Sobolev Norm Learning Rates for Conditional Mean Embeddings

Prem Talwai
ORC, MIT

Ali Shameli
IDSS, MIT

David Simchi-Levi
IDSS, MIT

Abstract

We develop novel learning rates for conditional mean embeddings by applying the theory of interpolation for reproducing kernel Hilbert spaces (RKHS). We derive explicit, adaptive convergence rates for the sample estimator under the misspecifed setting, where the target operator is not Hilbert-Schmidt or bounded with respect to the input/output RKHSs. We demonstrate that in certain parameter regimes, we can achieve uniform convergence rates in the output RKHS. We hope our analyses will allow the much broader application of conditional mean embeddings to more complex ML/RL settings involving infinite dimensional RKHSs and continuous state spaces.

1 INTRODUCTION

In the past decade, several studies have explored a new framework for embedding conditional distributions in reproducing kernel Hilbert spaces (RKHS). This approach seeks to represent a conditional distribution as an RKHS element, and thereby reduce the computation of conditional expectations to the evaluation of kernel inner products. Unlike other distribution learning approaches, which often involve density estimation and expensive numerical analysis, the conditional mean embedding (CME) framework exploits the popular kernel trick to allow distributions to be learned directly and efficiently from sample information, and do not require the target distribution to possess a density function. The broad generalizability and computational levity of conditional embeddings have led them to find many applications in reinforcement learning, hypothesis testing, and nonparametric inference (Fukumizu et al., 2007, 2009; Grünewälder et al., 2012b; Song et al., 2010), where conditional relationships are often of pertinent interest.

A central issue involved in the conditional embedding framework is the performance of the sample estimator. Despite their successful application, there has been a limited study of optimal learning rates for conditional mean embeddings. Several foundational works (Song et al., 2010, 2009) established the consistency of the sample embedding estimator, exploring its convergence rate to a "true" embedding in the RKHS norm. These works framed the act of conditioning as a linear operator between two Hilbert spaces, which mapped features of the independent variable in the input space to the mean embeddings of their respective conditional distributions in the output feature space. Under certain smoothness conditions on the underlying distribution, Song et al. (2010) demonstrated convergence of the sample estimator in the Hilbert-Schmidt norm. Although these works introduced a regularization parameter to tackle the ill-conditioning of the sample covariance operator, the learning task was not explicitly framed as a regularized regression problem, with the consistency of the sample estimator only implicitly depending on the polynomial decay of the regularizer. Later, Grünewälder et al. (2012a) explicitly formulated conditional embeddings as the solution of a vector-valued Tikhonov-regularized regression problem. Here, the learning target was framed as a Hilbert space-valued function acting directly on the independent variable. Drawing from the rich theory of regularized regression (Caponnetto and De Vito, 2007), they derived near-optimal learning rates for kernels whose spectrum exhibits polynomial decay. However, their analysis required the compactness of the input set and the target Hilbert space to be finite-dimensional, an assumption which is violated by several common kernels.

In recent years, there have been several attempts to further relax the hypotheses of the previous two approaches — namely the requirement of a

Proceedings of the 25^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

finite-dimensional output RKHS and the compactness of the true conditional mean operator (well-specification). These approaches have sought a measure-theoretic interpretation of conditional mean embeddings as Hilbert-valued Bochner-measurable random variables in either an operator or vector-valued RKHS. Klebanov et al. (2020) demonstrates almost sure consistency for centered operators under relatively weak assumptions, but only L^2 consistency in the more popular uncentered framework, providing no insight into learning rates in either case. Park and Muandet (2020) abandons the operator framework, and seeks to directly extend the vector-valued regression approach from Grünewälder et al. (2012a) to infinite-dimensional RKHS, but similarly only demonstrates consistency in the general setting, and must further assume the well-specified setting to provide an concrete learning rate for the surrogate risk. Moreover, the latter approach sacrifices the operator interpretation of the conditional embedding, which has recently found an elegant connection to transfer operators in dynamical systems theory, and their associated data-driven spectral techniques (Mollenhauer et al., 2020; Klus et al., 2018).

In this paper, we aim to address these gaps by deriving novel adaptive learning rates for conditional mean embeddings in the *misspecified* setting, that elucidate the relationship between the properties of the kernel class and target measure. In particular, we seek to capture the interplay between kernel complexity (as measured by eigenvalue decay/summability) and the continuity of the hypothesis class in establishing uniform convergence rates. We apply the theory of interpolation spaces for RKHS (Fischer and Steinwart, 2020; Steinwart and Scovel, 2012) to significantly relax the aforementioned source conditions, and simply require that the target "conditioning function" lie in some intermediate fractional space between the input RKHS and L^2. To the best of our knowledge, this is the first work to establish uniform convergence rates in the misspecified setting. Our approach is also distinct from existing operator-based methods in that we do not require the the target conditional mean operator to be Hilbert-Schmidt, but simply bounded on the aforementioned interpolation space. Our generalized notion of boundedness also significantly attenuates the need to explicitly verify this continuity condition, which can often be difficult and unintuitive, and was a key motivator of the regression approach adapted by Grünewälder et al. (2012a); Park and Muandet (2020). Moreover, our analysis does not make any assumptions on the dimensionality of the input/output RKHS or the compactness of the latent spaces. These relaxations do introduce a slight tradeoff of requiring the polynomial eigendecay of the covariance operator, a standard assumption in regularized least-squares problems (Lin et al., 2020; Lin and Cevher, 2020; Caponnetto and De Vito, 2007). In a sense, our approach hybridizes the two aforementioned frameworks — namely, like Song et al. (2010) we construct conditional embeddings as operators, and characterize the convergence of the sample estimator via the spectral structure of the target embedding operator. However, we seek inspiration from the regression formulation of Grünewälder et al. (2012a) and likewise try to extrapolate approaches from scalar-valued kernel regression to our operator learning problem.

2 MODEL AND PRELIMINARIES

2.1 Problem Statement

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ be a dataset of n independent, identically distributed observations sampled from a distribution P. Our goal is to learn the conditional distribution $P(Y|X)$, where $(X, Y) \sim P$. Here, we study a learning strategy based on conditional mean embeddings (Song et al., 2010), which seek to represent conditional distributions as operators between an input and output RKHS. Formally, let $\mathcal{H}_K$ be a separable RKHS on $\mathcal{X}$ with bounded measurable kernel $k(\cdot,\cdot)$ and $\mathcal{H}_L$ be a separable RKHS on $\mathcal{Y}$ with measurable kernel $l(\cdot,\cdot)$. Then, according to Song et al. (2009), we define a conditional mean embedding $C_{Y|X} : \mathcal{H}_K \to \mathcal{H}_L$ as follows:

Definition 2.1. The conditional mean embedding operator $C_{Y|X} : \mathcal{H}_K \to \mathcal{H}_L$ is defined such that:

- $\mu_{Y|x} \equiv \mathbb{E}_{Y|x}[l(Y,\cdot)] = C_{Y|X} k(x,\cdot)$
- $\mathbb{E}_{Y|x}[g(\cdot)] = \langle g, \mu_{Y|x}\rangle_L$ for all $g \in \mathcal{H}_L$ and $x \in \mathcal{X}$

Essentially, the operator $C_{Y|X}$ performs the action of conditioning on some $x \in \mathcal{X}$, which is represented by its feature mapping $k(x,\cdot) \in \mathcal{H}_K$. The output $\mu_{Y|x} \in \mathcal{H}_L$ then represents the conditional distribution $P(\cdot|x)$ in the output feature space $\mathcal{H}_L$— evaluating the conditional expectation of some function $g \in \mathcal{H}_L$ simply reduces to taking its inner product with $\mu_{Y|x}$. Thus, in a sense, $\mu_{Y|x}$ can be interpreted as a generalization of the "density" of $P(\cdot|x)$, although it is important to note that distributions do not need to possess Lebesgue densities to be represented via a CME.

It is important to note that, implicit in Definition 2.1 is the assumption that the function $g_f(\cdot) = \mathbb{E}[f(Y)|X = \cdot]$ is contained in $\mathcal{H}_K$ for every $f \in \mathcal{H}_L$. This is a strong assumption, and forms the so-called "well-specified" scenario treated exhaustively in the literature (see e.g Song et al. (2009, 2010)). It is violated in several common cases, such as when X and Y are independent and $\mathcal{H}_K$ is a Gaussian RKHS, which does not contain the constant functions $g_f(\cdot)$ for any $f \in \mathcal{H}_L$ (see Corollary 4.44 in Steinwart and Christmann (2008); our Lemma D.6 demonstrates that constants *are* included in every interpolation space, however). A key feature of our analysis will involve relaxing this assumption by replacing $\mathcal{H}_K$ in Definition 2.1 with a larger *interpolation* space $\mathcal{H}_K^\beta$ that lies "between" $\mathcal{H}_K$ and $L^2(P_X)$ (defined rigorously in the following section). Hence, our framework proves robust as long there exists some such fractional space that contains $g_f(\cdot)$ for every $f \in \mathcal{H}_L$ — in section 3, we demonstrate how our learning rates depend on the smoothness of this space.

We also define the uncentered kernel covariance $C_{XX} = \mathbb{E}_X[k(X,\cdot) \otimes k(X,\cdot)]$ and cross-covariance $C_{YX} = \mathbb{E}_{YX}[l(Y,\cdot) \otimes k(X,\cdot)]$ operators. Note here that $\otimes$ may be interpreted as a tensor product, i.e. C_{XX}, for example, may be alternatively expressed as: $C_{XX} = \mathbb{E}_X[k(X,\cdot)\langle k(X,\cdot),\cdot\rangle_K]$, if we wish to make the action of C_{XX} on $\mathcal{H}_K$ more explicit. It can be easily shown (Klebanov et al., 2020) that $C_{Y|X} = (C_{XX}^\dagger C_{XY})^*$, when $C_{Y|X}$ exists (where $\dagger$ denotes the pseudo-inverse and $*$ the adjoint).

In practice, we do not have access to the true covariance operators C_{XX} and C_{YX}, and hence use the regularized sample CME $\hat{C}_{Y|X}^\lambda = \hat{C}_{YX}(\hat{C}_{XX} + \lambda)^{-1}$, where $\lambda > 0$ and the empirical operators $\hat{C}_{YX}$ and $\hat{C}_{YX}$ are defined precisely like their population counterparts (with $\mathbb{E}_{YX}[\cdot]$ replaced by the empirical expectation $\mathbb{E}_{\mathcal{D}}[\cdot]$). Grünewälder et al. (2012a) demonstrated that $\hat{C}_{Y|X}^\lambda$ solves the following regularized least-squares problem:

$$\arg\min_{T:\mathcal{H}_K \to \mathcal{H}_L} \frac{1}{n}\sum_{i=1}^{n} ||l(y_i,\cdot) - T[k(x_i,\cdot)]||_L^2 + \lambda ||T||_{\text{HS}}^2 \quad (1)$$

Traditionally, the sample complexity of solutions to (1) has been analyzed through the lens of vector-valued regression (e.g. Park and Muandet (2020); Grünewälder et al. (2012a)). In earlier works, the consistency of the sample CME was demonstrated via a spectral characterization (Song et al., 2010) that imposed strong compactness conditions on $C_{Y|X}$. Here, we develop an integral operator approach towards the analysis of (1) that seeks to significantly weaken the spectral conditions on $C_{Y|X}$ through the use of interpolation spaces — our approach is strongly motivated by Fischer and Steinwart (2020) where integral operator techniques were successfully applied towards the analysis of scalar-valued kernel regression problems. In section 3, we notably demonstrate that we can achieve the same learning rates derived in Fischer and Steinwart (2020) for our operator regression problem, under weaker smoothness conditions.

Remark (Proofs)**.** All proofs can be found in the supplementary appendices.

Remark (Notation)**.** In the remainder of this paper, we define $\hat{C}_{Y|X} \equiv \hat{C}_{YX}(\hat{C}_{XX} + \lambda I)^{-1}$, $C_{Y|X}^\lambda \equiv C_{YX}(C_{XX} + \lambda I)^{-1}$, and $\mu_{Y|x} = \mathbb{E}_{Y|X=x}[l(Y,\cdot)]$, $\hat{\mu}_{Y|x} = \hat{C}_{Y|X}(k(x,\cdot))$, and $\mu_{Y|x}^\lambda = C_{Y|X}^\lambda(k(x,\cdot))$. Note that when denoting the sample conditional embedding $\hat{C}_{Y|X}$ we suppress the dependence on λ and the number of samples n, as these are typically understood from context. Moreover, for any two Banach spaces A and B, we denote by $\mathcal{L}(A,B)$ the set of all continuous linear operators between A and B. For any $T \in \mathcal{L}(A,B)$, $||T||$ denotes the operator norm given by $||T|| = \sup_{||x||_A \leq 1} ||Tx||_B$ and $||T||_{\text{HS}}$ denotes a Hilbert-Schmidt norm. Occasionally, we denote this operator norm as $||\cdot||_{A\to B}$ in order to make the domain and codomain more explicit. Finally, we use the symbol $\preccurlyeq$ to denote the Loewner (semidefinite) order between positive semidefinite operators (i.e. $A \preccurlyeq B$ iff $B - A$ is positive semidefinite).

2.2 Mathematical Preliminaries

We first summarize the theory of interpolation spaces between $\mathcal{H}_K$ and $\mathcal{L}^2(\nu)$ (where $\nu = P_X$ is the marginal distribution on $\mathcal{X}$). Consider the injective imbedding $I_\nu : \mathcal{H}_K \to L^2(\nu)$ of $\mathcal{H}_K$ into $\mathcal{L}^2(\nu)$. Let $S_\nu = I_\nu^*$ be its adjoint. Then, it can be shown that S_ν is an integral operator given by:

$$S_\nu(f) = \int_{\mathcal{X}} k(x,\cdot) f(y) d\nu(y)$$

Using S_ν and I_ν, we construct the following positive self-adjoint operators on $\mathcal{H}_K$ and $\mathcal{L}^2(\nu)$, respectively:

$$C_\nu = S_\nu I_\nu = I_\nu^* I_\nu$$
$$T_\nu = I_\nu S_\nu = I_\nu I_\nu^*$$

We observe that C_ν and T_ν are nuclear (see Lemma 2.2/2.3 in Steinwart and Scovel (2012)), and that C_ν coincides with our uncentered cross-covariance operator C_{XX}. In our discussion/analyses below we typically only use the notation C_{XX} when considering expansions of the operators $\hat{C}_{Y|X}, C_{Y|X}$, or $C_{Y|X}^\lambda$ in order to remain consistent with literature (when the latter operators are abbreviated as in this sentence, we instead use C_ν). Since, T_ν is nuclear and self-adjoint, it admits a spectral representation:

$$T_\nu = \sum_{j=1}^{\infty} \mu_j e_j \langle e_j, \cdot\rangle_{L^2(\nu)}$$

where $\{\mu_j\}_{j=1}^{\infty} \in (0,\infty)$ are nonzero eigenvalues of T_ν (ordered nonincreasingly) and $\{e_j\}_{j=1}^{\infty} \subset L^2(\nu)$ form an orthonormal system of corresponding eigenfunctions. Note that formally, the elements e_j of $L^2(\nu)$ are equivalence classes $[e_j]_\nu$ whose members only differ on a set of ν-measure zero— notationally, we consider this formalism to be understood here and simply write e_j to refer to elements in both $\mathcal{H}_K$, $L^2(\nu)$, and their interpolation spaces (with the residence of e_j understood from context). We define the interpolation spaces $\mathcal{H}_K^\alpha$ as:

Definition 2.2. For $\alpha > 0$, we define the space $\mathcal{H}_K^\alpha$:

$$\mathcal{H}_K^\alpha = \Big\{ f = \sum_i a_i(\mu_i^{\frac{\alpha}{2}} e_i) : \{a_i\}_{i=1}^{\infty} \in \ell^2 \Big\}$$

with inner product:

$$\Big\langle \sum_i a_i(\mu_i^{\frac{\alpha}{2}} e_i), \sum_i b_i(\mu_i^{\frac{\alpha}{2}} e_i) \Big\rangle_{\mathcal{H}_K^\alpha} = \sum_i a_i b_i$$

We observe that, if $\alpha > \beta$, $\mathcal{H}_K^\alpha \subset \mathcal{H}_K^\beta \subset L^2(\nu)$, with $\mathcal{H}_K^1 = \mathcal{H}_K$. Note it is easy to see that $\{\mu_i^{\frac{\alpha}{2}} e_i\}_{i=1}^{\infty}$ is an orthonormal basis for $\mathcal{H}_K^\alpha$. We also observe that if:

$$\sum_{i=1}^{\infty} \mu_i^\alpha e_i^2(x) < \infty \ \ \forall\, x \in \mathcal{X} \tag{2}$$

then $\mathcal{H}_K^\alpha$ can be viewed as an RKHS whose reproducing kernel is equivalent to that of the integral operator T_ν^α on $L^2(\nu)$ (Proposition 4.2 in Steinwart and Scovel (2012)). Even, when (2) is not satisfied, we denote this kernel as $k^\alpha(x,y) = \sum_i \mu_i^\alpha e_i(x)e_i(y)$, and write $||k^\alpha||_\infty = \sup_{x\in\mathcal{X}} \sum_{i=1}^{\infty} \mu_i^\alpha e_i^2(x)$, if the latter quantity is finite. Hence, we may identify $\mathcal{H}_K^\alpha \cong \overline{\text{ran } T_\nu^{\frac{\alpha}{2}}}$. A detailed development of RKHS interpolation spaces can be found in Steinwart and Scovel (2012).

2.3 Conditional Embeddings on Interpolation Spaces

We develop the notion of conditional embeddings on interpolation spaces. We begin with the following definition:

Definition 2.3. Let $T : \mathcal{H}_K^\beta \to \mathcal{H}_L$ be a (possibly unbounded) operator for some $\beta > 0$ and let $\gamma \in (0,\beta)$. Let $I_{\beta,\gamma,\nu} : \mathcal{H}^\beta \to \mathcal{H}^\gamma$ be the canonical embedding. We define the operator norm $||\cdot||_\gamma$:

$$||T||_{\beta,\gamma} = ||T \circ I^*_{\beta,\gamma,\nu}||_{\mathcal{H}_K^\gamma \to \mathcal{H}_L}$$

where both norms may possibly be infinite. When $\beta = 1$, we simply write $||\cdot||_\gamma$

Our definition of the interpolation norm is motivated by the following observation:

Lemma 1. Suppose $C_{Y|X} : \mathcal{H}_K \to \mathcal{H}_L$ is well-defined according to Definition 2.1. Then, for any $\beta \in (0,1)$, $C_{Y|X} \circ I^*_{1,\beta,\nu}$ is the conditional mean embedding from $\mathcal{H}_K^\beta$ to $\mathcal{H}_L$ (by Definition 2.1 with $\mathcal{H}_K$ replaced by $\mathcal{H}_K^\beta$).

When the conditional mean embedding from $\mathcal{H}_K^\beta$ to $\mathcal{H}_L$ is well-defined, we denote it as $C^\beta_{Y|X}$. Note, implicit in this definition of $C^\beta_{Y|X}$ is the assumption that $\mathcal{H}_K^\beta$ is indeed an RKHS, i.e. it satisfies condition (2). Thus, from Lemma 1, we observe that if $C_{Y|X}$ and $C^\beta_{Y|X}$ are well-defined, then $||C_{Y|X}||_\beta = ||C^\beta_{Y|X}||$. The following result further elaborates the relationship between the operator norms $||\cdot||$ and $||\cdot||_\gamma$:

Lemma 2. Let $T : \mathcal{H}_K^\beta \to \mathcal{H}_L$ be an operator. Then, for any $\gamma \in (0,\beta)$, we have that:

$$||T||_{\beta,\gamma} = ||T \circ C^{\frac{1}{2}}_{\beta,\gamma,\nu}||_{\mathcal{H}_K^\beta \to \mathcal{H}_L}$$

where $C_{\beta,\gamma,\nu} = I^*_{\beta,\gamma,\nu} I_{\beta,\gamma,\nu}$

Our motivation behind introducing the Sobolev norms in Definition 2.3 stems from our desire to study operator convergence over the interpolation spaces $\mathcal{H}_K^\beta$. A distinguishing feature of our analysis is that we do not assume the existence of the CME $C_{Y|X}$ over $\mathcal{H}_K$, but merely over some interpolant $C^\beta_{Y|X}$ ($\beta \in (0,2)$), which maps $k^\beta(x,\cdot) \in H_K^\beta$ to $\mu_{Y|x}$ (we are primarily interested in the misspecified setting $0 < \beta < 1$). Since, we cannot approximate $C^\beta_{Y|X}$ directly (as the exponent β is typically unknown), we construct the regularized approximation $C^\lambda_{Y|X} \in \mathcal{L}(\mathcal{H}_K, \mathcal{H}_L)$ (which is always well-defined and bounded) and "pushback" to $\mathcal{H}_K^\beta$ via the composition $C^\lambda_{Y|X} \circ I^*_{1,\beta,\nu}$. We observe that $(C^\lambda_{Y|X} \circ I^*_{1,\beta,\nu})k^\beta(x,\cdot) = C^\lambda_{Y|X} k(x,\cdot) = \mu^\lambda_{Y|x}$, i.e. $C^\lambda_{Y|X} \circ I^*_{1,\beta,\nu}$ maps the canonical "feature" $k^\beta(x,\cdot)$ in $\mathcal{H}_K^\beta$ to the regularized mean embedding $\mu^\lambda_{Y|x} \in \mathcal{H}_L$.

Thus, the use of Sobolev norms here is primarily a *mathematical construction employed to compare operators defined over different domains* — in applications, we are mainly interested in estimating $||\mu_{Y|x} - \hat{\mu}_{Y|x}||_L$, i.e. the distance between the sample and true embeddings of the conditional distribution $P(\cdot|x)$ in the output RKHS $\mathcal{H}_L$. Bounding the latter distance provides insight into the sample error involved in computing the conditional expectation of a function $g \in \mathcal{H}_L$, as

$$\begin{aligned}|\langle g, \hat{\mu}_{Y|x}\rangle_L - \mathbb{E}[g(Y)|x]| &= |\langle g, \hat{\mu}_{Y|x}\rangle_L - \langle g, \mu_{Y|x}\rangle_L| \\ &\leq ||\mu_{Y|x} - \hat{\mu}_{Y|x}||_L ||g||_L\end{aligned}$$

(note that $\langle g, \hat{\mu}_{Y|x}\rangle_L$ is typically *not* an expectation of g with respect to some distribution, but simply an approximation of the true expectation $\mathbb{E}[g(Y)|x]$; see Grünewälder et al. (2012b) for more details). If $\mathcal{H}_K^\beta$ is continuously embedded in $L^\infty(\mathcal{X})$ (i.e. k^β is bounded), then we can obtain uniform bounds on $||\mu_{Y|x} - \hat{\mu}_{Y|x}||_L$ over all $x \in \mathcal{X}$, by estimating the operator distance $||C^\lambda_{Y|X} \circ I^*_{1,\beta,\nu} - C^\beta_{Y|x}||$. Indeed, we have:

$$\begin{aligned} ||\mu_{Y|x} - \hat{\mu}_{Y|x}||_L &= ||(\hat{C}_{Y|X} \circ I^*_{1,\beta,\nu})k^\beta(x,\cdot) - (C^\beta_{Y|X})k^\beta(x,\cdot)||_L \\ &\leq ||\hat{C}_{Y|X} \circ I^*_{1,\beta,\nu} - C^\beta_{Y|X}||||k^\beta(x,\cdot)||_\beta \\ &\leq ||k^\beta||_\infty||\hat{C}_{Y|X} \circ I^*_{1,\beta,\nu} - C^\beta_{Y|X}|| \end{aligned} \tag{3}$$

In the following section, we discuss the various parameter regimes in which such bounds are attainable.

Remark (Abuse of Notation)**.** In light of Lemma 1, for the remainder of the paper, when $C^\beta_{Y|X}$ is well-defined, we abuse notation and simply write $||\hat{C}_{Y|X} - C_{Y|X}||_\beta$ to express $||\hat{C}_{Y|X} \circ I^*_{1,\beta,\nu} - C^\beta_{Y|X}||$, even when $C_{Y|X}$ is not well-defined/bounded, in order to make explicit the distance between a sample estimator and its "true" value. Similarly, for any $\gamma < \beta$, we define $||\hat{C}_{Y|X} - C_{Y|X}||_\gamma$ as $||\hat{C}_{Y|X} \circ I^*_{1,\gamma,\nu} - C^\beta_{Y|X} \circ I^*_{\beta,\gamma,\nu}||$

2.4 Assumptions

We state some assumptions similar to those in Fischer and Steinwart (2020) — namely, we impose conditions on the decay of the eigenvalues of T_ν, the boundedness of a kernel interpolant, the conditional kernel moments of our target distribution, and the boundedness of $C^\beta_{Y|X}$. Below, we discuss how the latter assumption is weaker than the direct generalization of its analogous hypothesis in Fischer and Steinwart (2020) for scalar-valued regression.

Assumption 1. There exists a $0 < p < 1$ such that $ci^{-\frac{1}{p}} \leq \mu_i \leq Ci^{-\frac{1}{p}}$, for some $c, C > 0$

Assumption 2. There exists a $0 < p < \alpha \leq 1$ such that the inclusion map $i : H^\alpha_K \hookrightarrow L^\infty(\nu)$ is continuous, with $||i|| = ||k^\alpha|| \leq A$ for some $A > 0$ (we define α as the smallest value satisfying these conditions)

Assumption 3. There exists a $0 < p < \beta < 2$ such that $||C^\beta_{Y|X}|| \leq B < \infty$

Assumption 4. There exists a trace-class operator $V : \mathcal{H}_L \to \mathcal{H}_L$ and scalar $R > 0$, such that for every $x \in \mathcal{X}$ and $p \geq 1$:

$$\mathbb{E}_{Y|x}\Big[\Big((L(Y,\cdot)-\mu_{Y|x})\otimes(L(Y,\cdot)-\mu_{Y|x})\Big)^p\Big] \preccurlyeq \frac{(2p)!R^{2p-2}}{2}V \tag{4}$$

2.4.1 Discussion/Comparison of Assumptions

Although they are listed separately here, assumptions 1 and 2 are indeed highly related as they both (implicitly) impose conditions on the summability of powers of eigenvalues of T_ν. Indeed, under an additional assumption of uniform boundedness of the eigenfunctions e_i, assumptions 1 and 2 can be shown to be equivalent for certain domains of α and p. A comprehensive discussion of the relationship between these two assumptions can be found in Fischer and Steinwart (2020); Steinwart and Scovel (2012).

Assumption 3 characterizes the continuity of the "true" conditional embedding operator. Note, a distinctive feature of our approach is that we not only allow the CME to exist over any fractional RKHS $\mathcal{H}^\beta_K$, but additionally only require that the CME is *bounded* over this space. This contrasts strongly with existing operator-theoretic literature (Song et al., 2009, 2010) where the CME is required to be Hilbert-Schmidt (or equivalently belong to a product RKHS in the regression formulation of Park and Muandet (2020)) in order to achieve explicit learning rates. The significance of this relaxation can be seen in the trivial example when $Y = X$ and $1 - \beta \leq \frac{p}{2}$: indeed here, it can be easily seen that $C^\beta_{Y|X} = I^*_{1,\beta,\nu}$, and hence $||C^\beta_{Y|X}|| = \mu_1^{1-\beta} < \infty$ while $||C^\beta_{Y|X}||^2_{\text{HS}} = \sum_{i=1}^\infty \mu_i^{2(1-\beta)} \geq \sum_{i=1}^\infty i^{-1} = \infty$. More, generally, it can be shown that if $C^\beta_{Y|X}$ exists, then it is automatically bounded when $\mathbb{E}_Y[l(Y,Y)] < \infty$, i.e. when l is bounded (see Lemma 3 below). Note, however, compared to the scalar regression case in Fischer and Steinwart (2020), this introduces the condition that $p < \beta$, which is trivially satisfied when $\mathcal{H}^\beta_K$ is an RKHS (Proposition 4.4 in Steinwart and Scovel (2012)). We are able to remove the latter condition if we require $C^\beta_{Y|X}$ to be Hilbert-Schmidt, which would be a direct generalization of the source condition in Fischer and Steinwart (2020) (this trade-off is directly indicated in the remark following the proof of Lemma 6 in Appendix B).

We are primarily interested in the misspecified/"hard learning" scenario when $0 < \beta < 1$, the regime where the conditional embedding is *not* bounded over the RKHS $\mathcal{H}_K$, as this is where our framework improves on the related literature. Note that since the interpolation spaces are descending, the regime with $1 \leq \beta \leq 2$ simply collapses to $\beta = 1$, which has already been analyzed in Song et al. (2010). We only include this regime here, to demonstrate that we can generalize the learning rates from Fischer and Steinwart (2020) almost exactly. We now demonstrate the relationship between Assumption 3 and the hard-learning scenario:

Lemma 3. Assumption 3 is equivalent to $\sup_{||f||_L \leq 1} ||\mathbb{E}[f(Y)|X = \cdot]||_{H_K^\beta} = B < \infty$ for some β with $0 < p < \beta < 2$. If $\mathbb{E}_Y[l(Y, Y)] < \infty$ and $C_{Y|X}^\beta$ exists, then Assumption 3 is automatic.

Lemma 3 captures the generality of our approach — notice that unlike the classical framework of CME, we do not require $\mathbb{E}[f(Y)|X = \cdot] \in \mathcal{H}_K$ for $f \in \mathcal{H}_L$. Indeed, by Lemma 3, $\mathbb{E}[f(Y)|X = \cdot] \in \mathcal{H}_K^\beta$ must only lie in a $|| \cdot ||_{\mathcal{H}_K^\beta}$ ball of radius B for all unit vectors $f \in \mathcal{H}_L$. Intuitively, the act of conditioning on $\mathcal{X}$ must map $\mathcal{H}_L$ continuously into $\mathcal{H}_K^\beta$, which is strictly larger than $\mathcal{H}_K$ for $\beta \in (0, 1)$ — the misspecified setting.

Recall that in the previous section, we demonstrate that uniform convergence rates are attainable when $\alpha < \beta$ (note that Assumption 2 automatically qualifies $\mathcal{H}_K^\alpha$ as a bounded RKHS). This setting is attainable in many common settings — for example, when k is a Matérn kernel of order $\gamma > 0$ on a bounded open subset $\mathcal{X} \subset \mathbb{R}^d$ with strong Lipschitz boundary, k^α is bounded for all $\alpha \in \left(\frac{2\gamma+d}{d}, 1\right)$ (see e.g. Example 4.8 in Steinwart (2019)). Moreover, in this scenario, the condition $p < \beta$ translates to requiring $\mathcal{H}_K^\beta \cong W^s(\mathcal{X})$ for $s > \frac{d}{2}$.

Assumption 4 may be viewed as an "operator subexponential" condition that controls the norm of the conditional operator MGF. Like Assumption 3, Assumption 4 can be weakened to $\mathbb{E}_{Y|x}[||l(Y, \cdot) - \mathbb{E}_{Y|x}[l(Y, \cdot)]||^{2p}] \leq \frac{(2p)!R^{2p-2}}{2}\sigma^2$ (for $\sigma \in \mathbb{R}$) if Assumption 3 is replaced with a stronger Hilbert-Schmidt criterion (which would be the natural generalization of the corresponding assumptions from Fischer and Steinwart (2020) to operator-valued RKHS). However, Lemma 4 demonstrates that Assumption 4 is satisfied when the *output* RKHS also satisfies a variant of Assumptions 1/2, suggesting that the tradeoff we choose here indeed achieves more generality.

Lemma 4. Let π be a measure on $\mathcal{Y}$. Suppose $\mathcal{H}_L$ is compactly and injectively embedded in $L^2(\pi)$, l is bounded ($\sup_{y\in\mathcal{Y}} \sqrt{l(y, y)} = \ell < \infty$), and T_π has spectrum $\{(\eta_i, f_i)\}_{i=1}^\infty$ (where T_π is defined on $L^2(\pi)$ analogously to T_ν above). Then, if $\eta_i = \mathcal{O}\left(i^{-q^{-1}}\right)$ for $0 < q < 1$, and $K \equiv \sup_{y\in\mathcal{Y}} \sum_{i=1}^\infty \eta_i^\gamma f_i^2(y) < \infty$ for some $\gamma \in (0, 1-q)$, we have that Assumption 4 is satisfied for $R = 2\ell$ and $V = KC_\pi^{1-\gamma}$.

Thus, Lemma 4 demonstrates that we can reduce Assumption 4 to a condition on the RKHS $\mathcal{H}_L$, instead of a constraint on the *conditional distribution* $Y|x$. Moreover, although Assumptions 1 and 2 are more restrictive than the measure-theoretic frameworks of Mollenhauer et al. (2020) and Park and Muandet (2020), these hypotheses do not impose conditions on the conditional distribution $P(Y|X)$ or the CME, but instead prescribe the relationship between the kernel and the measure $\nu = P_X$. A crucial feature of our analysis involves establishing explicit learning rates that are *adaptive* to this relationship between kernel complexity (Assumptions 1 and 2) and the CME continuity (Assumption 3).

As mentioned previously, Assumptions 1 and 2 are borrowed directly from Fischer and Steinwart (2020). However, our assumption 3 requiring only boundedness of $C_{Y|X}^\beta$ is significantly weaker than the Hilbert-Schmidt condition that would result from a direct generalization of the source condition in Fischer and Steinwart (2020) to operator-valued RKHSs. Indeed, the use of a more general source condition in Assumption 3 and operator subexponentiality in Assumption 4 distinguishes our analysis from that of Fischer and Steinwart (2020) and requires the development of additional approximation machinery to obtain operator norm learning rates (see Appendix C and section 3.1)

2.4.2 Example: Markov Operators

To further demonstrate the generality of Assumption 3, we consider an example involving Markov transition operators, which have recently found an elegant connection to CMEs (see e.g. Mollenhauer and Koltai (2020); Mollenhauer et al. (2020)). Although this example is presented to provide a concrete comparison with existing applications of conditional embeddings (Grünewälder et al., 2012b; Lever et al., 2016), it should be noted that the argument applies to any setting where the input and output variables range over the same measure space (i.e. $\mathcal{X} = \mathcal{Y}$).

Let $\{X_t\}_{t\geq 0}$ be a Markov process on a state space $S \subset \mathbb{R}^d$. Fix $\tau > 0$, and let $p_\tau(y|x) = P(X_{t+\tau} = y|X_t = x)$ be the conditional transition density. Then, we may define the Koopman operator P_τ acting on an observable $\phi \in \mathcal{F}$ in some suitable function space $\mathcal{F}$ by:

$$(P_\tau \phi)(x) = \int_S p_\tau(y|x)\phi(y)dy$$

In applications, we are often interested in building empirical approximations to P_τ which typically require restricting the domain of P_τ to an amenable space. Recently, kernel methods have been proposed for this purpose (Klus et al., 2020; Mollenhauer et al., 2020; Klus et al., 2018), where $\mathcal{F}$ is set to some RKHS $\mathcal{H}_K$ typically over $L^2(\nu)$ (here ν is the stationary measure invariant under P_τ). Indeed, Klus et al. (2020) demonstrate that when $\mathcal{F}$ is an RKHS, the Koopman operator P_τ is simply the dual of the conditional mean embedding mapping $P(X_t = \cdot)$ to $P(X_{t+\tau} = \cdot)$, thereby enabling the straightforward application of CME machinery towards the empirical estimation of P_τ. However, their construction notably requires P_τ to be invariant over $\mathcal{H}_K$, which as mentioned in Mollenhauer et al. (2020) and Das and Giannakis (2020) is quite restrictive, and equivalent to the assumption that the CME of p_τ exists from $\mathcal{H}_K$ to $\mathcal{H}_K$. Notably, this assumption often introduces a model error by requiring the (possibly weak) approximation of P_τ in $\mathcal{L}(\mathcal{H}_K, \mathcal{H}_K)$. We observe that this assumption is significantly relaxed in our framework — indeed by Lemma 3, we require that P_τ merely be a bounded operator on $\mathcal{H}_K$ with range in $\mathcal{H}_K^\beta$ for some $\beta \in (0, 1]$ (the latter space being strictly larger than $\mathcal{H}_K$ when $\beta < 1$). Hence, we may apply the new misspecified learning rates developed here towards the data-driven estimation of the Koopman operator P_τ in much broader settings.

3 TECHNICAL CONTRIBUTIONS

We first present our main result in Theorem 5. As expected, we achieve a faster learning rate as $\gamma \to 0$, i.e. as the norm $||\cdot||_\gamma$ weakens.

Theorem 5. Suppose Assumptions 1-4 are satisfied, and that $\sup_{x\in\mathcal{X}} ||\mu_{Y|x}||_L \leq \tilde{C} < \infty$. Then, let $\lambda_n \asymp \left(\frac{\log^r n}{n}\right)^{\frac{1}{\max\{\alpha,\beta+p\}}}$ for some $r > 1$. Then there exists a constant $K > 0$ (independent of n and δ), such that for $0 < \gamma < \beta$:

$$||\hat{C}_{Y|X} - C_{Y|X}||_\gamma \leq K \log(\delta^{-1}) \left(\frac{n}{\log^r n}\right)^{-\frac{\beta-\gamma}{2\max\{\alpha,\beta+p\}}}$$

with probability $1 - 2\delta$.

The exponent $\frac{\beta-\gamma}{2\max\{\alpha,\beta+p\}}$ illustrates that the learning rate hinges quite naturally on the comparison between α and β. Intuitively, the exponent α characterizes the boundedness of our kernel, while β characterizes the boundedness of the conditional mean operator. The sizes of α and β are related inversely to specification, with our problem being more strongly specified as $\alpha \to 0$ and $\beta \to 1$. We therefore expect to achieve faster learning rates for low α and high β. Indeed, when $\alpha > 2\beta$ (i.e. the kernel is relatively poorly bounded), then $\alpha = \max\{\alpha, \beta + p\}$, which will limit the magnitude of the exponent and lead to a slow learning rate. Conversely, if $\beta > \alpha$, then we can bring our learning rate arbitrarily close to $\frac{\beta}{2(\beta+p)} \geq \frac{1}{4}$ (by Assumption 3). Note, in this regime, the Sobolev norm learning rate $||\cdot||_\gamma$ is only useful for establishing uniform convergence rates when $\gamma \geq \alpha$. Indeed, by (3), here we can obtain a uniform error bound in $||\hat{\mu}_{Y|x} - \mu_{Y|x}||_L$ for the sample conditional mean embedding $\hat{\mu}_{Y|x}$ over all $x \in \mathcal{X}$. Moreover, as we will see later in Lemma 6, the exponent $\alpha - \beta$ characterizes our ability to control the worst-case bias of our estimator $\sup_{x\in\mathcal{X}} ||\mu^\lambda_{Y|x} - \mu_{Y|x}||$ as $\lambda \to 0$, which likewise relates directly to the convergence of the sample embedding operator.

Remark. We note that the additional assumption $\sup_{x\in\mathcal{X}} ||\mu_{Y|x}||_L \leq \tilde{C}$ is not very restrictive, as this is easily satisfied when the kernel ℓ is bounded (recall we do not place an *a priori* assumption on the boundedness of the output kernel ℓ).

Corollary 5.1. Suppose the hypotheses of Theorem 5. Then, if $\beta > \alpha$, we obtain, with probability $1 - 2\delta$ and constant $K > 0$:

$$\sup_{x\in\mathcal{X}} ||\hat{\mu}_{Y|x} - \mu_{Y|x}||_L \leq K \log(\delta^{-1}) \left(\frac{n}{\log^r n}\right)^{-\frac{\beta-\alpha}{2(\beta+p)}}$$

We emphasize that we are able to achieve learning rates roughly matching those in Fischer and Steinwart (2020) for scalar-valued regression, despite only requiring the continuity/boundedness of our target $C^\beta_{Y|X}$ rather than the stronger smoothness source condition imposed on the regression function in Fischer and Steinwart (2020). Moreover, we note that we obtain a roughly similar $\frac{\log n}{n}$ base observed in Grünewälder et al. (2012a) for finite-dimensional RKHSs, which is unsurprising as the latter is based off the regularized learning rates of Caponnetto and De Vito (2007), which is foundational in the scalar-valued kernel regression literature.

3.1 Proof of Theorem 5

To estimate the error $||\hat{C}_{Y|X} - C_{Y|X}||_\gamma = ||\hat{C}_{YX}(\hat{C}_{XX} + \lambda I)^{-1} - C_{Y|X}||_\gamma$, we follow the standard procedure by separating into bias and variance terms, and bounding each term independently. Namely, we write:

$$||\hat{C}_{Y|X} - C_{Y|X}||_\gamma \leq ||\hat{C}_{Y|X} - C^\lambda_{Y|X}||_\gamma + ||C^\lambda_{Y|X} - C_{Y|X}||_\gamma \quad (5)$$

Our main tool will be Theorem 7, where we estimate the variance by notably applying the subexponential condition in Assumption 4 and a new operator Bernstein inequality derived in Lemma C.3 in the Appendix, which may be of independent interest. Lemma C.3 is crucial in our analysis, as it enables us to quantify the variance in (5) *directly in operator norm*, rather than embedding the operators in a product RKHS, which implicitly requires them to be Hilbert-Schmidt (see discussion in e.g. Park and Muandet (2020); Mollenhauer and Koltai (2020)).

Like in Fischer and Steinwart (2020), our variance bound in Theorem 7 is expressed implicitly in terms of the worst-case bias. Hence, we first discuss the estimation of this bias term $||C^{\lambda}_{Y|X} - C_{Y|X}||_{\gamma}$.

3.1.1 Bounding the Bias

In the following result, we seek to estimate several different measures of the bias between $\mu^{\lambda}_{Y|X}$ and $\mu_{Y|X}$, that relate to the various spectral properties of the covariance operators arising in Theorem 7. We will see that while the "average" bias $\mathbb{E}_X[||\mu^{\lambda}_{Y|X} - \mu_{Y|X}||^2_L]$ can always be shown to decay polynomially at order $\beta - p$ as $\lambda \to 0$, estimating the worst-case bias is less straightforward. Notably, we can only demonstrate the polynomial decay of the latter quantity when $\beta > \alpha$, i.e. the "nice" regime when the conditional embedding can be expressed as a continuous operator acting on bounded RKHS, leading to uniform convergence rates in the output space $\mathcal{H}_L$. When $\beta \leq \alpha$, we can merely bound this worst-case bias in a way sufficient to achieve the learning rates in Theorem 5.

We argue that imposing a continuity constraint on $C^{\beta}_{Y|X}$ in Assumption 3 is more natural for studying uniform convergence of $\hat{\mu}_{Y|x}$, rather than the stronger Hilbert-Schmidt criteria often imposed in vector-valued regression. Indeed, estimating the bias $||\mu^{\lambda}_{Y|x} - \mu_{Y|x}||_L$ and sample error $||\hat{\mu}_{Y|x} - \mu_{Y|x}||_L$ involve quantifying distances in the output RKHS $\mathcal{H}_L$, the codomain of the true ($C^{\beta}_{Y|X}$), sample ($\hat{C}_{Y|X}$), and regularized ($C^{\lambda}_{Y|X}$) conditional embedding operators. Since $\mu_{Y|x}, \hat{\mu}_{Y|x}$, and $\mu^{\lambda}_{Y|x}$ lie more specifically in the range of their respective operators ($C^{\beta}_{Y|X}, \hat{C}_{Y|X}$, and $C^{\lambda}_{Y|X}$), intuitively, it is sufficient to constrain the operator norms of the latter to obtain uniform upper bounds in $\mathcal{H}_L$. However, we must note that additionally requiring $C^{\beta}_{Y|X}$ to be Hilbert-Schmidt would allow us to achieve the polynomial decay of the expected bias in (6) for any $\beta \in (0, 2)$, without requiring $\beta > p$ as in Assumption 3 (elaborated in the remark following the proof of Lemma 6 in Appendix B). We view this tradeoff to be quite minor with respect to elimination of the Hilbert-Schmidt requirement on $C^{\beta}_{Y|X}$ (as discussed in section 2.4.1).

Lemma 6. Suppose Assumptions 1-4 and $\sup_{x\in\mathcal{X}} ||\mu_{Y|x}||_L \leq \tilde{C}$. Then, there exists a constant $D > 0$, such that for all $0 < \gamma < \beta < 2$:

$$\mathbb{E}_X[||\mu^{\lambda}_{Y|X} - \mu_{Y|X}||^2_L] \leq DB\lambda^{\beta-p} \tag{6}$$

$$M^2(\lambda) \equiv \sup_{x\in\mathcal{X}} ||\mu_{Y|x} - \mu^{\lambda}_{Y|x}||^2_L \leq (\tilde{C}^2 + ||k^{\alpha}||^2_{\infty}B^2)\lambda^{-(\alpha-\beta)_+} \tag{7}$$

$$M_{\lambda} \equiv ||\mathbb{E}[(\mu_{Y|x} - \mu^{\lambda}_{Y|x}) \otimes (\mu_{Y|x} - \mu^{\lambda}_{Y|x})]|| \leq B^2\lambda^{\beta} \tag{8}$$

$$||C^{\lambda}_{Y|X} - C_{Y|X}||_{\gamma} \leq B\lambda^{\frac{\beta-\gamma}{2}} \tag{9}$$

3.1.2 Bounding the Variance

We now present our primary estimate, where we demonstrate that for sufficiently large n, the "variance" of the sample CME (in $||\cdot||_{\gamma}$) can be estimated implicitly via the bias. Specifically, we use a new operator Bernstein inequality (detailed in Appendix C) and the framework of Theorem 16 in Fischer and Steinwart (2020), to demonstrate the concentration of $\hat{C}_{Y|X}$ around $C_{Y|X}$ for a fixed λ.

Theorem 7. Suppose Assumptions 1-4 hold. Let $\sigma^2 = \mathrm{tr}(V)$ (where V is defined in Assumption 4). Define:

$$\begin{aligned}
\mathcal{N}(\lambda) &= \mathrm{tr}(C_{\nu}(C_{\nu} + \lambda)^{-1}) \\
Q &= \max\{M(\lambda), R\} \\
g_{\lambda} &= \log\Big(2e\mathcal{N}(\lambda)\frac{||C_{\nu}|| + \lambda}{||C_{\nu}||}\Big) \\
\rho_{\lambda} &= \mathbb{E}\Big[(\mu^{\lambda}_{Y|X} - \mu_{Y|X}) \otimes (\mu^{\lambda}_{Y|X} - \mu_{Y|X})\Big] \\
\eta &= \max\Big\{\frac{(\sigma^2 + M^2(\lambda))||C_{\nu}||}{||C_{\nu}|| + \lambda}, ||\mathcal{N}(\lambda)V + \frac{||k^{\alpha}||^2_{\infty}}{\lambda^{\alpha}}\rho_{\lambda}||\Big\} \\
\beta(\delta) &= \log\Big(\frac{4((2\sigma^2 + M^2(\lambda))\mathcal{N}(\lambda) + \frac{||k^{\alpha}||^2_{\infty}}{\lambda^{\alpha}}\mathrm{tr}(\rho_{\lambda}))}{\eta\delta}\Big)
\end{aligned}$$

Then, for $n \geq 8||k^{\alpha}||^2_{\infty}\log(\delta^{-1})g_{\lambda}\lambda^{-\alpha}$:

$$||\hat{C}_{Y|X} - C^{\lambda}_{Y|X}||_{\gamma} \leq 3\lambda^{-\frac{\gamma}{2}}\Big(\frac{16Q||k^{\alpha}||_{\infty}\beta(\delta)}{\lambda^{\frac{\alpha}{2}}n} + 8\sqrt{\frac{\eta\beta(\delta)}{n}}\Big) \tag{10}$$

with probability $1 - 2\delta$

The proof of Theorem 5 then follows by substituting the bias estimates in Lemma 6 into (10), combining with the operator bias bound in (9), and considering the behavior of the resulting bound as $\lambda_n \asymp \left(\frac{\log^r n}{n}\right)^{\frac{1}{\max\{\alpha,\beta+p\}}}$ as $n \to \infty$. A full proof of these three results can be found in Appendix B.

4 DISCUSSION

In this paper, we derive novel learning rates for conditional mean embeddings under a new misspecified framework that significantly relaxes the Hilbert-Schmidt criteria currently required to guarantee uniform convergence on infinite-dimensional RKHS. This relaxation reduces the need to explicitly verify the smoothness of the learning target, which can often be difficult or counterintuitive. Our results hopefully enable the much broader application of existing ML/RL algorithms for conditional mean embeddings to more complex, misspecified settings involving infinite dimensional RKHS and continuous state spaces.

There are several remaining questions. Firstly, complementary lower bounds would be required for Theorem 5 to ensure the results presented here are indeed optimal. Given the ease in matching the upper bounds from the scalar-valued setting in Fischer and Steinwart (2020), we suspect that our learning rates are likewise optimal in this setting, however verifying this would require further analysis. A further interesting question involves exploring how the framework developed here may generalize to other regularization approaches, such as spectral regularization, or quantile/expectile regression.

Acknowledgements

We would like to thank Yunzong Xu for insightful discussions and comments, and acknowledge the support of the MIT Data Science Lab.

References

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3): 331–368, 2007.

Suddhasattwa Das and Dimitrios Giannakis. Koopman spectra in reproducing kernel hilbert spaces. *Applied and Computational Harmonic Analysis*, 49 (2):573–607, 2020.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *J. Mach. Learn. Res.*, 21:205–1, 2020.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *NIPS*, volume 20, pages 489–496, 2007.

Kenji Fukumizu, Francis R Bach, Michael I Jordan, et al. Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4):1871–1905, 2009.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimilano Pontil. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1803–1810, 2012a.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Massimilano Pontil, and Arthur Gretton. Modelling transition dynamics in mdps with rkhs embeddings. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1603–1610, 2012b.

Ilja Klebanov, Ingmar Schuster, and TJ Sullivan. A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, 2 (3):583–606, 2020.

Stefan Klus, Feliks Nüske, Péter Koltai, Hao Wu, Ioannis Kevrekidis, Christof Schütte, and Frank Noé. Data-driven model reduction and transfer operator approximation. *Journal of Nonlinear Science*, 28(3):985–1010, 2018.

Stefan Klus, Ingmar Schuster, and Krikamol Muandet. Eigendecompositions of transfer operators in reproducing kernel hilbert spaces. *Journal of Nonlinear Science*, 30(1):283–315, 2020.

Guy Lever, John Shawe-Taylor, Ronnie Stafford, and Csaba Szepesvári. Compressed conditional mean embeddings for model-based reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Junhong Lin and Volkan Cevher. Optimal convergence for distributed learning with stochastic gradient methods and spectral algorithms. *J. Mach. Learn. Res.*, 21:147–1, 2020.

Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020.

Ha Quang Minh. Some properties of gaussian reproducing kernel hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32(2):307–338, 2010.

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

Mattes Mollenhauer, Stefan Klus, Christof Schütte, and Péter Koltai. Kernel autocovariance operators of stationary processes: Estimation and convergence. *arXiv preprint arXiv:2004.00891*, 2020.

Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. *Advances in Neural Information Processing Systems*, 33, 2020.

Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968, 2009.

Le Song, Arthur Gretton, and Carlos Guestrin. Nonparametric tree graphical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 765–772, 2010.

Ingo Steinwart. Convergence types and rates in generic karhunen-loeve expansions with applications to sample path properties. *Potential Analysis*, 51(3):361–395, 2019.

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.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

Supplementary Material:
Sobolev Norm Learning Rates for Conditional Mean Embeddings

A Proofs for Sections 2.3 and 2.4

Proof of Lemma 1. We must demonstrate that $C_{Y|X} \circ I^*_{1,\beta,\nu} : \mathcal{H}^\beta_K \to \mathcal{H}_L$ satisfies the definition of the conditional mean embedding in Definition 2.1 where the input space is taken as $\mathcal{H}^\beta_K$ (instead of $\mathcal{H}_K$). Thus, we must show that $C_{Y|X} \circ I^*_{1,\beta,\nu} k^\beta(x,\cdot) = \mu_{Y|x}$. We first observe that, for any $f \in \mathcal{H}_K$ and $x \in \mathcal{X}$, we have:

$$\begin{aligned}
\langle I_{1,\beta,\nu} f, k^\beta(x,\cdot)\rangle_{\mathcal{H}^\beta_K} &= \langle f, k^\beta(x,\cdot)\rangle_{\mathcal{H}^\beta_K} \\
&= f(x) \\
&= \langle f, k(x,\cdot)\rangle_K
\end{aligned}$$

Hence, we have that $I^*_{1,\beta,\nu} k^\beta(x,\cdot) = k(x,\cdot)$. Therefore, by the definition of $C_{Y|X}$ in Definition 2.1, we have:

$$\begin{aligned}
(C_{Y|X} \circ I^*_{1,\beta,\nu}) k^\beta(x,\cdot) &= C_{Y|X} k(x,\cdot) \\
&= \mu_{Y|x}
\end{aligned}$$

and we obtain our result. □

Proof of Lemma 2. Since $\{\mu_i^{\frac{\beta}{2}} e_i\}_{i=1}^\infty$ is an orthonormal basis for $\mathcal{H}^\beta_K$, we may express any $f \in \mathcal{H}^\beta_K$ as $f = \sum_{i=1}^\infty \langle f, \mu_i^{\frac{\beta}{2}} e_i\rangle_{\mathcal{H}^\beta_K} \mu_i^{\frac{\beta}{2}} e_i$. Hence, we have:

$$\begin{aligned}
\langle f, C_{\beta,\gamma,\nu}(\mu_i^{\frac{\beta}{2}} e_i)\rangle_{\mathcal{H}^\beta_K} &= \langle f, I^*_{\beta,\gamma,\nu} I_{\beta,\gamma,\nu}(\mu_i^{\frac{\beta}{2}} e_i)\rangle_{\mathcal{H}^\beta_K} \\
&= \langle I_{\beta,\gamma,\nu} f, I_{\beta,\gamma,\nu}(\mu_i^{\frac{\beta}{2}} e_i)\rangle_{\mathcal{H}^\gamma_K} \\
&= \langle f, \mu_i^{\frac{\beta}{2}} e_i\rangle_{\mathcal{H}^\gamma_K} \\
&= \left\langle \sum_{i=1}^\infty \langle f, \mu_i^{\frac{\beta}{2}} e_i\rangle_{\mathcal{H}^\beta_K} \mu_i^{\frac{\beta}{2}} e_i, \mu_i^{\frac{\beta}{2}} e_i \right\rangle_{\mathcal{H}^\gamma_K} \\
&= \mu_i^{\beta-\gamma} \langle f, \mu_i^{\frac{\beta}{2}} e_i\rangle_{\mathcal{H}^\beta_K}
\end{aligned}$$

where the final step follows from the fact that $\{\mu_i^{\frac{\gamma}{2}} e_i\}_{i=1}^\infty$ is an orthonormal basis in $\mathcal{H}^\gamma_K$. Hence $C_{\beta,\gamma,\nu}$ is a positive self-adjoint operator on $\mathcal{H}^\beta_K$ with eigenvalues $\{\mu_i^{\beta-\gamma}\}_{i=1}^\infty$ and an orthonormal basis of eigenfunctions $\{\mu_i^{\frac{\beta}{2}} e_i\}_{i=1}^\infty$. Moreover, since $C_{\beta,\gamma,\nu} = I^*_{\beta,\gamma,\nu} I_{\beta,\gamma,\nu}$ by definition and $I_{\beta,\gamma,\nu}$ is the canonical embedding of H^β_K into H^γ_K, it follows that the action of $I^*_{\beta,\gamma,\nu} : H^\gamma_K \to \mathcal{H}^\beta_K$ can be characterized as:

$$I^*_{\beta,\gamma,\nu} e_i = \mu_i^{\beta-\gamma} e_i \quad \nu - \text{almost surely} \tag{11}$$

for all $i \in \mathbb{N}$. Now, let B_{ℓ^2} denote the unit ball in ℓ^2. Then, we have that for any linear operator $T : \mathcal{H}^\beta_K \to$

$\mathcal{H}_L$:

$$
\begin{aligned}
||T||_{\beta,\gamma} &= ||T \circ I^*_{\beta,\gamma,\nu}|| \\
&= \sup_{a \in B_{\ell^2}} \Big\| \sum_i a_i \mu_i^{\frac{\gamma}{2}} (T \circ I^*_{\beta,\gamma,\nu}) e_i \Big\| \\
&= \sup_{a \in B_{\ell^2}} \Big\| \sum_i \mu_i^{\beta - \frac{\gamma}{2}} a_i (T e_i) \Big\| \qquad \text{by (11)} \\
&= \sup_{a \in B_{\ell^2}} \Big\| \sum_i a_i \mu_i^{\frac{\beta-\gamma}{2}} T(\mu_i^{\frac{\beta}{2}} e_i) \Big\| \\
&= \sup_{a \in B_{\ell^2}} \Big\| \sum_i a_i (T \circ C^{\frac{1}{2}}_{\beta,\gamma,\nu}) \mu_i^{\frac{\beta}{2}} e_i \Big\| \\
&= ||T \circ C^{\frac{1}{2}}_{\beta,\gamma,\nu}||
\end{aligned}
$$

where again the last and second equalities follow from the fact that $\{\mu_i^{\frac{\beta}{2}} e_i\}_{i \in \mathbb{N}}$ and $\{\mu_i^{\frac{\gamma}{2}} e_i\}_{i \in \mathbb{N}}$ are orthonormal bases in $\mathcal{H}^\beta_K$ and $\mathcal{H}^\gamma_K$, respectively. □

The following result demonstrates that if $\mathbb{E}_Y[\ell(Y,Y)] < \infty$, then $C^\beta_{Y|X}$ is always bounded when it exists.

Lemma A.1. Suppose that $C^\beta_{Y|X}$ exists and $\mathbb{E}_Y[\ell(Y,Y)] < \infty$. Then, $C^\beta_{Y|X}$ is bounded.

Proof. Define $C_{\beta,XY} \equiv \mathbb{E}_{XY}[k^\beta(X,\cdot) \otimes l(Y,\cdot)]$. Note, that this operator is analogous to the cross-covariance operator C_{XY} defined in section 2.2, except that the feature vectors $k(x,\cdot)$ have been replaced by $k^\beta(x,\cdot)$, since $C_{\beta,XY}$ maps between the RKHS $\mathcal{H}_L$ and $\mathcal{H}^\beta_K$. Similarly, it is easy to see that the embedded covariance operator of X over $\mathcal{H}^\beta_K$ is simply $\mathbb{E}_X[k^\beta(X,\cdot) \otimes k^\beta(X,\cdot)]$ and equivalent to $C_{\beta,0,\nu} = I^*_{\beta,0,\nu} I_{\beta,0,\nu}$ (as defined in Lemma 2 and Definition 2.3; note here $I_{\beta,0,\nu}$ is simply the embedding of $\mathcal{H}^\beta_K$ in $L^2(\nu)$). Thus, by the discussion in section 2.2 and Klebanov et al. (2020), it follows that when $C^\beta_{Y|X}$ exists it is given by $(C^\dagger_{\beta,0,\nu} C_{\beta,XY})^*$. Hence, in order to demonstrate that $C^\beta_{Y|X}$ is bounded, we must only demonstrate that $C_{\beta,0,\nu}$ and $C_{\beta,XY}$ are bounded and then apply Theorem A.1 in Klebanov et al. (2020). It is clear that $C_{\beta,0,\nu}$ shares the same eigenvalues as T^β_ν (just as T_ν and C_ν), and hence $||C_{\beta,0,\nu}|| = \mu_1^\beta < \infty$. To see that $C_{\beta,XY} : \mathcal{H}_L \to \mathcal{H}^\beta_K$ is bounded, we note that, for $f \in \mathcal{H}_L$ with $||f||_L \leq 1$, we have:

$$
\begin{aligned}
||C_{\beta,XY} f||_{\mathcal{H}^\beta_K} &= \Big\| \mathbb{E}_{XY}[k^\beta(X,\cdot)\langle l(Y,\cdot), f\rangle_L] \Big\|_{\mathcal{H}^\beta_K} \\
&\leq ||f||_L \Big\| \mathbb{E}_{XY}[k^\beta(X,\cdot)\sqrt{l(Y,Y)}] \Big\|_{\mathcal{H}^\beta_K} && (12) \\
&\leq \mathbb{E}_{XY}[||k^\beta(X,\cdot)|| \sqrt{l(Y,Y)}] && (13) \\
&= \mathbb{E}_{XY}[\sqrt{k^\beta(X,X) l(Y,Y)}] \\
&\leq \sqrt{\mathbb{E}_X[k^\beta(X,X)] \mathbb{E}_Y[l(Y,Y)]} && (14) \\
&< \infty && (15)
\end{aligned}
$$

where (12) follows from Cauchy-Schwarz, (13) follows from Jensen's inequality and the fact that $||f||_L \leq 1$, (14) follows from Cauchy-Schwarz, and finally (15) follows from the assumption $\mathbb{E}_Y[\ell(Y,Y)] < \infty$ and the fact that $\mathbb{E}_X[k^\beta(X,X)] = \mathbb{E}_X[\sum_{i=1}^\infty \mu_i^\beta e_i^2(X)] < \infty$, since $\mathcal{H}^\beta_K$ is implicitly an RKHS (since the CME $C^\beta_{Y|X}$ is well-defined) and hence satisfies (2). Hence $C_{\beta,XY}$ is bounded, and our result follows from Theorem A.1 in Klebanov et al. (2020). Moreover, if Assumption 1 is satisfied for some $p > 1$, it follows that since $\infty > \mathbb{E}_X[k^\beta(X,X)] = \mathbb{E}_X[\sum_{i=1}^\infty \mu_i^\beta e_i^2(X)] = \sum_{i=1}^\infty \mu_i^\beta \geq c \sum_{i=1}^\infty i^{-p^{-1}\beta}$, that $\beta > p$. □

Proof of Lemma 3. Let $g_f(\cdot) = \mathbb{E}[f(Y)|X = \cdot]$. We observe that for every $x \in \mathcal{X}$:

$$g_f(x) = \mathbb{E}_{Y|x}[f(Y)] \tag{16}$$
$$= \langle f, \mu_{Y|x}\rangle_L \tag{17}$$
$$= \langle f, C^{\beta}_{Y|X} k^{\beta}(x,\cdot)\rangle_L \tag{18}$$
$$= \langle (C^{\beta}_{Y|X})^* f, k^{\beta}(x,\cdot)\rangle_{\mathcal{H}^{\beta}_K} \tag{19}$$

Since $g_f \in \mathcal{H}^{\beta}_K$ by assumption (recall this is implicit in the existence of $C^{\beta}_{Y|X} : \mathcal{H}^{\beta}_K \to \mathcal{H}_L$), we have that $g_f = (C^{\beta}_{Y|X})^* f$. The result then follows from:

$$\begin{aligned}
||C^{\beta}_{Y|X}||^2 &= \sup_{||f||_L \le 1} \sum_{i=1}^{\infty} \langle f, C^{\beta}_{Y|X} \mu_i^{\frac{\beta}{2}} e_i\rangle_L^2 \\
&= \sup_{||f||_L \le 1} \sum_{i=1}^{\infty} \langle (C^{\beta}_{Y|X})^* f, \mu_i^{\frac{\beta}{2}} e_i\rangle_{\mathcal{H}^{\beta}_K}^2 \\
&= \sup_{||f||_L \le 1} \sum_{i=1}^{\infty} \langle g_f, \mu_i^{\frac{\beta}{2}} e_i\rangle_{\mathcal{H}^{\beta}_K}^2 \\
&= \sup_{||f||_L \le 1} ||g_f||^2_{\mathcal{H}^{\beta}_K}
\end{aligned}$$

The second part of the lemma follows directly from Lemma A.1. □

Proof of Lemma 4. We first note that, here π may be any measure, and we only require that the compact imbedding $\mathcal{H}_L \hookrightarrow L^2(\pi)$ be injective (which ensures that $\{\eta_i^{\frac{1}{2}} f_i\}_{i=1}^{\infty}$ is indeed an orthonormal basis for $\mathcal{H}_L$ by Theorem 3.3 in Steinwart and Scovel (2012)) Let $g_f(x) = \mathbb{E}_{Y|x}[f(Y)]$, for $f \in \mathcal{H}_L$. Then, we have that:

$$\begin{aligned}
\mathbb{E}_{Y|x}\Big[\Big((l(Y,\cdot) - \mu_{Y|x}) \otimes (l(Y,\cdot) - \mu_{Y|x})\Big)^p\Big] &= \mathbb{E}_{Y|x}[||l(Y,\cdot) - \mu_{Y|x}||^{2p-2}(l(Y,\cdot) - \mu_{Y|x}) \otimes (l(Y,\cdot) - \mu_{Y|x})] \\
&\preccurlyeq (2\ell)^{2p-2}\mathbb{E}_{Y|x}[(l(Y,\cdot) - \mu_{Y|x}) \otimes (l(Y,\cdot) - \mu_{Y|x})] \qquad (20) \\
&\preccurlyeq (2\ell)^{2p-2}\mathbb{E}_{Y|x}[l(Y,\cdot) \otimes l(Y,\cdot)] \qquad (21)
\end{aligned}$$

where (20) follows from the fact that $\mu_{Y|x} = \mathbb{E}_{Y|x}[l(Y,\cdot)]$ by definition and $||l(y,\cdot)|| = \sqrt{l(y,y)} \le l$ by assumption. Now, since:

$$l(y,\cdot) = \sum_{i=1}^{\infty} \eta_i f_i(y) f_i$$

converges pointwise (Theorem 3.3 in Steinwart and Scovel (2012)), we have that for any $h \in \mathcal{H}_L$,

$$\begin{aligned}
\langle h, l(y,\cdot)\rangle_L^2 &= \Big\langle h, \sum_{i=1}^{\infty} \eta_i f_i(y) f_i \Big\rangle_L^2 \\
&\le \Big(\sum_{i=1}^{\infty} \eta_i^{\gamma} f_i^2(y)\Big)\Big(\sum_{i=1}^{\infty} \eta_i^{1-\gamma}\langle h, \eta_i^{\frac{1}{2}} f_i\rangle_L^2\Big) \\
&\le K\langle h, C_{\pi}^{1-\gamma} h\rangle_L \qquad (22)
\end{aligned}$$

where (22) follows from the fact that $K \equiv \sum_{i=1}^{\infty} \eta_i^{\gamma} f_i^2(y) < \infty$ by assumption, and in (22), C_π is defined analogously to C_ν in section 2.2. Hence, for all $y \in \mathcal{Y}$, $l(y,\cdot) \otimes l(y,\cdot) \preccurlyeq K C_{\pi}^{1-\gamma}$ and:

$$\mathbb{E}_{Y|x}\Big[\Big((l(Y,\cdot) - \mu_{Y|x}) \otimes (l(Y,\cdot) - \mu_{Y|x})\Big)^p\Big] \preccurlyeq (2\ell)^{2p-2}\mathbb{E}_{Y|x}[l(Y,\cdot) \otimes l(Y,\cdot)] \preccurlyeq K(2\ell)^{2p-2} C_{\pi}^{1-\gamma}$$

Finally, $\mathrm{tr}\Big(C_{\pi}^{1-\gamma}\Big) = \sum_i \eta_i^{1-\gamma} \asymp \sum_i i^{-q^{-1}(1-\gamma)} < \infty$ since $\gamma < 1 - q$. Hence, we obtain our result with $V = K C_{\pi}^{1-\gamma}$ and $R = 2\ell$. □

Remark (Assumptions in Lemma 4)**.** A particularly illustrative case of the assumption $\eta_i = \mathcal{O}\left(i^{-q^{-1}}\right)$ occurs when the η_i decay exponentially (such as when l is the Gaussian kernel and π is the Lebesgue measure), in which case it is easy to see that the decay condition holds for any $q \in (0,1)$. Moreover, we note that our boundedness condition $\left\|\sum_{i\in\mathbb{N}} \eta_i^{\gamma} f_i^2\right\|_{L^\infty(\mathcal{Y})} < \infty$ is significantly weaker than requiring the uniform boundedness of the eigenfunctions ($\sup_{i\in\mathbb{N}} ||f_i||_{L^\infty(\mathcal{Y})} < \infty$), the latter of which is often violated even for C^∞ kernels (see discussion in Steinwart and Scovel (2012) and Zhou (2002)). In fact, for the kernel in Example 1 of Zhou (2002), it can be shown that the requirement $\left\|\sum_{i\in\mathbb{N}} \eta_i^{\gamma} f_i^2\right\|_{L^\infty(\mathcal{Y})} < \infty$, is satisfied for any choice of $\gamma \in \left(\frac{\ln 8}{\ln 16}, 1\right)$, despite $||f_i||_{L^\infty(\mathcal{Y})}$ growing exponentially. *Most importantly, Lemma 4 demonstrates that we can replace the requirement on the conditional distribution $Y|X$ in Assumption 4 with a condition on $\mathcal{H}_L$ and thereby eliminate any constraints on $P(Y|X)$ in our hypotheses.*

B Proof of Theorem 5

Proof of Lemma 6. We first note that:

$$\begin{aligned}
\mu^{\lambda}_{Y|X} &= C_{YX}(C_{XX}+\lambda)^{-1}k(x,\cdot)\\
&= \mathbb{E}_{YX}[l(y,\cdot)\otimes k(x,\cdot)](C_{XX}+\lambda)^{-1}k(x,\cdot)\\
&= C^{\beta}_{Y|X}\mathbb{E}_X[k^{\beta}(x,\cdot)\otimes k(x,\cdot)](C_{XX}+\lambda)^{-1}k(x,\cdot) \qquad (23)
\end{aligned}$$

where (23) follows from the fact that $\mu_{Y|x} = \mathbb{E}_{Y|X=x}[l(Y,\cdot)] = C^{\beta}_{Y|X}k^{\beta}(x,\cdot)$ by the definition of the conditional embedding $C^{\beta}_{Y|X}$ on $\mathcal{H}^{\beta}_K$. We then observe that:

$$\begin{aligned}
\mathbb{E}_X[k^{\beta}(x,\cdot)\otimes k(x,\cdot)](C_{XX}+\lambda)^{-1}k(x,\cdot) &= \mathbb{E}_X\Big[\Big(\sum_{i=1}^{\infty}\mu_i^{\beta}e_i(X)e_i\Big)\otimes\Big(\sum_{i=1}^{\infty}\mu_i e_i(X)e_i\Big)\Big](C_{XX}+\lambda)^{-1}k(x,\cdot)\\
&= \Big(\sum_{i=1}^{\infty}\mu_i^{1+\beta}e_i\otimes e_i\Big)(C_{XX}+\lambda)^{-1}k(x,\cdot)\\
&= \Big(\sum_{i=1}^{\infty}\frac{\mu_i^{1+\beta}}{\mu_i+\lambda}e_i\otimes e_i\Big)k(x,\cdot)\\
&= \sum_{i=1}^{\infty}\frac{\mu_i^{1+\beta}}{\mu_i+\lambda}e_i(x)e_i
\end{aligned}$$

We thus have that:

$$\begin{aligned}
\mu^{\lambda}_{Y|X} - \mu_{Y|X} &= C^{\beta}_{Y|X}\Big(\mathbb{E}_X[k^{\beta}(x,\cdot)\otimes k(x,\cdot)](C_{XX}+\lambda)^{-1}k(x,\cdot)\Big) - C^{\beta}_{Y|X}k^{\beta}(x,\cdot)\\
&= C^{\beta}_{Y|X}\Big(\sum_{i=1}^{\infty}\frac{\mu_i^{1+\beta}}{\mu_i+\lambda}e_i(x)e_i - \sum_{i=0}^{\infty}\mu_i^{\beta}e_i(x)e_i\Big)\\
&= \sum_{i=1}^{\infty}\frac{\lambda}{\mu_i+\lambda}\cdot C^{\beta}_{Y|X}\mu_i^{\beta}e_i(x)e_i
\end{aligned}$$

Thus, we can write:

$$\begin{aligned}
\mathbb{E}_X[||\mu_{Y|X} - \mu^{\lambda}_{Y|X}||^2_L] &= \mathbb{E}_X\Big[\Big|\Big|\sum_{i=1}^{\infty}\frac{\lambda}{\lambda+\mu_i}C^{\beta}_{Y|X}\mu_i^{\beta}e_i(X)e_i\Big|\Big|^2_L\Big] \\
&= \mathbb{E}_X\Big[\Big|\Big|\sum_{i=1}^{\infty}\frac{\lambda\cdot\mu_i^{\frac{\beta}{2}}}{\lambda+\mu_i}C^{\beta}_{Y|X}\mu_i^{\frac{\beta}{2}}e_i(X)e_i\Big|\Big|^2_L\Big] \\
&\leq \lambda^2||C^{\beta}_{Y|X}||^2\mathbb{E}_X\Big[\sum_{i=1}^{\infty}\Big(\frac{\mu_i^{\frac{\beta}{2}}}{\lambda+\mu_i}\Big)^2 e_i^2(X)\Big] \\
&= \lambda^2||C^{\beta}_{Y|X}||^2\sum_{i=1}^{\infty}\Big(\frac{\mu_i^{\frac{\beta}{2}}}{\lambda+\mu_i}\Big)^2 \\
&\leq D\lambda^{\beta-p}||C^{\beta}_{Y|X}||^2
\end{aligned}$$

where the last line follows from Lemma D.2. Moreover, we have, for any $x \in X$:

$$\begin{aligned}
||\mu_{Y|x} - \mu^{\lambda}_{Y|x}||^2_L &= \Big|\Big|\sum_{i=1}^{\infty}\frac{\lambda}{\lambda+\mu_i}C^{\beta}_{Y|X}\mu_i^{\beta}e_i(x)e_i\Big|\Big|^2_L \\
&= \Big|\Big|\sum_{i=1}^{\infty}\frac{\lambda\cdot\mu_i^{\frac{\beta-\alpha}{2}}}{\lambda+\mu_i}\cdot\mu^{\frac{\alpha}{2}}e_i(x)\cdot C^{\beta}_{Y|X}\mu_i^{\frac{\beta}{2}}e_i\Big|\Big|^2_L \\
&\leq \Big(\sum_i\Big(\frac{\lambda\cdot\mu_i^{\frac{\beta-\alpha}{2}}}{\lambda+\mu_i}\Big)^2\mu^{\alpha}e_i^2(x)\Big)||C^{\beta}_{Y|X}||^2 \qquad (24) \\
&\leq \Big(\sup_i\Big(\frac{\lambda\cdot\mu_i^{\frac{\beta-\alpha}{2}}}{\lambda+\mu_i}\Big)^2\Big)\cdot\sum_i\mu^{\alpha}e_i^2(x)\cdot||C^{\beta}_{Y|X}||^2 \\
&\leq \lambda^{\beta-\alpha}||k^{\alpha}||^2_{\infty}||C^{\beta}_{Y|X}||^2
\end{aligned}$$

when $\beta > \alpha$ (here (24) follows from the fact that $\{\mu_i^{\frac{\beta}{2}}e_i\}_{i=1}^{\infty}$ is an orthonormal basis for $\mathcal{H}^{\beta}_K$ and the last line follows from Lemma 25 in Fischer and Steinwart (2020)). When $\beta < \alpha$, we have that:

$$\begin{aligned}
||\mu^{\lambda}_{Y|x}||^2_L &= \Big|\Big|\sum_{i=1}^{\infty}\frac{\mu_i}{\mu_i+\lambda}\cdot C^{\beta}_{Y|X}\mu_i^{\beta}e_i(x)e_i\Big|\Big|^2_L \\
&= \Big|\Big|\sum_{i=1}^{\infty}\frac{\mu_i^{1+\frac{\beta-\alpha}{2}}}{\lambda+\mu_i}\cdot\mu_i^{\frac{\alpha}{2}}e_i(x)\cdot C^{\beta}_{Y|X}\mu_i^{\frac{\beta}{2}}e_i\Big|\Big|^2_L \\
&= \Big(\sum_i\Big(\frac{\mu_i^{1+\frac{\beta-\alpha}{2}}}{\lambda+\mu_i}\Big)^2\mu_i^{\alpha}e_i^2(x)\Big)||C^{\beta}_{Y|X}||^2 \\
&\leq \lambda^{\beta-\alpha}||k^{\alpha}||^2_{\infty}||C^{\beta}_{Y|X}||^2
\end{aligned}$$

where again the last line follows from Lemma 25 in Fischer and Steinwart (2020). Thus, we have for all cases:

$$\begin{aligned}
||\mu_{Y|x} - \mu^{\lambda}_{Y|x}||_L &\leq ||\mu_{Y|x}||_L + ||\mu^{\lambda}_{Y|x}||_L \\
&\leq \tilde{C} + \lambda^{\frac{\beta-\alpha}{2}}||k^{\alpha}||_{\infty}||C^{\beta}_{Y|X}|| \\
&\leq (\tilde{C} + ||k^{\alpha}||_{\infty}||C^{\beta}_{Y|X}||)\lambda^{-\frac{(\alpha-\beta)_+}{2}}
\end{aligned}$$

where we have used the fact that we may assume the fixed $\lambda \leq 1$ (as the $\lambda_n \to 0$ in Theorem 5). Moreover, we have:

$$\begin{aligned}
||\mathbb{E}[(\mu_{Y|x} - \mu^{\lambda}_{Y|x}) \otimes (\mu_{Y|x} - \mu^{\lambda}_{Y|x})]||^2 &= \sup_{||f||_L \leq 1} \mathbb{E}[\langle f, \mu_{Y|x} - \mu^{\lambda}_{Y|x}\rangle_L^2] \\
&= \sup_{||f||_L \leq 1} \mathbb{E}\Big[\Big(\sum_{i=1}^{\infty} \frac{\lambda \cdot \mu_i^{\frac{\beta}{2}} e_i(X)}{\lambda + \mu_i} \langle f, C^{\beta}_{Y|X} \mu_i^{\frac{\beta}{2}} e_i\rangle_L\Big)^2\Big] \\
&\leq \sup_{||f||_L \leq 1} \sum_{i=1}^{\infty} \Big(\frac{\lambda \cdot \mu_i^{\frac{\beta}{2}}}{\lambda + \mu_i}\Big)^2 \langle f, C^{\beta}_{Y|X} \mu_i^{\frac{\beta}{2}} e_i\rangle_L^2 \qquad (25) \\
&\leq \Big(\sup_i \Big(\frac{\lambda \cdot \mu_i^{\frac{\beta}{2}}}{\lambda + \mu_i}\Big)^2\Big) \sup_{||f||_L \leq 1} \sum_{i=1}^{\infty} \langle f, C^{\beta}_{Y|X} \mu_i^{\frac{\beta}{2}} e_i\rangle_L^2 \qquad (26) \\
&\leq \lambda^{\beta} ||C^{\beta}_{Y|X}||^2 \qquad (27)
\end{aligned}$$

where (25) follows from the fact that $\mathbb{E}_X[e_i(X)e_j(X)] = \delta_{ij}$ (as $\{e_i\}_{i=1}^{\infty}$ is an orthonormal basis for $L^2(\nu)$), and the last step follows from $\{\mu_i^{\frac{\beta}{2}} e_i\}_{i=1}^{\infty}$ being an orthonormal basis in $\mathcal{H}_K^{\beta}$. For the final part of Lemma 6, we observe that like before:

$$\begin{aligned}
C^{\lambda}_{Y|X} &= C_{YX}(C_{XX} + \lambda)^{-1} \\
&= C^{\beta}_{Y|X} \mathbb{E}_X[k^{\beta}(x, \cdot) \otimes k(x, \cdot)](C_{XX} + \lambda)^{-1} \\
&= \sum_{i=1}^{\infty} \frac{\mu_i^{1+\beta}}{\mu_i + \lambda} C^{\beta}_{Y|X} e_i \otimes e_i \qquad (28)
\end{aligned}$$

Recall that:

$$||C^{\lambda}_{Y|X} - C_{Y|X}||_{\gamma} = ||C^{\lambda}_{Y|X} \circ I^*_{1,\gamma,\nu} - C^{\beta}_{Y|X} \circ I^*_{\beta,\gamma,\nu}||$$

by definition (see remark after section 2.3). Now, observe that for any element $f = \sum_i a_i \mu_i^{\frac{\gamma}{2}} e_i \in \mathcal{H}_K^{\gamma}$ with $\{a_i\}_{i=1}^{\infty} \in \ell^2$, we have that:

$$\begin{aligned}
(C^{\lambda}_{Y|X} \circ I^*_{1,\gamma,\nu}) f &= (C^{\lambda}_{Y|X} \circ I^*_{1,\gamma,\nu})\Big(\sum_i a_i \mu_i^{\frac{\gamma}{2}} e_i\Big) \\
&= C^{\lambda}_{Y|X}\Big(\sum_i a_i \mu_i^{1-\frac{\gamma}{2}} e_i\Big) \qquad (29) \\
&= \Big(\sum_{i=1}^{\infty} \frac{\mu_i^{1+\beta}}{\mu_i + \lambda} C^{\beta}_{Y|X} e_i \otimes e_i\Big)\Big(\sum_i a_i \mu_i^{1-\frac{\gamma}{2}} e_i\Big) \qquad (30) \\
&= \sum_{i=1}^{\infty} \frac{a_i \mu_i^{1+\beta-\frac{\gamma}{2}}}{\mu_i + \lambda} C^{\beta}_{Y|X} e_i \qquad (31)
\end{aligned}$$

where (29) follows from (11), (30) follows from (28) and noting that $\sum_i a_i \mu_i^{1-\frac{\gamma}{2}} e_i \in \mathcal{H}_K$, since $\mu_i \to 0$ (as C_{ν} is compact) and $\frac{1-\gamma}{2} > 0$ (as $\gamma < 1$ by assumption); and (31) follows from noting that $\{\mu_i^{\frac{1}{2}} e_i\}_{i=1}^{\infty}$ is an orthonormal basis in $\mathcal{H}_K$. Similarly, we have that:

$$\begin{aligned}
(C^{\beta}_{Y|X} \circ I^*_{\beta,\gamma,\nu}) f &= (C^{\beta}_{Y|X} \circ I^*_{\beta,\gamma,\nu})\Big(\sum_i a_i \mu_i^{\frac{\gamma}{2}} e_i\Big) \\
&= \sum_i a_i \mu_i^{\beta-\frac{\gamma}{2}} C^{\beta}_{Y|X} e_i
\end{aligned}$$

Thus, we have that:

$$
\begin{aligned}
||C^{\lambda}_{Y|X} - C_{Y|X}||_{\gamma} &= ||C^{\lambda}_{Y|X} \circ I^{*}_{1,\gamma,\nu} - C^{\beta}_{Y|X} \circ I^{*}_{\beta,\gamma,\nu}|| \\
&= \sup_{||(a_i)_{i=1}^{\infty}||_{\ell^2}=1} ||(C^{\lambda}_{Y|X} \circ I^{*}_{1,\gamma,\nu} - C^{\beta}_{Y|X} \circ I^{*}_{\beta,\gamma,\nu})\Big(\sum_i a_i \mu_i^{\frac{\beta}{2}} e_i\Big)||_L \\
&= \sup_{||(a_i)_{i=1}^{\infty}||_{\ell^2}=1} \Big|\Big| \sum_{i=1}^{\infty} \frac{a_i \lambda \cdot \mu_i^{\frac{\beta-\gamma}{2}}}{\mu_i + \lambda} \cdot C^{\beta}_{Y|X} \mu_i^{\frac{\beta}{2}} e_i \Big|\Big| \\
&\leq \Big(\sup_i \frac{\lambda \cdot \mu_i^{\frac{\beta-\gamma}{2}}}{\mu_i + \lambda}\Big) ||C^{\beta}_{Y|X}|| \\
&\leq \lambda^{\frac{\beta-\gamma}{2}} ||C^{\beta}_{Y|X}||
\end{aligned}
$$

□

Remark (Expected Bias for Hilbert-Schmidt $C^{\beta}_{Y|X}$)**.** Observe that when $C^{\beta}_{Y|X}$ is Hilbert-Schmidt, we have, by the above proof:

$$
\begin{aligned}
\mathbb{E}_X[||\mu_{Y|X} - \mu^{\lambda}_{Y|X}||_L^2] &= \mathbb{E}_X\Big[\Big|\Big| \sum_{i=1}^{\infty} \frac{\lambda}{\lambda + \mu_i} C^{\beta}_{Y|X} \mu_i^{\beta} e_i(X) e_i \Big|\Big|_L^2\Big] \\
&= \mathbb{E}_X\Big[\Big|\Big| \sum_{i=1}^{\infty} \frac{\lambda \cdot \mu_i^{\frac{\beta}{2}}}{\lambda + \mu_i} C^{\beta}_{Y|X} \mu_i^{\frac{\beta}{2}} e_i(X) e_i \Big|\Big|_L^2\Big] \\
&= \sum_{i=1}^{\infty} \Big(\frac{\lambda \cdot \mu_i^{\frac{\beta}{2}}}{\lambda + \mu_i}\Big)^2 ||C^{\beta}_{Y|X} \mu_i^{\frac{\beta}{2}} e_i||_L^2 \\
&\leq \lambda^{\beta} ||C^{\beta}_{Y|X}||_{\mathrm{HS}}
\end{aligned}
$$

where the last line follows from Lemma 25 in Fischer and Steinwart (2020) and the fact that $\{\mu_i^{\frac{\beta}{2}} e_i\}_{i=1}^{\infty}$ is an orthonormal basis of $\mathcal{H}_K^{\beta}$. Thus, when $C^{\beta}_{Y|X}$ is Hilbert-Schmidt, we can achieve polynomial decay of the expected bias for all $\beta \in (0, 2)$.

Proof of Theorem 7. We begin like in the proof of Theorem 16 in Fischer and Steinwart (2020). Namely, applying Lemma 2 we write:

$$
\begin{aligned}
||\hat{C}_{Y|X} - C^{\lambda}_{Y|X}||_{\gamma} &= ||(\hat{C}_{Y|X} - C^{\lambda}_{Y|X}) \circ C^{\frac{1}{2}}_{1,\gamma,\nu}|| && (32) \\
&= ||(\hat{C}_{Y|X} - C^{\lambda}_{Y|X}) \circ C_{XX}^{\frac{1-\gamma}{2}}|| && (33) \\
&= ||(\hat{C}_{YX}(\hat{C}_{XX} + \lambda)^{-1} - C_{YX}(C_{XX} + \lambda)^{-1}) C_{XX}^{\frac{1-\gamma}{2}}|| \\
&\leq ||(\hat{C}_{YX} - C_{YX}(C_{XX} + \lambda)^{-1}(\hat{C}_{XX} + \lambda))(C_{XX} + \lambda)^{-\frac{1}{2}}|| \cdot \\
&\quad ||(C_{XX} + \lambda)^{\frac{1}{2}} (\hat{C}_{XX} + \lambda)^{-1} (C_{XX} + \lambda)^{\frac{1}{2}}|| ||C_{XX}^{\frac{1-\gamma}{2}} (C_{XX} + \lambda)^{-\frac{1}{2}}|| && (34)
\end{aligned}
$$

where (32) follows from Lemma 2 and (33) follows from the fact that $C_{1,\gamma,\nu} = C_{\nu}^{1-\gamma} = C_{XX}^{1-\gamma}$, since $C_{1,\gamma,\nu}$ has eigenfunctions $\{\mu_i^{\frac{1}{2}} e_i\}_{i=1}^{\infty}$ and eigenvalues $\{\mu_i^{1-\gamma}\}_{i=1}^{\infty}$ (see proof of Lemma 2 in Appendix A). Note here, we have used the notation C_{XX} instead of C_{ν} to remain consistent with the expansions of $\hat{C}_{Y|X}$ and $C^{\lambda}_{Y|X}$ in the literature. We primarily focus on bounding the first factor on the RHS of (34), as the remaining factors can be estimated simply as discussed previously in Fischer and Steinwart (2020). To start, we again imitate the approach from the proof of Theorem 16 in Fischer and Steinwart (2020). Namely, we have:

$$\begin{aligned}\hat{C}_{YX} - C_{YX}(C_{XX}+\lambda)^{-1}(\hat{C}_{XX}+\lambda) &= \hat{C}_{YX} - C_{YX}(C_{XX}+\lambda)^{-1}(C_{XX}+\lambda+\hat{C}_{XX}-C_{XX}) \\ &= \hat{C}_{YX} - C_{YX} + C_{YX}(C_{XX}+\lambda)^{-1}(C_{XX}-\hat{C}_{XX}) \\ &= \hat{C}_{YX} - C_{YX}(C_{XX}+\lambda)^{-1}\hat{C}_{XX} - (C_{YX} - C_{YX}(C_{XX}+\lambda)^{-1}C_{XX}) \\ &= \hat{C}_{YX} - C_{YX}(C_{XX}+\lambda)^{-1}\hat{\mathbb{E}}[k(X,\cdot)\otimes k(X,\cdot)] - C_{YX} \\ &+ C_{YX}(C_{XX}+\lambda)^{-1}\mathbb{E}[k(X,\cdot)\otimes k(X,\cdot)] \\ &= \hat{\mathbb{E}}[(L(Y,\cdot)-\mu^{\lambda}_{Y|X})\otimes k(X,\cdot)] - \mathbb{E}[(L(Y,\cdot)-\mu^{\lambda}_{Y|X})\otimes k(X,\cdot)]\end{aligned}$$

We now wish to apply Lemma C.3 to bound this deviation. Let $h(X,\cdot) = (C_{XX}+\lambda)^{-\frac{1}{2}}k(X,\cdot)$. We first write:

$$(L(Y,\cdot)-\mu^{\lambda}_{Y|X})\otimes h(X,\cdot) = (L(Y,\cdot)-\mu_{Y|X})\otimes h(X,\cdot) + (\mu_{Y|X}-\mu^{\lambda}_{Y|X})\otimes h(X,\cdot)$$

Then, applying Corollary D.1.1, we can write:

$$\begin{aligned}\Big[\Big((L(Y,\cdot)-\mu^{\lambda}_{Y|X})\otimes h(X,\cdot)\Big)^*\Big((L(Y,\cdot)-\mu^{\lambda}_{Y|X})\otimes h(X,\cdot)\Big)\Big]^p &\preccurlyeq 2^{2p-1}\Big[||L(Y,\cdot)-\mu_{Y|X}||_L^{2p}\Big(h(X,\cdot)\otimes h(X,\cdot)\Big)^p \qquad (35)\\ &+ ||\mu^{\lambda}_{Y|X}-\mu_{Y|X}||_L^{2p}\Big(h(X,\cdot)\otimes h(X,\cdot)\Big)^p\Big] \qquad (36)\\ \Big[\Big((L(Y,\cdot)-\mu^{\lambda}_{Y|X})\otimes h(X,\cdot)\Big)\Big((L(Y,\cdot)-\mu^{\lambda}_{Y|X})\otimes h(X,\cdot)\Big)^*\Big]^p &\preccurlyeq 2^{2p-1}||h(X,\cdot)||_K^{2p}\Big[\Big((L(Y,\cdot)-\mu_{Y|X})\otimes(L(Y,\cdot)-\mu_{Y|X})\Big)^p \qquad (37)\\ &+\Big((\mu^{\lambda}_{Y|X}-\mu_{Y|X})\otimes(\mu^{\lambda}_{Y|X}-\mu_{Y|X})\Big)^p\Big] \qquad (38)\end{aligned}$$

Hence, we have four terms to consider. We begin first with the RHS of (35):

$$\begin{aligned}\mathbb{E}[||L(Y,\cdot)-\mu_{Y|X}||_L^{2p}\Big(h(X,\cdot)\otimes h(X,\cdot)\Big)^p] &\preccurlyeq \mathbb{E}[||L(Y,\cdot)-\mu_{Y|X}||_L^{2p}||h(X,\cdot)||_K^{2(p-1)}\Big(h(X,\cdot)\otimes h(X,\cdot)\Big)] \\ &= \mathbb{E}_X\Big[\mathbb{E}_{Y|X}\Big[||L(Y,\cdot)-\mu_{Y|X}||_L^{2p}\Big]\cdot||h(X,\cdot)||_K^{2(p-1)}\Big(h(X,\cdot)\otimes h(X,\cdot)\Big)\Big] \\ &\preccurlyeq \frac{R^{2p-2}(2p)!\sigma^2}{2}\mathbb{E}_X\Big[||h(X,\cdot)||_K^{2(p-1)}\Big(h(X,\cdot)\otimes h(X,\cdot)\Big)\Big] \qquad (39)\\ &\preccurlyeq \frac{||k^\alpha||_\infty^{2(p-1)}R^{2p-2}(2p)!\sigma^2}{2\lambda^{\alpha(p-1)}}\mathbb{E}_X\Big[h(X,\cdot)\otimes h(X,\cdot)\Big] \qquad (40)\\ &= \frac{||k^\alpha||_\infty^{2(p-1)}R^{2p-2}(2p)!\sigma^2}{2\lambda^{\alpha(p-1)}}C_{XX}(C_{XX}+\lambda)^{-1} \qquad (41)\end{aligned}$$

where we have taken the trace of both sides in Assumption 4 to obtain (39) and have applied Lemma D.3 to obtain (40). By a similar reasoning, we have:

$$\mathbb{E}\Big[||h(X,\cdot)||_L^{2p}\Big((L(Y,\cdot)-\mu_{Y|X})\otimes(L(Y,\cdot)-\mu_{Y|X})\Big)^p\Big] \preccurlyeq \frac{||k^\alpha||_\infty^{2p-2}R^{2p-2}(2p)!\mathcal{N}(\lambda)V}{2\lambda^{(p-1)\alpha}} \qquad (42)$$

for the RHS of (37) after again applying Assumption 4. Note the only difference between the adjoint moment in (42) and (41) is that we have taken the trace of $C_{XX}(C_{XX}+\lambda)^{-1}$ ($\mathcal{N}(\lambda)$) in the former instead of $\mathrm{tr}(V) = \sigma^2$). Now, for (36), we have:

$$\begin{aligned}\mathbb{E}\Big[\Big(||\mu_{Y|X}-\mu^{\lambda}_{Y|X}||_L^2 h(X,\cdot)\otimes h(X,\cdot)\Big)^p\Big] &= \mathbb{E}\Big[||\mu_{Y|X}-\mu^{\lambda}_{Y|X}||^{2p}||h(X,\cdot)||_K^{2p-2}\Big(h(X,\cdot)\otimes h(X,\cdot)\Big)\Big] \\ &\preccurlyeq \frac{M(\lambda)^{2p}||k^\alpha||_\infty^{2p-2}}{\lambda^{(p-1)\alpha}}C_{XX}(C_{XX}+\lambda)^{-1} \\ &\preccurlyeq \frac{(2p)!(M(\lambda))^{2p}||k^\alpha||_\infty^{2p-2}}{2\lambda^{(p-1)\alpha}}C_{XX}(C_{XX}+\lambda)^{-1}\end{aligned}$$

where we recall the definition of $M(\lambda)$ from Lemma 6. Finally for (38)

$$\mathbb{E}\Big[||\mu^{\lambda}_{Y|X} - \mu_{Y|X}||_L^{2(p-1)}||h(X,\cdot)||_K^{2p}\Big((\mu^{\lambda}_{Y|X} - \mu_{Y|X}) \otimes (\mu^{\lambda}_{Y|X} - \mu_{Y|X})\Big)\Big] \preccurlyeq \frac{(2p)!M(\lambda)^{2(p-1)}||k^{\alpha}||_{\infty}^{2p}}{2\lambda^{p\alpha}} \cdot \mathbb{E}\Big[(\mu^{\lambda}_{Y|X} - \mu_{Y|X}) \otimes (\mu^{\lambda}_{Y|X} - \mu_{Y|X})\Big]$$

Let $Q = M(\lambda) \vee R$ and $\rho_\lambda = \mathbb{E}\Big[(\mu_{Y|X} - \mu^{\lambda}_{Y|X}) \otimes (\mu_{Y|X} - \mu^{\lambda}_{Y|X})\Big]$. Then, we can apply Lemma C.3 with $\tilde{V} = 2(\sigma^2 + M^2(\lambda))C_{XX}(C_{XX}+\lambda)^{-1}$, $\tilde{W} = 2\mathcal{N}(\lambda)V + \frac{2||k^{\alpha}||_{\infty}^2}{\lambda^{\alpha}}p_\lambda$. Then, we have that, with probability $1-\delta$:

$$||(\hat{C}_{YX} - C_{YX}(C_{XX}+\lambda)^{-1}(\hat{C}_{XX}+\lambda))(C_{XX}+\lambda)^{-\frac{1}{2}}|| \leq \frac{16Q||k^{\alpha}||_{\infty}\beta(\delta)}{\lambda^{\frac{\alpha}{2}}n} + 8\sqrt{\frac{\eta\beta(\delta)}{n}}$$

where:

$$\rho_\lambda = \mathbb{E}\Big[(\mu^{\lambda}_{Y|X} - \mu_{Y|X}) \otimes (\mu^{\lambda}_{Y|X} - \mu_{Y|X})\Big]$$
$$\eta = \max\{(\sigma^2 + M^2(\lambda))||C_\nu||(||C_\nu|| + \lambda)^{-1}, ||\mathcal{N}(\lambda)V + \frac{||k^{\alpha}||_{\infty}^2}{\lambda^{\alpha}}\rho_\lambda||\}$$
$$\beta(\delta) = \log\Big(\frac{4((\sigma^2 + M^2(\lambda))\mathcal{N}(\lambda) + (\sigma^2\mathcal{N}(\lambda) + \frac{||k^{\alpha}||_{\infty}^2}{\lambda^{\alpha}}\mathbb{E}_X[||\mu^{\lambda}_{Y|X} - \mu_{Y|X}||_L^2]))}{\eta\delta}\Big)$$

The last term in (34) is bounded as follows:

$$||C_{XX}^{\frac{1-\gamma}{2}}(C_{XX}+\lambda)^{-\frac{1}{2}}|| \leq \sqrt{\sup_i \frac{\mu_i^{1-\gamma}}{\mu_i + \lambda}} \leq \lambda^{-\frac{\gamma}{2}}$$

Finally, for the middle term, we may follow the proof of Theorem 16 in Fischer and Steinwart (2020) exactly to obtain:

$$||(C_{XX}+\lambda)^{\frac{1}{2}}(\hat{C}_{XX}+\lambda)^{-1}(C_{XX}+\lambda)^{\frac{1}{2}}|| \leq 3$$

for $n \geq 8||k^{\alpha}||_{\infty}^2 \log(\delta^{-1})g_\lambda\lambda^{-\alpha}$ with probability $1-\delta$ (for brevity, we do not repeat this argument here) . Putting these together, we obtain our result. □

Proof of Theorem 5. We must first demonstrate there exists a $n_0 \in \mathbb{N}$, such that for all $n \geq n_0$, $n \geq 8||k^{\alpha}||_{\infty}^2\log(\delta^{-1})g_{\lambda_n}\lambda_n^{-\alpha}$ in order apply the result in Theorem 7. Since $\lambda_n \to 0$, we can let $\lambda_n \leq \min\{1, ||C_\nu||\}$, from which we obtain:

$$\begin{aligned}
\frac{8||k^{\alpha}||_{\infty}^2\log(\delta^{-1})g_{\lambda_n}\lambda_n^{-\alpha}}{n} &= \frac{8||k^{\alpha}||_{\infty}^2\log(\delta^{-1})\lambda_n^{-\alpha}}{n} \cdot \log\Big(2e\mathcal{N}(\lambda_n)\frac{||C_\nu|| + \lambda_n}{||C_\nu||}\Big) \\
&\leq \frac{8||k^{\alpha}||_{\infty}^2\log(\delta^{-1})\lambda_n^{-\alpha}}{n} \cdot \log 4M_1e\lambda_n^{-p} \qquad (43) \\
&= \frac{8||k^{\alpha}||_{\infty}^2\log 4M_1e \cdot \log(\delta^{-1})\lambda_n^{-\alpha}}{n} + \frac{8p||k^{\alpha}||_{\infty}^2\log(\delta^{-1})\lambda_n^{-\alpha}\log\lambda_n^{-1}}{n}
\end{aligned}$$

where (43) follows from Lemma D.4. Thus, in order to demonstrate $\frac{8||k^{\alpha}||_{\infty}^2\log(\delta^{-1})g_{\lambda_n}\lambda_n^{-\alpha}}{n} \to 0$, it is sufficient to show $\frac{\lambda_n^{-\alpha}\log\lambda_n^{-1}}{n} \to 0$. This follows from the fact that:

$$\frac{\lambda_n^{-\alpha}\log\lambda_n^{-1}}{n} \asymp \frac{(\log n)^{1-\frac{r\alpha}{\max\{\alpha,\beta+p\}}}}{n^{1-\frac{\alpha}{\max\{\alpha,\beta+p\}}}}$$

after substituting for λ_n, and observing that $\frac{(\log n)^{1-\frac{r\alpha}{\max\{\alpha,\beta+p\}}}}{n^{1-\frac{\alpha}{\max\{\alpha,\beta+p\}}}} \to 0$ as $n \to \infty$ since $r > 1$. We now estimate each term in (10). We first have that:

$$\begin{aligned}\beta(\delta) &= \log\Big(\frac{4((\sigma^2+M^2(\lambda_n))\mathcal{N}(\lambda_n)+\sigma^2\mathcal{N}(\lambda_n)+\frac{||k^\alpha||_\infty^2}{\lambda^\alpha}\mathbb{E}_X[||\mu^\lambda_{Y|X}-\mu_{Y|X}||_L^2])}{\eta\delta}\Big)\\
&\le \log\Big(\frac{4(\sigma^2+M^2(\lambda_n))\mathcal{N}(\lambda_n)}{(\sigma^2+M^2(\lambda))||C_\nu||(||C_\nu||+\lambda)^{-1}}+\frac{4\sigma^2\mathcal{N}(\lambda_n)+\frac{4||k^\alpha||_\infty^2}{\lambda^\alpha}\mathbb{E}_X[||\mu^\lambda_{Y|X}-\mu_{Y|X}||_L^2]}{||\mathcal{N}(\lambda_n)V+\frac{||k^\alpha||_\infty^2}{\lambda_n^\alpha}\rho_{\lambda_n}||}\Big)-\log\delta \qquad (44)\\
&\le \log\Big(\frac{4\sigma^2\mathcal{N}(\lambda_n)}{\mathcal{N}(\lambda_n)||V||}+\frac{4\mathcal{N}(\lambda_n)}{||C_\nu||(||C_\nu||+\lambda)^{-1}}+\frac{\frac{4||k^\alpha||_\infty^2}{\lambda^\alpha}\mathbb{E}_X[||\mu^\lambda_{Y|X}-\mu_{Y|X}||_L^2]}{\mathcal{N}(\lambda_n)||V||}\Big)-\log\delta \qquad (45)\\
&\le \log\Big(N_1\lambda_n^{-p}+N_2\lambda_n^{\beta-\alpha}\Big)-\log\delta \qquad (46)\end{aligned}$$

for $N_1=\frac{4M_1||C_\nu+\lambda||}{||C_\nu||}+\frac{4\sigma^2}{||V||}$ and $N_2=\frac{4||k^\alpha||_\infty^2 DB^2}{M_2||V||}$. Note, (44) follows from the definition of η and the fact that $\frac{\mathrm{tr}(A+B)}{\max\{||A||,||B||\}}\le\frac{\mathrm{tr}(A)}{||A||}+\frac{\mathrm{tr}(B)}{||B||}$ for any self-adjoint operators A and B; (45) follows from the sublinearity of the operator norm and the fact that $\rho_{\lambda_n}\succcurlyeq 0$; and (46) follows from applying Lemmas D.4, D.5, (6), and noting that we can restrict $\lambda_n\le 1$ (which allows the absorption of the constant term $\frac{4\sigma^2}{||V||}$ into N_1). Thus, it follows that $\beta(\delta)\le N_3\log(\delta^{-1}\cdot\lambda_n^{-\max\{\alpha-\beta,p\}})$ for some $N_3>0$. Moreover, we have that:

$$\begin{aligned}Q &= M(\lambda)\vee R\\
&\le (\tilde{C}+||k^\alpha||_\infty B)\lambda_n^{-\frac{(\alpha-\beta)_+}{2}}\vee R\\
&\le N_4\lambda_n^{-\frac{(\alpha-\beta)_+}{2}} \qquad (47)\end{aligned}$$

for $N_4=\max\{\tilde{C}+||k^\alpha||_\infty B,R\}$ (where the penultimate step follows from applying (7) and the final step follows since we can again assume $\lambda_n\le 1$). We also have:

$$\begin{aligned}\eta &= \max\{(\sigma^2+M^2(\lambda))||C_\nu||(||C_\nu||+\lambda)^{-1},||\mathcal{N}(\lambda)V+\frac{||k^\alpha||_\infty^2}{\lambda^\alpha}\rho_\lambda||\}\\
&\le \max\{(\sigma^2+N_4^2\lambda_n^{-(\alpha-\beta)_+})||C_\nu||(||C_\nu||+\lambda)^{-1},M_1\lambda_n^{-p}||V||+B^2||k^\alpha||_\infty^2\lambda_n^{-\alpha}\lambda_n^\beta\} \qquad (48)\\
&\le N_5\lambda_n^{-\max\{p,\alpha-\beta\}} \qquad (49)\end{aligned}$$

for some $N_5>0$. Note, in (48), we have applied (7), (8), and Lemma D.4. Thus, we have that:

$$\begin{aligned}||\hat{C}_{Y|X}-C^\lambda_{Y|X}||_\gamma &\le 3\lambda_n^{-\frac{\gamma}{2}}\Big(\frac{16Q||k^\alpha||_\infty\beta(\delta)}{\lambda_n^{\frac{\alpha}{2}}n}+8\sqrt{\frac{\eta\beta(\delta)}{n}}\Big)\\
&\le 24\lambda_n^{-\frac{\gamma}{2}}\Big(\frac{2N_4||k^\alpha||_\infty\beta(\delta)}{n\lambda_n^{\frac{\alpha+(\alpha-\beta)_+}{2}}}+\sqrt{\frac{N_5\beta(\delta)}{n\lambda_n^{\max\{p,\alpha-\beta\}}}}\Big) \qquad (50)\\
&\le 24\lambda_n^{-\frac{\gamma}{2}}\sqrt{\frac{\beta(\delta)}{n\lambda_n^{\max\{p,\alpha-\beta\}}}}\Big(2N_4||k^\alpha||_\infty\sqrt{\frac{\beta(\delta)}{n\lambda_n^{\alpha+(\alpha-\beta)_+-\max\{p,\alpha-\beta\}}}}+\sqrt{N_5}\Big)\end{aligned}$$

where (50) follows from (49) and (47). Then, like in the proof of Theorem 1 in Fischer and Steinwart (2020), we consider the inner factor for our two parameter regimes. When $p<\alpha-\beta$, we have that $\lambda_n\asymp\left(\frac{n}{\log^r n}\right)^{-\frac{1}{\alpha}}$, and thus:

$$\frac{\beta(\delta)}{n\lambda_n^{\alpha+(\alpha-\beta)_+-\max\{p,\alpha-\beta\}}}\le\frac{N_3\log(\delta^{-1}\cdot\lambda_n^{\beta-\alpha})}{n\lambda_n^\alpha}=\log(\delta^{-1})\cdot\mathcal{O}\Big(\frac{\log n}{\log^r n}\Big)$$

Thus, since $r>1$, it follows that $\frac{\beta(\delta)}{n\lambda_n^{\alpha+(\alpha-\beta)_+-\max\{p,\alpha-\beta\}}}\to 0$ when $p<\alpha-\beta$. Similarly, when $p>\alpha-\beta$, we have that $\lambda_n\asymp\left(\frac{n}{\log^r n}\right)^{-\frac{1}{\beta+p}}$, and:

$$\frac{\beta(\delta)}{n\lambda_n^{\alpha+(\alpha-\beta)_+-\max\{p,\alpha-\beta\}}}\le\frac{N_3\log(\delta^{-1}\cdot\lambda_n^{-p})}{n\lambda_n^{\alpha+(\alpha-\beta)_+-p}}=\log(\delta^{-1})\cdot\mathcal{O}\Big((\log n)^{1-\frac{r\alpha+r(\alpha-\beta)_+-rp}{\beta+p}}n^{-\Big(1-\frac{\alpha+(\alpha-\beta)_+-p}{\beta+p}\Big)}\Big)$$

Thus, since $1 - \frac{\alpha+(\alpha-\beta)_+-p}{\beta+p} > 0$ by the assumption that $p > \alpha - \beta$, we again have $\frac{\beta(\delta)}{n\lambda_n^{\alpha+(\alpha-\beta)_+-\max\{p,\alpha-\beta\}}} \to 0$ when $p > \alpha - \beta$. Hence, we can bound, $\sqrt{\frac{\beta(\delta)}{n\lambda_n^{\alpha+(\alpha-\beta)_+-\max\{p,\alpha-\beta\}}}}$ by $N_6^2\sqrt{\log(\delta^{-1})}$ for some constant $N_6 > 0$. Thus, putting this all together and combining with the bias bound $||C^\lambda_{Y|X} - C_{Y|X}||_\gamma \leq B\lambda^{\frac{\beta-\gamma}{2}}$ in (9), we obtain by (5):

$$\begin{aligned}
||\hat{C}_{Y|X} - C_{Y|X}||_\gamma &\leq ||\hat{C}_{Y|X} - C^\lambda_{Y|X}||_\gamma + ||C^\lambda_{Y|X} - C_{Y|X}||_\gamma \\
&\leq 24\lambda_n^{-\frac{\gamma}{2}}\sqrt{\frac{\beta(\delta)}{n\lambda_n^{\max\{p,\alpha-\beta\}}}}\left(2N_4||k^\alpha||_\infty\sqrt{\frac{\beta(\delta)}{n\lambda_n^{\alpha+(\alpha-\beta)_+-\max\{p,\alpha-\beta\}}}} + N_5\right) + B\lambda^{\frac{\beta-\gamma}{2}} \\
&\leq 24\lambda_n^{-\frac{\gamma}{2}}\sqrt{\frac{\log(\delta^{-1})\cdot\beta(\delta)}{n\lambda_n^{\max\{p,\alpha-\beta\}}}}\left(2N_4N_6||k^\alpha||_\infty + N_5\right) + B\lambda^{\frac{\beta-\gamma}{2}} \\
&= \lambda_n^{\frac{\beta-\gamma}{2}}\left(N_7\sqrt{\frac{\log(\delta^{-1})\beta(\delta)}{n\lambda_n^{\max\{\beta+p,\alpha\}}}} + B\right)
\end{aligned}$$

where we have set $N_7 = 24(2N_4N_6||k^\alpha||_\infty + N_5)$. Now, noting that $\lambda_n \asymp \left(\frac{n}{\log^r n}\right)^{-\frac{1}{\max\{\alpha,\beta+p\}}}$ by definition, we observe that:

$$\begin{aligned}
\frac{\beta(\delta)}{n\lambda_n^{\max\{\beta+p,\alpha\}}} &\leq \frac{\log(\delta^{-1}\cdot\lambda_n^{-\max\{\alpha-\beta,p\}})}{n\lambda_n^{\max\{\beta+p,\alpha\}}} \\
&\leq \frac{\log(\delta^{-1}) + \log(\lambda_n^{-\max\{\alpha-\beta,p\}})}{n\lambda_n^{\max\{\beta+p,\alpha\}}} \\
&\leq \frac{\log(\delta^{-1}) + \log(\lambda_n^{-\max\{\alpha,\beta+p\}})}{n\lambda_n^{\max\{\beta+p,\alpha\}}} \qquad (51) \\
&\leq \frac{\log(\delta^{-1}) + \log n - \log\log^r n}{n\cdot\frac{\log^r n}{n}} \\
&\leq \frac{\log(\delta^{-1}) + \log n}{\log^r n}
\end{aligned}$$

where (51) follows from the fact that $\lambda_n^{-\beta} \geq 1$ as $\lambda_n \to 0$. Hence, since $\delta < 1$ and $r > 1$, we have that $\frac{\beta(\delta)}{n\lambda_n^{\max\{\beta+p,\alpha\}}} = \mathcal{O}(\log(\delta^{-1}))$ as $n \to \infty$. Thus, we have, that there exists a $K > 0$ not depending on n or δ, such that:

$$||\hat{C}_{Y|X} - C_{Y|X}||_\gamma \leq K\log(\delta^{-1})\lambda_n^{\frac{\beta-\gamma}{2}}$$

with probability $1 - 2\delta$. □

C Concentration Bounds

Lemma C.1. Let $X_1, X_2, \ldots X_N$ be i.i.d self-adjoint operators on a Hilbert space $\mathcal{V}$, with:

$$\begin{aligned}
E[X_i] &= 0 \\
E[X_i^{2p}] &\preccurlyeq \frac{R^{2p-2}(2p)!}{2}V \quad \forall p \in \mathbb{N} \\
||V|| &= \sigma^2
\end{aligned}$$

where V is a trace-class operator. Let $\delta > 0$ and $\beta(\delta) = \log\left(\frac{4\text{tr}(V)}{\delta\sigma^2}\right)$. Then, for $t \geq \frac{2R}{N} + \frac{2^{\frac{3}{4}}\sigma}{\sqrt{N}}$ we have that:

$$\Big|\Big|\frac{1}{N}\sum_{i=1}^N X_i\Big|\Big| \leq \frac{4R\beta(\delta)}{N} + 2\sigma\sqrt{\frac{2\beta(\delta)}{N}}$$

with probability $1 - \delta$.

Proof. We first note that for odd $p \geq 1$ and $y \in \mathcal{V}$, we have $\mathbb{E}[\langle y, X_i^p y\rangle_{\mathcal{V}}] = \mathbb{E}[\langle X_i y, X_i^{p-1} y\rangle_{\mathcal{V}}] \leq \sqrt{\mathbb{E}[\langle y, X_i^2 y\rangle_{\mathcal{V}}]\mathbb{E}[\langle y, X_i^{2p-2} y\rangle_{\mathcal{V}}]} \leq \sqrt{\frac{R^{2p-4}(2p-2)!}{2}}\langle y, Vy\rangle \leq \frac{(2R)^{p-2}(p-1)!}{2}\langle y, \sqrt{8}Vy\rangle$. Thus, letting $S = 2R$, we have, by the usual construction:

$$\begin{aligned}
\mathbb{E}[e^{\theta X}] &= I + \sum_{j=2}^{\infty} \frac{\mathbb{E}[(\theta X)^j]}{j!} \\
&\preccurlyeq I + \sum_{j=2}^{\infty} \frac{\sqrt{8}(\theta S)^j V}{2S^2} \\
&= I + \frac{\sqrt{8}\theta^2 V}{2}\sum_{j=0}^{\infty}(\theta S)^j \\
&= I + \frac{\sqrt{8}\theta^2 V}{2(1-\theta S)} \\
&\preccurlyeq \exp\Big(\frac{\sqrt{8}\theta^2 V}{2(1-\theta S)}\Big) \qquad (52)
\end{aligned}$$

where the first equality follows by assumption. Let $g(\theta) = \frac{\theta^2}{2(1-\theta S)}$. Equipped with this result, we then have:

$$\begin{aligned}
P\Big(\Big|\Big|\frac{1}{N}\sum_{i=1}^{N} X_i\Big|\Big| > t\Big) &\leq \frac{\mathbb{E}\Big[\Big|\Big|e^{\frac{\theta}{N}\sum_{i=1}^N X_i} - \frac{\theta}{N}\sum_{i=1}^N X_i - I\Big|\Big|\Big]}{e^{\theta t} - \theta t - 1} \\
&\leq \frac{\mathbb{E}[\operatorname{tr}(e^{\frac{\theta}{N}\sum_{i=1}^N X_i} - I)]}{e^{\theta t} - \theta t - 1} \\
&\leq \frac{\operatorname{tr}(\exp\Big(\sum_{i=1}^N \log \mathbb{E}[e^{\frac{\theta}{N}X_i}]\Big) - I)}{e^{\theta t} - \theta t - 1} \qquad (53) \\
&\leq \frac{\operatorname{tr}(e^{\sqrt{8}Ng(N^{-1}\theta)V} - I)}{e^{\theta t} - \theta t - 1} \qquad (54) \\
&\leq \frac{\operatorname{tr}(V)}{||V||} \cdot \frac{e^{\sqrt{8}Ng(N^{-1}\theta)||V||} - 1}{e^{\theta t} - \theta t - 1} \qquad (55) \\
&\leq \frac{\operatorname{tr}(V)}{||V||} \cdot \frac{e^{\theta t}e^{\sqrt{8}Ng(N^{-1}\theta)||V|| - \theta t}}{e^{\theta t} - \theta t - 1}
\end{aligned}$$

where (53) follows from the iterative application of the operator concavity of $\operatorname{tr}(\exp(A + \log X))$ in X (see e.g. Tropp (2015)), (54) follows from applying (52), and (55) follows from Lemma 7.5.1 in Tropp (2015) and the observation that $f(t) = e^{\theta t} - 1$ is convex with $f(0) = 0$. Applying the bound $\frac{e^a}{e^a - a - 1} \leq 1 + \frac{3}{a^2}$ for $a \geq 0$ (see e.g. the proof of Theorem 7.7.1 in Tropp (2015)), we obtain:

$$\begin{aligned}
P\Big(\Big|\Big|\frac{1}{N}\sum_{i=1}^{N} X_i\Big|\Big| > t\Big) &\leq \frac{\operatorname{tr}(V)}{\sigma^2}\Big(1 + \frac{3}{\theta^2 t^2}\Big)e^{\sqrt{8}Ng(N^{-1}\theta)\sigma^2 - \theta t} \\
&\leq \frac{\operatorname{tr}(V)}{\sigma^2}\Big(1 + \frac{3(\sqrt{8}\sigma^2 + 2Rt)^2}{N^2 t^4}\Big)\exp\Big(-\frac{Nt^2}{2(\sqrt{8}\sigma^2 + 2Rt)}\Big)
\end{aligned}$$

after setting $\theta = \frac{Nt}{\sqrt{8}\sigma^2 + 2Rt}$, noting $S = 2R$, and observing that $\sqrt{8}N\sigma^2 g(N^{-1}\theta) - \theta t = \frac{Nt^2}{2(\sqrt{8}\sigma^2 + 2Rt)} - \frac{Nt^2}{\sqrt{8}\sigma^2 + 2Rt} \leq -\frac{Nt^2}{2(\sqrt{8}\sigma^2 + 2Rt)}$. We consider only the case where $Nt^2 \geq \sqrt{8}\sigma^2 + 2Rt$, noting like in Theorem 7.7.1 in Tropp (2015) that the Chernoff bound above is typically vacuous when this restriction is violated. Solving this quadratic inequality, we obtain the more amenable expression: $t \geq \frac{R}{N} + \sqrt{\frac{R^2}{N^2} + \frac{\sqrt{8}\sigma^2}{N}}$. Thus, applying the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we have that for $t \geq \frac{2R}{N} + \frac{8^{0.25}\sigma}{\sqrt{N}}$:

$$P\Big(\Big|\Big|\frac{1}{N}\sum_{i=1}^{N} X_i\Big|\Big| > t\Big) \leq \frac{4\operatorname{tr}(V)}{\sigma^2}\exp\Big(-\frac{Nt^2}{2(\sqrt{8}\sigma^2 + 2Rt)}\Big)$$

The result follows from setting the RHS equal to δ, solving for t using the quadratic formula, applying the triangle inequality to this solution, and noting that $\sqrt{2} \leq 2$, we obtain our result. $\square$

Remark. We emphasize the qualification $t \geq \frac{2R}{N} + \frac{2^{\frac{3}{4}}\sigma}{\sqrt{N}}$ in Lemma C.1 is not very restrictive as the derived Chernoff bound is typically vacuous when this restriction is violated (and therefore is avoided by choosing sufficiently small δ. For brevity, we therefore omit this restriction in the below generalizations).

Lemma C.2. Let $X_1, X_2, \ldots X_N$ be i.i.d operators from $\mathcal{V}$ to $\mathcal{W}$, with:

$$
\begin{aligned}
E[X_i] &= 0 \\
E[(X_i^* X_i)^p] &\preccurlyeq \frac{R^{2p-2}(2p)!}{2} V \quad \forall p \in \mathbb{N} \\
E[(X_i X_i^*)^p] &\preccurlyeq \frac{R^{2p-2}(2p)!}{2} W \quad \forall p \in \mathbb{N} \\
\max\{||V||, ||W||\} &= \sigma^2
\end{aligned}
$$

where V and W are trace-class operators on $\mathcal{V}$ and $\mathcal{W}$, respectively. Let $\delta > 0$ and $\beta(\delta) = \log\left(\frac{4\text{tr}(V+W)}{\delta\sigma^2}\right)$. Then, we have that:

$$
\Big|\Big|\frac{1}{N}\sum_{i=1}^N X_i\Big|\Big| \leq \frac{4R\beta(\delta)}{N} + 2\sigma\sqrt{\frac{2\beta(\delta)}{N}}
$$

with probability $1 - \delta$

Proof. We generalize the approach from Tropp (2015) — namely we define the "dilation" operator T_i on $\mathcal{W} \times \mathcal{V}$ that maps $T_i : (w, v) \mapsto (X_i^* v, X_i w)$, where X_i^* denotes the adjoint of X_i. Then, it is easy to see that T_i is self-adjoint. Moreover, we have that $\Big|\Big|\sum_i T_i\Big|\Big| = \Big|\Big|\sum_i X_i\Big|\Big|$. Thus, we can apply Lemma C.1 to the T_i. Indeed, observe that $T_i^2 : (w, v) \mapsto (X_i X_i^* w, X_i^* X_i v)$, and hence we have that $\mathbb{E}[T_i^{2p}] : (w, v) \mapsto (\mathbb{E}[(XX^*)^p]w, \mathbb{E}[(X^*X)^p]v)$. From this, we obtain that $\mathbb{E}[T_i^{2p}] \preccurlyeq \frac{R^{2p-2}(2p)!}{2} U$, where $U : (w, v) \mapsto (Wv, Vv)$. Our result then follows from Lemma C.1. $\square$

Lemma C.3. Let $\mathcal{H}_K$ and $\mathcal{H}_L$ be RKHSs on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $X_1, X_2, \ldots X_N$ be i.i.d rank-1 operators from $\mathcal{H}_K$ to $\mathcal{H}_L$, with:

$$
\begin{aligned}
E[(X_i^* X_i)^p] &\preccurlyeq \frac{R^{2p-2}(2p)!}{2} V \quad \forall p \in \mathbb{N} \\
E[(X_i X_i^*)^p] &\preccurlyeq \frac{R^{2p-2}(2p)!}{2} W \quad \forall p \in \mathbb{N} \\
\max\{||V||, ||W||\} &= \sigma^2
\end{aligned}
$$

where V and W are trace-class operators on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $\delta > 0$ and $\beta(\delta) = \log\left(\frac{4\text{tr}(V+W)}{\delta\sigma^2}\right)$. Then, we have that:

$$
\Big|\Big|\frac{1}{N}\sum_{i=1}^N X_i - \mathbb{E}[X]\Big|\Big| \leq \frac{8R\beta(\delta)}{N} + 4\sigma\sqrt{\frac{2\beta(\delta)}{N}}
$$

with probability $1 - \delta$

Proof. We observe that:

$$
\begin{aligned}
\mathbb{E}[((X_i - \mathbb{E}[X_i])^*(X_i - \mathbb{E}[X_i]))^p] &\preccurlyeq \mathbb{E}[||X_i - \mathbb{E}[X_i]||_{\text{HS}}^{2(p-1)}(X_i - \mathbb{E}[X_i])^*(X_i - \mathbb{E}[X_i])] \\
&\preccurlyeq 2^{2p-1}\Big(\mathbb{E}[||X_i||_{\text{HS}}^{2(p-1)} X_i^* X_i] + ||\mathbb{E}[X_i]||_{\text{HS}}^{2(p-1)}\mathbb{E}[X_i]^*\mathbb{E}[X_i]\Big) \qquad (56) \\
&\preccurlyeq 4^p \mathbb{E}[||X_i||_{\text{HS}}^{2(p-1)} X_i^* X_i] \qquad (57) \\
&= 4^p \mathbb{E}[(X_i^* X_i)^p]
\end{aligned}
$$

where (56) and (57) follow from the convexity of $||X||_{\text{HS}}^{2(p-1)}||Xf||_{\mathcal{Y}}^2$ for any $f \in \mathcal{H}_K$ by Lemma D.1 and the definition of the semidefinite order, and the last step follows from the fact that X_i is rank-1 and hence $||X_i||_{\text{HS}}^{2p-2}X_i^*X_i = (X_i^*X_i)^p$. We can show a similar conclusion for $\mathbb{E}[((X_i - \mathbb{E}[X_i])(X_i - \mathbb{E}[X_i])^*)^p]$ from which the result follows by Lemma C.2. □

D Auxiliary Results

Lemma D.1. Let $\mathcal{H}_1, \mathcal{H}_2$ be Hilbert spaces, and let $\mathcal{L}_{\text{HS}}(\mathcal{H}_1, \mathcal{H}_2)$ denote the space of Hilbert-Schmidt operators from $\mathcal{H}_1$ to $\mathcal{H}_2$. Then, for any $y \in \mathcal{H}_1$, the functions $f : \mathcal{H}_1 \to \mathbb{R}^+$ and $g : \mathcal{L}_{\text{HS}}(\mathcal{H}_1, \mathcal{H}_2) \to \mathbb{R}^+$ given by $f(x) = ||x||_{\mathcal{H}_1}^{2p}\langle y, x\rangle_{\mathcal{H}_1}^2$ and $g(A) = ||A||_{\text{HS}}^{2p}||Ay||_{\mathcal{H}_2}^2$ are convex for all $p \geq 1$ in the semidefinite order.

Proof. We must show that for any $x, z \in \mathcal{H}_1$ and $X, Z \in \mathcal{L}_{\text{HS}}(\mathcal{H}_1, \mathcal{H}_2)$, the functions $\tilde{f}(t) = f(x + tz)$ and $\tilde{g}(t) = g(X + tZ)$ are convex in $t \in \mathbb{R}$. Observing that $\tilde{f}$ and $\tilde{g}$ can be expressed as: $\tilde{f}(t) = (||z||_{\mathcal{H}_1}^2 t^2 + 2t\langle x, z\rangle_{\mathcal{H}_1} + ||x||_{\mathcal{H}_1}^2)^p(\langle y, z\rangle_{\mathcal{H}_1} t + \langle y, x\rangle_{\mathcal{H}_1})^2$ and $\tilde{g}(t) = (||Z||_{\text{HS}}^2 t^2 + 2\text{tr}(X^* \circ Z)t + ||X||_{\text{HS}}^2)^p(||Zy||_{\mathcal{H}_2}^2 t^2 + 2t\langle Xy, Zy\rangle_{\mathcal{H}_2} + ||Xy||_{\mathcal{H}_2}^2)$, the claim can be readily verified by taking second derivatives. □

Corollary D.1.1. Let $\mathcal{H}$ be a Hilbert space, and let $\mathcal{L}_1(\mathcal{H})$ be the space of rank-1 linear operators on $\mathcal{H}$. Then, the operator-valued function $g : \mathcal{H} \to \mathcal{L}_1(\mathcal{H})$ given by $g(u) = (u \otimes u)^p$ is convex for any $p \geq 1$.

Proof. By the definition of the semidefinite order, we have g is convex iff the real-valued function $f_y(u) = ||u||^{2p-2}\langle y, u\rangle^2$ is convex for all $y \in \mathcal{H}$. The latter follows from Lemma D.1. □

Lemma D.2. Suppose Assumption 1 holds. Then, if $\beta > p$, there exists a constant $D > 0$ that depends only on β and p, such that:

$$\sum_{i=1}^{\infty}\Big(\frac{\mu_i^{\frac{\beta}{2}}}{\mu_i + \lambda}\Big)^2 \leq D\lambda^{\beta-p-2}$$

Proof. We have that:

$$\begin{aligned}
\sum_{i=1}^{\infty}\Big(\frac{\mu_i^{\frac{\beta}{2}}}{\mu_i + \lambda}\Big)^2 &= \sum_{i=1}^{\infty}\Big(\frac{\mu_i^{\frac{\beta}{2}-1}}{1 + \lambda\mu_i^{-1}}\Big)^2 \\
&\leq \int_0^{\infty}\Big(\frac{(c^{-p}x)^{-\frac{p^{-1}\beta}{2}+p^{-1}}}{1 + \lambda(C^{-p}x)^{p^{-1}}}\Big)^2 dx && (58) \\
&= \lambda^{\beta-p-2}\int_0^{\infty}\Big(\frac{(c^{-p}y)^{-\frac{p^{-1}\beta}{2}+p^{-1}}}{1 + (C^{-p}y)^{p^{-1}}}\Big)^2 dy && (59)
\end{aligned}$$

where (58) follows from Assumption 1, and (59) follows after making the substitution $\lambda^p x = y$. Now, observe that:

$$\begin{aligned}
\int_0^{\infty}\Big(\frac{(c^{-p}y)^{-\frac{p^{-1}\beta}{2}+p^{-1}}}{1 + (C^{-p}y)^{p^{-1}}}\Big)^2 dy &= \int_0^{\infty}\Big(\frac{(c^{-p}y)^{-\frac{p^{-1}\beta}{2}}}{(c/C) + (c^{-p}y)^{-p^{-1}}}\Big)^2 dy \\
&= \int_0^{1}\Big(\frac{(c^{-p}y)^{-\frac{p^{-1}\beta}{2}}}{(c/C) + (c^{-p}y)^{-p^{-1}}}\Big)^2 dy + \int_1^{\infty}\Big(\frac{(c^{-p}y)^{-\frac{p^{-1}\beta}{2}}}{(c/C) + (c^{-p}y)^{-p^{-1}}}\Big)^2 dy \\
&\leq \int_0^{1}\Big(\frac{c}{C}\Big)^{\frac{\beta}{2}-1} dy + \frac{C^2}{c^{2-\beta}}\int_1^{\infty} y^{-p^{-1}\beta} dy && (60) \\
&= D \\
&< \infty && (61)
\end{aligned}$$

where (60) follows from the fact that $\frac{\beta}{2} < 1$ and Lemma 25 in Fischer and Steinwart (2020), and the last line follows from $\beta > p$. □

The following two results, which are from Fischer and Steinwart (2020) (and were originally discussed in Steinwart and Scovel (2012)), characterizes the boundedness of the kernel and the "effective dimension". We include them here for completeness .

Lemma D.3. Suppose $||k^\alpha||_\infty < \infty$. Then, we have that:

$$||(C_{XX}+\lambda)^{-\frac{1}{2}}k(X,\cdot)||_K \leq \lambda^{-\frac{\alpha}{2}}||k^\alpha||_\infty$$

Proof. From definition, we have that:

$$\begin{aligned}(C_{XX}+\lambda)^{-\frac{1}{2}}k(X,\cdot) &= \sum_i \sqrt{\frac{\mu_i}{\mu_i+\lambda}}\cdot e_i(x)(\mu_i^{\frac{1}{2}}e_i)\\ &= \sum_i \sqrt{\frac{\mu_i^{1-\alpha}}{\mu_i+\lambda}}\cdot \mu^{\frac{\alpha}{2}}e_i(x)(\mu_i^{\frac{1}{2}}e_i)\end{aligned}$$

Thus:

$$||(C_{XX}+\lambda)^{-\frac{1}{2}}k(X,\cdot)||_K^2 \leq \Big(\max_i \frac{\mu_i^{1-\alpha}}{\mu_i+\lambda}\Big)\sum_i \mu_i^\alpha e_i^2(x) \leq \lambda^{-\alpha}||k^\alpha||_\infty^2$$

by Lemma 25 in Fischer and Steinwart (2020). ☐

Lemma D.4. Suppose Assumption 1 holds. Then, there exists a $M_1 > 0$ such that:

$$\mathcal{N}(\lambda) = \mathrm{tr}\Big(C_\nu(C_\nu+\lambda)^{-1}\Big) \leq M_1\lambda^{-p}$$

Proof. See Lemma 11 in Fischer and Steinwart (2020) ☐

Lemma D.5. Suppose Assumption 1 holds. Then, there exists a $M_2 > 0$ such that:

$$\mathcal{N}(\lambda) \geq M_2\lambda^{-p}$$

Proof. We have that:

$$\begin{aligned}\mathcal{N}(\lambda) &= \sum_{i=1}^\infty \frac{\mu_i}{\mu_i+\lambda}\\ &\geq \sum_{i=1}^\infty \frac{ci^{-p^{-1}}}{Ci^{-p^{-1}}+\lambda}\\ &\geq \int_1^\infty \frac{cx^{-p^{-1}}}{Cx^{-p^{-1}}+\lambda}dx\\ &= \int_1^\infty \frac{c}{C+\lambda x^{p^{-1}}}dx\\ &= \lambda^{-p}\int_1^\infty \frac{c}{C+y^{p^{-1}}}dy\end{aligned}$$

where the last line follows from making the substitution $y = \lambda^p x$. Then, our result follows from observing $\int_1^\infty \frac{c}{C+y^{p^{-1}}} = M_2 < \infty$ since $p < 1$ by assumption. ☐

Lemma D.6. Let $\mathcal{H}_K$ be a Gaussian RKHS over $\mathbb{R}^d$ with kernel $K(x,y) = \exp\Big(-\frac{||x-t||^2}{\sigma^2}\Big)$ for some $\sigma > 0$. Then, for every $\beta \in (0,1)$, $\mathcal{H}_K^\beta$ contains constant functions.

Proof. We only treat the one-dimensional case $d = 1$ and note that the more general case follows easily from the argument of Steinwart and Christmann (2008). By Minh (2010), we have that:

$$\mathcal{H}_K = \Big\{f = e^{-\frac{x^2}{\sigma^2}}\sum_{k=0}^\infty w_k x^k : ||f||_K^2 \equiv \sum_{k=0}^\infty \frac{w_k^2\sigma^{2k}k!}{2^k} < \infty\Big\}$$

Thus, we have by definition:

$$\mathcal{H}_K^\beta = \left\{ f = e^{-\frac{x^2}{\sigma^2}} \sum_{k=0}^{\infty} w_k x^k : ||f||^2_{\mathcal{H}^\beta} \equiv \sum_{k=0}^{\infty} \frac{w_k^2 \sigma^{2\beta k} (k!)^\beta}{2^{\beta k}} < \infty \right\}$$

For any $c \in \mathbb{R}$, we have that:

$$c = e^{-\frac{x^2}{\sigma^2}} \cdot e^{\frac{x^2}{\sigma^2}} c = e^{-\frac{x^2}{\sigma^2}} \cdot \sum_{k=0}^{\infty} \frac{c x^{2k}}{k! \sigma^{2k}}$$

Thus, we may define $w_{2k} = \frac{c}{k!\sigma^{2k}}$ and $w_{2k+1} = 0$ for $k \in \mathbb{N}$. Therefore, we have:

$$\begin{aligned} ||g_f||^2_{\mathcal{H}^\beta} &= c^2 \sum_{k=0}^{\infty} \frac{\sigma^{4\beta k} ((2k)!)^\beta}{4^{\beta k} (k!)^2 \sigma^{4k}} \\ &= c^2 \sum_{k=0}^{\infty} \frac{\sigma^{4(\beta-1)k} ((2k)!)^\beta}{4^{\beta k} (k!)^2} \end{aligned}$$

Let $a_k = \frac{\sigma^{4(\beta-1)k}((2k)!)^\beta}{4^{\beta k}(k!)^2}$. Now applying Stirling's formula $n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$, we have (for sufficiently large k):

$$a_k \sim e^{2(1-\beta)k} (\pi k)^{\frac{\beta-2}{2}} \sigma^{4(\beta-1)k} k^{2(\beta-1)k}$$

Noting that since $\beta < 1$, for $k \geq \frac{2^{\frac{1}{2(1-\beta)}} e}{\sigma^2}$, we have $e^{2(1-\beta)k} (\pi k)^{\frac{\beta-2}{2}} \sigma^{4(\beta-1)k} k^{2(\beta-1)k} \leq \left(\frac{1}{2}\right)^k$. Thus, we have $||g_f||^2_{\mathcal{H}^\beta} = c^2 \sum_{k=0}^{\infty} a_k < \infty$, and we obtain our result. □