On Regularized Radon–Nikodym Differentiation

Duc Hoan Nguyen DUC.NGUYEN@RICAM.OEAW.AC.AT
Johann Radon Institute for Computational and Applied Mathematics
Austrian Academy of Sciences
Altenberger Straße 69, 4040 Linz, Austria

University of Lorraine,
CNRS, CRAN, Nancy, F-54000, France

Werner Zellinger WERNER.ZELLINGER@RICAM.OEAW.AC.AT
Sergei Pereverzyev SERGEI.PEREVERZYEV@OEAW.AC.AT
Johann Radon Institute for Computational and Applied Mathematics
Austrian Academy of Sciences
Altenberger Straße 69, 4040 Linz, Austria

Editor: Maxim Raginsky

Abstract

We discuss the problem of estimating Radon–Nikodym derivatives. This problem appears in various applications, such as covariate shift adaptation, likelihood-ratio testing, mutual information estimation, and conditional probability estimation. However, in many of the above applications one is interested in the pointwise evaluation of the Radon–Nikodym derivatives rather than in their approximation as elements of some spaces of functions, and this aspect has been left unexplored in the previous studies. To address the above problem, we employ the general regularization scheme in reproducing kernel Hilbert spaces. The convergence rate of the corresponding regularized algorithm is established by taking into account both the smoothness of the derivative and the capacity of the space in which it is estimated. This is done in terms of general source conditions and the regularized Christoffel functions. We also find that the reconstruction of Radon–Nikodym derivatives at any particular point can be done with higher order of accuracy as compared to the reported work available so far. Our theoretical results are illustrated by numerical simulations.

Keywords: Density ratio, Reproducing kernel Hilbert space, Radon–Nikodym differentiation

1. Introduction

This paper is focused on the use of regularized kernel methods in the context of estimating the ratio of two probability density functions, which can also be called the Radon–Nikodym derivative of the corresponding probability measures.

Recently the estimation of Radon–Nikodym derivatives has gained significant attention due to its potential applications in such tasks as covariate shift adaptation, outlier detection, divergence estimation, and conditional probability estimation. Here we may refer to (Sugiyama et al., 2012) and references therein. In order to address the above problem, various kernel-based approaches are available. In particular, several regularization schemes in reproducing Kernel Hilbert space (RKHS) have been discussed in (Nguyen et al., 2010; Kanamori et al., 2012; Que and Belkin, 2013; Schuster et al., 2020; Gizewski et al., 2022).

In these earlier works, it has been assumed that the estimated Radon–Nikodym derivative belongs to RKHS in which the regularization is performed. In tasks like inlier-based outlier detection or covariate shift adaptation by importance weighting, the above realizability assumption is not just technical but also conceptual, because in the above mentioned tasks one is only interested in the pointwise evaluation of the Radon–Nikodym derivatives, and needs the assumption of belonging to RKHS to treat such evaluation as a continuous functional.

Note that Tikhonov–Lavrentier regularization is one of the most studied techniques employed for the estimation of Radon–Nikodym derivatives in RKHS. It has given rise to a method proposed in (Kanamori et al., 2012), where the authors have also argued that it compares favorably with other approaches in terms of computational efficiency and numerical stability. On the other hand, from the regularization theory (e.g., Lu and Pereverzyev, 2013) we know that a drawback of Tikhonov–Lavrentier regularization is its comparatively low qualification resulting in an earlier saturation, which means that if the smoothness of the estimated quantities exceeds a certain level, level of saturation, the order of the accuracy of Tikhonov–Lavrentiev regularization is not improved.

One of the main findings of this work is that the smoothness of quantities estimated in the pointwise evaluation of the Radon–Nikodym derivative can exceed the level of saturation of Tikhonov–Lavrentier regularization even though the smoothness of the derivative, considered by itself, is below that level (see Section 5 for the details); this observation goes beyond previous studies (Gizewski et al., 2022; Que and Belkin, 2013; Kanamori et al., 2012).

Then the use of Tikhonov–Lavrentier regularization in the tasks where only point values of the Radon–Nikodym derivatives are of interest, can lead to an unnecessary loss of accuracy. In view of the study by (Kanamori et al., 2012), the above observation is an important and rather unexpected research finding, in our opinion. It comes from the analysis presented below. At the same time, our analysis not only points out the above limitation of the method proposed by in (Kanamori et al., 2012), but also indicates the way to overcome it.

The analysis below is performed in terms of the so-called source conditions and the regularized Christoffel function. The concept of source condition is widely used in the regularization theory (e.g., Mathé and Hofmann, 2008) for measuring the smoothness of the estimated quantities against the rate of decrease of their Fourier coefficients with respect to orthonormal systems of the singular value decompositions of operators from the regularized problems. In the context of learning theory, the source conditions are discussed at least from the inspiring paper by (Smale and Zhou, 2007). In the present study, we consider the source conditions generated by the kernel covariance operator formed using the kernel K of the considered RKHS and the probability measure staying in the denominator of the Radon–Nikodym derivative.

One more concept used in our work is the regularized Christoffel function, which is an extension of the classical notion of the Christoffel function from the orthogonal polynomial literature. The extension is done by replacing the polynomials of increasing degrees by functions in the considered RKHS with increasing norms. Then it characterizes in some sense, the capacity of the employed approximating space and provides a bound for the evaluation at a given point independently somehow of the function to be evaluated. Another interesting observation of our study is that one can again use the concept of source conditions to elucidate the relationship of the regularized Christoffel function with the chosen kernel K as well as the probability measure with respect to which the Radon–Nikodym derivative is to be taken.

From the previous studies, one knows that the convergence of algorithms for Radon–Nikodym differentiation is influenced not only by the smoothness of the approximated function but also by the capacity of the approximating space. Though there are several studies that employed a particular regularization technique, such as Tikhonov–Lavrentiev regularization, to the best of our knowledge there is no study considering more general regularization schemes and taking into account both the above-mentioned factors, i.e.smoothness and capacity. For example, in (Kanamori et al., 2012) and (Que and Belkin, 2013) (see Type I setting there) only the capacity of the approximating space has been incorporated into error estimations, and in (Gizewski et al., 2022) and (Schuster et al., 2020) only the smoothness has been considered.

Besides, since in some applications the point values of the Radon–Nikodym derivatives are of interest, it seems natural to study their approximation in spaces, where pointwise evaluations are well-defined. However, in (Kanamori et al., 2012) and (Que and Belkin, 2013) the approximation has been analyzed in the space of integrable functions, where this is not the case.

In the present paper, we aim to overcome the above limitations. More precisely, we study general regularization schemes and analyze their accuracy with respect to both the smoothness of the Radon–Nikodym derivative and the capacity of the RKHS in which it is estimated. This is done in terms of general source conditions and regularized Christoffel functions. We then establish accuracy bounds of the corresponding regularized algorithm in the norm of RKHS and pointwise. Finally, we present some numerical illustrations supporting our theoretical results.

2. Assumptions and Auxiliaries

In the problem of estimation of Radon–Nikodym derivatives, we consider two probability measures p and q on a space $\mathbf{X} \subset \mathbb{R}^d$. The information about the measures is only provided in the form of samples $X_p = \{x_1, x_2, \ldots, x_n\}$ and $X_q = \{x'_1, x'_2, \ldots, x'_m\}$ drawn independently and identically (i.i.d) from p and q respectively. Moreover, we assume that there is a function $\beta : \mathbf{X} \to [0, \infty)$, which can be viewed as the Radon–Nikodym derivative $\frac{dq}{dp}$ of the probability measure $q(x)$ with respect to the probability measure $p(x)$, and for any measurable set $A \subset \mathbf{X}$ it holds

$$\int_A dq(x) = \int_A \beta(x) dp(x).$$

Our goal is to approximate the Radon–Nikodym derivative $\beta(x) = \frac{dq}{dp}$ by some function $\hat{\beta}(x)$ based on the observed samples. As it has been already explained in Introduction, we in fact need a strategy that ensures a good pointwise approximation to the derivatives $\beta(x)$.

Then it seems to be logical to estimate $\beta(x)$ in the norm of some RKHS, in which pointwise evaluations are well-defined.

Let $\mathcal{H}_K$ be a reproducing Kernel Hilbert space with a positive-definite function $K : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ as reproducing kernel. We assume that K is a continuous and bounded function, such that for any $x \in \mathbf{X}$

$$\|K(\cdot, x)\|_{\mathcal{H}_K} = \langle K(\cdot, x), K(\cdot, x)\rangle_{\mathcal{H}_K}^{\frac{1}{2}} = [K(x, x)]^{\frac{1}{2}} \leq \kappa_0 < \infty.$$

Let $L_{2,\rho}$ be the space of square-integrable functions $f : \mathbf{X} \to \mathbb{R}$ with respect to the probability measure ρ. We define $J_q : \mathcal{H}_K \hookrightarrow L_{2,q}$ and $J_p : \mathcal{H}_K \hookrightarrow L_{2,p}$ as the inclusion operators, such that for instance, J_q assigns to a function $g \in \mathcal{H}_K$ the same function seen as an element of $L_{2,q}$. In the sequel, we distinguish two sample operators

$$S_{X_q} f = (f(x'_1), f(x'_2), \ldots, f(x'_m)) \in \mathbb{R}^m,$$
$$S_{X_p} f = (f(x_1), f(x_2), \ldots, f(x_n)) \in \mathbb{R}^n,$$

acting from $\mathcal{H}_K$ to $\mathbb{R}^m$ and $\mathbb{R}^n$, where the norms in later spaces are generated by m^{-1}-times and n^{-1}-times the standard Euclidean inner products, such that, for example, for $u = (u_1, u_2, \ldots, u_m), w = (w_1, w_2, \ldots, w_m) \in \mathbb{R}^m$,

$$\langle u, w\rangle_{\mathbb{R}^m} = \frac{1}{m}\sum_{j=1}^{m} u_j w_j, \qquad \|u\|_{\mathbb{R}^m} = \langle u, u\rangle_{\mathbb{R}^m}^{\frac{1}{2}} = \left(\frac{1}{m}\sum_{j=1}^{m} u_j^2\right)^{\frac{1}{2}}.$$

Then the adjoint operators $S^*_{X_q} : \mathbb{R}^m \to \mathcal{H}_K$ and $S^*_{X_p} : \mathbb{R}^n \to \mathcal{H}_K$ are given as

$$S^*_{X_q} u(\cdot) = \frac{1}{m}\sum_{j=1}^{m} K(\cdot, x'_j) u_j, \qquad u = (u_1, u_2, \ldots, u_m) \in \mathbb{R}^m,$$
$$S^*_{X_p} v(\cdot) = \frac{1}{n}\sum_{i=1}^{n} K(\cdot, x_i) v_i, \qquad v = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^n.$$

In the literature, various RKHS-based approaches are available for a Radon–Nikodym derivative estimation. Here we may refer to (Kanamori et al., 2012) and to references therein. As it can be seen from (Que and Belkin, 2013), and also from (Gizewski et al., 2022), conceptually, under the assumption that $\beta \in \mathcal{H}_K$, several of the above approaches can be derived from a regularization of an operator equation, which can be written in our terms as

$$J^*_p J_p \beta = J^*_q J_q \mathbf{1}. \tag{1}$$

Because of the compactness of the operator $J^*_p J_p$, its inverse $(J^*_p J_p)^{-1}$ cannot be a bounded operator in $\mathcal{H}_K$, which makes the Eq. 1 ill-posed. Here, $\mathbf{1}$ is the constant function that takes the value 1 everywhere, and almost without loss of generality, we assume that $\mathbf{1} \in \mathcal{H}_K$, because otherwise the kernel $K_1(x, x') = 1 + K(x, x')$ will, for example, be used to generate a suitable RKHS containing all constant functions.

Since there is no direct access to the measures p and q, the Eq. 1 is inaccessible as well, but the samples X_p and X_q allow us to access its empirical version

$$S^*_{X_p} S_{X_p} \beta = S^*_{X_q} S_{X_q} \mathbf{1}. \tag{2}$$

A regularization of Eq. 2 may serve as a starting point for several approaches of estimating the Radon–Nikodym derivative β. For example, as it has been observed in (Kanamori et al., 2012; Gizewski et al., 2022), the known kernel mean matching (KMM) method (Huang et al., 2006) can be viewed as the regularization of Eq. 1 by the method of quasi (least-squares) solutions, originally proposed by Valentin Ivanov (1963) and also known as Ivanov regularization (see, for example, (Oneto et al., 2016) and (Page and Grünewälder, 2019) for its use in the context of learning). At the same time, from Theorem 1 of Kanamori et al. (2012) it follows that the kernelized unconstrained least-squares importance fitting (KuLSIF) proposed in (Kanamori et al., 2012) is in fact the application of the Lavrentiev regularization scheme to the empirical version Eq. 2 of the Eq. 1, that is in KuLSIF we have

$$\hat{\beta} = \beta^{\lambda}_{\mathbf{X}} = (\lambda I + S^*_{X_p} S_{X_p})^{-1} S^*_{X_q} S_{X_q} \mathbf{1}. \tag{3}$$

As we have already mentioned in Introduction, early bounds of the accuracy of Radon–Nikodym numerical differentiation have relied only on the capacity of the approximating space. For example, in (Nguyen et al., 2010; Kanamori et al., 2012) the capacity of the underlying space $\mathcal{H}_K$ has been measured in terms of the so-called bracketing entropy, and in (Kanamori et al., 2012) the value of the regularization parameter λ in KuLSIF Eq. 2 has been chosen depending on that capacity measure. Note that, in such approach, there is no possibility of incorporating into the regularization the information about other factors, such as the smoothness of the approximated derivative β, which, as we know from (Gizewski et al., 2022), also influences the regularization accuracy, and therefore should be taken into account when choosing λ. For this reason, in the present study, we follow (Pauwels et al., 2018) and employ the concept of the regularized Christoffel function that allows direct incorporation of the regularization parameter λ into the definition of a capacity measure. In this way, the intention is to relate two factors influencing regularization accuracy (i.e., the smoothness and the capacity) in one parameter.

Consider the function

$$C_\lambda(x) = \left\langle K(\cdot, x), (\lambda I + J^*_p J_p)^{-1} K(\cdot, x) \right\rangle_{\mathcal{H}_K} = \left\| (\lambda I + J^*_p J_p)^{-\frac{1}{2}} K(\cdot, x) \right\|^2_{\mathcal{H}_K} \tag{4}$$

Note that in (Pauwels et al., 2018) the reciprocal of $C_\lambda(x)$, i.e.$\frac{1}{C_\lambda(x)}$, was called the regularized Christoffel function, but for the sake of simplicity, we will keep the same name also for Eq. 4. Note also that in the context of supervised learning where usually only one probability measure, say p, is involved, the expected value

$$\mathcal{N}(\lambda) = \int_{\mathbf{X}} C_\lambda(x) dp(x)$$

of $C_\lambda(x)$, called the effective dimension, has been employed by (Caponnetto and De Vito, 2007) as an indicator of the capacity of the approximating space.

At the same time, if more than one measure appears in the supervised learning context, as is for example the case in the analysis of Nyström subsampling (Rudi et al., 2015; Lu et al., 2019), then the C-norm of the regularized Christoffel function

$$\mathcal{N}_\infty(\lambda) := \sup_{x \in \mathbf{X}} C_\lambda(x) \tag{5}$$

is used in parallel with the effective dimension $\mathcal{N}(\lambda)$. This gives a hint that $\mathcal{N}_\infty(\lambda)$ could also be a suitable capacity measure for analyzing the accuracy of Radon–Nikodym numerical differentiation because this differentiation is intrinsically related with more than one measure.

We will need the following statement.

Lemma 1 *Let $b_0 > 0$ be such that $|\beta(x)| \leq b_0$ for every $x \in \mathbf{X}$. Then with probability at least $1 - \delta$ we have*

$$\left\|(\lambda I + J_p^* J_p)^{-\frac{1}{2}}(S_{X_p}^* S_{X_p}\beta - S_{X_q}^* S_{X_q}\mathbf{1})\right\|_{\mathcal{H}_K} \leq \left(1 + \sqrt{2\log\frac{2}{\delta}}\right)\sqrt{\mathcal{N}_\infty(\lambda)}\sqrt{\frac{b_0^2}{n} + \frac{1}{m}}.$$

The proof of Lemma 1 is based on Lemma 4 of Huang et al. (2006), which we formulate in our notations as follows

Lemma 2 *(Huang et al. (2006)) Let ϕ be a map from $\mathbf{X}$ into $\mathcal{H}_K$ such that $\|\phi(x)\|_{\mathcal{H}_K} \leq R$ for all $x \in \mathbf{X}$. Then with probability at least $1 - \delta$ it holds*

$$\left\|\frac{1}{m}\sum_{j=1}^{m}\phi(x_j') - \frac{1}{n}\sum_{i=1}^{n}\beta(x_i)\phi(x_i)\right\|_{\mathcal{H}_K} \leq \left(1 + \sqrt{2\log\frac{2}{\delta}}\right) R\sqrt{\frac{b_0^2}{n} + \frac{1}{m}}.$$

Now we can return to Lemma 1.
Proof We define a map $\phi : \mathbf{X} \to \mathcal{H}_K$ as $\phi(x) = (\lambda I + J_p^* J_p)^{-\frac{1}{2}} K(\cdot, x)$, $x \in \mathbf{X}$. It is clear that

$$\|\phi(x)\|_{\mathcal{H}_K} = \left\|(\lambda I + J_p^* J_p)^{-\frac{1}{2}} K(\cdot, x)\right\|_{\mathcal{H}_K} = \sqrt{C_\lambda(x)} \leq \sqrt{\mathcal{N}_\infty(\lambda)}.$$

Therefore, for the map ϕ the condition of the above Lemma 2 is satisfied with $R = \mathcal{N}_\infty(\lambda)$. Then directly from that lemma, we have

$$\begin{aligned}\left\|(\lambda I + J_p^* J_p)^{-\frac{1}{2}}(S_{X_p}^* S_{X_p}\beta - S_{X_q}^* S_{X_q}\mathbf{1})\right\|_{\mathcal{H}_K} &= \left\|\frac{1}{n}\sum_{i=1}^{n}\beta(x_i)\phi(x_i) - \frac{1}{m}\sum_{j=1}^{m}\phi(x_j')\right\|_{\mathcal{H}_K} \\ &\leq \left(1 + \sqrt{2\log\frac{2}{\delta}}\right)\left(\sqrt{\frac{b_0^2}{n} + \frac{1}{m}}\right)\sqrt{\mathcal{N}_\infty(\lambda)}.\end{aligned}$$

■

3. General Regularization Scheme and General Source Conditions

All of the available regularization methods have the potential to be employed for the regularization of Eq. 2. In particular, we will use a general regularization scheme to construct a family of approximate solutions $\beta_{\mathbf{X}}^{\lambda}$ of Eq. 1 as follows

$$\beta_{\mathbf{X}}^{\lambda} = g_\lambda(S_{X_p}^* S_{X_p}) S_{X_q}^* S_{X_q} \mathbf{1}, \tag{6}$$

where $\{g_\lambda\}$ is a family of positive valued functions parametrized by parameter $\lambda > 0$ called a regularization parameter.

Recall that if $A : \mathcal{H}_K \to \mathcal{H}_K$ is a linear compact self-adjoint non-negative operator with the spectral decomposition

$$A = \sum_i \sigma_i f_i \left\langle f_i, \cdot \right\rangle_{\mathcal{H}_K},$$

then for any function $g : [0, \|A\|_{\mathcal{H}_K \to \mathcal{H}_K}] \to [0, \infty)$ the operator $g(A)$ is formally defined as follows

$$g(A) = \sum_i g(\sigma_i) f_i \left\langle f_i, \cdot \right\rangle_{\mathcal{H}_K}.$$

Note that for computing the image $g(A)f$ of $f \in \mathcal{H}_K$ it is not always necessary to know the spectral decomposition of A. As an example, one can take the computation of the iterated Lavrentiev regularization discussed below.

3.1 General Regularization Scheme

From the regularization theory we know (e.g., Lu and Pereverzyev, 2013, def. 2.2) that a family of functions $g_\lambda : [0, c] \to [0, \infty)$ parametrized by positive parameter λ can give rise to a regularization algorithm if there are positive constants $\gamma_0, \gamma_{-\frac{1}{2}}, \gamma_{-1}$ for which the following holds:

$$\begin{aligned} \sup_{0<t\le c} |1 - t g_\lambda(t)| &\le \gamma_0, \\ \sup_{0<t\le c} \sqrt{t} |g_\lambda(t)| &\le \frac{\gamma_{-\frac{1}{2}}}{\sqrt{\lambda}}, \\ \sup_{0<t\le c} |g_\lambda(t)| &< \frac{\gamma_{-1}}{\lambda}. \end{aligned} \tag{7}$$

Here and in the sequel, we adopt the convention that c denotes a generic positive coefficient, which can vary from appearance to appearance and may only depend on basic parameters such as p, q, κ_0, b_0, and others introduced below.

The *qualification* of a regularization scheme indexed by g_λ is the maximal $s > 0$ for which

$$\sup_{0<t\le c} t^s |1 - t g_\lambda(t)| \le \gamma_s \lambda^s, \tag{8}$$

where γ_s does not depend on λ. Following Definition 2.3 of Lu and Pereverzyev (2013) we also say that the qualification s covers a non-decreasing function $\varphi : [0, c] \to [0, \infty)$, $\varphi(0) = 0$, if the function $t \to \frac{t^s}{\varphi(t)}$ is non-decreasing for $t \in (0, c]$. Note that the higher the qualification is the more rapidly increasing the function it can cover, and in this way, as it can be seen below, the more smoothness of approximated solutions can be utilized in the regularization.

Observe that the Lavrentiev regularization used in KuLSIF Eq. 3 corresponds to $g_\lambda(t) = (\lambda + t)^{-1}$ and has qualification $s = 1$. At the same time, one can increase qualification by using the idea of the iterative Laverentiev regularization (e.g., Pereverzyev, 2022, page 41). In this way, the approximate Radon–Nikodym derivative can be obtained iteratively as follows

$$\begin{aligned}
&\beta_{\mathbf{X}}^{\lambda,0} = 0,\\
&\beta_{\mathbf{X}}^{\lambda,l} = (\lambda I + S_{X_p}^* S_{X_p})^{-1}(S_{X_q}^* S_{X_q}\mathbf{1} + \lambda\beta_{\mathbf{X}}^{\lambda,l-1}), \quad l \in \mathbb{N}.
\end{aligned}$$

After k such iterations we obtained the approximation $\beta_{\mathbf{X}}^{\lambda} = \beta_{\mathbf{X}}^{\lambda,k}$ that can be represented in the form Eq. 6 with

$$g_\lambda(t) = g_{\lambda,k}(t) = \frac{1 - \frac{\lambda^k}{(\lambda+t)^k}}{t}.$$

The regularization indexed by $g_{\lambda,k}(t)$ has the qualification k that can be taken as large as desired. Moreover, for $g_\lambda(t) = g_{\lambda,k}(t)$ the requirements Eq. 7, Eq. 8 are satisfied with $\gamma_0 = 1, \gamma_{-\frac{1}{2}} = k^{\frac{1}{2}}, \gamma_{-1} = k, \gamma_k = 1$.

For the sake of shortness, we introduce the residual function

$$r_\lambda(t) := 1 - tg_\lambda(t),$$

for which Eq. 7 and Eq. 8 give the bounds $r_\lambda(t) \le \gamma_0$ and $|t^s r_\lambda(t)| \le \gamma_s\lambda^s$.

3.2 General Source Conditions

As mentioned in the previous section, the Eq. 1 is inaccessible, but the result of Mathé and Hofmann (2008) tells us that there always exists an element $\nu_q \in \mathcal{H}_K$ and a continuous, strictly increasing function $\varphi : [0, \|J_p^* J_p\|_{\mathcal{H}_K \to \mathcal{H}_K}] \to [0, \infty)$, which obeys $\varphi(0) = 0$, such that the solution of Eq. 1 allows for a representation in terms of the *source condition*:

$$\beta = \varphi(J_p^* J_p)\nu_q. \tag{9}$$

The function φ above is usually called the *index function.* Moreover, for every $\epsilon > 0$ one can find such φ that Eq. 9 holds true for ν_q with

$$\|\nu_q\|_{\mathcal{H}_K} \le (1 + \epsilon)\|\beta\|_{\mathcal{H}_K}.$$

It is worth mentioning that Corollary 1 of Mathé and Hofmann (2008) tell us that under all index functions φ in Eq. 9 there is no function with the highest decay rate to zero, or what is the same, there is no maximal smoothness of β with respect to the operator $J_p^* J_p$. This is because in Eq. 9 the element $\nu_q \in \mathcal{H}_K$ also allows for a representation $\nu_q = \varphi_1(J_p^* J_p)\nu_q'$ with some other index function φ_1 and $\nu_q' \in \mathcal{H}_K$, $\left\|\nu_q'\right\|_{\mathcal{H}_K} \le (1 + \epsilon)\left\|\nu_q\right\|_{\mathcal{H}_K}$. Then from Eq. 9 it follows that

$$\beta = \varphi(J_p^* J_p)\varphi_1(J_p^* J_p)\nu_q' = \varphi_2(J_p^* J_p)\nu_q',$$

where $\left\|\nu_q'\right\|_{\mathcal{H}_K} \leq (1+\epsilon)^2 \left\|\beta\right\|_{\mathcal{H}_K}$, and $\varphi_2(t) = \varphi(t)\varphi_1(t)$ is an index function with higher decay rate to zero than $\varphi(t)$ as $t \to 0$.

Note that since the operator $J_p^* J_p$ is not accessible, there is a reason to restrict ourselves to consideration of such index functions φ, which allow us to control perturbations in the operators involved in the definition of source conditions. A class of such index functions has been discussed in (Mathé and Pereverzev, 2003) and in (Bauer et al., 2007). Here we follow those studies. Namely, we consider the class $\mathcal{F} = \mathcal{F}(0, c)$ of index functions $\varphi : [0, c] \to \mathbb{R}_+$ allowing splitting $\varphi(t) = \vartheta(t)\psi(t)$ into monotone Lipschitz part $\vartheta, \vartheta(t) = 0$, with the Lipschitz constant equal to 1, and an operator monotone part $\psi, \psi(0) = 0$.

Recall that a function ψ is *operator monotone* on $[0, c]$ if for any pair of self-adjoint operators U, V with spectra in $[0, c]$ such that $U \leq V$ (i.e.$V - U$ is an non-negative operator) we have $\psi(U) \leq \psi(V)$.

An important property of an operator monotone index function, say φ, is that it keeps bounds on the error of approximation of one self-adjoint operator U with spectrum on $[0, c]$ by another such operator V. Namely, $\|\varphi(U) - \varphi(V)\| \leq c\varphi(\|U - V\|)$, where c depends only on φ. The opposite side of this property of the operator monotone index functions is that they cannot converge faster than linearly to zero, for details, see (e.g., Mathé and Pereverzev, 2003). This is the reason to consider the class $\mathcal{F} = \mathcal{F}(0, c)$ of index functions that are free from above limitation.

Examples of operator monotone index functions are $\psi(t) = t^\nu$, $\psi(t) = \log^{-\nu}\left(\frac{1}{t}\right)$, $\psi(t) = \log^{-\nu}\left(\log\frac{1}{t}\right), 0 < \nu \leq 1$, while an example of a function φ from the above defined class $\mathcal{F}$ is $\varphi(t) = t^r \log^{-\nu}\left(\frac{1}{t}\right), r > 1, 0 < \nu \leq 1$, since it can be split into a Lipschitz part $\vartheta(t) = t^r$ and an operator monotone part $\psi(t) = \log^{-\nu}\left(\frac{1}{t}\right)$.

We will need the result of Proposition 3.1 in (Pereverzyev, 2022), which we formulate in our notations as follows

Lemma 3 *Let $\varphi : [0, c] \to \mathbb{R}$, $\varphi(0) = 0$, be any non-decreasing index function. If the qualification s of the regularization indexed by a family $\{g_\lambda\}$ covers the function φ, then for any $\lambda \in (0, c]$ it holds*

$$\sup_{t\in[0,c]} |r_\lambda(t)\varphi(t)| \leq \gamma_{0,s}\varphi(\lambda),$$

where $\gamma_{0,s} = \max\{\gamma_0, \gamma_s\}$, and γ_0, γ_s are the coefficients appearing in Eq. 7 and in Eq. 8.

To estimate the regularized Christoffel functions we slightly generalize a source condition for kernel sections $K(\cdot, x)$ that has been used in various contexts in (Lu et al., 2019) and (De Vito et al., 2014).

Assumption 1 (Source condition for kernel) *Let $\xi : [0, \left\|J_p^* J_p\right\|_{\mathcal{H}_K \to \mathcal{H}_K}] \to [0, \infty)$ be an operator monotone index function such that the function ξ^2 is covered by qualification $s = 1$. Assume that for each $x \in \mathbf{X}$ there exists an element $v_x \in \mathcal{H}_K$ such that*

$$K(\cdot, x) = \xi(J_p^* J_p)v_x,$$

and the norms $\|v_x\|$ are uniformly bounded by a constant c that is independent of x.

Remark 4 *In fact, Assumption 1 is not so restrictive, because as noted, for example, in (Bauer et al., 2007), any function $f \in \mathcal{H}_K$ meets the source condition $f = (J_p^* J_p)^\mu v$ with*

$\mu > 0$ and $\|v\|_{\mathcal{H}_K} \leq c\,\|f\|_{\mathcal{H}_K}$. Observe that by definition $K(\cdot,x) \in \mathcal{H}_K$ and $\|K(\cdot,x)\|_{\mathcal{H}_K} \leq \kappa_0$. Therefore, for each $x \in \mathbf{X}$ there is μ_x and $v_{\mu_x} \in \mathcal{H}_K$ such that

$$K(\cdot,x) = (J_p^* J_p)^{\mu_x} v_{\mu_x}, \quad \|v_{\mu_x}\|_{\mathcal{H}_K} \leq c\kappa_0. \tag{10}$$

Then Assumption 1 just says that there is an operator monotone index function $\xi(t)$ such that the functions $h_{\mu_x} = t^{\mu_x}/\xi(t)$ are bounded on $[0, \left\|J_p^* J_p\right\|_{\mathcal{H}_K \to \mathcal{H}_K}]$ by some constant c_1, and $\xi^2(t)$ is covered by qualification $s = 1$. Indeed, if such $\xi(t)$ exists then

$$K(\cdot,x) = (J_p^* J_p)^{\mu_x} v_{\mu_x} = \xi(J_p^* J_p) h_{\mu_x}(J_p^* J_p) v_{\mu_x},$$

which means that Assumption 1 is satisfied with $v_x = h_{\mu_x}(J_p^* J_p) v_{\mu_x}$ and $\|v_x\|_{\mathcal{H}_K} \leq c_1 c\kappa_0$. Note that Assumption 3b of De Vito et al. (2014) assumes the above function $\xi(t)$ to be of the form $\xi(t) = t^a, a \in (0, \frac{1}{2}]$, which means that all μ_x in Eq. 10 are assumed to be bounded away from zero by $a > 0$. This condition can be made even weaker by assuming $\xi(t)$ to be, for example, of the form $\xi(t) = \log^{-a}(\frac{1}{t}), a > 0$.

Assumption 1 allows us to relate the concept of the source conditions with the concept of the regularized Christoffel functions.

Lemma 5 *Under Assumption 1,*

$$\mathcal{N}_\infty(\lambda) \leq c\frac{\xi^2(\lambda)}{\lambda}.$$

Proof This simply follows from Lemma 3, Assumption 1 and the fact that the Lavrentiev regularization associated with $g_\lambda(t) = (\lambda + t)^{-1}$ has the qualification $s = 1$. Indeed,

$$\begin{aligned}
\mathcal{N}_\infty(\lambda) &= \sup_{x \in \mathbf{X}} \left\|(\lambda I + J_p^* J_p)^{-\frac{1}{2}} K(\cdot,x)\right\|^2_{\mathcal{H}_K} \\
&= \sup_{x \in \mathbf{X}} \left\|(\lambda I + J_p^* J_p)^{-\frac{1}{2}} \xi(J_p^* J_p) v_x\right\|^2_{\mathcal{H}_K} \\
&\leq \sup_{x \in \mathbf{X}} \|v_x\|^2_{\mathcal{H}_K} \sup_{t \in \left[0, ||J_p^* J_p||\right]} |(\lambda + t)^{-\frac{1}{2}} \xi(t)|^2 \\
&\leq c \sup_t |(\lambda + t)^{-1} \xi^2(t)| \\
&\leq c\lambda^{-1} \sup_t \left| \left(1 - t(\lambda + t)^{-1}\right) \xi^2(t)\right| \\
&\leq c\lambda^{-1} \sup_t |r_\lambda(t) \xi^2(t)| \\
&\leq c\frac{\xi^2(\lambda)}{\lambda}.
\end{aligned}$$

■

Remark 6 *In (Pauwels et al., 2018), the asymptotic behavior of the regularized Christoffel functions $C_\lambda(x)$ as $\lambda \to 0$ has been analyzed for translation invariant kernels $K(x,t) = K(x - t)$. Our Lemma 5 can be viewed as an extension of that analysis based on the general source conditions on the kernel sections $K_x(\cdot) = K(\cdot,x) \in \mathcal{H}_K$.*

4. Error Estimates in RKHS

In this section, we discuss error estimates between β and $\beta_{\mathbf{X}}^{\lambda}$ for RKHS norm. To this end, we consider an auxiliary regularized approximation $\bar{\beta}^{\lambda}$ defined as follows

$$\bar{\beta}^{\lambda} = g_{\lambda}(S_{X_p}^{*}S_{X_p})S_{X_p}^{*}S_{X_p}\beta. \tag{11}$$

Then we decompose the error bound into two parts:

$$\left\|\beta - \beta_{\mathbf{X}}^{\lambda}\right\|_{\mathcal{H}_K} \le \left\|\beta - \bar{\beta}^{\lambda}\right\|_{\mathcal{H}_K} + \left\|\bar{\beta}^{\lambda} - \beta_{\mathbf{X}}^{\lambda}\right\|_{\mathcal{H}_K}. \tag{12}$$

We call the first term on the right-hand side of Eq. 12 the approximation error, and the second term the noise propagation error.

Following (Lu et al., 2020), we introduce the functions

$$\mathcal{B}_{n,\lambda} := \frac{2\kappa_0}{\sqrt{n}}\left(\frac{\kappa_0}{\sqrt{n\lambda}} + \sqrt{\mathcal{N}(\lambda)}\right), \tag{13}$$

$$\Upsilon(\lambda) := \left(\frac{\mathcal{B}_{n,\lambda}}{\sqrt{\lambda}}\right)^2 + 1, \tag{14}$$

which will be useful in the subsequent analysis.

Moreover, we need the following estimates from (Lu et al., 2020) that are valid with probability at least $1-\delta$ and can be written in our notations as

$$\left\|J_p^{*}J_p - S_{X_p}^{*}S_{X_p}\right\|_{\mathcal{H}_K\to\mathcal{H}_K} \le \frac{4\kappa_0^2}{\sqrt{n}}\log\frac{2}{\delta}, \tag{15}$$

$$\left\|(\lambda I + J_p^{*}J_p)^{-1/2}(J_p^{*}J_p - S_{X_p}^{*}S_{X_p})\right\|_{\mathcal{H}_K\to\mathcal{H}_K} \le \mathcal{B}_{n,\lambda}\log\frac{2}{\delta}, \tag{16}$$

$$\Xi := \left\|(\lambda I + J_p^{*}J_p)(\lambda I + S_{X_p}^{*}S_{X_p})^{-1}\right\|_{\mathcal{H}_K\to\mathcal{H}_K} \le 2\left[\left(\frac{\mathcal{B}_{n,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2 + 1\right]. \tag{17}$$

Proposition 7 (Approximation error bound) *1. If β meets source condition Eq. 9, where φ is an operator monotone index function, then*

$$\left\|\beta - \bar{\beta}^{\lambda}\right\|_{\mathcal{H}_K} \le c(\gamma_0 + \gamma_{-1})\Xi\varphi(\lambda), \quad 0 < \lambda \le \kappa_0.$$

2. If β meets source condition Eq. 9, where $\varphi = \vartheta\psi \in \mathcal{F}(0,c)$ with c is large enough and if the qualification of the regularization g_λ covers $\vartheta(t)t^{\frac{3}{2}}$, then

$$\left\|\beta - \bar{\beta}^{\lambda}\right\|_{\mathcal{H}_K} \le c\Xi\varphi(\lambda) + \psi(\kappa_0)\gamma_0\left\|J_p^{*}J_p - S_{X_p}^{*}S_{X_p}\right\|_{\mathcal{H}_K\to\mathcal{H}_K}.$$

Proof The Proposition 7 can be proved by repeating line by line the argument of the proof of Proposition 4.3 in (Lu et al., 2020), where the items denoted there as T, T_x, $f^{\dagger}$ and $\bar{f}_x^{\lambda}$ should be substituted by $J_p^{*}J_p$, $S_{X_p}^{*}S_{X_p}$, β, and $\bar{\beta}^{\lambda}$, respectively. ■

Remark 8 *In the report of an anonymous reviewer on the previous version of this study it has been noted that the requirement on the qualification of g_λ to cover $\vartheta(t)t^{\frac{3}{2}}$ seems to be unnecessary. Indeed, after a close examination of the proof of Proposition 4.3 in (Lu et al., 2020) we can conclude that part 2 of Proposition 7 can be proven under a weaker assumption, namely, that the qualification of the regularization g_λ, covers the index function $\vartheta(t)t$. At the same time, to guarantee the accuracy bounds of the pointwise evaluation of the Radon–Nikodym numerical differentiation proven below in Proposition 14 and Theorem 16 we indeed need a regularization with a qualification covering the index function $\vartheta(t)t^{\frac{3}{2}}$ (details are explained in Section 5). On the other hand, from the viewpoint of the iterated Lavrentiev regularization discussed in Section 3.1, if the qualification $s = k$ of the regularization $g_\lambda = g_{\lambda,k}(t) = \frac{1-\frac{\lambda^k}{(\lambda+t)^k}}{t}$ covers the index function $\vartheta(t)t$ (i.e., the function $t \to \frac{t^{k-1}}{\vartheta(t)}$ is non-decreasing), but does not cover the function $\vartheta(t)t^{\frac{3}{2}}$, then one can perform just one more iteration and construct the regularization $g_\lambda = g_{\lambda,k+1}(t)$, which qualification $s = k+1$ is sufficient to cover the function $\vartheta(t)t^{\frac{3}{2}}$.*

Proposition 9 (Noise propagation error bound) *Let $\beta^\lambda_{\mathbf{X}}, \bar{\beta}^\lambda$ be defined by Eq. 3, Eq. 11. Then with probability at least $1-\delta$ it holds*

$$\left\|\bar{\beta}^\lambda - \beta^\lambda_{\mathbf{X}}\right\|_{\mathcal{H}_K} \le c\left(\gamma^2_{-\frac{1}{2}} + \gamma^2_{-1}\right)^{\frac{1}{2}} \Xi^{\frac{1}{2}} \frac{1}{\sqrt{\lambda}} \left(m^{-\frac{1}{2}} + n^{-\frac{1}{2}}\right) \sqrt{\mathcal{N}_\infty(\lambda)} \left(\log^{\frac{1}{2}} \frac{1}{\delta}\right).$$

Proof From Eq. 17 and well-known Cordes inequality ($\|A^sB^s\| \le \|AB\|^s$ for $s \in [0,1]$ and arbitrary positive operators A, B on a Hilbert space), as well as from the fact that $(\lambda I + J_p^*J_p), (\lambda I + S^*_{X_p}S_{X_p})^{-1}$ are positive and self-adjoint operators, we have

$$\left\|(\lambda I + J_p^*J_p)^{1/2}(\lambda I + S^*_{X_p}S_{X_p})^{-1/2}\right\|_{\mathcal{H}_K \to \mathcal{H}_K} \le \Xi^{1/2}. \tag{18}$$

Then using Eq. 7 and Lemma 1, we can continue

$$\begin{aligned}
\left\|\bar{\beta}^\lambda - \beta^\lambda_{\mathbf{X}}\right\|_{\mathcal{H}_K} &= \left\|g_\lambda(S^*_{X_p}S_{X_p})(S^*_{X_q}S_{X_q}\mathbf{1} - S^*_{X_p}S_{X_p}\beta)\right\|_{\mathcal{H}_K} \\
&\le \left\|g_\lambda(S^*_{X_p}S_{X_p})(\lambda I + S^*_{X_p}S_{X_p})^{\frac{1}{2}}\right\|_{\mathcal{H}_K \to \mathcal{H}_K} \left\|(\lambda I + S^*_{X_p}S_{X_p})^{-\frac{1}{2}}(\lambda I + J_p^*J_p)^{\frac{1}{2}}\right\|_{\mathcal{H}_K \to \mathcal{H}_K} \\
&\quad \times \left\|(\lambda I + J_p^*J_p)^{-\frac{1}{2}}(S^*_{X_q}S_{X_q}\mathbf{1} - S^*_{X_p}S_{X_p}\beta)\right\|_{\mathcal{H}_K} \\
&\le \sup_{0<t\le c} |g_\lambda(t)(\lambda+t)^{\frac{1}{2}}| \Xi^{\frac{1}{2}} \left\|(\lambda I + J_p^*J_p)^{-\frac{1}{2}}(S^*_{X_p}S_{X_p}\beta - S^*_{X_q}S_{X_q}\mathbf{1})\right\|_{\mathcal{H}_K} \\
&\le (\gamma^2_{-\frac{1}{2}} + \gamma^2_{-1})^{\frac{1}{2}} \frac{1}{\sqrt{\lambda}} \Xi^{\frac{1}{2}} \left(1 + \sqrt{2\log\frac{2}{\delta}}\right) \left(\sqrt{\frac{b_0^2}{n} + \frac{1}{m}}\right) \sqrt{\mathcal{N}_\infty(\lambda)} \\
&\le c\left(\gamma^2_{-\frac{1}{2}} + \gamma^2_{-1}\right)^{\frac{1}{2}} \frac{1}{\sqrt{\lambda}} \Xi^{\frac{1}{2}} \left(m^{-\frac{1}{2}} + n^{-\frac{1}{2}}\right) \sqrt{\mathcal{N}_\infty(\lambda)} \left(\log^{\frac{1}{2}}\frac{1}{\delta}\right).
\end{aligned}$$

■

The next proposition summaries of Proposition 7 and Proposition 9.

Proposition 10 *If β meets source condition Eq. 9, where φ is an operator monotone index function, then with probability at least $1-\delta$ it holds*

$$\left\|\beta - \beta_{\mathbf{X}}^{\lambda}\right\|_{\mathcal{H}_K} \leq c\left(\Upsilon(\lambda)\varphi(\lambda) + \frac{1}{\sqrt{\lambda}}[\Upsilon(\lambda)]^{\frac{1}{2}}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}})\sqrt{\mathcal{N}_{\infty}(\lambda)}\right)\left(\log\frac{2}{\delta}\right)^2.$$

If β meets source condition Eq. 9, where $\varphi = \vartheta\psi \in \mathcal{F}(0,c)$ with c is large enough, and if the qualification of the regularization g_λ covers $\vartheta(t)t^{\frac{3}{2}}$ then with probability at least $1-\delta$ the total error allows for the bound

$$\left\|\beta - \beta_{\mathbf{X}}^{\lambda}\right\|_{\mathcal{H}_K} \leq c\left(\Upsilon(\lambda)\varphi(\lambda) + n^{-\frac{1}{2}} + \frac{1}{\sqrt{\lambda}}[\Upsilon(\lambda)]^{\frac{1}{2}}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}})\sqrt{\mathcal{N}_{\infty}(\lambda)}\right)\left(\log\frac{2}{\delta}\right)^2.$$

Proof We first prove the results for β meets source condition Eq. 9, where φ is an operator monotone index function. Using the error estimates in Proposition 7 and Proposition 9, and Eq. 13 - Eq. 17, we have

$$\begin{aligned}
\left\|\beta - \beta_{\mathbf{X}}^{\lambda}\right\|_{\mathcal{H}_K} \leq & c(\gamma_0 + \gamma_{-1})\left[\left(\frac{\mathcal{B}_{n,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2 + 1\right]\varphi(\lambda) \\
& + \left(\gamma_{-\frac{1}{2}}^2 + \gamma_{-1}^2\right)^{\frac{1}{2}}\frac{1}{\sqrt{\lambda}}\sqrt{\left(\frac{\mathcal{B}_{n,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2 + 1}\left(\log^{\frac{1}{2}}\frac{1}{\delta}\right)\left(m^{-\frac{1}{2}} + n^{-\frac{1}{2}}\right)\sqrt{\mathcal{N}_{\infty}(\lambda)} \\
\leq & c\left(\Upsilon(\lambda)\varphi(\lambda) + \frac{1}{\sqrt{\lambda}}[\Upsilon(\lambda)]^{\frac{1}{2}}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}})\sqrt{\mathcal{N}_{\infty}(\lambda)}\right)\left(\log\frac{2}{\delta}\right)^2.
\end{aligned}$$

Similarly, if β meets source condition Eq. 9, with $\varphi = \vartheta\psi \in \mathcal{F}(0,c)$ and the qualification of the regularization g_λ covers $\vartheta(t)t^{\frac{3}{2}}$, we have

$$\begin{aligned}
\left\|\beta - \beta_{\mathbf{X}}^{\lambda}\right\|_{\mathcal{H}_K} \leq & c\left[\left(\frac{\mathcal{B}_{n,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2 + 1\right]\varphi(\lambda) + \psi(\kappa_0)\gamma_0\frac{4\kappa_0^2}{\sqrt{n}}\log\frac{2}{\delta} \\
& + c\frac{1}{\sqrt{\lambda}}\sqrt{\left(\frac{\mathcal{B}_{n,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2 + 1}\left(\log^{\frac{1}{2}}\frac{1}{\delta}\right)\left(m^{-\frac{1}{2}} + n^{-\frac{1}{2}}\right)\sqrt{\mathcal{N}_{\infty}(\lambda)} \\
\leq & c\left(\Upsilon(\lambda)\varphi(\lambda) + n^{-\frac{1}{2}} + \frac{1}{\sqrt{\lambda}}[\Upsilon(\lambda)]^{\frac{1}{2}}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}})\sqrt{\mathcal{N}_{\infty}(\lambda)}\right)\left(\log\frac{2}{\delta}\right)^2.
\end{aligned}$$

■

We will also need the following statement proven in (Lu et al., 2020) as Lemma 4.6.

Lemma 11 (Lu et al. (2020)) *There exists λ_* satisfying $\mathcal{N}(\lambda_*)/\lambda_* = n$. For $\lambda_* \leq \lambda \leq \kappa_0$,*

$$\mathcal{B}_{n,\lambda} \leq \frac{2\kappa_0}{\sqrt{n}}\left(\sqrt{2}\kappa_0 + \sqrt{\mathcal{N}(\lambda)}\right). \tag{19}$$

This yields

$$\Upsilon(\lambda) \leq 1 + (4\kappa_0^2 + 2\kappa_0)^2 \tag{20}$$

and also

$$\mathcal{B}_{n,\lambda}\left(\mathcal{B}_{n,\lambda} + \sqrt{\lambda}\right) \leq (1+4\kappa_0)^4 \min\left\{\lambda, \sqrt{\frac{\kappa_0}{n}}\right\}, \tag{21}$$

for n large enough.

For $\lambda > \lambda_*$ we can make the statement of Proposition 10 more transparent.

Theorem 12 *Let K satisfy Assumption 1, and $\lambda \geq \lambda^*$. Then under the assumptions of Proposition 10, with probability at least $1-\delta$, it holds*

$$\left\|\beta - \beta_{\mathbf{X}}^{\lambda}\right\|_{\mathcal{H}_K} \leq c\left(\varphi(\lambda) + (m^{-\frac{1}{2}} + n^{-\frac{1}{2}})\frac{\xi(\lambda)}{\lambda}\right)\left(\log\frac{2}{\delta}\right)^2.$$

Consider $\theta_{\varphi,\xi}(t) = \frac{\varphi(t)t}{\xi(t)}$ and $\lambda = \lambda_{m,n} = \theta_{\varphi,\xi}^{-1}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}})$, then

$$\left\|\beta - \beta_{\mathbf{X}}^{\lambda}\right\|_{\mathcal{H}_K} \leq c\varphi\left(\theta_{\varphi,\xi}^{-1}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}})\right)\log^2\frac{1}{\delta}.$$

Remark 13 *As we already mentioned, the accuracy of the approximation Eq. 6 in RKHS has also been estimated in Theorem 2 of Gizewski et al. (2022). In our terms, the result of Gizewski et al. (2022) can be written as follows:*

$$\left\|\beta - \beta_{\mathbf{X}}^{\lambda}\right\|_{\mathcal{H}_K} \leq c\varphi\left(\theta_{\varphi}^{-1}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}})\right)\log\frac{1}{\delta}. \tag{22}$$

where $\theta_\varphi(t) = \varphi(t)t$. To simplify the comparison of Theorem 12 and Eq. 22, let us consider the case when β meets the source condition Eq. 9 with $\varphi(t) = t^\eta$. In this case the bound Eq. 22 can be reduced to

$$\left\|\beta - \beta_{\mathbf{X}}^{\lambda}\right\|_{\mathcal{H}_K} = O\left((m^{-\frac{1}{2}} + n^{-\frac{1}{2}})^{\frac{\eta}{\eta+1}}\right). \tag{23}$$

It is noteworthy that the error bound established in Theorem 2 of Gizewski et al. (2022) does not take into consideration the capacity of $\mathcal{H}_K$. Such an additional factor can be accounted for in terms of Assumption 1. Assume that K satisfies Assumption 1 with $\xi(t) = t^\varsigma, 0 < \varsigma \leq \frac{1}{2}$, then for $\lambda = \lambda_{m,n} = \theta_{\varphi,\xi}^{-1}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}})$, the bound in Theorem 12 gives

$$\left\|\beta - \beta_{\mathbf{X}}^{\lambda}\right\|_{\mathcal{H}_K} = O\left((m^{-\frac{1}{2}} + n^{-\frac{1}{2}})^{\frac{\eta}{\eta+1-\varsigma}}\right),$$

that is better than the order of accuracy given by Eq. 23. Then one can conclude that the bound in Theorem 12 obtained by our argument generalizes, specifies, and refines the results of Gizewski et al. (2022).

Recall that the bounds in Theorem 12 are valid for $\lambda > \lambda_*$. Using Lemma 5 and 11 one can prove that $\lambda = \lambda_{m,n}$ also satisfies the above inequality. The corresponding proof can be easily recovered from Figure 1.

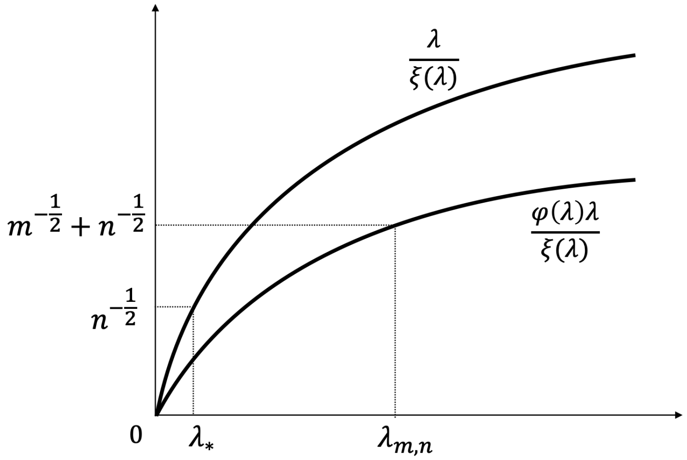


Figure 1: Relation between λ_* and $\lambda_{m,n}$.

5. Error Bounds for The Pointwise Evaluation

In this section, we discuss the error between point values of $\beta(x)$ and $\beta^\lambda_{\mathbf{X}}(x)$ for all $x \in \mathbf{X}$. In view of the reproducing property of K and Assumption 1 we have

$$\begin{aligned}|\beta(x) - \beta^\lambda_{\mathbf{X}}(x)| = \left| \left\langle K_x, \beta - \beta^\lambda_{\mathbf{X}} \right\rangle_{\mathcal{H}_K} \right| &= \left| \left\langle K(\cdot, x), \beta - \beta^\lambda_{\mathbf{X}} \right\rangle_{\mathcal{H}_K} \right| \\ &= \left| \left\langle \xi(J_p^* J_p) v_x, \beta - \beta^\lambda_{\mathbf{X}} \right\rangle_{\mathcal{H}_K} \right| \\ &\le c \left\| \xi(J_p^* J_p)(\beta - \beta^\lambda_{\mathbf{X}}) \right\|_{\mathcal{H}_K}. \end{aligned} \tag{24}$$

Similarly, we obtain

$$\begin{aligned} |\beta(x) - \bar{\beta}^\lambda(x)| &\le c \left\| \xi(J_p^* J_p)(\beta - \bar{\beta}^\lambda) \right\|_{\mathcal{H}_K}, \\ |\bar{\beta}^\lambda(x) - \beta^\lambda_{\mathbf{X}}(x)| &\le c \left\| \xi(J_p^* J_p)(\bar{\beta}^\lambda - \beta^\lambda_{\mathbf{X}}) \right\|_{\mathcal{H}_K}, \end{aligned}$$

that allows for the following decomposition of the error-bound

$$|\beta(x) - \beta^\lambda_{\mathbf{X}}(x)| \le c \left(\left\| \xi(J_p^* J_p)(\beta - \bar{\beta}^\lambda) \right\|_{\mathcal{H}_K} + \left\| \xi(J_p^* J_p)(\bar{\beta}^\lambda - \beta^\lambda_{\mathbf{X}}) \right\|_{\mathcal{H}_K} \right). \tag{25}$$

The bound Eq. 25 implies that in the pointwise evaluation setting, we in fact approximate the elements $\beta_\xi = \xi(J_p^* J_p)\beta$, which in view of Eq. 9 satisfy the source condition

$$\beta_\xi = \varphi_\xi(J_p^* J_p)\nu_q$$

with the index functions $\varphi_\xi(t) = \xi(t)\varphi(t)$. According to Assumption 1, $\xi^2(t)$ is covered by the qualification $s = 1$, which implies that the highest decay rate of $\xi(t)$ to zero as $t \to 0$ is $O(t^{\frac{1}{2}})$. Moreover, for $\varphi \in \mathcal{F}(0, c)$ we have $\varphi_\xi(t) = \xi(t)\vartheta(t)\psi(t)$, where $\psi(t)$ is assumed to be an operator monotone and, as it has been mentioned above with the reference to (Mathé and Pereverzev, 2003), $\psi(t)$ cannot converge to zero faster than linearly. Then the highest decay rate of $\varphi_\xi(t)$ to zero is $O(\vartheta(t)t^{\frac{3}{2}})$. Therefore, to guarantee that the qualification of a regularization g_λ is able to cover the index functions $\varphi_\xi(t) = \xi(t)\vartheta(t)\psi(t)$ appearing in the pointwise evaluation setting it is sufficient to require that the qualification of g_λ covers the index functions $\vartheta(t)t^{\frac{3}{2}}$, which is what is assumed in Proposition 7 and the next proposition estimating the first term in the right-hand side of Eq. 25.

Proposition 14 *Let Assumption 1 be satisfied. Assume also the conditions of Proposition 7. Then*

$$\left\|\xi(J_p^*J_p)(\beta - \bar{\beta}^\lambda)\right\|_{\mathcal{H}_K} \le c\xi(\lambda)\left(\Xi^{\frac{3}{2}}\varphi(\lambda) + (\gamma_0 + \gamma_{\frac{1}{2}})\Xi^{\frac{1}{2}}\psi(\kappa_0)\left\|J_p^*J_p - S_{X_p}^*S_{X_p}\right\|_{\mathcal{H}_K\to\mathcal{H}_K}\right).$$

for $\varphi \in \mathcal{F}(0,c)$, while for an operator monotone index function φ we have

$$\left\|\xi(J_p^*J_p)(\beta - \bar{\beta}^\lambda)\right\|_{\mathcal{H}_K} \le c(\gamma_0 + \gamma_{\frac{3}{2}})\Xi^{\frac{3}{2}}\xi(\lambda)\varphi(\lambda).$$

Proof The analysis below is based on a modification of arguments developed in (Lu et al., 2020) for estimating the $L_{2,p}$-norm of any function $f \in L_{2,p}$ in terms of $\left\|(J_p^*J_p)^{\frac{1}{2}}f\right\|_{\mathcal{H}_K}$. For the reader's convenience, we present this modification in detail.

First of all, directly from Lemma A.1 (Lu et al., 2020), it follows that, if g_λ is any regularization with qualification 1, then

$$\left\|r_\lambda(S_{X_p}^*S_{X_p})(\lambda I + S_{X_p}^*S_{X_p})^{\frac{1}{2}}\right\|_{\mathcal{H}_K\to\mathcal{H}_K} \le \sqrt{(\gamma_0^2 + \gamma_{\frac{1}{2}}^2)}\lambda^{\frac{1}{2}}. \tag{26}$$

If g_λ has qualification at least 3/2 then

$$\left\|r_\lambda(S_{X_p}^*S_{X_p})(\lambda I + S_{X_p}^*S_{X_p})^{\frac{3}{2}}\right\|_{\mathcal{H}_K\to\mathcal{H}_K} \le \sqrt{8(\gamma_0^2 + \gamma_{\frac{3}{2}}^2)}\lambda^{\frac{3}{2}}. \tag{27}$$

If β meets source condition Eq. 9, then for any $\lambda > 0$

$$\begin{aligned}\left\|\xi(J_p^*J_p)(\beta - \bar{\beta}^\lambda)\right\|_{\mathcal{H}_K} &= \left\|\xi(J_p^*J_p)(I - g_\lambda(S_{X_p}^*S_{X_p})S_{X_p}^*S_{X_p})\beta\right\|_{\mathcal{H}_K}\\ &= \left\|\xi(J_p^*J_p)(I - g_\lambda(x^*S_{X_p})S_{X_p}^*S_{X_p})\varphi(J_p^*J_p)v\right\|_{\mathcal{H}_K}\\ &\le \left\|\xi(J_p^*J_p)(\lambda I + J_p^*J_p)^{-\frac{1}{2}}\right\|_{\mathcal{H}_K\to\mathcal{H}_K}\left\|(\lambda I + J_p^*J_p)^{\frac{1}{2}}r_\lambda(S_{X_p}^*S_{X_p})\varphi(J_p^*J_p)v\right\|_{\mathcal{H}_K}. \end{aligned} \tag{28}$$

Now we are going to estimate each component on the right-hand side of Eq. 28. From the proof of Lemma 5, we have

$$\left\|\xi(J_p^*J_p)(\lambda I + J_p^*J_p)^{-\frac{1}{2}}\right\|_{\mathcal{H}_K\to\mathcal{H}_K} \le c\frac{\xi(\lambda)}{\sqrt{\lambda}}. \tag{29}$$

Observe also that

$$\begin{aligned}&\left\|(\lambda I + J_p^*J_p)^{\frac{1}{2}}r_\lambda(S_{X_p}^*S_{X_p})\varphi(J_p^*J_p)v\right\|_{\mathcal{H}_K}\\ &\quad\le \left\|(\lambda I + J_p^*J_p)^{\frac{1}{2}}r_\lambda(S_{X_p}^*S_{X_p})(\lambda I + J_p^*J_p)\right\|_{\mathcal{H}_K\to\mathcal{H}_K}\left\|(\lambda I + J_p^*J_p)^{-1}\varphi(J_p^*J_p)v\right\|_{\mathcal{H}_K}.\end{aligned}$$

Moreover, using Eq. 17 and the bounds Eq. 18 and Eq. 27, we get

$$
\begin{aligned}
&\left\|(\lambda I + J_p^*J_p)^{\frac{1}{2}} r_\lambda(S_{X_p}^*S_{X_p})(\lambda I + J_p^*J_p)\right\|_{\mathcal{H}_K\to\mathcal{H}_K} \\
&\quad\le \left\|(\lambda I + J_p^*J_p)^{\frac{1}{2}}(\lambda I + S_{X_p}^*S_{X_p})^{-\frac{1}{2}}(\lambda I + S_{X_p}^*S_{X_p})^{\frac{1}{2}} r_\lambda(S_{X_p}^*S_{X_p})(\lambda I + J_p^*J_p)\right\|_{\mathcal{H}_K\to\mathcal{H}_K} \\
&\quad\le \Xi^{\frac{1}{2}}\left\|(\lambda I + S_{X_p}^*S_{X_p})^{\frac{1}{2}} r_\lambda(S_{X_p}^*S_{X_p})(\lambda I + S_{X_p}^*S_{X_p})(\lambda I + S_{X_p}^*S_{X_p})^{-1}(\lambda I + J_p^*J_p)\right\|_{\mathcal{H}_K\to\mathcal{H}_K} \\
&\quad\le \Xi^{\frac{1}{2}}\left\|(\lambda I + S_{X_p}^*S_{X_p})^{\frac{1}{2}} r_\lambda(S_{X_p}^*S_{X_p})(\lambda I + S_{X_p}^*S_{X_p})\right\|_{\mathcal{H}_K\to\mathcal{H}_K}\Xi \\
&\quad= \left\|r_\lambda(S_{X_p}^*S_{X_p})(\lambda I + S_{X_p}^*S_{X_p})^{\frac{3}{2}}\right\|_{\mathcal{H}_K\to\mathcal{H}_K}\Xi^{\frac{3}{2}} \\
&\quad\le c\Xi^{\frac{3}{2}}\lambda^{\frac{3}{2}}\left(\gamma_0 + \gamma_{\frac{3}{2}}\right).
\end{aligned}
$$

Besides, using the same argument as in the proof of Lemma 5, for an operator monotone index function φ we have

$$
\left\|(\lambda I + J_p^*J_p)^{-1}\varphi(J_p^*J_p)v\right\|_{\mathcal{H}_K} \le c\frac{\varphi(\lambda)}{\lambda}\left\|v\right\|_{\mathcal{H}_K}. \tag{30}
$$

Thus,

$$
\left\|(\lambda I + J_p^*J_p)^{\frac{1}{2}} r_\lambda(S_{X_p}^*S_{X_p})\varphi(J_p^*J_p)v\right\|_{\mathcal{H}_K} \le c\left(\gamma_0 + \gamma_{\frac{3}{2}}\right)\Xi^{\frac{3}{2}}\lambda^{\frac{1}{2}}\varphi(\lambda).
$$

Substituting Eq. 29 and the above estimate into Eq. 28, we obtain the second bound of the proposition.

Now we turn to proving the first bound and assume that β meets Eq. 9 with $\varphi = \vartheta\psi$. Then we have

$$
\begin{aligned}
\left\|\xi(J_p^*J_p)(\beta - \bar{\beta}^\lambda)\right\|_{\mathcal{H}_K} \le{}& \left\|\xi(J_p^*J_p) r_\lambda(S_{X_p}^*S_{X_p})\vartheta(S_{X_p}^*S_{X_p})\psi(J_p^*J_p)v\right\|_{\mathcal{H}_K} \\
&+ \left\|\xi(J_p^*J_p) r_\lambda(S_{X_p}^*S_{X_p})(\vartheta(J_p^*J_p) - \vartheta(S_{X_p}^*S_{X_p}))\psi(J_p^*J_p)v\right\|_{\mathcal{H}_K}.
\end{aligned} \tag{31}
$$

Further, we estimate separately each term on the right-hand side of Eq. 31. By using Eq. 29 and Eq. 30, the first term is estimated as follows:

$$
\begin{aligned}
&\left\|\xi(J_p^*J_p) r_\lambda(S_{X_p}^*S_{X_p})\vartheta(S_{X_p}^*S_{X_p})\psi(J_p^*J_p)v\right\|_{\mathcal{H}_K} \\
&\quad\le \left\|\xi(J_p^*J_p)(\lambda I + J_p^*J_p)^{-\frac{1}{2}}\right\|_{\mathcal{H}_K\to\mathcal{H}_K}\left\|(\lambda I + J_p^*J_p)^{\frac{1}{2}} r_\lambda(S_{X_p}^*S_{X_p})\vartheta(S_{X_p}^*S_{X_p})\psi(J_p^*J_p)\right\|_{\mathcal{H}_K\to\mathcal{H}_K} \\
&\qquad\times\left\|(\lambda I + J_p^*J_p)^{-1}\psi(J_p^*J_p)v\right\|_{\mathcal{H}_K} \\
&\quad\le c\frac{\xi(\lambda)}{\sqrt{\lambda}}\left\|(\lambda I + J_p^*J_p)^{\frac{1}{2}} r_\lambda(S_{X_p}^*S_{X_p})\vartheta(S_{X_p}^*S_{X_p})(\lambda I + J_p^*J_p)\right\|_{\mathcal{H}_K\to\mathcal{H}_K}\frac{\psi(\lambda)}{\lambda}\left\|v\right\|_{\mathcal{H}_K} \\
&\quad\le c\frac{\xi(\lambda)\psi(\lambda)}{\lambda^{\frac{3}{2}}}\Xi^{\frac{1}{2}}\left\|(\lambda I + S_{X_p}^*S_{X_p})^{\frac{1}{2}} r_\lambda(S_{X_p}^*S_{X_p})\vartheta(S_{X_p}^*S_{X_p})(\lambda I + S_{X_p}^*S_{X_p})\right\|_{\mathcal{H}_K\to\mathcal{H}_K}\Xi \\
&\quad\le c\frac{\xi(\lambda)\psi(\lambda)}{\lambda^{\frac{3}{2}}}\Xi^{\frac{3}{2}}\vartheta(\lambda)\lambda^{\frac{3}{2}} \\
&\quad\le c\xi(\lambda)\varphi(\lambda)\Xi^{\frac{3}{2}},
\end{aligned} \tag{32}
$$

and with the use of Eq. 26 we can estimate the second term in Eq. 31 as

$$\begin{aligned}
&\left\|\xi(J_p^*J_p)r_\lambda(S_{X_p}^*S_{X_p})(\vartheta(J_p^*J_p)-\vartheta(S_{X_p}^*S_{X_p}))\psi(J_p^*J_p)v\right\|_{\mathcal{H}_K}\\
&\quad\le\left\|\xi(J_p^*J_p)(\lambda I+J_p^*J_p)^{-\frac{1}{2}}(\lambda I+J_p^*J_p)^{\frac{1}{2}}r_\lambda(S_{X_p}^*S_{X_p})(\vartheta(J_p^*J_p)-\vartheta(S_{X_p}^*S_{X_p}))\psi(J_p^*J_p)v\right\|_{\mathcal{H}_K}\\
&\quad\le c\frac{\xi(\lambda)}{\sqrt{\lambda}}\left\|(\lambda I+J_p^*J_p)^{\frac{1}{2}}r_\lambda(S_{X_p}^*S_{X_p})\right\|_{\mathcal{H}_K\to\mathcal{H}_K}\left\|(\vartheta(J_p^*J_p)-\vartheta(S_{X_p}^*S_{X_p}))\right\|_{\mathcal{H}_K\to\mathcal{H}_K}\\
&\quad\le c\frac{\xi(\lambda)}{\sqrt{\lambda}}\left\|(\lambda I+J_p^*J_p)^{\frac{1}{2}}r_\lambda(S_{X_p}^*S_{X_p})\right\|_{\mathcal{H}_K\to\mathcal{H}_K}\left\|J_p^*J_p-S_{X_p}^*S_{X_p}\right\|_{\mathcal{H}_K\to\mathcal{H}_K}\psi(\kappa_0)\left\|v\right\|_{\mathcal{H}_K}\\
&\quad\le c\frac{\xi(\lambda)}{\sqrt{\lambda}}\Xi^{\frac{1}{2}}\left\|(\lambda I+S_{X_p}^*S_{X_p})^{\frac{1}{2}}r_\lambda(S_{X_p}^*S_{X_p})\right\|_{\mathcal{H}_K\to\mathcal{H}_K}\left\|J_p^*J_p-S_{X_p}^*S_{X_p}\right\|_{\mathcal{H}_K\to\mathcal{H}_K}\\
&\quad\le c\xi(\lambda)\Xi^{\frac{1}{2}}\left(\gamma_0+\gamma_{\frac{1}{2}}\right)\left\|J_p^*J_p-S_{X_p}^*S_{X_p}\right\|_{\mathcal{H}_K\to\mathcal{H}_K}. \qquad (33)
\end{aligned}$$

Substituting Eq. 32 and Eq. 33 into Eq. 31, we obtain

$$\left\|\xi(J_p^*J_p)(\beta-\bar{\beta}^\lambda)\right\|_{\mathcal{H}_K}\le c\xi(\lambda)\left(\Xi^{\frac{3}{2}}\varphi(\lambda)+(\gamma_0+\gamma_{\frac{1}{2}})\Xi^{\frac{1}{2}}\left\|J_p^*J_p-S_{X_p}^*S_{X_p}\right\|_{\mathcal{H}_K\to\mathcal{H}_K}\right).$$

■

Next proposition estimates the second term in the right-hand side of Eq. 25.

Proposition 15 *Assume Assumption 1 be satisfied. Then it holds*

$$\left\|\xi(J_p^*J_p)(\bar{\beta}^\lambda-\beta_{\mathbf{X}}^\lambda)\right\|_{\mathcal{H}_K}\le c\frac{\xi(\lambda)}{\sqrt{\lambda}}\Xi(\gamma_{-1}+\gamma_0+1)\left\|(\lambda I+J_p^*J_p)^{-\frac{1}{2}}(S_{X_q}^*S_{X_q}\mathbf{1}-S_{X_p}^*S_{X_p}\beta)\right\|_{\mathcal{H}_K}.$$

Proof Using (9), (17) and (29), we derive

$$\begin{aligned}
\left\|\xi(J_p^*J_p)(\bar{\beta}^\lambda-\beta_{\mathbf{X}}^\lambda)\right\|_{\mathcal{H}_K}&\le\left\|\xi(J_p^*J_p)g_\lambda(S_{X_p}^*S_{X_p})(S_{X_q}^*S_{X_q}\mathbf{1}-S_{X_p}^*S_{X_p}\beta)\right\|_{\mathcal{H}_K}\\
&\le\left\|\xi(J_p^*J_p)(\lambda I+J_p^*J_p)^{-\frac{1}{2}}\right\|_{\mathcal{H}_K\to\mathcal{H}_K}\times\\
&\qquad\times\left\|(\lambda I+J_p^*J_p)^{\frac{1}{2}}g_\lambda(S_{X_p}^*S_{X_p})(S_{X_q}^*S_{X_q}\mathbf{1}-S_{X_p}^*S_{X_p}\beta)\right\|_{\mathcal{H}_K}\\
&\le c\frac{\xi(\lambda)}{\sqrt{\lambda}}\left\|(\lambda I+J_p^*J_p)^{\frac{1}{2}}g_\lambda(S_{X_p}^*S_{X_p})(\lambda I+J_p^*J_p)^{\frac{1}{2}}\right\|_{\mathcal{H}_K\to\mathcal{H}_K}\times\\
&\qquad\times\left\|(\lambda I+J_p^*J_p)^{-\frac{1}{2}}(S_{X_q}^*S_{X_q}\mathbf{1}-S_{X_p}^*S_{X_p}\beta)\right\|_{\mathcal{H}_K}\\
&\le c\frac{\xi(\lambda)}{\sqrt{\lambda}}\Xi^{\frac{1}{2}}\left\|g_\lambda(S_{X_p}^*S_{X_p})(\lambda I+S_{X_p}^*S_{X_p})\right\|_{\mathcal{H}_K}\Xi^{\frac{1}{2}}\times\\
&\qquad\times\left\|(\lambda I+J_p^*J_p)^{-\frac{1}{2}}(S_{X_q}^*S_{X_q}\mathbf{1}-S_{X_p}^*S_{X_p}\beta)\right\|_{\mathcal{H}_K}\\
&\le c\frac{\xi(\lambda)}{\sqrt{\lambda}}\Xi\sup_t\left|g_\lambda(t)(\lambda+t)\right|\left\|(\lambda I+J_p^*J_p)^{-\frac{1}{2}}(S_{X_q}^*S_{X_q}\mathbf{1}-S_{X_p}^*S_{X_p}\beta)\right\|_{\mathcal{H}_K}\\
&\le c\frac{\xi(\lambda)}{\sqrt{\lambda}}\Xi(\gamma_{-1}+\gamma_0+1)\left\|(\lambda I+J_p^*J_p)^{-\frac{1}{2}}(S_{X_q}^*S_{X_q}\mathbf{1}-S_{X_p}^*S_{X_p}\beta)\right\|_{\mathcal{H}_K}.
\end{aligned}$$

■

Now we can combine Eq. 25 with Propositions 14, 15 and with Lemma 1. Then the same argument as in the proof of Proposition 10 gives us the following statement.

Theorem 16 *Under the assumption of Propositions 14 and 15, for $\lambda > \lambda_*$ with probability at least $1-\delta$, for all $x \in \mathbf{X}$, we have*

$$|\beta(x) - \beta^{\lambda}_{\mathbf{X}}(x)| \leq c\xi(\lambda)\left(\varphi(\lambda) + (m^{-\frac{1}{2}} + n^{-\frac{1}{2}})\frac{\xi(\lambda)}{\lambda}\right)\left(\log\frac{2}{\delta}\right)^2,$$

and for $\lambda = \lambda_{m,n} = \theta^{-1}_{\varphi,\xi}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}})$,

$$|\beta(x) - \beta^{\lambda}_{\mathbf{X}}(x)| \leq c\xi(\lambda_{m,n})\varphi(\lambda_{m,n})\log^2\frac{1}{\delta}.$$

Remark 17 *Let us consider the same index functions $\varphi(t) = t^{\eta}$ and $\xi(t) = t^{\varsigma}$ as in Remark 13, where the accuracy of order $O\left((m^{-\frac{1}{2}} + n^{-\frac{1}{2}})^{\frac{\eta}{\eta+1-\varsigma}}\right)$ has been derived for Eq. 6. Under the same assumptions, Theorem 16 guarantees the accuracy of order $O\left((m^{-\frac{1}{2}} + n^{-\frac{1}{2}})^{\frac{\eta+\varsigma}{\eta+1-\varsigma}}\right)$. This illustrates that the reconstruction of the Radon–Nikodym derivative at any particular point can be done with much higher accuracy than its reconstruction as an element of RKHS. But let us stress that the above high order of accuracy is guaranteed when the qualification s of the used regularization scheme is higher than that of the Lavrentiev regularization or KuLSIF Eq. 3.*

6. Numerical Illustrations

In our examples, we simulate inputs $X_p = (x_1, x_2, \ldots, x_n)$ to be sampled from the normal distribution $p \sim N(2,5)$, while the inputs $X_q = (x'_1, x'_2, \ldots, x'_m)$ are sampled from the normal distribution $q \sim N(\mu_q, 0.5)$ with $\mu_q = \{2,3,4\}$. In this case, the Radon–Nikodym derivative $\beta = \frac{dq}{dp}$ is known to be

$$\beta(x) = \sqrt{10}e^{\frac{(x-2)^2 - 10(x-\mu_q)^2}{10}}.$$

In the algorithms described in Section 3, we choose the kernel as

$$K(x,x') = 1 + e^{-\frac{(x-x')^2}{2}},$$

which is a combination of a universal Gaussian kernel with a constant such that the corresponding space $\mathcal{H}_K$ contains all constant functions.

We are going to illustrate that to achieve a high order of accuracy for the reconstruction of the Radon–Nikodym derivative at any particular point, as it is guaranteed by Theorem 16, one needs to employ a regularization with the qualification that is higher than 1. For doing this we use a particular case of the general regularization scheme Eq. 6, namely the iterated Lavrentiev regularization, to compute the values of the approximate Radon–Nikodym derivative $\beta^\lambda_{\mathbf{X}} = (\beta^\lambda_1, \beta^\lambda_2, \cdots, \beta^\lambda_n)$ with $\beta^\lambda_i = \beta^\lambda_{\mathbf{X}}(x_i)$.

Recall that the k times iterated Lavrentiev regularization is indexed by the functions

$$g_\lambda(t) = g_{\lambda,k}(t) = \left(1 - \frac{\lambda^k}{(\lambda+t)^k}\right) t^{-1}, \tag{34}$$

and has the qualification $s = k$.

For $g_\lambda(t) = g_{\lambda,k}(t)$ the vector of values of the approximate Radon–Nikodym derivative $\beta^\lambda_{\mathbf{X}} = \beta^{\lambda,k}_{\mathbf{X}}$ given by Eq. 6, Eq. 34 is the k-th term of the sequence

$$\begin{aligned} &\beta^{\lambda,l}_{\mathbf{X},0} = 0, \\ &\beta^{\lambda,l}_{\mathbf{X}} = (n\lambda\mathbf{I} + \mathbf{K})^{-1}\left(n\lambda\beta^{\lambda,l-1}_{\mathbf{X}} + \bar{F}\right), \qquad l = 1, 2, \ldots, k. \end{aligned} \tag{35}$$

where $\mathbf{I}$ is n by n identity matrix, $\mathbf{K} = (K(x_i, x_j))^n_{i,j=1}$, and $\bar{F} = (F_i)^n_{i=1}$ with $F_i = \frac{n}{m}\sum_{j=1}^m K(x_i, x'_j)$.

The algorithm Eq. 35 has been implemented with $m = n = 100$ and $k = \{1, 2, 3, 5, 10\}$. The regularization parameter λ is chosen by the so-called quasi-optimality criterion (see, for example, Bauer and Reiß (2008), Kindermann et al. (2018)), $\bar{\lambda} \in \{\lambda_\iota = \lambda_0\varrho^\iota, \iota = 1, 2, \ldots, w\}$, $\varrho < 1$ such that for $\bar{\lambda} = \lambda_{\iota_0}$,

$$\left\|\beta^{\lambda_{\iota_0}}_{\mathbf{X}} - \beta^{\lambda_{\iota_0 -1}}_{\mathbf{X}}\right\|_{\mathbb{R}^n} = \min\left\{\left\|\beta^{\lambda_\iota}_{\mathbf{X}} - \beta^{\lambda_{\iota-1}}_{\mathbf{X}}\right\|_{\mathbb{R}^n}, \iota = 1, 2, \ldots, w\right\}.$$

Taking into consideration Theorem 16 and Figure 1, one can expect that $\bar{\lambda} \approx \lambda_{m,n} > (m^{-\frac{1}{2}} + n^{-\frac{1}{2}})$. Therefore, for $n = m = 100$ it is natural to look for $\bar{\lambda}$ within interval $[0.1, 0.9]$, and in our experiments we choose $\lambda_0 = 0.9$, $\varrho = \sqrt[9]{\frac{1}{9}}$, and $w = 9$, such that $\lambda_\iota \in [0.1, 0.9]$.

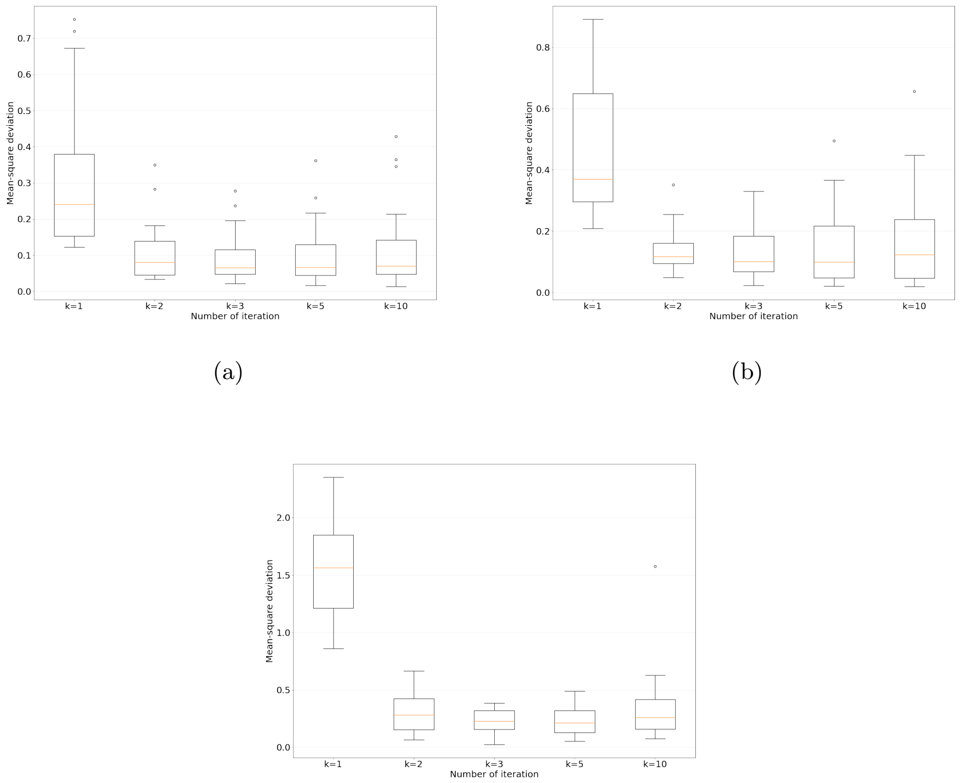
The performance of each implementation has been measured in terms of the mean-square deviation (MSD).

$$MSD = n^{-1}\sum_{i=1}^n \left(\beta(x_i) - \beta^{\lambda,k}_{\mathbf{X}}(x_i)\right)^2.$$

A summary of the performance over 20 simulations of $(x_i)^n_{i=1}$, $(x'_j)^m_{j=1}$ in all cases $\mu_q = \{2, 3, 4\}$ is presented in the form of box plots in Figure 2. It can be clear seen that in our examples the considered realization of the iterated Laventiev regularization outperforms its original version ($k = 1$). This supports a conclusion from Theorem 16 suggesting the use of high qualification regularization for pointwise evaluation of Radon–Nikodym derivation.

The performance of the algorithm in Eq. 35 for a particular simulation is displayed in Figure 3. In this figure, the exact values β are shown by the line, and the $\beta^{\lambda,1}_{\mathbf{X}}(x_i)$, $\beta^{\lambda,2}_{\mathbf{X}}(x_i)$, $\beta^{\lambda,3}_{\mathbf{X}}(x_i)$, $\beta^{\lambda,5}_{\mathbf{X}}(x_i)$, and $\beta^{\lambda,10}_{\mathbf{X}}(x_i)$ are denoted correspondingly by green triangles, red squares, cyan diamonds, yellow stars, and blue crosses.

Acknowledgments



(a)

(b)

(c)

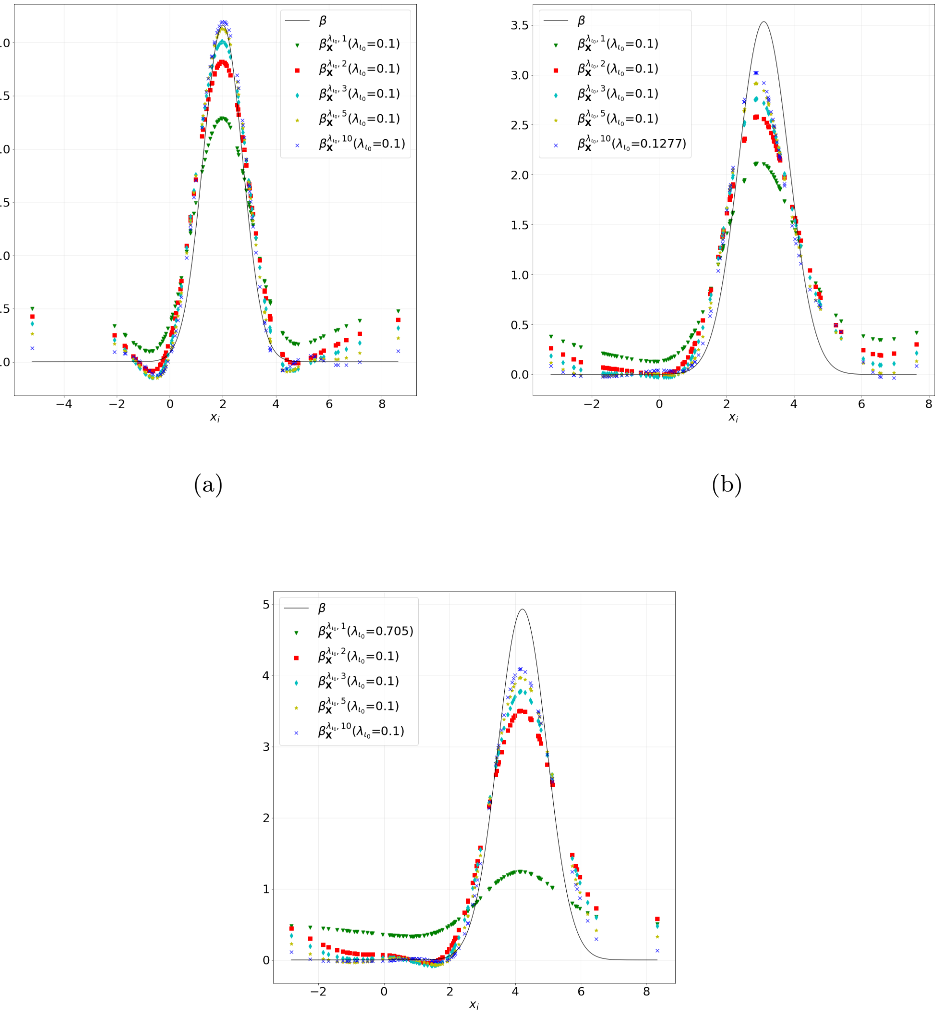
Figure 2: Mean-square deviation in examples with (a) $X_q \sim N(2, 0.5)$, (b) $X_q \sim N(3, 0.5)$, and (c) $X_q \sim N(4, 0.5)$.

The authors are very grateful to two anonymous referees for their valuable comments which improved the paper substantially. In particular, due to a comment of one of the referees, we have realized that Proposition 7 can be proved under the relaxed assumption (see the remark after Proposition 7).

The research reported in this paper has been funded by the Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK), the Federal Ministry for Digital and Economic Affairs (BMDW), and the Province of Upper Austria in the frame of the COMET–Competence Centers for Excellent Technologies Programme and the COMET Module S3AI managed by the Austrian Research Promotion Agency FFG.

References

Frank Bauer and Markus Reiß. Regularization independent of the noise level: an analysis of quasi-optimality. *Inverse Problems*, 24(5):055009, aug 2008. 10.1088/0266-5611/24/5/055009. URL `https://dx.doi.org/10.1088/0266-5611/24/5/055009`.



(a)

(b)

(c)

Figure 3: The performance of the algorithm for a particular simulation with (a) $X_q \sim N(2, 0.5)$, (b) $X_q \sim N(3, 0.5)$, and (c) $X_q \sim N(4, 0.5)$.

Frank Bauer, Sergei V. Pereverzyev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *J. Complex.*, 23:52–72, 2007.

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

Ernesto De Vito, Lorenzo Rosasco, and Alessandro Toigo. Learning sets with separating kernels. *Applied and Computational Harmonic Analysis*, 37(2):185–217, 2014. ISSN 1063-5203. https://doi.org/10.1016/j.acha.2013.11.003.

Elke R. Gizewski, Lukas Mayer, Bernhard A. Moser, Duc Hoan Nguyen, Sergiy Pereverzyev Jr, Sergei V. Pereverzyev, Natalia Shepeleva, and Werner Zellinger. On a regularization of unsupervised domain adaptation in RKHS. *Applied and Computational Harmonic Analysis*, 57:201–227, 2022. ISSN 1063-5203. https://doi.org/10.1016/j.acha.2021.12.002.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19, 2006.

Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86:335–367, 2012. 10.1007/s10994-011-5266-3.

Stefan Kindermann, Sergiy Pereverzyev Jr, and Andrey Pilipenko. The quasi-optimality criterion in the linear functional strategy. *Inverse Problems*, 34(7):075001, may 2018. 10.1088/1361-6420/aabe4f. URL `https://dx.doi.org/10.1088/1361-6420/aabe4f`.

Shuai Lu and Sergei Pereverzyev. *Regularization theory for ill-posed problems. Selected topics*. De Gruyter, 01 2013. ISBN 978-3-11-0286645-5.

Shuai Lu, Peter Mathé, and Sergiy Pereverzyev. Analysis of regularized Nyström subsampling for regression functions of low smoothness. *Analysis and Applications*, 17(06): 931–946, 2019. 10.1142/S0219530519500039.

Shuai Lu, Peter Mathé, and Sergei V. Pereverzev. Balancing principle in supervised learning for a general regularization scheme. *Applied and Computational Harmonic Analysis*, 48 (1):123–148, 2020. ISSN 1063-5203. https://doi.org/10.1016/j.acha.2018.03.001.

Peter Mathé and Bernd Hofmann. How general are general source conditions? *Inverse Problems*, 24(1):015009, jan 2008. 10.1088/0266-5611/24/1/015009. URL `https://dx.doi.org/10.1088/0266-5611/24/1/015009`.

Peter Mathé and Sergei V Pereverzev. Discretization strategy for linear ill-posed problems in variable hilbert scales. *Inverse Problems*, 19(6):1263, oct 2003. 10.1088/0266-5611/19/6/003.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010. 10.1109/TIT.2010.2068870.

Luca Oneto, Sandro Ridella, and Davide Anguita. Tikhonov, Ivanov and Morozov regularization for support vector machine learning. *Machine Learning*, 103(1):103–136, 2016.

Stephen Page and Steffen Grünewälder. Ivanov-regularised least-squares estimators over large rkhss and their interpolation spaces. *J. Mach. Learn. Res.*, 20(120):1–49, 2019.

Edouard Pauwels, Francis Bach, and Jean-Philippe Vert. Relating leverage scores and density using regularized christoffel functions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Sergei Pereverzyev. *An Introduction to Artificial Intelligence Based on Reproducing Kernel Hilbert Spaces.* Springer International Publishing, 05 2022. ISBN 978-3-030-98315-4. https://doi.org/10.1007/978-3-030-98316-1.

Qichao Que and Mikhail Belkin. Inverse density as an inverse problem: The Fredholm equation approach. *Advances in Neural Information Processing Systems*, 26:., 2013.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Ingmar Schuster, Mattes Mollenhauer, Stefan Klus, and Krikamol Muandet. Kernel conditional density operators. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 993–1004. PMLR, 26–28 Aug 2020.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007. 10.1007/s00365-006-0659-y. URL `https://doi.org/10.1007/s00365-006-0659-y`.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning.* Cambridge University Press, 2012. 10.1017/CBO9781139035613.