Learning linear operators: Infinite-dimensional regression as a well-behaved non-compact inverse problem

Mattes Mollenhauer[1]

Nicole Mücke[2]

T. J. Sullivan[3,4]

Lifting Inference with Kernel Embeddings (LIKE23)
Universität Bern, CH
26–30 June 2023

[1]Freie Universität Berlin, DE
[2]Technische Universität Braunschweig, DE
[3]**University of Warwick, UK**
[4]**Alan Turing Institute, UK**

For all the details:

arXiv:2211.08875

Problem statement

- Let X and Y be Bochner square-integrable random variables taking values in separable Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively, i.e. $(X, Y) \in L^2(\mathbb{P}; \mathcal{X} \times \mathcal{Y})$.
- We aim to solve the following regression problem:

$$\text{minimise } \mathbb{E}[\|Y - \theta X\|_{\mathcal{Y}}^2] \equiv \|Y - \theta X\|_{L^2(\mathbb{P};\mathcal{Y})}^2 \text{ w.r.t. } \theta \in L(\mathcal{X}, \mathcal{Y}), \tag{RP}$$

  where $L(\mathcal{X}, \mathcal{Y})$ is the Banach space of bounded linear operators from $\mathcal{X}$ into $\mathcal{Y}$.
- In practice, we will only have data points (X_i, Y_i), $i = 1, \dots, n$ — so we must think about empirical approximation and regularisation.
- **Moral of the talk:** From a regularisation standpoint, (RP) is "just as hard" as finite-dimensional regression in reasonable settings.

Motivating instances of the problem

- If at least one of $\mathcal{X}$ and $\mathcal{Y}$ has infinite dimension, then so too does the search space $L(\mathcal{X}, \mathcal{Y})$, and so (RP) is an infinite-dimensional regression problem.
- We are particularly motivated by the case of infinite-dimensional $\mathcal{Y}$, exemplified by relevant applications in
  - functional linear regression with functional response (Ramsay and Silverman, 2005);
  - non-parametric regression with vector-valued kernels (Caponnetto and De Vito, 2007) (more on this in a moment);
  - the conditional mean embedding (Park and Muandet, 2020; Li et al., 2022);
  - and inference for Hilbertian time series (Bosq, 2000).

Example: (Vector-valued) kernel regression 1

- Let $\mathcal{E}$ be a second-countable locally compact Hausdorff space equipped with its Borel σ-algebra $\mathcal{B}_{\mathcal{E}}$, and let $\mathcal{X}$ be an RKHS of $\mathbb{R}$-valued functions on $\mathcal{E}$ with reproducing kernel $k\colon \mathcal{E}^2 \to \mathbb{R}$ and canonical feature map $\varphi\colon \mathcal{E} \to \mathcal{X}$.
- Assume further that $(\mathcal{E}, \mathcal{B}_{\mathcal{E}})$ is equipped with a probability measure μ, with a compact embedding operator $i\colon \mathcal{X} \hookrightarrow L^2(\mu)$ (e.g. Christmann and Steinwart, 2008, Section 4.3).
- Let $\mathcal{Y}$ be another separable real Hilbert space. Consider $\mathcal{G} := \{A\varphi(\,\cdot\,) \mid A \in S_2(\mathcal{X}, \mathcal{Y})\}$; this is a vv-RKHS of $\mathcal{Y}$-valued functions with operator-valued reproducing kernel

$$\begin{aligned} K\colon \mathcal{E}^2 &\to L(\mathcal{Y}) \\ (x, x') &\mapsto k(x, x')\,\mathrm{Id}_{\mathcal{Y}} \end{aligned}$$

and we have a bounded linear embedding operator

$$I := i \otimes \mathrm{Id}_{\mathcal{Y}}\colon \mathcal{G} \cong \mathcal{X} \otimes \mathcal{Y} \hookrightarrow L^2(\mu) \otimes \mathcal{Y} \cong L^2(\mu; \mathcal{Y}).$$

As the embedding $i\colon \mathcal{X} \hookrightarrow L^2(\mu)$ is compact, the embedding $I := i \otimes \mathrm{Id}_{\mathcal{Y}}$ is compact $\iff \dim \mathcal{Y} < \infty$.

Example: (Vector-valued) kernel regression 2

- We now consider an $\mathcal{E}$-valued random variable ξ with law $\mathscr{L}(\xi) =: \mu$ on $(\mathcal{E}, \mathcal{B}_{\mathcal{E}})$ and a $\mathcal{Y}$-valued random variable Y, both defined on a common probability space.
- The nonlinear kernel regression problem

$$\min_{F \in \mathcal{G}} \mathbb{E}[\| Y - F(\xi) \|_{\mathcal{Y}}^2]$$

is equivalent to the (Hilbert–Schmidt) version of the linear regression problem (RP) with $X := \varphi(\xi)$:

$$\min_{\theta \in S_2(\mathcal{X}, \mathcal{Y})} \mathbb{E}[\| Y - \theta\varphi(\xi) \|_{\mathcal{Y}}^2].$$

Problem reformulation

The problem with infinite-dimensional regression

- Infinite-dimensional linear regression does not necessarily admit a minimiser!
- Assuming a well-specified linear model, i.e. the existence of a bounded linear operator $\theta_\star \colon \mathcal{X} \to \mathcal{Y}$ such that

$$Y = \theta_\star X + \varepsilon$$

  with an exogeneous $\mathcal{Y}$-valued noise variable ε satisfying $\mathbb{E}[\varepsilon | X] = 0$, (RP) is equivalent to the operator factorisation problem

$$C_{YX} = \theta C_{XX}, \qquad \theta \in L(\mathcal{X}, \mathcal{Y}), \tag{OFP}$$

  where $C_{YX} \in L(\mathcal{X}, \mathcal{Y})$ and $C_{XX} \in L(\mathcal{X}, \mathcal{X})$ are the covariance operators (Baker, 1973) associated with X and Y.
- Solubility of (OFP) is related to a well-known set of *range inclusion* and *operator majorisation* conditions due to Douglas (1966) and the *Moore–Penrose pseudoinverse* (Engl et al., 1996).

Recap: Tensor products and covariance operators

- For $y \in \mathcal{Y}$ and $x \in \mathcal{X}$, $y \otimes x \in L(\mathcal{X}, \mathcal{Y})$ is the rank-one operator

$$\mathcal{X} \ni v \mapsto (y \otimes x)(v) := \langle x, v \rangle_{\mathcal{X}}\, y \in \mathcal{Y}.$$

- The Hilbert tensor product $\mathcal{Y} \otimes \mathcal{X}$ is defined to be the completion of the linear span of all such rank-one operators w.r.t. $\langle y \otimes x, y' \otimes x' \rangle_{\mathcal{Y} \otimes \mathcal{X}} := \langle y, y' \rangle_{\mathcal{Y}} \langle x, x' \rangle_{\mathcal{X}}$.
- Note that $\mathcal{Y} \otimes \mathcal{X}$ is isometric with $S_2(\mathcal{X}, \mathcal{Y})$, the space of Hilbert–Schmidt operators; and also $L^2(\mathbb{P}; \mathcal{X}) \cong L^2(\mathbb{P}; \mathbb{R}) \otimes \mathcal{X}$.
- The (uncentred) covariance operators (Baker, 1973) of Y with X, and of X with itself, are given by

$$\mathbb{C}\mathrm{ov}[Y, X] := C_{YX} := \mathbb{E}[Y \otimes X] \in S_1(\mathcal{X}, \mathcal{Y}) = \{\text{trace-class op's}\} \quad \text{and}$$
$$\mathbb{C}\mathrm{ov}[X, X] := C_{XX} := \mathbb{E}[X \otimes X] \in S_1(\mathcal{X}).$$

- Note that $C_{YX}^* = C_{XY}$, and so C_{XX} is self-adjoint.
- The covariance operators are the unique operators satisfying

$$\mathbb{E}[\langle y, Y \rangle_{\mathcal{Y}} \langle x, X \rangle_{\mathcal{X}}] = \langle y, C_{YX} x \rangle_{\mathcal{Y}} \quad \text{for all } x \in \mathcal{X}, y \in \mathcal{Y}.$$

From operator factorisation to a non-compact linear inverse problem

- The operator factorisation problem (OFP)

$$C_{YX} = \theta C_{XX}, \qquad \theta \in L(\mathcal{X}, \mathcal{Y}), \tag{OFP}$$

can be reformulated in terms of a (potentially ill-posed) linear inverse problem

$$A_{C_{XX}}[\theta] = C_{YX}, \quad \theta \in L(\mathcal{X}, \mathcal{Y}) \tag{IP}$$

based on the (generally non-compact) forward operator $A_{C_{XX}} \colon L(\mathcal{X}, \mathcal{Y}) \to L(\mathcal{X}, \mathcal{Y})$,

$$A_{C_{XX}}[\theta] \coloneqq \theta C_{XX}.$$

- We call the operator $A_{C_{XX}}$ the precomposition operator associated with C_{XX}.
- Even in the misspecified case, the solution to the inverse problem (IP) still characterises the minimiser of the linear regression problem (RP)!

Spectral theory and regularisation

Naïve solution of the inverse problem

- The standard, naïve thing to do at this point would be to solve (IP)

$$A_{C_{XX}}[\theta] = C_{YX}, \quad \theta \in L(\mathcal{X}, \mathcal{Y})$$

using the Moore–Penrose pseudoinverse of $A_{C_{XX}}$:

$$\theta = A_{C_{XX}}^{\dagger}[C_{YX}].$$

- The problem is that $\dim \mathcal{Y} = \infty \implies A_{C_{XX}}$ is non-compact, in which case we have no good off-the-shelf spectral theory for $A_{C_{XX}}$, no pseudoinverse, etc.
- Fortunately, we can build a decent spectral theory for $A_{C_{XX}}$ if we focus on the Hilbert–Schmidt setting: we restrict the search to $\theta \in S_2(\mathcal{X}, \mathcal{Y})$ and use the fact that

$$A_{C_{XX}} \colon S_2(\mathcal{X}, \mathcal{Y}) \to S_2(\mathcal{X}, \mathcal{Y}).$$

Theorem 1 (Spectral decomposition)

Let $C \in S_2(\mathcal{X})$ be self-adjoint with spectral decomposition

$$C = \sum_{\lambda \in \sigma_{\mathrm{p}}(C)} \lambda P_{\mathrm{eig}_\lambda(C)} ,$$

where $P_{\mathrm{eig}_\lambda(C)} \colon \mathcal{X} \to \mathcal{X}$ is orthogonal projection onto $\mathrm{eig}_\lambda(C)$ and the above series expression converges in operator norm. Then the non-compact induced precomposition operator A_C on $S_2(\mathcal{X}, \mathcal{Y})$ has pure point spectrum and the spectral decomposition

$$A_C = \sum_{\lambda \in \sigma_{\mathrm{p}}(C)} \lambda P_{\mathcal{Y} \otimes \mathrm{eig}_\lambda(C)},$$

where $P_{\mathcal{Y} \otimes \mathrm{eig}_\lambda(C)} \colon S_2(\mathcal{X}, \mathcal{Y}) \to S_2(\mathcal{X}, \mathcal{Y})$ is orthogonal projection onto $\mathcal{Y} \otimes \mathrm{eig}_\lambda(C)$ and the above series converges in operator norm.

Spectral theory for precomposition operators 2

Corollary 2 (Compatibility with functional calculus)

Let $C = \sum_{\lambda \in \sigma_{\mathrm{p}}(C)} \lambda P_{\mathrm{eig}_\lambda(C)} \in S_2(\mathcal{X})$ be self-adjoint. If $g \colon \mathbb{R} \to \mathbb{R}$ is extended to act on self-adjoint Hilbert space operators with pure point spectrum in terms of their spectral decompositions via

$$g(C) := \sum_{\lambda \in \sigma_{\mathrm{p}}(C)} g(\lambda) P_{\mathrm{eig}_\lambda(C)},$$

then A_C as an operator on $S_2(\mathcal{X}, \mathcal{Y})$ satisfies

$$A_{g(C)} = g(A_C) = \sum_{\lambda \in \sigma_{\mathrm{p}}(C)} g(\lambda) P_{\mathcal{Y} \otimes \mathrm{eig}_\lambda(C)},$$

We will use this with $g = g_\alpha$ being some approximation — e.g. Tikhonov, spectral cutoff, ...— to the 'ideal' inverse $g(\lambda) = \lambda^{-1}$, yielding a regularised population solution to (IP):

$$\theta_\alpha := g_\alpha(A_{C_{XX}})[C_{YX}] = C_{YX} g_\alpha(C_{XX}).$$

Terminology for regularisation

- A family of functions $g_\alpha \colon [0, \infty) \to \mathbb{R}$, indexed by a regularisation parameter $\alpha > 0$, is a spectral regularisation strategy (Engl et al., 1996) if

(R1) $\sup_{\lambda \in [0,\infty)} |\lambda g_\alpha(\lambda)| \leqslant D$ for some constant D,

(R2) $\sup_{\lambda \in [0,\infty)} |1 - \lambda g_\alpha(\lambda)| \leqslant \gamma_0$ for some constant γ_0, and

(R3) $\sup_{\lambda \in [0,\infty)} |g_\alpha(\lambda)| < B\alpha^{-1}$, for some constant B.

- We write $r_\alpha(\lambda) \coloneqq 1 - \lambda g_\alpha(\lambda)$ for the residual associated to the regularisation scheme g_α.
- The qualification of g_α is the maximal q such that

$$\sup_{\lambda \in [0,\infty)} \lambda^q |r_\alpha(\lambda)| \equiv \sup_{\lambda \in [0,\infty)} \lambda^q |1 - \lambda g_\alpha(\lambda)| \leqslant \gamma_q \alpha^q$$

  for some constant γ_q which does not depend on α.
- Such assumptions are also common in learning theory (see e.g. Bauer et al., 2007; Gerfo et al., 2008; Dicker et al., 2017; Blanchard and Mücke, 2018).

Regularised empirical solutions

Empirical solutions

- X and Y are in practice only accessible through sample pairs $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ for $i = 1, \ldots, n$.
- For simplicity, we assume that these sample pairs are obtained i.i.d. from the joint law of (X, Y).
- We define the empirical covariance operators by

$$\widehat{C}_{XX} := \frac{1}{n} \sum_{i=1}^{n} X_i \otimes X_i \text{ and } \widehat{C}_{YX} := \frac{1}{n} \sum_{i=1}^{n} Y_i \otimes X_i.$$

- Note that $\widehat{C}_{XX}$ and $\widehat{C}_{XY}$ are $\mathbb{P}$-a.s. of rank at most n.
- We now analyse the regularised empirical solution

$$\widehat{\theta}_\alpha := g_\alpha(A_{\widehat{C}_{XX}})[\widehat{C}_{YX}] = \widehat{C}_{YX}\, g_\alpha(\widehat{C}_{XX}). \tag{EMP}$$

Error analysis

- We can obtain rates for Hilbert–Schmidt regression based on Hölder source conditions.
- We analyse the error $\theta_\star - \widehat{\theta}_\alpha$ associated with the regularised empirical solution $\widehat{\theta}_\alpha$.
- In particular, we are interested both in the Hilbert–Schmidt norm of this error and in the mean-square prediction error

$$\mathbb{E}\left[\left\|\left(\theta_\star - \widehat{\theta}_\alpha\right)X\right\|_{\mathcal{Y}}^2\right] \equiv \left\|\left(\theta_\star - \widehat{\theta}_\alpha\right)C_{XX}^{1/2}\right\|_{S_2(\mathcal{X},\mathcal{Y})}^2.$$

- To treat these in a unified way we will examine

$$\left\|\left(\theta_\star - \widehat{\theta}_\alpha\right)C_{XX}^{s}\right\|_{S_2(\mathcal{X},\mathcal{Y})} \text{ for } 0 \leqslant s \leqslant \tfrac{1}{2}.$$

Hölder source conditions

To establish quantitative convergence rates, we need a priori assumptions on the "smoothness" of the ground truth $\theta_\star$, a.k.a. "source conditions":

Assumption 3

We assume that the solution satisfies the Hölder source condition $\theta_\star \in \Omega(\nu, R)$, where

$$\Omega(\nu, R) := \left\{ A^{\nu}_{C_{XX}}[\theta] \,\middle|\, \theta \in S_2(\mathcal{X}, \mathcal{Y}), \|\theta\|_{S_2(\mathcal{X},\mathcal{Y})} \leqslant R \right\} \subseteq S_2(\mathcal{X}, \mathcal{Y}).$$

Lemma 4

The source condition $\theta_\star \in \Omega(\nu, R)$ holds if and only if the moment condition

$$\sum_{i \in I} \sup_{x \in \mathcal{X}} \frac{\left| \mathbb{E}[\langle x, X \rangle_{\mathcal{X}} \langle e_i, Y \rangle_{\mathcal{Y}}] \right|^2}{\| C_{XX}^{\nu+1} x \|_{\mathcal{X}}^2} \leqslant R^2$$

hold for some (indeed, any) complete orthonormal system $\{e_i\}_{i \in I}$ in $\mathcal{Y}$.

Decomposing the error 1/2

- Naïve error decomposition: $\mathbb{P}^{\otimes n}$-a.s. with respect to the samples $(X_i, Y_i)_{i=1}^n$,

$$\left\|\left(\theta_\star - \widehat{\theta}_\alpha\right) C_{XX}^s\right\|_{S_2(\mathcal{X},\mathcal{Y})} \leqslant \underbrace{\left\|\left(\theta_\star - \theta_\alpha\right) C_{XX}^s\right\|_{S_2(\mathcal{X},\mathcal{Y})}}_{=\text{approximation error}} + \underbrace{\left\|\left(\theta_\alpha - \widehat{\theta}_\alpha\right) C_{XX}^s\right\|_{S_2(\mathcal{X},\mathcal{Y})}}_{=\text{variance}}. \tag{3.1}$$

- However, this decomposition turns out to be less than ideal and instead we use:

$$\begin{aligned}\theta_\star - \widehat{\theta}_\alpha &= \theta_\star - \theta_\star \widehat{C}_{XX} g_\alpha\left(\widehat{C}_{XX}\right) + \theta_\star \widehat{C}_{XX} g_\alpha\left(\widehat{C}_{XX}\right) - \widehat{\theta}_\alpha \\ &= \theta_\star r_\alpha\left(\widehat{C}_{XX}\right) + \theta_\star \widehat{C}_{XX} g_\alpha\left(\widehat{C}_{XX}\right) - \widehat{C}_{YX} g_\alpha\left(\widehat{C}_{XX}\right) \\ &= \theta_\star r_\alpha\left(\widehat{C}_{XX}\right) + \left(\theta_\star \widehat{C}_{XX} - \widehat{C}_{YX}\right) g_\alpha\left(\widehat{C}_{XX}\right).\end{aligned}$$

- Hence, $\mathbb{P}^{\otimes n}$-a.s.,

$$\left\|\left(\theta_\star - \widehat{\theta}_\alpha\right) C_{XX}^s\right\|_{S_2(\mathcal{X},\mathcal{Y})} \leqslant \left\|\theta_\star r_\alpha\left(\widehat{C}_{XX}\right) C_{XX}^s\right\|_{S_2(\mathcal{X},\mathcal{Y})} + \left\|\left(\theta_\star \widehat{C}_{XX} - \widehat{C}_{YX}\right) g_\alpha\left(\widehat{C}_{XX}\right) C_{XX}^s\right\|_{S_2(\mathcal{X},\mathcal{Y})}.$$

Decomposing the error 2/2

- Hence, $\mathbb{P}^{\otimes n}$-a.s.,

$$\left\| \left(\theta_\star - \widehat{\theta}_\alpha\right) C_{XX}^s \right\|_{S_2(\mathcal{X},\mathcal{Y})} \leqslant \left\| \theta_\star r_\alpha\left(\widehat{C}_{XX}\right) C_{XX}^s \right\|_{S_2(\mathcal{X},\mathcal{Y})} + \left\| \left(\theta_\star \widehat{C}_{XX} - \widehat{C}_{YX}\right) g_\alpha\left(\widehat{C}_{XX}\right) C_{XX}^s \right\|_{S_2(\mathcal{X},\mathcal{Y})}. \tag{3.2}$$

- Again, we think of the two terms on the right-hand side of (3.2) as an approximation error and a variance term.
- Crucially, though, the approximation error in the decomposition (3.2) is random — as opposed to the deterministic approximation term in (3.1) — and both terms in (3.2) will be amenable to analysis using concentration-of-measure techniques.

Hilbert space concentration bounds

The key tool for us is a recent concentration inequality for Hilbert space-valued random variables:

Theorem 5 (Maurer and Pontil, 2021, Prop. 7.11)

Let $\xi, \xi_1, \ldots, \xi_n$ be i.i.d. random variables with joint law $\mathbb{P}^{\otimes n}$ taking values in a separable Hilbert space $\mathcal{H}$ such that $\mathbb{E}[\xi] = 0$ and the subexponential norm $\|\xi\|_{L_{\psi_1}(\mathbb{P};\mathcal{H})}$ is finite. Then, for all $\delta \in (0, \frac{1}{2}]$ and $n \geqslant \log(1/\delta)$, with $\mathbb{P}^{\otimes n}$-probability at least $1 - \delta$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i \right\|_{\mathcal{H}} \leqslant 8\sqrt{2}e \|\xi\|_{L_{\psi_1}(\mathbb{P};\mathcal{H})} \sqrt{\frac{\log(1/\delta)}{n}}.$$

Despite the large number of terms that we need to bound, we carefully reduce the number of independent appeals to Maurer and Pontil (2021) to a minimum of only two.

Subexponential and sub-Gaussian norms

- For a real-valued random variable ξ defined on $(\Omega, \mathcal{F}, \mathbb{P})$, we introduce the Banach spaces $L_{\psi_1}(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}) = L_{\psi_1}(\mathbb{P})$ and $L_{\psi_2}(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}) = L_{\psi_2}(\mathbb{P})$ via the norms

$$\text{subexponential:} \qquad \|\xi\|_{L_{\psi_1}(\mathbb{P})} := \sup_{1 \leqslant p < \infty} \frac{\|\xi\|_{L^p(\mathbb{P})}}{p},$$

$$\text{sub-Gaussian:} \qquad \|\xi\|_{L_{\psi_2}(\mathbb{P})} := \sup_{1 \leqslant p < \infty} \frac{\|\xi\|_{L^p(\mathbb{P})}}{p^{1/2}}.$$

- For ξ taking values in a separable Hilbert space $\mathcal{H}$:

$$\|\xi\|_{L_{\psi_1}(\mathbb{P};\mathcal{H})} := \| \, \|\xi\|_{\mathcal{H}} \, \|_{L_{\psi_1}(\mathbb{P})} = \sup_{1 \leqslant p < \infty} \frac{\|\xi\|_{L^p(\mathbb{P};\mathcal{H})}}{p}$$

and analogously for $\|\xi\|_{L_{\psi_2}(\mathbb{P};\mathcal{H})} := \|\|\xi\|_{\mathcal{H}}\|_{L_{\psi_2}(\mathbb{P})}$.

Convergence rates

Theorem 6 (Convergence rates under Hölder source conditions)

Suppose that g_α has qualification $q \geqslant \nu + s$. Suppose that $Y \in L_{\psi_2}(\mathbb{P}; \mathcal{Y})$, $X \in L_{\psi_2}(\mathbb{P}; \mathcal{X})$, $\theta_\star \in \Omega(\nu, R)$, and $0 < \alpha < 1$. Let $\delta \in (0, \frac{1}{e}]$ and $s \in [0, \frac{1}{2}]$. For the regularisation schedule

$$\alpha_n := \left(\frac{1}{\sqrt{n}}\right)^{\frac{1}{\nu+1}},$$

and for

$$n \geqslant n_0 := \max\left\{\|X\|^4_{L_{\psi_2}(\mathbb{P};\mathcal{X})}, \left(1152 e^2 \|X\|^4_{L_{\psi_2}(\mathbb{P};\mathcal{X})} \log(1/\delta)\right)^{\frac{1}{\nu}}\right\}^{1+\nu},$$

with $\mathbb{P}^{\otimes n}$-probability at least $1 - 2\delta$,

$$\left\|\left(\theta_\star - \widehat{\theta}_{\alpha_n}\right) C^s_{XX}\right\|_{S_2(\mathcal{X},\mathcal{Y})} \leqslant 3\bar{\kappa}\sqrt{\log(1/\delta)}\left(\frac{1}{\sqrt{n}}\right)^{\frac{s+\nu}{1+\nu}},$$

where $\bar{\kappa}$ is an explicit constant depending only on the regularisation scheme, the source condition, and the sub-Gaussian norms of X and Y.

Optimal rates and comparison to kernel setting

- The rates in Theorem 6 match those of kernel regression with scalar and finite-dimensional response variables under a Hölder source condition and with no additional assumptions on the eigenvalue decay of C_{XX} (Caponnetto and De Vito, 2007; Blanchard and Mücke, 2018; Lin et al., 2020).
- Minimax optimality of these rates is only derived by Caponnetto and De Vito (2007) and Blanchard and Mücke (2018) under the additional assumption that the eigenvalues of C_{XX} decay rapidly enough, which is an implicit assumption on the marginal distribution of X.
- To establish minimax optimality in our setting, we would have to repeat the standard arguments, e.g. apply a general reduction scheme in conjunction with Fano's method (Tsybakov, 2009).
- However, as discussed earlier, the Hilbert–Schmidt regression problem has scalar response kernel regression and some settings of kernel regression with vector-valued response as special cases.

Closing remarks

Open questions

- Can we obtain fast $1/n$ rates? This would require additional assumptions about the joint law of (X, Y). So far, this is only solved for the special case of the CME (Li et al., 2022).
- Solving (RP)/(IP) over the non-reflexive Banach space $L(\mathcal{X}, \mathcal{Y})$ — a simple yet really evil example is $\mathcal{X} = \mathcal{Y}$ and $\theta_\star = \mathrm{Id}$.
- Learning in $L(\mathcal{X}, \mathcal{Y})$ requires more general source conditions, something like $\theta_\star = \tilde{\theta} C_{XX}^{\nu}$ with $\tilde{\theta} \in L(\mathcal{X}, \mathcal{Y})$ implies $\theta_\star \in S_2(\mathcal{X}, \mathcal{Y})$ for $\nu > \frac{1}{2}$.
- For Banach space $\mathcal{X}$ and $\mathcal{Y}$, a suitable analogue of (IP) is needed. The Hilbert case uses $\mathrm{tr}(C_{XX}) = \mathbb{E}[\|X\|_{\mathcal{X}}^2]$ and derivative of squared norm.
- Extension to more general non-i.i.d. sample data, e.g. autoregression for stationary time series?

Thank You!

arXiv:2211.08875

References

C. R. Baker. Joint measures and cross-covariance operators. *Trans. Amer. Math. Soc.*, 186:273–289, 1973. doi:10.2307/1996566.

F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *J. Complexity*, 23(1):52–72, 2007. doi:10.1016/j.jco.2006.07.001.

G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. *Found. Comput. Math.*, 18:971–1013, 2018. doi:10.1007/s10208-017-9359-7.

D. Bosq. *Linear Processes in Function Spaces.* Springer, New York, 2000. doi:10.1007/978-1-4612-1154-9.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7(3):331–368, 2007. doi:10.1007/s10208-006-0196-8.

A. Christmann and I. Steinwart. *Support Vector Machines.* Springer, New York, 2008. doi:10.1007/978-0-387-77242-4.

L. H. Dicker, D. P. Foster, and D. Hsu. Kernel ridge vs. principal component regression: minimax bounds and the qualification of regularization operators. *Electron. J. Stat.*, 11(1):1022–1047, 2017. doi:10.1214/17-EJS1258.

R. G. Douglas. On majorization, factorization, and range inclusion of operators on Hilbert space. *Proc. Amer. Math. Soc.*, 17:413–415, 1966. doi:10.2307/2035178.

H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

L. L. Gerfo, L. Rosasco, F. Odone, E. D. Vito, and A. Verri. Spectral algorithms for supervised learning. *Neural Comput.*, 20(7):1873–1897, 2008. doi:10.1162/neco.2008.05-07-517.

Z. Li, D. Meunier, M. Mollenhauer, and A. Gretton. Optimal rates for regularized conditional mean embedding learning. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2022. To appear, arXiv:2208.01711.

J. Lin, A. Rudi, L. Rosasco, and V. Cevher. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. *Appl. Comput. Harmon. Anal.*, 48(3):868–890, 2020. doi:10.1016/j.acha.2018.09.009.

A. Maurer and M. Pontil. Concentration inequalities under sub-Gaussian and sub-exponential conditions. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 7588–7597. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/file/3e33b970f21d2fc65096871ea0d2c6e4-Paper.pdf`.

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

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005. doi:10.1007/b98888.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009. doi:10.1007/b13794. Revised and extended from the 2004 French original, Translated by V. Zaiats.