\frac{1}{2} \frac{1+\frac{d_o}{s_o}(\frac{s_x}{d_x}+\eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}} \|f\|_{L^2(\mathbb{R}^{d_x+d_o})}.$$

Proof. In contrast to the notation used in the rest of the document (with the exception of Lemma D.9), in this proof, we use \mathcal{F} to denote the unitary Fourier operator on $L^2(\mathbb{R}^{d_x+d_o})$, we use \mathcal{F}_x to denote the unitary Fourier operator on $L^2(\mathbb{R}^{d_x})$, and we use \mathcal{F}_o to denote the unitary Fourier operator on $L^2(\mathbb{R}^{d_o})$. We let $\otimes : L^2(\mathbb{R}^{d_x}) \times L^2(\mathbb{R}^{d_o}) \rightarrow L^2(\mathbb{R}^{d_x+d_o})$ denotes the tensor product mapping, where $(f \otimes g)(\mathbf{x}, \mathbf{o}) = f(\mathbf{x})g(\mathbf{o})$ for all $(\mathbf{x}, \mathbf{o}) \in \mathbb{R}^{d_x+d_o}$. By considering a convergent sequence to f_1 in $L^1(\mathbb{R}^{d_x}) \cap L^2(\mathbb{R}^{d_x})$, and similarly for f_2 , we can show that

$$\mathcal{F}[f \otimes g] = \mathcal{F}_x[f] \otimes \mathcal{F}_o[g]$$

for $f \in L^2(\mathbb{R}^{d_x})$ and $g \in L^2(\mathbb{R}^{d_o})$. At the start of Section D, we proved that for $f \in L^1(\mathbb{R}^{d_x+d_o})$, $f * \mathcal{F}^{-1}[\iota_{x, \gamma_x^{-1}}] \in L^2(\mathbb{R}^{d_x+d_o})$. Since $\|K_{\gamma_o}\|_{L^1(\mathbb{R}^{d_o})} = 1$ [Giné and Nickl, 2021, Section 4.1.2], we have

$$\|(f * \mathcal{F}^{-1}[\iota_{x, \gamma_x^{-1}}]) * K_{\gamma_o}\|_{L^2(\mathbb{R}^{d_x+d_o})}^2 \stackrel{(a)}{=} \int_{\mathbb{R}^{d_x}} \|(f * \mathcal{F}^{-1}[\iota_{x, \gamma_x^{-1}}])(\mathbf{x}, \cdot) * K_{\gamma_o}\|_{L^2(\mathbb{R}^{d_o})}^2 d\mathbf{x}$$

$$\begin{aligned} &\stackrel{(b)}{\leq} \int_{\mathbb{R}^{d_x}} \|(f * \mathcal{F}^{-1}[\iota_{x, \gamma_x^{-1}}])(\mathbf{x}, \cdot)\|_{L^2(\mathbb{R}^{d_o})}^2 \|K_{\gamma_o}\|_{L^1(\mathbb{R}^{d_o})}^2 \\ &= \|f * \mathcal{F}^{-1}[\iota_{x, \gamma_x^{-1}}]\|_{L^2(\mathbb{R}^{d_x+d_o})}^2. \end{aligned}$$

where in (a), $(f * \mathcal{F}^{-1}[\iota_{x, \gamma_x^{-1}}])(\mathbf{x}, \cdot)$ is a slice function of any representative of $f * \mathcal{F}^{-1}[\iota_{x, \gamma_x^{-1}}] \in L^2(\mathbb{R}^{d_x})$, and (b) follows from Young's Convolution Inequality. Hence we can apply \mathcal{F} to $f * \mathcal{F}^{-1}[\iota_{x, \gamma_x^{-1}}] \in L^2(\mathbb{R}^{d_x+d_o})$ and we find

$$\mathcal{F} \left[\mathcal{F}_x^{-1}[\iota_{x, \gamma_x^{-1}}] * K_{\gamma_o} * f \right] = (\iota_{x, \gamma_x^{-1}} \otimes \mathcal{F}_o[K_{\gamma_o}]) \cdot \mathcal{F}[f].$$

For $\phi(\mathbf{x}, \mathbf{o}) = \exp(-\gamma_x^{-2}\|\mathbf{x}\|_2^2) \cdot \exp(-\gamma_o^{-2}\|\mathbf{o}\|_2^2)$, its Fourier transform is

$$\mathcal{F}[\phi](\boldsymbol{\omega}_x, \boldsymbol{\omega}_o) = \pi^{d_x/2+d_o/2} \gamma_x^{d_x} \exp\left(-\frac{1}{4}\gamma_x^2\|\boldsymbol{\omega}_x\|_2^2\right) \cdot \gamma_o^{d_o} \exp\left(-\frac{1}{4}\gamma_o^2\|\boldsymbol{\omega}_o\|_2^2\right).$$

Hence, by definition of Gaussian RKHS norm [Wendland, 2004, Theorem 10.12],

$$\begin{aligned} &\left\| \mathcal{F}_x^{-1}[\iota_{x, \gamma_x^{-1}}] * K_{\gamma_o} * f \right\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 \\ &= \int_{\mathbb{R}^{d_x+d_o}} \frac{\left(\iota_{x, \gamma_x^{-1}}(\boldsymbol{\omega}_x) \cdot \mathcal{F}[f](\boldsymbol{\omega}_x, \boldsymbol{\omega}_o) \cdot \mathcal{F}_o[K_{\gamma_o}](\boldsymbol{\omega}_o) \right)^2}{\pi^{d_x/2+d_o/2} \gamma_x^{d_x} \exp\left(-\frac{1}{4}\gamma_x^2\|\boldsymbol{\omega}_x\|_2^2\right) \cdot \gamma_o^{d_o} \exp\left(-\frac{1}{4}\gamma_o^2\|\boldsymbol{\omega}_o\|_2^2\right)} d\boldsymbol{\omega}_x d\boldsymbol{\omega}_o \\ &\lesssim \gamma_x^{-d_x} \gamma_o^{-d_o} \int_{\mathbb{R}^{d_o}} \int_{\{\boldsymbol{\omega}_x : \|\boldsymbol{\omega}_x\|_2 \leq \gamma_x^{-1}\}} \frac{(\mathcal{F}[f](\boldsymbol{\omega}_x, \boldsymbol{\omega}_o) \cdot \mathcal{F}_o[K_{\gamma_o}](\boldsymbol{\omega}_o))^2}{\exp\left(-\frac{1}{4}\gamma_x^2\|\boldsymbol{\omega}_x\|_2^2\right) \cdot \exp\left(-\frac{1}{4}\gamma_o^2\|\boldsymbol{\omega}_o\|_2^2\right)} d\boldsymbol{\omega}_x d\boldsymbol{\omega}_o \\ &\lesssim \gamma_x^{-d_x} \gamma_o^{-d_o} \int_{\mathbb{R}^{d_x+d_o}} \exp\left(\frac{1}{4}\gamma_o^2\|\boldsymbol{\omega}_o\|_2^2\right) (\mathcal{F}[f](\boldsymbol{\omega}_x, \boldsymbol{\omega}_o) \cdot \mathcal{F}_o[K_{\gamma_o}](\boldsymbol{\omega}_o))^2 d\boldsymbol{\omega}_x d\boldsymbol{\omega}_o. \end{aligned}$$

Since $K_{\gamma_o}(\mathbf{o}) := \sum_{j=1}^r \binom{r}{j} (-1)^{1-j} \frac{1}{j^{d_o} \gamma_o^{d_o}} \left(\frac{2}{\pi}\right)^{\frac{d}{2}} \exp(-2j^{-2}\gamma_o^{-2}\|\mathbf{o}\|_2^2)$, so we have

$$\mathcal{F}_o[K_{\gamma_o}](\boldsymbol{\omega}_o) = \sum_{j=1}^r \binom{r}{j} (-1)^{1-j} \exp\left(-\frac{1}{8}j^2\gamma_o^2\|\boldsymbol{\omega}_o\|_2^2\right).$$

Consequently, we have

$$\begin{aligned} &\left\| \mathcal{F}_x^{-1}[\iota_{x, \gamma_x^{-1}}] * K_{\gamma_o} * f \right\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 \\ &\lesssim \gamma_x^{-d_x} \gamma_o^{-d_o} \int_{\mathbb{R}^{d_x+d_o}} \frac{\left(\mathcal{F}[f](\boldsymbol{\omega}_x, \boldsymbol{\omega}_o) \cdot \sum_{j=1}^r \binom{r}{j} (-1)^{1-j} \exp(-\frac{1}{8}j^2\gamma_o^2\|\boldsymbol{\omega}_o\|_2^2) \right)^2}{\exp(-\frac{1}{4}\gamma_o^2\|\boldsymbol{\omega}_o\|_2^2)} d\boldsymbol{\omega}_x d\boldsymbol{\omega}_o \\ &\leq \gamma_x^{-d_x} \gamma_o^{-d_o} 2^{2r} \int_{\mathbb{R}^{d_x+d_o}} (\mathcal{F}[f](\boldsymbol{\omega}_x, \boldsymbol{\omega}_o))^2 d\boldsymbol{\omega}_x d\boldsymbol{\omega}_o \\ &= \gamma_x^{-d_x} \gamma_o^{-d_o} 2^{2r} \|f\|_{L^2(\mathbb{R}^{d_x+d_o})}^2. \end{aligned}$$

where the last equality follows by Plancherel's Theorem. \square

Lemma D.8. Let $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_d) \in (0, 1)^d$. Let $k_{\boldsymbol{\gamma}}$ be the anisotropic Gaussian kernel on \mathbb{R}^d and let $\mathcal{H}_{X, \boldsymbol{\gamma}}$ be the RKHS associated with $k_{\boldsymbol{\gamma}}$. Then

$$\int_{\mathbb{R}^d} |\mathcal{F}[f](\boldsymbol{\omega})|^2 \exp\left(\frac{\|\boldsymbol{\omega} \odot \boldsymbol{\gamma}\|_2^2}{4}\right) d\boldsymbol{\omega} = 2^d \pi^{d/2} \left(\prod_{i=1}^d \gamma_i\right) \|f\|_{\mathcal{H}_{X, \boldsymbol{\gamma}}}^2$$

Proof. We define $\phi_{X,\gamma}(\mathbf{x}) = \exp(-\sum_{j=1}^d \gamma_j^{-2} x_j^2)$. Then

$$\begin{aligned}\mathcal{F}[\phi_{X,\gamma}](\boldsymbol{\omega}) &= \int_{\mathbb{R}^d} \phi_{X,\gamma}(\mathbf{x}) \exp(-i\langle \mathbf{x}, \boldsymbol{\omega} \rangle) d\mathbf{x} \\ &= \left(\prod_{i=1}^d \gamma_i \right) \int_{\mathbb{R}^d} \phi_{X,1}(\mathbf{x}) \exp(-i\langle \mathbf{x}, \boldsymbol{\omega} \odot \boldsymbol{\gamma} \rangle) d\mathbf{x} \\ &= \pi^{d/2} \left(\prod_{i=1}^d \gamma_i \right) \exp\left(-\frac{\|\boldsymbol{\omega} \odot \boldsymbol{\gamma}\|_2^2}{4}\right)\end{aligned}$$

where the last step follows from Wendland [2004][Theorem 5.20]. By Kanagawa et al. [2018][Theorem 2.4], we have

$$\begin{aligned}\|f\|_{\mathcal{H}_{X,\gamma}}^2 &= (2\pi)^{-d} \int_{\mathbb{R}^d} \frac{|\mathcal{F}[f](\boldsymbol{\omega})|^2}{\mathcal{F}[\phi_{X,\gamma}](\boldsymbol{\omega})} d\boldsymbol{\omega} \\ &= (2\pi)^{-d} \pi^{d/2} \int_{\mathbb{R}^d} |\mathcal{F}[f](\boldsymbol{\omega})|^2 \left(\prod_{i=1}^d \gamma_i \right)^{-1} \exp\left(\frac{\|\boldsymbol{\omega} \odot \boldsymbol{\gamma}\|_2^2}{4}\right) d\boldsymbol{\omega} \\ &= 2^{-d} \pi^{-d/2} \left(\prod_{i=1}^d \gamma_i \right)^{-1} \int |\mathcal{F}[f](\boldsymbol{\omega})|^2 \exp\left(\frac{\|\boldsymbol{\omega} \odot \boldsymbol{\gamma}\|_2^2}{4}\right) d\boldsymbol{\omega}\end{aligned}$$

whence the Lemma follows. \square

Lemma D.9. Suppose that Assumption 4.2 in the main text holds, and suppose $\gamma_x^{\delta_x} = \gamma_o^{\delta_o}$ for some positive constant δ_x, δ_o . Suppose $f \in \mathcal{H}_{\gamma_x, \gamma_o}$ with $\|f\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 \leq Cn$ for some constant $C > 0$. Then,

$$\begin{aligned}\|f\|_{L^2(P_{XO})} &\lesssim \gamma_x^{-d_x \eta_0} (\log(\gamma_x^{-\delta_x}))^{\frac{d_x \eta_0}{2}} \|Tf\|_{L^2(P_{ZO})} \\ &\quad + \gamma_x^{-d_x \eta_0} (\log(\gamma_x^{-\delta_x}))^{\frac{d_x \eta_0}{2}} \gamma_x^{\frac{1}{8} \delta_x} (n \gamma_x^{d_x} \gamma_o^{d_o})^{\frac{1}{2}}.\end{aligned}\tag{96}$$

In particular, for

$$\gamma_x = n^{-\frac{\frac{1}{d_x}}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}}, \quad \gamma_o = n^{-\frac{\frac{1}{s_o}(\frac{s_x}{d_x}+\eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}}.$$

and $\frac{1}{8} \delta_x = 2s_x + d_x \eta_1 + d_x \eta_0$, $\delta_x = \delta_o \frac{d_x}{s_o} (\frac{s_x}{d_x} + \eta_1)$. Then, we have

$$\|f\|_{L^2(P_{XO})} \lesssim \gamma_x^{-d_x \eta_0} (\log n)^{\frac{d_x \eta_0}{2}} \|Tf\|_{L^2(P_{ZO})} + n^{-\frac{\frac{s_x}{d_x}}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}}.$$

Proof. Define the following sets

$$\begin{aligned}I_{\mathbf{x}} &:= \left\{ \boldsymbol{\omega}_x \in \mathbb{R}^{d_x} : \|\boldsymbol{\omega}_x\|_2^2 \gamma_x^2 (\log(\gamma_x^{-\delta_x}))^{-1} \leq 1 \right\} \\ I_{\mathbf{o}} &:= \left\{ \boldsymbol{\omega}_o \in \mathbb{R}^{d_o} : \|\boldsymbol{\omega}_o\|_2^2 \gamma_o^2 (\log(\gamma_o^{-\delta_o}))^{-1} \leq 1 \right\} \\ I &:= I_{\mathbf{x}} \times I_{\mathbf{o}} \\ \tilde{I} &:= \left\{ (\boldsymbol{\omega}_x, \boldsymbol{\omega}_o) \in \mathbb{R}^{d_x+d_o} : \|\boldsymbol{\omega}_x\|_2^2 \gamma_x^2 (\log(\gamma_x^{-\delta_x}))^{-1} + \|\boldsymbol{\omega}_o\|_2^2 \gamma_o^2 (\log(\gamma_o^{-\delta_o}))^{-1} \leq 1 \right\}.\end{aligned}$$

Note that $\tilde{I} \subset I$. Let $\iota_x : \mathbb{R}^{d_x} \rightarrow \{0, 1\}$ denote the indicator function on I_x , let $\iota_o : \mathbb{R}^{d_o} \rightarrow \{0, 1\}$ denote the indicator function on I_o , let $\iota : \mathbb{R}^{d_x+d_o} \rightarrow \{0, 1\}$ denote the indicator function on I , which satisfies $\iota = \iota_x \cdot \iota_o$. Thus we have

$$\begin{aligned}
& \|f\|_{L^2(P_{XO})} \\
& \leq \|f * \mathcal{F}^{-1}(\iota)\|_{L^2(P_{XO})} + \|f * \mathcal{F}^{-1}(\iota) - f\|_{L^2(P_{XO})} \\
& \stackrel{(i)}{\leq} \gamma_x^{-d_x \eta_0} (\log(\gamma_x^{-\delta_x}))^{\frac{d_x \eta_0}{2}} \|T(f * \mathcal{F}^{-1}(\iota))\|_{L^2(P_{ZO})} + \|f * \mathcal{F}^{-1}(\iota) - f\|_{L^2(P_{XO})} \\
& \stackrel{(ii)}{\leq} \gamma_x^{-d_x \eta_0} (\log(\gamma_x^{-\delta_x}))^{\frac{d_x \eta_0}{2}} \|Tf\|_{L^2(P_{ZO})} \\
& \quad + \left(1 + \gamma_x^{-d_x \eta_0} (\log(\gamma_x^{-\delta_x}))^{\frac{d_x \eta_0}{2}}\right) \|f * \mathcal{F}^{-1}(\iota) - f\|_{L^2(P_{XO})} \\
& \lesssim \gamma_x^{-d_x \eta_0} (\log(\gamma_x^{-\delta_x}))^{\frac{d_x \eta_0}{2}} \|Tf\|_{L^2(P_{ZO})} + \gamma_x^{-d_x \eta_0} (\log(\gamma_x^{-\delta_x}))^{\frac{d_x \eta_0}{2}} \|f * \mathcal{F}^{-1}(\iota) - f\|_{L^2(\mathcal{X} \times \mathcal{O})}. \tag{97}
\end{aligned}$$

In the above derivations,

- (i) follows from [Assumption 4.2](#), and for $\mathcal{F}_x, \mathcal{F}_o$ defined in the proof of [Lemma D.7](#),

$$(\forall \mathbf{o} \in \mathcal{O}), \text{ supp}(\mathcal{F}_x[(f * \mathcal{F}^{-1}(\iota))(\cdot, \mathbf{o})]) = \text{ supp}(\mathcal{F}_x[(f * \mathcal{F}_o^{-1}[\iota_o])(\cdot, \mathbf{o})] \cdot \iota_x) \subseteq I_x,$$

where we note that the relevant Fourier transforms exist since $f \in \mathcal{H}_{\gamma_x, \gamma_o} \subset L^2(\mathbb{R}^{d_x+d_o})$.

- (ii) follows from a triangular inequality and a Jensen's inequality.

Next, we bound $(*) = \|f * \mathcal{F}^{-1}(\iota) - f\|_{L^2([0,1]^{d_x+d_o})}$. We have

$$\begin{aligned}
(*) &= \|f * \mathcal{F}^{-1}(\iota) - f\|_{L^2([0,1]^{d_x+d_o})} \leq \|f * \mathcal{F}^{-1}(\iota) - f\|_{L^2(\mathbb{R}^{d_x+d_o})} \\
&\stackrel{(a)}{=} \|\mathcal{F}[f * \mathcal{F}^{-1}(\iota) - f]\|_{L^2(\mathbb{R}^{d_x+d_o})} = \|\mathcal{F}[f]\iota - \mathcal{F}[f]\|_{L^2(\mathbb{R}^{d_x+d_o})} \\
&= \left(\int_{I^c} |\mathcal{F}[f](\omega)|^2 d\omega \right)^{\frac{1}{2}} \stackrel{(b)}{\leq} \left(\int_{\tilde{I}^c} |\mathcal{F}[f](\omega)|^2 d\omega \right)^{\frac{1}{2}},
\end{aligned}$$

where (a) holds by Plancherel's Theorem, (b) holds by $\tilde{I} \subseteq I$. Recall that for $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, $\mathbf{a} \odot \mathbf{b} := (a_1 b_1, \dots, a_d b_d) \in \mathbb{R}^d$. For $\omega = (\omega_x, \omega_o)$, we proceed from above to have,

$$\begin{aligned}
&\leq \left(\int_{\tilde{I}^c} |\mathcal{F}[f](\omega)|^2 \exp\left(\frac{\|(\omega_x \odot \gamma_x, \omega_o \odot \gamma_o)\|_2^2}{4}\right) \exp\left(-\frac{\|(\omega_x \odot \gamma_x, \omega_o \odot \gamma_o)\|_2^2}{4}\right) d\omega \right)^{\frac{1}{2}} \\
&\leq \sup_{\omega \in \tilde{I}^c} \exp\left(-\frac{\|(\omega_x \odot \gamma_x, \omega_o \odot \gamma_o)\|_2^2}{8}\right) \left(\int_{\mathbb{R}^{d_x+d_o}} |\mathcal{F}[f](\omega)|^2 \exp\left(\frac{\|(\omega_x \odot \gamma_x, \omega_o \odot \gamma_o)\|_2^2}{4}\right) d\omega \right)^{\frac{1}{2}} \\
&\stackrel{(c)}{\leq} \sup_{\omega \in \tilde{I}^c} \exp\left(-\frac{\|(\omega_x \odot \gamma_x, \omega_o \odot \gamma_o)\|_2^2}{8}\right) \left(\|f\|_{\mathcal{H}_{XO, \gamma_x, \gamma_o}}^2 2^{-\frac{d_x+d_o}{2}} \gamma_x^{d_x} \gamma_o^{d_o} \right)^{\frac{1}{2}} \\
&\lesssim \sup_{\|\omega_x\|_2^2 \gamma_x^2 (\log(\gamma_x^{-\delta_x}))^{-1} + \|\omega_o\|_2^2 \gamma_o^2 (\log(\gamma_o^{-\delta_o}))^{-1} \geq 1} \exp\left(-\frac{\gamma_x^2 \|\omega_x\|_2^2 + \gamma_o^2 \|\omega_o\|_2^2}{8}\right) (n \gamma_x^{d_x} \gamma_o^{d_o})^{\frac{1}{2}} \\
&\leq \exp\left(-\frac{1}{8} \log(\gamma_x^{-\delta_x})\right) (n \gamma_x^{d_x} \gamma_o^{d_o})^{\frac{1}{2}} \\
&\asymp \gamma_x^{\frac{1}{8} \delta_x} (n \gamma_x^{d_x} \gamma_o^{d_o})^{1/2},
\end{aligned}$$

where (c) holds by [Lemma D.8](#). Now we have proved the first claim of this lemma. Next, we proceed to prove the second claim by plugging in the specific values of δ_x, δ_o and γ_x, γ_o . Recall

that $\frac{1}{8}\delta_x = 2s_x + d_x\eta_1 + d_x\eta_0 + d_x$, we obtain

$$(*) \lesssim n^{-\frac{\delta_x}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}} \cdot n^{\frac{\frac{s_x}{d_x}+\eta_1}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}} = n^{\frac{-\frac{s_x}{d_x}-\eta_0-1}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}}.$$

We plug it back to Eq. (97) and obtain

$$\begin{aligned} \|f\|_{L^2(P_{XO})} &\lesssim \gamma_x^{-d_x\eta_0} (\log(\gamma_x^{-\delta_x}))^{\frac{d_x\eta_0}{2}} \|Tf\|_{L^2(P_{ZO})} \\ &\quad + \gamma_x^{-d_x\eta_0} (\log(\gamma_x^{-\delta_x}))^{\frac{d_x\eta_0}{2}} n^{\frac{-\frac{s_x}{d_x}-\eta_0-1}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}} \\ &\lesssim \gamma_x^{-d_x\eta_0} (\log n)^{\frac{d_x\eta_0}{2}} \|Tf\|_{L^2(P_{ZO})} + n^{\frac{\frac{s_x}{d_x}+1}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}} (\log n)^{\frac{d_x\eta_0}{2}} \\ &\lesssim \gamma_x^{-d_x\eta_0} (\log n)^{\frac{d_x\eta_0}{2}} \|Tf\|_{L^2(P_{ZO})} + n^{\frac{\frac{s_x}{d_x}}{1+2(\frac{s_x}{d_x}+\eta_1)+\frac{d_o}{s_o}(\frac{s_x}{d_x}+\eta_1)}}. \end{aligned}$$

So the proof is concluded. \square

Lemma D.10. Suppose Assumption 2.2 hold, and let

$$\gamma_x = n^{-\frac{\frac{1}{d_x}}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}}, \quad \gamma_o = n^{-\frac{\frac{1}{s_o}(\frac{s_x}{d_x}+\eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}}.$$

Then, for sufficiently large $n \geq 1$, we have

$$\|C_{FO}\| \geq a \left(\frac{\sqrt{\pi}}{4} \right)^{\frac{d_x+d_o}{2}} n^{-\frac{1}{2} \frac{1+\frac{d_o}{s_o}(\frac{s_x}{d_x}+\eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}}.$$

Proof. By definition of the operator norm, we have

$$\begin{aligned} &\|C_{FO}\|^2 \\ &= \sup_{\|f\|_{\mathcal{H}_{\gamma_x, \gamma_o}}=1} \langle f, \mathbb{E}[(F_*(Z, O) \otimes \phi_{\gamma_o}(O)) \otimes (F_*(Z, O) \otimes \phi_{\gamma_o}(O))]f \rangle_{\mathcal{H}_{\gamma_x, \gamma_o}} \\ &= \sup_{\|f\|_{\mathcal{H}_{\gamma_x, \gamma_o}}=1} \mathbb{E} \left[(\mathbb{E}[f(X, O)|Z, O])^2 \right] \\ &\geq \sup_{\|f\|_{\mathcal{H}_{\gamma_x, \gamma_o}}=1} (\mathbb{E}_{ZO \sim P_{ZO}} [\mathbb{E}[f(X, O)|Z, O]])^2 \\ &= \sup_{\|f\|_{\mathcal{H}_{\gamma_x, \gamma_o}}=1} \mathbb{E}[f(X, O)]^2 \\ &= \|\mu_{XO}\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2, \end{aligned}$$

where $\mu_{XO} := \mathbb{E}_{P_{XO}}[\phi_{\gamma_x}(X) \otimes \phi_{\gamma_o}(O)]$. Notice that

$$\begin{aligned} &\|\mu_{XO}\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 \\ &= \left\langle \iint_{\mathcal{O} \times \mathcal{X}} \phi_{\gamma_x}(\mathbf{x}) \otimes \phi_{\gamma_o}(\mathbf{o}) p_{XO}(\mathbf{x}, \mathbf{o}) \, d\mathbf{x} \, d\mathbf{o}, \right. \\ &\quad \left. \iint_{\mathcal{O} \times \mathcal{X}} \phi_{\gamma_x}(\mathbf{x}') \otimes \phi_{\gamma_o}(\mathbf{o}') p_{XO}(\mathbf{x}', \mathbf{o}') \, d\mathbf{x}' \, d\mathbf{o}' \right\rangle_{\mathcal{H}_{\gamma_x, \gamma_o}} \end{aligned}$$

$$\begin{aligned}
&= \iint_{\mathcal{O} \times \mathcal{X}} \iint_{\mathcal{O} \times \mathcal{X}} K_{\gamma_o}(\mathbf{o}, \mathbf{o}') K_{\gamma_x}(\mathbf{x}, \mathbf{x}') p_{XO}(\mathbf{x}, \mathbf{o}) p_{XO}(\mathbf{x}', \mathbf{o}') d\mathbf{x} d\mathbf{x}' d\mathbf{o} d\mathbf{o}' \\
&\stackrel{(a)}{\geq} a^2 \int_{[1/4, 3/4]^{d_o}} \int_{[1/4, 3/4]^{d_o}} K_{\gamma_o}(\mathbf{o}, \mathbf{o}') d\mathbf{o} d\mathbf{o}' \int_{[1/4, 3/4]^{d_x}} \int_{[1/4, 3/4]^{d_x}} K_{\gamma_x}(\mathbf{x}, \mathbf{x}') d\mathbf{x} d\mathbf{x}' \\
&\stackrel{(b)}{=} a^2 \left(\gamma_x \sqrt{\pi} \left[\frac{\operatorname{erf}\left(\frac{1}{2\gamma_x}\right)}{2} + \gamma_x \frac{\exp\left(-\frac{1}{4\gamma_x^2}\right) - 1}{\sqrt{\pi}} \right] \right)^{d_x} \\
&\quad \cdot \left(\gamma_o \sqrt{\pi} \left[\frac{\operatorname{erf}\left(\frac{1}{2\gamma_o}\right)}{2} + \gamma_o \frac{\exp\left(-\frac{1}{4\gamma_o^2}\right) - 1}{\sqrt{\pi}} \right] \right)^{d_o},
\end{aligned}$$

where $\operatorname{erf}(x) := \frac{1}{\sqrt{\pi}} \int_{-x}^x \exp(-t^2) dt$ is the standard error function of normal distribution. Step (a) holds by [Assumption 2.2](#) and step (b) holds by using [Lemma D.11](#). We plug in $\gamma_x = n^{-\frac{1}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}}$, $\gamma_o = n^{-\frac{\frac{1}{s_o}(\frac{s_x}{d_x}+\eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}}$ to obtain

$$\|\mu_{XO}\|_{\mathcal{H}_{\gamma_x, \gamma_o}}^2 \stackrel{(i)}{\geq} a^2 (\sqrt{\pi} \gamma_x / 4)^{d_x} (\sqrt{\pi} \gamma_o / 4)^{d_o} = a^2 (\sqrt{\pi} / 4)^{d_x+d_o} n^{-\frac{1+\frac{d_o}{s_o}(\frac{s_x}{d_x}+\eta_1)}{1+(2+\frac{d_o}{s_o})(\frac{s_x}{d_x}+\eta_1)}},$$

where (i) holds for sufficiently large $n \geq 1$ because $\operatorname{erf}(x) \geq \frac{1}{2}$ when $x \geq 1$. \square

Lemma D.11. *Let $k_\gamma : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be the Gaussian kernel with length scale $\gamma \in (0, 1]$. Then we have,*

$$\int_{[1/4, 3/4]^d} \int_{[1/4, 3/4]^d} k_\gamma(\mathbf{x}, \mathbf{x}') d\mathbf{x} d\mathbf{x}' = \left(\gamma \sqrt{\pi} \left[\frac{\operatorname{erf}\left(\frac{1}{2\gamma}\right)}{2} + \gamma \frac{\exp\left(-\frac{1}{4\gamma^2}\right) - 1}{\sqrt{\pi}} \right] \right)^{d_x}.$$

Here, $\operatorname{erf}(x) := \frac{1}{\sqrt{\pi}} \int_{-x}^x \exp(-t^2) dt$ is the standard error function.

Proof. Since $k_\gamma(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^d \exp\left(-\frac{(x_i - x'_i)^2}{\gamma^2}\right)$, it suffices to prove the following result

$$\int_{1/4}^{3/4} \int_{1/4}^{3/4} k_\gamma(x_i, x'_i) dx_i dx'_i = \gamma^2 \sqrt{\pi} \left[\frac{\operatorname{erf}\left(\frac{1}{2\gamma}\right)}{2\gamma} + \frac{\exp\left(-\frac{1}{4\gamma^2}\right) - 1}{\sqrt{\pi}} \right].$$

Notice that

$$\begin{aligned}
&\int_{1/4}^{3/4} \exp\left(-\frac{(x_i - x'_i)^2}{\gamma^2}\right) dx_i \\
&= \gamma \int_{(1/4 - x'_i)/\gamma}^{(3/4 - x'_i)/\gamma} \exp(-x^2) dx \\
&= \gamma \int_{-\infty}^{(3/4 - x'_i)/\gamma} \exp(-x^2) dx - \gamma \int_{-\infty}^{(1/4 - x'_i)/\gamma} \exp(-x^2) dx \\
&= \gamma \frac{\sqrt{\pi}}{2} \left(1 + \operatorname{erf}\left(\frac{3/4 - x'_i}{\gamma}\right) \right) - \gamma \frac{\sqrt{\pi}}{2} \left(1 - \operatorname{erf}\left(\frac{x'_i - 1/4}{\gamma}\right) \right)
\end{aligned}$$

$$= \gamma \frac{\sqrt{\pi}}{2} \left(\operatorname{erf} \left(\frac{3/4 - x'_i}{\gamma} \right) + \operatorname{erf} \left(\frac{x'_i - 1/4}{\gamma} \right) \right).$$

Therefore, we have

$$\begin{aligned} & \int_{1/4}^{3/4} \int_{1/4}^{3/4} \exp \left(-\frac{(x_i - x'_i)^2}{\gamma^2} \right) dx_i dx'_i \\ &= \gamma \frac{\sqrt{\pi}}{2} \int_{1/4}^{3/4} \operatorname{erf} \left(\frac{3/4 - x'_i}{\gamma} \right) + \operatorname{erf} \left(\frac{x'_i - 1/4}{\gamma} \right) dx'_i \\ &= \gamma^2 \sqrt{\pi} \int_0^{\frac{1}{2\gamma}} \operatorname{erf}(y) dy \\ &= \gamma^2 \sqrt{\pi} \left[\frac{\operatorname{erf} \left(\frac{1}{2\gamma} \right)}{2\gamma} + \frac{\exp \left(-\frac{1}{4\gamma^2} \right) - 1}{\sqrt{\pi}} \right]. \end{aligned}$$

where the last inequality holds by using the identity

$$\int \operatorname{erf}(x) dx = x \cdot \operatorname{erf}(x) + e^{-x^2}/\sqrt{\pi} + C.$$

□

E Proof of Theorem 4.2 in the main text

E.1 Relationship Between the NPIR-O Model and the NPIV-O Model

Following Chen and Reiss [2011, Section 3], we first establish that the NPIV-O model is no more informative than the reduced form nonparametric indirect regression with observed confounders (NPIR-O) model.

Definition 16 (Restricted NPIV-O model). *Let $\sigma_0 > 0$ be a finite constant. Recall \mathfrak{S} as defined in Assumption 2.3 in the main text. Let \mathcal{C} be a set of elements $(P_{\epsilon ZXO}, f)$ such that the following property holds: $(\forall f \in \mathfrak{S}) (\exists P_{\epsilon ZXO}, f) \in \mathcal{C}$ such that P_{ZY} is determined by $P_{\epsilon ZX}$ and f , and that*

$$Y_i - \mathbb{E}[Y_i | Z = \mathbf{z}_i, O = \mathbf{o}_i] = f(\mathbf{z}_i, \mathbf{o}_i) + \epsilon_i - (Tf)(\mathbf{z}_i, \mathbf{o}_i)$$

given $Z_i = \mathbf{z}_i, O_i = \mathbf{o}_i$ is $\mathcal{N}(0, \sigma^2(\mathbf{z}_i, \mathbf{o}_i))$ -distributed with $\sigma^2(\mathbf{z}_i, \mathbf{o}_i) \geq \sigma_0^2$.

For an NPIV-O model as defined in Definition 16, we specify the reduced form NPIR-O model as

$$Y_i = (Tf)(Z_i, O_i) + v_i, \quad i = 1, \dots, n$$

with (Z_i, O_i, v_i) i.i.d., $P_{v_i | Z_i = \mathbf{z}_i, O_i = \mathbf{o}_i} = \mathcal{N}(0, \sigma^2(\mathbf{z}_i, \mathbf{o}_i))$, $f \in \mathfrak{S}$ the unknown structural function, and $T : L^2(P_{XO}) \rightarrow L^2(P_{ZO})$ a known operator satisfying Assumption 4.1. The observations corresponding to the NPIR are $\{(Y_i, \mathbf{z}_i, \mathbf{o}_i)\}_{i=1}^n$.

Definition 17 (NPIR-O model class). *Let \mathcal{C} be as defined in Definition 16. The NPIR-O model class \mathcal{C}_0 consists of all model parameters $(P_{Z' O'}, \sigma(\cdot, \cdot), f)$ such that $(\exists (P_{\epsilon ZX}, f) \in \mathcal{C})$ with the following properties: $P_{ZO} = P_{Z' O'}$, $\sigma^2(\mathbf{z}, \mathbf{o}) \geq \sigma_0^2 > 0$, the conditional law $P_{X|Z,O}$ is prescribed according to T , and $P_{\epsilon|ZOX}$ is arbitrary among the conditions imposed in \mathcal{C} .*

The following Lemma is Chen [2007, Lemma 1], by augmenting relevant variables to include observed confounders O .

Lemma E.1. *The NPIR-O model is more informative than the NPIV-O model in the sense that for each estimator \hat{f}_n for the NPIV-O model, there is an estimator \tilde{f}_n for the NPIR-O model with*

$$\sup_{(P_{ZO}, \sigma(\cdot, \cdot), f) \in \mathcal{C}_0} \mathbb{E}_{(P_{ZO}, \sigma(\cdot, \cdot), f)} [\|\tilde{f}_n - f\|_{L^2(P_{XO})}^2] \leq \sup_{(P_{\epsilon ZXO}, f) \in \mathcal{C}} \mathbb{E}_{(P_{\epsilon ZXO}, f)} [\|\hat{f}_n - f\|_{L^2(P_{XO})}^2].$$

In this section, we provide a lower bound for the NPIR-O model class defined in Definition 17, which by the above discussion implies a lower bound for the (restricted) NPIV-O model class defined in Definition 16.

E.2 The Lower Bound for NPIR-O Model

Step One. We take \mathfrak{m} to be the smallest even integer such that $\mathfrak{m} > s_x \vee s_o$. To help us construct $f_{\mathbf{v}}$, we need to introduce several functions. We define

$$M_{\mathfrak{K}, -\frac{\mathfrak{m}}{2}}(\mathbf{x}) := \prod_{i=1}^{d_x} \iota_{\mathfrak{m}} \left(2^{\lfloor \mathfrak{K} \frac{s}{s_x} \rfloor} x_i + \frac{\mathfrak{m}}{2} \right), \quad M_{\mathfrak{K}, \ell_o}(\mathbf{o}) := \prod_{j=1}^{d_o} \iota_{\mathfrak{m}} \left(2^{\lfloor \mathfrak{K} \frac{s}{s_o} \rfloor} o_j - \ell_{o,j} \right). \quad (98)$$

Define

$$\mathcal{L} := \left\{ (\ell_x, \ell_o) : \ell_x \in \left\{ 0, 1, \dots, \left\lfloor \frac{0.8\pi}{\zeta} 2^{\frac{\mathfrak{K}s}{s_x}} \right\rfloor \right\}^{d_x}, \ell_o \in \left(m\mathbb{Z} \cap \left\{ 1, \dots, 2^{\lfloor \mathfrak{K} \frac{s}{s_o} \rfloor} \right\} \right)^{d_o} \right\}, \quad (99)$$

then we can compute the size of \mathcal{L} as

$$|\mathcal{L}| \asymp \zeta^{-d_x} 2^{\mathfrak{K}s \left(\frac{d_x}{s_x} + \frac{d_o}{s_o} \right)} \quad (100)$$

For $(\ell_x, \ell_o) \in \mathcal{L}$, define $\mathbb{1}_{\ell_x}$ as an indicator function of the set

$$I_{\ell_x} := \bigtimes_{j=1}^{d_x} \left[1.1\pi + \zeta \ell_{x,j} 2^{-\mathfrak{K} \frac{s}{s_x}}, 1.1\pi + (\zeta \ell_{x,j} + 1) 2^{-\mathfrak{K} \frac{s}{s_x}} \right], \quad (101)$$

Next, we define

$$\Omega_{\mathfrak{K}, \ell_x}(\mathbf{x}) := \left(M_{0, -\frac{\mathfrak{m}}{2}} * \mathcal{F}^{-1}[\mathbb{1}_{\ell_x}] \right) \left(2^{\frac{\mathfrak{K}s}{s_x}} \mathbf{x} \right), \quad \Omega_{\mathfrak{K}(\ell_x, \ell_o)}(\mathbf{x}, \mathbf{o}) := \Omega_{\mathfrak{K}, \ell_x}(\mathbf{x}) M_{\mathfrak{K}, \ell_o}(\mathbf{o}). \quad (102)$$

Now we are ready to define $f_{\mathbf{v}}$. Note that a vector $\mathbf{v} \in \{0, 1\}^{|\mathcal{L}|}$ canonically associates to each point (ℓ_x, ℓ_o) a value $\beta_{\mathbf{v}(\ell_x, \ell_o)} \in \{0, 1\}$. We define

$$f_{\mathbf{v}} := \epsilon_0 2^{-\mathfrak{K}s \left(1 - \frac{d_x}{2s_x} \right)} \sum_{\ell_x, \ell_o} \beta_{\mathbf{v}(\ell_x, \ell_o)} \Omega_{\mathfrak{K}(\ell_x, \ell_o)}, \quad (103)$$

where $\epsilon_0 > 0$ is a fixed scalar to be chosen later, to ensure that $\|f_{\mathbf{v}}\|_{B_{2, \infty}^{s_x, s_o}(\mathbb{R}^{d_x+d_o})} \leq 1$, thus ensuring that $f_{\mathbf{v}}$ satisfies Assumption 2.3. The function class $\mathfrak{F} := \{f_{\mathbf{v}}, \mathbf{v} \in \{0, 1\}^{|\mathcal{L}|}\}$. For each $f_{\mathbf{v}}$, consider the joint data generating distribution P_{ZXOY} specified as follows:

1. The marginal distribution P_{ZO} is the product of independent distributions P_Z and P_O , where both are uniform distributions on $[0, 1]^{d_z}$ and $[0, 1]^{d_o}$.

2. The marginal distribution P_{XO} is the product of independent distributions P_X and P_O , where P_O is the uniform distribution on $[0, 1]^{d_o}$, and P_X is supported on $[-1/2, 1/2]^{d_x}$ and admits the following density function:

$$p_X(\mathbf{x}) \propto \prod_{i=1}^{d_x} g(x_i), \quad g(x_i) := \exp\left(-\frac{2}{1-4x_i^2}\right) \mathbb{1}_{x_i \in [-1/2, 1/2]}. \quad (104)$$

3. The conditional distribution $P_{X|Z,O}$ satisfies Assumption 2.2 and it induces an operator $T : L^2(P_{XO}) \rightarrow L^2(P_{ZO})$ that satisfies Assumption 4.3.
4. The conditional distribution $P_{Y|Z=\mathbf{z}, O=\mathbf{o}}$ is a Gaussian distribution $\mathcal{N}((Tf_{\mathbf{v}})(\mathbf{z}, \mathbf{o}), \sigma^2)$ for any $\mathbf{z} \in \mathcal{Z}$ and $\mathbf{o} \in \mathcal{O}$.

Step Two. Now we are going to present some properties of the basis $\Omega_{\mathfrak{K}(\ell_x, \ell_o)}$, which will be used later on to prove some properties of $f_{\mathbf{v}}$. By Young's Convolution Theorem, we have

$$\left\| M_{0, -\frac{\mathfrak{m}}{2}} * \mathcal{F}^{-1}[\mathbb{1}_{\ell_x}] \right\|_{L^2(\mathbb{R}^{d_x})} \leq \|M_{0, -\frac{\mathfrak{m}}{2}}\|_{L^1(\mathbb{R}^{d_x})} \|\mathbb{1}_{\ell_x}\|_{L^2(\mathbb{R}^{d_x})} < \infty,$$

hence its Fourier transform is well-defined. We have

$$\begin{aligned} \mathcal{F}[\Omega_{\mathfrak{K}, \ell_x}](\boldsymbol{\omega}_x) &= 2^{-\mathfrak{K} \frac{s d_x}{s_x}} \mathcal{F} \left[M_{0, -\frac{\mathfrak{m}}{2}} * \mathcal{F}^{-1}[\mathbb{1}_{\ell_x}] \right] \left(2^{-\frac{\mathfrak{K}_s}{s_x}} \boldsymbol{\omega}_x \right) \\ &= 2^{-\mathfrak{K} \frac{s d_x}{s_x}} \cdot \mathcal{F} \left[M_{0, -\frac{\mathfrak{m}}{2}} \right] \left(2^{-\frac{\mathfrak{K}_s}{s_x}} \boldsymbol{\omega}_x \right) \cdot \mathbb{1}_{\ell_x} \left(2^{-\frac{\mathfrak{K}_s}{s_x}} \boldsymbol{\omega}_x \right) \\ &= \mathcal{F} \left[M_{\mathfrak{K}, -\frac{\mathfrak{m}}{2}} \right] (\boldsymbol{\omega}_x) \cdot \mathbb{1}_{\ell_x} \left(2^{-\frac{\mathfrak{K}_s}{s_x}} \boldsymbol{\omega}_x \right). \end{aligned} \quad (105)$$

Also, since $\mathcal{F}[\iota_m](\omega) = \exp(-\mathfrak{m}i\omega/2) \cdot \sin(\omega/2)^{\mathfrak{m}} \cdot (\omega/2)^{-\mathfrak{m}}$, we have

$$\mathcal{F} \left[M_{0, -\frac{\mathfrak{m}}{2}} \right] (\boldsymbol{\omega}_x) = \prod_{i=1}^{d_x} \frac{\sin(\omega_{i,x}/2)^{\mathfrak{m}}}{(\omega_{i,x}/2)^{\mathfrak{m}}}. \quad (106)$$

Now we can see that we pick the location vector to be $-\frac{\mathfrak{m}}{2}$ such that the Fourier transform of $M_{0, -\frac{\mathfrak{m}}{2}}$ is a real valued function. Note that since $\zeta \geq 1$, we have $(\zeta \ell_{x,j} + 1)2^{-\mathfrak{K} \frac{s}{s_x}} \leq \zeta(\ell_{x,j} + 1)2^{-\mathfrak{K} \frac{s}{s_x}}$ for $j \in \{1, \dots, d_x\}$, hence $I_{\ell_x} \cap I_{\ell'_x} = \emptyset$ if $\ell_x \neq \ell'_x$ which means that the support of $\mathcal{F}[\Omega_{\mathfrak{K}, \ell_x}]$ is disjoint for $\ell_x \neq \ell'_x$. Also, note that

$$\text{supp}(M_{\mathfrak{K}, \ell_o}) = \bigtimes_{j=1}^{d_o} [\ell_{o,j} 2^{-\mathfrak{K} \frac{s}{s_o}}, (\ell_{o,j} + \mathfrak{m}) 2^{-\mathfrak{K} \frac{s}{s_o}}]$$

and $\ell_{o,j}$ are multiples of \mathfrak{m} by definition of \mathcal{L} in Eq. (99), we have $\text{supp}(M_{\mathfrak{K}, \ell_o}) \cap \text{supp}(M_{k, \ell'_o}) = \emptyset$ for any $\ell_o \neq \ell'_o$.

Step Three. In this step, we are going to prove that $f_{\mathbf{v}} \in B_{2,\infty}^{s_x, s_o}(\mathbb{R}^{d_x + d_o})$ and its Besov norm is bounded by ϵ_0 . We have, ($\forall \mathbf{o} \in \mathcal{O}$),

$$\begin{aligned} \mathcal{F}[f_{\mathbf{v}}(\cdot, \mathbf{o})](\boldsymbol{\omega}_x) &= \epsilon_0 2^{-\mathfrak{K} \frac{s}{2s_x} \left(1 - \frac{d_x}{2s_x} \right)} \sum_{\ell_x, \ell_o} \beta_{\mathbf{v}(\ell_x, \ell_o)} \mathcal{F}[\Omega_{\mathfrak{K}(\ell_x, \ell_o)}(\cdot, \mathbf{o})](\boldsymbol{\omega}_x) \\ &= \epsilon_0 2^{-\mathfrak{K} \frac{s}{2s_x} \left(1 - \frac{d_x}{2s_x} \right)} \sum_{\ell_x, \ell_o} \beta_{\mathbf{v}(\ell_x, \ell_o)} \mathcal{F}[\Omega_{\mathfrak{K}, \ell_x}](\boldsymbol{\omega}_x) \cdot M_{\mathfrak{K}, \ell_o}(\mathbf{o}) \end{aligned}$$

$$\begin{aligned}
&= \epsilon_0 2^{-\tilde{\kappa}_s(1-\frac{d_x}{2s_x})} \sum_{\ell_x, \ell_o} \beta_{\mathbf{v}(\ell_x, \ell_o)} 2^{-\tilde{\kappa} \frac{s d_x}{s_x}} \mathcal{F}[M_{0, -\frac{m}{2}}] \left(2^{-\tilde{\kappa} \frac{s}{s_x}} \boldsymbol{\omega}_x \right) \cdot \mathbb{1}_{\ell_x} \left(2^{-\tilde{\kappa} \frac{s}{s_x}} \boldsymbol{\omega}_x \right) \cdot M_{\tilde{\kappa}, \ell_o}(\mathbf{o}) \\
&= \epsilon_0 2^{-\tilde{\kappa}_s(1-\frac{d_x}{2s_x})} \sum_{\ell_o} \mathcal{F}[M_{\tilde{\kappa}, -\frac{m}{2}}](\boldsymbol{\omega}_x) \cdot M_{\tilde{\kappa}, \ell_o}(\mathbf{o}) \cdot \left(\sum_{\ell_x} \beta_{\mathbf{v}(\ell_x, \ell_o)} \mathbb{1}_{\ell_x} \left(2^{-\tilde{\kappa} \frac{s}{s_x}} \boldsymbol{\omega}_x \right) \right). \tag{107}
\end{aligned}$$

Note that, for fixed ℓ_o , $\boldsymbol{\omega}_x \mapsto \sum_{\ell_x} \beta_{\mathbf{v}(\ell_x, \ell_o)} \mathbb{1}_{\ell_x}(2^{-\tilde{\kappa} \frac{s}{s_x}} \boldsymbol{\omega}_x)$ is the indicator function on

$$\bigcup_{\ell_x: \beta_{\mathbf{v}(\ell_x, \ell_o)} > 0} \bigtimes_{j=1}^{d_x} \left[1.1\pi 2^{\tilde{\kappa} \frac{s}{s_x}} + \zeta \ell_{x,j}, 1.1\pi 2^{\tilde{\kappa} \frac{s}{s_x}} + (\zeta \ell_{x,j} + 1) \right],$$

The above observation as applied to the right hand side of Eq. (107) means that we can apply Lemma E.3 to obtain

$$\|f_{\mathbf{v}}\|_{B_{2,q}^{s_x, s_o}(\mathbb{R}^{d_x+d_o})} \lesssim \epsilon_0 2^{-\tilde{\kappa}_s(1-\frac{d_x}{2s_x})} \cdot 2^{\tilde{\kappa}_s(1-\frac{d_x}{2s_x})} = \epsilon_0. \tag{108}$$

Hence we can choose ϵ_0 to be a small scalar such that $\|f_{\mathbf{v}}\|_{B_{2,q}^{s_x, s_o}(\mathbb{R}^{d_x+d_o})} \leq 1$. The rest of the conditions that $f_{\mathbf{v}} \in L^\infty(\mathbb{R}^{d_x+d_o}) \cap L^1(\mathbb{R}^{d_x+d_o}) \cap C^0(\mathbb{R}^{d_x+d_o})$ are all trivial to verify. Therefore, $f_{\mathbf{v}}$ satisfies Assumption 2.3.

Step Four. By the Gilbert-Varshamov Bound Tsybakov [2008][Lemma 2.9], for $|\mathcal{L}| \geq 8$, there exists a subset $V_{\tilde{\kappa}}$ of $\{0, 1\}^{|\mathcal{L}|}$ such that $\mathbf{0} \in V_{\tilde{\kappa}}$ and

$$\sum_{\ell_x, \ell_o} |\beta_{\mathbf{v}(\ell_x, \ell_o)} - \beta_{\mathbf{v}'(\ell_x, \ell_o)}|^2 \geq \frac{|\mathcal{L}|}{8} \tag{109}$$

for $\mathbf{v} \neq \mathbf{v}' \in V_{\tilde{\kappa}}$ and $|V_{\tilde{\kappa}}| \geq 2^{\frac{|\mathcal{L}|}{8}}$. Note that since $\sqrt{p_X}$ is an even function over $[-1/2, 1/2]^{d_x}$, its Fourier transform q is a real valued function. Define

$$q(\boldsymbol{\omega}_x) := \mathcal{F}[\sqrt{p_X}](\boldsymbol{\omega}_x), \quad \tilde{q}(\boldsymbol{\omega}_x) := \mathcal{F}[\sqrt{p_X}](\boldsymbol{\omega}_x) \mathbb{1}_{[-\zeta/3, \zeta/3]^{d_x}}(\boldsymbol{\omega}_x). \tag{110}$$

For coefficients $\alpha_{\mathbf{v}(\ell_x, \ell_o)} \in \mathbb{R}$, we have

$$\begin{aligned}
&\left\| \sum_{\ell_x, \ell_o} \alpha_{\mathbf{v}(\ell_x, \ell_o)} \Omega_{\tilde{\kappa}(\ell_x, \ell_o)} \right\|_{L^2(P_{X,O})}^2 \\
&= \int_{[0,1]^{d_o}} \int_{\mathbb{R}^{d_x}} \sum_{\substack{\ell_x, \ell_o \\ \tilde{\ell}_x, \tilde{\ell}_o}} \alpha_{\mathbf{v}(\ell_x, \ell_o)} \alpha_{\mathbf{v}(\tilde{\ell}_x, \tilde{\ell}_o)} \Omega_{\tilde{\kappa}, \ell_x}(\mathbf{x}) \Omega_{k \tilde{\ell}_x}(\mathbf{x}) M_{\tilde{\kappa}, \ell_o}(\mathbf{o}) M_{k \tilde{\ell}_o}(\mathbf{o}) p_X(\mathbf{x}) d\mathbf{x} d\mathbf{o} \\
&= \int_{\mathbb{R}^{d_x}} \sum_{\substack{\ell_x, \ell_o \\ \tilde{\ell}_x, \tilde{\ell}_o}} \alpha_{\mathbf{v}(\ell_x, \ell_o)} \alpha_{\mathbf{v}(\tilde{\ell}_x, \tilde{\ell}_o)} \Omega_{\tilde{\kappa}, \ell_x}(\mathbf{x}) \Omega_{k \tilde{\ell}_x}(\mathbf{x}) p_X(\mathbf{x}) \left(\int_{[0,1]^{d_o}} M_{\tilde{\kappa}, \ell_o}(\mathbf{o}) M_{k \tilde{\ell}_o}(\mathbf{o}) d\mathbf{o} \right) d\mathbf{x} \\
&\stackrel{(a)}{=} \sum_{\ell_o} \|M_{\tilde{\kappa}, \ell_o}\|_{L^2(\mathbb{R}^{d_o})}^2 \int_{\mathbb{R}^{d_x}} \left(\sum_{\ell_x} \alpha_{\mathbf{v}(\ell_x, \ell_o)} \Omega_{\tilde{\kappa}, \ell_x}(\mathbf{x}) \sqrt{p(\mathbf{x})} \right)^2 d\mathbf{x} \\
&\stackrel{(b)}{=} \sum_{\ell_o} \|M_{\tilde{\kappa}, \ell_o}\|_{L^2(\mathbb{R}^{d_o})}^2 \int_{\mathbb{R}^{d_x}} \left(\sum_{\ell_x} \alpha_{\mathbf{v}(\ell_x, \ell_o)} \mathcal{F}[\Omega_{\tilde{\kappa}, \ell_x} \sqrt{p}](\boldsymbol{\omega}_x) \right)^2 d\boldsymbol{\omega}_x
\end{aligned}$$

$$= 2^{-\frac{\mathfrak{K}_s d_o}{s_o}} \|M_{00}\|_{L^2(\mathbb{R}^{d_o})}^2 \int_{\mathbb{R}^{d_x}} \sum_{\ell_o} \left(\sum_{\ell_x} \alpha_{\mathbf{v}(\ell_x, \ell_o)} \mathcal{F}[\Omega_{\mathfrak{K}, \ell_x} \sqrt{p}] (\omega_x) \right)^2 d\omega_x. \quad (111)$$

where (a) follows from the fact that $M_{\mathfrak{K}, \ell_o}$ have pairwise disjoint support verified in Step Two above; and (b) follows from the Plancherel's Theorem. By Eq. (111), we have

$$\begin{aligned} & \|f_{\mathbf{v}} - f_{\mathbf{v}'}\|_{L^2(P_{XO})}^2 \\ &= \epsilon_0 2^{-2\mathfrak{K}_s \left(1 - \frac{d_x}{2s_x}\right)} 2^{-\frac{\mathfrak{K}_s d_o}{s_o}} \|M_{00}\|_{L^2(\mathbb{R}^{d_o})}^2 \\ &\quad \cdot \int_{\mathbb{R}^{d_x}} \sum_{\ell_o} \left(\sum_{\ell_x} (\beta_{\mathbf{v}(\ell_x, \ell_o)} - \beta_{\mathbf{v}'(\ell_x, \ell_o)}) \mathcal{F}[\Omega_{\mathfrak{K}, \ell_x} \sqrt{p_X}] (\omega_x) \right)^2 d\omega_x \\ &\geq \epsilon_0 2^{-2\mathfrak{K}_s \left(1 - \frac{d_x}{2s_x}\right)} 2^{-\frac{\mathfrak{K}_s d_o}{s_o}} \|M_{00}\|_{L^2(\mathbb{R}^{d_o})}^2 ((A)^2/2 - (B)^2), \end{aligned} \quad (112)$$

where the last step follows from the reverse triangular inequality $(a + b)^2 \geq \frac{a^2}{2} - b^2$, and we define

$$\begin{aligned} (A)^2 &:= \sum_{\ell_o} \left\| \sum_{\ell_x} (\beta_{\mathbf{v}(\ell_x, \ell_o)} - \beta_{\mathbf{v}'(\ell_x, \ell_o)}) \mathcal{F}[\Omega_{\mathfrak{K}, \ell_x}] * \tilde{q} \right\|_{L^2(\mathbb{R}^{d_x})}^2 \\ (B)^2 &:= \sum_{\ell_o} \left\| \sum_{\ell_x} (\beta_{\mathbf{v}(\ell_x, \ell_o)} - \beta_{\mathbf{v}'(\ell_x, \ell_o)}) \mathcal{F}[\Omega_{\mathfrak{K}, \ell_x}] * (q - \tilde{q}) \right\|_{L^2(\mathbb{R}^{d_x})}^2. \end{aligned}$$

In order to lower bound Eq. (112), we need to lower bound $(A)^2$ and upper bound $(B)^2$. First, we are going to lower bound $(A)^2$. From Eq. (105), we know that the support of $\mathcal{F}[\Omega_{\mathfrak{K}, \ell_x}]$ is

$$\tilde{I}_{\ell_x} := \bigtimes_{j=1}^{d_x} \left[1.1\pi 2^{\mathfrak{K}_s \frac{s}{s_x}} + \zeta \ell_{x,j}, 1.1\pi 2^{\mathfrak{K}_s \frac{s}{s_x}} + (\zeta \ell_{x,j} + 1) \right]. \quad (113)$$

Since the support of \tilde{q} is $[-\zeta/3, \zeta/3]^{d_x}$, by the standard fact that $\text{supp}(f * g) \subseteq \text{supp}(f) + \text{supp}(g)$, we have

$$\begin{aligned} \text{supp}(\mathcal{F}[\Omega_{\mathfrak{K}, \ell_x}] * \tilde{q}) &\subseteq \bigtimes_{j=1}^{d_x} \left[1.1\pi 2^{\mathfrak{K}_s \frac{s}{s_x}} + \zeta(\ell_{x,j} - 1/3), 1.1\pi 2^{\mathfrak{K}_s \frac{s}{s_x}} + (\zeta(\ell_{x,j} + 1/3) + 1) \right] \\ &=: \Lambda_{\mathfrak{K}} \end{aligned} \quad (114)$$

Note that for $\zeta \geq 3$, we have $\zeta(\ell_j + 1/3) + 1 \leq \zeta(\ell_j + 1 - 1/3)$, hence $\text{supp}(\mathcal{F}[\Omega_{\mathfrak{K}, \ell_x}] * \tilde{q})$ is pairwise disjoint with respect to different ℓ_x . Hence we obtain that

$$(A)^2 = \sum_{\ell_o} \sum_{\ell_x} (\beta_{\mathbf{v}(\ell_x, \ell_o)} - \beta_{\mathbf{v}'(\ell_x, \ell_o)})^2 \|\mathcal{F}[\Omega_{\mathfrak{K}, \ell_x}] * \tilde{q}\|_{L^2(\mathbb{R}^{d_x})}^2.$$

Notice that

$$\begin{aligned} & \|\mathcal{F}[\Omega_{\mathfrak{K}, \ell_x}] * \tilde{q}\|_{L^2(\mathbb{R}^{d_x})}^2 \\ &= \int_{\Lambda_{\mathfrak{K}}} \left| \int_{\mathbb{R}^{d_x}} \mathcal{F}[\Omega_{\mathfrak{K}, \ell_x}](\omega'_x) \tilde{q}(\omega_x - \omega'_x) d\omega'_x \right|^2 d\omega_x \end{aligned}$$

$$\begin{aligned}
&= 2^{-\frac{2\tilde{\kappa}_s d_x}{s_x}} \int_{\Lambda_{\tilde{\kappa}}} \left| \int_{\mathbb{R}^{d_x}} \mathcal{F}[M_{0, -\frac{m}{2}}] \left(2^{-\frac{\tilde{\kappa}_s}{s_x}} \omega'_x \right) \cdot \mathbb{1}_{\ell_x} \left(2^{-\frac{\tilde{\kappa}_s}{s_x}} \omega'_x \right) \cdot \tilde{q}(\omega_x - \omega'_x) d\omega'_x \right|^2 d\omega_x \\
&\stackrel{(i)}{\geq} 2^{-\frac{2\tilde{\kappa}_s d_x}{s_x}} \int_{\tilde{I}_{\ell_x}} \left| \int_{\mathbb{R}^{d_x}} \mathcal{F}[M_{0, -\frac{m}{2}}] \left(2^{-\frac{\tilde{\kappa}_s}{s_x}} \omega'_x \right) \cdot \mathbb{1}_{\ell_x} \left(2^{-\frac{\tilde{\kappa}_s}{s_x}} \omega'_x \right) \cdot \tilde{q}(\omega_x - \omega'_x) d\omega'_x \right|^2 d\omega_x \\
&\stackrel{(ii)}{=} 2^{-\frac{2\tilde{\kappa}_s d_x}{s_x}} \int_{\tilde{I}_{\ell_x}} \left| \int_{\tilde{I}_{\ell_x}} \mathcal{F}[M_{0, -\frac{m}{2}}] \left(2^{-\frac{\tilde{\kappa}_s}{s_x}} \omega'_x \right) \cdot \tilde{q}(\omega_x - \omega'_x) d\omega'_x \right|^2 d\omega_x. \tag{115}
\end{aligned}$$

(i) above holds because $\tilde{I}_{\ell_x} \subset \Lambda_{\tilde{\kappa}}$; and (ii) holds because of the indicator function. Notice, for any $\omega_x, \omega'_x \in \tilde{I}_{\ell_x}$, we have for $j \in \{1, \dots, d_x\}$, $1 \geq \omega_{x,j} - \omega'_{x,j} \geq -1$. Since for any $-1 \leq t \leq 1$, $i \in \{1, \dots, d_x\}$ and g defined in Eq. (104), there is $\mathcal{F}[\sqrt{g}](t) = \int_{-1/2}^{1/2} \sqrt{g}(x_i) \exp(-x_i t) dx_i = \int_{-1/2}^{1/2} \sqrt{g}(x_i) \cos(x_i t) dx_i > 0$. So we have, for $\zeta > 3$,

$$\tilde{q}(\omega_x - \omega'_x) = q(\omega_x - \omega'_x) = \mathcal{F}[\sqrt{p_X}](\omega_x - \omega'_x) = \prod_{i=1}^{d_x} \mathcal{F}[\sqrt{g}](\omega_{x,i} - \omega'_{x,i}) > 0.$$

Since both $\mathcal{F}[M_{0, -\frac{m}{2}}](2^{-\frac{\tilde{\kappa}_s}{s_x}} \omega'_x)$ and $\tilde{q}(\omega_x - \omega'_x)$ are positive and real, we continue from Eq. (115) to have

$$\geq 2^{-\frac{2\tilde{\kappa}_s d_x}{s_x}} \left(\inf_{\omega_x \in \tilde{I}_{\ell_x}} \mathcal{F}[M_{0, -\frac{m}{2}}](\omega_x) \right)^2 \cdot \underbrace{\int_{\tilde{I}_{\ell_x}} \left(\int_{\tilde{I}_{\ell_x}} \tilde{q}(\omega_x - \omega'_x) d\omega'_x \right)^2 d\omega_x}_{(*)}.$$

Notice that in the integration in the term (*) above, only the difference of $\omega_x - \omega'_x$ show up. So we can obtain

$$\begin{aligned}
(*) &= \int_{[0,1]^{d_x}} \left(\int_{[0,1]^{d_x}} \tilde{q}(\omega_x - \omega'_x) d\omega'_x \right)^2 d\omega_x \\
&= \prod_{i=1}^{d_x} \int_0^1 \left(\int_0^1 \tilde{q}_i(\omega_{x,i} - \omega'_{x,i}) d\omega'_{x,i} \right)^2 d\omega_{x,i} \\
&= \prod_{i=1}^{d_x} \int_0^1 \left(\int_{\omega_{x,i}-1}^{\omega_{x,i}} \tilde{q}_i(\omega''_{x,i}) d\omega''_{x,i} \right)^2 d\omega_{x,i} \\
&= \prod_{i=1}^{d_x} \int_0^1 \left(\int_{\omega_{x,i}-1}^{\omega_{x,i}} q_i(\omega''_{x,i}) d\omega''_{x,i} \right)^2 d\omega_{x,i}.
\end{aligned}$$

The second last equality holds by change of variables, and the last equality holds because $\tilde{q}_i(\omega) = q_i(\omega)$ for $\omega \in [-\zeta/3, \zeta/3]$. So (*) is a strictly positive constant independent of ζ, k, n . Plugging it back to above, we obtain

$$\|\mathcal{F}[\Omega_{\tilde{\kappa}, \ell_x}] * \tilde{q}\|_{L^2(\mathbb{R}^{d_x})}^2 \geq 2^{-\frac{2\tilde{\kappa}_s d_x}{s_x}} \left(\inf_{\omega_x \in \tilde{I}_{\ell_x}} \mathcal{F}[M_{0, -\frac{m}{2}}](\omega_x) \right)^2 \cdot (*).$$

Notice that, since $I_{\ell_x} \subseteq [1.1\pi, 1.95\pi]^{d_x}$ and we use Eq. (106) to obtain, since $M_{0, -\frac{m}{2}}$ is an even function and thus has real-valued Fourier transform

$$\inf_{\omega_x \in I_{\ell_x}} \left| \mathcal{F}[M_{0, -\frac{m}{2}}](\omega_x) \right|^2 = \inf_{\omega_x \in I_{\ell_x}} \left| \prod_{i=1}^{d_x} \frac{\sin(\omega_{x,i}/2)^m}{(\omega_{x,i}/2)^m} \right|^2 \geq \inf_{\omega \in [1.1\pi, 1.95\pi]} \left(\frac{\sin(\omega/2)}{\omega/2} \right)^{2md_x} > 0.$$

Define the following positive constant, independent of ζ , k or n ,

$$C_\chi := (*) \cdot \inf_{\omega \in [1.1\pi, 1.95\pi]} \left(\frac{\sin(\omega/2)}{\omega/2} \right)^{2\mathfrak{m}d_x}. \quad (116)$$

We thus have

$$\|\mathcal{F}[\Omega_{\mathfrak{K}, \ell_x}] * \tilde{q}\|_{L^2(\mathbb{R}^{d_x})}^2 \geq C_\chi 2^{-2\mathfrak{K}_s \frac{d_x}{s_x}}. \quad (117)$$

Therefore,

$$(A)^2 \geq \sum_{\ell_o, \ell_x} (\beta_{\mathbf{v}(\ell_x, \ell_o)} - \beta_{\mathbf{v}'(\ell'_x, \ell'_o)})^2 C_\chi 2^{-\frac{2\mathfrak{K}_s d_x}{s_x}} \stackrel{(a)}{\geq} \frac{C_\chi}{8} |\mathcal{L}| 2^{-\frac{2\mathfrak{K}_s d_x}{s_x}} \stackrel{(b)}{\gtrsim} \zeta^{-d_x} 2^{-\mathfrak{K}_s \left(\frac{d_x}{s_x} - \frac{d_o}{s_o} \right)}.$$

where (a) follows from Eq. (109) and (b) follows from Eq. (100).

Next, we are going to upper bound (B). We have

$$\begin{aligned} (B)^2 &\leq \sum_{\ell_o} \left(\sum_{\ell_x} \|\mathcal{F}[\Omega_{\mathfrak{K}, \ell_x}] * (q - \tilde{q})\|_{L^2(\mathbb{R}^{d_x})} \right)^2 \\ &\stackrel{(a)}{\leq} \|q - \tilde{q}\|_{L^1(\mathbb{R}^{d_x})}^2 \sum_{\ell_o} \left(\sum_{\ell_x} \|\mathcal{F}[\Omega_{\mathfrak{K}, \ell_x}]\|_{L^2(\mathbb{R}^{d_x})} \right)^2 \\ &\stackrel{(b)}{\leq} \|q - \tilde{q}\|_{L^1(\mathbb{R}^{d_x})}^2 \sum_{\ell_o} \left(\sum_{\ell_x} \|M_{\mathfrak{K}, -\frac{m}{2}}\|_{L^2(\mathbb{R}^{d_x})} \right)^2 \\ &\stackrel{(c)}{=} \|q - \tilde{q}\|_{L^1(\mathbb{R}^{d_x})}^2 \sum_{\ell_o} \left(\sum_{\ell_x} 2^{-\mathfrak{K} \frac{s d_x}{2 s_x}} \|M_{0, -\frac{m}{2}}\|_{L^2(\mathbb{R}^{d_x})} \right)^2 \\ &= \|q - \tilde{q}\|_{L^1(\mathbb{R}^{d_x})}^2 \sum_{\ell_o} \left(\sum_{\ell_x} 1 \right)^2 2^{-\mathfrak{K} \frac{s d_x}{s_x}} \|M_{0, -\frac{m}{2}}\|_{L^2(\mathbb{R}^{d_x})}^2 \\ &\lesssim \|q - \tilde{q}\|_{L^1(\mathbb{R}^{d_x})}^2 \zeta^{-2d_x} 2^{\mathfrak{K}_s \left(\frac{2d_x}{s_x} + \frac{d_o}{s_o} \right)} 2^{-\mathfrak{K} \frac{s d_x}{s_x}} \|M_{0, -\frac{m}{2}}\|_{L^2(\mathbb{R}^{d_x})}^2 \\ &\stackrel{(d)}{\leq} \|q - \tilde{q}\|_{L^1(\mathbb{R}^{d_x})}^2 \zeta^{-\frac{1}{2}d_x} 2^{\frac{\mathfrak{K}_s d_o}{s_o}} 2^{\frac{\mathfrak{K}_s d_x}{s_x}} \|M_{0, -\frac{m}{2}}\|_{L^2(\mathbb{R}^{d_x})}^2 \end{aligned} \quad (118)$$

In the above derivations, (a) holds by Young's convolution inequality, (b) holds by Plancherel's Theorem and Eq. (105), (c) holds by the change of variables $\mathbf{x} \leftarrow 2^{-\frac{\mathfrak{K}_s}{s_x}} \mathbf{x}$, and (d) holds for $\zeta \geq 1$. We have

$$\begin{aligned} \|q - \tilde{q}\|_{L^1(\mathbb{R}^{d_x})} &= \int_{\mathbb{R}^{d_x}} |q(\boldsymbol{\omega}_x) - \tilde{q}(\boldsymbol{\omega}_x)| \, d\boldsymbol{\omega}_x \\ &\stackrel{(i)}{=} 2^{d_x} \prod_{i=1}^{d_x} \int_{\zeta/3}^{\infty} |\mathcal{F}[\sqrt{g}](\boldsymbol{\omega}_{x,i})| \, d\boldsymbol{\omega}_{x,i} \\ &\stackrel{(ii)}{\asymp} \prod_{i=1}^{d_x} \int_{\zeta/3}^{\infty} \boldsymbol{\omega}_{x,i}^{-3/4} \exp(-\sqrt{\boldsymbol{\omega}_{x,i}}) \, d\boldsymbol{\omega}_x \end{aligned}$$

$$\begin{aligned}
&\leq \zeta^{-\frac{3}{4}d_x} \prod_{i=1}^{d_x} \int_{\zeta}^{\infty} \exp(-\sqrt{\omega_{x,i}}) d\omega_{x,i} \\
&= \zeta^{-\frac{3}{4}d_x} 2^{d_x} \left(1 + \zeta^{1/2}\right)^{d_x} \exp\left(-d_x \zeta^{\frac{1}{2}}\right) \\
&\leq 2^{2d_x} \zeta^{-\frac{1}{4}d_x} \exp\left(-d_x \zeta^{\frac{1}{2}}\right), \tag{119}
\end{aligned}$$

where in (i) we use the definition of \tilde{q} in Eq. (110); and in (ii) we use the asymptotic decay of the Fourier transform of the bump function [Johnson, 2015, Section 2].

$$|\mathcal{F}[\sqrt{g}](\omega)| \asymp \omega^{-3/4} \exp(-\sqrt{\omega}), \quad \omega \gg 1. \tag{120}$$

Hence, we plug Eq. (119) back to Eq. (118) and we find that

$$\begin{aligned}
(B)^2 &\lesssim \zeta^{-\frac{1}{2}d_x} 2^{\frac{\mathfrak{K}_s d_o}{s_o} + \frac{\mathfrak{K}_s d_x}{s_x}} \left(2^{2d_x} \zeta^{-\frac{1}{4}d_x} \exp\left(-d_x \zeta^{\frac{1}{2}}\right)\right)^2 \|M_{0,-\frac{m}{2}}\|_{L^2(\mathbb{R}^{d_x})}^2 \\
&\lesssim 2^{\frac{\mathfrak{K}_s d_o}{s_o} + \frac{\mathfrak{K}_s d_x}{s_x}} \zeta^{-d_x} \exp\left(-2d_x \zeta^{\frac{1}{2}}\right). \tag{121}
\end{aligned}$$

Hence in order for $(B)^2 \leq (A)^2/4$ to hold, a sufficient condition on ζ is given by the following inequality, up to some constants,

$$\begin{aligned}
2^{\frac{\mathfrak{K}_s d_o}{s_o} + \frac{\mathfrak{K}_s d_x}{s_x}} \zeta^{-d_x} \exp\left(-2d_x \zeta^{\frac{1}{2}}\right) &\lesssim \frac{1}{4} \zeta^{-d_x} 2^{-\frac{\mathfrak{K}_s d_x}{s_x}} 2^{\frac{\mathfrak{K}_s d_o}{s_o}} \\
\iff \exp\left(-2d_x \zeta^{\frac{1}{2}}\right) &\lesssim 2^{-\frac{2\mathfrak{K}_s d_x}{s_x}}. \tag{122}
\end{aligned}$$

Hence for sufficiently large $\zeta = \zeta(n)$, such that Eq. (122) is satisfied, we have $(A) \geq (B)/2$, thus

$$\begin{aligned}
\|f_{\mathbf{v}} - f_{\mathbf{v}'}\|_{L^2(P_{XO})}^2 &\geq \epsilon_0^2 2^{-2\mathfrak{K}_s\left(1-\frac{d_x}{2s_x}\right)} 2^{-\frac{\mathfrak{K}_s d_o}{s_o}} \|M_{0,-\frac{m}{2}}\|_{L^2(\mathbb{R}^{d_o})}^2 \frac{1}{4}(A)^2 \\
&\gtrsim \epsilon_0^2 2^{-2\mathfrak{K}_s\left(1-\frac{d_x}{2s_x}\right)} 2^{-\frac{\mathfrak{K}_s d_o}{s_o}} \zeta^{-d_x} 2^{-\frac{\mathfrak{K}_s d_x}{s_x}} 2^{\frac{\mathfrak{K}_s d_o}{s_o}} \\
&= \epsilon_0^2 2^{-2\mathfrak{K}_s} \zeta^{-d_x}. \tag{123}
\end{aligned}$$

Step Five Recall the distributions $P_{f_{\mathbf{v}}}$ defined above, and recall that $\mathbf{0} \in V_{\mathfrak{K}}$ for $V_{\mathfrak{K}}$ defined in Eq. (109). In this step, we are going to show that, for $P_{f_{\mathbf{v}}}^{\otimes n} := P_{f_{\mathbf{v}}} \otimes \cdots \otimes P_{f_{\mathbf{v}}}$ which is a probability distribution over $(\mathcal{Z} \times \mathcal{O} \times \mathbb{R})^n$, it satisfies

$$\frac{1}{|V_{\mathfrak{K}}|} \sum_{\mathbf{v} \in V_{\mathfrak{K}}} \text{KL}\left(P_{f_{\mathbf{v}}}^{\otimes n}, P_{f_{\mathbf{0}}}^{\otimes n}\right) \lesssim \epsilon_0^2 n 2^{-2\mathfrak{K} \frac{s d_x}{s_x} \eta_1 - 2\mathfrak{K}_s}.$$

Notice that KL divergence tensorizes over independent copies at each dimension, we have that $\text{KL}(P_{f_{\mathbf{0}}}^{\otimes n} \| P_{f_{\mathbf{v}}}^{\otimes n}) = n \text{KL}(P_{f_{\mathbf{0}}} \| P_{f_{\mathbf{v}}})$, so we are going to study $\text{KL}(P_{f_{\mathbf{0}}} \| P_{f_{\mathbf{v}}})$ as follows.

$$\begin{aligned}
\text{KL}(P_{f_{\mathbf{v}}}, P_{f_{\mathbf{0}}}) &= \mathbb{E}_{(\mathbf{z}, \mathbf{o}) \sim P_{ZO}} [\text{KL}(P_{f_{\mathbf{v}}}(\cdot | \mathbf{z}, \mathbf{o}), P_{f_{\mathbf{0}}}(\cdot | \mathbf{z}, \mathbf{o}))] \\
&\stackrel{(a)}{=} \frac{\|T f_{\mathbf{v}}\|_{L^2(P_{ZO})}^2}{2\sigma^2} \\
&\stackrel{(b)}{\leq} 2^{-2\mathfrak{K} \frac{s d_x}{s_x} \eta_1} \frac{\|f_{\mathbf{v}}\|_{L^2(P_{ZO})}^2}{2\sigma^2} \tag{124}
\end{aligned}$$

In the above chain of derivations, we use Blanchard and Mücke [2018][Proposition 6.2] in (a); and in (b) we use the Assumption 4.3 along with the fact from Eq. (107) that the support of the Fourier transform of $f_{\mathbf{v}}$ is indeed in high frequency.

$$\text{supp}(\mathcal{F}[f_{\mathbf{v}}]) = \cup_{\ell_x, \ell_o \in \mathcal{L}} \tilde{I}_{\ell_x} \subseteq \left[1.1\pi \cdot 2^{\tilde{\kappa} \frac{s}{s_x}}, 1.9\pi \cdot 2^{\tilde{\kappa} \frac{s}{s_x}} \right]^{d_x}.$$

Next notice that, by Eq. (111), we have

$$\begin{aligned} & \|f_{\mathbf{v}}\|_{L^2(P_{X_O})}^2 \\ & \stackrel{(a)}{\leq} \epsilon_0^2 2^{-2\tilde{\kappa}s(1-\frac{d_x}{2s_x})} 2^{-\frac{\tilde{\kappa}_s d_o}{s_o}} \|M_{0,\mathbf{0}}\|_{L^2(\mathbb{R}^{d_o})}^2 \int_{\mathbb{R}^{d_x}} \sum_{\ell_o} \left(\sum_{\ell_x} \beta_{\mathbf{v}(\ell_x, \ell_o)} \mathcal{F}[\Omega_{\tilde{\kappa}, \ell_x} \sqrt{p}](\omega_x) \right)^2 d\omega_x \\ & \leq \epsilon_0^2 2^{-2\tilde{\kappa}s(1-\frac{d_x}{2s_x})} 2^{-\frac{\tilde{\kappa}_s d_o}{s_o}} \|M_{0,\mathbf{0}}\|_{L^2(\mathbb{R}^{d_o})}^2 \int_{\mathbb{R}^{d_x}} \sum_{\ell_o} \left(\sum_{\ell_x} |\mathcal{F}[\Omega_{\tilde{\kappa}, \ell_x} \sqrt{p}](\omega_x)| \right)^2 d\omega_x \\ & \stackrel{(b)}{\lesssim} \epsilon_0^2 2^{-2\tilde{\kappa}s(1-\frac{d_x}{2s_x})} \int_{\mathbb{R}^{d_x}} \left(\sum_{\ell_x} |\mathcal{F}[\Omega_{\tilde{\kappa}, \ell_x} \sqrt{p}](\omega_x)| \right)^2 d\omega_x \\ & \leq \epsilon_0^2 2^{-2\tilde{\kappa}s(1-\frac{d_x}{2s_x})} \left\| \sum_{\ell_x} |\mathcal{F}[\Omega_{\tilde{\kappa}, \ell_x}] * (q - \tilde{q} + \tilde{q})| \right\|_{L^2(\mathbb{R}^{d_x})}^2 \\ & \stackrel{(c)}{\leq} \epsilon_0^2 2^{-2\tilde{\kappa}s(1-\frac{d_x}{2s_x})} \left(2 \left\| \sum_{\ell_x} |\mathcal{F}[\Omega_{\tilde{\kappa}, \ell_x}] * (q - \tilde{q})| \right\|_{L^2(\mathbb{R}^{d_x})}^2 + 2 \left\| \sum_{\ell_x} |\mathcal{F}[\Omega_{\tilde{\kappa}, \ell_x}] * \tilde{q}| \right\|_{L^2(\mathbb{R}^{d_x})}^2 \right). \end{aligned} \quad (125)$$

In the above derivations, (a) holds by Eq. (111), (b) holds by $\sum_{\ell_o} 1 \lesssim 2^{\frac{\tilde{\kappa}_s d_o}{s_o}}$, (c) holds by triangular inequality. The first term in Eq. (125) has already been upper bounded in Eq. (121). Hence for sufficiently large ζ that satisfies Eq. (122),

$$\begin{aligned} & \left\| \sum_{\ell_x} |\mathcal{F}[\Omega_{\tilde{\kappa}, \ell_x}] * (q - \tilde{q})| \right\|_{L^2(\mathbb{R}^{d_x})}^2 \lesssim 2^{\frac{\tilde{\kappa}_s d_x}{s_x}} \zeta^{-\frac{d_x}{2}} \exp(-2d_x \zeta^{\frac{1}{2}}) \\ & \lesssim 2^{\frac{\tilde{\kappa}_s d_x}{s_x}} 2^{-\frac{2\tilde{\kappa}_s d_x}{s_x}} = 2^{-\frac{\tilde{\kappa}_s d_x}{s_x}}. \end{aligned}$$

The second term in Eq. (125) can be upper bounded by

$$\begin{aligned} & \left\| \sum_{\ell_x} |\mathcal{F}[\Omega_{\tilde{\kappa}, \ell_x}] * \tilde{q}| \right\|_{L^2(\mathbb{R}^{d_x})}^2 \stackrel{(a)}{\lesssim} \sum_{\ell_x} \|\mathcal{F}[\Omega_{\tilde{\kappa}, \ell_x}] * \tilde{q}\|_{L^2(\mathbb{R}^{d_x})}^2 \\ & \stackrel{(b)}{\leq} \sum_{\ell_x} \|\mathcal{F}[\Omega_{\tilde{\kappa}, \ell_x}]\|_{L^2(\mathbb{R}^{d_x})}^2 \cdot \|\tilde{q}\|_{L^1(\mathbb{R}^{d_x})}^2 \\ & \stackrel{(c)}{\leq} \sum_{\ell_x} \left\| \mathcal{F}[M_{\tilde{\kappa}, -\frac{m}{2}}] \right\|_{L^2(\tilde{I}_{\ell_x})}^2 \cdot \|\tilde{q}\|_{L^1(\mathbb{R}^{d_x})}^2 \\ & \leq \left\| \mathcal{F}[M_{\tilde{\kappa}, -\frac{m}{2}}] \right\|_{L^2(\mathbb{R}^{d_x})}^2 \cdot \|\tilde{q}\|_{L^1(\mathbb{R}^{d_x})}^2 \\ & \stackrel{(d)}{=} \left\| M_{\tilde{\kappa}, -\frac{m}{2}} \right\|_{L^2(\mathbb{R}^{d_x})}^2 \cdot \|\tilde{q}\|_{L^1(\mathbb{R}^{d_x})}^2 \end{aligned}$$

$$\begin{aligned} &\stackrel{(e)}{=} 2^{-\frac{\mathfrak{K}_s d_x}{s_x}} \left\| M_{0, -\frac{m}{2}} \right\|_{L^2(\mathbb{R}^{d_x})}^2 \|\tilde{q}\|_{L^1(\mathbb{R}^{d_x})}^2 \\ &\lesssim 2^{-\frac{\mathfrak{K}_s d_x}{s_x}}. \end{aligned}$$

In the above chain of derivations, (a) holds because $\mathcal{F}[\Omega_{\mathfrak{K}, \ell_x}] * \tilde{q}$ have disjoint support for different ℓ_x proved in Eq. (114); (b) holds by Young's convolution inequality; (c) holds by Eq. (105); (d) holds by the fact that I_{ℓ_x} are pairwise disjoint and Plancherel's Theorem and (e) holds by change of variables $\mathbf{x} \leftarrow 2^{-\frac{\mathfrak{K}_s}{s_x}} \mathbf{x}$. Therefore, combining the upper bound on the above two terms, we obtain

$$\|f_{\mathbf{v}}\|_{L^2(P_{XO})}^2 \lesssim \epsilon_0^2 2^{-2\mathfrak{K}_s \left(1 - \frac{d_x}{2s_x}\right)} 2^{-\frac{\mathfrak{K}_s d_x}{s_x}} = \epsilon_0^2 2^{-2\mathfrak{K}_s}. \quad (126)$$

We plug the upper bound on $\|f_{\mathbf{v}}\|_{L^2(P_{XO})}^2$ back to Eq. (124) to obtain

$$\text{KL}(P_{f_{\mathbf{v}}}, P_{f_0}) \lesssim \epsilon_0^2 \sigma^{-2} 2^{-2\mathfrak{K} \frac{s d_x}{s_x} \eta_1 - 2\mathfrak{K}_s}. \quad (127)$$

And thus,

$$\frac{1}{|V_{\mathfrak{K}}|} \sum_{\mathbf{v} \in V_{\mathfrak{K}}} \text{KL}\left(P_{f_{\mathbf{v}}}^{\otimes n}, P_{f_0}^{\otimes n}\right) \leq \epsilon_0^2 n \sigma^{-2} 2^{-2\mathfrak{K} \frac{s d_x}{s_x} \eta_1 - 2\mathfrak{K}_s}. \quad (128)$$

Step Six In this step, we are going to show that, for any measurable learning method $(\mathbf{z}_i, \mathbf{o}_i, y_i)_{i=1}^n =: D \mapsto \hat{f}_D$, there is a distribution P among $P_{f_{\mathbf{v}}}$ with $\mathbf{v} \in V_{\mathfrak{K}}$ which is difficult to learn for the considered learning method. We define a measurable mapping

$$\Psi : \left([0, 1]^{d_z} \times [0, 1]^{d_o} \times \mathbb{R}\right)^n \rightarrow V_{\mathfrak{K}}, \quad \Psi(D) := \underset{\mathbf{v} \in V_{\mathfrak{K}}}{\operatorname{argmin}} \left\| \hat{f}_D - f_{\mathbf{v}} \right\|_{L^2(P_{XO})}. \quad (129)$$

For $\mathbf{v} \in V_{\mathfrak{K}}$ and $D \in \left([0, 1]^{d_z} \times [0, 1]^{d_o} \times \mathbb{R}\right)^n$ with $\Psi(D) \neq \mathbf{v}$, we start from (123) to have

$$\begin{aligned} \epsilon_0 2^{-\mathfrak{K}_s} \zeta^{-\frac{d_x}{2}} &\lesssim \|f_{\Psi(D)} - f_{\mathbf{v}}\|_{L^2(P_{XO})} \\ &\leq \|f_{\Psi(D)} - \hat{f}_D\|_{L^2(P_{XO})} + \|\hat{f}_D - f_{\mathbf{v}}\|_{L^2(P_{XO})} \\ &\leq 2 \|\hat{f}_D - f_{\mathbf{v}}\|_{L^2(P_{XO})}. \end{aligned}$$

Consequently, for all $\mathbf{v} \in V_{\mathfrak{K}}$ we find

$$P_{f_{\mathbf{v}}}^{\otimes n}(D : \Psi(D) \neq \mathbf{v}) \leq P_{f_{\mathbf{v}}}^{\otimes n}\left(D : \left\| \hat{f}_D - f_{\mathbf{v}} \right\|_{L^2(P_{XO})} \gtrsim \epsilon_0 2^{-\mathfrak{K}_s} \zeta^{-\frac{d_x}{2}}\right).$$

Therefore, we have

$$\begin{aligned} &\max_{\mathbf{v} \in V_{\mathfrak{K}}} P_{f_{\mathbf{v}}}^{\otimes n}\left(D : \left\| \hat{f}_D - f_{\mathbf{v}} \right\|_{L^2(P_{XO})} \gtrsim \epsilon_0 2^{-\mathfrak{K}_s} \zeta^{-\frac{d_x}{2}}\right) \\ &\geq \max_{\mathbf{v} \in V_{\mathfrak{K}}} P_{f_{\mathbf{v}}}^{\otimes n}(D : \Psi(D) \neq \mathbf{v}) \\ &\stackrel{(a)}{\gtrsim} \frac{\sqrt{|V_{\mathfrak{K}}|}}{\sqrt{|V_{\mathfrak{K}}|} + 1} \left(1 - \frac{\epsilon_0^2 n 2^{-2\mathfrak{K} \frac{s d_x}{s_x} \eta_1} 2^{-2\mathfrak{K}_s}}{\log |V_{\mathfrak{K}}|} - \frac{1}{2 \log |V_{\mathfrak{K}}|}\right) \end{aligned}$$

$$\begin{aligned}
&\stackrel{(b)}{\geq} \frac{\sqrt{|V_{\mathfrak{K}}|}}{\sqrt{|V_{\mathfrak{K}}|} + 1} \left(1 - \frac{\epsilon_0^2 n 2^{-2\mathfrak{K}_{sx}} 2^{-2\mathfrak{K}_s}}{\zeta^{-d_x} 2^{\mathfrak{K}_{sx} \frac{s_x}{s_o} + \mathfrak{K}_{so} \frac{d_o}{s_o}}} - \frac{1}{2 \log |V_{\mathfrak{K}}|} \right) \\
&\geq \frac{\sqrt{|V_{\mathfrak{K}}|}}{\sqrt{|V_{\mathfrak{K}}|} + 1} \left(1 - \epsilon_0^2 \zeta^{d_x} n 2^{-\mathfrak{K}_{sx} \left(2d_x \eta_1 + 2s_x + d_x + \frac{d_o}{s_o} s_x \right)} - \frac{1}{2 \log |V_{\mathfrak{K}}|} \right). \tag{130}
\end{aligned}$$

(a) holds by Fischer and Steinwart [2020, Theorem 20] and Eq. (128). (b) holds by the construction of $V_{\mathfrak{K}}$ in Eq. (109) with $|V_{\mathfrak{K}}| \geq 2^{\frac{c}{s}}$ so $\log |V_{\mathfrak{K}}| \gtrsim \zeta^{-d_x} 2^{\mathfrak{K}_s \left(\frac{d_x}{s_x} + \frac{d_o}{s_o} \right)}$. Next, we choose ζ and k as the following function of n :

$$\zeta = \left(\frac{1}{2} \frac{1}{2d_x \eta_1 + 2s_x + d_x + \frac{d_o}{s_o} s_x} \log n \right)^2, \quad 2^{-\mathfrak{K}_s} = n^{-\frac{s_x}{2d_x \eta_1 + 2s_x + d_x + \frac{d_o}{s_o} s_x}} \zeta^{-s_x}. \tag{131}$$

The choice of ζ ensures that Eq. (122) is satisfied for sufficiently large $n \geq 1$. The choice of k as a function of n ensures that we can proceed from Eq. (130) to obtain

$$\begin{aligned}
&\max_{\mathbf{v} \in V_{\mathfrak{K}}} P_{\mathbf{f}_{\mathbf{v}}}^{\otimes n} \left(D : \left\| \hat{f}_D - f_{\mathbf{v}} \right\|_{L^2(P_{XO})} \geq \epsilon_0 n^{-\frac{s_x}{2s_x + 2d_x \eta_1 + d_x + \frac{d_o}{s_o} s_x}} (\log n)^{-2s_x - d_x} \right) \\
&\gtrsim \frac{\sqrt{|V_{\mathfrak{K}}|}}{\sqrt{|V_{\mathfrak{K}}|} + 1} \left(1 - \epsilon_0^2 - \frac{1}{2 \log |V_{\mathfrak{K}}|} \right) \\
&\geq 1 - 2\epsilon_0^2.
\end{aligned}$$

The last inequality holds for sufficiently large n because $|V_{\mathfrak{K}}| \rightarrow \infty$ as $n \rightarrow \infty$. We set $2\epsilon_0^2 = \tau^2$ and relabelling the constants, to obtain

$$\begin{aligned}
\inf_{D \mapsto \hat{f}_D} \sup_{f \in B_{2,q}^{s_x, s_o}(\mathbb{R}^{d_x + d_o})} \left\| \hat{f}_D - f \right\|_{L^2(P_{XO})} &\geq \inf_{D \mapsto \hat{f}_D} \max_{\mathbf{v} \in V_{\mathfrak{K}}} \left\| \hat{f}_D - f_{\mathbf{v}} \right\|_{L^2(P_{XO})} \\
&\geq \tau n^{-\frac{s_x}{2s_x + 2d_x \eta_1 + d_x + \frac{d_o}{s_o} s_x}} (\log n)^{-2s_x - d_x} \\
&= \tau n^{-\frac{s_x}{1 + 2(\frac{s_x}{d_x} + \eta_1) + \frac{d_o}{s_o} \frac{s_x}{d_x}}} (\log n)^{-2s_x - d_x},
\end{aligned}$$

holds for sufficiently large n with probability $\geq 1 - C\tau^2$.

E.3 Auxiliary Results

Lemma E.2. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$, $f \in B_{2,q}^s(\mathbb{R}^d)$ and $K \in L^1(\mathbb{R}^d)$, then

$$\|K * f\|_{B_{2,q}^s(\mathbb{R}^d)} \leq \|K\|_{L^1(\mathbb{R}^d)} \|f\|_{B_{2,q}^s(\mathbb{R}^d)}$$

Proof. Define $(\tau_{\mathbf{h}} f)(\mathbf{x}) := f(\mathbf{x} + \mathbf{h})$. We have

$$\tau_{\mathbf{h}}(K * f)(\mathbf{x}) = (K * f)(\mathbf{x} + \mathbf{h}) = (K * (\tau_{\mathbf{h}} f))(\mathbf{x}).$$

Since $\Delta_{\mathbf{h}}^r = (\tau_{\mathbf{h}} - \text{id})^r$, we have $\Delta_{\mathbf{h}}^r(K * f) = K * (\Delta_{\mathbf{h}}^r f)$. Thus for any $r > \max\{s_1, \dots, s_d\}$,

$$\omega_{r,2} \left(K * f, t^{\frac{1}{s_1}}, \dots, t^{\frac{1}{s_d}}, \mathbb{R}^d \right) = \sup_{0 < |h_i| < t^{\frac{1}{s_i}}} \|\Delta_{\mathbf{h}}^r(K * f)\|_{L^2(\mathbb{R}^d)}$$

$$\begin{aligned}
&= \sup_{0 < |h_i| < t^{\frac{1}{s_i}}} \|K * \Delta_{\mathbf{h}}^r(f)\|_{L^2(\mathbb{R}^d)} \\
&\stackrel{(i)}{\leq} \sup_{0 < |h_i| < t^{\frac{1}{s_i}}} \|K\|_{L^1(\mathbb{R}^d)} \|\Delta_{\mathbf{h}}^r f\|_{L^2(\mathbb{R}^d)} \\
&= \|K\|_{L^1(\mathbb{R}^d)} \omega_{r,2} \left(f, t^{\frac{1}{s_1}}, \dots, t^{\frac{1}{s_d}}, \mathbb{R}^d \right),
\end{aligned} \tag{132}$$

where we use Young's convolution inequality in (i). Hence

$$\begin{aligned}
|K * f|_{B_{2,q}^s(\mathbb{R}^d)} &= \left(\int_0^1 \left[t^{-1} \omega_{r,2} \left(K * f, t^{\frac{1}{s_1}}, \dots, t^{\frac{1}{s_d}}, \mathbb{R}^d \right) \right]^q \frac{dt}{t} \right)^{\frac{1}{q}} \\
&\leq \|K\|_{L^1(\mathbb{R}^d)} \left(\int_0^1 \left[t^{-1} \omega_{r,2} \left(f, t^{\frac{1}{s_1}}, \dots, t^{\frac{1}{s_d}}, \mathbb{R}^d \right) \right]^q \frac{dt}{t} \right)^{\frac{1}{q}} \\
&= \|K\|_{L^1(\mathbb{R}^d)} |f|_{B_{2,q}^s(\mathbb{R}^d)}.
\end{aligned}$$

Hence

$$\begin{aligned}
\|K * f\|_{B_{2,q}^s(\mathbb{R}^d)} &= \|K * f\|_{L^2(\mathbb{R}^d)} + |K * f|_{B_{2,q}^s(\mathbb{R}^d)} \\
&\stackrel{(i)}{\leq} \|K\|_{L^1(\mathbb{R}^d)} \|f\|_{L^2(\mathbb{R}^d)} + \|K\|_{L^1(\mathbb{R}^d)} |f|_{B_{2,q}^s(\mathbb{R}^d)} = \|K\|_{L^1(\mathbb{R}^d)} \|f\|_{B_{2,q}^s(\mathbb{R}^d)},
\end{aligned}$$

where (i) is due to Young's convolution inequality. \square

Lemma E.3. Let $M_{k,-\frac{m}{2}} : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$, $M_{\mathfrak{K},\ell_o} : \mathbb{R}^{d_o} \rightarrow \mathbb{R}$ be defined in Eq. (98). Let $\mathbb{1}_{S(\ell_o)} : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$ be an indicator function over a measurable set $S(\ell_o) \subset \mathbb{R}^{d_x}$ for each ℓ_o in Eq. (99). Define $f(\mathbf{x}, \mathbf{o}) := \sum_{\ell_o} (M_{k,-\frac{m}{2}} * \mathcal{F}^{-1}[\mathbb{1}_{S(\ell_o)}])(\mathbf{x}) \cdot M_{\mathfrak{K},\ell_o}(\mathbf{o})$. Then we have $\|f\|_{B_{2,\infty}^{s_x,s_o}(\mathbb{R}^{d_x+d_o})} \lesssim 2^{\mathfrak{K}_S - \mathfrak{K} \frac{s}{s_x} \frac{d_x}{2}}$.

Proof. By definition of Besov norm in Leisner [2003][Eq. (2.2)], for any $r > \max\{s_x, s_o\}$, we have

$$\|f\|_{B_{2,\infty}^{s_x,s_o}(\mathbb{R}^{d_x+d_o})} \asymp \underbrace{\sup_{0 < t < 1} \left[t^{-1} \omega_{r,2}^{\mathbf{x}}(f, t^{1/s_x}, \dots, t^{1/s_x}) \right]}_{(I)} + \underbrace{\sup_{0 < t < 1} \left[t^{-1} \omega_{r,2}^{\mathbf{o}}(f, t^{1/s_o}, \dots, t^{1/s_o}) \right]}_{(II)},$$

where $\omega_{r,2}^{\mathbf{x}}$ (respectively $\omega_{r,2}^{\mathbf{o}}$) denotes the *partial* modulus of smoothness in the \mathbf{x} (respectively \mathbf{o})-direction, and we restrict the supremum from $t \in (0, \infty)$ to $t \in (0, 1)$ by DeVore and Lorentz [1993, Theorem 10.1]. Specifically, define $(\tau_{\mathbf{h}} g)(\mathbf{x}) := g(\mathbf{x} + \mathbf{h})$ and $\Delta_{\mathbf{h}}^r := (\tau_{\mathbf{h}} - \text{id})^r$, we write

$$\omega_{r,2}^{\mathbf{x}}(f, t^{1/s_x}, \dots, t^{1/s_x}) := \sup_{|h_i| \leq t^{1/s_x}} \left(\int_{\mathbb{R}^{d_o}} \int_{\mathbb{R}^{d_x}} |\Delta_{\mathbf{h}}^r(f(\cdot, \mathbf{o}))(\mathbf{x})|^2 d\mathbf{x} d\mathbf{o} \right)^{\frac{1}{2}},$$

and $\omega_{r,2}^{\mathbf{o}}$ is defined similarly. We first bound (I). We have

$$\tau_{\mathbf{h}}(K * g)(\mathbf{x}) = (K * g)(\mathbf{x} + \mathbf{h}) = (K * (\tau_{\mathbf{h}} g))(\mathbf{x}).$$

We thus have $\Delta_{\mathbf{h}}^r(K * g) = K * (\Delta_{\mathbf{h}}^r g)$. Thus we have

$$\omega_{r,2}^{\mathbf{x}}(f, t^{1/s_x}, \dots, t^{1/s_x})$$

$$\begin{aligned}
&= \sup_{|h_i| \leq t^{1/s_x}} \left(\int_{\mathbb{R}^{d_o}} \int_{\mathbb{R}^{d_x}} \left| \sum_{\ell_o} ((\Delta_{\mathbf{h}}^r M_{k, -\frac{m}{2}}) * \mathcal{F}^{-1}[\mathbb{1}_{S(\ell_o)}])(\mathbf{x}) M_{\mathfrak{K}, \ell_o}(\mathbf{o}) \right|^2 d\mathbf{x} d\mathbf{o} \right)^{\frac{1}{2}} \\
&= \sup_{|h_i| \leq t^{1/s_x}} \left(\sum_{\ell_o, \ell'_o} \int_{\mathbb{R}^{d_x}} ((\Delta_{\mathbf{h}}^r M_{k, -\frac{m}{2}}) * \mathcal{F}^{-1}[\mathbb{1}_{S(\ell_o)}])(\mathbf{x}) \overline{((\Delta_{\mathbf{h}}^r M_{k, -\frac{m}{2}}) * \mathcal{F}^{-1}[\mathbb{1}_{S(\ell'_o)}])(\mathbf{x})} d\mathbf{x} \right. \\
&\quad \cdot \left. \int_{\mathbb{R}^{d_o}} M_{\mathfrak{K}, \ell_o}(\mathbf{o}) M_{k, \ell'_o}(\mathbf{o}) d\mathbf{o} \right)^{\frac{1}{2}} \\
&\stackrel{(a)}{=} \sup_{|h_i| \leq t^{1/s_x}} \left(\sum_{\ell_o} \|\mathcal{F}^{-1}[\mathbb{1}_{S(\ell_o)}] * (\Delta_{\mathbf{h}}^r M_{k, -\frac{m}{2}})\|_{L^2(\mathbb{R}^{d_x})}^2 \|M_{\mathfrak{K}, \ell_o}\|_{L^2(\mathbb{R}^{d_o})}^2 \right)^{\frac{1}{2}} \\
&\stackrel{(b)}{\leq} \sup_{|h_i| \leq t^{1/s_x}} \left(\sum_{\ell_o} \|\Delta_{\mathbf{h}}^r M_{k, -\frac{m}{2}}\|_{L^2(\mathbb{R}^{d_x})}^2 \|M_{\mathfrak{K}, \ell_o}\|_{L^2(\mathbb{R}^{d_o})}^2 \right)^{\frac{1}{2}} \\
&= \left(\sup_{|h_i| \leq t^{1/s_x}} \|\Delta_{\mathbf{h}}^r M_{k, -\frac{m}{2}}\|_{L^2(\mathbb{R}^{d_x})} \right) \left(\sum_{\ell_o} \|M_{\mathfrak{K}, \ell_o}\|_{L^2(\mathbb{R}^{d_o})}^2 \right)^{\frac{1}{2}} \\
&\stackrel{(c)}{\leq} \sup_{|h_i| \leq t^{1/s_x}} \|\Delta_{\mathbf{h}}^r M_{k, -\frac{m}{2}}\|_{L^2(\mathbb{R}^{d_x})} \\
&= \omega_{r,2}(M_{k, -\frac{m}{2}}, t^{1/s_x}, \dots, t^{1/s_x}).
\end{aligned}$$

In the above derivations,

- we use $\text{supp}(M_{\mathfrak{K}, \ell_o}) \cap \text{supp}(M_{\mathfrak{K}, \ell'_o}) = \emptyset$ if $\ell_o \neq \ell'_o$ in (a),
- we use in (b) the inequality

$$\begin{aligned}
&\|\mathcal{F}^{-1}[\mathbb{1}_{S(\ell_o)}] * (\Delta_{\mathbf{h}}^r M_{k, -\frac{m}{2}})\|_{L^2(\mathbb{R}^{d_x})}^2 = \|\mathbb{1}_{S(\ell_o)} \cdot \mathcal{F}[\Delta_{\mathbf{h}}^r M_{k, -\frac{m}{2}}]\|_{L^2(\mathbb{R}^{d_x})}^2 \\
&\leq \|\mathcal{F}[\Delta_{\mathbf{h}}^r M_{k, -\frac{m}{2}}]\|_{L^2(\mathbb{R}^{d_x})}^2 = \|\Delta_{\mathbf{h}}^r M_{k, -\frac{m}{2}}\|_{L^2(\mathbb{R}^{d_x})}^2,
\end{aligned} \tag{133}$$

- we use $\sum_{\ell_o} \|M_{\mathfrak{K}, \ell_o}\|_{L^2(\mathbb{R}^{d_o})}^2 \leq 2^{\frac{\mathfrak{K}_s d_o}{s_o}} 2^{-\frac{\mathfrak{K}_s d_o}{s_o}} \|M_{00}\|_{L^2(\mathbb{R}^{d_o})}^2 \leq 1$ in (c). Indeed, note that $\|M_{00}\|_{L^2(\mathbb{R}^{d_o})}^2 = \|\iota_m\|_{L^2(\mathbb{R})}^{2d_o} \leq \|\iota_m\|_{L^1(\mathbb{R})}^{2d_o} = 1$.

Thus we have

$$(I) \leq \sup_{t>0} \left[t^{-1} \omega_{r,2} \left(M_{k, -\frac{m}{2}}, t^{\frac{1}{s_x}}, \dots, t^{\frac{1}{s_x}} \right) \right] \asymp |M_{k, -\frac{m}{2}}|_{B_{2,\infty}^{s_x}(\mathbb{R}^{d_x})} \asymp 2^{\mathfrak{K}_s \left(1 - \frac{d_x}{2s_x} \right)},$$

where we bound (I) by an isotropic Besov norm, and we use the sequential Besov norm equivalence [DeVore and Popov, 1988][Theorem 5.1] (see also Leisner [2003, Theorem 3.4] applied to $B_{2,\infty}^{s_x}(\mathbb{R}^{d_x})$) in the last equality. Now we bound (II). Notice that

$$\begin{aligned}
&\omega_{r,2}^{\mathbf{o}}(f, t^{1/s_o}, \dots, t^{1/s_o}) \\
&\stackrel{(a)}{\lesssim} \omega_{s_o,2}^{\mathbf{o}}(f, t^{1/s_o}, \dots, t^{1/s_o})
\end{aligned}$$

$$\begin{aligned}
&\stackrel{(b)}{\lesssim} 2^{\frac{\mathfrak{K}_S}{s_o}} \omega_{s_o, 2}^{\mathbf{o}} \left(f, 2^{-\frac{\mathfrak{K}_S}{s_o}} t^{1/s_o}, \dots, 2^{-\frac{\mathfrak{K}_S}{s_o}} t^{1/s_o} \right) \\
&= 2^{\frac{\mathfrak{K}_S}{s_o}} \sup_{|h_i| \leq 2^{-\frac{\mathfrak{K}_S}{s_o} t^{1/s_o}}} \left(\int_{\mathbb{R}^{d_x+d_o}} \left| \sum_{\ell_o} (\mathcal{F}^{-1}[\mathbb{1}_{S(\ell_o)}] * M_{k, -\frac{m}{2}})(\mathbf{x}) \cdot (\Delta_{\mathbf{h}}^{s_o} M_{\mathfrak{K}, \ell_o})(\mathbf{o}) \right|^2 d\mathbf{x} d\mathbf{o} \right)^{\frac{1}{2}} \\
&= 2^{\frac{\mathfrak{K}_S}{s_o}} \sup_{|h_i| \leq 2^{-\frac{\mathfrak{K}_S}{s_o} t^{1/s_o}}} \left(\sum_{\ell_o, \ell'_o} \int_{\mathbb{R}^{d_x}} (\mathcal{F}^{-1}[\mathbb{1}_{S(\ell_o)}] * M_{k, -\frac{m}{2}})(\mathbf{x}) \cdot \overline{\mathcal{F}^{-1}[\mathbb{1}_{S(\ell'_o)}] * M_{k, -\frac{m}{2}}}(\mathbf{x}) d\mathbf{x} \right. \\
&\quad \left. \cdot \int_{\mathbb{R}^{d_o}} (\Delta_{\mathbf{h}}^{s_o} M_{\mathfrak{K}, \ell_o})(\mathbf{o}) \cdot (\Delta_{\mathbf{h}}^{s_o} M_{k, \ell'_o})(\mathbf{o}) d\mathbf{o} \right)^{\frac{1}{2}} \\
&\stackrel{(c)}{=} 2^{\frac{\mathfrak{K}_S}{s_o}} \sup_{|h_i| \leq 2^{-\frac{\mathfrak{K}_S}{s_o} t^{1/s_o}}} \left(\sum_{\ell_o} \|\mathcal{F}^{-1}[\mathbb{1}_{S(\ell_o)}] * M_{k, \ell_o}\|_{L^2(\mathbb{R}^{d_x})}^2 \cdot \|\Delta_{\mathbf{h}}^{s_o} M_{\mathfrak{K}, \ell_o}\|_{L^2(\mathbb{R}^{d_o})}^2 \right)^{\frac{1}{2}} \\
&\stackrel{(d)}{\leq} 2^{\frac{\mathfrak{K}_S}{s_o}} \sup_{|h_i| \leq 2^{-\frac{\mathfrak{K}_S}{s_o} t^{1/s_o}}} \left(\sum_{\ell_o} \|M_{k, -\frac{m}{2}}\|_{L^2(\mathbb{R}^{d_x})}^2 \cdot \|\Delta_{\mathbf{h}}^{s_o} M_{\mathfrak{K}, \ell_o}\|_{L^2(\mathbb{R}^{d_o})}^2 \right)^{\frac{1}{2}} \\
&= 2^{\frac{\mathfrak{K}_S}{s_o}} \|M_{k, -\frac{m}{2}}\|_{L^2(\mathbb{R}^{d_x})} \left(\sup_{|h_i| \leq 2^{-\frac{\mathfrak{K}_S}{s_o} t^{1/s_o}}} \sum_{\ell_o} \|\Delta_{\mathbf{h}}^{s_o} M_{\mathfrak{K}, \ell_o}\|_{L^2(\mathbb{R}^{d_o})}^2 \right)^{\frac{1}{2}} \\
&\stackrel{(e)}{=} 2^{\frac{\mathfrak{K}_S}{s_o}} \|M_{k, -\frac{m}{2}}\|_{L^2(\mathbb{R}^{d_x})} \left(\sup_{|h_i| \leq 2^{-\frac{\mathfrak{K}_S}{s_o} t^{1/s_o}}} \left\| \sum_{\ell_o} (\Delta_{\mathbf{h}}^{s_o} M_{\mathfrak{K}, \ell_o}) \right\|_{L^2(\mathbb{R}^{d_o})}^2 \right)^{\frac{1}{2}} \\
&\stackrel{(f)}{=} 2^{\frac{\mathfrak{K}_S}{s_o}} \|M_{k, -\frac{m}{2}}\|_{L^2(\mathbb{R}^{d_x})} \left(\sup_{|h_i| \leq 2^{-\frac{\mathfrak{K}_S}{s_o} t^{1/s_o}}} \left\| \Delta_{\mathbf{h}}^{s_o} \left(\sum_{\ell_o} M_{\mathfrak{K}, \ell_o} \right) \right\|_{L^2(\mathbb{R}^{d_o})}^2 \right)^{\frac{1}{2}} \\
&= 2^{\frac{\mathfrak{K}_S}{s_o}} \|M_{k, -\frac{m}{2}}\|_{L^2(\mathbb{R}^{d_x})} \left(\omega_{s_o, 2} \left(\sum_{\ell_o} M_{\mathfrak{K}, \ell_o}, 2^{-\frac{\mathfrak{K}_S}{s_o} t^{\frac{1}{s_o}}}, \dots, 2^{-\frac{\mathfrak{K}_S}{s_o} t^{\frac{1}{s_o}}} \right)^2 \right)^{\frac{1}{2}}.
\end{aligned}$$

In the above derivations,

- we use $r = s_x \vee s_o \geq s_o$ and Minkowski's inequality in (a) (this is also sometimes referred to as the reverse Marchaud inequality, see [Kolomoitsev and Tikhonov \[2020\]](#)[Property 8]),
- we use [Leisner \[2003\]](#)[Theorem 2.1.1] in (b), which we can apply since $2^{\frac{\mathfrak{K}_S}{s_o}} \geq 1$,
- we deduce from Eq. (98) that

$$\text{supp}(M_{k\ell_o}) = \bigtimes_{j=1}^{d_o} \left[2^{-\left\lfloor \frac{\mathfrak{K}_S}{s_o} \right\rfloor} \ell_{o,j}, 2^{-\left\lfloor \frac{\mathfrak{K}_S}{s_o} \right\rfloor} (\mathfrak{m} + \ell_{o,j}) \right].$$

We also see that

$$\text{supp}(\Delta_{\mathbf{h}}^{s_o} M_{k\ell_o}) = \bigtimes_{j=1}^{d_o} \left[2^{-\left\lfloor \frac{\mathfrak{K}_S}{s_o} \right\rfloor} \ell_{o,j} - s_o h_j, 2^{-\left\lfloor \frac{\mathfrak{K}_S}{s_o} \right\rfloor} (\mathfrak{m} + \ell_{o,j}) \right]$$

We deduce that a sufficient condition to guarantee that

$$\text{supp}(\Delta_{\mathbf{h}}^{s_o} M_{\mathfrak{K}, \ell_o}) \cap \text{supp}(\Delta_{\mathbf{h}}^{s_o} M_{k, \ell'_o}) = \emptyset$$

if $\ell_o \neq \ell'_o$ is given by

$$(\forall j = 1, \dots, d_o), |h_j| \leq 2^{-\lfloor \frac{\mathfrak{K}_S}{s_o} \rfloor} \frac{\mathfrak{m} - 1}{s_o}.$$

Since $\mathfrak{m} - 1 \geq s_o$ by choice, and we have for all $i = 1, \dots, d_o$, $|h_i| \leq 2^{-\frac{\mathfrak{K}_S}{s_o}} t^{1/s_o} \leq 2^{-\frac{\mathfrak{K}_S}{s_o}}$, this sufficient condition is satisfied. Hence

$$\langle \Delta_{\mathbf{h}}^{s_o} M_{\mathfrak{K}, \ell_o}, \Delta_{\mathbf{h}}^{s_o} M_{k, \ell'_o} \rangle_{L^2(\mathbb{R}^{d_o})} = \|\Delta_{\mathbf{h}}^{s_o} M_{\mathfrak{K}, \ell_o}\|_{L^2(\mathbb{R}^{d_o})}^2 \delta_{\ell_o, \ell'_o}. \quad (134)$$

- we use Plancherel's Theorem in (d), in a similar way as in Eq. (133),
- we use Eq. (134) again for step (e),
- we use linearity of $\Delta_{\mathbf{h}}^{s_o}$ for step (f).

Thus we have shown that, for $q = \infty$,

$$\begin{aligned} (II) &\lesssim 2^{\mathfrak{K}_S} \|M_{k, -\frac{\mathfrak{m}}{2}}\|_{L^2(\mathbb{R}^{d_x})} \sup_{t>0} \left[t^{-1} \omega_{s_o, 2} \left(\sum_{\ell_o} M_{\mathfrak{K}, \ell_o}, 2^{-\frac{\mathfrak{K}_S}{s_o}} t^{\frac{1}{s_o}}, \dots, 2^{-\frac{\mathfrak{K}_S}{s_o}} t^{\frac{1}{s_o}} \right) \right] \\ &\lesssim \|M_{k, -\frac{\mathfrak{m}}{2}}\|_{L^2(\mathbb{R}^{d_x})} \sup_{t>0} \left[(t 2^{-\mathfrak{K}_S})^{-1} \omega_{s_o, 2} \left(\sum_{\ell_o} M_{\mathfrak{K}, \ell_o}, 2^{-\frac{\mathfrak{K}_S}{s_o}} t^{\frac{1}{s_o}}, \dots, 2^{-\frac{\mathfrak{K}_S}{s_o}} t^{\frac{1}{s_o}} \right) \right] \\ &\stackrel{(a)}{\lesssim} \|M_{k, -\frac{\mathfrak{m}}{2}}\|_{L^2(\mathbb{R}^{d_x})} \sup_{t>0} \left[\int_t^\infty w^{-s_o} \omega_{s_o+1} \left(\sum_{\ell_o} M_{\mathfrak{K}, \ell_o}, w, \dots, w \right) \frac{dw}{w} \right] \\ &\leq \|M_{k, -\frac{\mathfrak{m}}{2}}\|_{L^2(\mathbb{R}^{d_x})} \int_0^\infty w^{-s_o} \omega_{s_o+1} \left(\sum_{\ell_o} M_{\mathfrak{K}, \ell_o}, w, \dots, w \right) \frac{dw}{w} \\ &\asymp \|M_{k, -\frac{\mathfrak{m}}{2}}\|_{L^2(\mathbb{R}^{d_x})} \left\| \sum_{\ell_o} M_{\mathfrak{K}, \ell_o} \right\|_{B_{2,1}^{s_o}(\mathbb{R}^{d_o})} \\ &\stackrel{(b)}{\lesssim} 2^{-\frac{\mathfrak{K}_S d_x}{2 s_x}} 2^{\frac{\mathfrak{K}_S}{s_o} (s_o - d_o / 2)} \left(\sum_{\ell_o} 1 \right)^{\frac{1}{2}} \\ &\lesssim 2^{\mathfrak{K}_S \left(1 - \frac{d_x}{2 s_x} \right)}, \end{aligned}$$

where we use the Marchaud-type estimate DeVore and Lorentz [1993, Chapter 2, Eq. (10.3)] in (a), we use the sequential Besov norm equivalence DeVore and Popov [1988][Theorem 5.1] (see also [Leisner, 2003, Theorem 3.3.3]) in (b). \square