Nonparametric Instrumental Regression via Kernel Methods is Minimax Optimal

Abstract

We provide a theoretical convergence analysis of ***kernel-based nonparametric instrumental variable*** (NPIV) regression (Singh et al., 2019), a nonparametric approach to IV regression using kernel features in two stages. First, we relax the assumption that the structural function is uniquely identified through the instrument by proposing a minimum norm solution. This relaxation is crucial, as the uniqueness assumption is often violated in practical scenarios. Additionally, we introduce a novel approach to characterize the smoothness of the target function that does not rely on the instrument, instead leveraging a new description of the projected subspace size, which is closely related to the link condition in inverse learning literature. With the subspace size description and under standard kernel learning assumptions, we derive, for the first time, the minimax optimal learning rate for kernel NPIV even when the solution is non-unique. Our results demonstrate that the strength of the instrument, measured by the choice of the kernel, is essential in achieving efficient learning.

1 Introduction

In this paper, we consider the ***nonparametric instrumental variable*** (NPIV) regression problem (Angrist and Imbens, 1995; Newey and Powell, 2003; Darolles et al., 2011). Specifically, let X, Y and Z be three random variables. In the context of IV estimation, X stands for endogenous variables, Y is an outcome variable, and Z stands for exogenous instrument variables. The model can be described as follows

$$Y = h_0(X) + U, \quad \mathbb{E}(U|Z) = 0, \tag{1}$$

where h_0 is the structural function defined in the L_2-space with respect to the marginal distribution of X, denoted as $L_2(X)$. U represents the unmeasured, confounding noise. Alternatively, NPIV is characterized through the following functional equation (Darolles et al., 2011; Bennett et al., 2023b):

$$\mathcal{T}h = r_0, \tag{2}$$

where $r_0(Z) \doteq \mathbb{E}[Y|Z]$ and $\mathcal{T} : L_2(X) \to L_2(Z)$ is the bounded linear operator that maps every $h \in L_2(X)$ to $\mathbb{E}[h(X)|Z] \in L_2(Z)$. The aim of the NPIV estimation is to solve for this possibly under-determined system. The structural function h_0 is uniquely identified as a solution of Eq. (2) if and only if the operator $\mathcal{T}$ is injective (Darolles et al., 2011).

NPIV estimation has many applications, such as causal inference (Angrist and Imbens, 1995; Newey and Powell, 2003), missing data problems (Wang et al., 2014; Miao et al., 2015) and reinforcement learning (Liao et al., 2021; Uehara et al., 2021; Xu et al., 2020). However, NPIV estimation, as an ill-posed inverse problem, is notoriously difficult. For example, the solution of the IV problem often might not be unique given the instrument at hand. Specifically, uniqueness is often violated if instrumental variables are under identified (see e.g., Andrews and

Stock, 2005; Andrews et al., 2019). Hence, many estimation methods require the unique solution assumption (Darolles et al., 2011; Chen and Reiss, 2011; Singh et al., 2019). Moreover, even when the solution is unique, a small perturbation of the outcome can cause the estimator to be infinitely far from the true solution, as $\mathcal{T}^{-1}$ is typically not continuous (Carrasco et al., 2007). There has been a surge in interest to address these difficulties. Existing attempts proposed for NPIV regression can be largely categorized into two classes: conditional moment methods (Muandet et al., 2020; Liao et al., 2020; Dikkala et al., 2020; Bennett et al., 2019, 2023a,c) and two-stage estimation methods (Newey and Powell, 2003; Carrasco et al., 2007; Horowitz, 2011; Darolles et al., 2011; Chen and Christensen, 2015; Chen, 2007; Singh et al., 2019; Xu et al., 2020; Hartford et al., 2017).

Conditional moment methods consider a saddle-point optimization problem of the form $\min_{f\in\mathcal{H}}\max_{g\in\mathcal{G}}\mathcal{L}(f,g)$ for some function classes $\mathcal{H}$ and $\mathcal{G}$, and risk function $\mathcal{L}$. While conditional moment methods can relax the uniqueness assumption on the structural function, obtaining estimators from samples are often difficult since the solutions are typically saddle-points. Moreover, many estimators studied in this direction cannot obtain convergence in the metric of interest: the $L_2(X)$ metric (Bennett et al., 2023a). Finally, the minimax optimality of conditional moment methods is under-explored due to a lack of lower bounds. On the other hand, two-stage methods split the NPIV regression into the following steps. First, depending on the specific method, Stage 1 estimates either the conditional expectation operator $\mathcal{T}$ or the conditional density $P(X \mid Z)$ estimation. Second, Stage 2 then performs a regression of the outcome on the estimators obtained in Stage 1. When both stages involve a least-squares problem, a two-stage method is called a two-stage least-squares (2SLS) regression. Comparing with conditional moment methods, two-stage estimation provides more stable algorithms since it avoids saddle-point optimization. However, existing two-stage methods suffer from the following drawbacks. First, two-stage methods often require the solution of the NPIV to be unique, which may not hold in practical settings. Second, most of the existing two-stage algorithms can only obtain the upper bound on the learning risk, while the information theoretical lower bound is unknown. Third, the instances where minimax optimal learning rate is obtained are restricted to sieve and wavelet estimators (Chen and Reiss, 2011; Chen and Christensen, 2015), there is a need to extend minimax analysis to a wider class of estimators.

In this paper, we focus on the kernel-based 2SLS estimation proposed by Singh et al. (2019). The theory developped in Singh et al. (2019) suffers from the following limitations. First, identification of the structural function is required; second, the learning risk is studied in a pseudo-metric rather than the $L_2(X)$-metric; third, the smoothness of the structural function is expressed with respect to the instrument and is therefore uninterpretable. In light of this, we address the above concerns by making the following contribution.

- *Removing Uniqueness:* As we mentioned before, most NPIV algorithms, including kernel NPIV in Singh et al. (2019), require h_0 to be uniquely identified. In this paper, we relax the uniqueness assumption through introducing the minimum norm solution.
- *Subspace Measure:* For general 2SLS problems including NPIV estimation, Stage 2 learning is often performed in a subspace of the proposed hypothesis space. As such, quantifying the size of the subspace is vital to study the theoretical properties of estimators from 2SLS, in particular kernel NPIV. In light of this, we propose a novel way to measure the size of certain subspaces to study the behavior of the learning risk. Our results reveal that the larger the subspace relative to the constructed hypothesis, the faster the learning rate.
- *Minimax Optimality:* With the help of the subspace measure, we obtain a refined analysis of the learning risk for kernel 2SLS estimators, under standard kernel regression assumptions. In particular, we show that kernel NPIV is minimax optimal for a wide class of estimators under mild assumptions, contrasting with existing methods that only study specific estimators such as sieve or wavelet-based estimators. Moreover, we demonstrate that the kernel NPIV estimator is minimax optimal even when the solution is not unique. To our knowledge, this is the first result that achieve this. All our bound hold in the strong $L_2(X)$-norm. This is in contrast with kernel NPIV in Singh et al. (2019) and other NPIV algorithms (Dikkala et al., 2020, see e.g.,) where they only achieve guarantees in the pseudo-metric $\|\mathcal{T}(\cdot)\|_{L_2(Z)}$.

1.1 Related Works

NPIV is an important estimation problem that has received intensive research interest. Regression-based NPIV learning includes series-based estimators (Ai and Chen, 2003; Hall and Horowitz, 2005; Blundell et al., 2007; Darolles et al., 2011; Chen and Pouzo, 2012; Florens et al., 2011), kernel-based estimators (Hall and Horowitz, 2005; Horowitz, 2007; Singh et al., 2019) and neural networks based estimators (Hartford et al., 2017; Xu et al., 2020; Li et al., 2024b). Recently, there has been a growing interest in employing the min-max optimization approach to solve the NPIV problem, notable examples include Lewis and Syrgkanis (2018); Bennett et al. (2019); Dikkala et al. (2020); Liao et al. (2020); Muandet et al. (2020); Bennett et al. (2022, 2023b).

While there are many algorithms proposed to address the NPIV estimation problem, theoretical understanding of these algorithms is less explored. Among the above mentioned works, Darolles et al. (2011); Singh et al. (2019); Xu et al. (2020); Dikkala et al. (2020); Liao et al. (2020); Bennett et al. (2022); Li et al. (2024b) provide consistency results on the algorithms, either in L_2 rate or in the pseudo-metric sense. However, the learning rates obtained in these papers are often slow with deeper questions such as whether the obtained learning rate is optimal and what is the information-theoretic lower bound remaining unknown.

In econometric literature, Horowitz (2011); Chen and Reiss (2011); Chen and Christensen (2018) demonstrate that their estimator can achieve the minimax optimal rate. However, these papers often suffer the following two problems: first, they require strong assumptions on the instrument such that the structural function can be uniquely identified. As we discussed, uniqueness can be easily violated in many practical scenarios. Secondly, their convergence rate derivation often relies on restrictive assumptions on the data distribution such as uniform marginal distribution. We finally note that when deep learning is embedded into conditional moment methods or 2SLS there is a need to study optimization guarantees. Works in this direction are obtained in Petrulionyte et al. (2024) where they study the joint optimization of 2SLS and in Chen et al. (2024) where they study online stochastic optimization for a conditional moment approach.

In our paper, we show that with or without the uniqueness assumption and under very general kernel learning setting, we can achieve minimax optimal learning rate for a large class of kernel estimators.

2 Background

We introduce the notation and necessary basics of RKHS theory in this section. Most of the background is established in Li et al. (2022a, 2024a); Meunier et al. (2024) and we provide this for ease of reference.

2.1 Notation and Tensor Product of Hilbert Spaces

Throughout the paper, we consider three random variables: X (the covariate), Y (the outcome) and Z (the instrument). Y is defined on $\mathbb{R}$ while X and Z are defined respectively on the second countable locally compact Hausdorff spaces E_X and E_Z endowed with their respective Borel σ-field $\mathcal{F}_{E_X}$ and $\mathcal{F}_{E_Z}$. We let $(\Omega, \mathcal{F}, \mathbb{P})$ be the underlying probability space with expectation operator $\mathbb{E}$. Let P be the pushforward of $\mathbb{P}$ under (X, Y, Z) and π_W for $W \in \{X, Y, Z, (X, Y), (Z, Y), (X, Z)\}$ denotes the marginal distributions. We use the Markov kernel $p : E_Z \times \mathcal{F}_{E_X} \to \mathbb{R}_+$ to define the conditional distribution:

$$\mathbb{P}[X \in A | Z = z] = \int_A p(z, dx),$$

for all $z \in E_Z$ and events $A \in \mathcal{F}_{E_X}$. We denote the space of real-valued Lebesgue square-integrable functions on $(E_X, \mathcal{F}_{E_X})$ with respect to π_X as $L_2(E_X, \mathcal{F}_{E_X}, \pi_X)$, abbreviated $L_2(X)$, and similarly for π_Z as $L_2(E_Z, \mathcal{F}_{E_Z}, \pi_Z)$, abbreviated $L_2(Z)$. We introduce some notation related to linear operators on Hilbert spaces and vector-valued integration; formal definitions can be found in Appendix A for completeness, or we refer the reader to Weidmann (1980); Diestel and Uhl (1977). Let H be a separable real Hilbert space with inner product $\langle\cdot,\cdot\rangle_H$. $L_2(E_Z, \mathcal{F}_{E_Z}, \pi_Z; H)$, abbreviated $L_2(Z; H)$, is the space of strongly $\mathcal{F}_{E_Z} - \mathcal{F}_H$ measurable and Bochner 2-integrable functions from E_Z to H equipped with the norm $\|\cdot\|^2_{L_2(Z;H)} = \int_{E_Z} \|\cdot\|^2_H \, \mathrm{d}\pi_Z$. We write $\mathcal{L}(H, H')$ as the Banach space of bounded linear operators from H to another Hilbert space H', equipped with the operator norm $\|\cdot\|_{H\to H'}$. When $H = H'$, we simply write $\mathcal{L}(H)$ instead. We write $S_2(H, H')$ as the Hilbert space of Hilbert-Schmidt operators from H to H' and $S_1(H, H')$ as the Banach space of trace class operators (see Appendix A for a complete definition). For two Hilbert spaces H, H', we say that H is (continuously) embedded in H' and denote it as $H \hookrightarrow H'$ if H can be interpreted as a vector subspace of H' and the inclusion operator $i : H \to H'$ performing the change of norms with $ix = x$ for $x \in H$ is continuous; and we say that H is isometrically isomorphic to H' and denote it as $H \simeq H'$ if there is a linear isomorphism between H and H' which is an isometry. For any linear operator A, $\mathcal{R}(A)$ denotes its range and $\mathcal{N}(A)$ its null space. For any bounded linear operator A, A^* denotes its adjoint. For any subspace $M \subseteq H$, $M^\perp$ denotes the orthogonal complement.

Tensor Product of Hilbert Spaces (Aubin, 2000, Section 12)**:** Denote $H \otimes H'$ the tensor product of Hilbert spaces H, H'. The element $x \otimes x' \in H \otimes H'$ is treated as the linear rank-one operator $x \otimes x' : H' \to H$ defined by $y' \to \langle y', x'\rangle_{H'} x$ for $y' \in H'$. Based on this identification, the tensor product space $H \otimes H'$ is isometrically isomorphic to the space of Hilbert-Schmidt operators from H' to H, i.e., $H \otimes H' \simeq S_2(H', H)$. We will hereafter not make the distinction between these two spaces, and treat them as being identical.

Remark 1 (Aubin, 2000, Theorem 12.6.1)**.** *Consider the Bochner space $L_2(Z; H)$ where H is a separable Hilbert space. One can show that $L_2(Z; H)$ is isometrically identified with the tensor product space $H \otimes L_2(Z)$, and we denote as Ψ the isometric isomorphism between the two spaces. See Appendix A for more details.*

2.2 Reproducing Kernel Hilbert Spaces and Conditional Mean Embedding

Scalar-valued Reproducing Kernel Hilbert Space (RKHS). We let $k_X : E_X \times E_X \to \mathbb{R}$ be a symmetric and positive definite kernel function and $\mathcal{H}_X$ be a vector space of functions from E_X to $\mathbb{R}$, endowed with a Hilbert space structure via an inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_X}$. We say k_X is a reproducing kernel of $\mathcal{H}_X$ if and only if for all $x \in E_X$ we have $k_X(\cdot, x) \in \mathcal{H}_X$ and for all $x \in E_X$ and $f \in \mathcal{H}_X$, we have $f(x) = \langle f, k_X(x,\cdot)\rangle_{\mathcal{H}_X}$. A space $\mathcal{H}_X$ which possesses a reproducing kernel is called a reproducing kernel Hilbert space (RKHS; Berlinet and Thomas-Agnan, 2011). We denote the canonical feature map of $\mathcal{H}_X$ as $\phi_X(x) = k_X(\cdot, x)$. Similarly for E_Z, we consider a RKHS $\mathcal{H}_Z$ with symmetric and positive definite kernel $k_Z : E_Z \times E_Z \to \mathbb{R}$ and canonical feature map denoted as ϕ_Z.

Assumption 1. *We require some technical assumptions on the previously defined RKHSs and kernels:*

1. *$\mathcal{H}_X$ and $\mathcal{H}_Z$ are separable, this is satisfied if k_X, k_Z are continuous, given that E_X, E_Z are separable*[1]*;*
2. *$k_X(\cdot, x)$ and $k_Z(\cdot, z)$ are measurable for π_X-almost all $x \in E_X$ and π_Z-almost all $z \in E_Z$;*
3. *$k_X(x, x) \leqslant \kappa_X^2$ for π_X-almost all $x \in E_X$ and $k_Z(z, z) \leqslant \kappa_Z^2$ for π_Z-almost all $z \in E_Z$.*

The above assumptions are not restrictive in practice, as well-known kernels such as the Gaussian, Laplace and Matérn kernels satisfy them on $\mathbb{R}^d$ (Sriperumbudur et al., 2011). We now introduce some facts about the interplay between $\mathcal{H}_X$ and $L_2(X)$, which has been extensively studied by Smale and Zhou (2004, 2005), De Vito et al. (2006) and Steinwart and Scovel (2012). We first define the (not necessarily injective) embedding $\mathcal{I}_X : \mathcal{H}_X \to L_2(X)$, mapping a function $f \in \mathcal{H}_X$ to its π_X-equivalence class $[f]_X$. The embedding is a well-defined compact operator since its Hilbert-Schmidt norm can be bounded as (Steinwart and Scovel, 2012,

[1] This follows from Steinwart and Christmann (2008, Lemma 4.33). Note that the lemma requires separability of E_X, E_Z, which is satisfied since we assume that E_X, E_Z are second countable locally compact Hausdorff spaces.

Lemma 2.2 & 2.3) $\|\mathcal{I}_X\|_{S_2(\mathcal{H}_X,L_2(X))} \leqslant \kappa_X$. The adjoint operator $S_X \doteq \mathcal{I}_X^* : L_2(X) \to \mathcal{H}_X$ is an integral operator with respect to the kernel k_X, i.e. for $f \in L_2(X)$ and $x \in E_X$ we have (Steinwart and Christmann, 2008, Theorem 4.27)

$$(S_X f)(x) = \int_{E_X} k_X(x,x') f(x') \,\mathrm{d}\pi_X(x').$$

Next, we define the self-adjoint, positive semi-definite and trace class integral operators

$$L_X \doteq \mathcal{I}_X S_X : L_2(X) \to L_2(X) \quad \text{and} \quad C_X \doteq S_X \mathcal{I}_X : \mathcal{H}_X \to \mathcal{H}_X.$$

By the spectral theorem for self-adjoint compact operators, there exists an at most countable index set I, a non-increasing sequence $(\mu_{X,i})_{i\in I} > 0$, and a family $(e_{X,i})_{i\in I} \in \mathcal{H}_X$, such that $([e_{X,i}]_X)_{i\in I}$ is an orthonormal basis (ONB) of $\overline{\mathcal{R}(\mathcal{I}_X)} \subseteq L_2(X)$, $(\mu_{X,i}^{1/2} e_i)_{i\in I}$ is an ONB of $\mathcal{N}(\mathcal{I}_X)^\perp \subseteq \mathcal{H}_X$, and we have

$$L_X = \sum_{i\in I} \mu_{X,i} \langle \cdot, [e_{X,i}]_X \rangle_{L_2(X)} [e_{X,i}]_X, \qquad C_X = \sum_{i\in I} \mu_{X,i} \langle \cdot, \mu_{X,i}^{\frac{1}{2}} e_{X,i} \rangle_{\mathcal{H}_X} \mu_{X,i}^{\frac{1}{2}} e_{X,i}. \tag{3}$$

We similarly define $\mathcal{I}_Z, S_Z, L_Z, C_Z, (\mu_{Z,i})_{i\in I}, (e_{Z,i})_{i\in I}$ for the RKHS $\mathcal{H}_Z$.

Vector-valued Reproducing Kernel Hilbert Space (vRKHS). Let $K : E_Z \times E_Z \to \mathcal{L}(\mathcal{H}_X)$ be an operator valued positive definite kernel (Carmeli et al., 2006, Definition 2.2). Fix $z \in E_Z$, and $h \in \mathcal{H}_X$, then $(K_z h)(\cdot) \doteq K(\cdot, z)h$ defines a function from E_Z to $\mathcal{H}_X$. The completion of

$$\mathcal{G}_{\text{pre}} \doteq \operatorname{span}\{K_z h \mid z \in E_Z, h \in \mathcal{H}_X\}$$

with inner product on $\mathcal{G}_{\text{pre}}$ defined on the elementary elements as $\langle K_z h, K_{z'} h' \rangle_{\mathcal{G}} \doteq \langle h, K(z,z') h' \rangle_{\mathcal{Y}}$, defines a vRKHS denoted as $\mathcal{G}$. For a more complete overview of the vector-valued reproducing kernel Hilbert space, we refer the reader to Carmeli et al. (2006), Carmeli et al. (2010) and Li et al. (2023, Section 2). In the following, we will denote $\mathcal{G}$ as the vRKHS induced by the kernel $K : E_Z \times E_Z \to \mathcal{L}(\mathcal{H}_X)$ with

$$K(z,z') \doteq k_Z(z,z') \operatorname{Id}_{\mathcal{H}_X}, \quad z, z' \in E_Z. \tag{4}$$

We emphasize that this family of kernels is the de-facto standard for high- and infinite-dimensional applications (Grünewälder et al., 2012b,a; Park and Muandet, 2020; Ciliberto et al., 2016, 2020; Singh et al., 2019; Mastouri et al., 2021; Kostic et al., 2022, 2023) due to the crucial *representer theorem* which gives a closed-form solution to estimators derived from this family of kernels.

Remark 2 (General multiplicative kernel). *Without loss of generality, we provide our results for the vRKHS $\mathcal{G}$ induced by the operator-valued kernel given by $K = k_Z \operatorname{Id}_{\mathcal{H}_X}$. However, with suitably adjusted constants in the assumptions, our results transfer directly to the more general vRKHS $\widetilde{\mathcal{G}}$ induced by the more general operator-valued kernel*

$$\widetilde{K}(z,z') \doteq k_Z(z,z') Q, \quad z, z' \in E_Z,$$

where $Q : \mathcal{H}_X \to \mathcal{H}_X$ is any positive-semidefinite self-adjoint operator. The precise characterization of the adjusted constants is given by Li et al. (2023, Section 4.1).

An important property of $\mathcal{G}$ is that it is isometrically isomorphic to the space of Hilbert-Schmidt operators between $\mathcal{H}_Z$ and $\mathcal{H}_X$ (Li et al., 2023, Corollary 1). Similarly to the scalar case, we can map every element in $\mathcal{G}$ into its π_Z−equivalence class in $L_2(Z; \mathcal{H}_X)$ and we use the shorthand notation $[F] = [F]_{Z;X}$ (see Definition 6 in Appendix A for more details).

Theorem 1 (vRKHS isomorphism). *(Li et al., 2023, Corollary 1) For every function $F \in \mathcal{G}$ there exists a unique operator $C \in S_2(\mathcal{H}_Z, \mathcal{H}_X)$ such that $F(\cdot) = C\phi_Z(\cdot) \in \mathcal{H}_X$ with $\|C\|_{S_2(\mathcal{H}_Z,\mathcal{H}_X)} = \|F\|_{\mathcal{G}}$ and vice versa. Hence $\mathcal{G} \simeq S_2(\mathcal{H}_Z, \mathcal{H}_X)$ and we denote the isometric isomorphism between $S_2(\mathcal{H}_Z, \mathcal{H}_X)$ and $\mathcal{G}$ as $\bar{\Psi}$. It follows that $\mathcal{G}$ can be written as $\mathcal{G} = \{F : E_Z \to \mathcal{H}_X \mid F = C\phi_Z(\cdot), C \in S_2(\mathcal{H}_Z, \mathcal{H}_X)\}$.*

Conditional Mean Embedding: A particular advantage of kernel methods is its convenience of operating probability distributions (see e.g., Muandet et al., 2017; Sejdinovic et al., 2013). In particular, kernel methods allow us to deal with the conditional distributions through conditional mean embedding, defined in Park and Muandet (2020); Klebanov et al. (2020).

Definition 1. *The $\mathcal{H}_X$-valued conditional mean embedding (CME) for the conditional distribution of X given Z, is defined as*

$$F_*(\cdot) \doteq \int_{E_X} \phi_X(x) p(\cdot, dx) = \mathbb{E}\left[\phi_X(X) | Z = \cdot\right] \in L_2(Z; \mathcal{H}_X).$$

By the reproducing property, we have $\mathbb{E}[f(X)|Z = z] = \langle f, F_*(z)\rangle_{\mathcal{H}_X}$, for all $f \in \mathcal{H}_X$ and $z \in E_Z$. Therefore, the CME allows us to easily compute conditional expectations through an inner product, which as we will see below is a crucial step in kernel NPIV. The approximation of F_* with Tikhonov regularization (also known as vector-valued kernel ridge regression) is a key concept in kernel methods and is extensively studied in Park and Muandet (2020); Li et al. (2022b, 2023), where learning F_* is formulated as the following optimization problem at the population level:

$$F_\xi \doteq \underset{F \in \mathcal{G}}{\arg\min}\, \mathbb{E}_{X,Z} \left\|\phi_X(X) - F(Z)\right\|^2_{\mathcal{H}_X} + \xi \|F\|^2_{\mathcal{G}}, \tag{5}$$

which can be expressed in closed-form as

$$F_\xi(\cdot) \doteq C_\xi \phi_Z(\cdot) \in \mathcal{G}, \qquad C_\xi \doteq C_{XZ}\left(C_Z + \xi \operatorname{Id}_{\mathcal{H}_Z}\right)^{-1} \in S_2(\mathcal{H}_Z, \mathcal{H}_X).$$

Tikhonov regularization in Eq. (5) is known to exhibit the ***saturation effect*** where it fails to benefit from high smoothness of the target function F_*. This was recently verified in Meunier et al. (2024) in the context of vector-valued regression. To avoid saturation, we therefore generalize Tikhonov regularization to general spectral regularization (see also Mollenhauer and Koltai (2020); Mollenhauer et al. (2022)). General spectral regularization typically starts with defining a filter function, i.e., a function on an interval which is applied on self-adjoint operators to each individual eigenvalue via the spectral calculus.

Definition 2 (Filter function)**.** *Let $\Lambda \subseteq \mathbb{R}^+$. A family of functions $g_\xi : [0, \infty) \to [0, \infty)$ indexed by $\xi \in \Lambda$ is called a filter with qualification $\rho \geqslant 0$ if it satisfies the following two conditions:*

1. *There exists a positive constant E such that, for all $\xi \in \Lambda$*

$$\sup_{\theta \in [0,1]} \sup_{x \in [0, \kappa_Z^2]} \xi^{1-\theta} x^\theta g_\xi(x) \leqslant E$$

2. *There exists a positive constant $\omega_\rho < \infty$ such that*

$$\sup_{\theta \in [0,\rho]} \sup_{\xi \in \Lambda} \sup_{x \in [0, \kappa_Z^2]} |1 - g_\xi(x)x| x^\theta \xi^{-\theta} \leqslant \omega_\rho.$$

One may think of the above definition as a class of functions approximating the inversion map $x \mapsto 1/x$ while still being defined for $x = 0$ in a reasonable way. As examples, with $g_\xi(x) = (x + \xi)^{-1}$, we retrieve ridge regression with $\rho = 1$ and with $g_\xi(x) = x^{-1}\mathbb{1}[x \geqslant \xi]$ we obtain kernel principal component regression with $\rho = +\infty$. We refer to Appendix A.3 for other examples of spectral methods such as gradient descent, iterated Tikhonov and gradient flow.

Given a filter function g_ξ, we call $g_\xi(C_Z)$ the regularized inverse of C_Z. We may think of the regularized inverse as approximating the *pseudoinverse* of C_Z (see e.g. Engl et al. (2000)) when $\xi \to 0$. We define the regularized population solution with filter function g_ξ as

$$C_\xi \doteq C_{XZ} g_\xi(C_Z) \in S_2(\mathcal{H}_Z, \mathcal{H}_X), \qquad F_\xi(\cdot) \doteq C_\xi \phi_Z(\cdot) \in \mathcal{G}. \tag{6}$$

Empirical solution: Given a dataset $\mathcal{D}_1 = \{(\tilde{z}_i, \tilde{x}_i)\}_{i=1}^m$, the empirical analogue of Eq. (6) is

$$\hat{C}_\xi \doteq \hat{C}_{XZ} g_\xi(\hat{C}_Z), \qquad \hat{F}_\xi(\cdot) \doteq \hat{C}_\xi \phi_Z(\cdot) \in \mathcal{G}, \tag{7}$$

where $\hat{C}_{XZ}$, $\hat{C}_Z$ are empirical covariance operators define as

$$\hat{C}_Z \doteq \frac{1}{m}\sum_{i=1}^m \phi_Z(\tilde{z}_i) \otimes \phi_Z(\tilde{z}_i) \qquad \hat{C}_{XZ} \doteq \frac{1}{m}\sum_{i=1}^m \phi_X(\tilde{x}_i) \otimes \phi_Z(\tilde{z}_i).$$

The closed-form formula is obtained in Meunier et al. (2024) using a generalization of the ***representer theorem***. It will allow us to obtain a closed-form expression for our estimator of the structural function (see Section 3.1).

3 Instrumental Variable Regression with RKHS

In this section, we illustrate how kernel-based algorithms can solve the NPIV problem. We recall that a solution to NPIV satisfies the following functional equation:

$$\mathcal{T} h = r_0.$$

A common condition required in the NPIV literature is that $r_0 \in \mathcal{R}(\mathcal{T})$, i.e., the true structural function is in the range of the conditional expectation operator. It simply states that there exists at least one solution satisfying the integral equation. However, even equipped with this assumption, NPIV remains a challenge because it is an ill-posed inverse problem. In particular, the solution to the above equation might not be unique. To address this, most of existing literature assumes injective $\mathcal{T}$, enforcing the solution to be unique. However, injectivity is a strong assumption as it may not hold in many practical settings. Specifically, uniqueness is violated if we have weak instrumental variables (see e.g., Andrews and Stock, 2005; Andrews et al., 2019). For example, when both X and Z are discrete random variables and the cardinality of X exceeds Z, we cannot have injective $\mathcal{T}$. Kallus et al. (2021) further demonstrates various examples where uniqueness is violated in the proximal causal inference setting. When $\mathcal{T}$ is not injective, the NPIV problem becomes even more ill-posed and obtaining an L_2-convergence rate is much more difficult. In particular, it is even not clear whether existing estimators converge to any particular solution, since there might be infinitely many solutions. Consequently, recent works (see e.g., Chen, 2021; Bennett et al., 2023a,c; Li et al., 2024b) in NPIV propose to consider the solution that achieves the least L_2-norm, which is shown to be uniquely defined (e.g., Lemma 5 Bennett et al., 2023c). Following this direction, in our work, we propose to target the solution with the minimum RKHS norm. Specifically, let $\tilde{\mathcal{T}} \doteq \mathcal{T} \circ \mathcal{I}_X : \mathcal{H}_X \to L_2(Z)$ and denote

$$\mathcal{N}_{r_0}(\tilde{\mathcal{T}}) \doteq \{h \in \mathcal{H}_X : \tilde{\mathcal{T}} h = r_0\}$$

Assumption 2 (Well-specifiedness of solutions)**.** $\mathcal{N}_{r_0}(\tilde{\mathcal{T}}) \neq \varnothing$.

The RKHS $\mathcal{H}_X$ encompasses our a priori belief on the properties that should satisfy the structural function. Assumption 2 states that there is at least one function in $\mathcal{H}_X$ satisfying the integral equation. Note that Assumption 2 is stronger than $r_0 \in \mathcal{R}(\mathcal{T})$ as $\mathcal{H}_X$ can be seen as a subset of $L_2(X)$. However, for a universal RKHS, $\mathcal{H}_X$ is dense in $L_2(X)$ under L_∞ norm (see e.g., Steinwart and Christmann, 2008, Chapter 4.6). Since $\mathcal{T}$ is not guaranteed to be injective, $\tilde{\mathcal{T}}$ is also not guaranteed to be injective. The minimum RKHS norm solution is then defined as

$$h_* \doteq \operatorname*{arg\,min}_{h \in \mathcal{N}_{r_0}(\tilde{\mathcal{T}})} \|h\|_{\mathcal{H}_X}. \tag{8}$$

h_* can also be seen as the pseudo-inverse of the linear system: $h_* = \left(\tilde{\mathcal{T}}\right)^\dagger r_0$. The next proposition shows that h_* is uniquely defined, the proof is postponed to Appendix B.

Proposition 1. *Under Assumption 2, h_* uniquely exists and $\{h_*\} = \mathcal{N}(\tilde{\mathcal{T}})^{\perp} \cap \mathcal{N}_{r_0}(\tilde{\mathcal{T}})$.*

Remark 3. *We remark that our construction guarantees the uniqueness of the target solution for both injective and non-injective $\mathcal{T}$ i.e., h_*. We will see that the kernel-based NPIV algorithm ensures its estimator converging to h_* under both scenarios. In this way, we relax the strong injectivity assumption imposed in the classical NPIV literature (see e.g., Newey and Powell, 2003; Chen and Reiss, 2011; Darolles et al., 2011, and references therein).*

3.1 The Kernel NPIV Estimator

Under Assumption 2, $h_* \in \mathcal{H}_X$, and using the reproducing property, we see that for all $z \in E_Z$,

$$\left[\tilde{\mathcal{T}} h_*\right](Z) = \mathbb{E}\left[\langle h_*, \phi_X(X)\rangle_{\mathcal{H}_X} \mid Z\right] = \langle h_*, F_*(Z)\rangle_{\mathcal{H}_X}. \tag{9}$$

Singh et al. (2019) then suggests a two-stage least-squares estimation procedure with a sample splitting strategy, where we use m samples for the stage 1 regression and n samples for the stage 2 regression.

1. estimate F_* with vector-valued regression using dataset $\mathcal{D}_1 \doteq \{(\tilde{z}_i, \tilde{x}_i)\}_{i=1}^m$ and spectral regularization;
2. estimate h_* through regressing Y on $F_*(Z)$ using dataset $\mathcal{D}_2 \doteq \{(z_i, y_i)\}_{i=1}^n$.

Instead of using Tiknohov regularization for stage 1 as in Singh et al. (2019), below we employ a learning procedure with general spectral algorithms for stage 1.

Stage 1. Using $\mathcal{D}_1$ we apply Eq. (7) to obtain the empirical estimator $\hat{F}_\xi$.

Stage 2. The algorithm at the population level can be written as

$$\begin{aligned} h_\lambda &\doteq \underset{h \in \mathcal{H}_X}{\arg\min}\, \mathbb{E}_{Y,Z}\left(Y - \tilde{\mathcal{T}} h(Z)\right)^2 + \lambda \|h\|^2_{\mathcal{H}_X} \\ &= \underset{h \in \mathcal{H}_X}{\arg\min}\, \mathbb{E}_{Y,Z}\left(Y - \langle h, F_*(Z)\rangle_{\mathcal{H}_X}\right)^2 + \lambda \|h\|^2_{\mathcal{H}_X} \end{aligned} \tag{10}$$

Empirically, we use the estimated $\hat{F}_\xi$ from stage 1 to learn h_* with $\mathcal{D}_2$

$$\hat{h}_\lambda \doteq \underset{h \in \mathcal{H}_X}{\arg\min}\, \frac{1}{n}\sum_{i=1}^n \left(y_i - \left\langle h_*, \hat{F}_\xi(z_i)\right\rangle_{\mathcal{H}_X}\right)^2 + \lambda \|h\|^2_{\mathcal{H}_X}.$$

Kernel NPIV admits a closed-form solution as derived in Singh et al. (2019). We provide a new version with spectral algorithms in Appendix D.

We also introduce $\bar{h}_\lambda$, a theoretical estimator for stage 2 that would access the true CME:

$$\bar{h}_\lambda = \underset{h \in \mathcal{H}_X}{\arg\min}\, \frac{1}{n}\sum_{i=1}^n \left(y_i - \langle h, F_*(z_i)\rangle_{\mathcal{H}_X}\right)^2 + \lambda \|h\|^2_{\mathcal{H}_X}, \tag{11}$$

Remark 4 (Spectral Algorithm)**.** *We remark that one can employ the spectral regularization for stage 2, instead of Tikhonov regularization. However, the interplay between the qualification of the spectral regularization with our smoothness assumptions (see Assumption SRCX below) is far from trivial. We therefore leave this investigation for future work. A first step in that direction is obtained in Bennett et al. (2023c) where they study how iterated Tikhonov regularization can be incorporated in a conditional moment model.*

4 Measure of Subspace Size

Our next step is to characterize the behavior of the finite sample estimator $\hat{h}_\lambda$, i.e., $\|\hat{h}_\lambda - h_*\|_{L_2(X)}$. To this end, we first note that stage 1 in 2SLS aims at estimating the CME F_*. The CME F_* induces a subspace of $\mathcal{H}_X$. Specifically, if we introduce the following covariance operator

$$C_F \doteq \mathbb{E}[F_*(Z) \otimes F_*(Z)], \tag{12}$$

we then have $C_F = \tilde{\mathcal{T}}^* \tilde{\mathcal{T}}$ and $\overline{\mathcal{R}(C_F)} \subseteq \overline{\mathcal{R}(C_X)}$ since $C_F \preceq C_X$ by Jensen's inequality[2]. Therefore, by Proposition 1, $h_* \in \mathcal{N}(\tilde{\mathcal{T}})^\perp = \overline{\mathcal{R}(\tilde{\mathcal{T}}^*)} = \overline{\mathcal{R}(C_F)} \subseteq \overline{\mathcal{R}(C_X)}$. As estimating h_* is conducted in $\overline{\mathcal{R}(C_F)}$, the size of the subspace $\overline{\mathcal{R}(C_F)}$ plays a pivotal role in the stage 2 regression learning. Intuitively, if the subspace is large with respect to $\mathcal{H}_X$, we should expect a fast learning rate in stage 2. As such, we propose a novel measure of subspace size to formalize this intuition.

Assumption 3. *Define $\gamma_0, \gamma_1 \in [1, +\infty)$. We say that C_F and C_X satisfy the link condition, written* LINK(γ_0, γ_1) *if:*

$$\begin{aligned} R_0 \|C_X^{\gamma_0/2} f\|_{\mathcal{H}_X} &\leqslant \|C_F^{1/2} f\|_{\mathcal{H}_X}, \quad \forall f \in \overline{\mathcal{R}(C_F)}, \\ \|C_F^{1/2} f\|_{\mathcal{H}_X} &\leqslant R_1 \|C_X^{\gamma_1/2} f\|_{\mathcal{H}_X}, \quad \forall f \in \mathcal{H}_X, \end{aligned} \tag{LINK}$$

for some universal constants $R_0, R_1 \geqslant 0$.

Notice that, since $C_F \preceq C_X$, we can always take $\gamma_1 = 1$. However, there are some settings where it is not possible to find $\gamma_0 < +\infty$, for example if C_X and C_F share the same eigenvectors with respective eigenvalues $\lambda_{X,i} = i^{-2}$ and $\lambda_{F,i} = e^{-i}$, $i \geqslant 1$. This specific scenario is refereed to as the "severely ill-posed" setting and is tackled for example in (Theorem 1 Chen and Reiss, 2011). In this work, however, we only focus on the mildly ill-posed setting.

Remark 5 (Subspace size)**.** *In kernel regression, we often use the eigenvalue decay rate of C_X to describe the size of the RKHS $\mathcal{H}_X$ (see e.g., Steinwart et al., 2009; Fischer and Steinwart, 2020, and references therein). The faster the decay rate, the smaller the space $\mathcal{H}_X$. The eigenvalue decay rate therefore determines the kernel regression learning rate.*

In our definition, the pair (γ_0, γ_1) characterizes the eigenvalue decay rate of C_F, relative to the decay rate of C_X. When both γ_0 and γ_1 are large, the eigenvalue decay is fast. The pair (γ_0, γ_1) hence can be understood as describing precisely the size of the subspace $\overline{\mathcal{R}(C_F)}$. We will see in next section that the final learning rate of kernel NPIV depends on both γ_0 and γ_1. Finally, since the important parameters are γ_0 and γ_1 in describing the subspace size, from now on, we will without loss of generality assume that $R_0 = R_1 = 1$.

Remark 6 (Link condition in inverse problem)**.** *We point out that Assumption (LINK) is closely related to the link condition widely used in the inverse problem literature (see e.g., Chen and Reiss, 2011; Nair et al., 2005, and references therein). In particular, Theorem 2 in Chen and Reiss (2011) essentially makes the following link assumption to obtain minimax optimal rate*

$$\|C_F^{1/2} f\|_{\mathcal{H}_X} = M_1 \|\varphi(C_X)^{1/2} f\|_{\mathcal{H}_X}, \quad \forall f \in \mathcal{H}_X \tag{13}$$

where $\varphi(\cdot) : \mathbb{R} \to \mathbb{R}$ is a function that applies pointwise to the spectrum of C_X. For example, we retrieve our link assumption if $\varphi(x) = x^{\gamma_1}$ and $\gamma_0 = \gamma_1$. Therefore, the link conditions used in the inverse problem is similar to our (LINK). However, we do point out two important differences, which make our assumption less restrictive.

First, in all previous works, C_F is assumed to be injective. This is because these works impose that the solution to NPIV is unique. However, we do not have such a restriction. This is because we aim to study the convergence

[2] $\preceq$ denotes the partial order of self-adjoint operators.

of NPIV for both the injective and non-injective case. In order to do so, we adapt the link condition to take into account potential non-injectivity. Specifically, for the inequality involving γ_0, we only require that the condition holds for the functions $f \in \overline{\mathcal{R}(C_F)}$. In this way, we enable the study of the convergence for non-injective C_F.

Secondly, as pointed out in Chen and Reiss (2011), there might not be a single function φ such that the equality holds in (13). Even when C_F is injective, it might be the case that we need φ_1 and φ_2 to upper and lower bound $\|C_F^{1/2} f\|_{\mathcal{H}_X}$. This is precisely why we need the (γ_0, γ_1) pair to characterize the eigenspectrum of C_F with respect to C_X.

The next proposition demonstrates some key properties of our proposed subspace size measure and plays a pivotal role in deriving the kernel NPIV learning rate. The proof is postponed to Appendix B.

Proposition 2. *Let P_F be the orthogonal projection on $\overline{\mathcal{R}(C_F)}$. Under Assumption (LINK) with $\gamma_0, \gamma_1 \in [1, +\infty)$, we have the following properties*

a) $1 \leqslant \gamma_1 \leqslant \gamma_0$.

b) γ_0, γ_1 *can be equivalently characterized respectively as* $P_F C_X^{\gamma_0} P_F \preceq C_F$ *and* $C_F \preceq C_X^{\gamma_1}$.

c) γ_0, γ_1 *can be equivalently characterized respectively as* $\mathcal{R}(C_F^{1/2}) \subseteq \mathcal{R}(C_X^{\gamma_1/2})$ *and* $\mathcal{R}\left(\left(P_F C_X^{\gamma_0} P_F\right)^{1/2}\right) \subseteq \mathcal{R}(C_F^{1/2})$.

5 Minimax Optimal Learning Rates

In this section, we establish the minimax optimality of the kernel NPIV estimator for both injective and non-injective $\mathcal{T}$. Before listing our assumptions below, we first briefly introduce the interpolation space- a concept which is often used to deal with misspecified learning. The readers are referred to Appendix A.4 for full details.

We start with scalar-valued functions. Given $\beta \geqslant 0$ and a squared-integrable scalar-valued function $f \in L_2(Z)$, the $\beta-$interpolation norm is defined as

$$\|f\|_\beta \doteq \|L_Z^{-\beta/2} f\|_{L_2(Z)}.$$

The subset of $f \in L_2(Z)$ for which $\|f\|_\beta < +\infty$ is denoted $[\mathcal{H}_Z]^\beta$. $[\mathcal{H}_X]^\beta \subseteq L_2(X)$ is defined similarly with L_X.

Vector-valued interpolation norms and spaces introduced in Li et al. (2022a) generalize the above interpolation space definitions to spaces of vector-valued functions. Given $\beta \geqslant 0$ and a vector-valued function $F \in L_2(Z; \mathcal{H}_X)$ since $L_2(Z; \mathcal{H}_X)$ is isometric to $S_2(L_2(Z), \mathcal{H}_X)$ (see e.g., Corollary 1 in Li et al., 2023), there is an operator $C \in S_2(L_2(Z), \mathcal{H}_X)$ such that $\|F\|_{L_2(Z;\mathcal{H}_X)} = \|C\|_{S_2(L_2(Z),\mathcal{H}_X)}$. The vector-valued $\beta-$interpolation norm is then defined as

$$\|F\|_\beta \doteq \|C\|_\beta \doteq \|C L_Z^{-\beta/2}\|_{S_2(L_2(Z),\mathcal{H}_X)}. \tag{14}$$

The space of $F \in L_2(Z; \mathcal{H}_X)$ such that $\|F\|_\beta < +\infty$ is denoted $[\mathcal{G}]^\beta$. For details regarding vector-valued interpolation spaces, we refer to Appendix A.4.

5.1 Assumptions for Stage 1

The analysis of stage 1 convergence is essentially studied in Li et al. (2022a, 2024a). We here provide the results to ease our discussion later.

Assumption 4 (Eigenvalue decay for stage 1)**.** *For some constants $c_1 > 0$ and $p_Z \in (0,1]$ and for all $i \in I$,*

$$\mu_{Z,i} \leqslant c_1 i^{-1/p_Z}. \tag{EVDZ}$$

Assumption 5 (Embedding into L_∞ for stage 1)**.** *There exists $\alpha_Z \in [p_Z, 1]$ such that the inclusion map $\mathcal{I}_Z^{\alpha_Z,\infty} : [\mathcal{H}]_Z^{\alpha_Z} \hookrightarrow L_\infty(Z)$ is continuous and there is a constant $A_Z > 0$ such that,*

$$\|\mathcal{I}_Z^{\alpha_Z,\infty}\|_{[\mathcal{H}_Z]^{\alpha_Z} \to L_\infty(Z)} = A_Z. \tag{EMBZ}$$

Assumption 6 (Source condition for stage 1)**.** *There exists $\beta_Z \geqslant \alpha_Z$ and a constant $B_Z \geqslant 0$ such that,*

$$\|F_*\|_{\beta_Z} = \|C_* L_Z^{-\frac{\beta_Z}{2}}\|_{S_2(L_2(Z),\mathcal{H}_X)} \leqslant B_Z, \tag{SRCZ}$$

where $C_* \doteq \Psi^{-1}(F_*) \in S_2(L_2(Z),\mathcal{H}_X)$ (see Remark 1 for the definition of Ψ).

(EVDZ) is a classical assumption that characterizes the size of the RKHS $\mathcal{H}_Z$ equipped with the marginal distribution π_Z. (SRCZ) characterizes the smoothness of the target function F_*. Property (EMBZ) is referred to as the *embedding property* in Fischer and Steinwart (2020). It can be shown that it holds if and only if there exists a constant $A_Z \geqslant 0$ with $\sum_{i \in I} \mu_i^\alpha e_{Z,i}^2(z) \leqslant A_Z^2$ for π-almost all $z \in E_Z$ (Fischer and Steinwart, 2020, Theorem 9). Since we assume k_Z to be bounded, the embedding property always hold true with $\alpha_Z = 1$. Furthermore, (EMBZ) implies a polynomial eigenvalue decay of order $1/\alpha_Z$, which is why we take $\alpha_Z \geqslant p_Z$. For an in-depth discussion of these assumptions, we refer the reader to Li et al. (2023). Under (EVDZ), (SRCZ), (EMBZ), Meunier et al. (2024) demonstrates that the estimator in Eq. (7) converges to F_*. The following informal result is from Meunier et al. (2024, Theorem 4), we refer to Theorem 15 in Appendix H for the formal statement. The L_2−rate is tight as it matches the lower bound provided in Li et al. (2023). Finally, it is important to note that the convergence with minimax rate can still be maintained in the misspecified regime, indeed if $\alpha_Z \leqslant \beta_Z < 1$, then $F_* \notin \mathcal{G}$.

Theorem 2. *Let g_ξ be a filter function with qualification $\rho \geqslant 1$ used to build the estimator $\hat{F}_\xi$ on $\mathcal{D}_1$ with Eq. (7). Let Assumptions 1, (EVDZ), (SRCZ) and (EMBZ) hold with $\beta_Z \in (\alpha_Z, 2\rho]$ and $0 < p_Z \leqslant \alpha_Z \leqslant 1$. Taking $\xi_m = \Theta\left(m^{-\frac{1}{\beta_Z + p_Z}}\right)$, there are constants $J, J' > 0$ such that with high probability,*

$$\left\|\hat{F}_\xi - F_*\right\|_{L_2(Z,\mathcal{H}_X)}^2 \leqslant J m^{-\frac{\beta_Z}{\beta_Z + p_Z}} \quad \& \quad \left\|\hat{F}_\xi - F_*\right\|_{L_\infty(Z,\mathcal{H}_X)}^2 \leqslant J' m^{-\frac{\beta_Z - \alpha_Z}{\beta_Z + p_Z}}.$$

We draw a comparison between Theorem 2 and the stage 1 rate achieved in Singh et al. (2019) in the context of Tikhonov regularization. Theorem 2 is more general in the following ways. First, instead of assuming well-specified setting in Singh et al. (2019), Theorem 2 allows for a more general misspecified setting where $F_* \notin \mathcal{G}$. Secondly, obtained rates in Singh et al. (2019) are in RKHS norm ($\|\cdot\|_{\mathcal{G}}$) only. Finally, learning rates in Singh et al. (2019) is slow in the sense that they did not consider the eigenvalue decay of the operator C_Z (which corresponds to setting $p_Z = 1$ in (EVDZ)).

5.2 Assumptions for Stage 2

Assumption 7 (Eigenvalue decay for stage 2)**.** *For some constants $c_2 > 0$ and $p_X \in (0,1]$ and for all $i \in I$,*

$$\mu_{X,i} \leqslant c_2 i^{-1/p_X}. \tag{EVDX}$$

Assumption 8 (Source condition for stage 2)**.** *There exists $\beta_X \geqslant 1$ and a constant $B_X \geqslant 0$ such that*

$$\|h_*\|_{\beta_X} = \left\|L_X^{-\frac{\beta_X}{2}} h_*\right\|_{L_2(X)} \leqslant B_X. \tag{SRCX}$$

Assumption 9 (MOM)**.** *There are constants $\sigma, L > 0$ such that*

$$\mathbb{E}\left[|Y - \mathbb{E}[h_*(X) \mid Z]|^m \mid Z\right] \leqslant \frac{1}{2} m! \sigma^2 L^{m-2}. \tag{MOM}$$

(EVDX) and (SRCX) plays the same role as for stage 1, the former charaterizes the size of the space $\mathcal{H}_X$ equipped with the marginal distribution π_X while the latter characterizes the smoothness of the target funciton h_*. Note that (SRCX) can be equivalently stated as $\left\| C_X^{-\frac{\beta_X - 1}{2}} h_* \right\|_{\mathcal{H}_X} \leqslant B_X$. Finally, the (MOM) condition on the conditional distribution is a Bernstein moment condition used to control the noise of the observations (see Caponnetto and De Vito, 2007; Fischer and Steinwart, 2020 for more details).

Remark 7 (Misspecified Setting)**.** *Our* (SRCX) *requires $\beta_X \geqslant 1$, corresponding to the well-specified case, e.g., $h_* \in \mathcal{H}_X$. We conjecture that convergence of $\hat{h}_\lambda$ to h_* is still achievable in the mis-specified setting where $h_* \notin \mathcal{H}_X$. However, a rigorous theoretical investigation is left as future work.*

Under these assumptions, our next theorem provides an upper bound on the learning risk $\|\hat{h}_\lambda - h_*\|_{L_2(X)}$. The proof is in Appendix E.

Theorem 3. *Let Assumptions 1, 2,* (EVDX)*,* (SRCX)*,* (MOM) *and* (LINK) *hold with $p_X \in (0, 1]$ and $1 \leqslant \beta_X \leqslant \gamma_0 + 1$ and let Assumptions* (SRCZ) *and* (EMBZ) *hold with $\alpha_Z \leqslant \beta_Z$. For any $\tau \geqslant 1$, $\lambda > 0$ and sufficiently large m and n, with P^n-probability over $1 - 12e^{-\tau}$*

$$\begin{aligned} \|\hat{h}_\lambda - h_*\|_{L_2(X)} \leqslant J_0 \lambda^{\frac{1 - c_F}{2\gamma_0} - 1} \left(\left\| F_* - \hat{F}_\xi \right\|_{L_2(Z; \mathcal{H}_X)} + \frac{\left\| \hat{F}_\xi - F_* \right\|_{\alpha_Z}}{\sqrt{n}} \right) \left(\left\| \bar{h}_\lambda \right\|_{\mathcal{H}_X} + 1 \right) \\ + J_1 \left(\lambda^{\frac{\beta_X}{2\gamma_0}} + \sqrt{1 + \frac{1}{n \lambda^{1 - \frac{p_X}{\gamma_1}}}} \sqrt{\frac{1}{n \lambda^{1 - \frac{1}{\gamma_0} + \frac{p_X}{\gamma_1}}}} \right) \end{aligned}$$

where J_0, J_1 only depend on $\sigma, L, A_Z, B_Z, \kappa_Z, \alpha_Z, \beta_Z, p_X, \kappa_X, B_X$, and $c_F \doteq 1_{\mathcal{N}(C_F) \neq \{0\}}$.

Theorem 3 provides a detailed upper bound of $\|\hat{h}_\lambda - h_*\|_{L_2(X)}$. While the first term on the r.h.s corresponds to the generalization error in stage 1, the second term is due to stage 2. The mathematical statement hidden behind "for sufficiently large m and n is provided in Appendix E, Theorem 6. We provide specific learning rates depending on the ratio of stage 1 and 2 samples in the next corollary.

Corollary 1. *Let the assumptions of Theorem 3 hold together with Assumption* (EVDZ)*. Let $a > 0$ control the trade-off between stage 1 and stage 2 samples: $m = n^a$. Let $\xi_m = \Theta\left(m^{-\frac{1}{\beta_Z + p_Z}}\right)$. We consider two different scenarios.*

Case A. $\alpha_Z(\beta_X - 1 + 2\gamma_0 + c_F) \leqslant \beta_Z(\beta_X - 1 + \gamma_0 + \frac{\gamma_0}{\gamma_1} p_X)$.

- $a \geqslant \frac{\beta_Z + p_Z}{\beta_Z} \frac{\beta_X - 1 + 2\gamma_0 + c_F}{\beta_X - 1 + \gamma_0 + \frac{\gamma_0}{\gamma_1} p_X}$, $\lambda = \Theta\left(n^{-\frac{\gamma_0}{\beta_X - 1 + \gamma_0 + \frac{\gamma_0}{\gamma_1} p_X}}\right)$, $\|\hat{h}_\lambda - h\|^2_{L_2(X)} = O_P\left(n^{-\frac{\beta_X}{\beta_X - 1 + \gamma_0 + \frac{\gamma_0}{\gamma_1} p_X}}\right)$.
- $a \leqslant \frac{\beta_Z + p_Z}{\beta_Z} \frac{\beta_X - 1 + 2\gamma_0 + c_F}{\beta_X - 1 + \gamma_0 + \frac{\gamma_0}{\gamma_1} p_X}$, $\lambda = \Theta\left(n^{-a \cdot \frac{\beta_Z}{\beta_Z + p_Z} \cdot \frac{2\gamma_0}{\beta_X - 1 + 2\gamma_0 + c_F}}\right)$, $\|\hat{h}_\lambda - h\|^2_{L_2(X)} = O_P\left(n^{-a \cdot \frac{\beta_Z}{\beta_Z + p_Z} \cdot \frac{\beta_X}{\beta_X - 1 + 2\gamma_0 + c_F}}\right)$.

Case B. $\alpha_Z(\beta_X - 1 + 2\gamma_0 + c_F) \geqslant \beta_Z(\beta_X - 1 + \gamma_0 + \frac{\gamma_0}{\gamma_1} p_X)$

- $a \geqslant \frac{\beta_Z+p_Z}{\beta_Z-\alpha_Z}\frac{\gamma_0\left(1-\frac{p_X}{\gamma_1}\right)+c_F}{\beta_X-1+\gamma_0+\frac{\gamma_0}{\gamma_1}p_X}$, $\lambda = \Theta\left(n^{-\frac{\gamma_0}{\beta_X-1+\gamma_0+\frac{\gamma_0}{\gamma_1}p_X}}\right)$, $\|\hat{h}_\lambda - h\|^2_{L_2(X)} = O_P\left(n^{-\frac{\beta_X}{\beta_X-1+\gamma_0+\frac{\gamma_0}{\gamma_1}p_X}}\right)$.
- $\frac{\beta_Z+p_Z}{\alpha_Z} \leqslant a \leqslant \frac{\beta_Z+p_Z}{\beta_Z-\alpha_Z}\frac{\gamma_0\left(1-\frac{p_X}{\gamma_1}\right)+c_F}{\beta_X-1+\gamma_0+\frac{\gamma_0}{\gamma_1}p_X}$, $\lambda = \Theta\left(n^{-\frac{a(\beta_Z-\alpha_Z)+1+p_z}{2\beta_Z+p_Z}\cdot\frac{2\gamma_0}{\beta_X-1+2\gamma_0+c_F}}\right)$,
  $\|\hat{h}_\lambda - h\|^2_{L_2(X)} = O_P\left(n^{-\frac{a(\beta_Z-\alpha_Z)+1+p_z}{2\beta_Z+p_Z}\cdot\frac{\beta_X}{\beta_X-1+2\gamma_0+c_F}}\right)$.
- $a \leqslant \frac{\beta_Z+p_Z}{\beta_Z}\frac{\beta_X-1+2\gamma_0+c_F}{\beta_X-1+\gamma_0+\frac{\gamma_0}{\gamma_1}p_X}$, $\lambda = \Theta\left(n^{-a\cdot\frac{\beta_Z}{\beta_Z+p_Z}\cdot\frac{2\gamma_0}{\beta_X-1+2\gamma_0+c_F}}\right)$, $\|\hat{h}_\lambda - h\|^2_{L_2(X)} = O_P\left(n^{-a\cdot\frac{\beta_Z}{\beta_Z+p_Z}\cdot\frac{\beta_X}{\beta_X-1+2\gamma_0+c_F}}\right)$.

Corollary 1 demonstrates that by choosing optimal regularization parameters and given enough stage 1 data, kernel NPIV achieves $O_P(n^{-\frac{\beta_X}{\beta_X-1+\gamma_0+\frac{\gamma_0}{\gamma_1}p_X}})$ learning rate. This learning rate recovers the classical minimax optimal learning rate for kernel ridge regression: $O_P(n^{-\frac{\beta_X}{\beta_X+p_X}})$, when $\gamma_0 = 1$ (since $\gamma_0 \geqslant \gamma_1 \geqslant 1$). However, for general γ_0, we can see that the learning rate of kernel NPIV is slower than that of kernel regression. Our minimax lower bound below affirms that kernel NPIV achieves a strictly worse rate than kernel ridge regression. Before we state the minimax lower bound, we introduce the following additional assumption

Assumption 10. *For some constants $c_2, c_3 > 0$ and $p_X \in (0,1]$ and for all $i \in I$,*

$$c_3 i^{-1/p_X} \leqslant \mu_{X,i} \leqslant c_2 i^{-1/p_X} \tag{EVDX+}$$

The above assumption is needed so that we can have a precise description of the hypothesis space $\mathcal{H}_X$.

Theorem 4. *Let k_X be a kernel on E_X such that Assumption 1 hold and π_X be a probability distribution on E_X such that (EVDX+) holds with $0 < p_X \leqslant 1$. Then for all parameters $1 \leqslant \beta_X$ and all constants $\sigma, L, B_X \geqslant 0$, there exist constants $J_0, J, \theta > 0$ such that for all learning methods $D \to \hat{h}_D$ ($D \doteq \{(x_i, z_i, y_i)\}_{i=1}^n$), all $\tau > 0$, and all sufficiently large $n \geqslant 1$ there is a distribution over variables (X, Z, Y) inducing a model of the form of Eq. (1), used to sample D, with marginal distribution π on E_X, such that (SRCX) with respect to B_X, β_X and (MOM) with respect to σ, L are satisfied, and with P^n-probability not less than $1 - J_0\tau^{1/\theta}$,*

$$\left\|\hat{h}_D - h_*\right\|^2_{L_2(X)} \geqslant \tau^2 J n^{-\frac{\beta_X}{\beta_X+\gamma_1-1+p_X}}.$$

Remark 8 (Minimax Optimal)**.** *Theorem 4 states that under standard kernel learning assumptions, no learning method can achieve a learning rate better than $n^{-\frac{\beta_X}{\beta_X+\gamma_1-1+p_X}}$. In particular, we can see that when $\gamma_0 = \gamma_1$, we obtain mimimax optimal learning rate for kernel NPIV*

$$n^{-\frac{\beta_X}{\beta_X+\gamma_1-1+p_X}}.$$

To our knowledge, this is the first minimax optimal rate obtained for kernel two-stage regression in the context of instrumental variable regression.

Beyond kernel methods, previous works such as Chen and Reiss (2011); Hall and Horowitz (2005); Hoffmann and Reiss (2008) have achieved similar minimax optimal rates in the NPIV setting. Our approach is significantly less restrictive for the following two reasons however. First, while Chen and Reiss (2011); Hall and Horowitz (2005); Hoffmann and Reiss (2008) require the operator $\mathcal{T}$ to be injective, we are the first to establish that kernel NPIV methods achieve minimax optimal rate under both injective and non-injective $\mathcal{T}$. Moreover, previous works often impose stringent conditions on the density of π_Z and the conditional density $p(x \mid z)$. For instance, Chen and Reiss (2011) assume that π_Z is continuous, bounded away from zero, and that the eigenvalues of C_F are bounded and not vanishing. In contrast, our method imposes no assumptions on these density functions, and the eigenvalue of C_F is only controlled by in terms of (γ_0, γ_1) and C_X, which allows the vanishing case.

6 Conclusion

In conclusion, we provide a comprehensive theoretical analysis of the kernel nonparametric instrumental variables (NPIV) regression method. Our key contributions include relaxing the injectivity assumption of the operator $\mathcal{T}$ by introducing the minimum Reproducing Kernel Hilbert Space (RKHS) norm solution, which ensures that the kernel NPIV algorithm converges to the minimum norm solution even when $\mathcal{T}$ is not injective.

We also introduce a novel measure of subspace size, characterized by the parameters γ_0 and γ_1, which plays a crucial role in determining the learning rate of the kernel NPIV estimator. These parameters describe the eigenvalue decay rate of the covariance operators defined through conditional mean embedding, which in turn influence the efficiency of the learning process. Finally our theoretical results demonstrate that the proposed kernel NPIV method achieves a minimax optimal learning rate under standard assumptions.

Our study highlights the limitations and potential inefficiencies of the kernel NPIV method compared to kernel ridge regression. Our findings suggest that the inefficiency is proportional to the subspace size measure defined by the (γ_0, γ_1) pair, indicating that a data adaptive methods for Stage 1 learning such as deep neural network based algorithms (Xu et al., 2020) could potentially lead to improved learning rates.

References

Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. ***Econometrica***, 71(6):1795–1843, 2003.

Donald Andrews and James H Stock. Inference with weak instruments. ***National Bureau of Economic Research Cambridge, Mass., USA***, 2005.

Isaiah Andrews, James H Stock, and Liyang Sun. Weak instruments in instrumental variables regression: Theory and practice. ***Annual Review of Economics***, 11:727–753, 2019.

Joshua Angrist and Guido Imbens. Identification and estimation of local average treatment effects, 1995.

Jean-Pierre Aubin. ***Applied Functional Analysis***. John Wiley & Sons, Inc., 2nd edition, 2000.

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. ***Advances in neural information processing systems***, 32, 2019.

Andrew Bennett, Nathan Kallus, Xiaojie Mao, Whitney Newey, Vasilis Syrgkanis, and Masatoshi Uehara. Inference on strongly identified functionals of weakly identified functions. ***arXiv preprint arXiv:2208.08291***, 2022.

Andrew Bennett, Nathan Kallus, Xiaojie Mao, Whitney Newey, Vasilis Syrgkanis, and Masatoshi Uehara. Inference on strongly identified functionals of weakly identified functions. In ***The Thirty Sixth Annual Conference on Learning Theory***, pages 2265–2265. PMLR, 2023a.

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 Thirty Sixth Annual Conference on Learning Theory***, pages 2291–2318. PMLR, 2023b.

Andrew Bennett, Nathan Kallus, Xiaojie Mao, Whitney Newey, Vasilis Syrgkanis, and Masatoshi Uehara. Source condition double robust inference on functionals of inverse problems. ***arXiv preprint arXiv:2307.13793***, 2023c.

Alain Berlinet and Christine Thomas-Agnan. ***Reproducing kernel Hilbert spaces in probability and statistics***. Springer Science & Business Media, 2011.

Rajendra Bhatia. ***Matrix analysis***, volume 169. Springer Science & Business Media, 2013.

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.

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.

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.

Pattrawut Chansangiam. Operator monotone functions: Characterizations and integral representations. ***arXiv preprint arXiv:1305.2471***, 2013.

Qihui Chen. Robust and optimal estimation for partially linear instrumental variables models with partial identification. ***Journal of econometrics***, 221(2):368–380, 2021.

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. ***Handbook of econometrics***, 6:5549–5632, 2007.

Xiaohong Chen and Timothy Christensen. Optimal sup-norm rates, adaptivity and inference in nonparametric instrumental variables estimation. ***Cowles Foundation discussion paper***, 2015.

Xiaohong Chen and Timothy M Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. ***Quantitative Economics***, 9(1):39–84, 2018.

Xiaohong Chen and Demian Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. ***Econometrica***, 80(1):277–321, 2012.

Xiaohong Chen and Markus Reiss. On rate optimality for ill-posed inverse problems in econometrics. ***Econometric Theory***, 27(3):497–521, 2011.

Xuxing Chen, Abhishek Roy, Yifan Hu, and Krishnakumar Balasubramanian. Stochastic optimization algorithms for instrumental variable regression with streaming data. ***arXiv preprint arXiv:2405.19463***, 2024.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. ***Advances in Neural Information Processing Systems***, 29, 2016.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A general framework for consistent structured prediction with implicit loss embeddings. ***The Journal of Machine Learning Research***, 21(1):3852–3918, 2020.

Serge Darolles, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. Nonparametric instrumental regression. ***Econometrica***, 79(5):1541–1565, 2011.

Ernesto De Vito, Lorenzo Rosasco, and Andrea Caponnetto. Discretization error analysis for tikhonov regularization. ***Analysis and Applications***, 4(01):81–99, 2006.

Joe Diestel and J.J. Uhl. ***Vector Measures***. American Mathematical Society, 1977.

Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. ***Advances in Neural Information Processing Systems***, 33:12248–12262, 2020.

Heinz Werner Engl, Martin Hanke, and A. Neubauer. ***Regularization of Inverse Problems***. Kluwer, 2000.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. ***J. Mach. Learn. Res.***, 21:205–1, 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.

S. Grünewälder, G. Lever, Ll. Baldassarre, M. Pontil, and A. Gretton. Modelling transition dynamics in mdps with rkhs embeddings. In ***Proceedings of the 29th International Conference on Machine Learning***, pages 535–542, New York, NY, USA, 2012a. Omnipress.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimilano Pontil. Conditional mean embeddings as regressors—supplementary. ***arXiv preprint arXiv:1205.4656***, 2012b.

Peter Hall and Joel L Horowitz. Nonparametric methods for inference in the presence of instrumental variables. ***Annals of Statistics***, 33(6):2904–2929, 2005.

Frank Hansen and Gert K Pedersen. Jensen's operator inequality. ***Bulletin of the London Mathematical Society***, 35(4):553–564, 2003.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In ***International Conference on Machine Learning***, pages 1414–1423. PMLR, 2017.

Erhard Heinz. Beiträge zur störungstheorie der spektralzerleung. ***Mathematische Annalen***, 123(1):415–438, 1951.

Marc Hoffmann and Markus Reiss. Nonlinear estimation for linear inverse problems with error in the operator. ***Annals of Statistics***, 36(1):310–336, 2008.

Joel L Horowitz. Asymptotic normality of a nonparametric instrumental variables estimator. ***International Economic Review***, 48(4):1329–1349, 2007.

Joel L Horowitz. Applied nonparametric instrumental variables estimation. ***Econometrica***, 79(2):347–394, 2011.

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.

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.

Ilja Klebanov, Björn Sprungk, and Timothy John Sullivan. The linear conditional expectation in Hilbert space. ***Bernoulli***, 27(4):2267–2299, 2021.

Vladimir Kostic, Pietro Novelli, Andreas Maurer, Carlo Ciliberto, Lorenzo Rosasco, and Massimiliano Pontil. Learning dynamical systems via Koopman operator regression in reproducing kernel Hilbert spaces. ***Advances in Neural Information Processing Systems***, 35:4017–4031, 2022.

Vladimir Kostic, Karim Lounici, Pietro Novelli, and Massimiliano Pontil. Koopman operator learning: Sharp spectral rates and spurious eigenvalues. ***arXiv preprint arXiv:2302.02004***, 2023.

Greg Lewis and Vasilis Syrgkanis. Adversarial generalized method of moments. ***arXiv preprint arXiv:1803.07164***, 2018.

Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Optimal rates for regularized conditional mean embedding learning. In ***Advances in Neural Information Processing Systems***, volume 35, pages 4433–4445, 2022a.

Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Optimal rates for regularized conditional mean embedding learning. In ***Advances in Neural Information Processing Systems***, 2022b.

Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Towards optimal Sobolev norm rates for the vector-valued regularized least-squares algorithm. ***arXiv preprint arXiv:2312.07186***, 2023.

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 deepiv 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. ***Advances in Neural Information Processing Systems***, 33:8947–8958, 2020.

Luofeng Liao, Zuyue Fu, Zhuoran Yang, Yixin Wang, Mladen Kolar, and Zhaoran Wang. Instrumental variable value iteration for causal offline reinforcement learning. ***arXiv preprint arXiv:2102.09907***, 2021.

Junhong Lin and Volkan Cevher. Optimal distributed learning with multi-pass stochastic gradient methods. In ***International Conference on Machine Learning***, pages 3092–3101. PMLR, 2018.

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.

Sébastien Loustau. Inverse statistical learning. ***Electronic Journal of Statistics***, 7:2065–2097, 2013.

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 ***International Conference on Machine Mearning***, pages 7512–7523. PMLR, 2021.

Dimitri Meunier, Zikai Shen, Mattes Mollenhauer, Arthur Gretton, and Zhu Li. Optimal rates for vector-valued spectral regularization learning algorithms. ***arXiv preprint arXiv:2405.14778***, 2024.

Wang Miao, Lan Liu, Eric Tchetgen Tchetgen, and Zhi Geng. Identification, doubly robust estimation, and semiparametric efficiency theory of nonignorable missing data with a shadow variable. ***arXiv preprint arXiv:1509.02556***, 2015.

Mattes Mollenhauer and Péter Koltai. Nonparametric approximation of conditional expectation operators. ***arXiv preprint arXiv:2012.12917***, 2020.

Mattes Mollenhauer, Nicole Mücke, and TJ Sullivan. Learning linear operators: Infinite-dimensional regression as a well-behaved non-compact inverse problem. ***arXiv preprint arXiv:2211.08875***, 2022.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. ***Foundations and Trends® in Machine Learning***, 10(1-2): 1–141, 2017.

Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. Dual instrumental variable regression. ***Advances in Neural Information Processing Systems***, 33:2710–2721, 2020.

Nicole Mücke, Gergely Neu, and Lorenzo Rosasco. Beating sgd saturation with tail-averaging and minibatching. ***Advances in Neural Information Processing Systems***, 32, 2019.

M Thamban Nair, Sergei V Pereverzev, and Ulrich Tautenhahn. Regularization in hilbert scales under general smoothing conditions. ***Inverse Problems***, 21(6):1851, 2005.

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. ***Advances in Neural Information Processing Systems***, 33:21247–21259, 2020.

Ieva Petrulionyte, Julien Mairal, and Michael Arbel. Functional bilevel optimization for machine learning. ***arXiv preprint arXiv:2403.20233***, 2024.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In ***Advances in Neural Information Processing Systems***, pages 1657–1665, 2015.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. ***The Annals of Statistics***, pages 2263–2291, 2013.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. ***Advances in Neural Information Processing Systems***, 32, 2019.

Steve Smale and Ding-Xuan Zhou. Shannon sampling and function reconstruction from point values. ***Bulletin of the American Mathematical Society***, 41(3):279–305, 2004.

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.

Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. ***Journal of Machine Learning Research***, 12(Jul):2389–2410, 2011.

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(3):363–417, 2012.

Ingo Steinwart, Don R Hush, Clint Scovel, et al. Optimal rates for regularized least squares regression. In ***COLT***, pages 79–93, 2009.

Alexandre B Tsybakov. ***Introduction to Nonparametric Estimation***. Springer, 2009.

Masatoshi Uehara, Masaaki Imaizumi, Nan Jiang, Nathan Kallus, Wen Sun, and Tengyang Xie. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. ***arXiv preprint arXiv:2102.02981***, 2021.

Sheng Wang, Jun Shao, and Jae Kwang Kim. An instrumental variable approach for identification and estimation with nonignorable nonresponse. ***Statistica Sinica***, pages 1097–1116, 2014.

Joachim Weidmann. ***Linear Operators in Hilbert Spaces***. Springer, 1980.

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***, 2020.

Appendices

In Section A, we provide additional background material. In Section B, we provide the proof of Proposition 1. In Section C, we provide the proof of Proposition 2. In Section D, we provide a closed-form expression for the Kernel NPIV estimator $\hat{h}_\lambda$. In Section E, we provide a proof sketch (Section E.1) followed by the detailed proof (Section E.2) for the upper bound presented in Theorem 3. In Section F, we prove the lower bound given in Theorem 4. In Section G, we provide additional bounds used in the main proofs. Finally, in Section H, we collect some technical supporting results.

A Additional Background

A.1 Hilbert spaces and linear operators

Definition 3 (Bochner L_q–spaces, e.g. Diestel and Uhl (1977))**.** *Let H be a separable Hilbert space. For $1 \leqslant q \leqslant \infty$, $L_q(E_Z, \mathcal{F}_{E_Z}, \pi_Z; H)$, abbreviated $L_q(Z;H)$, is the space of strongly $\mathcal{F}_{E_Z} - \mathcal{F}_H$ measurable and Bochner q-integrable functions from E_Z to H, with the norms*

$$\|F\|^q_{L_q(Z;H)} = \int_{E_Z} \|F\|^q_H \, \mathrm{d}\pi_Z, \quad 1 \leqslant q < \infty, \qquad \|F\|_{L_\infty(Z;H)} = \inf\left\{C \geqslant 0 : \pi_Z\{\|F\|_H > C\} = 0\right\}.$$

Definition 4 (q-Schatten class, e.g. Weidmann (1980))**.** *Let H, H' be separable Hilbert spaces. For $1 \leqslant q \leqslant \infty$, $S_q(H,H')$, abbreviated $S_q(H)$ if $H = H'$, is the Banach space of all compact operators Q from H to H' such that $\|Q\|_{S_q(H,H')} \doteq \|(\sigma_i(Q))_{i\in I}\|_{\ell_q}$ is finite. Here $\|(\sigma_i(Q))_{i\in I}\|_{\ell_q}$ is the ℓ_q–sequence space norm of the sequence of the strictly positive singular values of Q indexed by the at most countable set I. For $q = 2$, we retrieve the space of Hilbert-Schmidt operators, for $q = 1$ we retrieve the space of Trace Class operators, and for $q = +\infty$, $\|\cdot\|_{S_\infty(H,H')}$ corresponds to the operator norm $\|\cdot\|_{H\to H'}$.*

Definition 5 (Tensor Product of Hilbert Spaces, Aubin (2000))**.** *Let H, H' be Hilbert spaces. The Hilbert space $H \otimes H'$ is the completion of the algebraic tensor product with respect to the norm induced by the inner product $\langle x_1 \otimes x_1', x_2 \otimes x_2'\rangle_{H\otimes H'} = \langle x_1, x_2\rangle_H \langle x_1', x_2'\rangle_{H'}$ for $x_1, x_2 \in H$ and $x_1', x_2' \in H'$ defined on the elementary tensors of $H \otimes H'$. This definition extends to $\operatorname{span}\{x \otimes x' | x \in H, x' \in H'\}$ and finally to its completion. The space $H \otimes H'$ is separable whenever both H and H' are separable. If $\{e_i\}_{i\in I}$ and $\{e_j'\}_{j\in J}$ are orthonormal basis in H and H' respectively, $\{e_i \otimes e_j'\}_{i\in I, j\in J}$ is an orthonormal basis in $H \otimes H'$.*

Theorem 5 (Isometric Isomorphism between $L_2(Z;H)$ and $S_2(L_2(Z),H)$, Theorem 12.6.1 Aubin (2000))**.** *Let H be a separable Hilbert space. The Bochner space $L_2(Z;H)$ is isometrically isomorphic to $S_2(L_2(Z),H)$ and the isometric isomorphism is realized by the map $\Psi : S_2(L_2(Z),H) \to L_2(Z;H)$ acting on elementary tensors as $\Psi(f \otimes h) = (\omega \to f(\omega)h)$.*

A.2 RKHS embedding into L_2

Recall that $\mathcal{I}_Z : \mathcal{H}_Z \to L_2(Z)$ is the embedding that maps every function in $\mathcal{H}_Z$ into its π_Z-equivalence class in $L_2(Z)$ and that we used the shorthand notation $[f]_Z = \mathcal{I}_Z(f)$ for all $f \in \mathcal{H}_Z$. We define similarly $\mathcal{I}_{Z;X} : \mathcal{G} \to L_2(Z;\mathcal{H}_X)$ as the embedding that maps every function in $\mathcal{G}$ into their π_Z-equivalence class in $L_2(Z;\mathcal{H}_X)$.

Definition 6 (Embedding $\mathcal{G}$ into $L_2(Z;\mathcal{H}_X)$)**.** *Let $\mathcal{I}_{Z;X} \doteq \mathrm{Id}_{\mathcal{H}_X} \otimes \mathcal{I}_Z$ be the tensor product of the operator $\mathrm{Id}_{\mathcal{H}_X}$ with the operator $\mathcal{I}_Z$ (see Aubin (2000, Definition 12.4.1.) for the definition of tensor product of operators). $\mathcal{I}_{Z;X}$ maps every function in $\mathcal{G}$ into their π_Z-equivalence class in $L_2(Z;\mathcal{H}_X)$. We then use the shorthand notation $[F]_{Z;X} = \mathcal{I}_{Z;X}(F)$ for all $F \in \mathcal{G}$.*

A.3 Spectral regularization

1. *Ridge regression.* From the Tikhonov filter function $g_\xi(x) = (x+\xi)^{-1}$, we obtain the ridge regression algorithm introduced in Eq. (5). In this case, we have $E = \rho = \omega_\rho = 1$.

2. *Gradient Descent.* From the Landweber iteration filter function given by

$$g_k(x) \doteq \tau \sum_{i=0}^{k-1} (1-\tau x)^i \text{ for } k \doteq 1/\xi, k \in \mathbb{N}$$

we obtain the gradient descent scheme with constant step size $\tau > 0$, which corresponds to the population gradient iteration given by $F_{k+1} \doteq F_k - \frac{\tau}{2} \nabla_F \left(\mathbb{E}_{X,Z} \left\| \phi_X(X) - F(Z) \right\|_{\mathcal{H}_X}^2 \right) (F_k)$ for $k \in \mathbb{N}$. In this case, we have $E = 1$ and arbitrary qualification with $\omega_\rho = 1$ whenever $0 < \rho \leqslant 1$ and $\omega_\rho = \rho^\rho$ otherwise. Gradient schemes with more complex update rules can be expressed in terms of filter functions as well (Mücke et al., 2019; Lin and Cevher, 2018; Lin et al., 2020).

3. *Kernel principal component regression.* The truncation filter function $g_\xi(x) = x^{-1} \mathbb{1}[x \geqslant \xi]$ yields kernel principal component regression, corresponding to a hard thresholding of eigenvalues at a truncation level ξ. In this case we have $E = \omega_\rho = 1$ for arbitrary qualification ρ.

4. *Iterated Tikhonov.* Mixture between Landweber iteration and Tikhonov regularization. Unlike Tikhonov regularization which has finite qualification and cannot exploit the regularity of the solution beyond a certain regularity level, iterated Tikhonov overcomes this problem by means of the following regularization: $g_{\xi,\nu}(x) = \frac{(x+\xi)^\nu - \xi^\nu}{x(x+\xi)^\nu}$ with $\nu > 0$. In this case we have $E = \omega_\rho = 1$ and $\rho = \nu$. For $\nu = 1$, we retrieve the standard Tikhonov regularization and for $\nu \in \mathbb{N}$ we can show that applying $g_{\xi,\nu}$ corresponds to the following iterative procedure:

$$\begin{aligned} g_{\xi,1} &= (x+\xi)^{-1} \\ g_{\xi,k} &= (1 + \xi g_{\xi,k-1}) g_{\xi,1}. \end{aligned}$$

5. *Gradient Flow.* If we fix the total distance in the Landweber iteration to $\xi^{-1} := \tau k$ and take $\tau \to 0_+$, we obtain the gradient flow filter function $g_\xi(x) = (1 - e^{-\frac{x}{\xi}}) x^{-1}$. In this case we have $E = 1$ and $\omega_\rho = (\tau/e)^\tau$ for arbitrary qualification ρ.

A.4 Interpolation spaces

The interpolation spaces $[\mathcal{H}_Z]^\beta$, $[\mathcal{H}_X]^\beta$ and $[\mathcal{G}]^\beta$ introduced previously correspond to the Hilbert scale generated by the operator L_Z, L_X and $\mathrm{Id}_{\mathcal{H}_X} \otimes L_Z$ respectively (see e.g. Steinwart and Scovel (2012) and Fischer and Steinwart (2020)). We know give more details on their construction.

For $\beta \geqslant 0$, we define the β-interpolation space (Steinwart and Scovel, 2012) by

$$[\mathcal{H}_Z]^\beta \doteq \left\{ \sum_{i \in I} a_i \mu_{Z,i}^{\beta/2} [e_{Z,i}]_Z : (a_i)_{i \in I} \in \ell_2 \right\} \subseteq L_2(Z),$$

equipped with the inner product

$$\left\langle \sum_{i \in I} a_i (\mu_{Z,i}^{\beta/2} [e_{Z,i}]_Z), \sum_{i \in I} b_i (\mu_{Z,i}^{\beta/2} [e_{Z,i}]_Z) \right\rangle_\beta = \sum_{i \in I} a_i b_i.$$

The β-interpolation space is a separable Hilbert space with ONB $\left(\mu_{Z,i}^{\beta/2} [e_{Z,i}]_Z \right)_{i \in I}$. For $\beta = 0$, we have $[\mathcal{H}_Z]^0 = \overline{\mathcal{R}(\mathcal{I}_Z)} \subseteq L_2(Z)$ with $\| \cdot \|_0 = \| \cdot \|_{L_2(Z)}$. For $\beta = 1$, we have $[\mathcal{H}_Z]^1 = \mathcal{R}(\mathcal{I}_Z)$ and $[\mathcal{H}_Z]^1$ is isometrically isomorphic to the closed subspace $(\mathcal{N}(\mathcal{I}_Z))^\perp$ of $\mathcal{H}_Z$ via $\mathcal{I}_Z$, i.e. $\|[f]_Z\|_1 = \|f\|_{\mathcal{H}_Z}$ for $f \in (\mathcal{N}(\mathcal{I}_Z))^\perp$. For $0 < \beta < \alpha$, we have

$$[\mathcal{H}_Z]^\alpha \hookrightarrow [\mathcal{H}_Z]^\beta \hookrightarrow [\mathcal{H}_Z]^0 \subseteq L_2(Z).$$

For $\beta > 0$ and $f \in \overline{\mathcal{R}(\mathcal{I}_Z)}$, the β-interpolation space is given by the image of the fractional integral operator,

$$[\mathcal{H}_Z]^\beta = \mathcal{R}(L_Z^{\beta/2}) \quad \text{and} \quad \|f\|_\beta = \|L_Z^{-\beta/2} f\|_{L_2(Z)}.$$

For a vector-valued function $F \in L_2(Z; \mathcal{H}_X)$ since $L_2(Z; \mathcal{H}_X)$ is isometric to $S_2(L_2(Z), \mathcal{H}_X)$, there is an operator $C \in S_2(L_2(Z), \mathcal{H}_X)$ such that $\|F\|_{L_2(Z;\mathcal{H}_X)} = \|C\|_{S_2(L_2(Z),\mathcal{H}_X)}$. For $C \in S_2(\overline{\mathcal{R}(\mathcal{I}_Z)}, \mathcal{H}_X)$, we define the vector-valued β−interpolation norm as

$$\|F\|_\beta \doteq \|C\|_\beta \doteq \|C L_Z^{-\beta/2}\|_{S_2(L_2(Z),\mathcal{H}_X)}.$$

For details regarding vector-valued interpolation spaces, we refer to Li et al. (2022a, 2023).

Remark 9. *The interpolation space $[\mathcal{H}_X]^\beta$ is defined similarly to $[\mathcal{H}_Z]^\beta$.*

B Proof of Proposition 1

By Assumption 2, $\mathcal{N}_{r_0}(\tilde{\mathcal{T}}) \subseteq \mathcal{H}_X$ is not empty. Fix h an element of $\mathcal{N}_{r_0}(\tilde{\mathcal{T}})$. Since $\mathcal{H}_X = \mathcal{N}(\tilde{\mathcal{T}}) \oplus \mathcal{N}(\tilde{\mathcal{T}})^\perp$ there exists a unique pair $(h', h'') \in \mathcal{N}(\tilde{\mathcal{T}})^\perp \times \mathcal{N}(\tilde{\mathcal{T}})$ such that $h = h' + h''$. Since $h \in \mathcal{N}_{r_0}(\tilde{\mathcal{T}})$ and $h'' \in \mathcal{N}(\tilde{\mathcal{T}})$, we have:

$$r_0 = \tilde{\mathcal{T}} h = \tilde{\mathcal{T}} h' + \tilde{\mathcal{T}} h'' = \tilde{\mathcal{T}} h'.$$

Therefore $h' \in \mathcal{N}_{r_0}(\tilde{\mathcal{T}})$. Furthermore, $\|h\|^2_{\mathcal{H}_X} = \|h'\|^2_{\mathcal{H}_X} + \|h''\|^2_{\mathcal{H}_X} \geqslant \|h'\|_{\mathcal{H}_X}$. This proves that the minimum norm solution in $\mathcal{H}_X$ exists and is uniquely defined as h' and belongs to $\mathcal{N}(\tilde{\mathcal{T}})^\perp \cap \mathcal{N}_{r_0}(\tilde{\mathcal{T}})$. To conclude, we show that $\mathcal{N}(\tilde{\mathcal{T}})^\perp \cap \mathcal{N}_{r_0}(\tilde{\mathcal{T}})$ only contains one element. Assume that there exists $h, \tilde{h} \in \mathcal{N}(\tilde{\mathcal{T}})^\perp \cap \mathcal{N}_{r_0}(\tilde{\mathcal{T}})$, then $\tilde{\mathcal{T}}(h - \tilde{h}) = r_0 - r_0 = 0$, therefore $h - \tilde{h} \in \mathcal{N}(\tilde{\mathcal{T}})$. But since we also have $h - \tilde{h} \in \mathcal{N}(\tilde{\mathcal{T}})^\perp$, it implies $h = \tilde{h}$.

C Proof of Proposition 2

Proposition 3. *Let P_F be the orthogonal projection on $\overline{\mathcal{R}(C_F)}$. Under Assumption (LINK) with $\gamma_0, \gamma_1 \in [1, +\infty)$, we have the following properties*

a). $1 \leqslant \gamma_1 \leqslant \gamma_0$.

b). γ_0, γ_1 can be equivalently characterized respectively as $P_F C_X^{\gamma_0} P_F \preceq C_F$ and $C_F \preceq C_X^{\gamma_1}$.

c). γ_0, γ_1 can be equivalently characterized respectively as $\mathcal{R}(C_F^{1/2}) \subseteq \mathcal{R}(C_X^{\gamma_1/2})$ and $\mathcal{R}\left(\left(P_F C_X^{\gamma_0} P_F\right)^{1/2}\right) \subseteq \mathcal{R}(C_F^{1/2})$.

d). For any $\tau \in [0, 1]$, we have $P_F C_X^{\tau\gamma_0} P_F \preceq C_F^\tau \preceq C_X^{\tau\gamma_1}$.

Proof. For Part a), by (LINK), for all $f \in \overline{\mathcal{R}(C_F)}$, $\langle f, C_X^{\gamma_0} f\rangle_{\mathcal{H}_X} \leqslant \langle f, C_X^{\gamma_1} f\rangle_{\mathcal{H}_X}$. Let us reason by contradiction and assume that we have $\gamma_1 > \gamma_0$, we can then find $f \in \overline{\mathcal{R}(C_F)}$ such that $\langle f, C_X^{\gamma_1} f\rangle_{\mathcal{H}_X} < \langle f, C_X^{\gamma_0} f\rangle_{\mathcal{H}_X}$ which yields a contradiction. We hence have $\gamma_1 \leqslant \gamma_0$.

Part b) is obtained directly by definition of γ_0 and P_F.

To prove c), we apply Proposition 6.

Part d). Note that γ_1 is characterized by $C_F \preceq C_X^{\gamma_1}$. Therefore, the second inequality is obtained using the fact that the map $x \mapsto x^\tau$ is operator monotone for $\tau \in [0,1]$ (Löwner-Heniz theorem Heinz (1951)).

For the first inequality, since by part b), γ_0 is characterized by $P_F C_X^{\gamma_0} P_F \preceq C_F$, using again the Löwner-Heniz theorem (Heinz, 1951), we obtain that $\forall f \in \mathcal{H}_X$ and any $\tau \in (0,1)$,

$$\langle f, \left(P_F C_X^{\gamma_0} P_F\right)^\tau f\rangle_{\mathcal{H}_X} \leqslant \langle f, C_F^\tau f\rangle_{\mathcal{H}_X}.$$

Choosing $f = P_F g$ for any $g \in \mathcal{H}_X$, we then have

$$\langle P_F g, \left(P_F C_X^{\gamma_0} P_F\right)^\tau P_F g\rangle_{\mathcal{H}_X} \leqslant \langle P_F g, C_F^\tau P_F g\rangle_{\mathcal{H}_X}. \tag{15}$$

This proves that $P_F \left(P_F C_X^{\gamma_0} P_F\right)^\tau P_F \preceq P_F C_F^\tau P_F$. Furthermore, by Jensen's Operator Inequality (Theorem 2.1 Hansen and Pedersen (2003)), for any operator concave function h, every self-adjoint operator A with spectrum in I and every λ in I, we have

$$Ph(PAP + \lambda(1-P))P \geqslant Ph(A)P.$$

We apply this result to $\lambda = 0$, $A = C_X^{\gamma_0}$ and $h(x) = x^\tau$, which is operator concave for $\tau \in [0,1]$ (Example 3.6 i Chansangiam (2013), Bhatia (2013)). Applying this to Eq. (15), we obtain

$$P_F \left(C_X^{\gamma_0}\right)^\tau P_F \preceq P_F \left(P_F C_X^{\gamma_0} P_F\right)^\tau P_F \preceq P_F C_F^\tau P_F = C_F^\tau.$$

□

D Explicit Solutions of Kernel NPIV

The closed-form solution for kernel NPIV is already studied in (Algorithm 1 Singh et al., 2019). However, they employ the Tikhonov regularization for both stages. We here provide the closed-form solution where we allow general regularization scheme for stage 1. Stage 2, on the other hand, remains with Tikhonov regularization.

Stage 1. Recall that in Stage 1, we obtain the following estimator with $\mathcal{D}_1$ and $\xi > 0$ (see Eq. (7) and the definition of $\mathcal{G}$): $\hat{F}_\xi(\cdot) = \hat{C}_{X|Z,\xi}\phi_Z(\cdot)$, with

$$\hat{C}_{X|Z,\xi} = \frac{1}{m}\mathbf{\Phi}_{\tilde{X}}^T \mathbf{\Phi}_{\tilde{Z}} g_\xi\left(\frac{1}{m}\mathbf{\Phi}_{\tilde{Z}}^T \mathbf{\Phi}_{\tilde{Z}}\right), \tag{16}$$

with

$$\begin{aligned} \mathbf{\Phi}_{\tilde{Z}} : \mathcal{H}_Z \to \mathbb{R}^m \qquad & \mathbf{\Phi}_{\tilde{Z}} = [\phi_Z(\tilde{z}_1), \dots, \phi_Z(\tilde{z}_m)]^* \\ \mathbf{\Phi}_{\tilde{X}} : \mathcal{H}_X \to \mathbb{R}^m \qquad & \mathbf{\Phi}_{\tilde{X}} = [\phi_X(\tilde{x}_1), \dots, \phi_X(\tilde{x}_m)]^* \end{aligned}$$

The solution can also be written in the following dual form (see Meunier et al. (2024)):

$$\hat{C}_{X|Z,\xi} = \frac{1}{m}\mathbf{\Phi}_{\tilde{X}}^T g_\xi\left(\frac{\mathbf{K}_{\tilde{Z}\tilde{Z}}}{m}\right)\mathbf{\Phi}_{\tilde{Z}},$$

where we introduce the Gram matrix

$$\mathbf{K}_{\tilde{Z}\tilde{Z}} = \mathbf{\Phi}_{\tilde{Z}}\mathbf{\Phi}_{\tilde{Z}}^T, \qquad [\mathbf{K}_{\tilde{Z}\tilde{Z}}]_{ij} = \langle \phi_Z(\tilde{z}_i), \phi_Z(\tilde{z}_j)\rangle_{\mathcal{H}_Z} \qquad i,j \in [m].$$

Stage 2. Recall from Eq. (21) that

$$\hat{h}_\lambda = \left(\frac{1}{n}\mathbf{\Phi}_{\hat{F}}^T\mathbf{\Phi}_{\hat{F}} + \lambda\,\mathrm{Id}\right)^{-1}\frac{1}{n}\mathbf{\Phi}_{\hat{F}}^T Y = \left(\hat{C}_{\hat{F}} + \lambda\,\mathrm{Id}\right)^{-1}\frac{1}{n}\mathbf{\Phi}_{\hat{F}}^T Y,$$

which we can write in dual form as follows:

$$\hat{h}_\lambda = \mathbf{\Phi}_{\hat{F}}^T\left[\mathbf{F} + n\lambda\,\mathrm{Id}\right]^{-1} Y, \qquad Y = [y_1, \dots, y_n]^T \in \mathbb{R}^n$$

$$\mathbf{F}_{ij} = [\mathbf{\Phi}_{\hat{F}}\mathbf{\Phi}_{\hat{F}}^T]_{ij} = \langle \hat{F}_\xi(z_i), \hat{F}_\xi(z_j)\rangle_{\mathcal{H}_X} \qquad i,j \in [n]$$

By Eq. (16) and using $\hat{F}_\xi(\cdot) = \hat{C}_{X|Z,\xi}\phi_Z(\cdot)$, we obtain, $\mathbf{F}$ and $\mathbf{\Phi}_{\hat{F}}^T$ in closed form:

$$\begin{aligned}
\mathbf{F} &= \mathbf{K}_{\tilde{Z}\tilde{Z}}\left[\mathbf{K}_{\tilde{Z}\tilde{Z}} + m\xi\,\mathrm{Id}\right]^{-1}\mathbf{K}_{\tilde{X}\tilde{X}}\left[\mathbf{K}_{\tilde{Z}\tilde{Z}} + m\xi\,\mathrm{Id}\right]^{-1}\mathbf{K}_{\tilde{Z}\tilde{Z}}\\
\mathbf{\Phi}_{\hat{F}}^T &= \mathbf{K}_{\tilde{Z}\tilde{Z}}\left[\mathbf{K}_{\tilde{Z}\tilde{Z}} + m\xi\,\mathrm{Id}\right]^{-1}\mathbf{\Phi}_{\tilde{X}}\\
\mathbf{K}_{\tilde{X}\tilde{X}} &= \mathbf{\Phi}_{\tilde{X}}\mathbf{\Phi}_{\tilde{X}}^T, \qquad [\mathbf{K}_{\tilde{X}\tilde{X}}]_{ij} = \langle \phi_{\tilde{X}}(\tilde{x}_i), \phi_{\tilde{X}}(\tilde{x}_j)\rangle_{\mathcal{H}_X} \qquad i,j \in [n]
\end{aligned}$$

Therefore, introducing $\mathbf{J} \doteq \mathbf{K}_{\tilde{Z}\tilde{Z}}\left[\mathbf{K}_{\tilde{Z}\tilde{Z}} + m\xi\,\mathrm{Id}\right]^{-1}$, for a new test point $x \in E_X$, we have,

$$\hat{h}_\lambda(x) = \langle \hat{h}_\lambda, \phi_X(x)\rangle_{\mathcal{H}_X} = Y^T\left[\mathbf{J}\mathbf{K}_{\tilde{X}\tilde{X}}\mathbf{J} + n\lambda\,\mathrm{Id}\right]^{-1}\mathbf{\Phi}_{\hat{F}}\phi_X(x),$$

with

$$\mathbf{\Phi}_{\hat{F}}\phi_X(x) = (\langle \hat{F}_\xi(z_i), \phi_X(x)\rangle_{\mathcal{H}_X})_{i=1}^n = \mathbf{J}\mathbf{\Phi}_{\tilde{X}}\phi_X(x), \qquad \mathbf{\Phi}_{\tilde{X}}\phi_X(x) = (k_X(\tilde{x}_i, x))_{i=1}^n.$$

E Proof of Theorem 3

In this section we prove Theorem 3, which we give in full detail in Theorem 6 below. We prove a more general version by bounding the error $\hat{h}_\lambda - h_*$ in γ−norm ($\gamma \in [0,1]$), see Section 5.1. For $\gamma = 0$, we retrieve the L_2−norm and for $\gamma = 1$ we retrieve the $\mathcal{H}_X$−norm.

Theorem 6. *For $\tau \geqslant 1$ and $\lambda > 0$, we define*

$$\begin{aligned}
\mathcal{N}_F(\lambda) &\doteq \mathrm{Tr}\left(C_F\left(C_F + \lambda\,\mathrm{Id}_{\mathcal{H}_X}\right)^{-1}\right)\\
g_\lambda &\doteq \log\left(2e\mathcal{N}_F(\lambda)\frac{\|C_F\|_{\mathcal{H}_X\to\mathcal{H}_X} + \lambda}{\|C_F\|_{\mathcal{H}_X\to\mathcal{H}_X}}\right)\\
A_{\lambda,\tau} &\doteq 8\tau g_\lambda \kappa_X^2\lambda^{-1},
\end{aligned} \tag{17}$$

Let Assumptions 1, 2, (EVDX), (SRCX), (MOM) and (LINK) hold with $p_X \in (0,1]$ and $1 \leqslant \beta_X \leqslant \gamma_0 + 1$ and let Assumptions (SRCZ) and (EMBZ) hold with $\alpha_Z \leqslant \beta_Z$. For sufficiently large m and n such that

$$\begin{aligned}
&n \geqslant 8\tau g_\lambda \kappa_X^2\lambda^{-1}\\
&\frac{J\sqrt{\tau}\|F_* - \hat{F}_\xi\|_{\alpha_Z}}{\lambda\sqrt{n}} \leqslant \frac{1}{12},\\
&\frac{J\|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)}}{\lambda} \leqslant \frac{1}{12},\\
&\|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)} \leqslant 1 \qquad \|F_* - \hat{F}_\xi\|_{\alpha_Z} \leqslant 1,
\end{aligned} \tag{18}$$

where J depends on $A_Z, B_Z, \kappa_Z, \alpha_Z, \beta_Z$ is given in , with P^n-probability over $1 - 12e^{-\tau}$

$$
\begin{aligned}
\|\hat{h}_\lambda - h_*\|_\gamma \leqslant J_0 \tau \lambda^{\omega_\gamma - 1} &\left(\left\|F_* - \hat{F}_\xi\right\|_{L_2(Z;\mathcal{H}_X)} + \frac{\left\|\hat{F}_\xi - F_*\right\|_{\alpha_Z}}{\sqrt{n}} \right) \left(\left\|\bar{h}_\lambda\right\|_{\mathcal{H}_X} + 1 \right) \\
&+ J_1 \left(\lambda^{\frac{\beta_X - \gamma}{2\gamma_0}} + \tau \sqrt{1 + \frac{1}{n\lambda^{1 - \frac{p_X}{\gamma_1}}}} \sqrt{\frac{1}{n\lambda^{1 - \frac{1-\gamma}{\gamma_0} + \frac{p_X}{\gamma_1}}}} \right)
\end{aligned}
$$

where J_0, J_1 only depend on $\sigma, L, A_Z, B_Z, \kappa_Z, \alpha_Z, \beta_Z, p_X, \kappa_X, B_X$ and $\omega_\gamma \doteq \frac{1-\gamma}{2\gamma_0} 1_{\mathcal{N}(C_F)} = \{0\}$.

E.1 Analysis Outline

Fix $\gamma \in [0,1]$. The starting point is the following decomposition:

$$
\begin{aligned}
\|\hat{h}_\lambda - h_*\|_\gamma &= \|\hat{h}_\lambda - \bar{h}_\lambda + \bar{h}_\lambda - h_*\|_\gamma \\
&\leqslant \underbrace{\|\hat{h}_\lambda - \bar{h}_\lambda\|_\gamma}_{\text{Stage 1 Error}} + \underbrace{\|\bar{h}_\lambda - h_*\|_\gamma}_{\text{Stage 2 Error}} .
\end{aligned}
$$

The stage 1 error measures the loss of performance in using features $\hat{F}_\xi$ instead of F_*. This quantity will be bounded by a function of m (the number of samples for stage 1), via the difference $\hat{F}_\xi - F_*$, and n (the number of samples for stage 2). On the other hand, stage 2 error only depends on n and measures how well we approximate h_* by regressing Y on $F_*(Z)$.

E.1.1 Stage 1 Error

We start with the observation that we always have

$$
\|\hat{h}_\lambda - \bar{h}_\lambda\|_\gamma \leqslant \|C_X^{\frac{1-\gamma}{2}} (\hat{h}_\lambda - \bar{h}_\lambda)\|_{\mathcal{H}_X} \leqslant \kappa_X^{1-\gamma} \|\hat{h}_\lambda - \bar{h}_\lambda\|_{\mathcal{H}_X} \leqslant \lambda^{-1/2} \kappa_X^{1-\gamma} \|(C_F + \lambda\,\mathrm{Id})^{1/2} \left(\hat{h}_\lambda - \bar{h}_\lambda\right) \|_{\mathcal{H}_X}, \tag{19}
$$

where we used $\|C_X\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant \kappa_X^2$ and Fischer and Steinwart (2020, Lemma 12). Alternatively, we would like to use the Assumption (LINK), however, we generally cannot ensure that $\hat{h}_\lambda \in \overline{\mathcal{R}(C_F)}$ except if $\overline{\mathcal{R}(C_F)} = \mathcal{H}_X$, i.e. $\mathcal{N}(C_F) = \{0\}$. In that case, we are guaranteed that $\hat{h}_\lambda \in \overline{\mathcal{R}(C_F)}$ and by Assumption (LINK) combined with Proposition 3 d), we have,

$$
\|\hat{h}_\lambda - \bar{h}_\lambda\|_\gamma = \|C_X^{\frac{1-\gamma}{2}} (\hat{h}_\lambda - \bar{h}_\lambda)\|_{\mathcal{H}_X} \leqslant \|C_F^{\frac{1-\gamma}{2\gamma_0}} (\hat{h}_\lambda - \bar{h}_\lambda)\|_{\mathcal{H}_X} \leqslant \lambda^{\frac{1-\gamma}{2\gamma_0} - \frac{1}{2}} \|(C_F + \lambda\,\mathrm{Id})^{1/2} \left(\hat{h}_\lambda - \bar{h}_\lambda\right) \|_{\mathcal{H}_X}, \tag{20}
$$

where we used Lemma 4 to obtain $\|C_F^{\frac{1-\gamma}{2\gamma_0}} (C_F + \lambda\,\mathrm{Id})^{-1/2}\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant \lambda^{\frac{1-\gamma}{2\gamma_0} - \frac{1}{2}}$. To go further, we use that $\hat{h}_\lambda, \bar{h}_\lambda$, admit the following closed-form expressions:

$$
\begin{aligned}
\hat{h}_\lambda &= \left(\frac{1}{n} \mathbf{\Phi}_{\hat{F}}^* \mathbf{\Phi}_{\hat{F}} + \lambda\,\mathrm{Id} \right)^{-1} \frac{1}{n} \mathbf{\Phi}_{\hat{F}}^* Y = \left(\hat{C}_{\hat{F}} + \lambda\,\mathrm{Id} \right)^{-1} \frac{1}{n} \mathbf{\Phi}_{\hat{F}}^* Y \qquad (21) \\
\bar{h}_\lambda &= \left(\frac{1}{n} \mathbf{\Phi}_F^* \mathbf{\Phi}_F + \lambda\,\mathrm{Id} \right)^{-1} \frac{1}{n} \mathbf{\Phi}_F^* Y = \left(\hat{C}_F + \lambda\,\mathrm{Id} \right)^{-1} \frac{1}{n} \mathbf{\Phi}_F^* Y, \qquad (22)
\end{aligned}
$$

where,

$$
\begin{aligned}
\mathbf{\Phi}_{\hat{F}} &: \mathcal{H}_X \to \mathbb{R}^n \qquad & \mathbf{\Phi}_{\hat{F}} &= [\hat{F}_\xi(z_1), \ldots, \hat{F}_\xi(z_n)]^* \\
\mathbf{\Phi}_F &: \mathcal{H}_X \to \mathbb{R}^n \qquad & \mathbf{\Phi}_F &= [F_*(z_1), \ldots, F_*(z_n)]^*,
\end{aligned}
$$

and,

$$\hat{C}_{\hat{F}} = \frac{1}{n}\mathbf{\Phi}_{\hat{F}}^*\mathbf{\Phi}_{\hat{F}} = \frac{1}{n}\sum_{i=1}^{n}\hat{F}_\xi(z_i)\otimes\hat{F}_\xi(z_i) \qquad \hat{C}_F = \frac{1}{n}\mathbf{\Phi}_F^*\mathbf{\Phi}_F = \frac{1}{n}\sum_{i=1}^{n}F_*(z_i)\otimes F_*(z_i).$$

Let us define $c_{\lambda,\gamma} \doteq \kappa_X^{1-\gamma}\lambda^{-1/2}1_{\mathcal{N}(C_F)\neq\{0\}} + \lambda^{\frac{1-\gamma}{2\gamma_0}-\frac{1}{2}}1_{\mathcal{N}(C_F)=\{0\}}$. We therefore have, combining Eq. (19), Eq. (20), Eq. (21) and Eq. (22),

$$\begin{aligned}\|\hat{h}_\lambda - \bar{h}_\lambda\|_\gamma &\leqslant c_{\lambda,\gamma}\left\|(C_F+\lambda\,\mathrm{Id})^{1/2}\left(\left(\hat{C}_{\hat{F}}+\lambda\,\mathrm{Id}\right)^{-1}\frac{1}{n}\mathbf{\Phi}_{\hat{F}}^*Y - \left(\hat{C}_F+\lambda\,\mathrm{Id}\right)^{-1}\frac{1}{n}\mathbf{\Phi}_F^*Y\right)\right\|_{\mathcal{H}_X}\\ &\leqslant c_{\lambda,\gamma}\left(S_{-1}+S_0\right),\end{aligned}$$

where,

$$S_{-1} \doteq \left\|(C_F+\lambda\,\mathrm{Id})^{1/2}\left(\hat{C}_{\hat{F}}+\lambda\,\mathrm{Id}\right)^{-1}\left(\frac{1}{n}\mathbf{\Phi}_{\hat{F}}^*Y - \frac{1}{n}\mathbf{\Phi}_F^*Y\right)\right\|_{\mathcal{H}_X} \tag{23}$$

$$S_0 \doteq \left\|(C_F+\lambda\,\mathrm{Id})^{1/2}\left(\left(\hat{C}_{\hat{F}}+\lambda\,\mathrm{Id}\right)^{-1}\frac{1}{n}\mathbf{\Phi}_F^*Y - \left(\hat{C}_F+\lambda\,\mathrm{Id}\right)^{-1}\frac{1}{n}\mathbf{\Phi}_F^*Y\right)\right\|_{\mathcal{H}_X}. \tag{24}$$

S_{-1} and S_0 are bounded respectively in Theorem 11 and Theorem 12. Putting them together, we obtain the following bound for the stage 1 error.

Theorem 7. *Let Assumptions 1, (MOM), (SRCZ) and (EMBZ) hold with $\alpha_Z \leqslant \beta_Z$. Let $\gamma \in [0,1]$. For sufficiently large m and n such that Eq. (18) hold, we have with P^n-probability over $1-8e^{-\tau}$, for some $\tau \geqslant 1$,*

$$\|\hat{h}_\lambda - \bar{h}_\lambda\|_\gamma \leqslant c_0\tau\lambda^{\omega_\gamma-1}\left(\left\|F_* - \hat{F}_\xi\right\|_{L_2(Z;\mathcal{H}_X)} + \frac{\left\|\hat{F}_\xi - F_*\right\|_{\alpha_Z}}{\sqrt{n}}\right)\left(\left\|\bar{h}_\lambda\right\|_{\mathcal{H}_X}+1\right),$$

with $\omega_\gamma \doteq \frac{1-\gamma}{2\gamma_0}1_{\mathcal{N}(C_F)} = \{0\}$ and c_0 depending on $\sigma, L, A_Z, B_Z, \kappa_Z, \alpha_Z$ and β_Z.

Therefore, we have upper bounded the stage 1 error by the error $\hat{F}_\xi - F_*$. This error has been thoroughly investigated in Meunier et al. (2024) and we can directly plug-in their results (see Theorem 15).

E.1.2 Stage 2 Error

To analyze the stage 2 error, we assume a known conditional distribution for $X \mid Z$ (or equivalently a known F_* or $\mathcal{T}$). Recall that the NPIV model is given by

$$Y = h_*(X) + U, \qquad \mathbb{E}[U|Z] = 0. \tag{NPIV}$$

As $\mathcal{T}$ is known, we can reformulate (NPIV) to obtain the ***nonparametric indirect regression (NPIR)*** model (Chen and Reiss, 2011),

$$Y = \mathcal{T}h_*(Z) + \xi, \qquad \mathbb{E}[\xi|Z] = 0, \tag{NPIR}$$

where $\xi \doteq h_*(X) - \mathcal{T}h_*(Z) + U$. (NPIR) was first used by Chen and Reiss (2011) to obtain a lower bound for (NPIV). In this section, we use (NPIR) to study the stage 2 error in a straightforward manner. (NPIR) is a typical ***inverse statistical learning problem***, where we observe the image of a function through a known linear operator (here $\mathcal{T}$) at i.i.d. random design points, superposed with an additive noise (see Loustau (2013) and references therein). When restricted to the class of functions $\mathcal{H}_X$ for h_* (and under Assumption 2), using the reproducing property, Eq. (NPIR) reduces to

$$Y = \langle h_*, F_*(Z)\rangle_{\mathcal{H}_X} + \xi, \quad \mathbb{E}(\xi|Z) = 0.$$

In this context of learning with RKHSs it was obtained in Blanchard and Mücke (2018) that the inverse statistical learning problem can be converted to a "standard" statistical learning problem by defining a new RKHS $\mathcal{H}_F$. This will then allow us to apply standard kernel ridge regression results, for example Fischer and Steinwart (2020), to bound the stage 2 error.

To give the reader some insights on how the new RKHS is constructed, we first recall that the ideal estimator for stage 2 (i.e. the estimator built knowing the true CME F_*) is (see Eq. (11))

$$\bar{h}_\lambda = \arg\min_{h \in \mathcal{H}_X} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \langle h, F_*(z_i) \rangle_{\mathcal{H}_X}\right)^2 + \lambda \|h\|_{\mathcal{H}_X}^2 .$$

While the above looks like a kernel ridge regression problem with kernel $k_F(z, z') = \langle F_*(z), F_*(z') \rangle_{\mathcal{H}_X}$, the crucial difference here is that $\mathcal{H}_X$ is not the RKHS generated by k_F. In order to apply standard kernel ridge regression results, we need to find the corresponding RKHS. To this end, consider the operator $V : \mathcal{H}_X \to \mathbb{R}^{\mathcal{Z}}, h \mapsto (z \mapsto \langle h, F_*(z) \rangle_{\mathcal{H}_X})$ and the subset of $\mathbb{R}^{\mathcal{Z}}$

$$\mathcal{H}_F \doteq \{f : \mathcal{Z} \to \mathbb{R} \mid \exists h \in \mathcal{H}_X \text{ with } f(z) = \langle h, F_*(z) \rangle_{\mathcal{H}_X} \text{ for all } z \in \mathcal{Z}\} = \mathcal{R}(V),$$

The objective for stage 2 can be re-written as

$$\arg\min_{h \in \mathcal{H}_X} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \langle h, F_*(z_i) \rangle_{\mathcal{H}_X}\right)^2 + \lambda \|h\|_{\mathcal{H}_X}^2 \Leftrightarrow \arg\min_{h \in \mathcal{H}_X} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - Vh(z_i)\right)^2 + \lambda \|h\|_{\mathcal{H}_X}^2 .$$

By making the change of variable $r = Vh$, we obtain the following minimization objective

$$\bar{r}_\lambda = \arg\min_{r \in \mathcal{H}_F} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - r(z_i)\right)^2 + \lambda \|V^\dagger r\|_{\mathcal{H}_X}^2, \tag{25}$$

where $\bar{r}_\lambda = V\bar{h}_\lambda$ and $V^\dagger$ denotes the pseudo-inverse. Introducing the norm

$$\|r\|_{\mathcal{H}_F} \doteq \|V^\dagger r\|_{\mathcal{H}_X} = \inf\left\{\|h\|_{\mathcal{H}_X} : h \in \mathcal{H}_X \text{ with } r = Vh\right\},$$

we obtain that Eq. (25) can be written as

$$\bar{r}_\lambda = \arg\min_{r \in \mathcal{H}_F} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - r(z_i)\right)^2 + \lambda \|r\|_{\mathcal{H}_F}^2, \tag{26}$$

which now looks like a standard kernel ridge regression objective. To verify this, we need to show that $\mathcal{H}_F$ is indeed a RKHS. Fortunately, this was studied by Steinwart and Christmann (2008) (see also Blanchard and Mücke (2018) where this result is applied to inverse regression) and we recall the result in our own notations for completeness.

Theorem 8 (Theorem 4.21 Steinwart and Christmann (2008))**.** *$\mathcal{H}_F$ endowed with the norm $\|\cdot\|_{\mathcal{H}_F}$ is the unique RKHS associated to the kernel k_F. Furthermore, the operator V is a partial isometry from $\mathcal{H}_X$ to $\mathcal{H}_F$, i.e. an isometry on the orthogonal of its kernel: for all $h, h' \in \mathcal{N}(V)^\perp$, $\langle h, h' \rangle_{\mathcal{H}_X} = \langle Vh, Vh' \rangle_{\mathcal{H}_F}$ or equivalently for all $r, r' \in \mathcal{H}_F$, $\langle r, r' \rangle_{\mathcal{H}_F} = \langle V^\dagger r, V^\dagger r' \rangle_{\mathcal{H}_X}$.*

To summarize, we have reduced the inverse regression model (NPIR) into the least-squares problem

$$Y = r_0(Z) + \xi, \qquad \mathbb{E}[\xi \mid Z] = 0,$$

with the goal of approximating $r_0 \in L_2(Z)$ with kernel ridge regression (estimator $\bar{r}_\lambda$ in Eq. (26)) from data $\mathcal{D}_2 \doteq \{(z_i, y_i)\}_{i=1}^n$.

The generalization error of kernel ridge regression have been thoroughly investigated (Smale and Zhou, 2004, 2005, 2007; Caponnetto and De Vito, 2007). The key quantities that control the upper bounds are the effective dimension, controlling the size of the space, and the smoothness of the target function r_0 relatively to the covariance in $(\mathcal{H}_F, \pi_Z)$. The key difference with the standard kernel ridge regression setting is that our assumptions (SRCX,EVDX) are expressed with respect to C_X the covariance in $(\mathcal{H}_X, \pi_X)$. However, the link condition (LINK) allows us to draw a link between $(\mathcal{H}_F, \pi_Z)$ and $(\mathcal{H}_X, \pi_X)$. Our analysis departs from Blanchard and Mücke (2018) where assumptions on the smoothness of r_0 and the effective dimension are made directly on $(\mathcal{H}_F, \pi_Z)$.

The first step is to identify the covariance operator in $(\mathcal{H}_F, \pi_Z)$. Let us first look at the canonical feature map in $\mathcal{H}_F$. For all $z \in E_Z$ we denote the canonical feature map as $k_{F,z} \doteq (z' \mapsto k_F(z, z'))$. By definition of k_F it is clear that for all $z \in E_Z$, $k_{F,z} = VF_*(z)$. And therefore the covariance associated to the pair $(\mathcal{H}_F, \pi_Z)$ is given by

$$\Sigma_F \doteq \mathbb{E}[k_{F,Z} \otimes k_{F,Z}] = V\mathbb{E}[F_*(Z) \otimes F_*(Z)]V^* = VC_FV^*.$$

We recall that Assumption (LINK) relates C_X to C_F, both operators acting on $\mathcal{H}_X$. The next proposition shows that C_F shares the same eigenspectrum with Σ_F and their eigenvectors are related through V, therefore with Assumption (LINK) we can relate C_X to Σ_F.

Proposition 4 (Link between C_F and Σ_F)**.** *Let $\{(\mu_{F,i}, \sqrt{\mu_{F,i}}e_{F,i})\}_{i\in I}$ be the eigensystem of C_F (see Definition 9), then $\{(\mu_{F,i}, \sqrt{\mu_{F,i}}Ve_{F,i})\}_{i\in I}$ is the eigensystem of Σ_F. Therefore, for all $\lambda > 0$, $\mathcal{N}_\Sigma(\lambda) \doteq \operatorname{Tr}\left((\Sigma_F + \lambda \operatorname{Id}_{\mathcal{H}_F})^{-1}\Sigma_F\right) = \operatorname{Tr}\left((C_F + \lambda \operatorname{Id}_{\mathcal{H}_X})^{-1}C_F\right) = \mathcal{N}_F(\lambda)$ and $\|\Sigma_F\|_{\mathcal{H}_F \to \mathcal{H}_F} = \|C_F\|_{\mathcal{H}_X \to \mathcal{H}_X}$.*

Proof. Let us define $h_{F,i} \coloneqq \sqrt{\mu_{F,i}}Ve_{F,i}$ for all $i \in I$. $(h_{F,i}, \mu_{F,i})_{i\in I}$ is the eigensystem of C_F such that, $\mu_{F,i} \neq 0$ and $\langle h_{F,i}, h_{F,j}\rangle_{\mathcal{H}_X} = \delta_{ij}$ for all $i, j \in I$. $(h_{F,i})_{i\in I}$ forms an ONB of $\ker(C_F)^\perp$. It is easy to verify that $(Vh_{F,i}, \mu_{F,i})_{i\in I}$ is the eigensystem of Σ_F since for all $i \in I$

$$\Sigma_F V h_{F,i} = VC_FV^*Vh_{F,i} = VC_Fh_{F,i} = \mu_{F,i}Vh_{F,i},$$

and for any $i, j \in I$, by the partial isometry property of V,

$$\langle Vh_{F,i}, Vh_{F,j}\rangle_{\mathcal{H}_F} = \langle h_{F,i}, h_{F,j}\rangle_{\mathcal{H}_X} = \delta_{ij}.$$

Note that we can apply the partial isometry property as $h_{F,i}, h_{F,j} \in \ker(C_F)^\perp \subseteq \ker(V)^\perp$. □

To apply Theorem 14, that studies the generalization error of kernel ridge regression, to $\mathcal{H}_F$, we bound the effective dimension $\mathcal{N}_\Sigma(\cdot)$ and characterize the smoothness of r_0 with respect to Σ_F. For the former, under Assumption (LINK) combined with (EVDX), by Lemma 3 and Proposition 4,

$$\mathcal{N}_\Sigma(\lambda) \leqslant D\lambda^{-\frac{p_X}{\gamma_1}}, \quad \lambda \geqslant 0. \tag{27}$$

In the next proposition, we show that r_0 satisfies a source condition with respect to $(\mathcal{H}_F, \pi_Z)$ with parameter $\beta_F \doteq \frac{\beta_X - 1}{\gamma_0} + 1$.

Theorem 9. *Under Assumptions (LINK) with parameter $\gamma_0 \geqslant 1$ and (SRCX) with parameter $1 \leqslant \beta_X \leqslant \gamma_0 + 1$,*

$$\|\Sigma_F^{-\frac{\beta_F - 1}{2}} r_0\|_{\mathcal{H}_F} \leqslant B_X,$$

with $\beta_F = \frac{\beta_X - 1}{\gamma_0} + 1 \in [1, 2]$.

Proof. We first notice that r_0 is almost surely an element of $\mathcal{H}_F$. Indeed, recall that $h_* \in \mathcal{H}_X$ and $r_0 = \mathbb{E}[h_*(X) \mid Z]$, therefore, by the reproducing property, for π_Z−almost all $z \in E_Z$, $r_0(z) = \langle h_*, F_*(z)\rangle_{\mathcal{H}_X} = Vh_*(z)$. By Assumption (LINK) and Proposition 3 d) applied with $\tau = (\beta_X - 1)/\gamma_0$ we have $P_FC_X^{\beta_X - 1}P_F \preceq C_F^{\frac{\beta_X - 1}{\gamma_0}}$. Note that $\tau \in [0,1]$ since $1 \leqslant \beta_X \leqslant \gamma_0 + 1$. As we have $h_* \in \overline{\mathcal{R}(C_F)}$ and $h_* \in \mathcal{R}(C_X^{\frac{\beta_X-1}{2}})$, we can apply Proposition 7 with $A = C_X^{\beta_X-1}$ and $B = C_F^{\frac{\beta_X-1}{\gamma_0}}$ to obtain,

$$\|C_F^{-\frac{\beta_X-1}{2\gamma_0}} h_*\|_{\mathcal{H}_X} \leqslant \|C_X^{-\frac{\beta_X-1}{2}} h_*\|_{\mathcal{H}_X} \leqslant B_X.$$

To conclude, by Proposition 4 and using $V^\dagger r_0 = h_*$, it is readily seen that $\|C_F^{-\frac{\beta_X-1}{2\gamma_0}} h_*\|_{\mathcal{H}_X} = \|\Sigma_F^{-\frac{\beta_X-1}{2\gamma_0}} r_0\|_{\mathcal{H}_F}$. □

We are now ready to apply Theorem 14 to $\mathcal{H}_F$.

Theorem 10 (Stage 2 error)**.** *Let Assumptions 1, 2,* (EMBZ)*,* (SRCZ)*,* (EVDX)*,* (SRCX)*,* (MOM) *and* (LINK) *hold with* $1 \leqslant \beta_X \leqslant \gamma_0 + 1$ *and* $0 < \alpha_Z \leqslant \beta_Z$*. Then for the abbreviations*

$$g_\lambda \doteq \log\left(2e\mathcal{N}_F(\lambda)\frac{\|C_F\|_{\mathcal{H}_X\to\mathcal{H}_X} + \lambda}{\|C_F\|_{\mathcal{H}_X\to\mathcal{H}_X}}\right)$$
$$A_{\lambda,\tau} \doteq 8\tau g_\lambda \kappa_Z^{\beta_Z-\alpha_Z} A_Z^2 B_Z^2 \lambda^{-1},$$

and $\tau \geqslant 1, 0 < \lambda \leqslant 1, \gamma \in [0,1]$*, and* $n \geqslant A_{\lambda,\tau}$*, with* P^n*-probability not less than* $1 - 4e^{-\tau}$*, the stage 2 error can be bounded as*

$$\|\bar{h}_\lambda - h_*\|_\gamma \leqslant c_1\left(\lambda^{\frac{\beta_X-\gamma}{2\gamma_0}} + \tau\sqrt{1 + \frac{1}{n\lambda^{1-\frac{p_X}{\gamma_1}}}}\sqrt{\frac{1}{n\lambda^{1-\frac{1-\gamma}{\gamma_0}+\frac{p_X}{\gamma_1}}}}\right)$$

where c_1 *only depends on* $B_X, \sigma, L, \kappa_Z, A_Z, B_Z, \alpha_Z, \beta_Z$ *and* $\|h_*\|_{\mathcal{H}_X}$*.*

Proof. We check assumptions 1., 2., 3. and 4. from Theorem 14, applied to $\mathcal{H}_F$ for the estimation of $r_0 = \mathbb{E}[Y \mid Z]$.

- Assumption 1. By Lemma 12, under (EMBZ) and (SRCZ), for π_Z−almost every $z \in E_Z$, $k_F(z,z) \leqslant \|F_*(z)\|^2_{\mathcal{H}_X} \leqslant \kappa_F^2$ with $\kappa_F = \kappa_Z^{\frac{\beta_Z-\alpha_Z}{2}} A_Z B_Z$. Therefore we take $\kappa = \kappa_F$ in Theorem 14.
- Assumption 2. is satisfied by Assumption (MOM).
- Assumption 3. is satisfied with $p = p_X/\gamma_1$ by Eq. (27).
- Assumption 4. is satisfied with $\beta = \beta_F = \frac{\beta_X-1}{\gamma_0} + 1 \in [1,2]$ and $B = B_X$ by Theorem 9.

By Theorem 14, we therefore obtain, for the abbreviations

$$g_\lambda \doteq \log\left(2e\mathcal{N}_\Sigma(\lambda)\frac{\|\Sigma_F\|_{\mathcal{H}_F\to\mathcal{H}_F} + \lambda}{\|\Sigma_F\|_{\mathcal{H}_F\to\mathcal{H}_F}}\right) = \log\left(2e\mathcal{N}_F(\lambda)\frac{\|C_F\|_{\mathcal{H}_X\to\mathcal{H}_X} + \lambda}{\|C_F\|_{\mathcal{H}_X\to\mathcal{H}_X}}\right)$$
$$A_{\lambda,\tau} \doteq 8\tau g_\lambda \kappa_Z^{\beta_Z-\alpha_Z} A_Z^2 B_Z^2 \lambda^{-1},$$

and $0 \leqslant \theta \leqslant 1, \tau \geqslant 1, 0 < \lambda \leqslant 1$, and $n \geqslant A_{\lambda,\tau}$, that the following bound is satisfied with P^n-probability not less than $1 - 4e^{-\tau}$,

$$\|\Sigma_F^{\frac{1-\theta}{2}}(\bar{r}_\lambda - r_0)\|^2_{\mathcal{H}_F} \leqslant J\left(\lambda^{\beta_F-\theta} + \frac{\tau^2}{n\lambda^{\theta+\frac{p_X}{\gamma_1}}}\left(1 + \frac{1}{n\lambda^{1-\frac{p_X}{\gamma_1}}}\right)\right), \tag{28}$$

where J is a constant depending on $B_X, \sigma, L, \kappa_Z, A_Z, B_Z, \alpha_Z, \beta_Z$ and $\|r_0\|_{\mathcal{H}_F}$. Note that $\|r_0\|_{\mathcal{H}_F} = \|h_*\|_{\mathcal{H}_X}$.

Finally, by Assumption (LINK),

$$\|\bar{h}_\lambda - h_*\|_\gamma = \|C_X^{\frac{1-\gamma}{2}}(\bar{h}_\lambda - h_*)\|_{\mathcal{H}_X} \leqslant \|C_F^{\frac{1-\gamma}{2\gamma_0}}(\bar{h}_\lambda - h_*)\|_{\mathcal{H}_X} = \|\Sigma_F^{\frac{1-\gamma}{2\gamma_0}}(\bar{r}_\lambda - r_0)\|_{\mathcal{H}_F},$$

where the last equality follows from Proposition 4 and using $\bar{h}_\lambda = V^\dagger \bar{r}_\lambda, h_* = V^\dagger r_0$. Plugging $\theta = 1 - \frac{1-\gamma}{\gamma_0}$ and $\beta_F = \frac{\beta_X - 1}{\gamma_0} + 1$ and $\frac{\beta_X - 1}{\gamma_0} + 1$ in Eq. (28) concludes the proof. □

Theorem 3 is then obtained by combining Theorem 7 and 10, which study the stage 1 and 2 error respectively.

E.2 Detailed Proof

E.2.1 Stage 1 Error

The following theorem provides an upper bound on Eq. (23), term S_{-1}.

Theorem 11. *Let Assumptions 1, (MOM), (SRCZ) and (EMBZ) hold with $\alpha_Z \leqslant \beta_Z$, we have with P^n-probability over $1 - 8e^{-\tau}$, for some $\tau > 0$, and sufficiently large m and n such that Eq. (18) is satisfied,*

$$S_{-1} \leqslant c\frac{\tau}{\sqrt{\lambda}}\left(\frac{\|\hat{F}_\xi - F_*\|_{\alpha_Z}}{\sqrt{n}} + \|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)}\right),$$

with c depending on $\sigma, L, A_Z, B_Z, \kappa_Z, \alpha_Z$ and β_Z.

Proof. We start with the following decomposition,

$$\begin{aligned}
&\left\|(C_F + \lambda \mathrm{Id})^{1/2}\left(\hat{C}_{\hat{F}} + \lambda \mathrm{Id}\right)^{-1}\left(\frac{1}{n}\mathbf{\Phi}_{\hat{F}}^* Y - \frac{1}{n}\mathbf{\Phi}_{F_*}^* Y\right)\right\|_{\mathcal{H}_X} \\
\leqslant &\left\|(C_F + \lambda \mathrm{Id})^{1/2}\left(\hat{C}_{\hat{F}} + \lambda \mathrm{Id}\right)^{-1/2}\right\|_{\mathcal{H}_X \to \mathcal{H}_X}\left\|\left(\hat{C}_{\hat{F}} + \lambda \mathrm{Id}\right)^{-1/2}\left(\hat{C}_F + \lambda \mathrm{Id}\right)^{1/2}\right\|_{\mathcal{H}_X \to \mathcal{H}_X} && (29) \\
\cdot &\left\|\left(\hat{C}_F + \lambda \mathrm{Id}\right)^{-1/2}\left(\frac{1}{n}\mathbf{\Phi}_{\hat{F}}^* Y - \frac{1}{n}\mathbf{\Phi}_{F_*}^* Y\right)\right\|_{\mathcal{H}_X} && (30)
\end{aligned}$$

Apply Lemma 10, we obtain with probability over $1 - 6e^{-\tau}$ and sufficiently large m and n such that the constraints in Eq. (18) are satisfied,

$$\left\|(C_F + \lambda \mathrm{Id})^{1/2}\left(\hat{C}_{\hat{F}} + \lambda \mathrm{Id}\right)^{-1/2}\right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant 3.$$

For the second term in Eq. (29), under the constraints of Eq. (18), by Lemma 9, with probability over $1 - 4e^{-\tau}$

$$\left\|\left(\hat{C}_{\hat{F}} + \lambda \mathrm{Id}\right)^{-1/2}\left(\hat{C}_F + \lambda \mathrm{Id}\right)^{1/2}\right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant \sqrt{\frac{6}{5}},$$

and therefore

$$\text{Eq. (29)} \leqslant 3\sqrt{\frac{6}{5}} \leqslant 4.$$

For Eq. (30)

$$\begin{aligned}
&\left\| \left(\hat{C}_F + \lambda \mathrm{Id}\right)^{-1/2} \left(\frac{1}{n}\mathbf{\Phi}_{\hat{F}}^* Y - \frac{1}{n}\mathbf{\Phi}_{F_*}^* Y\right) \right\|_{\mathcal{H}_X} \\
&= \left\| \left(\hat{C}_F + \lambda \mathrm{Id}\right)^{-1/2} \frac{1}{n} \left(\mathbf{\Phi}_{\hat{F}} - \mathbf{\Phi}_{F_*}\right)^* Y \right\|_{\mathcal{H}_X} \\
&= \left\| \left(\hat{C}_F + \lambda \mathrm{Id}\right)^{-1/2} \frac{1}{n} \left(\mathbf{\Phi}_{\hat{F}} - \mathbf{\Phi}_{F_*}\right)^* \left(Y - \mathbf{\Phi}_{F_*} h_* + \mathbf{\Phi}_{F_*} h_*\right) \right\|_{\mathcal{H}_X} \\
&\leqslant \underbrace{\left\| \left(\hat{C}_F + \lambda \mathrm{Id}\right)^{-1/2} \frac{1}{n} \left(\mathbf{\Phi}_{\hat{F}} - \mathbf{\Phi}_{F_*}\right)^* \left(Y - \mathbf{\Phi}_{F_*} h_*\right) \right\|_{\mathcal{H}_X}}_{I} + \underbrace{\left\| \left(\hat{C}_F + \lambda \mathrm{Id}\right)^{-1/2} \frac{1}{n} \left(\mathbf{\Phi}_{\hat{F}} - \mathbf{\Phi}_{F_*}\right)^* \mathbf{\Phi}_{F_*} h_* \right\|_{\mathcal{H}_X}}_{II}. \qquad (31)
\end{aligned}$$

For term I, we notice that

$$\begin{aligned}
&\left\| \left(\hat{C}_F + \lambda \mathrm{Id}\right)^{-1/2} \frac{1}{n} \left(\mathbf{\Phi}_{\hat{F}} - \mathbf{\Phi}_{F_*}\right)^* \left(Y - \mathbf{\Phi}_{F_*} h_*\right) \right\|_{\mathcal{H}_X} \\
&\leqslant \lambda^{-1/2} \left\| \frac{1}{n} \left(\mathbf{\Phi}_{\hat{F}} - \mathbf{\Phi}_{F_*}\right)^* \left(Y - \mathbf{\Phi}_{F_*} h_*\right) \right\|_{\mathcal{H}_X} \\
&= \lambda^{-1/2} \left\| \frac{1}{n} \sum_{i=1}^{n} \left(\hat{F}_\xi(z_i) - F_*(z_i)\right) \left(y_i - \langle h_*, F_*(z_i)\rangle_{\mathcal{H}_X}\right) \right\|_{\mathcal{H}_X},
\end{aligned}$$

Let us define $\theta(z, y) = \left(\hat{F}_\xi(z) - F_*(z)\right)\left(y - \langle h_*, F_*(z)\rangle_{\mathcal{H}_X}\right)$ and note that $\theta(Z, Y)$ is centered. For $m \geqslant 2$,

$$\begin{aligned}
\mathbb{E}\left\|\theta(Z,Y)\right\|_{\mathcal{H}_X}^m &= \int_{E_Z} \left\|\hat{F}_\xi(z) - F_*(z)\right\|_{\mathcal{H}_X}^m \int_{\mathbb{R}} (y - \langle h_*, F_*(z)\rangle_{\mathcal{H}_X})^m P(\mathrm{d}y \mid z) \mathrm{d}\pi_Z(z) \\
&\leqslant \left(A_Z \|\hat{F}_\xi - F_*\|_{\alpha_Z}\right)^m \int_{\mathbb{R}} (y - \langle h_*, F_*(z)\rangle_{\mathcal{H}_X})^m P(\mathrm{d}y \mid z),
\end{aligned}$$

where we used Lemma 12. Using (MOM) yields

$$\int_{\mathbb{R}} (y - \langle h_*, F_*(z)\rangle_{\mathcal{H}_X})^m P(\mathrm{d}y \mid z) \leqslant \frac{1}{2} m! \sigma^2 L^{m-2}.$$

We therefore have

$$\mathbb{E}\left\|\theta(Z,Y)\right\|_{\mathcal{H}_X}^m \leqslant \frac{1}{2} m! \left(\sigma A_Z \|\hat{F}_\xi - F_*\|_{\alpha_Z}\right)^2 \left(L A_Z \|\hat{F}_\xi - F_*\|_{\alpha_Z}\right)^{m-2}.$$

Using Theorem 16, we have with P^n-probability over $1 - 2e^{-\tau}$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \theta(z_i, y_i) \right\|_{\mathcal{H}_X} \leqslant \sqrt{32} \frac{\tau}{\sqrt{n}} \left(\sigma A_Z \|\hat{F}_\xi - F_*\|_{\alpha_Z} + \frac{L A_Z \|\hat{F}_\xi - F_*\|_{\alpha_Z}}{\sqrt{n}} \right), \qquad (32)$$

and therefore,

$$I \leqslant \sqrt{32} \lambda^{-1/2} A_Z \|\hat{F}_\xi - F_*\|_{\alpha_Z} \frac{\tau(\sigma + L)}{\sqrt{n}}.$$

For term II, using the same proof technique as in Lemma 7, we have, with P^n probability over $1 - 2e^{-\tau}$,

$$\left\| \left(\hat{C}_F + \lambda \mathrm{Id}\right)^{-1/2} \frac{1}{n} \left(\mathbf{\Phi}_{\hat{F}} - \mathbf{\Phi}_{F_*}\right)^* \mathbf{\Phi}_{F_*} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant \frac{\kappa_Z^{\frac{\beta_Z - \alpha_Z}{2}} A_Z B_Z}{\sqrt{\lambda}} \left(A_Z \|\hat{F}_\xi - F_*\|_{\alpha_Z} \sqrt{\frac{\tau}{n}} + \|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)} \right).$$

This further implies that, with probability at least $1 - 8e^{-\tau}$,

$$\begin{aligned}
S_{-1} &\leqslant 4 \left(\sqrt{32} A_Z \|\hat{F}_\xi - F_*\|_{\alpha_Z} \frac{\tau(\sigma + L)}{\sqrt{\lambda n}} + \frac{\kappa_Z^{\frac{\beta_Z - \alpha_Z}{2}} A_Z B_Z}{\sqrt{\lambda}} \left(A_Z \|\hat{F}_\xi - F_*\|_{\alpha_Z} \sqrt{\frac{\tau}{n}} + \|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)} \right) \right) \\
&\leqslant c \frac{\tau}{\sqrt{\lambda}} \left(\frac{\|\hat{F}_\xi - F_*\|_{\alpha_Z}}{\sqrt{n}} + \|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)} \right),
\end{aligned}$$

with c depending on $\sigma, L, A_Z, B_Z, \kappa_Z, \alpha_Z$ and β_Z. $\square$

The following theorem provides an upper bound on Eq. (24), term S_0.

Theorem 12. *Let Assumptions 1,* (SRCZ) *and* (EMBZ) *hold with $\alpha_Z \leqslant \beta_Z$, for sufficiently large n and m such that Eq.* (18) *holds, we have with P^n-probability over $1 - 6e^{-\tau}$*

$$S_0 \leqslant c' \frac{\tau}{\sqrt{\lambda}} \left(\frac{\|\hat{F}_\xi - F_*\|_{\alpha_Z}}{\sqrt{n}} + \|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)} \right) \|\bar{h}_\lambda\|_{\mathcal{H}_X} ,$$

with c' depending on $A_Z, B_Z, \alpha_Z, \beta_Z$ and κ_Z.

Proof.

$$\begin{aligned}
&\left\| (C_F + \lambda \operatorname{Id})^{1/2} \left(\left(\hat{C}_{\hat{F}} + \lambda \operatorname{Id} \right)^{-1} \frac{1}{n} \mathbf{\Phi}^*_{F_*} Y - \left(\hat{C}_F + \lambda \operatorname{Id} \right)^{-1} \frac{1}{n} \mathbf{\Phi}^*_{F_*} Y \right) \right\|_{\mathcal{H}_X} \\
&= \left\| (C_F + \lambda \operatorname{Id})^{1/2} \left(\left(\hat{C}_{\hat{F}} + \lambda \operatorname{Id} \right)^{-1} - \left(\hat{C}_F + \lambda \operatorname{Id} \right)^{-1} \right) \frac{1}{n} \mathbf{\Phi}^*_{F_*} Y \right\|_{\mathcal{H}_X} \\
&= \left\| (C_F + \lambda \operatorname{Id})^{1/2} \left(\hat{C}_{\hat{F}} + \lambda \operatorname{Id} \right)^{-1} \left(\hat{C}_{\hat{F}} - \hat{C}_F \right) \left(\hat{C}_F + \lambda \operatorname{Id} \right)^{-1} \frac{1}{n} \mathbf{\Phi}^T_{F_*} Y \right\|_{\mathcal{H}_X} \\
&\leqslant \lambda^{-1/2} \left\| (C_F + \lambda \operatorname{Id})^{1/2} \left(\hat{C}_{\hat{F}} + \lambda \operatorname{Id} \right)^{-1/2} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \left\| \hat{C}_{\hat{F}} - \hat{C}_F \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \left\| \bar{h}_\lambda \right\|_{\mathcal{H}_X} .
\end{aligned}$$

For the first term, by Lemma 11, for any $\tau \geqslant 1$, $\lambda > 0$ and sufficiently large m and n such that the constraints of Eq. (18) are satisfied, with probability at least $1 - 2e^{-\tau}$,

$$\left\| \left(\hat{C}_F + \lambda \operatorname{Id} \right)^{-1/2} (C_F + \lambda \operatorname{Id})^{1/2} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant 2.$$

For the second term, by Lemma 7, under the assumptions that $\|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)} \leqslant 1$, and $\|F_* - \hat{F}_\xi\|_{\alpha_Z} \leqslant 1$ with P^n-probability over $1 - 4e^{-\tau}$, it holds that

$$\left\| \hat{C}_F - \hat{C}_{\hat{F}} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant J \left(\sqrt{\frac{\tau}{n}} \|F_* - \hat{F}_\xi\|_{\alpha_Z} + \|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)} \right),$$

where J depends on $A_Z, B_Z, \alpha_Z, \beta_Z$ and κ_Z. $\square$

F Proof of Theorem 4

We adopt a strategy similar to the one presented in Chen and Reiss (2011). The proof of the lower bound consists of two steps. First, we re-introduced the nonparametric indirect regression (NPIR) model used to bound the stage 2 error in Section E.1.2. We show that estimators for NPIR have lower risk compared to the NPIV model. We then provide a lower bound for the NPIR model, where the target function belongs to a RKHS. By comparing this to our upper bound for Kernel NPIV, we identify the settings in which minimax optimality is achieved.

In this section, as we focus on deriving a lower bound, we assume that $\mathcal{T}$ is an injective operator. This assumption corresponds to the completeness assumption commonly used in the NPIV literature (Newey and Powell, 2003).

Assumption 11 (Completeness)**.** *Let X, Z be two random variables taking values in E_X and E_Z respectively. We say that (X, Z) satisfies the completeness assumption if the operator $\mathcal{T}$ defined in Eq.* (2) *is injective.*

Note that under the completeness assumption, there is no distinction between the structural function (denoted h_0) and the minimum norm solution to the NPIV model (denoted h_*).

F.1 Relationship Between the NPIR Model and the NPIV Model

Recall that a NPIV model depends on a joint distribution $\pi_{X,Y,Z}$, a structural function $h_0 \in L_2(X)$ and takes the following form:

$$Y = h_0(X) + U, \qquad \mathbb{E}[U|Z] = 0$$

Under the completeness assumption (Assumption 11) and the assumption that $r_0 \in \mathcal{R}(\mathcal{T})$, h_0 is identified as the unique solution to the integral equation given in Eq. (2): $r_0 = \mathcal{T}h_0$. We define $\tilde{\mathcal{F}}$, as the set of models (NPIV) with $(\pi_{X,Y,Z}, h_0)$ such that $r_0 \in \mathcal{R}(\mathcal{T})$ and Assumption 11 hold. We saw in Section E.1.2 that when $\mathcal{T}$ is known, (NPIV) can be reformulated as the NPIR model

$$Y = \mathcal{T}h_0(Z) + \xi, \qquad \xi = h_0(X) - \mathcal{T}h_0(Z) + U,$$

where $\mathbb{E}[\xi|Z] = 0$. To obtain a lower bound for (NPIR), we make the following restriction on the set $\tilde{\mathcal{F}}$.

Definition 7. *Let $\sigma_0 > 0$ be a finite constant and $\mathcal{C}$ be a subset of $L_2(X)$. Let $\mathcal{F}$ be the subset of $\tilde{\mathcal{F}}$ such that for all $h_0 \in \mathcal{C}$, there is a joint probability distribution $\pi_{X,Y,Z}$ with $(\pi_{X,Y,Z}, h_0) \in \mathcal{F}$ such that $\pi_{Y,Z}$ is determined by $\pi_{X,Y,Z}$ and h_0, and such that*

$$\xi \doteq Y - \mathbb{E}[Y \mid Z] = h_0(X) - \mathcal{T}h_0(Z) + U$$

given Z is $\mathcal{N}(0, \sigma^2(Z))$-distributed with $\sigma^2(Z) \geqslant \sigma_0^2$ almost surely.

Example 1. *A simple example of building an element in $\mathcal{F}$ is to take Z from an arbitrary probability distribution π_Z, followed by generating X according to a conditional density of X given Z. We then sample ξ from $\mathcal{N}(0, \sigma^2(Z))$ and define*

$$U \doteq \mathcal{T}h_0(Z) - h_0(X) + \xi.$$

For each (NPIV) model $(\pi_{X,Y,Z}, h_0)$ in $\mathcal{F}$, an (NPIR) model is built, assuming that the operator $\mathcal{T}$ is known. We formally define the NPIR class as follows.

Definition 8. *The NPIR model class $\mathcal{F}_0$ consists of all model parameters $(\pi_{Z'}, \sigma^2(\cdot), h_0)$ such that there is a corresponding $(\pi_{X,Y,Z}, h_0) \in \mathcal{F}$ with the following properties: $\pi_Z = \pi_{Z'}$, $\sigma^2(Z) \geqslant \sigma_0^2$ almost surely, the conditional distribution of X given Z is fully prescribed according to $\mathcal{T}$, and the distribution of U given (X, Z) is arbitrary among the conditions imposed in $\mathcal{F}$.*

We now demonstrate that in order to obtain a lower bound for (NPIV), it suffices to prove a lower bound for (NPIR). Given data $(X_i, Y_i, Z_i)_{i=1}^n$, let $\hat{h}_n((X_i, Y_i, Z_i)_{i=1}^n)$ be an estimator for the NPIV model. Note that knowing $\mathcal{T}$ is equivalent to knowing the conditional law of X given Z. Let us call the observations in the NPIR model $(Y_i, Z_i)_{i=1}^n$ generated by some $(\pi_{Z'}, \sigma^2(\cdot), h_0) \in \mathcal{F}_0$. We then generate artificially i.i.d. observations $(X_i)_{i=1}^n$ according to the conditional law of $X \mid Z = z$ with $Z = z_i$. Then the observations $(X_i, Y_i, Z_i)_{i=1}^n$ follow the law of some $(\pi_{X,Y,Z}, h_0) \in \mathcal{F}$ by definition of $\mathcal{F}_0$ and $\mathcal{F}$. Consequently, the estimator $\tilde{h}((Y_i, Z_i)_{i=1}^n) \doteq \hat{h}_n((X_i, Y_i, Z_i)_{i=1}^n)$ for the NPIR model is such that $\tilde{h} - h_0$ under $(\pi_{Z'}, \sigma^2(\cdot), h_0) \in \mathcal{F}_0$ is equal to $\hat{h} - h_0$ under $(\pi_{X,Y,Z}, h_0) \in \mathcal{F}$. This argument shows that (NPIV) in $\mathcal{F}$ is statistically more demanding than learning (NPIR) in $\mathcal{F}_0$. This discussion is adapted from the proof of Chen and Reiss (2011, Lemma 1).

F.2 The Lower Bound for Kernel NPIR Model

In this section, we provide a lower bound for the NPIR model, as by Section F.1, it implies a lower bound for the NPIV model. Our proof differs from Chen and Reiss (2011) in two ways. Firstly, the proof of Chen and

Reiss (2011, Theorem 1) is based on Assouad's cube technique (Tsybakov, 2009, Lemma 2.12) and only provides a lower bound in expectation. To obtain a lower bound with high probability, we instead rely on Tsybakov (2009, Theorem 2.5) and follow the same template as Fischer and Steinwart (2020) for plain least-squares regression. Secondly, our proof specializes to the RKHS learning setting, considering both L_2-rates and rates in the interpolation norms. We take $\mathcal{C} = \mathcal{H}_X$ in Definition 7 and we fix parameters $\beta_X \geqslant 0$, $\gamma \in [0,1]$ and $B_X, \sigma^2, L > 0$.

Let $\mathcal{H}_X$ be a RKHS of functions form E_X to $\mathbb{R}$ equipped with a marginal distribution π_X such that Assumption (EVDX+) is satisfied. We consider elements $(\pi_Z, \sigma^2(\cdot), h_0) \in \mathcal{F}_0$ with π_Z and $\sigma^2(\cdot)$ fixed such that $\sigma^2(\cdot) \geqslant \sigma_0 = \min\{\sigma, L\}$ while $h_0 \in \mathcal{H}_X$ is a free parameter that we will choose to obtain the lower bound. Given $h \in \mathcal{H}_X$ we write P_h the joint probability for (Y, Z), to indicate the dependence on h.

We recall (see Eq. (3)), that C_X admits the following decomposition:

$$C_X = \mathbb{E}[\phi_X(X) \otimes \phi_X(X)] = \sum_{i \geqslant 1} \mu_{X,i} \sqrt{\mu_{X,i}} e_{X,i} \otimes \sqrt{\mu_{X,i}} e_{X,i},$$

where $\{\sqrt{\mu_{X,i}} e_{X,i}\}_{i \geqslant 1}$ is an ONB of $\overline{\mathcal{R}(C_X)} \subseteq \mathcal{H}_X$ and $\{[e_{X,i}]\}_{i \geqslant 1}$ is an ONB of $\overline{\mathcal{R}(\mathcal{I}_X)} \subseteq L_2(X)$. We recall the definition of the Kullback-Leibler divergence. For two probability measures P_1, P_2 on some measurable space $(\Omega, \mathcal{A})$ the Kullback-Leibler divergence is given by

$$KL(P_1, P_2) := \int_\Omega \log\left(\frac{\mathrm{d}P_1}{\mathrm{d}P_2}\right) \mathrm{d}P_1$$

if $P_1 \ll P_2$ and otherwise $K(P_1, P_2) := \infty$

For $\omega = \{\omega_i\}_{i \geqslant 1}$ with $\omega_i \in \{-1, +1\}$, $0 < \epsilon \leqslant 1$ and $m \in \mathbb{N}$, we consider the following function in $\mathcal{H}_X$:

$$h_\omega \doteq 2\left(\frac{8\varepsilon}{m}\right)^{1/2} \sum_{i=1}^m \omega_i \mu_{X,i+m}^{\gamma/2} e_{X,i+m}.$$

We distinguish the following steps to obtain the lower bound.

- Step 1: Control the separation in $\gamma-$norm between the different h_ω;
- Step 2: Control the KL divergence between NPIR models induced by the different h_ω;
- Step 3: Check that h_ω satisfy the conditions (SRCX) with parameters β_X and B_X and (MOM);

Step 1. Assume that $\sum_{i=1}^m (\omega_i - \omega_i')^2 \geqslant m/8$ (this will be ensured later by Lemma 1). Then,

$$\|h_\omega - h_{\omega'}\|_\gamma^2 = \frac{32\epsilon}{m} \sum_{i=1}^m (\omega_i - \omega_i')^2 \geqslant 4\epsilon.$$

Step 2. Recall that in the NPIR model, given $h \in \mathcal{H}_X$, for all $z \in E_Z$, $P_h(\cdot \mid z) = \mathcal{N}(\langle h, F_*(z)\rangle_{\mathcal{H}_X}, \sigma^2(z))$ and $\sigma^2(z) \geqslant \sigma_0^2$. For $h, h' \in \mathcal{H}_X$, we therefore have,

$$\begin{aligned} KL(P_h, P_{h'}) &= \int_{E_Z} KL(P_h(\cdot \mid z), P_{h'}(\cdot \mid z)) d\pi_Z(z) \\ &= \frac{1}{2} \int_{E_Z} \frac{\langle h - h', F_*(z)\rangle_{\mathcal{H}_X}^2}{\sigma^2(z)} d\pi_Z(z) \\ &\leq \frac{1}{2\sigma_0^2} \int_{E_Z} \langle h - h', F_*(z)\rangle_{\mathcal{H}_X}^2 d\pi_Z(z) \\ &= \frac{1}{2\sigma_0^2} \|C_F^{1/2}(h - h')\|_{\mathcal{H}_X}^2. \end{aligned}$$

By Assumptions (LINK) and (EVDX), we have

$$\begin{aligned}\|C_F^{1/2}(h_\omega - h_{\omega'})\|^2_{\mathcal{H}_X} &\leqslant \|C_X^{\gamma_1/2}(h_\omega - h_{\omega'})\|^2_{\mathcal{H}_X} \\ &= \frac{32\epsilon}{m}\sum_{i=1}^m (\omega_i - \omega'_i)^2 \mu_{X,i+m}^{\gamma_1 - 1 + \gamma} \\ &\leqslant 32\epsilon\mu_{X,m}^{\gamma_1 - 1+\gamma} \\ &\leqslant 32 c_2^{\gamma_1 - 1 + \gamma}\epsilon m^{-\frac{\gamma_1 - 1 + \gamma}{p_X}}\end{aligned}$$

and therefore,

$$KL(P_{h_\omega}^{\otimes^n}, P_{h_{\omega'}}^{\otimes^n}) \leqslant \frac{n}{2\sigma_0^2} 32 c_2^{\gamma_1 - 1+\gamma}\epsilon m^{-\frac{\gamma_1 - 1 + \gamma}{p_X}}.$$

Step 3. For any $h \in \mathcal{H}_X$, Assumption (MOM) is satisfied with parameters σ and L. Indeed, under the NPIR model, for any $z \in E_Z$, $Y - \langle h, F_*(z)\rangle_{\mathcal{H}_X} \sim \mathcal{N}(0, \sigma^2(z))$ and the conclusion can be obtained as in Fischer and Steinwart (2020, Lemma 21). We next consider the condition $\|C_X^{-\frac{\beta_X - 1}{2}} h_\omega\|_{\mathcal{H}_X} \leqslant B_X$. Using Assumption (EVDX+), we get

$$\|C_X^{-\frac{\beta_X - 1}{2}} h_\omega\|^2_{\mathcal{H}_X} = \frac{32\epsilon}{m}\sum_{i=1}^m \omega_i^2 \mu_{X,i+m}^{-(\beta_X - \gamma)} \leqslant 32\epsilon\mu_{X,2m}^{-(\beta_X-\gamma)} \leqslant 32 c_1^{-(\beta_X - \gamma)} 2^{\frac{\beta_X - \gamma}{p_X}} \epsilon m^{\frac{\beta_X - \gamma}{p_X}} \leqslant B_X^2$$

for $m \leqslant U\epsilon^{-\frac{p_X}{\beta_X - \gamma}}$ with $U \doteq \left(B_X^2/32\right)^{\frac{p_X}{\beta_X-\gamma}} c_1^{p_X}/2$. We have proved the following: for $u = \frac{p_X}{\beta_X - \gamma}$, for all $0 \leqslant \beta_X$, there are constants $U > 0$ and $0 < \epsilon \leqslant 1$ such that for all $m \leqslant U\epsilon^{-u}$ the function h_ω satisfies the bound $\|C_X^{-\frac{\beta_X - 1}{2}} h_\omega\|_{\mathcal{H}_X} \leqslant B_X$ for all $\omega \in \{0,1\}^m$.

To conclude we use the following theorem that is derived from Tsybakov (2009, Proposition 2.3) and (Fischer and Steinwart, 2020, Theorem 20).

Theorem 13. *Let $M \geqslant 2, (\Omega, \mathcal{A})$ be a measurable space, $P_0, P_1, \ldots, P_M$ be probability measures on $(\Omega, \mathcal{A})$ with $P_j \ll P_0$ for all $j = 1, \ldots, M$, and $0 < \alpha_* < \infty$ with*

$$\frac{1}{M}\sum_{j=1}^M KL\left(P_j, P_0\right) \leqslant \alpha_*.$$

Then, for all measurable functions $\Psi : \Omega \to \{0, 1, \ldots, M\}$, the following bound is satisfied

$$\max_{j=0,1,\ldots,M} P_j(\omega \in \Omega : \Psi(\omega) \neq j) \geqslant \frac{\sqrt{M}}{1+\sqrt{M}}\left(1 - \frac{3\alpha_*}{\log(M)} - \frac{1}{2\log(M)}\right).$$

Two obtain the distributions $P_0, P_1, \ldots, P_M$ we use the following lemma (Tsybakov, 2009, Lemma 2.9).

Lemma 1 (Gilbert-Varshamov Bound). *For $m \geqslant 8$ there exists some $M \geqslant 2^{m/8}$ and some binary strings $\omega^{(0)}, \ldots, \omega^{(M)} \in \{0,1\}^m$ with $\omega^{(0)} = (0, \ldots, 0)$ and*

$$\sum_{i=1}^m \left(\omega_i^{(j)} - \omega_i^{(k)}\right)^2 \geqslant m/8$$

for all $j \neq k$, where $\omega^{(k)} = \left(\omega_1^{(k)}, \ldots, \omega_m^{(k)}\right)$.

Define $\epsilon_0 \doteq \min\left\{\epsilon, (U/9)^{1/u}\right\}$ and $m_\epsilon \doteq \lfloor U\epsilon^{-u}\rfloor$. Now, we fix an $n \geqslant 1$ and an $0 < \epsilon \leqslant \epsilon_0$. Since $m_\epsilon \geqslant 9$, the Gilbert-Varshamov Bound Lemma yields at least $M_\epsilon \doteq \left\lceil 2^{m_\epsilon/8}\right\rceil \geqslant 3$ binary strings $\omega^{(0)}, \omega^{(1)}, \ldots, \omega^{(M_\epsilon)} \in \{0,1\}^{m_\epsilon}$

satisfying the Gilbert-Varshamov Bound. For $j = 0, 1, \dots, M_\epsilon$, the corresponding functions $h_j \doteq h_{\omega(j)}$ satisfy the bound $\|C_X^{-\frac{\beta_X - 1}{2}} h_j\|_{\mathcal{H}_X} \leqslant B_X$. Due to the definitions of M_ϵ, m_ϵ and $m_\epsilon \geqslant 9$ we get $8U/9\epsilon^{-u} \leqslant m_\epsilon \leqslant U\epsilon^{-u}$ and

$$2^{U/9\epsilon^{-u}} \leqslant 2^{m_\epsilon/8} \leqslant M_\epsilon \leqslant 2^{m_\epsilon/4} \leqslant 2^{U/3\epsilon^{-u}}.$$

We can simplify it as $2^{C_2\epsilon^{-u}} \leqslant M_\epsilon \leqslant 2^{3C_2\epsilon^{-u}}$ with $C_2 \doteq U/9$. We have,

$$\frac{1}{M_\epsilon} \sum_{j=1}^{M_\epsilon} KL(P_j^{\otimes^n}, P_0^{\otimes^n}) \leqslant \frac{n}{\sigma_0^2} 16 c_1^{\gamma_1 - 1 + \gamma} \epsilon m_\epsilon^{-\frac{\gamma_1 - 1 + \gamma}{p_X}}.$$

Furthermore, using $m_\epsilon \geqslant 8U/9\epsilon^{-u}$ we find

$$\frac{1}{M_\epsilon} \sum_{j=1}^{M_\epsilon} KL\left(P_j^n, P_0^n\right) \leqslant C' n \varepsilon^{1 + \frac{\gamma_1 - 1 + \gamma}{\beta_X - \gamma}} =: \alpha_*$$

with $C' \doteq \frac{16 c_1^{\gamma_1 - 1 + \gamma} 9^{\frac{\gamma_1 - 1 + \gamma}{p_X}}}{\sigma_0^2 (8U)^{\frac{\gamma_1 - 1 + \gamma}{p_X}}}$. For a measurable function $\Psi : (E_Z \times \mathbb{R})^n \to \{0, 1, \dots, M_\epsilon\}$, since $M_\epsilon \geqslant 2^{C_2\epsilon^{-u}}$, it yields

$$\begin{aligned} \max_{j=0,1,\dots,M_\epsilon} P_j^n(D : \Psi(D) \neq j) &\geqslant \frac{\sqrt{M_\epsilon}}{1 + \sqrt{M_\epsilon}} \left(1 - \frac{3C' n \epsilon^{1 + \frac{\gamma_1 - 1 + \gamma}{\beta_X - \gamma}}}{\log(M_\epsilon)} - \frac{1}{2 \log(M_\epsilon)} \right) \\ &\geqslant \frac{\sqrt{M_\epsilon}}{1 + \sqrt{M_\epsilon}} \left(1 - \frac{3C'}{C_2 \log(2)} n \epsilon^{1 + \frac{\gamma_1 - 1 + \gamma}{\beta_X - \gamma} + u} - \frac{1}{2 \log(M_\epsilon)} \right). \end{aligned}$$

Since $1 + \frac{\gamma_1 - 1 + \gamma}{\beta_X - \gamma} + u = \frac{\beta_X - 1 + \gamma_1 + p_X}{\beta_X - \gamma}$, we get

$$\max_{j=0,1,\dots,M_\epsilon} P_j^n(D : \Psi(D) \neq j) \geqslant \frac{\sqrt{M_\epsilon}}{1 + \sqrt{M_\epsilon}} \left(1 - C_1 n \epsilon^{\frac{\beta_X - 1 + \gamma_1 + p_X}{\beta_X - \gamma}} - \frac{1}{2 \log(M_\epsilon)} \right). \tag{33}$$

for $C_1 \doteq \frac{3C'}{C_2 \log(2)}$. To conclude the proof we follow the general reduction scheme from Tsybakov (2009, Section 2.2). Let $D \mapsto h_D$ be a (measurable) learning method for NPIR. Set $r \doteq \frac{\beta_X - \gamma}{\beta_X + \gamma_1 + p_X - 1}$, and fix $\tau > 0$ and $n \geqslant 1$ with $\epsilon_n \doteq \tau n^{-r} \leqslant \epsilon_0$. It remains to show that there is a distribution P which is difficult to learn for the considered learning method. For $\epsilon = \epsilon_n$, we take the previous possible candidates $P_0, P_1, \dots, P_{M_n}$, with $M_n \doteq M_{\epsilon_n}$. Next, we estimate the left hand side of the inequality in Eq. (33). To this end, we consider the measurable function $\Psi : (E_Z \times \mathbb{R})^n \to \{0, 1, \dots, M_n\}$ defined by

$$\Psi(D) \doteq \operatorname*{argmin}_{j=0,1,\dots,M_n} \|h_D - h_j\|_\gamma.$$

For $j \in \{0, 1, \dots, M_n\}$ and $D \in (E_Z \times \mathbb{R})^n$ with $\Psi(D) \neq j$ we have

$$2\sqrt{\epsilon_n} \leq \left\|h_{\Psi(D)} - h_j\right\|_\gamma \leqslant \left\|h_{\Psi(D)} - h_D\right\|_\gamma + \|h_D - h_j\|_\gamma \leqslant 2\|h_D - h_j\|_\gamma.$$

Consequently, for all $j = 0, 1, \dots, M_n$ we find

$$P_j^n(D : \Psi(D) \neq j) \leqslant P_j^n\left(D : \|h_D - h_j\|_\gamma^2 \geqslant \varepsilon_n\right).$$

Therefore, there is h_* in $\{h_0, \dots, h_{M_n}\}$ such that,

$$\begin{aligned} P^n\left(D : \|h_D - h_*\|_\gamma^2 \geqslant \varepsilon_n\right) &\geqslant \max_{j=0,1,\dots,M_n} P^n(D : \Psi(D) \neq j) \\ &\geqslant \frac{\sqrt{M_n}}{\sqrt{M_n} + 1} \left(1 - C_1 \tau^{1/r} - \frac{1}{2 \log(M_n)} \right). \end{aligned}$$

Since $M_n \to \infty$ for $n \to \infty$ we can choose n sufficiently large such that the right hand side is bounded from below by $1 - 2C_1\tau^{1/r}$.

G Some Bounds

Definition 9 (Spectral decomposition for C_F)**.** *By Jensen's inequality, $0 \preceq C_F \preceq C_X$ and therefore C_F,defined in Eq. (12), is a Trace class operator since $\mathrm{Tr}(\mathrm{C_F}) \leqslant \mathrm{Tr}(\mathrm{C_X}) < +\infty$. Therefore there exists an eigensystem $\{(\mu_{F,i}, \sqrt{\mu_{F,i}}e_{F,i})\}_{i\in I}$ with positive summable eigenvalues such that*

$$C_F = \sum_{i\in I} \mu_{F,i} \sqrt{\mu_{F,i}} e_{F,i} \otimes \sqrt{\mu_{F,i}} e_{F,i},$$

where $\{\sqrt{\mu_{F,i}}e_{F,i}\}_{i\in I}$ is an ONB of $\overline{\mathcal{R}(C_F)} \subseteq \mathcal{H}_X$.

Lemma 2. *Let C_F be defined as in Eq. (12) with eigensystem $\{(\mu_{F,i}, \sqrt{\mu_{F,i}}e_{F,i})\}_{i\in I}$ given in Definition 9. Let Assumption (LINK) hold. We then have, for all $i \in I$, $\mu_{F,i} \leqslant \mu_{X,i}^{\gamma_1}$.*

Proof. The proof follows from Lemma 17. □

Lemma 3. *Define the following quantity, for $\lambda > 0$,*

$$\mathcal{N}_F(\lambda) = \mathrm{Tr}\left(C_F\left(C_F + \lambda\,\mathrm{Id}_{\mathcal{H}_X}\right)^{-1}\right)$$

and let Assumption (EVDX) hold. There is a constant $D > 0$, independent of λ, such that the following inequality is satisfied,

$$\mathcal{N}_F(\lambda) \leqslant D\lambda^{-\frac{p_X}{\gamma_1}}.$$

Proof. From Lemma 2, and using Assumption (EVDX), we see that $\mu_{F,i} \leqslant \mu_{X,i}^{\gamma_1} \leqslant c_2 i^{-\frac{\gamma_1}{p_X}}$ for all $i \in I$. Applying Fischer and Steinwart (2020, Lemma 11) with p_X/γ_1 yields the final result. □

Lemma 4. *Let Assumption 1 hold. Then for all $\lambda > 0$ and $\theta \geqslant 1$, we have the following bound:*

$$\|C_F^{\frac{1}{2\theta}}(C_F + \lambda\,\mathrm{Id})^{-1/2}\|_{\mathcal{H}_X\to\mathcal{H}_X} \leqslant \lambda^{\frac{1}{2\theta}-\frac{1}{2}}.$$

Proof.

$$\begin{aligned}\|C_F^{\frac{1}{2\theta}}(C_F + \lambda\,\mathrm{Id})^{-1/2}\|_{\mathcal{H}_X\to\mathcal{H}_X} &= \sqrt{\sup_{i\in I} \frac{\lambda_{F,i}^{\frac{1}{\theta}}}{\lambda_{F,i} + \lambda}} \\ &\leqslant \lambda^{\frac{1}{2\theta}-\frac{1}{2}},\end{aligned}$$

where in the inequality we used Lemma 13 with $\theta^{-1} \leqslant 1$ since $\theta \geqslant 1$. □

Lemma 5. *Let Assumptions 1, (SRCX) and (LINK) hold with $0 \leqslant \beta_X \leqslant \gamma_0 + 1$. Then for all $\lambda > 0$ and $\theta \in [0,1]$, we have the following bound:*

$$\|C_F^{\frac{1-\theta}{2}}(C_F + \lambda\,\mathrm{Id})^{-1}h_*\|_{\mathcal{H}_X} \leqslant B_X\lambda^{\frac{\beta_X-1}{2\gamma_0}+\frac{1-\theta}{2}-1}.$$

Proof. Using (SRCX) and the fact that $h_* \in \overline{\mathcal{R}(C_F)}$, we have,

$$\begin{aligned}
\|C_F^{\frac{1-\theta}{2}}(C_F + \lambda \operatorname{Id})^{-1} h_*\|_{\mathcal{H}_X} &\leqslant B_X \|C_F^{\frac{1-\theta}{2}}(C_F + \lambda \operatorname{Id})^{-1} P_F C_X^{\frac{\beta_X - 1}{2}}\|_{\mathcal{H}_X \to \mathcal{H}_X} \\
&= B_X \|C_X^{\frac{\beta_X - 1}{2}} P_F (C_F + \lambda \operatorname{Id})^{-1} C_F^{\frac{1-\theta}{2}}\|_{\mathcal{H}_X \to \mathcal{H}_X} \\
&= B_X \sup_{h \in \mathcal{H}_X, \|h\|_{\mathcal{H}_X} = 1} \|C_X^{\frac{\beta_X - 1}{2}} P_F (C_F + \lambda \operatorname{Id})^{-1} C_F^{\frac{1-\theta}{2}} h\|_{\mathcal{H}_X} \\
&\leqslant B_X \sup_{h \in \mathcal{H}_X, \|h\|_{\mathcal{H}_X} = 1} \|C_F^{\frac{\beta_X - 1}{2\gamma_0}} P_F (C_F + \lambda \operatorname{Id})^{-1} C_F^{\frac{1-\theta}{2}} h\|_{\mathcal{H}_X} \\
&= B_X \sup_{i \in I} \frac{\lambda_{F,i}^{\frac{\beta_X - 1}{2\gamma_0} + \frac{1-\theta}{2}}}{\lambda_{F,i} + \lambda} \\
&\leqslant B_X \lambda^{\frac{\beta_X - 1}{2\gamma_0} + \frac{1-\theta}{2} - 1},
\end{aligned}$$

where in the second inequality, we used Assumption (LINK) combined with Lemma 14 for $\omega = \frac{\beta_X - 1}{\gamma_0} \leqslant 1$, and in the last inequality we used Lemma 13 with $\frac{\beta_X - 1}{2\gamma_0} + \frac{1-\theta}{2} \leqslant \frac{\beta_X - 1 + \gamma_0}{2\gamma_0} \leqslant 1$ since $\beta_X \leqslant \gamma_0 + 1$. □

Lemma 6. *Let g_λ be defined as in Eq. (17),*

$$g_\lambda \doteq \log\left(2e\mathcal{N}_F(\lambda) \frac{\|C_F\|_{\mathcal{H}_X \to \mathcal{H}_X} + \lambda}{\|C_F\|_{\mathcal{H}_X \to \mathcal{H}_X}}\right).$$

Then, for $\tau \geqslant 1, \lambda > 0$, and $n \geqslant 1$, the following operator norm bound is satisfied with P^n-probability over $1 - 2e^{-\tau}$,

$$\left\|(C_F + \lambda)^{-1/2}\left(C_F - \hat{C}_F\right)(C_F + \lambda)^{-1/2}\right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant \frac{4\kappa_X^2 \tau g_\lambda}{3n\lambda} + \sqrt{\frac{2\kappa_X^2 \tau g_\lambda}{n\lambda}}.$$

For $n \geqslant 8\tau g_\lambda \kappa_X^2 \lambda^{-1}$, with probability over $1 - 2e^{-\tau}$,

$$\left\|(C_F + \lambda)^{-1/2}\left(C_F - \hat{C}_F\right)(C_F + \lambda)^{-1/2}\right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant \frac{2}{3}.$$

Proof. The bound is obtained directly from Fischer and Steinwart (2020, Lemma 17) applied to C_F with $\alpha = 1$ and using that almost surely $\|F(Z)\|_{\mathcal{H}_X} \leqslant \mathbb{E}[\|\phi_X(X)\|_{\mathcal{H}_X} \mid Z] \leqslant \kappa_X$. For $n \geqslant 8\tau g_\lambda \kappa_X^2 \lambda^{-1}$, we obtain that with probability over $1 - 2e^{-\tau}$,

$$\begin{aligned}
\left\|(C_F + \lambda \operatorname{Id})^{-1/2}\left(C_F - \hat{C}_F\right)(C_F + \lambda \operatorname{Id})^{-1/2}\right\|_{\mathcal{H}_X \to \mathcal{H}_X} &\leqslant \frac{4\kappa_X^2 \tau g_\lambda}{3n\lambda} + \sqrt{\frac{2\kappa_X^2 \tau g_\lambda}{n\lambda}} \\
&\leqslant \frac{4}{3} \cdot \frac{1}{8} + \sqrt{2 \cdot \frac{1}{8}} = \frac{2}{3}.
\end{aligned}$$

□

Lemma 7. *Let Assumptions (SRCZ) and (EMBZ) hold with $\alpha_Z \leqslant \beta_Z$. Under the assumptions that $\|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)} \leqslant 1$, and $\|F_* - \hat{F}_\xi\|_{\alpha_Z} \leqslant 1$ with P^n-probability over $1 - 4e^{-\tau}$, it holds that*

$$\|\hat{C}_F - \hat{C}_{\hat{F}}\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant J\left(\sqrt{\frac{\tau}{n}} \|F_* - \hat{F}_\xi\|_{\alpha_Z} + \|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)}\right),$$

where J depends on $A_Z, B_Z, \alpha_Z, \beta_Z$ and κ_Z.

Proof. We start with the following decomposition, for any $z \in E_Z$,

$$\begin{aligned}
&F_*(z) \otimes F_*(z) - \hat{F}_\xi(z) \otimes \hat{F}_\xi(z) \\
&= F_*(z) \otimes F_*(z) - F_*(z) \otimes \hat{F}_\xi(z) + F_*(z) \otimes \hat{F}_\xi(z) - \hat{F}_\xi(z) \otimes \hat{F}_\xi(z) \\
&= F_*(z) \otimes \left(F_*(z) - \hat{F}_\xi(z)\right) + \left(F_*(z) - \hat{F}_\xi(z)\right) \otimes \hat{F}_\xi(z) \\
&= F_*(z) \otimes \left(F_*(z) - \hat{F}_\xi(z)\right) + \left(F_*(z) - \hat{F}_\xi(z)\right) \otimes F_*(z) - \left(F_*(z) - \hat{F}_\xi(z)\right) \otimes \left(F_*(z) - \hat{F}_\xi(z)\right). \qquad (34)
\end{aligned}$$

Therefore,

$$\hat{C}_F - \hat{C}_{\hat{F}} = \frac{1}{n}\sum_{i=1}^{n}\left(F_*(z_i) \otimes \left(F_*(z_i) - \hat{F}_\xi(z_i)\right) + \left(F_*(z_i) - \hat{F}_\xi(z_i)\right) \otimes F_*(z_i) - \left(F_*(z_i) - \hat{F}_\xi(z_i)\right) \otimes \left(F_*(z_i) - \hat{F}_\xi(z_i)\right)\right),$$

we then have,

$$\begin{aligned}
\left\|\hat{C}_F - \hat{C}_{\hat{F}}\right\| \quad \mathcal{H}_X \to \mathcal{H}_X &\leqslant \left\|\frac{1}{n}\sum_{i=1}^{n} F_*(z_i) \otimes \left(F_*(z_i) - \hat{F}_\xi(z_i)\right)\right\|_{\mathcal{H}_X \to \mathcal{H}_X} + \left\|\frac{1}{n}\sum_{i=1}^{n}\left(F_*(z_i) - \hat{F}_\xi(z_i)\right) \otimes F_*(z_i)\right\|_{\mathcal{H}_X \to \mathcal{H}_X} \\
&+ \left\|\frac{1}{n}\sum_{i=1}\left(F_*(z_i) - \hat{F}_\xi(z_i)\right) \otimes \left(F_*(z_i) - \hat{F}_\xi(z_i)\right)\right\|_{\mathcal{H}_X \to \mathcal{H}_X} \\
&\leqslant \frac{2}{n}\sum_{i=1}^{n}\left\|F_*(z_i)\right\|_{\mathcal{H}_X}\left\|F_*(z_i) - \hat{F}_\xi(z_i)\right\|_{\mathcal{H}_X} + \frac{1}{n}\sum_{i=1}^{n}\left\|F_*(z_i) - \hat{F}_\xi(z_i)\right\|_{\mathcal{H}_X}^2 \\
&\leqslant \frac{2\kappa_Z^{\frac{\beta_Z - \alpha_Z}{2}} A_Z B_Z}{n}\sum_{i=1}^{n}\left\|F_*(z_i) - \hat{F}_\xi(z_i)\right\|_{\mathcal{H}_X} + \frac{1}{n}\sum_{i=1}^{n}\left\|F_*(z_i) - \hat{F}_\xi(z_i)\right\|_{\mathcal{H}_X}^2 \\
&\leqslant \frac{2\kappa_Z^{\frac{\beta_Z - \alpha_Z}{2}} A_Z B_Z}{n}\sum_{i=1}^{n}\left(\left\|F_*(z_i) - \hat{F}_\xi(z_i)\right\|_{\mathcal{H}_X} - \mathbb{E}\left[\left\|F_*(z_i) - \hat{F}_\xi(z_i)\right\|_{\mathcal{H}_X}\right] + \mathbb{E}\left[\left\|F_*(z_i) - \hat{F}_\xi(z_i)\right\|_{\mathcal{H}_X}\right]\right) \\
&+ \frac{1}{n}\sum_{i=1}^{n}\left(\left\|F_*(z_i) - \hat{F}_\xi(z_i)\right\|_{\mathcal{H}_X}^2 - \|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)}^2 + \|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)}^2\right),
\end{aligned}$$

where we used Lemma 12 in the second inequality. To conclude, we will apply Hoeffding inequality (conditionally on $\mathcal{D}_1$) to $X_i \doteq \left\|F_*(z_i) - \hat{F}_\xi(z_i)\right\|_{\mathcal{H}_X}, i = 1, \ldots, n$. Note that by Jensen's inequality, for $i = 1, \ldots, n$, $\mathbb{E}[X_i] \leqslant \|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)}$, and by (EMBZ), we have,

$$X_i \leqslant A_Z \|F_* - \hat{F}_\xi\|_{\alpha_Z},$$

almost surely. Therefore, by Hoeffding inequality, with P^n probability over $1 - 2e^{-\tau}$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left\|F_*(z_i) - \hat{F}_\xi(z_i)\right\|_{\mathcal{H}_X} - \mathbb{E}\left[\left\|F_*(z_i) - \hat{F}_\xi(z_i)\right\|_{\mathcal{H}_X}\right]\right| \leqslant A_Z \|F_* - \hat{F}_\xi\|_{\alpha_Z}\sqrt{\frac{\tau}{n}}.$$

And a similar reasoning gives us that with P^n probability over $1 - 2e^{-\tau}$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left\|F_*(z_i) - \hat{F}_\xi(z_i)\right\|_{\mathcal{H}_X}^2 - \|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)}^2\right| \leqslant A_Z^2 \|F_* - \hat{F}_\xi\|_{\alpha_Z}^2\sqrt{\frac{\tau}{n}}.$$

Under the assumptions that $\|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)} \leqslant 1$ and $\|F_* - \hat{F}_\xi\|_{\alpha_Z} \leqslant 1$, we conclude with an union bound that with P^n probability over $1 - 4e^{-\tau}$,

$$\left\|\hat{C}_F - \hat{C}_{\hat{F}}\right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant J\left(\sqrt{\frac{\tau}{n}}\|F_* - \hat{F}_\xi\|_{\alpha_Z} + \|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)}\right),$$

where J depends on $A_Z, B_Z, \alpha_Z, \beta_Z$ and κ_Z. □

Lemma 8. *Let Assumptions* (SRCZ) *and* (EMBZ) *hold with $\alpha_Z \leqslant \beta_Z$. For any $\tau \geqslant 1$, $\lambda > 0$ and sufficiently large m and n such that Eq.* (18) *is satisfied, with probability over $1 - 6e^{-\tau}$, we have*

$$\left\| (C_F + \lambda \operatorname{Id})^{-1/2} \left(C_F - \hat{C}_{\hat{F}} \right) (C_F + \lambda \operatorname{Id})^{-1/2} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant \frac{5}{6}.$$

Proof. Let $A_\lambda \doteq (C_F + \lambda \operatorname{Id})^{-1/2} \left(C_F - \hat{C}_{\hat{F}} \right) (C_F + \lambda \operatorname{Id})^{-1/2}$. We have,

$$\begin{aligned} \|A_\lambda\|_{\mathcal{H}_X \to \mathcal{H}_X} &= \left\| (C_F + \lambda \operatorname{Id})^{-1/2} \left(C_F - \hat{C}_F + \hat{C}_F - \hat{C}_{\hat{F}} \right) (C_F + \lambda \operatorname{Id})^{-1/2} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \\ &\leqslant \left\| (C_F + \lambda \operatorname{Id})^{-1/2} \left(C_F - \hat{C}_F \right) (C_F + \lambda \operatorname{Id})^{-1/2} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} + \left\| (C_F + \lambda \operatorname{Id})^{-1/2} \left(\hat{C}_{\hat{F}} - \hat{C}_F \right) (C_F + \lambda \operatorname{Id})^{-1/2} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \\ &\leqslant \left\| (C_F + \lambda \operatorname{Id})^{-1/2} \left(C_F - \hat{C}_F \right) (C_F + \lambda \operatorname{Id})^{-1/2} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} + \lambda^{-1} \left\| \left(\hat{C}_{\hat{F}} - \hat{C}_F \right) \right\|_{\mathcal{H}_X \to \mathcal{H}_X}. \end{aligned}$$

Apply Lemma 6 to the first term, for $\tau \geqslant 1, \lambda > 0$ and $n \geqslant 8\tau g_\lambda \kappa_X^2 \lambda^{-1}$, with probability over $1 - 2e^{-\tau}$,

$$\left\| (C_F + \lambda \operatorname{Id})^{-1/2} \left(C_F - \hat{C}_F \right) (C_F + \lambda \operatorname{Id})^{-1/2} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant \frac{2}{3}.$$

For the second term, we apply Lemma 7, under the assumptions that $\|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)} \leqslant 1$ and $\|F_* - \hat{F}_\xi\|_{\alpha_Z} \leqslant 1$, with P^n-probability over $1 - 4e^{-\tau}$, it holds

$$\left\| \hat{C}_F - \hat{C}_{\hat{F}} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant J \left(\sqrt{\frac{\tau}{n}} \|F_* - \hat{F}_\xi\|_{\alpha_Z} + \|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)} \right).$$

Under the constraints of Eq. (18), it implies that with probability over $1 - 6e^{-\tau}$,

$$\|A_\lambda\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant \frac{5}{6},$$

□

Lemma 9. *Let Assumptions* (SRCZ) *and* (EMBZ) *hold with $\alpha_Z \leqslant \beta_Z$. For any $\tau \geqslant 1$, $\lambda > 0$ and sufficiently large m and n such that the constraints of Eq.* (18) *are satisfied, it holds with probability at least $1 - 4e^{-\tau}$ that*

$$\left\| \left(\hat{C}_{\hat{F}} + \lambda \operatorname{Id} \right)^{-1/2} \left(\hat{C}_F + \lambda \operatorname{Id} \right)^{1/2} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant \sqrt{\frac{6}{5}}.$$

Proof. By Lemma 16, we obtain that

$$\left\| \left(\hat{C}_{\hat{F}} + \lambda \operatorname{Id} \right)^{-1/2} \left(\hat{C}_F + \lambda \operatorname{Id} \right)^{1/2} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant (1 - t)^{-1/2},$$

where $t = \left\| \left(\hat{C}_F + \lambda \operatorname{Id} \right)^{-1/2} \left(\hat{C}_F - \hat{C}_{\hat{F}} \right) \left(\hat{C}_F + \lambda \operatorname{Id} \right)^{-1/2} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant \lambda^{-1} \left\| \hat{C}_F - \hat{C}_{\hat{F}} \right\|_{\mathcal{H}_X \to \mathcal{H}_X}$. By Lemma 7, under the assumptions that $\|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)} \leqslant 1$ and $\|F_* - \hat{F}_\xi\|_{\alpha_Z} \leqslant 1$, with P^n-probability over $1 - 4e^{-\tau}$, it holds

$$\left\| \hat{C}_F - \hat{C}_{\hat{F}} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant J \left(\sqrt{\frac{\tau}{n}} \|F_* - \hat{F}_\xi\|_{\alpha_Z} + \|F_* - \hat{F}_\xi\|_{L_2(Z;\mathcal{H}_X)} \right).$$

Under the constraints of Eq. (18), it implies that with probability over $1 - 4e^{-\tau}$, $t \leqslant \frac{1}{6}$, and therefore,

$$\left\| \left(\hat{C}_{\hat{F}} + \lambda \operatorname{Id} \right)^{-1/2} \left(\hat{C}_F + \lambda \operatorname{Id} \right)^{1/2} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant \sqrt{\frac{6}{5}}.$$

□

Lemma 10. *Let Assumptions (SRCZ) and (EMBZ) hold with $\alpha_Z \leqslant \beta_Z$. For any $\tau \geqslant 1$, $\lambda > 0$ and sufficiently large m and n such that the constraints of Eq. (18) are satisfied, it holds with probability at least $1 - 6e^{-\tau}$ that*

$$\left\| (C_F + \lambda \operatorname{Id})^{1/2} \left(\hat{C}_{\hat{F}} + \lambda \operatorname{Id} \right)^{-1/2} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant 3.$$

Proof. We use Lemma 16 to obtain

$$\left\| (C_F + \lambda \operatorname{Id})^{1/2} \left(\hat{C}_{\hat{F}} + \lambda \operatorname{Id} \right)^{-1/2} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant (1-t)^{-1/2},$$

when $t \doteq \|A_\lambda\|_{\mathcal{H}_X \to \mathcal{H}_X} < 1$, with $A_\lambda \doteq (C_F + \lambda \operatorname{Id})^{-1/2} \left(C_F - \hat{C}_{\hat{F}} \right) (C_F + \lambda \operatorname{Id})^{-1/2}$. By Lemma 8, under the constraints of Eq. (18), with probability over $1 - 6e^{-\tau}$, $t \leqslant 5/6$, and therefore,

$$\left\| (C_F + \lambda \operatorname{Id})^{1/2} \left(\hat{C}_{\hat{F}} + \lambda \operatorname{Id} \right)^{-1/2} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant \sqrt{6} \leqslant 3.$$

□

Lemma 11. *For any $\tau \geqslant 1$, $\lambda > 0$ and sufficiently large m and n such that the constraints of Eq. (18) are satisfied, it holds with probability at least $1 - 2e^{-\tau}$ that*

$$\left\| \left(\hat{C}_F + \lambda \operatorname{Id} \right)^{-1/2} (C_F + \lambda \operatorname{Id})^{1/2} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant 2.$$

Proof. By Lemma 16,

$$\left\| \left(\hat{C}_F + \lambda \operatorname{Id} \right)^{-1/2} (C_F + \lambda \operatorname{Id})^{1/2} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant (1-t)^{-1/2},$$

with $t = \left\| (C_F + \lambda \operatorname{Id})^{-1/2} \left(C_F - \hat{C}_F \right) (C_F + \lambda \operatorname{Id})^{-1/2} \right\|_{\mathcal{H}_X \to \mathcal{H}_X}$. By Lemma 6, for $\tau \geqslant 1, \lambda > 0$ and $n \geqslant 8\tau g_\lambda \kappa_X^2 \lambda^{-1}$, with probability over $1 - 2e^{-\tau}$,

$$\left\| (C_F + \lambda \operatorname{Id})^{-1/2} \left(C_F - \hat{C}_F \right) (C_F + \lambda \operatorname{Id})^{-1/2} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant \frac{2}{3},$$

and therefore,

$$\left\| \left(\hat{C}_F + \lambda \operatorname{Id} \right)^{-1/2} (C_F + \lambda \operatorname{Id})^{1/2} \right\|_{\mathcal{H}_X \to \mathcal{H}_X} \leqslant \sqrt{3} \leqslant 2.$$

□

H Auxiliary Results

The following theorem is adapted from Fischer and Steinwart (2020) and studies the generalization error of kernel ridge regression.

Theorem 14. *Consider a pair of random variables (Y, Z) defined on $\mathbb{R} \times E_Z$ with conditional mean function $f_* := \mathbb{E}[Y \mid Z]$ and marginal distribution π_Z for Z. Let H be a RKHS of functions from E_Z to $\mathbb{R}$ with kernel $k(z, z') = \langle k_z, k_{z'} \rangle_H$, where $k_z \in H$ denotes the canonical feature map. Given $\lambda > 0$ and $(z_i, y_i)_{i=1}^n$ independently sampled from the same distribution as (Y, Z), let $\bar{f}_\lambda$ be the kernel ridge estimator:*

$$\bar{f}_\lambda = \operatorname*{arg\,min}_{f \in H} \frac{1}{n} \sum_{i=1}^n (y_i - f(z_i))^2 + \lambda \|f\|_H^2.$$

Finally let $\Sigma = \mathbb{E}[k_Z \otimes k_Z]$ denotes the covariance associated to (H, π_Z), and $\mathcal{N}_\Sigma(\lambda) = \operatorname{Tr}((\Sigma + \lambda \operatorname{Id}_H)^{-1} \Sigma)$, $\lambda > 0$ denotes the effective dimension. Assume the following:

1. *For π_Z−almost all $z \in E_Z$, $k(z,z) \leqslant \kappa$;*
2. *There exist $\sigma, L > 0$ such that for all $m \geqslant 2$, $\mathbb{E}[(Y - f_*(Z))^m \mid Z] \leqslant \frac{1}{2} m! \sigma^2 L^{m-2}$, π_Z−almost surely;*
3. *There exist $p \in (0,1]$ and a constant $D > 0$ such that $\mathcal{N}_\Sigma(\lambda) \leqslant D\lambda^{-p}$;*
4. *There exists $\beta \in [1,2]$ such that $\|\Sigma^{-\frac{\beta-1}{2}} f_*\|_H \leqslant B$.*

Then for the abbreviations

$$\begin{aligned} g_\lambda &\doteq \log\left(2e\mathcal{N}_\Sigma(\lambda)\frac{\|\Sigma\|_{H\to H} + \lambda}{\|\Sigma\|_{H\to H}}\right) \\ A_{\lambda,\tau} &\doteq 8\tau g_\lambda \kappa^2 \lambda^{-1}, \end{aligned} \tag{35}$$

and $0 \leqslant \theta \leqslant 1, \tau \geqslant 1, 0 < \lambda \leqslant 1$, and $n \geqslant A_{\lambda,\tau}$, the following bound is satisfied with P^n-probability not less than $1 - 4e^{-\tau}$,

$$\|\Sigma^{\frac{1-\theta}{2}}\left(\bar{f}_\lambda - f_*\right)\|_H^2 \leqslant J\left(\lambda^{\beta-\theta} + \frac{\tau^2}{n\lambda^{\theta+p}}\left(1 + \frac{1}{n\lambda^{1-p}}\right)\right),$$

where J is a constant depending on B, σ, L, D, κ and $\|f_*\|_H$.

Proof. Let $\|\cdot\|_\theta$ be the interpolation norm introduced in Fischer and Steinwart (2020, Section 2). By Fischer and Steinwart (2020, Lemma 12) and the definition of the interpolation norm, for any element $f \in H$ and $\theta \geqslant 0$, we have $\|f\|_\theta = \|\Sigma^{\frac{1-\theta}{2}} f\|_H$. Therefore, assumption 4. with $\beta \geqslant 1$ is equivalent to assumption (SRC) in Fischer and Steinwart (2020) (i.e. $\|f_*\|_\beta \leqslant B$ in their notation). As k is almost surely bounded we take $\alpha = 1$ and $\|k^\alpha\|_\infty = \kappa$ (see Fischer and Steinwart (2020, Eq. (16))) in assumption (EMB) in Fischer and Steinwart (2020).

Combining an approximation-estimation error decomposition (Fischer and Steinwart, 2020, Eq. (16)), Lemma 14 (for the approximation error) and Theorem 16 (for the estimation error) from Fischer and Steinwart (2020), we obtain that for $0 \leqslant \theta \leqslant 1, \tau \geqslant 1$, and $n \geqslant A_{\lambda,\tau}$, the following bound is satisfied with P^n-probability not less than $1 - 4e^{-\tau}$,

$$\|\Sigma^{\frac{1-\theta}{2}}\left(\bar{f}_\lambda - f_*\right)\|_H^2 \leqslant B^2\lambda^{\beta-\theta} + \frac{576\tau^2}{n\lambda^\theta}\left(\sigma^2\mathcal{N}_\Sigma(\lambda) + \kappa^2\frac{\|f_* - f_\lambda\|_{L_2(Z)}^2}{\lambda} + 2\kappa^2\frac{\max\{L^2, \|f_* - f_\lambda\|_{L_\infty(Z)}^2\}}{n\lambda}\right),$$

where $f_\lambda \doteq \arg\min_{f\in H} \mathbb{E}\left[(Y - f(Z))^2\right] + \lambda\|f\|_H^2$. Firstly, by Fischer and Steinwart (2020, Lemma 14),

$$\|f_* - f_\lambda\|_{L_2(Z)}^2 \leqslant \|f_*\|_\beta^2 \lambda^\beta \leqslant B\lambda^\beta.$$

Secondly, the expression of f_λ can be simplified as follows, using $f_* \in H$ and the reproducing property,

$$\begin{aligned} f_\lambda &= \arg\min_{f\in H} \mathbb{E}\left[(f_*(Z) - f(Z))\right]^2 + \lambda\|f\|_H^2 \\ &= \arg\min_{f\in H} \mathbb{E}\left[\langle f_* - f, k_Z\rangle_H^2\right] + \lambda\|f\|_H^2 \\ &= \arg\min_{f\in H} \|\Sigma^{1/2}(f_* - f)\|_H^2 + \lambda\|f\|_H^2 \\ &= \arg\min_{f\in H} \langle(\Sigma + \lambda\,\mathrm{Id}_H)f, f\rangle_H - 2\langle\Sigma f_*, f\rangle_H \\ &= (\Sigma + \lambda\,\mathrm{Id}_{\mathcal{H}})^{-1}\Sigma f_*. \end{aligned}$$

Therefore, $f_* - f_\lambda = \lambda(\Sigma + \lambda\,\mathrm{Id}_H)^{-1} f_*$, and we obtain the following upper bound,

$$\|f_* - f_\lambda\|_{L_\infty(Z)} \leqslant \kappa\,\|f_* - f_\lambda\|_H = \lambda\kappa\left\|(\Sigma + \lambda\,\mathrm{Id}_H)^{-1} f_*\right\|_H \leqslant \kappa\,\|f_*\|_H.$$

Thirdly, by assumption 3., $\mathcal{N}_\Sigma(\lambda) \leqslant D\lambda^{-p}$. We finally obtain that,

$$\begin{aligned}\|\Sigma^{\frac{1-\theta}{2}}\left(\bar{f}_\lambda - f_*\right)\|_H^2 &\leqslant B^2\lambda^{\beta-\theta} + \frac{576\tau^2}{n\lambda^\theta}\left(\sigma^2 D\lambda^{-p} + \kappa^2 B^2\lambda^{\beta-1} + 2\kappa^2\frac{\max\{L^2,\kappa^2\|f_*\|_H^2\}}{n\lambda}\right)\\ &\leqslant B^2\lambda^{\beta-\theta} + \frac{576\tau^2}{n\lambda^{\theta+p}}\left(\sigma^2 D + \kappa^2 B^2\lambda^{\beta-1+p} + 2\kappa^2\frac{\max\{L^2,\kappa^2\|f_*\|_H^2\}}{n\lambda^{1-p}}\right)\\ &\leqslant B^2\lambda^{\beta-\theta} + \frac{576\tau^2}{n\lambda^{\theta+p}}\left(\sigma^2 D + \kappa^2 B^2 + 2\kappa^2\frac{\max\{L^2,\kappa^2\|f_*\|_H^2\}}{n\lambda^{1-p}}\right)\end{aligned}$$

where in the last step, we use $\lambda \leqslant 1$ with $\beta \geqslant 1$. □

Proposition 5. *Let $h_*, h_\lambda, \bar{h}_\lambda$, be defined in Eq. (8), Eq. (10) and Eq. (11) respectively. Then $h_*, h_\lambda \in \overline{\mathcal{R}(C_F)}$ and almost surely $\bar{h}_\lambda \in \overline{\mathcal{R}(C_F)}$.*

Proof. It follows from Lemma 15 and Corollary 2. □

Lemma 12. *Let $F \in \mathcal{G}$, then, for π_Z-almost all $z \in E_Z$, $\|F(z)\|_{\mathcal{H}_X} \leqslant \kappa_Z\|F\|_{\mathcal{G}}$.*

Alternatively, if (EMBZ) holds and F satisfies (SRCZ) with $\alpha_Z \leqslant \beta_Z$, then for π_Z-almost all $z \in E_Z$, $\|F(z)\|_{\mathcal{H}_X} \leqslant A_Z\|F\|_{\alpha_Z} \leqslant \kappa_Z^{\frac{\beta_Z-\alpha_Z}{2}} A_Z B_Z$.

Proof. By Theorem 1, since $F \in \mathcal{G}$, there is an operator $C \in S_2(\mathcal{H}_Z, \mathcal{H}_X)$ such that for all $z \in E_Z$, $F(z) = C\phi_Z(z)$ and $\|F\|_{\mathcal{G}} = \|C\|_{S_2}$. Therefore, for π_Z-almost all $z \in E_Z$

$$\|F(z)\|_{\mathcal{H}_X} = \|C\phi_Z(z)\|_{\mathcal{H}_X} \leqslant \kappa_Z\|C\|_{\mathcal{H}_Z\to\mathcal{H}_X} \leqslant \kappa_Z\|C\|_{S_2} = \kappa_Z\|F\|_{\mathcal{G}},$$

where we used Assumption 1: $k_Z(z,z) \leqslant \kappa_Z^2$ for π_Z-almost all $z \in E_Z$.

Under (EMBZ), it is shown in Lemma 4 Li et al. (2022a) that for all functions $F : E_Z \to \mathcal{H}_X$ such that $\|F\|_{\alpha_Z} < +\infty$,

$$\|F\|_{L_\infty(Z;\mathcal{H}_X)} \leqslant A_Z\|F\|_{\alpha_Z}.$$

To conclude we show that since F satisfies (SRCZ) with $\alpha_Z \leqslant \beta_Z$ then $\|F\|_{\alpha_Z} \leqslant \|F\|_{\beta_Z}$. Indeed, since $F \in L_2(Z;\mathcal{H}_X)$, by Remark 1, there is an operator $C \in S_2(\overline{\mathcal{R}(L_Z)}, \mathcal{H}_X)$ such that $F = \Psi(C)$ and by Eq. (14), for any $\theta \geqslant 0$,

$$\|F\|_\theta = \|CL_Z^{-\theta/2}\|_{S_2(L_2(Z),\mathcal{H}_X)}.$$

Since $C \in S_2(\overline{\mathcal{R}(L_Z)}, \mathcal{H}_X)$, exploiting the spectral decomposition of L_Z (see Eq. (3)) and using the fact that $\{\sqrt{\mu_{X,i}}e_{X,i} \otimes [e_{Z,j}]\}_{i\in I_X, j\in I_Z}$ is an ONB of $S_2(\overline{\mathcal{R}(L_Z)}, \mathcal{H}_X)$ (see Definition 5), we have

$$\begin{aligned}\|F\|_{\alpha_Z}^2 &= \sum_{i\in I_X}\sum_{j\in I_Z}\mu_{Z,i}^{-\alpha_Z}\langle C, \sqrt{\mu_{X,i}}e_{X,i} \otimes [e_{Z,j}]\rangle_{S_2}^2\\ &= \frac{1}{\kappa_Z^{\alpha_Z}}\sum_{i\in I_X}\sum_{j\in I_Z}\left(\frac{\kappa_Z}{\mu_{Z,i}}\right)^{\alpha_Z}\langle C, \sqrt{\mu_{X,i}}e_{X,i} \otimes [e_{Z,j}]\rangle_{S_2}^2\\ &\leqslant \frac{1}{\kappa_Z^{\alpha_Z}}\sum_{i\in I_X}\sum_{j\in I_Z}\left(\frac{\kappa_Z}{\mu_{Z,i}}\right)^{\beta_Z}\langle C, \sqrt{\mu_{X,i}}e_{X,i} \otimes [e_{Z,j}]\rangle_{S_2}^2\\ &= \kappa_Z^{\beta_Z-\alpha_Z}\|F\|_{\beta_Z}^2.\end{aligned}$$

□

The following theorem provides convergence guarantees for learning the conditional mean embedding F_*.

Theorem 15 (Theorem 4 Meunier et al. (2024)). *Let g_ξ be a filter function with qualification $\rho \geqslant 1$ used to build the estimator $\hat{F}_\xi$ on $\mathcal{D}_1$ with Eq. (7). Let Assumptions 1, (EVDZ) and (EMBZ) hold with $0 < p_Z \leqslant \alpha_Z \leqslant 1$. With $0 \leqslant \gamma \leqslant 1$, if (SRCZ) is satisfied with $\max\{\gamma, \alpha\} < \beta_Z \leqslant 2\rho$, we have, taking $\xi_m = \Theta\left(m^{-\frac{1}{\beta_Z + p_Z}}\right)$, that there is a constant $J > 0$ independent of $m \geqslant 1$ and $\tau \geqslant 1$ such that*

$$\left\|\hat{F}_\xi - F_*\right\|_\gamma^2 \leqslant \tau^2 J m^{-\frac{\beta_Z - \gamma}{\beta_Z + p_Z}}$$

is satisfied for sufficiently large $m \geqslant 1$ with P^m-probability not less than $1 - 4e^{-\tau}$. In particular, by Assumption (EMBZ),

$$\left\|\hat{F}_\xi - F_*\right\|_{L_\infty(Z;\mathcal{H}_X)}^2 \leqslant A_Z^2 \left\|\hat{F}_\xi - F_*\right\|_{\alpha_Z}^2 \leqslant \tau^2 A_Z J m^{-\frac{\beta_Z - \alpha_Z}{\beta_Z + p_Z}}.$$

Lemma 13 (Lemma 25 Fischer and Steinwart (2020)). *For $\lambda > 0$ and $0 \leqslant \alpha \leqslant 1$, let the function $f_{\lambda,\alpha} : [0, \infty) \to \mathbb{R}$ be defined by $f_{\lambda,\alpha}(t) \doteq t^\alpha/(\lambda + t)$. Then,*

$$\sup_{t \geqslant 0} f_{\lambda,\alpha}(t) \leqslant \lambda^{\alpha-1}.$$

In the remainder of this section, **we fix H a separable Hilbert space**.

Lemma 14 (Löwner-Heniz theorem Heinz (1951)). *$x \mapsto x^\omega$ is operator monotone for $\omega \in [0,1]$. Therefore for any two bounded positive semidefinite operators A, B acting on H, if $A \leqslant B$, then for any $\omega \in [0,1]$, $A^\omega \leqslant B^\omega$.*

Proposition 6. *Let A, B be two bounded, self-adjoint operators acting on H.*

1. *If there is a constant $c < +\infty$, such that $\|Ax\|_H \leqslant c\|Bx\|_H$ for all $x \in H$, then $\mathcal{R}(A) \subseteq \mathcal{R}(B)$ and $\|B^\dagger A\|_{H \to H} \leqslant c$;*
2. *If $\mathcal{R}(A) \subseteq \mathcal{R}(B)$, then $B^\dagger A$ is a well-defined bounded operator on H and $\|Ax\|_H \leqslant c\|Bx\|_H$ for all $x \in H$ with $c = \|B^\dagger A\|_{H \to H}$.*

For details on the pseudo-inverse $B^\dagger$, see Engl et al. (2000).

Proof.

1. Consider the operator S_0 defined on $\mathcal{R}(B)$ by $S_0(Bx) = Ax$. The operator S_0 is well-defined since by assumption the condition $Bx = 0$ implies $Ax = 0$. Therefore, if $x, x' \in H$ are such that $Bx = Bx'$, then $Ax = Ax'$. Moreover, using the assumption again, S_0 is bounded and $\|S_0\|_{H \to H} \leqslant c$. Hence S_0 extends uniquely to a bounded operator $S_1 : \overline{\mathcal{R}(B)} \to \overline{\mathcal{R}(A)}$ with $\|S_1\|_{H \to H} = \|S_0\|_{H \to H}$. Let S be the operator defined by $Sy = S_1 y$ if $y \in \overline{\mathcal{R}(B)}$ and $Sy = 0$ if $y \in \overline{\mathcal{R}(B)}^\perp$. Then S is a bounded operator satisfying $SB = S_0 B = A$. Hence $A = BS^*$. Thus
$$\mathcal{R}(A) \subseteq \mathcal{R}(B).$$
We conclude by observing that $S^* = B^\dagger A$ and using $\|S^*\|_{H \to H} = \|S\|_{H \to H} \leqslant \|S_1\|_{H \to H} \leqslant c$.
2. Under the assumption that $\mathcal{R}(A) \subseteq \mathcal{R}(B)$, $Q \doteq B^\dagger A$ is well-defined, bounded and such that $A = BQ$ (Theorem A.1 Klebanov et al., 2021). Therefore $A = Q^* B$ which implies that for all $x \in H$, $\|Ax\|_H \leqslant \|Q^*\|_H \|Bx\|_H = \|Q\|_H \|Bx\|_H$.

□

Lemma 15. *Let X be a a random variable taking values in H and admitting a Trace class covariance operator $C = \mathbb{E}[X \otimes X]$. Then, for all $f \in H$,*

$$f \in \mathcal{N}(C) \iff \langle f, X \rangle_H = 0 \quad \textit{almost surely.}$$

Proof. Assume $f \in \mathcal{N}(C)$, then

$$0 = \langle f, Cf \rangle_H = \mathbb{E}[\langle f, X \rangle_H^2],$$

and therefore $\langle f, X \rangle_H = 0$ a.e. The reverse direction is treated similarly, assume that $\langle f, X \rangle_H = 0$ a.e., then re-using

$$0 = \mathbb{E}[\langle f, X \rangle_H^2] = \langle f, Cf \rangle_H = \|C^{1/2} f\|_H^2$$

it implies $f \in \mathcal{N}(C^{1/2}) \subseteq \mathcal{N}(C)$, which concludes the proof. □

Corollary 2. *Let X be a a random variable taking values in H and admitting a Trace class covariance operator $C = \mathbb{E}[X \otimes X]$. Let $X_1, \ldots, X_n$ be i.i.d random variables with the same law as X. Let $\hat{C}_n \doteq \frac{1}{n} \sum_{i=1}^n X_i \otimes X_i$ be the empirical covariance operator. Then,*

$$\mathbb{P}^n \left(\overline{\mathcal{R}\left(\hat{C}_n\right)} \subseteq \overline{\mathcal{R}\left(C\right)} \right) = 1.$$

Proof. It is equivalent to show that almost surely, $\mathcal{N}(C) \subseteq \mathcal{N}(\hat{C}_n)$. Let us take $f \in \mathcal{N}(C)$. We have,

$$\langle f, \hat{C}_n f \rangle_H = \frac{1}{n} \sum_{i=1}^n \langle f, X_i \rangle_H^2 = 0,$$

almost surely, where we used Lemma 15. This implies that $f \in \mathcal{N}(\hat{C}_n^{1/2}) \subseteq \mathcal{N}(\hat{C}_n)$, which concludes the proof. □

Theorem 16 (Theorem 26 Fischer and Steinwart (2020) - Bernstein's Inequality)**.** *Let $(\Omega, \mathcal{B}, P)$ be a probability space and $\xi : \Omega \to H$ be a random variable with*

$$\mathbb{E}_P \|\xi\|_H^m \leqslant \frac{1}{2} m! \tilde{\sigma}^2 \tilde{L}^{m-2}$$

for all $m \geqslant 2$. Then, for $\tau \geqslant 1$ and $n \geqslant 1$, the following concentration inequality is satisfied

$$P^n \left((\omega_1, \ldots, \omega_n) \in \Omega^n : \left\| \frac{1}{n} \sum_{i=1}^n \xi\left(\omega_i\right) - \mathbb{E}_P \xi \right\|_H^2 \geqslant 32 \frac{\tau^2}{n} \left(\tilde{\sigma}^2 + \frac{\tilde{L}^2}{n} \right) \right) \leqslant 2e^{-\tau}$$

In particular, the above condition is satisfied if the following two bounds hold,

$$\|\xi\|_{\mathcal{H}} \leqslant \tilde{L} \text{ a.s.}$$
$$\mathbb{E}\left[\|\xi\|_{\mathcal{H}}^2\right] \leqslant \tilde{\sigma}^2.$$

Lemma 16 (Proposition 7 Rudi et al. (2015))**.** *Let A, B be two bounded positive semidefinite operators acting on H and $\lambda > 0$. Then,*

$$\left\|(A + \lambda \operatorname{Id}_H)^{-1/2} B^{1/2}\right\|_{\mathcal{H} \to \mathcal{H}} \leqslant \left\|(A + \lambda \operatorname{Id}_H)^{-1/2} (B + \lambda \operatorname{Id}_H)^{1/2}\right\|_{\mathcal{H} \to \mathcal{H}} \leqslant (1 - \beta)^{-1/2},$$

when

$$\beta = \left\|(B + \lambda \operatorname{Id}_H)^{-1/2} (B - A) (B + \lambda \operatorname{Id}_H)^{-1/2}\right\|_{\mathcal{H} \to \mathcal{H}} < 1.$$

Lemma 17. *Let A, B be compact, self-adjoint operators acting on H, whose positive eigenvalues are listed in decreasing order:*

$$\lambda_1(A) \geqslant \lambda_2(A) \geqslant \cdots > 0$$
$$\lambda_1(B) \geqslant \lambda_2(B) \geqslant \cdots > 0.$$

Then if $A \leqslant B$, for $i \geqslant 1$, $\lambda_i(A) \leqslant \lambda_i(B)$.

Proof. By the Min-Max Theorem, for all $i \geqslant 1$,

$$\lambda_i(B) = \min_{S_{i-1}} \max_{x \in S_{i-1}^{\perp}, \|x\|=1} \langle Bx, x \rangle_H \geqslant \min_{S_{i-1}} \max_{x \in S_{i-1}^{\perp}, \|x\|=1} \langle Ax, x \rangle_H = \lambda_i(A),$$

where the minimum is taken over all subspaces $S_{i-1} \subset H$ of dimension $n-1$. □

Proposition 7. *Let A, B be two compact self-adjoint positive semi-definite operators acting on an Hilbert space H and let P be the orthogonal projection on $\overline{\mathcal{R}(B)}$. If $PAP \leqslant B$, then for all $\delta > 0$,*

$$P\left(B + \delta I_H\right)^{-1} P \leqslant P\left(A + \delta I_H\right)^{-1} P.$$

Furthermore, if $f \in \overline{\mathcal{R}(B)}$ and $f \in \mathcal{R}(A^{1/2})$, we have

$$\langle f, B^{\dagger} f \rangle_H \leqslant \langle f, A^{\dagger} f \rangle_H.$$

Proof. For any $t, \alpha > 0$ define $C_{t,\alpha} \doteq B + tP + \alpha P_{\perp}$. Then if $t(\alpha - \|A\|) \geqslant \|A\|^2$, we have $A \leqslant C_{t,\alpha}$. Indeed, for all $f \in H$,

$$\begin{aligned}
\langle f, (C_{t,\alpha} - A) f \rangle_H &= \langle Pf + P_{\perp} f, (C_{t,\alpha} - A)(Pf + P_{\perp} f) \rangle_H \\
&= \langle f, Bf \rangle_H + t\langle Pf, Pf \rangle_{\mathcal{H}} - \langle f, PAPf \rangle_{\mathcal{H}} - 2\langle Pf, AP_{\perp} f \rangle_H + \alpha \langle P_{\perp} f, P_{\perp} f \rangle_H - \langle P_{\perp} f, AP_{\perp} f \rangle_H \\
&\geq t\|Pf\|_H^2 - 2\|A\| \|Pf\|_H \|P_{\perp} f\|_H + (\alpha - \|A\|)\|P_{\perp} f\|_H^2 \\
&= t\|Pf\|_H^2 - 2\|A\| \|Pf\|_H \|P_{\perp} f\|_H + \frac{\|A\|^2}{t}\|P_{\perp} f\|_H^2 - \frac{\|A\|^2}{t}\|P_{\perp} f\|_H^2 + (\alpha - \|A\|)\|P_{\perp} f\|_H^2 \\
&= \left(\sqrt{t}\|Pf\|_H - \frac{\|A\|}{\sqrt{t}}\|P_{\perp} f\|_H\right)^2 - \frac{\|A\|^2}{t}\|P_{\perp} f\|_H^2 + (\alpha - \|A\|)\|P_{\perp} f\|_H^2 \\
&\geqslant 0.
\end{aligned}$$

where the last inequality follows from $t(\alpha - \|A\|) \geqslant \|A\|^2$. Since B is compact self-adjoint positive semi-definite, it admits a decomposition

$$B = \sum_{i \geqslant 1} \omega_i b_i \otimes b_i,$$

where for all $i \geqslant 1$, (ω_i, b_i) are pairs of eigenvalues and eigenvectors of B such that $\omega_i > 0$ and $\{b_i\}_{i \geqslant 1}$ forms a orthonormal basis of $\overline{\mathcal{R}(B)}$. Therefore, on one hand,

$$P(B + tP + \delta I_{\mathcal{H}})^{-1} P = P\left(\sum_{i \geqslant 1} \frac{1}{\delta + t + \omega_i} b_i \otimes b_i + \frac{1}{\delta} P_{\perp}\right) P = \sum_{i \geqslant 1} \frac{1}{\delta + t + \omega_i} b_i \otimes b_i,$$

and on the other hand,

$$P(C_{t,\alpha} + \delta I_{\mathcal{H}})^{-1} P = P\left(\sum_{i \geqslant 1} \frac{1}{\delta + t + \omega_i} b_i \otimes b_i + \frac{1}{\delta + \alpha} P_{\perp}\right) P = \sum_{i \geqslant 1} \frac{1}{\delta + t + \omega_i} b_i \otimes b_i.$$

It follows that, for $t(\alpha - \|A\|) \geqslant \|A\|^2$,

$$P(B + tP + \delta I_{\mathcal{H}})^{-1} P = P(C_{t,\alpha} + \delta I_{\mathcal{H}})^{-1} P \leq P(A + \delta I_{\mathcal{H}})^{-1} P.$$

Let $t \to 0^+$, the result follows: $P(B + \delta I_{\mathcal{H}})^{-1} P \leq P(A + \delta I_{\mathcal{H}})^{-1} P$.

For the second part, let us consider $f \in \overline{\mathcal{R}(B)}$. Then $Pf = f$ and

$$\langle f, (B + \delta I_{\mathcal{H}})^{-1} f \rangle_H \leqslant \langle f, (A + \delta I_{\mathcal{H}})^{-1} f \rangle_H,$$

by the first part of the proposition. Under the assumption that $f \in \mathcal{R}(A^{1/2})$, $\|(A^{1/2})^{\dagger} f\|_H < +\infty$ and taking the limit with $\delta \to 0^+$ gives the final result. □