Journal of Machine Learning Research 18 (2017) 1-59 Submitted 1/16; Revised 4/17; Published 7/17

Density Estimation in Infinite Dimensional Exponential Families

Bharath Sriperumbudur BKS18@PSU.EDU
Department of Statistics, Pennsylvania State University
University Park, PA 16802, USA.

Kenji Fukumizu FUKUMIZU@ISM.AC.JP
The Institute of Statistical Mathematics
10-3 Midoricho, Tachikawa, Tokyo 190-8562 Japan.

Arthur Gretton ARTHUR.GRETTON@GMAIL.COM
ORCID 0000-0003-3169-7624
Gatsby Computational Neuroscience Unit, University College London
Sainsbury Wellcome Centre, 25 Howland Street, London W1T 4JG, UK

Aapo Hyvärinen A.HYVARINEN@UCL.AC.UK
Gatsby Computational Neuroscience Unit, University College London
Sainsbury Wellcome Centre, 25 Howland Street, London W1T 4JG, UK

Revant Kumar RKUMAR74@GATECH.EDU
College of Computing, Georgia Institute of Technology
801 Atlantic Drive, Atlanta, GA 30332, USA.

Editor: Ingo Steinwart

Abstract

In this paper, we consider an infinite dimensional exponential family $\mathcal{P}$ of probability densities, which are parametrized by functions in a reproducing kernel Hilbert space $\mathcal{H}$, and show it to be quite rich in the sense that a broad class of densities on $\mathbb{R}^d$ can be approximated arbitrarily well in Kullback-Leibler (KL) divergence by elements in $\mathcal{P}$. Motivated by this approximation property, the paper addresses the question of estimating an unknown density p_0 through an element in $\mathcal{P}$. Standard techniques like maximum likelihood estimation (MLE) or pseudo MLE (based on the method of sieves), which are based on minimizing the KL divergence between p_0 and $\mathcal{P}$, do not yield practically useful estimators because of their inability to efficiently handle the log-partition function. We propose an estimator $\hat{p}_n$ based on minimizing the *Fisher divergence*, $J(p_0\|p)$ between p_0 and $p \in \mathcal{P}$, which involves solving a simple finite-dimensional linear system. When $p_0 \in \mathcal{P}$, we show that the proposed estimator is consistent, and provide a convergence rate of $n^{-\min\left\{\frac{2}{3},\frac{2\beta+1}{2\beta+2}\right\}}$ in Fisher divergence under the smoothness assumption that $\log p_0 \in \mathcal{R}(C^\beta)$ for some $\beta \geq 0$, where C is a certain Hilbert-Schmidt operator on $\mathcal{H}$ and $\mathcal{R}(C^\beta)$ denotes the image of C^β. We also investigate the misspecified case of $p_0 \notin \mathcal{P}$ and show that $J(p_0\|\hat{p}_n) \to \inf_{p\in\mathcal{P}} J(p_0\|p)$ as $n \to \infty$, and provide a rate for this convergence under a similar smoothness condition as above. Through numerical simulations we demonstrate that the proposed estimator outperforms the non-parametric kernel density estimator, and that the advantage of the proposed estimator grows as d increases.

Keywords: density estimation, exponential family, Fisher divergence, kernel density estimator, maximum likelihood, interpolation space, inverse problem, reproducing kernel Hilbert space, Tikhonov regularization, score matching

1. Introduction

Exponential families are among the most important classes of parametric models studied in statistics, and include many common distributions such as the normal, exponential, gamma, and Poisson. In its "natural form", the family generated by a probability density q_0 (defined over $\Omega \subseteq \mathbb{R}^d$) and *sufficient statistic*, $T : \Omega \to \mathbb{R}^m$ is defined as

$$\mathscr{P}_{\text{fin}} := \left\{ p_\theta(x) = q_0(x) e^{\theta^T T(x) - A(\theta)},\, x \in \Omega \,:\, \theta \in \Theta \subset \mathbb{R}^m \right\} \tag{1}$$

where $A(\theta) := \log \int_\Omega e^{\theta^T T(x)} q_0(x)\, dx$ is the cumulant generating function (also called the log-partition function), $\Theta \subset \{\theta \in \mathbb{R}^m : A(\theta) < \infty\}$ is the *natural parameter space* and θ is a finite-dimensional vector called the *natural parameter*. Exponential families have a number of properties that make them extremely useful for statistical analysis (see Brown, 1986 for more details).

In this paper, we consider an infinite dimensional generalization (Canu and Smola, 2005; Fukumizu, 2009) of (1),

$$\mathcal{P} = \left\{ p_f(x) = e^{f(x) - A(f)} q_0(x),\, x \in \Omega : f \in \mathcal{F} \right\},$$

where the function space $\mathcal{F}$ is defined as

$$\mathcal{F} = \left\{ f \in \mathcal{H} : e^{A(f)} < \infty \right\}, \text{ with } A(f) := \log \int_\Omega e^{f(x)} q_0(x)\, dx$$

being the cumulant generating function, and $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ a reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950) with k as its reproducing kernel. While various generalizations are possible for different choices of $\mathcal{F}$ (e.g., an Orlicz space as in Pistone and Sempi, 1995), the connection of $\mathcal{P}$ to the natural exponential family in (1) is particularly enlightening when $\mathcal{H}$ is an RKHS. This is due to the reproducing property of the kernel, $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$, through which $k(x, \cdot)$ takes the role of the sufficient statistic. In fact, it can be shown (see Section 3 and Example 1 for more details) that every $\mathscr{P}_{\text{fin}}$ is generated by $\mathcal{P}$ induced by a finite dimensional RKHS $\mathcal{H}$, and therefore the family $\mathcal{P}$ with $\mathcal{H}$ being an infinite dimensional RKHS is a natural infinite dimensional generalization of $\mathscr{P}_{\text{fin}}$. Furthermore, this generalization is particularly interesting as in contrast to $\mathscr{P}_{\text{fin}}$, it can be shown that $\mathcal{P}$ is a rich class of densities (depending on the choice of k and therefore $\mathcal{H}$) that can approximate a broad class of probability densities arbitrarily well (see Propositions 1, 13 and Corollary 2). This generalization is not only of theoretical interest, but also has implications for statistical and machine learning applications. For example, in Bayesian non-parametric density estimation, the densities in $\mathcal{P}$ are chosen as prior distributions on a collection of probability densities (e.g., see van der Vaart and van Zanten, 2008). $\mathcal{P}$ has also found applications in nonparametric hypothesis testing (Gretton et al., 2012; Fukumizu et al., 2008) and dimensionality reduction (Fukumizu et al., 2004, 2009) through the mean and covariance operators, which are obtained as the first and second Fréchet derivatives of $A(f)$ (see Fukumizu, 2009, Section 1.2.3). Recently, the infinite dimensional exponential family, $\mathcal{P}$ has been used to develop a gradient-free adaptive MCMC algorithm based on Hamiltonian Monte Carlo (Strathmann et al., 2015) and also has been used in the context of learning the structure of graphical models (Sun et al., 2015).

Motivated by the richness of the infinite dimensional generalization and its statistical applications, it is of interest to model densities by $\mathcal{P}$, and therefore the goal of this paper is to estimate unknown densities by elements in $\mathcal{P}$ when $\mathcal{H}$ is an infinite dimensional RKHS. Formally, given i.i.d. random samples $(X_a)_{a=1}^n$ drawn from an unknown density p_0, the goal is to estimate p_0 through $\mathcal{P}$. Throughout the paper, we refer to case of $p_0 \in \mathcal{P}$ as *well-specified*, in contrast to the *misspecified* case where $p_0 \notin \mathcal{P}$. The setting is useful because $\mathcal{P}$ is a rich class of densities that can approximate a broad class of probability densities arbitrarily well, hence it may be widely used in place of non-parametric density estimation methods (e.g., kernel density estimation (KDE)). In fact, through numerical simulations, we show in Section 6 that estimating p_0 through $\mathcal{P}$ performs better than KDE, and that the advantage of the proposed estimator grows with increasing dimensionality.

In the finite-dimensional case where $\theta \in \Theta \subset \mathbb{R}^m$, estimating p_θ through maximum likelihood (ML) leads to solving elegant likelihood equations (Brown, 1986, Chapter 5). However, in the infinite dimensional case (assuming $p_0 \in \mathcal{P}$), as in many non-parametric estimation methods, a straightforward extension of maximum likelihood estimation (MLE) suffers from the problem of ill-posedness (Fukumizu, 2009, Section 1.3.1). To address this problem, Fukumizu (2009) proposed a method of sieves involving pseudo-MLE by restricting the infinite dimensional manifold $\mathcal{P}$ to a series of finite-dimensional submanifolds, which enlarge as the sample size increases, i.e., $p_{\hat{f}^{(l)}}$ is the density estimator with

$$\hat{f}^{(l)} = \arg \max_{f\in\mathcal{F}^{(l)}} \frac{1}{n}\sum_{a=1}^{n} f(X_a) - A(f), \tag{2}$$

where $\mathcal{F}^{(l)} = \{f \in \mathcal{H}^{(l)} : e^{A(f)} < \infty\}$ and $(\mathcal{H}^{(l)})_{l=1}^\infty$ is a sequence of finite-dimensional subspaces of $\mathcal{H}$ such that $\mathcal{H}^{(l)} \subset \mathcal{H}^{(l+1)}$ for all $l \in \mathbb{N}$. While the consistency of $p_{\hat{f}^{(l)}}$ is proved in Kullback-Leibler (KL) divergence (Fukumizu, 2009, Theorem 6), the method suffers from many drawbacks that are both theoretical and computational in nature. On the theoretical front, the consistency in Fukumizu (2009, Theorem 6) is established by assuming a decay rate on the eigenvalues of the covariance operator (see (A-2) and the discussion in Section 1.4 of Fukumizu (2009) for details), which is usually difficult to check in practice. Moreover, it is not clear which classes of RKHS should be used to obtain a consistent estimator (Fukumizu, 2009, (A-1)) and the paper does not provide any discussion about the convergence rates. On the practical side, the estimator is not attractive as it can be quite difficult to construct the sequence $(\mathcal{H}^{(l)})_{l=1}^\infty$ that satisfies the assumptions in Fukumizu (2009, Theorem 6). In fact, the impracticality of the estimator, $\hat{f}^{(l)}$ is accentuated by the difficulty in efficiently handling $A(f)$ (though it can be approximated by numerical integration).

A related work was carried out by Barron and Sheu (1991)—also see references therein—where the goal is to estimate a density, p_0 by approximating its logarithm as an expansion in terms of basis functions, such as polynomials, splines or trigonometric series. Similar to

Fukumizu (2009), Barron and Sheu proposed the ML estimator $p_{\hat{f}_m}$, where

$$\hat{f}_m = \arg\max_{f \in \mathcal{F}_m} \frac{1}{n} \sum_{a=1}^{n} f(X_a) - A(f)$$

and $\mathcal{F}_m$ is the linear space of dimension m spanned by the chosen basis functions. Under the assumption that $\log p_0$ has square-integrable derivatives up to order r, they showed that $KL(p_0 \| p_{\hat{f}_m}) = O_{p_0}(n^{-2r/(2r+1)})$ with $m = n^{1/(2r+1)}$ for each of the approximating families, where $KL(p\|q) = \int p(x) \log(p(x)/q(x))\, dx$ is the KL divergence between p and q. Similar work was carried out by Gu and Qiu (1993), who assumed that $\log p_0$ lies in an RKHS, and proposed an estimator based on penalized MLE, with consistency and rates established in Jensen-Shannon divergence. Though these results are theoretically interesting, these estimators are obtained via a procedure similar to that in Fukumizu (2009), and therefore suffers from the practical drawbacks discussed above.

The discussion so far shows that the MLE approach to learning $p_0 \in \mathcal{P}$ results in estimators that are of limited practical interest. To alleviate this, one can treat the problem of estimating $p_0 \in \mathcal{P}$ in a completely non-parametric fashion by using KDE, which is well-studied (Tsybakov, 2009, Chapter 1) and easy to implement. This approach ignores the structure of $\mathcal{P}$, however, and is known to perform poorly for moderate to large d (Wasserman, 2006, Section 6.5) (see also Section 6 of this paper).

1.1 Score Matching and Fisher Divergence

To counter the disadvantages of KDE and pseudo/penalized-MLE, in this paper, we propose to use the *score matching method* introduced by Hyvärinen (2005, 2007). While MLE is based on minimizing the KL divergence, the score matching method involves minimizing the *Fisher divergence* (also called the Fisher information distance; see Definition 1.13 in Johnson (2004)) between two continuously differentiable densities, p and q on an open set $\Omega \subseteq \mathbb{R}^d$, given as

$$J(p\|q) = \frac{1}{2} \int_\Omega p(x) \left\| \nabla \log p(x) - \nabla \log q(x) \right\|_2^2 \, dx, \tag{3}$$

where $\nabla \log p(x) = (\partial_1 \log p(x), \ldots, \partial_d \log p(x))$ with $\partial_i \log p(x) := \frac{\partial}{\partial x_i} \log p(x)$. Fisher divergence is closely related to the KL divergence through de Bruijn's identity (Johnson, 2004, Appendix C) and it can be shown that $KL(p\|q) = \int_0^\infty J(p_t\|q_t)\, dt$, where $p_t = p * N(0, tI_d)$, $q_t = q * N(0, tI_d)$, $*$ denotes the convolution, and $N(0, tI_d)$ denotes a normal distribution on $\mathbb{R}^d$ with mean zero and diagonal covariance with $t > 0$ (see Proposition B.1 for a precise statement; also see Theorem 1 in Lyu, 2009). Moreover, convergence in Fisher divergence is a stronger form of convergence than that in KL, total variation and Hellinger distances (see Lemmas E.2 & E.3 in Johnson, 2004 and Corollary 5.1 in Ley and Swan, 2013).

To understand the advantages associated with the score matching method, let us consider the problem of density estimation where the data generating distribution (say p_0) belongs to $\mathscr{P}_{\text{fin}}$ in (1). In other words, given random samples $(X_a)_{a=1}^n$ drawn i.i.d. from $p_0 := p_{\theta_0}$, the goal is to estimate θ_0 as $\hat{\theta}_n$, and use $p_{\hat{\theta}_n}$ as an estimator of p_0. While the

MLE approach is well-studied and enjoys nice statistical properties in asymptopia (i.e., asymptotically unbiased, efficient, and normally distributed), the computation of $\hat{\theta}_n$ can be intractable in many situations as discussed above. In particular, this is the case for $p_\theta(x) = \frac{r_\theta(x)}{A(\theta)}$ where $r_\theta \geq 0$ for all $\theta \in \Theta$, $A(\theta) = \int_\Omega r_\theta(x)\,dx$, and the functional form of r is known (as a function of θ and x); yet we do not know how to easily compute A, which is often analytically intractable. In this setting (which is exactly the setting of this paper), assuming p_θ to be differentiable (w.r.t. x), and $\int_\Omega p_0(x)\|\nabla \log p_\theta(x)\|_2^2\,dx < \infty, \forall\, \theta \in \Theta$, $J(p_0\|p_\theta) =: J(\theta)$ in (3) reduces to

$$J(\theta) = \sum_{i=1}^{d} \int_\Omega p_0(x) \left(\frac{1}{2} \left(\partial_i \log p_\theta(x)\right)^2 + \partial_i^2 \log p_\theta(x) \right) dx + \frac{1}{2} \int_\Omega p_0(x) \left\| \nabla \log p_0(x) \right\|_2^2 \, dx, \tag{4}$$

through integration by parts (see Hyvärinen, 2005, Theorem 1), under appropriate regularity conditions on p_0 and p_θ for all $\theta \in \Theta$. Here $\partial_i^2 \log p_\theta(x) := \frac{\partial^2}{\partial x_i^2} \log p_\theta(x)$. The main advantage of the objective in (3) (and also (4)) is that when it is applied to the situation discussed above where $p_\theta(x) = \frac{r_\theta(x)}{A(\theta)}$, $J(\theta)$ is independent of $A(\theta)$, and an estimate of θ_0 can be obtained by simply minimizing the empirical counterpart of $J(\theta)$, given by

$$J_n(\theta) := \frac{1}{n} \sum_{a=1}^{n} \sum_{i=1}^{d} \left(\frac{1}{2} \left(\partial_i \log p_\theta(X_a)\right)^2 + \partial_i^2 \log p_\theta(X_a) \right) + \frac{1}{2} \int_\Omega p_0(x) \left\| \nabla \log p_0(x) \right\|_2^2 \, dx.$$

Since $J_n(\theta)$ is also independent of $A(\theta)$, $\hat{\theta}_n = \arg\min_{\theta\in\Theta} J_n(\theta)$ may be easily computable, unlike the MLE. We would like to highlight that while the score matching approach may have computational advantages over MLE, it only estimates p_θ up to the scaling factor $A(\theta)$, and therefore requires the approximation or computation of $A(\theta)$ through numerical integration to estimate p_θ. Note that this issue (of computing $A(\theta)$ through numerical integration) exists even with MLE, but not with KDE. In score matching, however, numerical integration is needed only once, while MLE would typically require a functional form of the log-partition function which is approximated through numerical integration at every step of an iterative optimization algorithm (for example, see (2)), thus leading to major computational savings. An important application that does not require the computation of $A(\theta)$ is in finding modes of the distribution, which has recently become very popular in image processing (Comaniciu and Meer, 2002), and has already been investigated in the score matching framework (Sasaki et al., 2014). Similarly, in sampling methods such as sequential Monte Carlo (Doucet et al., 2001), it is often the case that the evaluation of unnormalized densities is sufficient to calculate required importance weights.

1.2 Contributions

(i) We present an estimate of $p_0 \in \mathcal{P}$ in the well-specified case through the minimization of Fisher divergence, in Section 4. First, we show that estimating $p_0 := p_{f_0}$ using the score matching method reduces to estimating f_0 by solving a simple finite-dimensional linear system (Theorems 4 and 5). Hyvärinen (2007) obtained a similar result for $\mathscr{P}_{\text{fin}}$ where the estimator is obtained by solving a linear system, which in the case of Gaussian family

matches the MLE (Hyvärinen, 2005). The estimator obtained in the infinite dimensional case is not a simple extension of its finite-dimensional counterpart, however, as the former requires an appropriate regularizer (we use $\|\cdot\|^2_{\mathcal{H}}$) to make the problem well-posed. We would like to highlight that to the best of our knowledge, the proposed estimator is the first practically computable estimator of p_0 with consistency guarantees (see below).

(ii) In contrast to Hyvärinen (2007) where no guarantees on consistency or convergence rates are provided for the density estimator in $\mathscr{P}_{\text{fin}}$, we establish in Theorem 6 the consistency and rates of convergence for the proposed estimator of f_0, and use these to prove consistency and rates of convergence for the corresponding plug-in estimator of p_0 (Theorems 7 and B.2), even when $\mathcal{H}$ is infinite dimensional. Furthermore, while the estimator of f_0 (and therefore p_0) is obtained by minimizing the Fisher divergence, the resultant density estimator is also shown to be consistent in KL divergence (and therefore in Hellinger and total-variation distances) and we provide convergence rates in all these distances.

Formally, we show that the proposed estimator $\hat{f}_n$ is converges as

$$\|f_0-\hat{f}_n\|_{\mathcal{H}}=O_{p_0}(n^{-\alpha}),\ KL(p_0\|p_{\hat{f}_n})=O_{p_0}(n^{-2\alpha}) \text{ and } J(p_0\|p_{\hat{f}_n})=O_{p_0}\left(n^{-\min\left\{\frac{2}{3},\frac{2\beta+1}{2\beta+2}\right\}}\right)$$

if $f_0\in\mathcal{R}(C^\beta)$ for some $\beta>0$, where $\mathcal{R}(A)$ denotes the range or image of an operator A, $\alpha=\min\{\frac{1}{4},\frac{\beta}{2\beta+2}\}$, and $C:=\sum_{i=1}^d\int_\Omega \partial_i k(x,\cdot)\otimes\partial_i k(x,\cdot)\,p_0(x)\,dx$ is a Hilbert-Schmidt operator on $\mathcal{H}$ (see Theorem 4) with k being the reproducing kernel and $\otimes$ denoting the tensor product. When $\mathcal{H}$ is a finite-dimensional RKHS, we show that the estimator enjoys parametric rates of convergence, i.e.,

$$\|f_0-\hat{f}_n\|_{\mathcal{H}}=O_{p_0}(n^{-1/2}),\ KL(p_0\|p_{\hat{f}_n})=O_{p_0}(n^{-1}) \text{ and } J(p_0\|p_{\hat{f}_n})=O_{p_0}(n^{-1}).$$

Note that the convergence rates are obtained under a non-classical smoothness assumption on f_0, namely that it lies in the image of certain fractional power of C, which reduces to a more classical assumption if we choose k to be a Matérn kernel (see Section 2 for its definition), as it induces a Sobolev space. In Section 4.2, we discuss in detail the smoothness assumption on f_0 for the Gaussian (Example 2) and Matérn (Example 3) kernels. Another interesting point to observe is that unlike in the classical function estimation methods (e.g., kernel density estimation and regression), the rates presented above for the proposed estimator tend to saturate for $\beta>1$ ($\beta>\frac{1}{2}$ w.r.t. J), with the best rate attained at $\beta=1$ ($\beta=\frac{1}{2}$ w.r.t. J), which means the smoothness of f_0 is not fully captured by the estimator. Such a saturation behavior is well-studied in the inverse problem literature (Engl et al., 1996) where it has been attributed to the choice of regularizer. In Section 4.3, we discuss alternative regularization strategies using ideas from Bauer et al. (2007), which covers non-parametric least squares regression: we show that for appropriately chosen regularizers, the above mentioned rates hold for any $\beta>0$, and do not saturate for the aforementioned ranges of β (see Theorem 9).

(iii) In Section 5, we study the problem of density estimation in the misspecified setting, i.e., $p_0\notin\mathcal{P}$, which is not addressed in Hyvärinen (2007) and Fukumizu (2009). Using a more sophisticated analysis than in the well-specified case, we show in Theorem 12 that $J(p_0\|p_{\hat{f}_n})\to\inf_{p\in\mathcal{P}}J(p_0\|p)$ as $n\to\infty$. Under an appropriate smoothness assumption

on $\log \frac{p_0}{q_0}$ (see the statement of Theorem 12 for details), we show that $J(p_0 \| p_{\hat{f}_n}) \to 0$ as $n \to \infty$ along with a rate for this convergence, even though $p_0 \notin \mathcal{P}$. However, unlike in the well-specified case, where the consistency is obtained not only in J but also in other distances, we obtain convergence only in J for the misspecified case. Note that while Barron and Sheu (1991) considered the estimation of p_0 in the misspecified setting, the results are restricted to the approximating families consisting of polynomials, splines, or trigonometric series. Our results are more general, as they hold for abstract RKHSs.

(iv) In Section 6, we present preliminary numerical results comparing the proposed estimator with KDE in estimating a Gaussian and mixture of Gaussians, with the goal of empirically evaluating performance as d gets large for a fixed sample size. In these two estimation problems, we show that the proposed estimator outperforms KDE, and the advantage grows as d increases. Inspired by this preliminary empirical investigation, our proposed estimator (or computationally efficient approximations) has been used by Strathmann et al. (2015) in a gradient-free adaptive MCMC sampler, and by Sun et al. (2015) for graphical model structure learning. These applications demonstrate the practicality and performance of the proposed estimator.

Finally, we would like to make clear that our principal goal is not to construct density estimators that improve uniformly upon KDE, but to provide a novel flexible modeling technique for approximating an unknown density by a rich parametric family of densities, with the parameter being infinite dimensional, in contrast to the classical approach of finite dimensional approximation.

Various notations and definitions that are used throughout the paper are collected in Section 2. The proofs of the results are provided in Section 8, along with some supplementary results in an appendix.

2. Definitions & Notation

We introduce the notation used throughout the paper. Define $[d] := \{1, \ldots, d\}$. For $a := (a_1, \ldots, a_d) \in \mathbb{R}^d$ and $b := (b_1, \ldots, b_d) \in \mathbb{R}^d$, $\|a\|_2 := \sqrt{\sum_{i=1}^d a_i^2}$ and $\langle a, b \rangle := \sum_{i=1}^d a_i b_i$. For $a, b > 0$, we write $a \lesssim b$ if $a \le \gamma b$ for some positive universal constant γ. For a topological space $\mathcal{X}$, $C(\mathcal{X})$ (*resp.* $C_b(\mathcal{X})$) denotes the space of all continuous (*resp.* bounded continuous) functions on $\mathcal{X}$. For a locally compact Hausdorff space $\mathcal{X}$, $f \in C(\mathcal{X})$ is said to *vanish at infinity* if for every $\epsilon > 0$ the set $\{x : |f(x)| \ge \epsilon\}$ is compact. The class of all continuous f on $\mathcal{X}$ which vanish at infinity is denoted as $C_0(\mathcal{X})$. For open $\mathcal{X} \subset \mathbb{R}^d$, $C^1(\mathcal{X})$ denotes the space of continuously differentiable functions on $\mathcal{X}$. For $f \in C_b(\mathcal{X})$, $\|f\|_\infty := \sup_{x \in \mathcal{X}} |f(x)|$ denotes the supremum norm of f. $M_b(\mathcal{X})$ denotes the set of all finite Borel measures on $\mathcal{X}$. For $\mu \in M_b(\mathcal{X})$, $L^r(\mathcal{X}, \mu)$ denotes the Banach space of r-power ($r \ge 1$) μ-integrable functions. For $\mathcal{X} \subset \mathbb{R}^d$, we will use $L^r(\mathcal{X})$ for $L^r(\mathcal{X}, \mu)$ if μ is a Lebesgue measure on $\mathcal{X}$. For $f \in L^p(\mathcal{X}, \mu)$, $\|f\|_{L^r(\mathcal{X},\mu)} := \left(\int_{\mathcal{X}} |f|^r \, d\mu\right)^{1/r}$ denotes the L^r-norm of f for $1 \le r < \infty$ and we denote it as $\| \cdot \|_{L^r(\mathcal{X})}$ if $\mathcal{X} \subset \mathbb{R}^d$ and μ is the Lebesgue measure. The convolution $f * g$ of two measurable functions f and g on $\mathbb{R}^d$ is defined as

$$(f * g)(x) := \int_{\mathbb{R}^d} f(y) g(x - y) \, dy,$$

provided the integral exists for all $x \in \mathbb{R}^d$. The Fourier transform of $f \in L^1(\mathbb{R}^d)$ is defined as

$$f^\wedge(y) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(x)\, e^{-i\langle y,x\rangle}\, dx$$

where i denotes the imaginary unit $\sqrt{-1}$.

In the following, for the sake of completeness and simplicity, we present definitions restricted to Hilbert spaces. Let H_1 and H_2 be abstract Hilbert spaces. A map $S : H_1 \to H_2$ is called a *linear operator* if it satisfies $S(\alpha x) = \alpha Sx$ and $S(x+x') = Sx + Sx'$ for all $\alpha \in \mathbb{R}$ and $x, x' \in H_1$, where $Sx := S(x)$. A linear operator S is said to be *bounded*, i.e., the image SB_{H_1} of B_{H_1} under S is bounded if and only if there exists a constant $c \in [0,\infty)$ such that for all $x \in H_1$ we have $\|Sx\|_{H_2} \le c\|x\|_{H_1}$, where $B_{H_1} := \{x \in H_1 : \|x\|_{H_1} \le 1\}$. In this case, the *operator norm* of S is defined as $\|S\| := \sup\{\|Sx\|_{H_2} : x \in B_{H_1}\}$. Define $\mathcal{L}(H_1, H_2)$ be the space of bounded linear operators from H_1 to H_2. $S \in \mathcal{L}(H_1, H_2)$ is said to be *compact* if $\overline{SB_{H_1}}$ is a compact subset in H_2. The *adjoint operator* $S^* : H_2 \to H_1$ of $S \in \mathcal{L}(H_1, H_2)$ is defined by $\langle x, S^* y\rangle_{H_1} = \langle Sx, y\rangle_{H_2}$, $x \in H_1$, $y \in H_2$. $S \in \mathcal{L}(H) := \mathcal{L}(H, H)$ is called *self-adjoint* if $S^* = S$ and is called *positive* if $\langle Sx, x\rangle_H \ge 0$ for all $x \in H$. $\alpha \in \mathbb{R}$ is called an *eigenvalue* of $S \in \mathcal{L}(H)$ if there exists an $x \neq 0$ such that $Sx = \alpha x$ and such an x is called the *eigenvector* of S and α. For compact, positive, self-adjoint $S \in \mathcal{L}(H)$, $S^r : H \to H$, $r \ge 0$ is called a *fractional power* of S and $S^{1/2}$ is the *square root* of S, which we write as $\sqrt{S} := S^{1/2}$. An operator $S \in \mathcal{L}(H_1, H_2)$ is *Hilbert-Schmidt* if $\|S\|_{HS} := (\sum_{j\in J} \|Se_j\|^2_{H_2})^{1/2} < \infty$ where $(e_j)_{j\in J}$ is an arbitrary orthonormal basis of separable Hilbert space H_1. $S \in \mathcal{L}(H_1, H_2)$ is said to be of *trace class* if $\sum_{j\in J}\langle (S^*S)^{1/2} e_j, e_j\rangle_{H_1} < \infty$. For $x \in H_1$ and $y \in H_2$, $x \otimes y$ is an element of the tensor product space $H_1 \otimes H_2$ which can also be seen as an operator from H_2 to H_1 as $(x \otimes y)z = x\langle y, z\rangle_{H_2}$ for any $z \in H_2$. $\mathcal{R}(S)$ denotes the *range space (or image)* of S.

A real-valued symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a positive definite (pd) kernel if, for all $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and all $x_1, \ldots, x_n \in \mathcal{X}$, we have $\sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) \ge 0$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $(x, y) \mapsto k(x, y)$ is a *reproducing kernel* of the Hilbert space $(\mathcal{H}_k, \langle\cdot,\cdot\rangle_{\mathcal{H}_k})$ of functions if and only if (i) $\forall\, y \in \mathcal{X}$, $k(y,\cdot) \in \mathcal{H}_k$ and (ii) $\forall\, y \in \mathcal{X}$, $\forall\, f \in \mathcal{H}_k$, $\langle f, k(y,\cdot)\rangle_{\mathcal{H}_k} = f(y)$ hold. If such a k exists, then $\mathcal{H}_k$ is called a *reproducing kernel Hilbert space*. Since $\langle k(x,\cdot), k(y,\cdot)\rangle_{\mathcal{H}_k} = k(x, y)$, $\forall\, x, y \in \mathcal{X}$, it is easy to show that every reproducing kernel (r.k.) k is symmetric and positive definite. Some examples of kernels that appear throughout the paper are: *Gaussian kernel*, $k(x, y) = \exp(-\sigma\|x-y\|_2^2)$, $x, y \in \mathbb{R}^d$, $\sigma > 0$ that induces the following *Gaussian RKHS*,

$$\mathcal{H}_k = \mathcal{H}_\sigma := \Big\{ f \in L^2(\mathbb{R}^d) \cap C(\mathbb{R}^d) \;:\; \int |f^\wedge(\omega)|^2 e^{\|\omega\|_2^2/4\sigma}\, d\omega < \infty \Big\},$$

the *inverse multiquadric kernel*, $k(x, y) = (1 + \|\frac{x-y}{c}\|_2^2)^{-\beta}$, $x, y \in \mathbb{R}^d$, $\beta > 0$, $c \in (0,\infty)$ and the *Matérn kernel*, $k(x, y) = \frac{2^{1-\beta}}{\Gamma(\beta)}\|x-y\|_2^{\beta-d/2}\mathfrak{K}_{d/2-\beta}(\|x-y\|_2)$, $x, y \in \mathbb{R}^d$, $\beta > d/2$ that induces the Sobolev space, H_2^β,

$$\mathcal{H}_k = H_2^\beta := \Big\{ f \in L^2(\mathbb{R}^d) \cap C(\mathbb{R}^d) \;:\; \int (1 + \|\omega\|_2^2)^\beta |f^\wedge(\omega)|^2\, d\omega < \infty \Big\},$$

where Γ is the Gamma function, and $\mathfrak{K}_v$ is the modified Bessel function of the third kind of order v (v controls the smoothness of k).

For any real-valued function f defined on open $\mathcal{X} \subset \mathbb{R}^d$, f is said to be m-times continuously differentiable if for $\alpha \in \mathbb{N}_0^d$ with $|\alpha| := \sum_{i=1}^d \alpha_i \le m$, $\partial^\alpha f(x) = \partial_1^{\alpha_1} \dots \partial_d^{\alpha_d} f(x) = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} f(x)$ exists. A kernel k is said to be m-times continuously differentiable if $\partial^{\alpha,\alpha} k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ exists and is continuous for all $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \le m$ where $\partial^{\alpha,\alpha} := \partial_1^{\alpha_1} \dots \partial_d^{\alpha_d} \partial_{1+d}^{\alpha_1} \dots \partial_{2d}^{\alpha_d}$. Corollary 4.36 in Steinwart and Christmann (2008) and Theorem 1 in Zhou (2008) state that if $\partial^{\alpha,\alpha} k$ exists and is continuous, then $\partial^\alpha k(x,\cdot) = \partial_1^{\alpha_1} \dots \partial_d^{\alpha_d} k(x,\cdot) = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} k((x_1,\dots,x_d),\cdot) \in \mathcal{H}_k$ with $x = (x_1,\dots,x_d)$ and for every $f \in \mathcal{H}_k$, we have $\partial^\alpha f(x) = \langle \partial^\alpha k(x,\cdot), f \rangle_{\mathcal{H}_k}$ and $\partial^{\alpha,\alpha} k(x,x') = \langle \partial^\alpha k(x,\cdot), \partial^\alpha k(x',\cdot) \rangle_{\mathcal{H}_k}$.

Given two probability densities, p and q on $\Omega \subset \mathbb{R}^d$, the Kullback-Leibler divergence (KL) and Hellinger distance (h) are defined as $KL(p\|q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx$ and $h(p,q) = \|\sqrt{p} - \sqrt{q}\|_{L^2(\Omega)}$ respectively. We refer to $\|p-q\|_{L^1(\Omega)}$ as the total variation (TV) distance between p and q.

3. Approximation of Densities by $\mathcal{P}$

In this section, we first show that every finite dimensional exponential family, $\mathscr{P}_{\text{fin}}$ is generated by the family $\mathcal{P}$ induced by a finite dimensional RKHS, which naturally leads to the infinite dimensional generalization of $\mathscr{P}_{\text{fin}}$ when $\mathcal{H}$ is an infinite dimensional RKHS. Next, we investigate the approximation properties of $\mathcal{P}$ in Proposition 1 and Corollary 2 when $\mathcal{H}$ is an infinite dimensional RKHS.

Let us consider a r-parameter exponential family, $\mathscr{P}_{\text{fin}}$ with sufficient statistic $T(x) := (T_1(x),\dots,T_r(x))$ and construct a Hilbert space, $\mathcal{H} = \text{span}\{T_1(x),\dots,T_r(x)\}$. It is easy to verify that $\mathcal{P}$ induced by $\mathcal{H}$ is exactly the same as $\mathscr{P}_{\text{fin}}$ since any $f \in \mathcal{H}$ can be written as $f(x) = \sum_{i=1}^r \theta_i T_i(x)$ for some $(\theta_i)_{i=1}^r \subset \mathbb{R}$. In fact, by defining the inner product between $f = \sum_{i=1}^r \theta_i T_i$ and $g = \sum_{i=1}^r \gamma_i T_i$ as $\langle f,g \rangle_{\mathcal{H}} := \sum_{i=1}^r \theta_i \gamma_i$, it follows that $\mathcal{H}$ is an RKHS with the r.k. $k(x,y) = \langle T(x), T(y) \rangle_{\mathbb{R}^r}$ since $\langle f, k(x,\cdot) \rangle_{\mathcal{H}} = \sum_{i=1}^r \theta_i T_i(x) = f(x)$. Based on this equivalence between $\mathscr{P}_{\text{fin}}$ and $\mathcal{P}$ induced by a finite dimensional RKHS, it is therefore clear that $\mathcal{P}$ induced by a infinite dimensional RKHS is a strict generalization to $\mathscr{P}_{\text{fin}}$ with $k(\cdot,x)$ playing the role of a sufficient statistic.

Example 1 *The following are some popular examples of probability distributions that belong to $\mathscr{P}_{\text{fin}}$. Here we show the corresponding RKHSs $(\mathcal{H}, k)$ that generate these distributions. In some of these examples, we choose $q_0(x) = 1$ and ignore the fact that q_0 is a probability distribution as assumed in the definition of $\mathcal{P}$.*

Exponential: $\Omega = \mathbb{R}_{++} := \mathbb{R}_+ \backslash \{0\}$, $k(x,y) = xy$.

Normal: $\Omega = \mathbb{R}$, $k(x,y) = xy + x^2y^2$.

Beta: $\Omega = (0,1)$, $k(x,y) = \log x \log y + \log(1-x)\log(1-y)$.

Gamma: $\Omega = \mathbb{R}_{++}$, $k(x,y) = \log x \log y + xy$.

Inverse Gaussian: $\Omega = \mathbb{R}_{++}$, $k(x,y) = xy + \frac{1}{xy}$.

Poisson: $\Omega = \mathbb{N} \cup \{0\}$, $k(x,y) = xy$, $q_0(x) = (x!\, e)^{-1}$.

Binomial: $\Omega = \{0, \ldots, m\}$, $k(x, y) = xy$, $q_0(x) = 2^{-m}\binom{m}{c}$.

While Example 1 shows that all popular probability distributions are contained in $\mathcal{P}$ for an appropriate choice of finite-dimensional $\mathcal{H}$, it is of interest to understand the richness of $\mathcal{P}$ (i.e., what class of distributions can be approximated arbitrarily well by $\mathcal{P}$?) when $\mathcal{H}$ is an infinite dimensional RKHS. This is addressed by the following result, which is proved in Section 8.1.

Proposition 1 *Define*

$$\mathcal{P}_0 := \left\{\pi_f(x) = e^{f(x)-A(f)}q_0(x),\, x \in \Omega : f \in C_0(\Omega)\right\}$$

where $\Omega \subseteq \mathbb{R}^d$ is locally compact Hausdorff. Suppose $k(x, \cdot) \in C_0(\Omega)$, $\forall\, x \in \Omega$ and

$$\int\int k(x, y)\, d\mu(x)\, d\mu(y) > 0,\ \forall\, \mu \in M_b(\Omega)\backslash\{0\}. \tag{5}$$

Then $\mathcal{P}$ is dense in $\mathcal{P}_0$ w.r.t. Kullback-Leibler divergence, total variation (L^1 norm) and Hellinger distances. In addition, if $q_0 \in L^1(\Omega) \cap L^r(\Omega)$ for some $1 < r \leq \infty$, then $\mathcal{P}$ is also dense in $\mathcal{P}_0$ w.r.t. L^r norm.

A sufficient condition for $\Omega \subseteq \mathbb{R}^d$ to be locally compact Hausdorff is that it is either open or closed. Condition (5) is equivalent to k being c_0-universal (Sriperumbudur et al., 2011, p. 2396). If $k(x, y) = \psi(x - y)$, $x, y \in \Omega = \mathbb{R}^d$ where $\psi \in C_b(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$, then (5) can be shown to be equivalent to $\text{supp}(\psi^\wedge) = \mathbb{R}^d$ (Sriperumbudur et al., 2011, Proposition 5). Examples of kernels that satisfy the conditions in Proposition 1 include the Gaussian, Matérn and inverse multiquadrics. In fact, any compactly supported non-zero $\psi \in C_b(\mathbb{R}^d)$ satisfies the assumptions in Proposition 1 as $\text{supp}(\psi^\wedge) = \mathbb{R}^d$ (Sriperumbudur et al., 2010, Corollary 10). Though $\mathcal{P}_0$ is still a parametric family of densities indexed by a Banach space (here $C_0(\Omega)$), the following corollary (proved in Section 8.2) to Proposition 1 shows that a broad class of continuous densities are contained in $\mathcal{P}_0$ and therefore can be approximated arbitrarily well in L^r norm ($1 \leq r \leq \infty$), Hellinger distance, and KL divergence by $\mathcal{P}$.

Corollary 2 *Let $q_0 \in C(\Omega)$ be a probability density such that $q_0(x) > 0$ for all $x \in \Omega$, where $\Omega \subseteq \mathbb{R}^d$ is locally compact Hausdorff. Suppose there exists a constant ℓ such that for any $\epsilon > 0$, $\exists\, R > 0$ that satisfies $|\frac{p(x)}{q_0(x)} - \ell| \leq \epsilon$ for any x with $\|x\|_2 > R$. Define*

$$\mathcal{P}_c := \left\{p \in C(\Omega) : \int_\Omega p(x)\, dx = 1, p(x) \geq 0, \forall\, x \in \Omega \text{ and } \frac{p}{q_0} - \ell \in C_0(\Omega)\right\}.$$

Suppose $k(x, \cdot) \in C_0(\Omega)$, $\forall\, x \in \Omega$ and (5) holds. Then $\mathcal{P}$ is dense in $\mathcal{P}_c$ w.r.t. KL divergence, TV and Hellinger distances. Moreover, if $q_0 \in L^1(\Omega) \cap L^r(\Omega)$ for some $1 < r \leq \infty$, then $\mathcal{P}$ is also dense in $\mathcal{P}_c$ w.r.t. L^r norm.

By choosing Ω to be compact and q_0 to be a uniform distribution on Ω, Corollary 2 reduces to an easily interpretable result that any continuous density p_0 on Ω can be approximated arbitrarily well by densities in $\mathcal{P}$ in KL, Hellinger and L^r ($1 \leq r \leq \infty$) distances.

Similar to the results so far, an approximation result for $\mathcal{P}$ can also be obtained w.r.t. Fisher divergence (see Proposition 13). Since this result is heavily based on the notions and results developed in Section 5, we defer its presentation until that section. Briefly, this result states that if $\mathcal{H}$ is sufficiently rich (i.e., dense in an appropriate class of functions), then any $p \in C^1(\Omega)$ with $J(p\|q_0) < \infty$ can be approximated arbitrarily well by elements in $\mathcal{P}$ w.r.t. Fisher divergence, where $q_0 \in C^1(\Omega)$.

4. Density Estimation in $\mathcal{P}$: Well-specified Case

In this section, we present our score matching estimator for an unknown density $p_0 := p_{f_0} \in \mathcal{P}$ (well-specified case) from i.i.d. random samples $(X_a)_{a=1}^n$ drawn from it. This involves choosing the minimizer of the (empirical) Fisher divergence between p_0 and $p_f \in \mathcal{P}$ as the estimator, $\hat{f}$ which we show in Theorem 5 to be obtained by solving a simple finite-dimensional linear system. In contrast, we would like to remind the reader that the MLE is infeasible in practice due to the difficulty in handling $A(f)$. The consistency and convergence rates of $\hat{f} \in \mathcal{F}$ and the plug-in estimator $p_{\hat{f}}$ are provided in Section 4.1 (see Theorems 6 and 7). Before we proceed, we list the assumptions on p_0, q_0 and $\mathcal{H}$ that we need in our analysis.

(A) Ω is a non-empty open subset of $\mathbb{R}^d$ with a piecewise smooth boundary $\partial\Omega := \overline{\Omega}\backslash\Omega$, where $\overline{\Omega}$ denotes the closure of Ω.

(B) p_0 is continuously extendible to $\overline{\Omega}$. k is twice continuously differentiable on $\Omega \times \Omega$ with continuous extension of $\partial^{\alpha,\alpha}k$ to $\overline{\Omega} \times \overline{\Omega}$ for $|\alpha| \leq 2$.

(C) $\partial_i\partial_{i+d}k(x,x)p_0(x) = 0$ for $x \in \partial\Omega$ and $\sqrt{\partial_i\partial_{i+d}k(x,x)}p_0(x) = o(\|x\|_2^{1-d})$ as $x \in \Omega$, $\|x\|_2 \to \infty$ for all $i \in [d]$.

(D) *(ε-Integrability)* For some $\varepsilon \geq 1$ and $\forall\, i \in [d]$, $\partial_i\partial_{i+d}k(x,x)$, $\sqrt{\partial_i^2\partial_{i+d}^2k(x,x)}$ and $\sqrt{\partial_i\partial_{i+d}k(x,x)}\partial_i \log q_0(x) \in L^\varepsilon(\Omega, p_0)$, where $q_0 \in C^1(\Omega)$.

Remark 3 (i) *Ω being a subset of $\mathbb{R}^d$ along with k being continuous ensures that $\mathcal{H}$ is separable (Steinwart and Christmann, 2008, Lemma 4.33). The twice differentiability of k ensures that every $f \in \mathcal{H}$ is twice continuously differentiable (Steinwart and Christmann, 2008, Corollary 4.36).* **(C)** *ensures that J in (3) is equivalent to the one in (4) through integration by parts on Ω (see Corollary 7.6.2 in Duistermaat and Kolk, 2004 for integration by parts on bounded subsets of $\mathbb{R}^d$ which can be extended to unbounded Ω through a truncation and limiting argument) for densities in $\mathcal{P}$. In particular,* **(C)** *ensures that $\int_\Omega \partial_i f(x)\partial_i p_0(x)\,dx = -\int_\Omega \partial_i^2 f(x)p_0(x)\,dx$ for all $f \in \mathcal{H}$ and $i \in [d]$, which will be critical to prove the representation in Theorem 4(ii), upon which rest of the results depend. The decay condition in* **(C)** *can be weakened to $\sqrt{\partial_i\partial_{i+d}k(x,x)}p_0(x) = o(\|x\|_2^{1-\overline{d}})$ as $x \in \Omega$, $\|x\|_2 \to \infty$ for all $i \in [d]$ if Ω is a (possibly unbounded) box where $\overline{d} = \#\{i \in [d] | (a_i, b_i)$ is unbounded$\}$.*

(ii) *When $\varepsilon = 1$, the first condition in* **(D)** *ensures that $J(p_0\|p_f) < \infty$ for any $p_f \in \mathcal{P}$. The other two conditions ensure the validity of the alternate representation for $J(p_0\|p_f)$ in (4) which will be useful in constructing estimators of p_0 (see Theorem 4). Examples of* *kernels that satisfy* **(D)** *are the Gaussian, Matérn (with* $\beta > \max\{2, d/2\}$*), and inverse multiquadric kernels, for which it is easy to show that there exists* q_0 *that satisfies* **(D)**.

(iii) (Identifiability) *The above list of assumptions do not include the identifiability condition that ensures* $p_{f_1} = p_{f_2}$ *if and only if* $f_1 = f_2$*. It is clear that if constant functions are included in* $\mathcal{H}$*, i.e.,* $1 \in \mathcal{H}$*, then* $p_f = p_{f+c}$ *for any* $c \in \mathbb{R}$*. On the other hand, it can be shown that if* $1 \notin \mathcal{H}$ *and* $\text{supp}(q_0) = \Omega$*, then* $p_{f_1} = p_{f_2} \Leftrightarrow f_1 = f_2$*. A sufficient condition for* $1 \notin \mathcal{H}$ *is* $k \in C_0(\Omega \times \Omega)$*. We do not explicitly impose the identifiability condition as a part of our blanket assumptions because the assumptions under which consistency and rates are obtained in Theorem 7 automatically ensure identifiability.*

Under these assumptions, the following result—proved in Section 8.3—shows that the problem of estimating p_0 through the minimization of Fisher divergence reduces to the problem of estimating f_0 through a weighted least squares minimization in $\mathcal{H}$ (see parts (i) and (ii)). This motivates the minimization of the regularized empirical weighted least squares (see part (iv)) to obtain an estimator $f_{\lambda,n}$ of f_0, which is then used to construct the plug-in estimate $p_{f_{\lambda,n}}$ of p_0.

Theorem 4 *Suppose* **(A)**–**(D)** *hold with* $\varepsilon = 1$*. Then* $J(p_0\|p_f) < \infty$ *for all* $f \in \mathcal{F}$*. In addition, the following hold.*

(i) For all $f \in \mathcal{F}$*,*

$$J(f) := J(p_0\|p_f) = \frac{1}{2}\left\langle f - f_0, C(f - f_0)\right\rangle_{\mathcal{H}}, \tag{6}$$

where $C : \mathcal{H} \to \mathcal{H}$*,* $C := \int_\Omega p_0(x) \sum_{i=1}^d \partial_i k(x,\cdot) \otimes \partial_i k(x,\cdot)\, dx$ *is a trace-class positive operator with*

$$Cf = \int_\Omega p_0(x) \sum_{i=1}^d \partial_i k(x,\cdot) \partial_i f(x)\, dx.$$

(ii) Alternatively,

$$J(f) = \frac{1}{2}\langle f, Cf\rangle_{\mathcal{H}} + \langle f, \xi\rangle_{\mathcal{H}} + J(p_0\|q_0)$$

where

$$\xi := \int_\Omega p_0(x) \sum_{i=1}^d \left(\partial_i k(x,\cdot)\partial_i \log q_0(x) + \partial_i^2 k(x,\cdot)\right)\, dx \in \mathcal{H}$$

and f_0 *satisfies* $Cf_0 = -\xi$*.*

(iii) For any $\lambda > 0$*, a unique minimizer* f_λ *of* $J_\lambda(f) := J(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$ *over* $\mathcal{H}$ *exists and is given by*

$$f_\lambda = -(C + \lambda I)^{-1}\xi = (C + \lambda I)^{-1} C f_0.$$

(iv) ***(Estimator of*** f_0***)*** *Given samples* $(X_a)_{a=1}^n$ *drawn i.i.d. from* p_0*, for any* $\lambda > 0$*, the unique minimizer* $f_{\lambda,n}$ *of* $\hat{J}_\lambda(f) := \hat{J}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$ *over* $\mathcal{H}$ *exists and is given by*

$$f_{\lambda,n} = -(\hat{C} + \lambda I)^{-1}\hat{\xi},$$

where $\hat{J}(f) := \frac{1}{2}\langle f, \hat{C}f\rangle_{\mathcal{H}} + \langle f, \hat{\xi}\rangle_{\mathcal{H}} + J(p_0\|q_0)$, $\hat{C} := \frac{1}{n}\sum_{a=1}^{n}\sum_{i=1}^{d} \partial_i k(X_a,\cdot)\otimes\partial_i k(X_a,\cdot)$ *and*

$$\hat{\xi} := \frac{1}{n}\sum_{a=1}^{n}\sum_{i=1}^{d}\left(\partial_i k(X_a,\cdot)\partial_i \log q_0(X_a) + \partial_i^2 k(X_a,\cdot)\right).$$

An advantage of the alternate formulation of $J(f)$ in Theorem 4(ii) over (6) is that it provides a simple way to obtain an empirical estimate of $J(f)$—by replacing C and ξ by their empirical estimators, $\hat{C}$ and $\hat{\xi}$ respectively—from finite samples drawn i.i.d. from p_0, which is then used to obtain an estimator of f_0. Note that the empirical estimate of $J(f)$, i.e., $\hat{J}(f)$ depends only on $\hat{C}$ and $\hat{\xi}$ which in turn depend on the known quantities, k and q_0, and therefore $f_{\lambda,n}$ in Theorem 4(iv) should in principle be computable. In practice, however, it is not easy to compute the expression for $f_{\lambda,n} = -(\hat{C}+\lambda I)^{-1}\hat{\xi}$ as it involves solving an infinite dimensional linear system. In Theorem 5 (proved in Section 8.4), we provide an alternative expression for $f_{\lambda,n}$ as a solution of a simple finite-dimensional linear system (see (7) and (8)), using the general representer theorem (see Theorem A.2). It is interesting to note that while the solution to $J(f)$ in Theorem 4(ii) is obtained by solving a non-linear system, $Cf_0 = -\xi$ (the system is non-linear as C depends on p_0 which in turn depends on f_0), its estimator $f_{\lambda,n}$ proposed in Theorem 4, is obtained by solving a simple linear system. In addition, we would like to highlight the fact that the proposed estimator, $f_{\lambda,n}$ is precisely the Tikhonov regularized solution (which is well-studied in the theory of linear inverse problems) to the ill-posed linear system $\hat{C}f = -\hat{\xi}$. We further discuss the choice of regularizer in Section 4.3 using ideas from the inverse problem literature.

An important remark we would like to make about Theorem 4 is that though $J(f)$ in (6) is valid only for $f\in\mathcal{F}$, as it is obtained from $J(p_0\|p_f)$ where $p_0, p_f\in\mathcal{P}$, the expression $\langle f-f_0, C(f-f_0)\rangle_{\mathcal{H}}$ is valid for any $f\in\mathcal{H}$, as it is finite under the assumption that **(D)** holds with $\varepsilon = 1$. Therefore, in Theorem 4(iii, iv), f_λ and $f_{\lambda,n}$ are obtained by minimizing J_λ and $\hat{J}_\lambda$ over $\mathcal{H}$ instead of over $\mathcal{F}$, as the latter does not yield a nice expression (unlike f_λ and $f_{\lambda,n}$, respectively). However, there is no guarantee that $f_{\lambda,n}\in\mathcal{F}$, and so the density estimator $p_{f_{\lambda,n}}$ may not be valid. While this is not an issue when studying the convergence of $\|f_{\lambda,n}-f_0\|_{\mathcal{H}}$ (see Theorem 6), the convergence of $p_{f_{\lambda,n}}$ to p_0 (in various distances) needs to be handled slightly differently depending on whether the kernel is bounded or not (see Theorems 7 and B.2). Note that when the kernel is bounded, we obtain $\mathcal{F} = \mathcal{H}$, which implies $p_{f_{\lambda,n}}$ is valid.

Theorem 5 (Computation of $f_{\lambda,n}$) *Let* $f_{\lambda,n} = \arg\inf_{f\in\mathcal{H}}\hat{J}_\lambda(f)$, *where* $\hat{J}_\lambda(f)$ *is defined in Theorem 4(iv) and* $\lambda > 0$. *Then*

$$f_{\lambda,n} = -\frac{\hat{\xi}}{\lambda} + \sum_{a=1}^{n}\sum_{i=1}^{d}\beta_{(a-1)d+i}\partial_i k(X_a,\cdot), \tag{7}$$

where $\hat{\xi}$ *is defined in Theorem 4(iv) and* $\boldsymbol{\beta} = (\beta_{(a-1)d+i})_{a,i}$ *is obtained by solving*

$$(\boldsymbol{G}+n\lambda I)\,\boldsymbol{\beta} = \frac{1}{\lambda}\boldsymbol{h} \tag{8}$$

with $(\boldsymbol{G})_{(a-1)d+i,(b-1)d+j} = \partial_i \partial_{j+d} k(X_a, X_b)$ *and*

$$(\boldsymbol{h})_{(a-1)d+i} = \langle \hat{\xi}, \partial_i k(X_a, \cdot)\rangle_{\mathcal{H}} = \frac{1}{n}\sum_{b=1}^{n}\sum_{j=1}^{d} \partial_i \partial^2_{j+d} k(X_a, X_b) + \partial_i \partial_{j+d} k(X_a, X_b) \partial_j \log q_0(X_b).$$

We would like to highlight that though $f_{\lambda,n}$ requires solving a simple linear system in (8), it can still be computationally intensive when d and n are large as $\boldsymbol{G}$ is a $nd \times nd$ matrix. This is still a better scenario than that of MLE, however, since computationally efficient methods exist to solve large linear systems such as (8), whereas MLE can be intractable due to the difficulty in handling the log-partition function (though it can be approximated). On the other hand, MLE is statistically well-understood, with consistency and convergence rates established in general for the problem of density estimation (van de Geer, 2000) and in particular for the problem at hand (Fukumizu, 2009). In order to ensure that $f_{\lambda,n}$ and $p_{f_{\lambda,n}}$ are statistically useful, in the following section, we investigate their consistency and convergence rates under some smoothness conditions on f_0.

4.1 Consistency and Rate of Convergence

In this section, we prove the consistency of $f_{\lambda,n}$ (see Theorem 6(i)) and $p_{f_{\lambda,n}}$ (see Theorems 7 and B.2). Under the smoothness assumption that $f_0 \in \mathcal{R}(C^\beta)$ for some $\beta > 0$, we present convergence rates for $f_{\lambda,n}$ and $p_{f_{\lambda,n}}$ in Theorems 6(ii), 7 and B.2. In reference to the following results, for simplicity we suppress the dependence of λ on n by defining $\lambda := \lambda_n$ where $(\lambda_n)_{n\in\mathbb{N}} \subset (0,\infty)$.

Theorem 6 (Consistency and convergence rates for $f_{\lambda,n}$) *Suppose* **(A)**–**(D)** *with* $\varepsilon = 2$ *hold.*

(i) If $f_0 \in \overline{\mathcal{R}(C)}$, *then* $\|f_{\lambda,n} - f_0\|_{\mathcal{H}} \stackrel{p_0}{\to} 0$ *as* $\lambda \to 0$, $\lambda\sqrt{n} \to \infty$ *and* $n \to \infty$.

(ii) If $f_0 \in \mathcal{R}(C^\beta)$ *for some* $\beta > 0$, *then for* $\lambda = n^{-\max\left\{\frac{1}{4}, \frac{1}{2(\beta+1)}\right\}}$,

$$\|f_{\lambda,n} - f_0\|_{\mathcal{H}} = O_{p_0}\left(n^{-\min\left\{\frac{1}{4}, \frac{\beta}{2(\beta+1)}\right\}}\right) \ \text{as} \ n \to \infty.$$

(iii) If $\|C^{-1}\| < \infty$, *then for* $\lambda = n^{-\frac{1}{2}}$, $\|f_{\lambda,n} - f_0\|_{\mathcal{H}} = O_{p_0}(n^{-1/2})$ *as* $n \to \infty$.

Remark *(i)* While Theorem 6 (proved in Section 8.5) provides an asymptotic behavior for $\|f_{\lambda,n} - f_0\|_{\mathcal{H}}$ under conditions that depend on p_0 (and are therefore not easy to check in practice), a non-asymptotic bound on $\|f_{\lambda,n} - f_0\|_{\mathcal{H}}$ that holds for all $n \geq 1$ can be obtained under stronger assumptions through an application of Bernstein's inequality in separable Hilbert spaces. For the sake of simplicity, we provided asymptotic results which are obtained through an application of Chebyshev's inequality.

(ii) The proof of Theorem 6(i) involves decomposing $\|f_{\lambda,n} - f_0\|_{\mathcal{H}}$ into an estimation error part, $\mathcal{E}(\lambda, n) := \|f_{\lambda,n} - f_\lambda\|_{\mathcal{H}}$, and an approximation error part, $\mathcal{A}_0(\lambda) := \|f_\lambda - f_0\|_{\mathcal{H}}$, where $f_\lambda = (C + \lambda I)^{-1} C f_0$. While $\mathcal{E}(\lambda, n) \to 0$ as $\lambda \to 0$, $\lambda\sqrt{n} \to \infty$ and $n \to \infty$ without any assumptions on f_0 (see the proof in Section 8.5 for details), it is not reasonable to expect

$\mathcal{A}_0(\lambda) \to 0$ as $\lambda \to 0$ without assuming $f_0 \in \overline{\mathcal{R}(C)}$. This is because, if f_0 lies in the null space of C, then f_λ is zero irrespective of λ and therefore cannot approximate f_0.

(iii) The condition $f_0 \in \overline{\mathcal{R}(C)}$ is difficult to check in practice as it depends on p_0 (which in turn depends on f_0). However, since the null space of C is just constant functions if the kernel is bounded and $supp(q_0) = \Omega$ (see Lemma 14 in Section 8.6 for details), assuming $1 \notin \mathcal{H}$ yields that $\overline{\mathcal{R}(C)} = \mathcal{H}$ and therefore consistency can be attained under conditions that are easy to impose in practice. As mentioned in Remark 3(iii), the condition $1 \notin \mathcal{H}$ ensures identifiability and a sufficient condition for it to hold is $k \in C_0(\Omega \times \Omega)$, which is satisfied by Gaussian, Matérn and inverse multiquadric kernels.

(iv) It is well known that convergence rates are possible only if the quantity of interest (here f_0) satisfies some additional conditions. In function estimation, this additional condition is classically imposed by assuming f_0 to be sufficiently smooth, e.g., f_0 lies in a Sobolev space of certain smoothness. By contrast, the smoothness condition in Theorem 6(ii) is imposed in an indirect manner by assuming $f_0 \in \mathcal{R}(C^\beta)$ for some $\beta > 0$—so that the results hold for abstract RKHSs and not just Sobolev spaces—which then provides a rate, with the best rate being $n^{-1/4}$ that is attained when $\beta \geq 1$. While such a condition has already been used in various works (Caponnetto and Vito, 2007; Smale and Zhou, 2007; Fukumizu et al., 2013) in the context of non-parametric least squares regression, we explore it in more detail in Proposition 8, and Examples 2 and 3. Note that this condition is common in the inverse problem theory (see Engl, Hanke, and Neubauer, 1996), and it naturally arises here through the connection of $f_{\lambda,n}$ being a Tikhonov regularized solution to the ill-posed linear system $\hat{C}f = -\hat{\xi}$. An interesting observation about the rate is that it does not improve with increasing β (for $\beta > 1$), in contrast to the classical results in function estimation (e.g., kernel density estimation and kernel regression) where the rate improves with increasing smoothness. This issue is discussed in detail in Section 4.3.

(v) Since $\|C^{-1}\| < \infty$ only if $\mathcal{H}$ is finite-dimensional, we recover the parametric rate of $n^{-1/2}$ in a finite-dimensional situation with an automatic choice for λ as $n^{-1/2}$.While Theorem 6 provides statistical guarantees for parameter convergence, the question of primary interest is the convergence of $p_{f_{\lambda,n}}$ to p_0. This is guaranteed by the following result, which is proved in Section 8.6.

Theorem 7 (Consistency and rates for $p_{f_{\lambda,n}}$) *Suppose* **(A)**–**(D)** *with $\varepsilon = 2$ hold and $\|k\|_\infty := \sup_{x\in\Omega} k(x,x) < \infty$. Assume* $\mathrm{supp}(q_0) = \Omega$. *Then the following hold:*

(i) For any $1 < r \leq \infty$ with $q_0 \in L^1(\Omega) \cap L^r(\Omega)$,

$$\|p_{f_{\lambda,n}} - p_0\|_{L^r(\Omega)} \to 0,\ h(p_{f_{\lambda,n}}, p_0) \to 0,\ KL(p_0\|p_{f_{\lambda,n}}) \to 0 \ \text{ as } \ \lambda\sqrt{n} \to \infty,\ \lambda \to 0 \text{ and } n \to \infty.$$

In addition, if $f_0 \in \mathcal{R}(C^\beta)$ for some $\beta > 0$, then for $\lambda = n^{-\max\left\{\frac{1}{4}, \frac{1}{2(\beta+1)}\right\}}$,

$$\|p_{f_{\lambda,n}} - p_0\|_{L^r(\Omega)} = O_{p_0}(\theta_n),\ h(p_0, p_{f_{\lambda,n}}) = O_{p_0}(\theta_n),\ KL(p_0\|p_{f_{\lambda,n}}) = O_{p_0}(\theta_n^2)$$

as $n \to \infty$ where $\theta_n := n^{-\min\left\{\frac{1}{4}, \frac{\beta}{2(\beta+1)}\right\}}$.

(ii) $J(p_0\|p_{f_{\lambda,n}}) \to 0$ as $\lambda n \to \infty$, $\lambda \to 0$ and $n \to \infty$. In addition, if $f_0 \in \mathcal{R}(C^\beta)$ for some

$\beta \geq 0$, then for $\lambda = n^{-\max\left\{\frac{1}{3}, \frac{1}{2(\beta+1)}\right\}}$,

$$J(p_0 \| p_{f_{\lambda,n}}) = O_{p_0}\left(n^{-\min\left\{\frac{2}{3}, \frac{2\beta+1}{2(\beta+1)}\right\}}\right) \quad \text{as } n \to \infty.$$

(iii) If $\|C^{-1}\| < \infty$, then $\theta_n = n^{-\frac{1}{2}}$ and $J(p_0\|p_{f_{\lambda,n}}) = O_{p_0}(n^{-1})$ with $\lambda = n^{-\frac{1}{2}}$.

Remark *(i)* Comparing the results of Theorem 6(i) and Theorem 7(i) (for L^r, Hellinger and KL divergence), we would like to highlight that while the conditions on λ and n match in both the cases, the latter does not require $f_0 \in \overline{\mathcal{R}(C)}$ to ensure consistency. While $f_0 \in \overline{\mathcal{R}(C)}$ can be imposed in Theorem 7 to attain consistency, we replaced this condition with $supp(q_0) = \Omega$—a simple and easy condition to work with—which along with the boundedness of the kernel ensures that for any $f_0 \in \mathcal{H}$, there exists $\tilde{f}_0 \in \overline{\mathcal{R}(C)}$ such that $p_{\tilde{f}_0} = p_0$ (see Lemma 14).

(ii) In contrast to the results in L^r, Hellinger and KL divergence, consistency in J can be obtained with λ converging to zero at a rate faster than in these results. In addition, one can obtain rates in J with $\beta = 0$, i.e., no smoothness assumption on f_0, while no rates are possible in other distances (the latter might also be an artifact of the proof technique, as these results are obtained through an application of Theorem 6(ii) in Lemma A.1) which is due to the fact that the convergence in these other distances is based on the convergence of $\|f_{\lambda,n} - f_0\|_{\mathcal{H}}$, which in turn involves convergence of $\mathcal{A}_0(\lambda) := \|f_\lambda - f_0\|_{\mathcal{H}}$ to zero while the convergence in J is controlled by $\mathcal{A}_{\frac{1}{2}}(\lambda) := \|\sqrt{C}(f_\lambda - f_0)\|_{\mathcal{H}}$ which can be shown to behave as $O(\sqrt{\lambda})$ as $\lambda \to 0$, without requiring any assumptions on f_0 (see Proposition A.3). Indeed, as a further consequence, the rate of convergence in J is faster than in other distances.

(iii) An interesting aspect in Theorem 7 is that $p_{f_{\lambda,n}}$ is consistent in various distances such as L^r, Hellinger and KL, despite being obtained by minimizing a different loss function, i.e., J. However, we will see in Section 5 that such nice results are difficult to obtain in the misspecified case, where consistency and rates are provided only in J.

While Theorem 7 addresses the case of bounded kernels, the case of unbounded kernels requires a technical modification. The reason for this modification, as alluded to in the discussion following Theorem 4, is due to the fact that $f_{\lambda,n}$ may not be in $\mathcal{F}$ when k is unbounded, and therefore the corresponding density estimator, $p_{f_{\lambda,n}}$ may not be well-defined. In order to keep the main ideas intact, we discuss the unbounded case in detail in Section B.2 in Appendix B.

4.2 Range Space Assumption

While Theorems 6 and 7 are satisfactory from the point of view of consistency, we believe the presented rates are possibly not minimax optimal since these rates are valid for any RKHS that satisfies the conditions **(A)**–**(D)** and does not capture the smoothness of k (and therefore the corresponding $\mathcal{H}$). In other words, the rates presented in Theorems 6 and 7 should depend on the decay rate of the eigenvalues of C which in turn effectively captures the smoothness of $\mathcal{H}$. However, we are not able to obtain such a result—see the remark following the proof of Theorem 6 for a discussion. While these rates do not reflect the intrinsic smoothness of $\mathcal{H}$, they are obtained under the smoothness assumption, i.e., *range space condition* that $f_0 \in \mathcal{R}(C^\beta)$ for some $\beta > 0$. This condition is quite different from the classical smoothness conditions that appear in non-parametric function estimation. While the range space assumption has been made in various earlier works (e.g., Caponnetto and Vito (2007); Smale and Zhou (2007); Fukumizu et al. (2013) in the context of non-parametric least square regression), in the following, we investigate the implicit smoothness assumptions that it makes on f_0 in our context. To this end, first it is easy to show (see the proof of Proposition B.3 in Section B.3) that

$$\mathcal{R}(C^\beta) = \left\{ \sum_{i\in I} c_i \phi_i \,:\, \sum_{i \in I} c_i^2 \alpha_i^{-2\beta} < \infty \right\}, \tag{9}$$

where $(\alpha_i)_{i\in I}$ are the positive eigenvalues of C, $(\phi_i)_{i\in I}$ are the corresponding eigenvectors that form an orthonormal basis for $\mathcal{R}(C)$, and I is an index set which is either finite (if $\mathcal{H}$ is finite-dimensional) or $I = \mathbb{N}$ with $\lim_{i\to\infty} \alpha_i = 0$ (if $\mathcal{H}$ is infinite dimensional). From (9) it is clear that larger the value of β, the faster is the decay of the Fourier coefficients $(c_i)_{i\in I}$, which in turn implies that the functions in $\mathcal{R}(C^\beta)$ are smoother. Using (9), an interpretation can be provided for $\mathcal{R}(C^\beta)$ ($\beta > 0$ and $\beta \notin \mathbb{N}$) as interpolation spaces (see Section A.5 for the definition of interpolation spaces) between $\mathcal{R}(C^{\lceil \beta \rceil})$ and $\mathcal{R}(C^{\lfloor \beta \rfloor})$ where $\mathcal{R}(C^0) := \mathcal{H}$ (see Proposition B.3 for details). While it is not completely straightforward to obtain a sufficient condition for $f_0 \in \mathcal{R}(C^\beta)$, $\beta \in \mathbb{N}$, the following result provides a necessary condition for $f_0 \in \mathcal{R}(C)$ (and therefore a necessary condition for $f_0 \in \mathcal{R}(C^\beta)$, $\forall\, \beta > 1$) for translation invariant kernels on $\Omega = \mathbb{R}^d$, whose proof is presented in Section 8.7.

Proposition 8 (Necessary condition) *Suppose $\psi, \phi \in C_b(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$ are positive definite functions on $\mathbb{R}^d$ with Fourier transforms $\psi^\wedge$ and $\phi^\wedge$ respectively. Let $\mathcal{H}$ and $\mathcal{G}$ be the RKHSs associated with $k(x,y) = \psi(x-y)$ and $l(x,y) = \phi(x-y), x, y \in \mathbb{R}^d$ respectively. For $1 \le r \le 2$, suppose the following hold:*

(i) $\int_{\mathbb{R}^d} \|\omega\|_2^2 \psi^\wedge(\omega)\, d\omega < \infty$; (ii) $\left\| \frac{\phi^\wedge}{\psi^\wedge} \right\|_\infty < \infty$; (iii) $\frac{\|\cdot\|_2^2 (\psi^\wedge)^2}{\phi^\wedge} \in L^{\frac{r}{2-r}}(\mathbb{R}^d)$; (iv) $q_0 \in L^r(\mathbb{R}^d)$.

Then $f_0 \in \mathcal{R}(C)$ implies $f_0 \in \mathcal{G} \subset \mathcal{H}$.

In the following, we apply the above result in two examples involving Gaussian and Matérn kernels to get insights into the range space assumption.

Example 2 (Gaussian kernel) *Let $\psi(x) = e^{-\sigma\|x\|^2}$ with $\mathcal{H}_\sigma$ as its corresponding RKHS (see Section 2 for its definition). By Proposition 8, it is easy to verify that $f_0 \in \mathcal{R}(C)$ implies $f_0 \in \mathcal{H}_\alpha \subset \mathcal{H}_\sigma$ for $\frac{\sigma}{2} < \alpha \le \sigma$. Since $\mathcal{H}_\beta \subset \mathcal{H}_\gamma$ for $\beta < \gamma$ (i.e., Gaussian RKHSs are nested), $f_0 \in \mathcal{R}(C)$ ensures that f_0 lies in $\mathcal{H}_{\frac{\sigma}{2}+\epsilon}$ for arbitrary small $\epsilon > 0$.*

Example 3 (Matérn kernel) *Let $\psi(x) = \frac{2^{1-s}}{\Gamma(s)} \|x\|_2^{s-\frac{d}{2}} \mathfrak{K}_{d/2-s}(\|x\|_2)$, $x \in \mathbb{R}^d$ with $H_2^s(\mathbb{R}^d)$ as its corresponding RKHS (see Section 2 for its definition) where $s > \frac{d}{2}$. By Proposition 8, we have that for $q_0 \in L^1(\mathbb{R}^d)$, if $f_0 \in \mathcal{R}(C)$, then $f_0 \in H_2^\alpha(\mathbb{R}^d) \subset H_2^s(\mathbb{R}^d)$ for $1 + \frac{d}{2} < s \le$*

$\alpha < 2s - 1 - \frac{d}{2}$. Since $H_2^\delta(\mathbb{R}^d) \subset H_2^\gamma(\mathbb{R}^d)$ for $\gamma < \delta$ (i.e., Sobolev spaces are nested), this means f_0 lies in $H_2^{2s-1-\frac{d}{2}-\epsilon}(\mathbb{R}^d)$ for arbitrarily small $\epsilon > 0$, i.e., f_0 has at least $2s - 1 - \lceil \frac{d}{2} \rceil$ weak-derivatives. By the minimax theory (Tsybakov, 2009, Chapter 2), it is well known that for any $\alpha > \delta \geq 0$,

$$\inf_{\hat{f}_n} \sup_{f_0 \in H_2^\alpha(\mathbb{R}^d)} \|\hat{f}_n - f_0\|_{H_2^\delta(\mathbb{R}^d)} \asymp n^{-\frac{\alpha-\delta}{2(\alpha-\delta)+d}}, \tag{10}$$

where the infimum is taken over all possible estimators. Here $a_n \asymp b_n$ means that for any two sequences $a_n, b_n > 0$, a_n/b_n is bounded away from zero and infinity as $n \to \infty$. Suppose $f_0 \notin H_2^\alpha(\mathbb{R}^d)$ for $\alpha \geq 2s - 1 - \frac{d}{2}$, which means $f_0 \in H_2^{2s-1-\frac{d}{2}-\epsilon}(\mathbb{R}^d)$ for arbitrarily small $\epsilon > 0$. This implies that the rate of $n^{-1/4}$ obtained in Theorem 6 is minimax optimal if $\mathcal{H}$ is chosen to be $H_2^{1+d+\epsilon}(\mathbb{R}^d)$ (i.e., choose $\alpha = 2s - 1 - \frac{d}{2} - \epsilon$ and $\delta = s$ in (10) and solve for s by equating the exponent in the r.h.s. of (10) to $-\frac{1}{4}$). Similarly, it can be shown that if $q_0 \in L^2(\mathbb{R}^d)$, then the rate of $n^{-1/4}$ in Theorem 6 is minimax optimal if $\mathcal{H}$ is chosen to be $H_2^{1+\frac{d}{2}+\epsilon}(\mathbb{R}^d)$. This example also explains away the dimension independence of the rate provided by Theorem 6 by showing that the dimension effect is captured in the relative smoothness of f_0 w.r.t. $\mathcal{H}$.

While Example 3 provides some understanding about the minimax optimality of $f_{\lambda,n}$ under additional assumptions on f_0, the problem is not completely resolved. In the following section, however, we show that the rate in Theorem 6 is not optimal for $\beta > 1$, and that improved rates can be obtained by choosing the regularizer appropriately.

4.3 Choice of Regularizer

We understand from the characterization of $\mathcal{R}(C^\beta)$ in (9) that larger β values yield smoother functions in $\mathcal{H}$. However, the smoothness of $f_0 \in \mathcal{R}(C^\beta)$ for $\beta > 1$ is not captured in the rates in Theorem 6(ii), where the rate saturates at $\beta = 1$ providing the best possible rate of $n^{-1/4}$ (irrespective of the size of β). This is unsatisfactory on the part of the estimator, as it does not effectively capture the smoothness of f_0, i.e., the estimator is not adaptive to the smoothness of f_0. We remind the reader that the estimator $f_{\lambda,n}$ is obtained by minimizing the regularized empirical Fisher divergence (see Theorem 4(iv)) yielding $f_{\lambda,n} = -(\hat{C} + \lambda I)^{-1}\hat{\xi}$, which can be seen as a heuristic to solve the (non-linear) inverse problem $Cf_0 = -\xi$ (see Theorem 4(ii)) from finite samples, by replacing C and ξ with their empirical counterparts. This heuristic, which ensures that the finite sample inverse problem is well-posed, is popular in inverse problem literature under the name of Tikhonov regularization (Engl et al., 1996, Chapter 5). Note that Tikhonov regularization helps to make the ill-posed inverse problem a well-posed one by approximating α^{-1} by $(\alpha + \lambda)^{-1}$, $\lambda > 0$, where α^{-1} appears as the inverse of the eigenvalues of C while computing C^{-1}. In other words, if $\hat{C}$ is invertible, then an estimate of f_0 can be obtained as $\hat{f}_n = -\hat{C}^{-1}\hat{\xi}$, i.e., $\hat{f}_n = -\sum_{i \in I} \frac{\langle \hat{\xi}, \hat{\phi}_i \rangle_{\mathcal{H}}}{\hat{\alpha}_i} \hat{\phi}_i$, where $(\hat{\alpha}_i)_{i \in I}$ and $(\hat{\phi}_i)_{i \in I}$ are the eigenvalues and eigenvectors of $\hat{C}$ respectively. However, $\hat{C}$ being a rank n operator defined on $\mathcal{H}$ (which can be infinite dimensional) is not invertible and therefore the regularized estimator is constructed as $f_{\lambda,n} = -g_\lambda(\hat{C})\hat{\xi}$ where $g_\lambda(\hat{C})$ is defined through functional calculus (see Engl, Hanke, and

Neubauer, 1996, Section 2.3) as

$$g_\lambda(\hat{C}) = \sum_{i \in I} g_\lambda(\hat{\alpha}_i)\langle \cdot, \hat{\phi}_i \rangle_{\mathcal{H}} \hat{\phi}_i$$

with $g_\lambda : \mathbb{R}_+ \to \mathbb{R}$ and $g_\lambda(\alpha) := (\alpha + \lambda)^{-1}$. Since the Tikhonov regularization is well-known to saturate (as explained above)—see Engl et al. (1996, Sections 4.2 and 5.1) for details—, better approximations to α^{-1} have been used in the inverse problems literature to improve the rates by using g_λ other than $(\cdot + \lambda)^{-1}$ where $g_\lambda(\alpha) \to \alpha^{-1}$ as $\lambda \to 0$. In the statistical context, Rosasco et al. (2005) and Bauer et al. (2007) have used the ideas from Engl et al. (1996) in non-parametric regression for learning a square integrable function from finite samples through regularization in RKHS. In the following, we use these ideas to construct an alternate estimator for f_0 (and therefore for p_0) that appropriately captures the smoothness of f_0 by providing a better convergence rate when $\beta > 1$. To this end, we need the following assumption—quoted from Engl et al. (1996, Theorems 4.1–4.3 and Corollary 4.4) and Bauer et al. (2007, Definition 1)—that is standard in the theory of inverse problems.

(E) There exists finite positive constants A_g, B_g, C_g, η_0 and $(\gamma_\eta)_{\eta \in (0,\eta_0]}$ (all independent of $\lambda > 0$) such that $g_\lambda : [0, \chi] \to \mathbb{R}$ satisfies:

(a) $\sup_{\alpha \in \mathcal{D}} |\alpha g_\lambda(\alpha)| \leq A_g$, (b) $\sup_{\alpha \in \mathcal{D}} |g_\lambda(\alpha)| \leq \frac{B_g}{\lambda}$, (c) $\sup_{\alpha \in \mathcal{D}} |1 - \alpha g_\lambda(\alpha)| \leq C_g$ and (d) $\sup_{\alpha \in \mathcal{D}} |1 - \alpha g_\lambda(\alpha)|\alpha^\eta \leq \gamma_\eta \lambda^\eta$, $\forall\, \eta \in (0, \eta_0]$ where $\mathcal{D} := [0, \chi]$ and $\chi := d \sup_{x \in \Omega, i \in [d]} \partial_i \partial_{i+d} k(x, x) < \infty$.

The constant η_0 is called the *qualification* of g_λ which is what determines the point of saturation of g_λ. We show in Theorem 9 that if g_λ has a finite qualification, then the resultant estimator cannot fully exploit the smoothness of f_0 and therefore the rate of convergence will suffer for $\beta > \eta_0$. Given g_λ that satisfies **(E)**, we construct our estimator of f_0 as

$$f_{g,\lambda,n} = -g_\lambda(\hat{C})\hat{\xi}.$$

Note that the above estimator can be obtained by using the data dependent regularizer, $\frac{1}{2}\langle f, ((g_\lambda(\hat{C}))^{-1} - \hat{C})f \rangle_{\mathcal{H}}$ in the minimization of $\hat{J}(f)$ defined in Theorem 4(iv), i.e.,

$$f_{g,\lambda,n} = \arg \inf_{f \in \mathcal{H}} \hat{J}(f) + \frac{1}{2}\langle f, ((g_\lambda(\hat{C}))^{-1} - \hat{C})f \rangle_{\mathcal{H}}.$$

However, unlike $f_{\lambda,n}$ for which a simple form is available in Theorem 5 by solving a linear system, we are not able to obtain such a nice expression for $f_{g,\lambda,n}$. The following result (proved in Section 8.8) presents an analog of Theorems 6 and 7 for the new estimators, $f_{g,\lambda,n}$ and $p_{f_{g,\lambda,n}}$.

Theorem 9 (Consistency and convergence rates for $f_{g,\lambda,n}$ and $p_{f_{g,\lambda,n}}$) *Suppose* **(A)**–**(E)** *hold with* $\varepsilon = 2$.

(i) If $f_0 \in \mathcal{R}(C^\beta)$ *for some* $\beta > 0$, *then for any* $\lambda \geq n^{-1/2}$,

$$\|f_{g,\lambda,n} - f_0\|_{\mathcal{H}} = O_{p_0}(\theta_n),$$

where $\theta_n := n^{-\min\left\{\frac{\beta}{2(\beta+1)}, \frac{\eta_0}{2(\eta_0+1)}\right\}}$ *with* $\lambda = n^{-\max\left\{\frac{1}{2(\beta+1)}, \frac{1}{2(\eta_0+1)}\right\}}$. *In addition, if* $\|k\|_\infty < \infty$, *then for any* $1 < r \le \infty$ *with* $q_0 \in L^1(\Omega) \cap L^r(\Omega)$,

$$\|p_{f_{g,\lambda,n}} - p_0\|_{L^r(\Omega)} = O_{p_0}(\theta_n),\ h(p_0, p_{f_{g,\lambda,n}}) = O_{p_0}(\theta_n)\ \text{ and }\ KL(p_0\|p_{f_{g,\lambda,n}}) = O_{p_0}(\theta_n^2).$$

(ii) If $f_0 \in \mathcal{R}(C^\beta)$ *for some* $\beta \ge 0$, *then for any* $\lambda \ge n^{-1/2}$,

$$J(p_0\|p_{f_{g,\lambda,n}}) = O_{p_0}\left(n^{-\frac{\min\{2\beta+1, 2\eta_0\}}{\min\{2\beta+2, 2\eta_0+1\}}}\right)$$

with $\lambda = n^{-\frac{1}{\min\{2\beta+2, 2\eta_0+1\}}}$.

(iii) If $\|C^{-1}\| < \infty$, *then for any* $\lambda \ge n^{-1/2}$,

$$\|f_{g,\lambda,n} - f_0\|_{\mathcal{H}} = O_{p_0}(\theta_n)\ \text{ and }\ J(p_0\|p_{f_{g,\lambda,n}}) = O_{p_0}(\theta_n^2)$$

with $\theta_n = n^{-\frac{1}{2}}$ *and* $\lambda = n^{-\frac{1}{\min\{2, 2\eta_0\}}}$.

Theorem 9 shows that if g_λ has infinite qualification, then smoothness of f_0 is fully captured in the rates and as $\beta \to \infty$, we attain $O_{p_0}(n^{-1/2})$ rate for $\|f_{g,\lambda,n} - f_0\|_{\mathcal{H}}$ in contrast to $n^{-1/4}$ (similar improved rates are also obtained for $p_{f_{g,\lambda,n}}$ in various distances) in Theorem 6. In the following example, we present two choices of g_λ that improve on Tikhonov regularization. We refer the reader to Rosasco et al. (2005, Section 3.1) for more examples of g_λ.

Example 4 (Choices of g_λ) *(i) Tikhonov regularization involves* $g_\lambda(\alpha) = (\alpha + \lambda)^{-1}$ *for which it is easy to verify that* $\eta_0 = 1$ *and therefore the rates saturate at* $\beta = 1$, *leading to the results in Theorems 6 and 7.*

(ii) Showalter's method and spectral cut-off use

$$g_\lambda(\alpha) = \frac{1 - e^{-\alpha/\lambda}}{\alpha}\ \text{ and }\ g_\lambda(\alpha) = \begin{cases} \frac{1}{\alpha}, & \alpha \ge \lambda \\ 0, & \alpha < \lambda \end{cases}$$

respectively for which it is easy to verify that $\eta_0 = +\infty$ *(see Engl, Hanke, and Neubauer, 1996, Examples 4.7 & 4.8 for details) and therefore improved rates are obtained for* $\beta > 1$ *in Theorem 9 compared to that of Tikhonov regularization.*

5. Density Estimation in $\mathcal{P}$: Misspecified Case

In this section, we analyze the misspecified case where $p_0 \notin \mathcal{P}$, which is a more reasonable case than the well-specified one, as in practice it is not easy to check whether $p_0 \in \mathcal{P}$. To this end, we consider the same estimator $p_{f_{\lambda,n}}$ as considered in the well-specified case where $f_{\lambda,n}$ is obtained from Theorem 5. The following result shows that $J(p_0\|p_{f_{\lambda,n}}) \to \inf_{p\in\mathcal{P}} J(p_0\|p)$ as $\lambda \to 0$, $\lambda n \to \infty$ and $n \to \infty$ under the assumption that there exists $f^* \in \mathcal{F}$ such that $J(p_0\|p_{f^*}) = \inf_{p\in\mathcal{P}} J(p_0\|p)$. We present the result for bounded kernels although it can be easily extended to unbounded kernels as in Theorem B.2. Also, the presented result for Tikhonov regularization extends easily to $p_{f_{g,\lambda,n}}$ using the ideas in the proof of Theorem 9. Note that unlike in the well-specified case where convergence in other distances can be shown even though the estimator is constructed from J, it is difficult to show such a result in the misspecified case.

Theorem 10 *Let p_0, $q_0 \in C^1(\Omega)$ be probability densities such that $J(p_0\|q_0) < \infty$ where Ω satisfies* **(A)**. *Assume that* **(B)**, **(C)** *and* **(D)** *with $\varepsilon = 2$ hold. Suppose $\|k\|_\infty < \infty$, $\mathrm{supp}(q_0) = \Omega$ and there exists $f^* \in \mathcal{F}$ such that*

$$J(p_0\|p_{f^*}) = \inf_{p\in\mathcal{P}} J(p_0\|p).$$

Then for an estimator $p_{f_{\lambda,n}}$ constructed from random samples $(X_a)_{a=1}^n$ drawn i.i.d. from p_0, where $f_{\lambda,n}$ is defined in (7)—also see Theorem 4(iv)—with $\lambda > 0$, we have

$$J(p_0\|p_{f_{\lambda,n}}) \to \inf_{p\in\mathcal{P}} J(p_0\|p) \ \ \text{as } \lambda \to 0,\ \lambda n \to \infty \text{ and } n \to \infty.$$

In addition, if $f^* \in \mathcal{R}(C^\beta)$ for some $\beta \geq 0$, then

$$\sqrt{J(p_0\|p_{f_{\lambda,n}})} \leq \sqrt{\inf_{p\in\mathcal{P}} J(p_0\|p)} + O_{p_0}\left(n^{-\min\left\{\frac{1}{3},\frac{2\beta+1}{4(\beta+1)}\right\}}\right)$$

with $\lambda = n^{-\max\left\{\frac{1}{3},\frac{1}{2(\beta+1)}\right\}}$. If $\|C^{-1}\| < \infty$, then for $\lambda = n^{-\frac{1}{2}}$,

$$\sqrt{J(p_0\|p_{f_{\lambda,n}})} \leq \sqrt{\inf_{p\in\mathcal{P}} J(p_0\|p)} + O_{p_0}(n^{-1/2}).$$

with $\lambda = n^{-\frac{1}{2}}$.

While the above result is useful and interesting, the assumption about the existence of f^* is quite restrictive. This is because if p_0 (which is not in $\mathcal{P}$) belongs to a family $\mathcal{Q}$ where $\mathcal{P}$ is dense in $\mathcal{Q}$ w.r.t. J, then there is no $f \in \mathcal{H}$ that attains the infimum, i.e., f^* does not exist and therefore the proof technique employed in Theorem 10 will fail. In the following, we present a result (Theorem 12) that does not require the existence of f^* but attains the same result as in Theorem 10, but requiring a more complicated proof. Before we present Theorem 12, we need to introduce some notation.

To this end, let us return to the objective function under consideration,

$$J(p_0\|p_f) = \frac{1}{2}\int_\Omega p_0(x)\left\|\nabla \log \frac{p_0}{p_f}\right\|_2^2 dx = \frac{1}{2}\int_\Omega p_0(x)\sum_{i=1}^d \left(\partial_i f_\star - \partial_i f\right)^2 dx,$$

where $f_\star = \log \frac{p_0}{q_0}$ and $p_0 \notin \mathcal{P}$. Define

$$\mathcal{W}_2(\Omega, p_0) := \left\{f \in C^1(\Omega) \,:\, \partial^\alpha f \in L^2(\Omega, p_0),\, \forall\,|\alpha| = 1\right\}.$$

This is a reasonable class of functions to consider as under the condition $J(p_0\|q_0) < \infty$, it is clear that $f_\star \in \mathcal{W}_2(\Omega, p_0)$. Endowed with a semi-norm,

$$\|f\|_{\mathcal{W}_2}^2 := \sum_{|\alpha|=1} \|\partial^\alpha f\|_{L^2(\Omega,p_0)}^2,$$

$\mathcal{W}_2(\Omega, p_0)$ is a vector space of functions, from which a normed space can be constructed as follows. Let us define $f, f' \in \mathcal{W}_2(\Omega, p_0)$ to be equivalent, i.e., $f \sim f'$, if $\|f - f'\|_{\mathcal{W}_2} = 0$.

In other words, $f \sim f'$ if and only if f and f' differ by a constant p_0-almost everywhere. Now define the quotient space $\mathcal{W}_2^\sim(\Omega, p_0) := \{[f]_\sim : f \in \mathcal{W}_2(\Omega, p_0)\}$ where $[f]_\sim := \{f' \in \mathcal{W}_2(\Omega, p_0) : f \sim f'\}$ denotes the equivalence class of f. Defining $\|[f]_\sim\|_{\mathcal{W}_2^\sim} := \|f\|_{\mathcal{W}_2}$, it is easy to verify that $\|\cdot\|_{\mathcal{W}_2^\sim}$ defines a norm on $\mathcal{W}_2^\sim(p_0)$. In addition, endowing the following bilinear form on $\mathcal{W}_2^\sim(\Omega, p_0)$

$$\langle [f]_\sim, [g]_\sim \rangle_{\mathcal{W}_2^\sim} := \int_\Omega p_0(x) \sum_{|\alpha|=1} (\partial^\alpha f)(x)(\partial^\alpha g)(x)\, dx$$

makes it a pre-Hilbert space. Let $W_2(\Omega, p_0)$ be the Hilbert space obtained by completion of $\mathcal{W}_2^\sim(\Omega, p_0)$. As shown in Proposition 11 below, under some assumptions, a continuous mapping $I_k : \mathcal{H} \to W_2(\Omega, p_0), f \mapsto [f]_\sim$ can be defined, which is injective modulo constant functions. Since addition of a constant does not contribute to p_f, the space $W_2(\Omega, p_0)$ can be regarded as a parameter space extended from $\mathcal{H}$. In addition to I_k, Proposition 11 (proved in Section 8.10) describes the adjoint of I_k and relevant self-adjoint operators, which will be useful in analyzing $p_{f_{\lambda,n}}$ in Theorem 12.

Proposition 11 *Let* $\mathrm{supp}(q_0) = \Omega$ *where* $\Omega \subset \mathbb{R}^d$ *is non-empty and open. Suppose* k *satisfies* **(B)** *and* $\partial_i \partial_{i+d} k(x, x) \in L^1(\Omega, p_0)$ *for all* $i \in [d]$*. Then* $I_k : \mathcal{H} \to W_2(\Omega, p_0)$*,* $f \mapsto [f]_\sim$ *defines a continuous mapping with the null space* $\mathcal{H} \cap \mathbb{R}$*. The adjoint of* I_k *is* $S_k : W_2(\Omega, p_0) \to \mathcal{H}$ *whose restriction to* $\mathcal{W}_2^\sim(\Omega, p_0)$ *is given by*

$$S_k[h]_\sim(y) = \int_\Omega \sum_{i=1}^d \partial_i k(x, y) \partial_i h(x)\, p_0(x)\, dx, \qquad [h]_\sim \in \mathcal{W}_2^\sim(\Omega, p_0),\ y \in \Omega.$$

In addition, I_k *and* S_k *are Hilbert-Schmidt and therefore compact. Also,* $E_k := S_k I_k$ *and* $T_k := I_k S_k$ *are compact, positive and self-adjoint operators on* $\mathcal{H}$ *and* $W_2(\Omega, p_0)$ *respectively where*

$$E_k g(y) = \int_\Omega \sum_{i=1}^d \partial_i k(x, y) \partial_i g(x) p_0(x)\, dx, \qquad g \in \mathcal{H},\ y \in \Omega$$

and the restriction of T_k *to* $\mathcal{W}_2^\sim(\Omega, p_0)$ *is given by*

$$T_k[h]_\sim = \left[\int_\Omega \sum_{i=1}^d \partial_i k(x, \cdot) \partial_i h(x)\, p_0(x)\, dx \right]_\sim, \qquad [h]_\sim \in \mathcal{W}_2^\sim(\Omega, p_0).$$

Note that for $[h]_\sim \in \mathcal{W}_2^\sim(\Omega, p_0)$, the derivatives $\partial_i h$ do not depend on the choice of a representative element almost surely w.r.t. p_0, and thus the above integrals are well defined. Having constructed $W_2(\Omega, p_0)$, it is clear that $J(p_0\|p_f) = \frac{1}{2}\|[f_\star]_\sim - I_k f\|_{W_2}^2$, which means estimating p_0 is equivalent to estimating $f_\star \in W_2(\Omega, p_0)$ by $f \in \mathcal{F}$. With all these preparations, we are now ready to present a result (proved in Section 8.11) on consistency and convergence rate for $p_{f_{\lambda,n}}$ without assuming the existence of f^*.

Theorem 12 *Let* $p_0,\ q_0 \in C^1(\Omega)$ *be probability densities such that* $J(p_0\|q_0) < \infty$*. Assume that* **(A)**–**(D)** *hold with* $\varepsilon = 2$ *and* $\chi := d \sup_{x \in \Omega, i \in [d]} \partial_i \partial_{i+d} k(x, x) < \infty$*. Then the following* *hold.*

(i) As $\lambda \to 0$, $\lambda n \to \infty$ *and* $n \to \infty$, $J(p_0\|p_{f_{\lambda,n}}) \to \inf_{p\in\mathcal{P}} J(p_0\|p)$.

(ii) Define $f_\star := \log \frac{p_0}{q_0}$. *If* $[f_\star]_\sim \in \overline{\mathcal{R}(T_k)}$, *then*

$$J(p_0\|p_{f_{\lambda,n}}) \to 0 \ \ \textit{as} \ \ \lambda \to 0,\, \lambda n \to \infty \ \textit{and} \ n \to \infty.$$

In addition, if $[f_\star]_\sim \in \mathcal{R}(T_k^\beta)$ *for some* $\beta > 0$, *then for* $\lambda = n^{-\max\left\{\frac{1}{3},\frac{1}{2\beta+1}\right\}}$

$$J(p_0\|p_{f_{\lambda,n}}) = O_{p_0}\left(n^{-\min\left\{\frac{2}{3},\frac{2\beta}{2\beta+1}\right\}}\right)$$

. *(iii) If* $\|E_k^{-1}\| < \infty$ *and* $\|T_k^{-1}\| < \infty$, *then* $J(p_0\|p_{f_{\lambda,n}}) = O_{p_0}\left(n^{-1}\right)$ *with* $\lambda = n^{-\frac{1}{2}}$.

Remark *(i)* The result in Theorem 12(ii) is particularly interesting as it shows that $[f_\star]_\sim \in W_2(\Omega, p_0)\backslash I_k(\mathcal{H})$ can be consistently estimated by $f_{\lambda,n} \in \mathcal{H}$, which in turn implies that certain $p_0 \notin \mathcal{P}$ can be consistently estimated by $p_{f_{\lambda,n}} \in \mathcal{P}$. In particular, if S_k is injective, then $I_k(\mathcal{H})$ is dense in $W_2(\Omega, p_0)$ w.r.t. $\|\cdot\|_{W_2}$, which implies $\inf_{p\in\mathcal{P}} J(p_0||p) = 0$ though there does not exist $f^* \in \mathcal{H}$ for which $J(p_0||p_{f^*}) = 0$. While Theorem 10 cannot handle this situation, (i) and (ii) in Theorem 12 coincide showing that $p_0 \notin \mathcal{P}$ can be consistently estimated by $p_{f_{\lambda,n}} \in \mathcal{P}$. While the question of when $I_k(\mathcal{H})$ is dense in $W_2(\Omega, p_0)$ is open, we refer the reader to Section B.4 for a related discussion.

(ii) Replicating the proof of Theorem 4.6 in Steinwart and Scovel (2012), it is easy to show that for all $0 < \gamma < 1$, $\mathcal{R}(T_k^{\gamma/2}) = [W_2(\Omega, p_0), I_k(\mathcal{H})]_{\gamma,2}$, where the r.h.s. is an interpolation space obtained through the real interpolation of $W_2(\Omega, p_0)$ and $I_k(\mathcal{H})$ (see Section A.5 for the notation and definition). Here $I_k(\mathcal{H})$ is endowed with the Hilbert space structure by $I_k(\mathcal{H}) \cong \mathcal{H}/\mathcal{H}\cap\mathbb{R}$. This interpolation space interpretation means that, for $\beta \geq \frac{1}{2}$, $\mathcal{R}(T_k^\beta) \subset \mathcal{H}$ modulo constant functions. It is nice to note that the rates in Theorem 12(ii) for $\beta \geq \frac{1}{2}$ match with the rates in Theorem 7 (i.e., the well-specified case) w.r.t. J for $0 \leq \beta \leq \frac{1}{2}$. We highlight the fact that $\beta = 0$ corresponds to $\mathcal{H}$ in Theorem 7 whereas $\beta = \frac{1}{2}$ corresponds to $\mathcal{H}$ in Theorem 12(ii) and therefore the range of comparison is for $\beta \geq \frac{1}{2}$ in Theorem 12(ii) versus $0 \leq \beta \leq \frac{1}{2}$ in Theorem 7. In contrast, Theorem 10 is very limited as it only provides a rate for the convergence of $J(p_0||p_{f_{\lambda,n}})$ to $\inf_{p\in\mathcal{P}} J(p_0||p)$ assuming that f^* is sufficiently smooth.

Based on the observation (i) in the above remark that $\inf_{p\in\mathcal{P}} J(p_0\|p) = 0$ if $I_k(\mathcal{H})$ is dense in $W_2(\Omega, p_0)$ w.r.t. $\|\cdot\|_{W_2}$, it is possible to obtain an approximation result for $\mathcal{P}$ (similar to those discussed in Section 3) w.r.t. Fisher divergence as shown below, whose proof is provided in Section 8.12.

Proposition 13 *Suppose* $\Omega \subset \mathbb{R}^d$ *is non-empty and open. Let* $q_0 \in C^1(\Omega)$ *be a probability density and*

$$\mathcal{P}_{\mathrm{FD}} := \left\{p \in C^1(\Omega) : \int_\Omega p(x)\, dx = 1, p(x) \geq 0, \forall\, x \in \Omega \text{ and } J(p\|q_0) < \infty\right\}.$$

For any $p \in \mathcal{P}_{\mathrm{FD}}$, *if* $I_k(\mathcal{H})$ *is dense in* $W_2(\Omega, p)$ *w.r.t.* $\|\cdot\|_{W_2}$, *then for every* $\epsilon > 0$, *there exists* $\tilde{p} \in \mathcal{P}$ *such that* $J(p\|\tilde{p}) \leq \epsilon$.

6. Numerical Simulations

We have proposed an estimator of p_0 that is obtained by minimizing the regularized empirical Fisher divergence and presented its consistency along with convergence rates. As discussed in Section 1, however one can simply ignore the structure of $\mathcal{P}$ and estimate p_0 in a completely non-parametric fashion, for example using the kernel density estimator (KDE). In fact, consistency and convergence rates of KDE are also well-studied (Tsybakov, 2009, Chapter 1) and the kernel density estimator is very simple to compute—requiring only $O(n)$ computations—compared to the proposed estimator, which is obtained by solving a linear system of size $nd \times nd$. This raises questions about the applicability of the proposed estimator in practice, though it is very well known that KDE performs poorly for moderate to large d (Wasserman, 2006, Section 6.5). In this section, we numerically demonstrate that the proposed score matching estimator performs significantly better than the KDE, and in particular, that the advantage with the proposed estimator grows as d gets large. Note further that the maximum likelihood approach of Barron and Sheu (1991) and Fukumizu (2009) does not yield estimators that are practically feasible, and therefore to the best of our knowledge, the proposed estimator is the only viable estimator for estimating densities through $\mathcal{P}$.

In the following, we consider two simple scenarios of estimating a multivariate normal and mixture of normals using the proposed estimator and demonstrate the superior performance of the proposed estimator over KDE. Inspired by this preliminary empirical investigation, recently, the proposed estimator has been explored in two concrete applications of gradient-free adaptive MCMC sampler (Strathmann et al., 2015) and graphical model structure learning (Sun et al., 2015) where the superiority of working with the infinite dimensional family is demonstrated. We would like to again highlight that the goal of this work is not to construct density estimators that improve upon KDE but to provide a novel modeling technique of approximating an unknown density by a rich parametric family of densities with the parameter being infinite dimensional in contrast to the classical approach of finite dimensional approximation.

We consider the problems of estimating a standard normal distribution on $\mathbb{R}^d$, $N(0, I_d)$ and mixture of Gaussians,

$$p_0(x) = \frac{1}{2}\phi_d(x; \alpha\mathbf{1}_n, I_d) + \frac{1}{2}\phi_d(x; \beta\mathbf{1}_n, I_d)$$

through the score matching approach and KDE, and compare their estimation accuracies. Here $\phi_d(x; \mu, \Sigma)$ is the p.d.f. of $N(\mu, \Sigma I_d)$. By choosing the kernel, $k(x, y) = \exp(-\frac{\|x-y\|_2^2}{2\sigma^2}) + r(x^T y + c)^2$, which is a Gaussian plus polynomial of degree 2, it is easy to verify that Gaussian distributions lie in $\mathcal{P}$, and therefore the first problem considers the well-specified case while the second problem deals with the misspecified case. In our simulations, we chose $r = 0.1$, $c = 0.5$, $\alpha = 4$ and $\beta = -4$. The base measure of the exponential family is $N(0, 10^2 I_d)$. The bandwidth parameter σ is chosen by cross-validation (CV) of the objective function $\hat{J}_\lambda$ (see Theorem 4(iv)) within the parameter set $\{0.1, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6\} \times \sigma_*$, where σ_* is the median of pairwise distances of data, and the regularization parameter λ is set as $\lambda = 0.1 \times n^{-1/3}$ with sample size n. For KDE, the Gaussian kernel is used for the smoothing kernel, and the bandwidth parameter is chosen by CV from

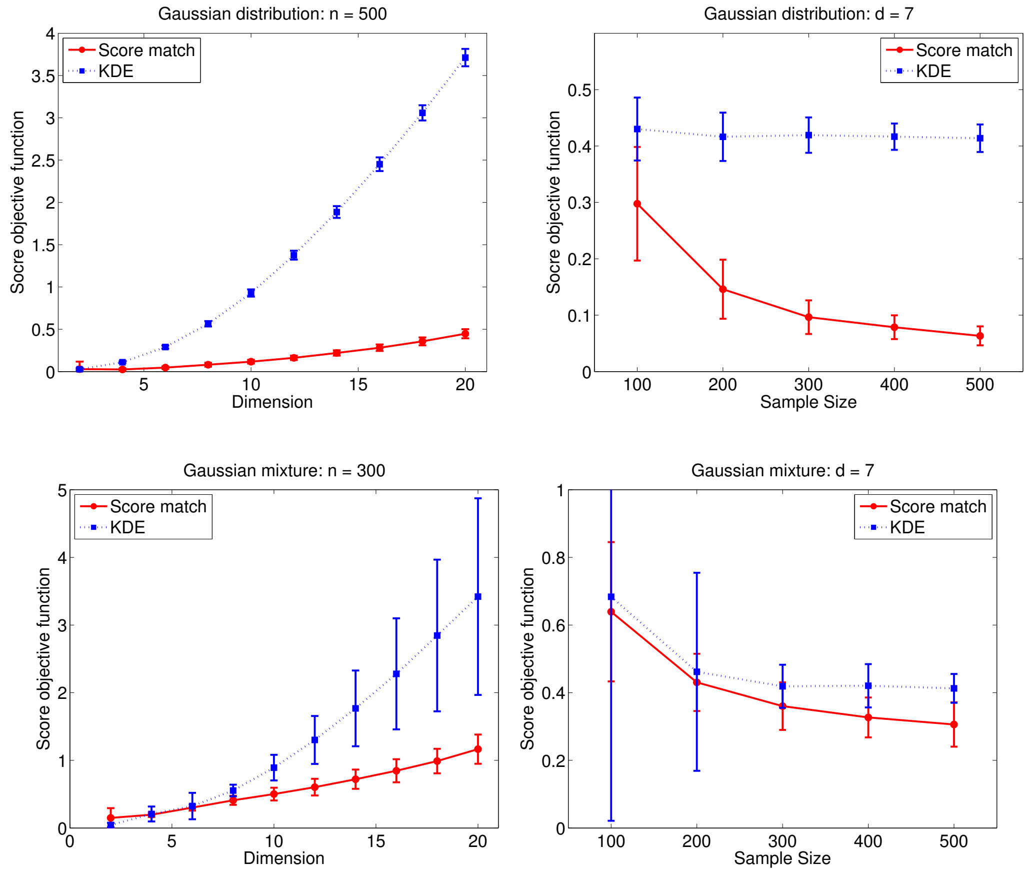

Figure 1: Experimental comparisons with the score objective function: proposed method and kernel density estimator

$\{0.02, 0.04, 0.06, 0.08, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0\} \times \sigma_*$; where for both the methods, 5-fold CV is applied. Since it is difficult to accurately estimate the normalization constant in the proposed method, we use two methods to evaluate the accuracy of estimation. One is the objective function for the score matching method,

$$\tilde{J}(p) = \sum_{i=1}^{d} \int_\Omega \left(\frac{1}{2} |\partial_i \log p(x)|^2 + \partial_i^2 \log p(x) \right) p_0(x) dx,$$

and the other is correlation of the estimator with the true density function,

$$\mathrm{Cor}(p, p_0) := \frac{\mathbb{E}_R[p(X)p_0(X)]}{\sqrt{\mathbb{E}_R[p(X)^2]\mathbb{E}_R[p_0(X)^2]}},$$

where R is a probability distribution. For R, we use the empirical distribution based on 10000 random samples drawn i.i.d. from $p_0(x)$.

Figures 1 and 2 show the score objective function ($\tilde{J}(p)$) and the correlation ($\mathrm{Cor}(p, p_0)$) (along with their standard deviation as error bars) of the proposed estimator and KDE for the tasks of estimating a Gaussian and a mixture of Gaussians, for different sample sizes

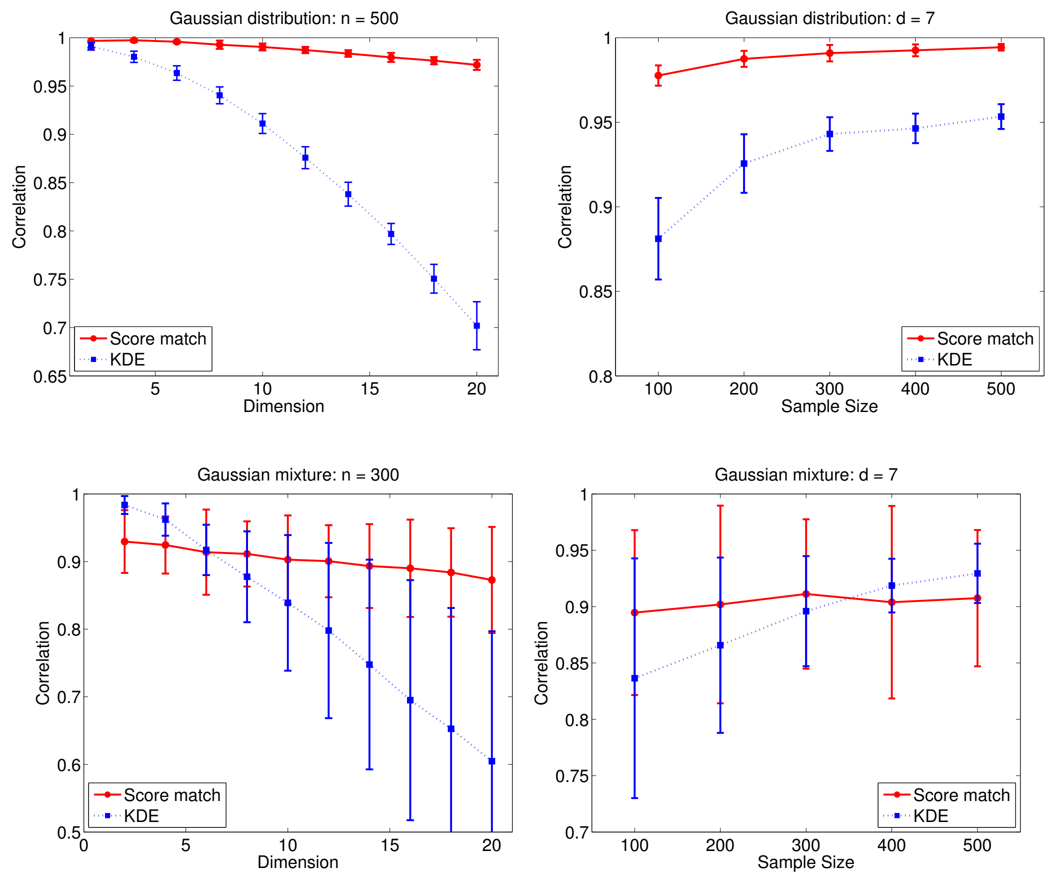

Figure 2: Experimental comparisons with the correlation: proposed method and kernel density estimator

(n) and dimensions (d). From the figures, we see that the proposed estimator outperforms (i.e., lower function values) KDE in all the cases except the low dimensional cases ($(n, d) = (500, 2)$ for the Gaussian, and $(n, d) = (300, 2), (300, 4)$ for the Gaussian mixture). In the case of the correlation measure, the score matching method yields better results (i.e., higher correlation) besides in the Gaussian mixture cases of $d = 2, 4, 6$ (Fig.2, lower-left) and some cases of $d = 7$ (lower-right). The proposed method shows an increased advantage over KDE as the dimensionality increases, thereby demonstrating the advantage of the proposed estimator for high dimensional data.

7. Summary & Discussion

We have considered an infinite dimensional generalization, $\mathcal{P}$, of the finite-dimensional exponential family, where the densities are indexed by functions in a reproducing kernel Hilbert space (RKHS), $\mathcal{H}$. We showed that $\mathcal{P}$ is a rich object that can approximate a large class of probability densities arbitrarily well in Kullback-Leibler divergence, and addressed the main question of estimating an unknown density, p_0 from finite samples drawn i.i.d. from it, in well-specified ($p_0 \in \mathcal{P}$) and misspecified ($p_0 \notin \mathcal{P}$) settings. We proposed a density estimator based on minimizing the regularized version of the empirical Fisher divergence, which results in solving a simple finite-dimensional linear system. Our estimator provides a computationally efficient alternative to maximum likelihood based estimators, which suffer from the computational intractability of the log-partition function. The proposed estimator is also shown to empirically outperform the classical kernel density estimator, with advantage increasing as the dimension of the space increases. In addition to these computational and empirical results, we have established the consistency and convergence rates under certain smoothness assumptions (e.g., $\log p_0 \in \mathcal{R}(C^\beta)$) for both well-specified and misspecified scenarios.

Three important questions still remain open in this work which we intend to address in our future work. First, the assumption $\log p_0 \in \mathcal{R}(C^\beta)$ is not well understood. Though we presented a necessary condition for this assumption (with $\beta = 1$) to hold for bounded continuous translation invariant kernels on $\mathbb{R}^d$, obtaining a sufficient condition can throw light on the minimax optimality of the proposed estimator. Another alternative is to directly study the minimax optimality of the rates for $0 < \beta \leq 1$ (for $\beta > 1$, we showed that the above mentioned rates can be improved by an appropriate choice of the regularizer) by obtaining minimax lower bounds under the source condition $\log p_0 \in \mathcal{R}(C^\beta)$ and the eigenvalue decay rate of C, using the ideas in DeVore et al. (2004). Second, the proposed estimator depends on the regularization parameter, which in turn depends on the smoothness scale β. Since β is not known in practice, it is therefore of interest to construct estimators that are adaptive to unknown β. Third, since the proposed estimator is computationally expensive as it involves solving a linear system of size $nd \times nd$, it is important to study either alternate estimators or efficient implementations of the proposed estimator to improve the applicability of the method.

8. Proofs

We provide proofs of the results presented in Sections 3–5.

8.1 Proof of Proposition 1

Sriperumbudur et al. (2011, Proposition 5) showed that $\mathcal{H}$ is dense in $C_0(\Omega)$ w.r.t. uniform norm if and only if k satisfies (5). Therefore, the denseness in L^1, KL and Hellinger distances follow trivially from Lemma A.1. For L^r norm ($r > 1$), the denseness follows by using the bound $\|p_f - p_g\|_{L^r(\Omega)} \leq 2e^{2\|f-g\|_\infty} e^{2\|f\|_\infty} \|f - g\|_\infty \|q_0\|_{L^r(\Omega)}$ obtained from Lemma A.1(i) with $f \in C_0(\Omega)$ and $g \in \mathcal{H}$. ■

8.2 Proof of Corollary 2

For any $p \in \mathcal{P}_c$, define $p_\delta := \frac{p+\delta q_0}{1+\delta}$. Note that $p_\delta(x) > 0$ for all $x \in \Omega$ and $\|p - p_\delta\|_{L^r(\Omega)} = \frac{\delta\|p-q_0\|_{L^r(\Omega)}}{1+\delta}$, implying that $\lim_{\delta\to 0} \|p - p_\delta\|_{L^r(\Omega)} = 0$ for any $1 \leq r \leq \infty$. This means, for any $\epsilon > 0$, $\exists \delta_\epsilon > 0$ such that for any $0 < \theta < \delta_\epsilon$, we have $\|p - p_\theta\|_{L^r(\Omega)} \leq \epsilon$, where $p_\theta(x) > 0$ for all $x \in \Omega$.

Define $f := \log \frac{p_\theta}{q_0} - c_\theta$ where $c_\theta := \log \frac{\ell+\theta}{1+\theta}$. It is clear that $f \in C(\Omega)$ since $p, q \in C(\Omega)$. Fix any $\eta > 0$ and define

$$A := \{x : f(x) \geq \eta\} = \left\{x : \frac{p(x)}{q_0(x)} - \ell \geq (\ell+\theta)\left(e^\eta - 1\right)\right\}.$$

Since $\frac{p}{q_0} - \ell \in C_0(\Omega)$, it is clear that A is compact and so $f \in C_0(\Omega)$. Also, it is easy to verify that $p_\theta = e^{f-A(f)}q_0$ which implies $p_\theta \in \mathcal{P}_0$, where $\mathcal{P}_0$ is defined in Proposition 1. This means, for any $\epsilon > 0$, there exists $p_g \in \mathcal{P}$ such that $\|p_\theta - p_g\|_{L^r(\Omega)} \leq \epsilon$ under the assumption that $q_0 \in L^1(\Omega) \cap L^r(\Omega)$. Therefore $\|p - p_g\|_{L^r(\Omega)} \leq 2\epsilon$ for any $1 \leq r \leq \infty$, which proves the denseness of $\mathcal{P}$ in $\mathcal{P}_c$ w.r.t. L^r norm for any $1 \leq r \leq \infty$. Since $h(p,q) \leq \sqrt{\|p-q\|_{L^1(\Omega)}}$ for any probability densities p, q, the denseness in Hellinger distance follows.

We now prove the denseness in KL divergence by noting that

$$\begin{aligned} KL(p\|p_\delta) &= \int_{\{p>0\}} p \log \frac{p+p\delta}{p+q_0\delta}\,dx \leq \int_{\{p>0\}} p\left(\frac{p+p\delta}{p+q_0\delta} - 1\right)dx \\ &= \delta \int_{p>0} (p-q_0)\frac{p}{p+q_0\delta}\,dx \leq \delta\|p-q_0\|_{L^1(\Omega)} \leq 2\delta, \end{aligned}$$

which implies $\lim_{\delta\to 0} KL(p\|p_\delta) = 0$. This implies, for any $\epsilon > 0$, $\exists \delta_\epsilon > 0$ such that for any $0 < \theta < \delta_\epsilon$, $KL(p\|p_\theta) \leq \epsilon$. Arguing as above, we have $p_\theta \in \mathcal{P}_0$, i.e., there exists $f \in C_0(\Omega)$ such that $p_\theta = \frac{e^f q_0}{\int e^f q_0\,dx}$. Since $\mathcal{H}$ is dense in $C_0(\Omega)$, for any $f \in C_0(\Omega)$ and any $\epsilon > 0$, there exists $g \in \mathcal{H}$ such that $\|f-g\|_\infty \leq \epsilon$. For $p_g \in \mathcal{P}$, since $\int p \log \frac{p_\theta}{p_g}\,dx \leq \left\|\log \frac{p_\theta}{p_g}\right\|_\infty \leq 2\|f-g\|_\infty \leq 2\epsilon$, we have

$$KL(p\|p_g) = \int_\Omega p \log \frac{p}{p_g}\,dx = \int_\Omega p \log \frac{p}{p_\theta}\,dx + \int_\Omega p \log \frac{p_\theta}{p_g}\,dx \leq 3\epsilon$$

and the result follows. ■

8.3 Proof of Theorem 4

(i) By the reproducing property of $\mathcal{H}$, since $\partial_i f(x) = \langle f, \partial_i k(x,\cdot)\rangle_{\mathcal{H}}$ for all $i \in [d]$, it is easy to verify that

$$\begin{aligned} J(f) &= \frac{1}{2}\int_\Omega p_0(x) \sum_{i=1}^d \langle f - f_0, \partial_i k(x,\cdot)\rangle_{\mathcal{H}}^2\,dx \\ &= \frac{1}{2}\int_\Omega p_0(x) \sum_{i=1}^d \langle f - f_0, (\partial_i k(x,\cdot) \otimes \partial_i k(x,\cdot))(f-f_0)\rangle_{\mathcal{H}}\,dx \\ &= \frac{1}{2}\int_\Omega p_0(x)\,\langle f - f_0, C_x(f-f_0)\rangle_{\mathcal{H}}\,dx, \end{aligned} \tag{11}$$

where in the second line, we used $\langle a, b\rangle_H^2 = \langle a,b\rangle_H\langle a,b\rangle_H = \langle a, (b\otimes b)a\rangle_H$ for $a, b \in H$ with H being a Hilbert space and

$$C_x := \sum_{i=1}^d \partial_i k(x,\cdot) \otimes \partial_i k(x,\cdot). \tag{12}$$

Observe that for all $x \in \Omega$, C_x is a Hilbert-Schmidt operator as $\|C_x\|_{HS} \le \sum_{i=1}^d \|\partial_i k(x,\cdot)\|^2_{\mathcal{H}} = \sum_{i=1}^d \partial_i\partial_{i+d}k(x,x) < \infty$ and $(f-f_0)\otimes(f-f_0)$ is also Hilbert-Schmidt as $\|(f-f_0)\otimes(f-f_0)\|_{HS} = \|f-f_0\|^2_{\mathcal{H}} < \infty$. Therefore, (11) is equivalent to

$$J(f) = \frac{1}{2}\int_\Omega p_0(x)\,\langle (f-f_0)\otimes(f-f_0), C_x\rangle_{HS}\;dx.$$

Since the first condition in **(D)** implies $\int_\Omega \|C_x\|_{HS} p_0(x)\,dx < \infty$, C_x is p_0-integrable in the Bochner sense (see Diestel and Uhl, 1977, Definition 1 and Theorem 2), and therefore it follows from Diestel and Uhl (1977, Theorem 6) that

$$J(f) = \frac{1}{2}\left\langle (f-f_0)\otimes(f-f_0), \int_\Omega C_x\; p_0(x)\,dx\right\rangle_{HS},$$

where $C := \int_\Omega C_x\; p_0(x)\,dx$ is the Bochner integral of C_x, thereby yielding (6).

We now show that C is trace-class. Let $(e_l)_{l\in\mathbb{N}}$ be an orthonormal basis in $\mathcal{H}$ (a countable ONB exists as $\mathcal{H}$ is separable—see Remark 3(i)). Define $B := \sum_l \langle Ce_l, e_l\rangle_{\mathcal{H}}$ so that

$$\begin{aligned} B &= \sum_l \int_\Omega \langle e_l, C_x e_l\rangle_{\mathcal{H}} p_0(x)\,dx = \sum_l \int_\Omega \sum_{i=1}^d \langle e_l, \partial_i k(x,\cdot)\rangle^2_{\mathcal{H}}\, p_0(x)\,dx \\ &\stackrel{(*)}{=} \int_\Omega \sum_{i\in[d],l} \langle e_l, \partial_i k(x,\cdot)\rangle^2_{\mathcal{H}}\, p_0(x)\,dx \stackrel{(**)}{=} \int_\Omega \sum_{i=1}^d \|\partial_i k(x,\cdot)\|^2_{\mathcal{H}}\, p_0(x)\,dx < \infty, \end{aligned}$$

which means C is trace-class and therefore compact. Here, we used monotone convergence theorem in $(*)$ and Parseval's identity in $(**)$. Note that C is positive since $\langle f, Cf\rangle_{\mathcal{H}} = \int_\Omega p_0(x)\,\|\nabla f\|_2^2\;dx \ge 0$, $\forall\, f \in \mathcal{H}$.

(ii) From (6), we have $J(f) = \frac{1}{2}\langle f, Cf\rangle_{\mathcal{H}} - \langle f, Cf_0\rangle_{\mathcal{H}} + \frac{1}{2}\langle f_0, Cf_0\rangle_{\mathcal{H}}$. Using $\partial_i f_0(x) = \partial_i \log p_0(x) - \partial_i \log q_0(x)$ for all $i \in [d]$, we obtain that for any $f \in \mathcal{H}$,

$$\begin{aligned} \langle f, Cf_0\rangle_{\mathcal{H}} &= \int_\Omega p_0(x)\sum_{i=1}^d \partial_i f(x)\partial_i f_0(x)\,dx \\ &= \int_\Omega \sum_{i=1}^d \partial_i f(x)\partial_i p_0(x)\,dx - \int_\Omega p_0(x)\sum_{i=1}^d \partial_i f(x)\partial_i \log q_0(x)\,dx \\ &\stackrel{(b)}{=} -\int_\Omega p_0(x)\sum_{i=1}^d \partial_i^2 f(x)\,dx - \int_\Omega p_0(x)\sum_{i=1}^d \partial_i f(x)\partial_i \log q_0(x)\,dx \\ &= -\int_\Omega p_0(x)\left\langle f, \overbrace{\sum_{i=1}^d \partial_i^2 k(x,\cdot) + \partial_i k(x,\cdot)\partial_i \log q_0(x)}^{\xi_x}\right\rangle_{\mathcal{H}} dx \stackrel{(c)}{=} \langle f, -\xi\rangle_{\mathcal{H}}, \end{aligned} \quad (13)$$

where (b) follows from integration by parts under **(C)** and the equality in (c) is valid as ξ_x is Bochner p_0-integrable under **(D)** with $\varepsilon = 1$. Therefore $Cf_0 = -\xi$. For the third term, $\langle f_0, Cf_0\rangle_{\mathcal{H}} = \int_\Omega p_0(x)\sum_{i=1}^d (\partial_i f_0(x))^2\;dx$ and the result follows.

(iii) Define $c_0 := J(p_0\|q_0)$. For any $\lambda > 0$, it is easy to verify that

$$J_\lambda(f) = \frac{1}{2}\|(C+\lambda I)^{1/2} f + (C+\lambda I)^{-1/2}\xi\|^2_{\mathcal{H}} - \frac{1}{2}\langle \xi, (C+\lambda I)^{-1}\xi\rangle_{\mathcal{H}} + c_0.$$

Clearly, $J_\lambda(f)$ is minimized if and only if $(C+\lambda I)^{1/2} f = -(C+\lambda I)^{-1/2}\xi$ and therefore $f_\lambda = -(C+\lambda I)^{-1}\xi$ is the unique minimizer of $J_\lambda(f)$.

(iv) Since (iv) is similar to (iii) with C replaced by $\hat{C}$ and ξ replaced by $\hat{\xi}$, we obtain $f_{\lambda,n} = (\hat{C}+\lambda I)^{-1}\hat{\xi}$. ■

8.4 Proof of Theorem 5

We prove the result based on the general representer theorem (Theorem A.2). From Theorem 4(iv), we have

$$\begin{aligned} f_{\lambda,n} &= \arg\inf_{f\in\mathcal{H}} \frac{1}{2}\langle f, \hat{C} f\rangle_{\mathcal{H}} + \langle f, \hat{\xi}\rangle_{\mathcal{H}} + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}} \\ &= \arg\inf_{f\in\mathcal{H}} \frac{1}{2n}\sum_{a=1}^n\sum_{i=1}^d \langle f, \partial_i k(X_a,\cdot)\rangle^2_{\mathcal{H}} + \langle f, \hat{\xi}\rangle_{\mathcal{H}} + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}} \\ &= \arg\inf_{f\in\mathcal{H}} V(\langle f,\phi_1\rangle_{\mathcal{H}},\ldots,\langle f,\phi_{nd}\rangle_{\mathcal{H}},\langle f,\phi_{nd+1}\rangle_{\mathcal{H}}) + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}}, \end{aligned}$$

where $V(\theta_1,\ldots,\theta_{nd},\theta_{nd+1}) := \frac{1}{2n}\sum_{a=1}^n\sum_{i=1}^d \theta^2_{(a-1)d+i} + \theta_{nd+1}$, $\phi_{(a-1)d+i} := \partial_i k(X_a,\cdot)$, $a \in [n]$, $i \in [d]$ and $\phi_{nd+1} := \hat{\xi}$. Therefore, it follows from Theorem A.2 that

$$f_{\lambda,n} = \delta\hat{\xi} + \sum_{a=1}^n\sum_{i=1}^d \beta_{(a-1)d+i}\phi_{(a-1)d+i} \tag{14}$$

where δ and $\boldsymbol{\beta}$ satisfy

$$\lambda\begin{pmatrix}\boldsymbol{\beta}\\ \delta\end{pmatrix} + \nabla V\left(\boldsymbol{K}\begin{pmatrix}\boldsymbol{\beta}\\ \delta\end{pmatrix}\right) = 0 \tag{15}$$

with $\boldsymbol{K} = \begin{pmatrix}\boldsymbol{G} & \boldsymbol{h}\\ \boldsymbol{h}^T & \|\hat{\xi}\|^2_{\mathcal{H}}\end{pmatrix}$. Since $\nabla V\begin{pmatrix}\boldsymbol{z}\\ t\end{pmatrix} = \begin{pmatrix}\frac{1}{n}\boldsymbol{z}\\ 1\end{pmatrix}$, (15) reduces to $\lambda\delta + 1 = 0$ and $\lambda\boldsymbol{\beta} + \frac{1}{n}\boldsymbol{G}\boldsymbol{\beta} + \frac{\delta}{n}\boldsymbol{h} = 0$ yielding $\delta = -\frac{1}{\lambda}$ and $(\frac{1}{n}\boldsymbol{G} + \lambda I)\boldsymbol{\beta} = \frac{1}{n\lambda}\boldsymbol{h}$. ■

Remark Instead of using the general representer theorem (Theorem A.2), it is possible to see that the standard representer theorem (Kimeldorf and Wahba, 1971; Schölkopf et al., 2001) gives a similar, but slightly different linear system, and the solutions are the same if $\boldsymbol{K}$ is non-singular. The general representer theorem yields that $\boldsymbol{\beta}$ and δ are solution to $\boldsymbol{F}\begin{pmatrix}\boldsymbol{\beta}\\ \delta\end{pmatrix} = \begin{pmatrix}\boldsymbol{0}\\ 1\end{pmatrix}$, where $\boldsymbol{F} = \begin{pmatrix}\frac{1}{n}\boldsymbol{G} + \lambda I & \frac{1}{n}\boldsymbol{h}\\ \boldsymbol{0}^T & \lambda\end{pmatrix}$. On the other hand, by using the standard representer theorem, it is easy to show that $f_{\lambda,n}$ has the form in (14) with δ and $\boldsymbol{\beta}$ being solution to $\boldsymbol{K}\boldsymbol{F}\begin{pmatrix}\boldsymbol{\beta}\\ \delta\end{pmatrix} = \boldsymbol{K}\begin{pmatrix}\boldsymbol{0}\\ 1\end{pmatrix}$. Clearly, both the solutions match if $\boldsymbol{K}$ is invertible while the latter has many solutions if $\boldsymbol{K}$ is not invertible.

8.5 Proof of Theorem 6

Consider

$$f_{\lambda,n} - f_\lambda = -(\hat{C}+\lambda I)^{-1}\Big(\hat{\xi} + (\hat{C}+\lambda I)f_\lambda\Big) \overset{(*)}{=} -(\hat{C}+\lambda I)^{-1}\Big(\hat{\xi} + \hat{C}f_\lambda + C(f_0 - f_\lambda)\Big)$$
$$= (\hat{C}+\lambda I)^{-1}(C-\hat{C})(f_\lambda - f_0) - (\hat{C}+\lambda I)^{-1}(\hat{\xi}+\hat{C}f_0)$$
$$= (\hat{C}+\lambda I)^{-1}(C-\hat{C})(f_\lambda - f_0) - (\hat{C}+\lambda I)^{-1}(\hat{\xi}-\xi) + (\hat{C}+\lambda I)^{-1}(C-\hat{C})f_0,$$

where we used $\lambda f_\lambda = C(f_0 - f_\lambda)$ in $(*)$. Define $S_1 := \|(\hat{C}+\lambda I)^{-1}(C-\hat{C})(f_\lambda - f_0)\|_{\mathcal{H}}$, $S_2 := \|(\hat{C}+\lambda I)^{-1}(\hat{\xi}-\xi)\|_{\mathcal{H}}$ and $S_3 := \|(\hat{C}+\lambda I)^{-1}(C-\hat{C})f_0\|_{\mathcal{H}}$ so that

$$\|f_{\lambda,n} - f_0\|_{\mathcal{H}} \le \|f_{\lambda,n} - f_\lambda\|_{\mathcal{H}} + \|f_\lambda - f_0\|_{\mathcal{H}} \le S_1 + S_2 + S_3 + \mathcal{A}_0(\lambda), \tag{16}$$

where $\mathcal{A}_0(\lambda) := \|f_\lambda - f_0\|_{\mathcal{H}}$. We now bound S_1, S_2 and S_3 using Proposition A.4. Note that $C = \int_\Omega C_x\, p_0(x)\, dx$ where C_x is defined in (12) is a positive, self-adjoint, trace-class operator and **(D)** (with $\varepsilon = 2$) implies that

$$\int_\Omega \|C_x\|^2_{HS} p_0(x)\, dx \le \int_\Omega \left(\sum_{i=1}^d \|\partial_i k(x,\cdot)\|^2_{\mathcal{H}}\right)^2 p_0(x)\, dx \le d \sum_{i=1}^d \int_\Omega \|\partial_i k(x,\cdot)\|^4_{\mathcal{H}}\, p_0(x)\, dx < \infty.$$

Therefore, by Proposition A.4(i,iii),

$$S_1 \le \|(\hat{C}+\lambda I)^{-1}\| \|(C-\hat{C})(f_\lambda - f_0)\|_{\mathcal{H}} = O_{p_0}\left(\frac{\mathcal{A}_0(\lambda)}{\lambda\sqrt{n}}\right) \tag{17}$$

and

$$S_2 \le \|(\hat{C}+\lambda I)^{-1}\| \|\hat{\xi}-\xi\|_{\mathcal{H}} = O_{p_0}\left(\frac{1}{\lambda\sqrt{n}}\right), \tag{18}$$

where by using the technique in the proof of Proposition A.4(i), we show below that $\|\hat{\xi} - \xi\|_{\mathcal{H}} = O_{p_0}(n^{-1/2})$. Note that $\mathbb{E}_{p_0}\|\hat{\xi}-\xi\|^2_{\mathcal{H}} = \frac{\int_\Omega \|\xi_x\|^2_{\mathcal{H}} p_0(x)\, dx - \|\xi\|^2_{\mathcal{H}}}{n} \le \frac{\int_\Omega \|\xi_x\|^2_{\mathcal{H}} p_0(x)\, dx}{n}$, where $\xi_x \in \mathcal{H}$ is defined in (13) and **(D)** (with $\varepsilon = 2$) implies that $\int_\Omega \|\xi_x\|^2_{\mathcal{H}} p_0(x)\, dx < \infty$. Therefore $\|\hat{\xi}-\xi\|_{\mathcal{H}} = O_{p_0}(n^{-1/2})$ follows from an application of Chebyshev's inequality. Again using Proposition A.4(i,iii), we obtain that

$$S_3 \le \|(\hat{C}+\lambda I)^{-1}\| \|(C-\hat{C})f_0\|_{\mathcal{H}} = O_{p_0}\left(\frac{1}{\lambda\sqrt{n}}\right). \tag{19}$$

Using the bounds in S_1, S_2 and S_3 in (16), we obtain

$$\|f_{\lambda,n} - f_0\|_{\mathcal{H}} = O_{p_0}\left(\frac{1}{\lambda\sqrt{n}} + \frac{\mathcal{A}_0(\lambda)}{\lambda\sqrt{n}}\right) + \mathcal{A}_0(\lambda). \tag{20}$$

(i) By Proposition A.3(i), we have that $\mathcal{A}_0(\lambda) \to 0$ as $\lambda \to 0$ if $f_0 \in \overline{\mathcal{R}(C)}$. Therefore, it follows from (20) that $\|f_{\lambda,n} - f_0\|_{\mathcal{H}} \to 0$ as $\lambda \to 0$, $\lambda\sqrt{n} \to \infty$ and $n \to \infty$.

(ii) If $f_0 \in \mathcal{R}(C^\beta)$ for $\beta > 0$, it follows from Proposition A.3(ii) that

$$\mathcal{A}_0(\lambda) \le \max\{1, \|C\|^{\beta-1}\}\|C^{-\beta}f_0\|_{\mathcal{H}}\lambda^{\min\{1,\beta\}}$$

and therefore the result follows by choosing $\lambda = n^{-\max\left\{\frac{1}{4}, \frac{1}{2(\beta+1)}\right\}}$.

(iii) Note that

$$S_1 = \|(\hat{C} + \lambda I)^{-1}(C - \hat{C})(f_\lambda - f_0)\|_{\mathcal{H}} \le \|C(\hat{C} + \lambda I)^{-1}\| \|C^{-1}\| \|(C - \hat{C})(f_\lambda - f_0)\|_{\mathcal{H}},$$

$$S_2 = \|(\hat{C} + \lambda I)^{-1}(\hat{\xi} - \xi)\|_{\mathcal{H}} \le \|C(\hat{C} + \lambda I)^{-1}\| \|C^{-1}\| \|\hat{\xi} - \xi\|_{\mathcal{H}},$$

$$S_3 = \|(\hat{C} + \lambda I)^{-1}(C - \hat{C})f_0\|_{\mathcal{H}} \le \|C(\hat{C} + \lambda I)^{-1}\| \|C^{-1}\| \|(C - \hat{C})f_0\|_{\mathcal{H}}$$

and

$$\mathcal{A}_0(\lambda) = \|f_\lambda - f_0\|_{\mathcal{H}} \le \|C^{-1}\| \|C(f_\lambda - f_0)\|_{\mathcal{H}}.$$

It follows from Proposition A.4(v) that $\|C(\hat{C}+\lambda I)^{-1}\| \lesssim 1$ for $n \ge \frac{c}{\lambda^2}$ where c is a sufficiently large constant that depends on $\sum_{i=1}^{d} \int_\Omega (\partial_i \partial_{i+d} k(x,x))^2 p_0(x)\, dx$ but not on n and λ. Using the bounds on $\|(C - \hat{C})(f_\lambda - f_0)\|_{\mathcal{H}}$, $\|\hat{\xi} - \xi\|_{\mathcal{H}}$ and $\|(C - \hat{C})f_0\|_{\mathcal{H}}$ from part (i) and the bound on $\|C(f_\lambda - f_0)\|_{\mathcal{H}}$ from Proposition A.3(ii), we therefore obtain

$$\|f_{\lambda,n} - f_0\|_{\mathcal{H}} \lesssim O_{p_0}\left(\frac{1}{\sqrt{n}}\right) + \lambda \tag{21}$$

as $n \to \infty$ and the result follows. ■

Remark Under slightly strong assumptions on the kernel, the bound on S_1 in (17) can be improved to obtain $S_1 = O_{p_0}(n^{-1/2})$ while the one on S_3 in (19) can be refined to obtain $S_3 = O_{p_0}\left(\sqrt{\frac{\mathcal{N}(\lambda)}{\lambda n}}\right)$ where $\mathcal{N}(\lambda) := Tr((C + \lambda I)^{-1}C)$ is the intrinsic dimension of $\mathcal{H}$. Using the fact that $\mathcal{N}(\lambda) \le \frac{1}{\lambda}$, it is easy to verify that the latter is an improved bound than the one in (19). In addition S_3 dominates S_1. However, if S_2 in (18) is not improved, then S_2 dominates S_3, thereby resulting in a bound that does not capture the smoothness of k (or the corresponding $\mathcal{H}$). Unfortunately, even with a refined analysis (not reported here), we are not able to improve the bound on S_2 wherein the difficulty lies with handling ξ.

8.6 Proof of Theorem 7

Before we prove the result, we present a lemma.

Lemma 14 *Suppose* $\sup_{x\in\Omega} k(x,x) < \infty$ *and* $\text{supp}(q_0) = \Omega$. *Then* $\mathcal{F} = \mathcal{H}$ *and for any* $f_0 \in \mathcal{H}$ *there exists* $\tilde{f}_0 \in \overline{\mathcal{R}(C)}$ *such that* $p_{\tilde{f}_0} = p_0$.

Proof Since $\sup_{x\in\Omega} k(x,x) < \infty$, it implies that, for every $f \in \mathcal{H}$, $\int_\Omega e^{f(x)} q_0(x)\, dx < \infty$ and hence $\mathcal{F} = \mathcal{H}$. Also, under the assumptions on k and q_0, it is easy to verify that $\text{supp}(p_0) = \Omega$, which implies

$$\mathcal{N}(C) = \left\{ f \in \mathcal{H} : \int_\Omega \|\nabla f\|_2^2\, p_0(x)\, dx = 0 \right\}$$

is either $\mathbb{R}$ or $\{0\}$, where $\mathcal{N}(C)$ denotes the null space of C. Let $\tilde{f}_0$ be the orthogonal projection of f_0 onto $\overline{\mathcal{R}(C)} = \mathcal{N}(C)^\perp$. Then $\tilde{f}_0 - f_0 \in \mathbb{R}$ and therefore $p_{\tilde{f}_0} = p_{f_0}$. ■

Proof of Theorem 7. From Theorem 4(iii), $f_\lambda = (C+\lambda I)^{-1}Cf_0 = (C+\lambda I)^{-1}C\tilde{f}_0$ where the second equality follows from the proof of Lemma 14. Now, carrying out the decomposition as in the proof of Theorem 6(i), we obtain $f_{\lambda,n} - f_\lambda = (\hat{C}+\lambda I)^{-1}(C-\hat{C})(f_\lambda - \tilde{f}_0) - (\hat{C}+\lambda I)^{-1}(\hat{\xi}-\xi) + (\hat{C}+\lambda I)^{-1}(C-\hat{C})\tilde{f}_0$ and therefore,

$$\|f_{\lambda,n} - \tilde{f}_0\|_{\mathcal{H}} \le \|(\hat{C}+\lambda I)^{-1}\| \left(\|(C-\hat{C})(f_\lambda - \tilde{f}_0)\|_{\mathcal{H}} + \|\xi - \hat{\xi}\|_{\mathcal{H}} + \|(C-\hat{C})\tilde{f}_0\|_{\mathcal{H}} \right) + \tilde{\mathcal{A}}_0(\lambda),$$

where $\tilde{\mathcal{A}}_0(\lambda) = \|f_\lambda - \tilde{f}_0\|_{\mathcal{H}}$. The bounds on these quantities follow those in the proof of Theorem 6(i) verbatim and so the consistency result in Theorem 6(i) holds for $\|f_{\lambda,n} - \tilde{f}_0\|_{\mathcal{H}}$. By Lemma 14, since $p_{f_0} = p_{\tilde{f}_0}$, it is sufficient to consider the convergence of $p_{f_{\lambda,n}}$ to $p_{\tilde{f}_0}$. Therefore, the convergence (along with rates) in L^r (for any $1 \le r \le \infty$), Hellinger and KL distances follow from using the bound $\|f_{\lambda,n} - \tilde{f}_0\|_\infty \le \sqrt{\|k\|_\infty}\|f_{\lambda,n} - \tilde{f}_0\|_{\mathcal{H}}$ (obtained through the reproducing property of k) in Lemma A.1 and invoking Theorem 6.

In the following, we obtain a bound on $J(p_0\|p_{f_{\lambda,n}}) = \frac{1}{2}\|\sqrt{C}(f_{\lambda,n} - f_0)\|^2_{\mathcal{H}}$. While one can trivially use the bound $\|\sqrt{C}(f_{\lambda,n} - f_0)\|^2_{\mathcal{H}} \le \|\sqrt{C}\|^2\|f_{\lambda,n} - f_0\|^2_{\mathcal{H}}$ to obtain a rate on $J(p_0\|p_{f_{\lambda,n}})$ through the result in Theorem 6(ii), a better rate can be obtained by carefully bounding $\|\sqrt{C}(f_{\lambda,n} - f_0)\|^2_{\mathcal{H}}$ as shown below. Consider

$$\|\sqrt{C}(f_{\lambda,n} - f_0)\|_{\mathcal{H}} \le \|\sqrt{C}(f_{\lambda,n} - f_\lambda)\|_{\mathcal{H}} + \tilde{\mathcal{A}}_{\frac{1}{2}}(\lambda) + \mathcal{A}^\star_{\frac{1}{2}}(\lambda)$$

$$\le \|\sqrt{C}(\hat{C}+\lambda I)^{-1}\| \left(\|(C-\hat{C})(f_\lambda - \tilde{f}_0)\|_{\mathcal{H}} + \|\xi - \hat{\xi}\|_{\mathcal{H}} + \|(C-\hat{C})\tilde{f}_0\|_{\mathcal{H}} \right) + \tilde{\mathcal{A}}_{\frac{1}{2}}(\lambda) + \mathcal{A}^\star_{\frac{1}{2}}(\lambda),$$

where $\tilde{\mathcal{A}}_{\frac{1}{2}}(\lambda) := \|\sqrt{C}(f_\lambda - \tilde{f}_0)\|_{\mathcal{H}}$ and $\mathcal{A}^\star_{\frac{1}{2}}(\lambda) := \|\sqrt{C}(\tilde{f}_0 - f_0)\|_{\mathcal{H}}$. It follows from Theorem 4(i) and Lemma 14 that $\mathcal{A}^\star_{\frac{1}{2}}(\lambda) = J(p_0\|p_{\tilde{f}_0}) = 0$. Also it follows from Proposition A.4(v) that $\|\sqrt{C}(\hat{C}+\lambda I)^{-1}\| \lesssim \frac{1}{\sqrt{\lambda}}$ for $n \ge \frac{c}{\lambda^2}$ where c is a large enough constant that does not depend on n and λ and depends only on $\sum_{i=1}^d \int \|\partial_i k(x,\cdot)\|^4_{\mathcal{H}}\, p_0(x)\, dx$. Using the bounds from the proof of Theorem 6(i) for the rest of the terms within parenthesis, we obtain

$$\|\sqrt{C}(f_{\lambda,n} - f_0)\|_{\mathcal{H}} \le O_{p_0}\left(\frac{1}{\sqrt{\lambda n}}\right) + \tilde{\mathcal{A}}_{\frac{1}{2}}(\lambda). \tag{22}$$

The consistency result therefore follows from Proposition A.3(i) by noting that $\tilde{\mathcal{A}}_{\frac{1}{2}}(\lambda) \to 0$ as $\lambda \to 0$. If $f_0 \in \mathcal{R}(C^\beta)$ for some $\beta \ge 0$, then Proposition A.3(ii) yields $\tilde{\mathcal{A}}_{\frac{1}{2}}(\lambda) \le \max\{1, \|C\|^{\beta-\frac{1}{2}}\}\lambda^{\min\{1,\beta+\frac{1}{2}\}}\|C^{-\beta}f_0\|_{\mathcal{H}}$ which when used in (22) provides the desired rate with $\lambda = n^{-\max\{\frac{1}{3},\frac{1}{2(\beta+1)}\}}$. If $\|C^{-1}\| < \infty$, then the result follows by noting $\|\sqrt{C}(f_{\lambda,n} - f_0)\|_{\mathcal{H}} \le \|\sqrt{C}\|\|f_{\lambda,n} - f_0\|_{\mathcal{H}}$ and invoking the bound in (21). ■

8.7 Proof of Proposition 8

Observation 1: By (Wendland, 2005, Theorem 10.12), we have

$$\mathcal{H} = \left\{ f \in L^2(\mathbb{R}^d) \cap C_b(\mathbb{R}^d) : \frac{f^\wedge}{\sqrt{\psi^\wedge}} \in L^2(\mathbb{R}^d) \right\},$$

where $f^\wedge$ is defined in L^2 sense. Since

$$\int_{\mathbb{R}^d} |f^\wedge(\omega)|\,d\omega \le \left(\int_{\mathbb{R}^d} \frac{|f^\wedge(\omega)|^2}{\psi^\wedge(\omega)}\,d\omega\right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^d} \psi^\wedge(\omega)\,d\omega\right)^{\frac{1}{2}} < \infty$$

where we used $\psi^\wedge \in L^1(\mathbb{R}^d)$ (see Wendland, 2005, Corollary 6.12), we have $f^\wedge \in L^1(\mathbb{R}^d)$. Hence Plancherel's theorem and continuity of f along with the inverse Fourier transform of $f^\wedge$ allow to recover any $f \in \mathcal{H}$ pointwise from its Fourier transform as

$$f(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{ix^T\omega} f^\wedge(\omega)\,d\omega,\; x \in \mathbb{R}^d. \tag{23}$$

Observation 2: Since $\psi^\wedge \in L^1(\mathbb{R}^d)$ and $\psi^\wedge \ge 0$, we have for all $j = 1,\ldots,d$,

$$\begin{aligned}\left(\int_{\mathbb{R}^d} |\omega_j|\psi^\wedge(\omega)\,d\omega\right)^2 &= \left(\int_{\mathbb{R}^d} \psi^\wedge(\omega)\,d\omega\right)^2 \left(\int_{\mathbb{R}^d} |\omega_j| \frac{\psi^\wedge(\omega)}{\int_{\mathbb{R}^d}\psi^\wedge(\omega)\,d\omega}\,d\omega\right)^2 \\ &\overset{(*)}{\le} \left(\int_{\mathbb{R}^d} \psi^\wedge(\omega)\,d\omega\right) \left(\int_{\mathbb{R}^d} |\omega_j|^2\psi^\wedge(\omega)\,d\omega\right) \\ &\le \left(\int_{\mathbb{R}^d} \psi^\wedge(\omega)\,d\omega\right) \left(\int_{\mathbb{R}^d} \|\omega\|^2\psi^\wedge(\omega)\,d\omega\right) \overset{(i)}{<} \infty,\end{aligned}$$

where we used Jensen's inequality in $(*)$. This means $\omega_j\psi^\wedge(\omega) \in L^1(\mathbb{R}^d)$, $\forall\, j \in [d]$ which ensures the existence of its Fourier transform and so

$$\partial_j\psi(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} (i\omega_j)\psi^\wedge(\omega)e^{ix^T\omega}\,d\omega,\; x \in \mathbb{R}^d,\; \forall\, j \in [d]. \tag{24}$$

Observation 3: For $g \in \mathcal{H}$, we have for all $j \in [d]$,

$$\begin{aligned}\int_{\mathbb{R}^d} |\omega_j||g^\wedge(\omega)|\,d\omega &\le \left(\int_{\mathbb{R}^d} \frac{|g^\wedge(\omega)|^2}{\psi^\wedge(\omega)}\,d\omega\right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^d} |\omega_j|^2\psi^\wedge(\omega)\,d\omega\right)^{\frac{1}{2}} \\ &\le \left(\int_{\mathbb{R}^d} \frac{|g^\wedge(\omega)|^2}{\psi^\wedge(\omega)}\,d\omega\right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^d} \|\omega\|^2\psi^\wedge(\omega)\,d\omega\right)^{\frac{1}{2}} \overset{(i)}{<} \infty,\end{aligned}$$

which implies $\omega_j g^\wedge(\omega) \in L^1(\mathbb{R}^d)$, $\forall\, j = 1,\ldots,d$. Therefore,

$$\partial_j g(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} (i\omega_j)g^\wedge(\omega)e^{ix^T\omega}\,d\omega,\; x \in \mathbb{R}^d,\; \forall\, j \in [d]. \tag{25}$$

Observation 4: For any $g \in \mathcal{G}$, we have

$$\int_{\mathbb{R}^d} \frac{|g^\wedge(\omega)|^2}{\psi^\wedge(\omega)}\,d\omega = \int_{\mathbb{R}^d} \frac{|g^\wedge(\omega)|^2}{\phi^\wedge(\omega)}\frac{\phi^\wedge(\omega)}{\psi^\wedge(\omega)}\,d\omega \le \|g\|_{\mathcal{G}}^2 \left\|\frac{\phi^\wedge}{\psi^\wedge}\right\|_\infty \overset{(ii)}{<} \infty,$$

which implies $g \in \mathcal{H}$, i.e., $\mathcal{G} \subset \mathcal{H}$.

We now use these observations to prove the result. Since $f_0 \in \mathcal{R}(C)$, there exists $g \in \mathcal{H}$ such that $f_0 = Cg$, which means

$$f_0(y) \;=\; \int_{\mathbb{R}^d} \sum_{j=1}^d \partial_j k(x,y)\,\partial_j\, p_0(x)\,dx$$

$$\overset{(24)}{=} \int_{\mathbb{R}^d} \sum_{j=1}^d \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i(x-y)^T\omega}(i\omega_j)\psi^\wedge(\omega)\, d\omega\, \partial_j g(x)\, p_0(x)\, dx$$

$$\overset{(\dagger)}{=} \int_{\mathbb{R}^d} \sum_{j=1}^d \left(\frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{ix^T\omega}\partial_j g(x)\, p_0(x)\, dx \right) (i\omega_j)\psi^\wedge(\omega)e^{-iy^T\omega}\, d\omega$$

$$\overset{(25)}{=} \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \sum_{j=1}^d \overline{\left(i(\cdot)_j g^\wedge * p_0^\wedge\right)(\omega)}(i\omega_j)\psi^\wedge(\omega)e^{-iy^T\omega}\, d\omega$$

which from (23) means $f_0^\wedge(\omega) = \sum_{j=1}^d \overline{\left(i(\cdot)_j g^\wedge * p_0^\wedge\right)(-\omega)}(-i\omega_j)\psi^\wedge(\omega)$ where we have invoked Fubini's theorem in (†) and $*$ represents the convolution. Define $\|\cdot\|_{L^r(\mathbb{R}^d)} := \|\cdot\|_r$ and $\theta := \frac{r}{r-1}$. Consider

$$\|f_0\|_{\mathcal{G}}^2 = \int_{\mathbb{R}^d} \frac{|f_0^\wedge(\omega)|^2}{\phi^\wedge(\omega)}\, d\omega = \int_{\mathbb{R}^d} \left| \sum_{j=1}^d \overline{\left(i(\cdot)_j g^\wedge * p_0^\wedge\right)(-\omega)}(i\omega_j) \right|^2 (\psi^\wedge)^2(\omega)(\phi^\wedge(\omega))^{-1}\, d\omega$$

$$\leq \int_{\mathbb{R}^d} \left(\sum_{j=1}^d \left| i(\cdot)_j g^\wedge * p_0^\wedge \right|(-\omega)|\omega_j| \right)^2 (\psi^\wedge)^2(\omega)(\phi^\wedge(\omega))^{-1}\, d\omega$$

$$\leq \int_{\mathbb{R}^d} \sum_{j=1}^d \left| i(\cdot)_j g^\wedge * p_0^\wedge \right|^2 (-\omega)\|\omega\|^2(\psi^\wedge)^2(\omega)(\phi^\wedge(\omega))^{-1}\, d\omega$$

$$\leq \left\| \sum_{j=1}^d \left| i\omega_j g^\wedge(\omega) * p_0^\wedge(\omega) \right|^2 (\cdot) \right\|_{\frac{\theta}{2}} \left\| \|\cdot\|^2(\psi^\wedge)^2(\cdot)(\phi^\wedge(\cdot))^{-1} \right\|_{\frac{r}{2-r}} \overset{(iii)}{<} \infty,$$

where in the following we show that $\sum_{j=1}^d \left| i\omega_j g^\wedge(\omega) * p_0^\wedge(\omega) \right|^2 (\cdot) \in L^{\frac{\theta}{2}}(\mathbb{R}^d)$, i.e.,

$$\left\| \sum_{j=1}^d \left| i\omega_j g^\wedge(\omega) * p_0^\wedge(\omega) \right|^2 (\cdot) \right\|_{\frac{\theta}{2}} \leq \sum_{j=1}^d \left\| \left| i\omega_j g^\wedge(\omega) * p_0^\wedge(\omega) \right|^2 (\cdot) \right\|_{\frac{\theta}{2}} = \sum_{j=1}^d \left\| i\omega_j g^\wedge(\omega) * p_0^\wedge(\omega) \right\|_\theta^2$$

$$\overset{(*)}{\leq} \sum_{j=1}^d \|i\omega_j g^\wedge(\omega)\|_1^2 \|p_0^\wedge\|_\theta^2 \overset{(**)}{\leq} \|p_0\|_r^2 \sum_{j=1}^d \|i\omega_j g^\wedge(\omega)\|_1^2 \overset{(\ddagger)}{<} \infty,$$

where we have invoked generalized Young's inequality (Folland, 1999, Proposition 8.9) in $(*)$, Hausdorff-Young inequality (Folland, 1999, p. 253) in $(**)$, and observation 3 combined with (iv) in (‡). This shows that $f_0 \in \mathcal{R}(C) \Rightarrow f_0 \in \mathcal{G}$, i.e., $\mathcal{R}(C) \subset \mathcal{G}$. ■

8.8 Proof of Theorem 9

To prove Theorem 9, we need the following lemma (De Vito et al., 2012, Lemma 5), which is due to Andreas Maurer.

Lemma 15 *Suppose A and B are self-adjoint Hilbert-Schmidt operators on a separable Hilbert space H with spectrum contained in the interval $[a, b]$, and let $(\sigma_i)_{i\in I}$ and $(\tau_j)_{j\in J}$ be* *the eigenvalues of A and B, respectively. Given a function $r : [a,b] \to \mathbb{R}$, if there exists a finite constant L such that $|r(\sigma_i)-r(\tau_j)| \le L|\sigma_i-\tau_j|$, $\forall\, i \in I$, $j \in J$, then $\|r(A)-r(B)\|_{HS} \le L\|A-B\|_{HS}$.*

Proof of Theorem 9. (i) The proof follows the ideas in the proof of Theorem 10 in Bauer et al. (2007), which is a more general result dealing with the smoothness condition, $f_0 \in \mathcal{R}(\Theta(C))$ where Θ is operator monotone. Recall that Θ is operator monotone on $[0,b]$ if for any pair of self-adjoint operators U, V with spectra in $[0,b]$ such that $U \le V$, we have $\Theta(U) \le \Theta(V)$, where "$\le$" is the partial ordering for self-adjoint operators on some Hilbert space H, which means for any $f \in H$, $\langle f, Uf\rangle_H \le \langle f, Vf\rangle_H$. In our case, we adapt the proof for $\Theta(C) = C^\beta$. Define $r_\lambda(\alpha) := g_\lambda(\alpha)\alpha - 1$. Since $f_0 \in \mathcal{R}(C^\beta)$, there exists $h \in \mathcal{H}$ such that $f_0 = C^\beta h$, which yields

$$\begin{aligned} f_{g,\lambda,n} - f_0 &= -g_\lambda(\hat{C})\hat{\xi} - f_0 = -g_\lambda(\hat{C})(\hat{\xi} + \hat{C} f_0) + r_\lambda(\hat{C})C^\beta h \\ &= -g_\lambda(\hat{C})(\hat{\xi}-\xi) + g_\lambda(\hat{C})(C-\hat{C})f_0 + r_\lambda(\hat{C})\hat{C}^\beta h + r_\lambda(\hat{C})(C^\beta - \hat{C}^\beta)h. \end{aligned} \tag{26}$$

so that

$$\begin{aligned} \|f_{g,\lambda,n} - f_0\|_{\mathcal{H}} \le \underbrace{\|g_\lambda(\hat{C})(\hat{\xi}-\xi)\|_{\mathcal{H}}}_{(A)} + \underbrace{\|g_\lambda(\hat{C})(\hat{C}-C)f_0\|_{\mathcal{H}}}_{(B)} + \underbrace{\|r_\lambda(\hat{C})C^\beta h\|_{\mathcal{H}}}_{(C)} \\ + \underbrace{\|r_\lambda(\hat{C})(C^\beta - \hat{C}^\beta)h\|_{\mathcal{H}}}_{(D)}. \end{aligned}$$

We now bound (A)–(D). Since $(A) \le \|g_\lambda(\hat{C})\|\|\hat{\xi}-\xi\|_{\mathcal{H}}$, we have $(A) = O_{p_0}\left(\frac{1}{\lambda\sqrt{n}}\right)$ where we used (b) in **(E)** and the bound on $\|\hat{\xi}-\xi\|_{\mathcal{H}}$ from the proof of Theorem 6(i). Similarly, $(B) \le \|g_\lambda(\hat{C})\|\|(\hat{C}-C)f_0\|_{\mathcal{H}}$ implies $(B) = O_{p_0}\left(\frac{1}{\lambda\sqrt{n}}\right)$ where (b) in **(E)** and Proposition A.4(i) are invoked. Also, (d) in **(E)** implies that

$$(C) \le \|r_\lambda(\hat{C})\hat{C}^\beta\|\|h\|_{\mathcal{H}} \le \max\{\gamma_\beta, \gamma_{\eta_0}\}\lambda^{\min\{\beta,\eta_0\}}\|C^{-\beta} f_0\|_{\mathcal{H}}.$$

(D) can be bounded as

$$(D) \le \|r_\lambda(\hat{C})\|\|C^\beta - \hat{C}^\beta\|\|C^{-\beta} f_0\|_{\mathcal{H}}.$$

We now consider two cases:

$\underline{\beta \le 1}$: Since $\alpha \mapsto \alpha^\theta$ is operator monotone on $[0,\chi]$ for $0 \le \theta \le 1$, by Theorem 1 in Bauer et al. (2007), there exists a constant c_θ such that $\|\hat{C}^\theta - C^\theta\| \le c_\theta\|\hat{C}-C\|^\theta \le c_\theta\|\hat{C}-C\|^\theta_{HS}$. We now obtain a bound on $\|\hat{C}-C\|_{HS}$. To this end, consider

$$\begin{aligned} \mathbb{E}\|\hat{C}-C\|^2_{HS} &= \mathbb{E}\|\hat{C}\|^2_{HS} - \|C\|^2_{HS} \\ &\le \frac{1}{n}\int \left\|\sum_{i=1}^d \partial_i k(x,\cdot)\otimes\partial_i k(x,\cdot)\right\|^2_{HS} p_0(x)\,dx \le \frac{d}{n}\sum_{i=1}^d \int \|\partial_i k(x,\cdot)\|^4\, p_0(x)\,dx, \end{aligned}$$

which by Chebyshev's inequality implies that

$$\|\hat{C}-C\|_{HS} = O_{p_0}(n^{-1/2})$$

and therefore $(D) = O_{p_0}(n^{-\beta/2})$. Since $\lambda \geq n^{-1/2}$, we have $(D) = O_{p_0}(\lambda^\beta)$.
$\underline{\beta > 1}$: Since $\alpha \mapsto \alpha^\theta$ is Lipschitz on $[0, \chi]$ for $\theta \geq 1$, by Lemma 15, $\|C^\beta - \hat{C}^\beta\| \leq \|C^\beta - \hat{C}^\beta\|_{HS} \leq \beta\chi^{\beta-1}\|C - \hat{C}\|_{HS}$ and therefore $(C) = O_{p_0}(n^{-1/2})$.

Collecting all the above bounds, we obtain

$$\|f_{g,\lambda,n} - f_0\|_{\mathcal{H}} \leq O_{p_0}\left(\frac{1}{\lambda\sqrt{n}}\right) + O_{p_0}\left(\lambda^{\min\{\beta,\eta_0\}}\right)$$

and the result follows. The proofs of the claims involving L^r, h and KL follow exactly the same ideas as in the proof of Theorem 7 by using the above bound on $\|f_{g,\lambda,n} - f_0\|_{\mathcal{H}}$ in Lemma A.1.

(ii) We now bound $J(p_0\|p_{f_{g,\lambda,n}}) = \|\sqrt{C}(f_{g,\lambda,n} - f_0)\|^2_{\mathcal{H}}$ as follows. Note that

$$\sqrt{C}(f_{g,\lambda,n} - f_0) = \underbrace{(\sqrt{C} - \sqrt{\hat{C}})(f_{g,\lambda,n} - f_0)}_{(I')} + \underbrace{\sqrt{\hat{C}}(f_{g,\lambda,n} - f_0)}_{(II')}.$$

We bound $\|(I')\|_{\mathcal{H}}$ as

$$\begin{aligned}\|(I')\|_{\mathcal{H}} \leq \|\sqrt{C} - \sqrt{\hat{C}}\|\|f_{g,\lambda,n} - f_0\|_{\mathcal{H}} &\leq c_{\frac{1}{2}}\sqrt{\|C - \hat{C}\|}_{HS}\|f_{g,\lambda,n} - f_0\|_{\mathcal{H}} \\ &= O_{p_0}\left(\frac{1}{\sqrt{\lambda n}}\right) + O_{p_0}\left(\lambda^{\min\{\beta,\eta_0\}+\frac{1}{2}}\right),\end{aligned}$$

where we used the fact that $\alpha \mapsto \sqrt{\alpha}$ is operator monotone along with $\lambda \geq n^{-1/2}$. Using (26), $\|(II')\|_{\mathcal{H}}$ can be bounded as

$$\begin{aligned}\|(II')\|_{\mathcal{H}} \leq \|\sqrt{\hat{C}}g_\lambda(\hat{C})\|\|\hat{\xi} + \hat{C}f_0\|_{\mathcal{H}} &+ \|\sqrt{\hat{C}}r_\lambda(\hat{C})\hat{C}^\beta\|\|C^{-\beta}f_0\|_{\mathcal{H}} \\ &+ \|\sqrt{\hat{C}}r_\lambda(\hat{C})\|\|C^\beta - \hat{C}^\beta\|\|C^{-\beta}f_0\|_{\mathcal{H}}\end{aligned}$$

where

$$\|\sqrt{\hat{C}}g_\lambda(\hat{C})\| \leq \sqrt{\frac{A_gB_g}{\lambda}},\ \ \|\sqrt{\hat{C}}r_\lambda(\hat{C})\hat{C}^\beta\| \leq (\gamma_{\beta+\frac{1}{2}} \vee \gamma_{\eta_0})\lambda^{\min\{\beta+\frac{1}{2},\eta_0\}}$$

and

$$\|\sqrt{\hat{C}}r_\lambda(\hat{C})\| \leq (\gamma_{\frac{1}{2}} \vee \gamma_{\eta_0})\lambda^{\min\{\frac{1}{2},\eta_0\}}$$

with $\|\hat{C}f_0 + \hat{\xi}\|$ and $\|C^\beta - \hat{C}^\beta\|$ bounded as in part (i) above. Here $(a \vee b) := \max\{a, b\}$. Combining $\|(I')\|_{\mathcal{H}}$ and $\|(II')\|_{\mathcal{H}}$, we obtain the required result.

(iii) The proof follows the ideas in the proof of Theorems 6 and 7. Consider $f_{g,\lambda,n} - f_0 = -g_\lambda(\hat{C})(\hat{C}f_0 + \hat{\xi}) + r_\lambda(\hat{C})f_0$ so that

$$\begin{aligned}\|f_{g,\lambda,n} - f_0\|_{\mathcal{H}} &\leq \|C^{-1}\|\|Cg_\lambda(\hat{C})(\hat{C}f_0 + \hat{\xi})\|_{\mathcal{H}} + \|C^{-1}\|\|Cr_\lambda(\hat{C})f_0\|_{\mathcal{H}} \\ &\leq \|C^{-1}\|\|\hat{C}f_0 + \hat{\xi}\|_{\mathcal{H}}\left(\|\hat{C}g_\lambda(\hat{C})\| + \|\hat{C} - C\|\|g_\lambda(\hat{C})\|\right) \\ &\quad + \|C^{-1}\|\|f_0\|_{\mathcal{H}}\left(\|\hat{C}r_\lambda(\hat{C})\| + \|\hat{C} - C\|\|r_\lambda(\hat{C})\|\right).\end{aligned}$$

Therefore $\|f_{g,\lambda,n} - f_0\|_{\mathcal{H}} = O_{p_0}(n^{-1/2}) + O\left(\lambda^{\min\{1,\eta_0\}}\right)$ where we used the fact that $\lambda \geq n^{-1/2}$ and the result follows. ■

8.9 Proof of Theorem 10

Before we analyze $J(p_0\|p_{f_{\lambda,n}})$, we need a small calculation for notational convenience. For any probability densities $p, q \in C^1$, it is clear that $\sqrt{2J(p\|q)} = \|\|\nabla \log p - \nabla \log q\|_2\|_{L^2(p)}$. We generalize this by defining

$$\sqrt{2J(p\|q\|\mu)} := \|\|\nabla \log p - \nabla \log q\|_2\|_{L^2(\mu)}\,.$$

Clearly, if $\mu = p$, then $J(p\|q\|\mu)$ matches with $J(p\|q)$. Therefore, for probability densities $p, q, r \in C^1$,

$$\sqrt{J(p\|r\|p)} \le \sqrt{J(p\|q\|p)} + \sqrt{J(q\|r\|p)}. \tag{27}$$

Based on (27), we have

$$\sqrt{\inf_{p\in\mathcal{P}} J(p_0\|p)} \le \sqrt{J(p_0\|p_{f_{\lambda,n}}\|p_0)} \le \sqrt{J(p_0\|p_{f^*}\|p_0)} + \sqrt{J(p_{f^*}\|p_{f_{\lambda,n}}\|p_0)}$$

$$\begin{aligned}
&= \sqrt{\inf_{p\in\mathcal{P}} J(p_0\|p\|p_0)} + \sqrt{J(p_{f^*}\|p_{f_{\lambda,n}}\|p_0)} \\
&= \sqrt{\inf_{p\in\mathcal{P}} J(p_0\|p)} + \frac{1}{\sqrt{2}}\sqrt{\langle f_{\lambda,n} - f^*, C(f_{\lambda,n} - f^*)\rangle_{\mathcal{H}}} \\
&= \sqrt{\inf_{p\in\mathcal{P}} J(p_0\|p)} + \frac{1}{\sqrt{2}}\|\sqrt{C}(f_{\lambda,n} - f^*)\|_{\mathcal{H}} \qquad (28)\\
&= \sqrt{\inf_{p\in\mathcal{P}} J(p_0\|p)} + \frac{1}{\sqrt{2}}\|\sqrt{C}(f_{\lambda,n} - f_\lambda)\|_{\mathcal{H}} + \frac{1}{\sqrt{2}}\mathcal{A}^*(\lambda),
\end{aligned}$$

where $\mathcal{A}^*(\lambda) = \|\sqrt{C}(f_\lambda - f^*)\|_{\mathcal{H}}$. The result simply follows from the proof of Theorem 7, where we showed that $\|\sqrt{C}(f_{\lambda,n} - f_\lambda)\|_{\mathcal{H}} = O_{p_0}\left(\frac{1}{\sqrt{\lambda n}}\right)$ and $\mathcal{A}^*(\lambda) = O(\lambda^{\min\{1,\beta+\frac{1}{2}\}})$ if $f^* \in \mathcal{R}(C^\beta)$ for $\beta \ge 0$ as $\lambda \to 0$, $n \to \infty$. When $\|C^{-1}\| < \infty$, we bound $\|\sqrt{C}(f_{\lambda,n} - f^*)\|_{\mathcal{H}}$ in (28) as $\|\sqrt{C}\|\|f_{\lambda,n} - f^*\|_{\mathcal{H}}$ where $\|f_{\lambda,n} - f^*\|_{\mathcal{H}}$ is in turn bounded as in (21). ■

8.10 Proof of Proposition 11

For $f \in \mathcal{H}$, we have

$$\|f\|^2_{\mathcal{W}_2} = \int_\Omega \sum_{i=1}^d (\partial_i f)^2\, p_0(x)\, dx \le \|f\|^2_{\mathcal{H}} \int_\Omega \sum_{i=1}^d \|\partial_i k(x,\cdot)\|^2_{\mathcal{H}}\, p_0(x)\, dx < \infty,$$

which means $f \in \mathcal{W}_2(\Omega, p_0)$ and therefore $[f]_\sim \in W_2(\Omega, p_0)$. Since $\|I_k f\|_{W_2} = \|[f]_\sim\|_{\mathcal{W}_2^\sim} = \|f\|_{\mathcal{W}_2} \le c\|f\|_{\mathcal{H}} < \infty$ where c is some constant, it is clear that I_k is a continuous map from $\mathcal{H}$ to $W_2(\Omega, p_0)$. The adjoint $S_k : W_2(\Omega, p_0) \to \mathcal{H}$ of $I_k : \mathcal{H} \to W_2(\Omega, p_0)$ is defined by the relation $\langle S_k f, g\rangle_{\mathcal{H}} = \langle f, I_k g\rangle_{W_2}$, $f \in W_2(\Omega, p_0)$, $g \in \mathcal{H}$. If $f := [h]_\sim \in \mathcal{W}_2^\sim(\Omega, p_0)$, then

$$\langle [h]_\sim, I_k g\rangle_{W_2} = \langle [h]_\sim, [g]_\sim\rangle_{\mathcal{W}_2^\sim} = \sum_{|\alpha|=1} \int_\Omega (\partial^\alpha h)(x)(\partial^\alpha g)(x)\, p_0(x)\, dx.$$

For $y \in \Omega$ and $g = k(\cdot, y)$, this yields

$$S_k[h]_\sim(y) = \langle S_k[h]_\sim, k(\cdot, y)\rangle_{\mathcal{H}} = \langle [h]_\sim, I_k k(\cdot, y)\rangle_{W_2} = \int_\Omega \sum_{i=1}^d \partial_i k(x, y)\partial_i h(x) p_0(x)\, dx.$$

We now show that I_k is Hilbert-Schmidt. Since $\mathcal{H}$ is separable, let $(e_l)_{l\geq 1}$ be an ONB of $\mathcal{H}$. Then we have

$$\begin{aligned}\sum_l \|I_k e_l\|^2_{W_2} = \sum_l \int_\Omega \sum_{i=1}^d (\partial_i e_l(x))^2\, p_0(x)\, dx &= \int_\Omega \sum_{i=1}^d \sum_l \langle e_l, \partial_i k(x, \cdot)\rangle^2_{\mathcal{H}}\, p_0(x)\, dx \\ &= \int_\Omega \sum_{i=1}^d \|\partial_i k(x, \cdot)\|^2_{\mathcal{H}}\, p_0(x)\, dx < \infty,\end{aligned}$$

which proves that I_k is Hilbert-Schmidt (hence compact) and therefore S_k is also Hilbert-Schmidt and compact. The other assertions about $S_k I_k$ and $I_k S_k$ are straightforward. ■

8.11 Proof of Theorem 12

By slight abuse of notation, $f_\star$ is used to denote $[f_\star]_\sim$ in the proof for simplicity. For $f \in \mathcal{F}$, we have

$$J(p_0\|p_f) = \frac{1}{2}\|I_k f - f_\star\|^2_{W_2} = \frac{1}{2}\langle E_k f, f\rangle_{\mathcal{H}} - \langle S_k f_\star, f\rangle_{\mathcal{H}} + \frac{1}{2}\|f_\star\|^2_{W_2}.$$

Since k satisfies **(C)** it is easy to verify that $\langle S_k f_\star, f\rangle_{\mathcal{H}} = \langle f, -\xi\rangle_{\mathcal{H}}$, $\forall\, f \in \mathcal{H}$ (see proof of Theorem 4(ii)). This implies $S_k f_\star = -\xi$ and

$$J(p_0\|p_f) = \frac{1}{2}\langle E_k f, f\rangle_{\mathcal{H}} + \langle f, \xi\rangle_{\mathcal{H}} + \frac{1}{2}\|f_\star\|^2_{W_2}, \tag{29}$$

where ξ is defined in Theorem 4(ii), and E_k is precisely the operator C defined in Theorem 4(ii). Following the proof of Theorem 4(ii), for $\lambda > 0$, it is easy to show that the unique minimizer of the regularized objective, $J(p_0\|p_f) + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}}$ exists and is given by

$$f_\lambda = -(E_k + \lambda I)^{-1}\xi = (E_k + \lambda I)^{-1} S_k f_\star. \tag{30}$$

We would like to reiterate that (29) and (30) also match with their counterparts in Theorem 4 and therefore as in Theorem 4(iv), an estimator of $f_\star$ is given by $f_{\lambda,n} = -(\hat{E}_k+\lambda I)^{-1}\hat{\xi}$. In other words, this is the same as in Theorem 4(iv) since $\hat{E}_k = \hat{C}$, and can be solved by a simple linear system provided in Theorem 5. Here $\hat{E}_k$ is the empirical estimator of E_k. Now consider

$$\begin{aligned}\sqrt{2\, J(p_0\|p_{f_{\lambda,n}})} = \|I_k f_{\lambda,n} - f_\star\|_{W_2} &\leq \|I_k(f_{\lambda,n} - f_\lambda)\|_{W_2} + \|I_k f_\lambda - f_\star\|_{W_2} \\ &= \|\sqrt{E_k}(f_{\lambda,n} - f_\lambda)\|_{\mathcal{H}} + \mathcal{B}(\lambda),\end{aligned} \tag{31}$$

where $\mathcal{B}(\lambda) := \|I_k f_\lambda - f_\star\|_{W_2}$. The proof now proceeds using the following decomposition, equivalent to the one used in the proof of Theorem 6(i), i.e.,

$$f_{\lambda,n} - f_\lambda = -(\hat{E}_k + \lambda I)^{-1}\hat{\xi} - f_\lambda$$

$$= -(\hat{E}_k + \lambda I)^{-1}(\hat{\xi} + \hat{E}_k f_\lambda + \lambda f_\lambda)$$
$$\overset{(\dagger)}{=} -(\hat{E}_k + \lambda I)^{-1}(\hat{\xi} + \hat{E}_k f_\lambda + S_k f_\star - E_k f_\lambda - \hat{S}_k f_\star + \hat{S}_k f_\star),$$

where we used (30) in (†). $\hat{S}_k f_\star$ is well-defined as it is the empirical version of the restriction of S_k to $\mathcal{W}_2^\sim(p_0)$. Since $S_k f_\star - E_k f_\lambda = S_k(f_\star - I_k f_\lambda)$ and $\hat{S}_k f_\star - \hat{E}_k f_\lambda = \hat{S}_k(f_\star - I_k f_\lambda)$, we have

$$f_{\lambda,n} - f_\lambda = -(\hat{E}_k + \lambda I)^{-1}(\hat{\xi} + \hat{S}_k f_\star) + (\hat{E}_k + \lambda I)^{-1}(\hat{S}_k - S_k)(f_\star - I_k f_\lambda)$$

and so

$$\|\sqrt{E_k}(f_{\lambda,n} - f_\lambda)\|_{\mathcal{H}} \leq \|\sqrt{E_k}(\hat{E}_k + \lambda I)^{-1}\| \left(\|\hat{\xi} + \hat{S}_k f_\star\|_{\mathcal{H}} + \|(\hat{S}_k - S_k)(f_\star - I_k f_\lambda)\|_{\mathcal{H}} \right). \quad (32)$$

It follows from Proposition A.4(v) that

$$\|\sqrt{E_k}(\hat{E}_k + \lambda I)^{-1}\| \lesssim \frac{1}{\sqrt{\lambda}} \quad (33)$$

for $n \geq \frac{c}{\lambda^2}$ where c is a sufficiently large constant that does not depend on n and λ. Following the proof of Proposition A.4(i), we have

$$\mathbb{E}\|\hat{\xi} + \hat{S}_k f_\star\|_{\mathcal{H}}^2 = \frac{n-1}{n}\|\xi + S_k f_\star\|_{\mathcal{H}}^2 + \frac{1}{n}\int_\Omega \left\| \sum_{i=1}^d \partial_i k(x,\cdot)\partial_i f_\star + \xi_x \right\|_{\mathcal{H}}^2 p_0(x)\,dx$$

wherein the first term is zero as $S_k f_\star + \xi = 0$ and since

$$\left\| \sum_{i=1}^d \partial_i k(x,\cdot)\partial_i f_\star + \xi_x \right\|_{\mathcal{H}}^2 \leq 2\|\xi_x\|_{\mathcal{H}}^2 + 2\chi\|\nabla f_\star\|_2^2,$$

the integral in the second term is finite because of **(D)** and $f^* \in W_2(\Omega, p_0)$. Therefore, an application of Chebyshev's inequality yields

$$\|\hat{\xi} + \hat{S}_k f_\star\|_{\mathcal{H}} = O_{p_0}(n^{-1/2}). \quad (34)$$

We now show that $\|(\hat{S}_k - S_k)(f_\star - I_k f_\lambda)\|_{\mathcal{H}} = O_{p_0}(\mathcal{B}(\lambda)n^{-1/2})$. To this end, define $g := f_\star - I_k f_\lambda$ and consider

$$\mathbb{E}_{p_0}\|\hat{S}_k g - S_k g\|_{\mathcal{H}}^2 = \frac{\int_\Omega \|\sum_{i=1}^d \partial_i k(x,\cdot)\partial_i g(x)\|_{\mathcal{H}}^2 p_0(x)\,dx - \|S_k g\|_{\mathcal{H}}^2}{n} \leq \frac{\chi}{n}\|g\|_{W_2}^2,$$

which therefore yields the claim through an application of Chebyshev's inequality. Using this along with (33) and (34) in (32), and using the resulting bound in (31) yields

$$\sqrt{2\,J(p_0\|p_{f_{\lambda,n}})} \leq O_{p_0}\left(\frac{1}{\sqrt{\lambda n}} + \frac{\mathcal{B}(\lambda)}{\sqrt{\lambda n}}\right) + \mathcal{B}(\lambda). \quad (35)$$

(i) We bound $\mathcal{B}(\lambda)$ as follows. First note that

$$\mathcal{B}(\lambda) = \|I_k(S_k I_k + \lambda I)^{-1} S_k f_\star - f_\star\|_{W_2} = \|(T_k + \lambda I)^{-1} T_k f_\star - f_\star\|_{W_2}$$

and so for any $h \in \mathcal{H}$, we have

$$\begin{aligned}\mathcal{B}(\lambda) &= \|(T_k+\lambda I)^{-1}T_k f_\star - f_\star\|_{W_2} \\ &\le \underbrace{\|((T_k+\lambda I)^{-1}T_k - I)(f_\star - I_k h)\|_{W_2}}_{(I)} + \underbrace{\|(T_k+\lambda I)^{-1}T_k I_k h - I_k h\|_{W_2}}_{(II)}. \end{aligned} \tag{36}$$

Since T_k is a self-adjoint compact operator, there exists $(\alpha_l)_{l\in\mathbb{N}}$ and ONB $(\phi_l)_{l\in\mathbb{N}}$ of $\overline{\mathcal{R}(T_k)}$ so that $T_k = \sum_l \alpha_l \langle \phi_l, \cdot\rangle_{W_2}\phi_l$. Let $(\psi_j)_{j\in\mathbb{N}}$ be the orthonormal basis of $\mathcal{N}(T_k)$. Then we have

$$\begin{aligned}(I)^2 &= \sum_l \left(\frac{\alpha_l}{\alpha_l+\lambda}-1\right)^2 \langle f_\star - I_k h, \phi_l\rangle^2_{W_2} + \sum_j \langle f_\star - I_k h, \psi_j\rangle^2_{W_2} \\ &\le \sum_l \langle f_\star - I_k h, \phi_l\rangle^2_{W_2} + \sum_j \langle f_\star - I_k h, \psi_j\rangle^2_{W_2} = \|f_\star - I_k h\|^2_{W_2}. \end{aligned} \tag{37}$$

From $(T_k+\lambda I)^{-1}T_k = I_k(E_k+\lambda I)^{-1}S_k$ and $S_k I_k h = E_k h$, we have

$$\begin{aligned}(II) &= \|I_k(E_k+\lambda I)^{-1}E_k h - I_k h\|_{W_2} \\ &= \|\sqrt{E_k}(E_k+\lambda I)^{-1}E_k h - \sqrt{E_k}h\|_{\mathcal{H}} \le \|h\|_{\mathcal{H}}\sqrt{\lambda}, \end{aligned} \tag{38}$$

where the inequality follows from Proposition A.3(ii). Using (37) and (38) in (36), we obtain $\mathcal{B}(\lambda) \le \|f_\star - I_k h\|_{W_2} + \|h\|_{\mathcal{H}}\sqrt{\lambda}$, using which in (35) yields

$$\sqrt{2\,J(p_0\|p_{f_{\lambda,n}})} \le \|f_\star - I_k h\|_{W_2} + O_{p_0}\left(\frac{1}{\sqrt{\lambda n}}\right) + \|h\|_{\mathcal{H}}\sqrt{\lambda}.$$

Since the above inequality holds for any $h \in \mathcal{H}$, we therefore have

$$\begin{aligned}\sqrt{2\,J(p_0\|p_{f_{\lambda,n}})} &\le \inf_{h\in\mathcal{H}}\left(\|f_\star - I_k h\|_{W_2} + \sqrt{\lambda}\|h\|_{\mathcal{H}}\right) + O_{p_0}\left(\frac{1}{\sqrt{\lambda n}}\right) \\ &= K(f_\star, \sqrt{\lambda}, W_2(p_0), I_k(\mathcal{H})) + O_{p_0}\left(\frac{1}{\sqrt{\lambda n}}\right) \end{aligned} \tag{39}$$

where the K-functional is defined in (A.6). Note that $I_k(\mathcal{H}) \cong \mathcal{H}/\mathcal{H}\cap\mathbb{R}$ is continuously embedded in $W(p_0)$. From (A.6), it is clear that the K-functional as a function of t is an infimum over a family of affine linear and increasing functions and therefore is concave, continuous and increasing w.r.t. t. This means, in (39), as $\lambda \to 0$,

$$K(f_\star, \sqrt{\lambda}, W_2(p_0), I_k(\mathcal{H})) \to \inf_{h\in\mathcal{H}} \|f_\star - I_k h\|_{W_2} = \sqrt{2\inf_{p\in\mathcal{P}} J(p_0\|p)}.$$

Since $J(p_0\|p_{f_{\lambda,n}}) \ge \inf_{p\in\mathcal{P}} J(p_0\|p)$, we have that $J(p_0\|p_{f_{\lambda,n}}) \to \inf_{p\in\mathcal{P}} J(p_0\|p)$ as $\lambda \to 0$, $\lambda n \to \infty$ and $n \to \infty$.

(ii) Recall $\mathcal{B}(\lambda)$ from (i). From Proposition A.3(i) it follows that $\mathcal{B}(\lambda) \to 0$ as $\lambda \to 0$ if $f_\star \in \overline{\mathcal{R}(T_k)}$. Therefore, (35) reduces to $\sqrt{2\,J(p_0\|p_{f_{\lambda,n}})} \le O_{p_0}\left(\frac{1}{\sqrt{\lambda n}}\right) + \mathcal{B}(\lambda)$ and the consistency result follows. If $f_\star \in \mathcal{R}(T_k^\beta)$ for some $\beta > 0$, then the rates follow from

Proposition A.3 by noting that $\mathcal{B}(\lambda) \le \max\{1, \|T_k\|^{\beta-1}\}\lambda^{\min\{1,\beta\}}\|T_k^{-\beta} f_\star\|_{W_2}$ and choosing $\lambda = n^{-\max\left\{\frac{1}{3},\frac{1}{2\beta+1}\right\}}$.

(iii) This simply follows from an analysis similar to the one used in the proof of Theorem 6(iii). ■

8.12 Proof of Proposition 13

For any $p \in \mathcal{P}_{\text{FD}}$, define $f := \log \frac{p}{q_0}$, which implies that $[f]_\sim \in W_2(p)$. Since $I_k(\mathcal{H})$ is dense in $W_2(p)$, we have for any $\epsilon > 0$, there exists $g \in \mathcal{H}$ such that $\|[f]_\sim - I_k g\|_{W_2} \le \sqrt{2\epsilon}$. For a given $g \in \mathcal{H}$, pick $p_g \in \mathcal{P}$. Therefore,

$$J(p\|p_g) = \frac{1}{2}\int_\Omega p(x)\,\|\nabla \log p - \nabla \log p_g\|_2^2\,dx = \frac{1}{2}\|[f]_\sim - I_k g\|_{W_2}^2 \le \epsilon$$

and the result follows. ■

Acknowledgments

We thank the action editor and two anonymous reviewers for their detailed comments that significantly improved the paper. Theorem A.2 and the proof are due to an anonymous reviewer. We also thank Dougal Sutherland for a careful reading of the paper which helped to fix minor errors. A part of the work was carried out while BKS was a Research Fellow in the Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge. BKS thanks Richard Nickl for many valuable comments and insightful discussions. KF is supported in part by MEXT Grant-in-Aid for Scientific Research on Innovative Areas 25120012.

A. Appendix: Technical Results

In this appendix, we present some technical results that are used in the proofs.

A.1 Bounds on Various Distances Between p_f and p_g

In the following result, claims (iii) and (iv) are quoted from Lemma 3.1 of van der Vaart and van Zanten (2008).

Lemma A.1 *Define $\mathcal{P}_\infty := \left\{p_f = e^{f-A(f)}q_0 \,:\, f \in \ell^\infty(\Omega)\right\}$, where q_0 is a probability density on $\Omega \subseteq \mathbb{R}^d$ and $\ell^\infty(\Omega)$ is the space of bounded measurable functions on Ω. Then for any $p_f, p_g \in \mathcal{P}_\infty$, we have*

(i) $\|p_f - p_g\|_{L^r(\Omega)} \le 2e^{2\|f-g\|_\infty} e^{2\min\{\|f\|_\infty,\|g\|_\infty\}}\|f - g\|_\infty \|q_0\|_{L^r(\Omega)}$ *for any* $1 \le r \le \infty$*;*

(ii) $\|p_f - p_g\|_{L^1(\Omega)} \le 2e^{\|f-g\|_\infty}\|f - g\|_\infty$*;*

(iii) $KL(p_f\|p_g) \le c\,\|f-g\|_\infty^2 e^{\|f-g\|_\infty}\left(1+\|f-g\|_\infty\right)$ *where* c *is a universal constant;*

(iv) $h(p_f,p_g) \le e^{\|f-g\|_\infty/2}\|f-g\|_\infty$.

Proof *(i)* Define $B(f) := \int e^f q_0\,dx$. Consider

$$\begin{aligned}
\|p_f-p_g\|_{L^r(\Omega)} &= \left\|\frac{e^f q_0}{B(f)} - \frac{e^g q_0}{B(g)}\right\|_{L^r(\Omega)} = \frac{\left\|e^f q_0 B(g) - e^g q_0 B(f)\right\|_{L^r(\Omega)}}{B(f)B(g)} \\
&= \frac{\left\|e^f q_0\left(B(g)-B(f)\right) + \left(e^f-e^g\right) q_0 B(f)\right\|_{L^r(\Omega)}}{B(f)B(g)} \\
&\le \frac{\left\|e^f q_0\left(B(g)-B(f)\right)\right\|_{L^r(\Omega)}}{B(f)B(g)} + \frac{\left\|\left(e^f-e^g\right) q_0 B(f)\right\|_{L^r(\Omega)}}{B(f)B(g)} \\
&\le \frac{|B(g)-B(f)|\,\|e^f q_0\|_{L^r(\Omega)}}{B(g)B(f)} + \frac{\left\|(e^f-e^g)q_0\right\|_{L^r(\Omega)}}{B(g)}.
\end{aligned} \tag{A.1}$$

Observe that

$$|B(f)-B(g)| \le \int_\Omega |e^f-e^g|q_0\,dx = \int_\Omega e^g|e^{f-g}-1|q_0\,dx \le e^{\|f-g\|_\infty}\|f-g\|_\infty B(g)$$

since $|e^{u-v}-1| \le |u-v|e^{|u-v|}$ for any $u,v\in\mathbb{R}$. Similarly,

$$\left\|(e^f-e^g)q_0\right\|_{L^r(\Omega)} \le e^{\|f-g\|_\infty}\|f-g\|_\infty\|e^g q_0\|_{L^r(\Omega)}.$$

Using these above, we obtain

$$\|p_f-p_g\|_{L^r(\Omega)} \le e^{\|f-g\|_\infty}\|f-g\|_\infty\left(\frac{\|e^f q_0\|_{L^r(\Omega)}}{B(f)} + \frac{\|e^g q_0\|_{L^r(\Omega)}}{B(g)}\right). \tag{A.2}$$

Since $\|e^f q_0\|_{L^r(\Omega)} \le e^{\|f\|_\infty}\|q_0\|_{L^r(\Omega)}$ and $B(f) \ge e^{-\|f\|_\infty}$, from (A.2) we obtain

$$\begin{aligned}
\|p_f-p_g\|_{L^r(\Omega)} &\le e^{\|f-g\|_\infty}\|f-g\|_\infty\|q_0\|_{L^r(\Omega)}\left(e^{2\|f\|_\infty}+e^{2\|g\|_\infty}\right). \\
&\le 2e^{\|f-g\|_\infty}\|f-g\|_\infty\|q_0\|_{L^r(\Omega)}e^{2\max\{\|f\|_\infty,\|g\|_\infty\}} \\
&\le 2e^{2\|f-g\|_\infty}\|f-g\|_\infty\|q_0\|_{L^r(\Omega)}e^{2\min\{\|f\|_\infty,\|g\|_\infty\}}
\end{aligned}$$

where we used $\max\{a,b\} \le \min\{a,b\}+|a-b|$ for $a,b\ge 0$ in the last line above.
(ii) This simply follows from (A.2) by using $r=1$. ■

A.2 General Representer Theorem

The following is the general representer theorem for abstract Hilbert spaces.

Theorem A.2 (General representer theorem) *Let H be a real Hilbert space and let $(\phi_i)_{i=1}^m \in H^m$. Suppose $J : H \to \mathbb{R}$ be such that $J(f) = V(\langle f, \phi_1\rangle_H, \ldots, \langle f, \phi_m\rangle_H)$, $f \in H$ where $V : \mathbb{R}^n \to \mathbb{R}$ is a convex differentiable function. Define*

$$f_\lambda = \arg \inf_{f\in H} J(f) + \frac{\lambda}{2}\|f\|_H^2,$$

where $\lambda > 0$. Then there exists $(\alpha_i)_{i=1}^m \in \mathbb{R}^m$ such that $f_\lambda = \sum_{i=1}^m \alpha_i\phi_i$ where $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_m)$ satisfies the following (possibly nonlinear) equation

$$\lambda\boldsymbol{\alpha} + \nabla V(\boldsymbol{K}\boldsymbol{\alpha}) = 0,$$

with $\boldsymbol{K}$ being a linear map on $\mathbb{R}^m$ and $(\boldsymbol{K})_{i,j} = \langle\phi_i, \phi_j\rangle_H$, $i \in [m]$, $j \in [m]$.

Proof Define $A : H \to \mathbb{R}^m$, $f \mapsto (\langle f, \phi_i\rangle_H)_{i=1}^m$. Then $f_\lambda = \arg\inf_{f\in H} V(Af) + \frac{\lambda}{2}\|f\|_H^2$. Therefore, Fermat's rule yields

$$\begin{aligned} 0 = A^*\nabla V(Af_\lambda) + \lambda f_\lambda \Leftrightarrow f_\lambda = A^*\left(-\frac{1}{\lambda}\nabla V(Af_\lambda)\right) \\ \Leftrightarrow (\exists\, \boldsymbol{\alpha} \in \mathbb{R}^m) \quad f_\lambda = A^*\boldsymbol{\alpha},\ \boldsymbol{\alpha} = -\frac{1}{\lambda}\nabla V(Af_\lambda) \\ \Leftrightarrow (\exists\, \boldsymbol{\alpha} \in \mathbb{R}^m) \quad f_\lambda = A^*\boldsymbol{\alpha},\ \boldsymbol{\alpha} = -\frac{1}{\lambda}\nabla V(AA^*\boldsymbol{\alpha}), \end{aligned}$$

where $A^* : \mathbb{R}^m \to H$ is the adjoint of A which can be obtained as follows. Note that

$$\langle Af, \boldsymbol{\alpha}\rangle = \sum_{i=1}^m \alpha_i\langle f, \phi_i\rangle_H = \left\langle f, \sum_{i=1}^m \alpha_i\phi_i\right\rangle_H \qquad (\forall\, f \in H)\ (\forall\, \boldsymbol{\alpha} \in \mathbb{R}^m)$$

and thus $A^*\boldsymbol{\alpha} = \sum_{i=1}^m \alpha_i\phi_i$. Therefore $AA^*\boldsymbol{\alpha} = \sum_{j=1}^m \alpha_j A\phi_j = \sum_{j=1}^m \alpha_j(\langle\phi_j, \phi_i\rangle_H)_{i=1}^m$, and so for every $i \in [m]$, $(AA^*\boldsymbol{\alpha})_i = \sum_{j=1}^m \langle\phi_j, \phi_i\rangle_H\alpha_j$ and hence $AA^* = \boldsymbol{K}$. ■

A.3 Bounds on Approximation Errors, $\mathcal{A}_0(\lambda)$ and $\mathcal{A}_{\frac{1}{2}}(\lambda)$

The following result is quite well-known in the linear inverse problem theory (Engl et al., 1996).

Proposition A.3 *Let C be a bounded, self-adjoint compact operator on a separable Hilbert space H. For $\lambda > 0$ and $f \in H$, define $f_\lambda := (C + \lambda I)^{-1}Cf$ and $\mathcal{A}_\theta(\lambda) := \|C^\theta(f_\lambda - f)\|_H$ for $\theta \geq 0$. Then the following hold.*

(i) For any $\theta > 0$, $\mathcal{A}_\theta(\lambda) \to 0$ as $\lambda \to 0$ and if $f \in \overline{\mathcal{R}(C)}$, then $\mathcal{A}_0(\lambda) \to 0$ as $\lambda \to 0$.

(ii) If $f \in \mathcal{R}(C^\beta)$ for $\beta \geq 0$ and $\beta + \theta > 0$, then

$$\mathcal{A}_\theta(\lambda) \leq \max\{1, \|C\|^{\beta+\theta-1}\}\lambda^{\min\{1,\beta+\theta\}}\|C^{-\beta}f\|_H.$$

Proof (i) Since C is bounded, compact, and self-adjoint, the Hilbert-Schmidt theorem (Reed and Simon, 1972, Theorems VI.16, VI.17) ensures that $C = \sum_l \alpha_l \phi_l \langle \phi_l, \cdot \rangle_H$, where $(\alpha_l)_{l\in\mathbb{N}}$ are the positive eigenvalues and $(\phi_l)_{l\in\mathbb{N}}$ are the corresponding unit eigenvectors that form an ONB for $\mathcal{R}(C)$. Let $\theta = 0$. Since $f \in \overline{\mathcal{R}(C)}$,

$$\mathcal{A}_0^2(\lambda) = \left\|(C+\lambda I)^{-1}Cf - f\right\|_H^2 = \left\|\sum_i \frac{\alpha_i}{\alpha_i+\lambda}\langle f,\phi_i\rangle_H \phi_i - \sum_i \langle f,\phi_i\rangle_H\phi_i\right\|_H^2$$
$$= \left\|\sum_i \frac{\lambda}{\alpha_i+\lambda}\langle f,\phi_i\rangle_H\phi_i\right\|_H^2 = \sum_i \left(\frac{\lambda}{\alpha_i+\lambda}\right)^2 \langle f,\phi_i\rangle_H^2 \to 0 \text{ as } \lambda\to 0$$

by the dominated convergence theorem. For any $\theta > 0$, we have

$$\mathcal{A}_\theta^2(\lambda) = \left\|C^\theta(C+\lambda I)^{-1}Cf - C^\theta f\right\|_H^2.$$

Let $f = f_R + f_N$ where $f_R \in \overline{\mathcal{R}(C^\theta)}$, $f_N \in \overline{\mathcal{R}(C^\theta)}^\perp$ if $0 < \theta \le 1$ and $f_R \in \overline{\mathcal{R}(C)}$, $f_N \in \overline{\mathcal{R}(C)}^\perp$ if $\theta \ge 1$. Then

$$\mathcal{A}_\theta^2(\lambda) = \left\|C^\theta(C+\lambda I)^{-1}Cf - C^\theta f\right\|_H^2 = \left\|C^\theta(C+\lambda I)^{-1}Cf_R - C^\theta f_R\right\|_H^2$$
$$= \left\|\sum_i \frac{\alpha_i^{1+\theta}}{\alpha_i+\lambda}\langle f_R,\phi_i\rangle_H\phi_i - \sum_i \alpha_i^\theta\langle f_R,\phi_i\rangle_H\phi_i\right\|_H^2$$
$$= \left\|\sum_i \frac{\lambda\alpha_i^\theta}{\alpha_i+\lambda}\langle f_R,\phi_i\rangle_H\phi_i\right\|_H^2 = \sum_i\left(\frac{\lambda\alpha_i^\theta}{\alpha_i+\lambda}\right)^2\langle f_R,\phi_i\rangle_H^2 \to 0$$

as $\lambda \to 0$.

(ii) If $f \in \mathcal{R}(C^\beta)$, then there exists $g \in H$ such that $f = C^\beta g$. This yields

$$\mathcal{A}_\theta^2(\lambda) = \left\|C^\theta(C+\lambda I)^{-1}Cf - C^\theta f\right\|_H^2 = \left\|C^\theta(C+\lambda I)^{-1}C^{\beta+1}g - C^{\theta+\beta}g\right\|_H^2$$
$$= \left\|\sum_i \frac{\lambda\alpha_i^{\theta+\beta}}{\alpha_i+\lambda}\langle g,\phi_i\rangle_H\phi_i\right\|_H^2 = \sum_i\left(\frac{\lambda\alpha_i^{\theta+\beta}}{\alpha_i+\lambda}\right)^2\langle g,\phi_i\rangle_H^2. \tag{A.3}$$

Suppose $0 < \beta+\theta < 1$. Then

$$\frac{\alpha_i^{\beta+\theta}\lambda}{\alpha_i+\lambda} = \left(\frac{\alpha_i}{\alpha_i+\lambda}\right)^{\beta+\theta}\left(\frac{\lambda}{\alpha_i+\lambda}\right)^{1-\theta-\beta}\lambda^{\beta+\theta} \le \lambda^{\beta+\theta}.$$

On the other hand, for $\beta+\theta \ge 1$, we have

$$\frac{\alpha_i^{\beta+\theta}\lambda}{\alpha_i+\lambda} = \left(\frac{\alpha_i}{\alpha_i+\lambda}\right)\alpha_i^{\beta+\theta-1}\lambda \le \|C\|^{\beta+\theta-1}\lambda.$$

Using the above in (A.3) yields the result. ■

A.4 Bound on the Norm of Certain Operators and Functions

The following result is used in many places throughout the paper. We would like to highlight that special cases of this result are known, e.g., see the proof of Theorem 4 in Caponnetto and Vito (2007) where concentration inequalities are obtained for the quantities in Proposition A.4 using Bernstein's inequality. Here, we provide asymptotic statements using Chebyshev's inequality.

Proposition A.4 *Let $\mathcal{X}$ be a topological space, H be a separable Hilbert space and $\mathcal{L}_2^+(H)$ be the space of positive, self-adjoint Hilbert-Schmidt operators on H. Define $R := \int_{\mathcal{X}} r(x)\, d\mathbb{P}(x)$ and $\hat{R} := \frac{1}{n}\sum_{a=1}^m r(X_a)$ where $\mathbb{P} \in M_+^1(\mathcal{X})$, $(X_a)_{a=1}^m \stackrel{i.i.d.}{\sim} \mathbb{P}$ and r is a $\mathcal{L}_2^+(H)$-valued measurable function on $\mathcal{X}$ satisfying $\int_{\mathcal{X}} \|r(x)\|_{HS}^2\, d\mathbb{P}(x) < \infty$. Define $g_\lambda := (R+\lambda I)^{-1}Rg$ for $g \in H$, $\lambda > 0$, and $\mathcal{A}_0(\lambda) := \|g_\lambda - g\|_H$. Let $\alpha \geq 0$ and $\theta > 0$. Then the following hold:*

(i) $\|(\hat{R}-R)(g_\lambda - g)\|_H = O_{\mathbb{P}}\left(\frac{\mathcal{A}_0(\lambda)}{\sqrt{m}}\right)$.

(ii) $\|R^\alpha(R+\lambda I)^{-\theta}\| \leq \lambda^{\alpha-\theta}$.

(iii) $\|\hat{R}^\alpha(\hat{R}+\lambda I)^{-\theta}\| \leq \lambda^{\alpha-\theta}$.

(iv) $\|(R+\lambda I)^{-\theta}(\hat{R}-R)\| = O_{\mathbb{P}}\left(\sqrt{\frac{1}{m\,\lambda^{2\theta}}}\right)$.

(v) $\|R^\alpha(\hat{R}+\lambda I)^{-1}\| \lesssim \lambda^{\alpha-1}$ *for $m \geq \frac{c}{\lambda^2}$ where is c is a sufficiently large constant that depends on $\int \|r(x)\|_{HS}^2\, d\mathbb{P}(x)$ but not on m and λ.*

Proof (i) Note that for any $f \in H$,

$$\mathbb{E}_{\mathbb{P}}\|(\hat{R}-R)f\|_H^2 = \mathbb{E}_{\mathbb{P}}\|\hat{R}f\|_H^2 + \|Rf\|_H^2 - 2\mathbb{E}_{\mathbb{P}}\langle \hat{R}f, Rf\rangle_H,$$

where $\mathbb{E}_{\mathbb{P}}\langle \hat{R}f, Rf\rangle_H = \frac{1}{n}\sum_{a=1}^n \mathbb{E}_{\mathbb{P}}\langle r(X_a)f, Rf\rangle_H = \frac{1}{n}\sum_{a=1}^n \mathbb{E}_{\mathbb{P}}\langle r(X_a), f\otimes Rf\rangle_{HS}$. Since $\int_{\mathcal{X}} \|r(x)\|_{HS}^2\, d\mathbb{P}(x) < \infty$, $r(x)$ is $\mathbb{P}$-integrable in the Bochner sense (see Diestel and Uhl, 1977, Definition 1 and Theorem 2), and therefore it follows from Diestel and Uhl (1977, Theorem 6) that $\mathbb{E}_{\mathbb{P}}\langle r(X_a), f\otimes Rf\rangle_{HS} = \langle \int_{\mathcal{X}} r(x)\, d\mathbb{P}(x), f\otimes Rf\rangle_{HS} = \|Rf\|_H^2$. Therefore,

$$\mathbb{E}_{\mathbb{P}}\|(\hat{R}-R)f\|_H^2 = \mathbb{E}_{\mathbb{P}}\|\hat{R}f\|_H^2 - \|Rf\|_H^2,$$

where

$$\mathbb{E}_{\mathbb{P}}\|\hat{R}f\|_H^2 = \mathbb{E}_{\mathbb{P}}\left\|\frac{1}{m}\sum_{a=1}^m r(X_a)f\right\|_H^2 = \frac{1}{m^2}\sum_{a,b=1}^m \mathbb{E}_{\mathbb{P}}\langle r(X_a)f, r(X_b)f\rangle_H.$$

Splitting the sum into two parts (one with $a=b$ and the other with $a \neq b$), it is easy to verify that $\mathbb{E}_{\mathbb{P}}\|\hat{R}f\|_H^2 = \frac{1}{m}\int_{\mathcal{X}} \|r(x)f\|_H^2\, d\mathbb{P}(x) + \frac{m-1}{m}\|Rf\|_H^2$, thereby yielding

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}\|(\hat{R}-R)f\|_H^2 = \frac{1}{m}\left(\int_{\mathcal{X}} \|r(x)f\|_H^2\, d\mathbb{P}(x) - \|Rf\|_H^2\right) &\leq \frac{1}{m}\int_{\mathcal{X}} \|r(x)f\|_H^2\, d\mathbb{P}(x)\\ &\leq \frac{\|f\|_H^2}{m}\int_{\mathcal{X}} \|r(x)\|_H^2\, d\mathbb{P}(x).\end{aligned}$$

Using $f = g_\lambda - g$, an application of Chebyshev's inequality yields the result.

(ii, iii) $\|R^\alpha(R+\lambda I)^{-\theta}\| = \sup_i \frac{\gamma_i^\alpha}{(\gamma_i+\lambda)^\theta} = \sup_i \left[\left(\frac{\gamma_i}{\gamma_i+\lambda}\right)^\alpha \frac{1}{(\gamma_i+\lambda)^{\theta-\alpha}}\right] \le \sup_i \frac{1}{(\gamma_i+\lambda)^{\theta-\alpha}} \le \lambda^{\alpha-\theta}$, where $(\gamma_i)_{i\in\mathbb{N}}$ are the eigenvalues of R. (iii) follows by replacing $(\gamma_i)_{i\in\mathbb{N}}$ with the eigenvalues of $\hat{R}$.

(iv) Since $\|(R+\lambda I)^{-\theta}(\hat{R}-R)\| \le \|(R+\lambda I)^{-\theta}(\hat{R}-R)\|_{HS}$, consider $\mathbb{E}_\mathbb{P}\|(R+\lambda I)^{-\theta}(\hat{R}-R)\|^2_{HS}$, which using the technique in the proof of (i), can be shown to be bounded as

$$\mathbb{E}_\mathbb{P}\|(R+\lambda I)^{-\theta}(\hat{R}-R)\|^2_{HS} \le \frac{1}{m}\int_\mathcal{X} \|(R+\lambda I)^{-\theta} r(x)\|^2_{HS}\, d\mathbb{P}(x). \tag{A.4}$$

Note that

$$\begin{aligned}\|(R+\lambda I)^{-\theta} r(x)\|^2_{HS} &= \langle (R+\lambda I)^{-\theta} r(x), (R+\lambda I)^{-\theta} r(x)\rangle_{HS} \\ &= \|(R+\lambda I)^{-2\theta}\|\mathrm{Tr}\,(r(x)r(x)) = \|(R+\lambda I)^{-2\theta}\|\|r(x)\|^2_{HS} \\ &\le \lambda^{-2\theta}\|r(x)\|^2_{HS},\end{aligned} \tag{A.5}$$

where the last inequality follows from (iii). Using (A.5) in (A.4), we obtain

$$\mathbb{E}_\mathbb{P}\|(R+\lambda I)^{-\theta}(\hat{R}-R)\|^2_{HS} \le \frac{1}{m\lambda^{2\theta}}\int_\mathcal{X} \|r(x)\|^2_{HS}\, d\mathbb{P}(x).$$

The result therefore follows by an application of Chebyshev's inequality.

(v) We use the idea in Step 2.1 of the proof of Theorem 4 in Caponnetto and Vito (2007), where $R^\alpha(\hat{R}+\lambda I)^{-1}$ is written equivalently as follows: Note that $\hat{R}+\lambda I = (\hat{R}-R)+(R+\lambda I)$, which implies

$$(\hat{R}+\lambda I)^{-1} = \left((\hat{R}-R)+(R+\lambda I)\right)^{-1} = (R+\lambda I)^{-1}\left(I-(R-\hat{R})(R+\lambda I)^{-1}\right)^{-1}$$

and so $R^\alpha(\hat{R}+\lambda I)^{-1} = R^\alpha(R+\lambda I)^{-1}\left(I-(R-\hat{R})(R+\lambda I)^{-1}\right)^{-1}$. Using the Von Neumann series representation, we have

$$R^\alpha(\hat{R}+\lambda I)^{-1} = R^\alpha(R+\lambda I)^{-1}\sum_{j=0}^\infty \left((R-\hat{R})(R+\lambda I)^{-1}\right)^j$$

so that

$$\begin{aligned}\|R^\alpha(\hat{R}+\lambda I)^{-1}\| &\le \|R^\alpha(R+\lambda I)^{-1}\|\sum_{j=0}^\infty \|(R-\hat{R})(R+\lambda I)^{-1}\|^j_{HS} \\ &\le \lambda^{\alpha-1}\sum_{j=0}^\infty \|(R-\hat{R})(R+\lambda I)^{-1}\|^j_{HS}.\end{aligned}$$

From the proof of (iv), we have that for any $\delta > 0$, with probability at least $1-\delta$, $\|(R-\hat{R})(R+\lambda I)^{-1}\|_{HS} \le \sqrt{\frac{\int_\mathcal{X}\|r(x)\|^2_{HS}\,d\mathbb{P}(x)}{m\lambda^2\delta}}$. Suppose $m \ge \frac{\int_\mathcal{X}\|r(x)\|^2_{HS}\,d\mathbb{P}(x)}{s^2\lambda^2\delta}$ where $s<1$. Then $\sum_{j=0}^\infty \|(R-\hat{R})(R+\lambda I)^{-1}\|^j_{HS} \le \sum_{j=0}^\infty s^j = \frac{1}{1-s}$. This means for $m \ge \frac{c}{\lambda^2}$ where c is sufficiently large, we obtain $\|R^\alpha(\hat{R}+\lambda I)^{-1}\| \lesssim \lambda^{\alpha-1}$. ■

A.5 Interpolation Space

In this section, we briefly recall the definition of interpolation spaces of the real method. To this end, let E_0 and E_1 be two arbitrary Banach spaces that are continuously embedded in some topological (Hausdorff) vector space $\mathcal{E}$. Then, for $x \in E_0 + E_1 := \{x_0 + x_1 : x_0 \in E_0,\, x_1 \in E_1\}$ and $t > 0$, the *K-functional* of the real interpolation method (see Bennett and Sharpley, 1988, Definition 1.1, p. 293) is defined by

$$K(x, t, E_0, E_1) := \inf\{\|x_0\|_{E_0} + t\|x_1\|_{E_1} \,:\, x_0 \in E_0,\, x_1 \in E_1,\, x = x_0 + x_1\}.$$

Suppose E and F are two Banach spaces that satisfy $F \hookrightarrow E$ (i.e., $F \subset E$ and the inclusion operator $\mathrm{id} : F \to E$ is continuous), then the K-functional reduces to

$$K(x, t, E, F) = \inf_{y \in F} \|x - y\|_E + t\|y\|_F. \tag{A.6}$$

The K-functional can be used to define interpolation norms, for $0 < \theta < 1$, $1 \le s \le \infty$ and $x \in E_0 + E_1$, as

$$\|x\|_{\theta,s} := \begin{cases} \left(\int \left(t^{-\theta} K(x,t)\right)^s t^{-1}\, dt\right)^{1/s}, & 1 \le s < \infty \\ \sup_{t>0} t^{-\theta} K(x,t), & s = \infty. \end{cases}$$

Moreover, the corresponding interpolation spaces (Bennett and Sharpley, 1988, Definition 1.7, p. 299) are defined as

$$[E_0, E_1]_{\theta,s} := \{x \in E_0 + E_1 \,:\, \|x\|_{\theta,s} < \infty\}.$$

B. Appendix: Miscellaneous Results

In this appendix, we present the proofs of some claims that we made in Sections 1, 4 and 5.

B.1 Relation between Fisher and Kullback-Leibler Divergences

The following result provides a relationship between Fisher and Kullback-Leibler divergences.

Proposition B.1 *Let p and q be probability densities defined on $\mathbb{R}^d$. Define $p_t := p * N(0, tI_d)$ and $q_t := q * N(0, tI_d)$ where $N(0, tI_d)$ denotes a normal distribution on $\mathbb{R}^d$ with mean zero and diagonal covariance with $t > 0$. Suppose p_t and q_t satisfy*

$$\partial_i p_t(x) \log p_t(x) = o\left(\|x\|_2^\alpha\right), \ \partial_i p_t(x) \log q_t(x) = o\left(\|x\|_2^\alpha\right) \text{ and } \partial_i \log q_t(x) p_t(x) = o\left(\|x\|_2^\alpha\right)$$

as $\|x\|_2 \to \infty$ for all $i \in [d]$ where $\alpha = 1 - d$. Then

$$KL(p\|q) = \int_0^\infty J(p_t\|q_t)\, dt, \tag{B.1}$$

where J is defined in (3).

Proof Under the conditions mentioned on p_t and q_t, it can be shown that

$$\frac{d}{dt}KL(p_t\|q_t) = -J(p_t\|q_t). \tag{B.2}$$

See Theorem 1 in Lyu (2009) for a proof. The above identity is a simple generalization of de Bruijn's identity that relates the Fisher information to the derivative of the Shannon entropy (see Cover and Thomas, 1991, Theorem 16.6.2). Integrating w.r.t. t on both sides of (B.2), we obtain $KL(p_t\|q_t)\Big|_{t=0}^{\infty} = -\int_0^\infty J(p_t\|q_t)\,dt$ which yields the equality in (B.1) as $KL(p_t\|q_t) \to 0$ as $t\to\infty$ and $KL(p_t\|q_t)\to KL(p\|q)$ as $t\to 0$. ■

B.2 Estimation of p_0: Unbounded k

To handle the case of unbounded k, in the following, we assume that there exists a positive constant M such that $\|f_0\|_{\mathcal{H}} \le M$, so that an estimator of f_0 can be constructed as

$$\breve{f}_{\lambda,n} = \arg\inf_{f\in\mathcal{H}} \hat{J}_\lambda(f) \;\text{ subject to }\; \|f\|_{\mathcal{H}} \le M, \tag{B.3}$$

where $\hat{J}_\lambda$ is defined in Theorem 4(iv). This modification yields a valid estimator $p_{\breve{f}_{\lambda,n}}$ as long as k satisfies $\int_\Omega e^{M\sqrt{k(x,x)}}q_0(x)\,dx < \infty$, since this implies $\breve{f}_{\lambda,n} \in \mathcal{F}$. The construction of $\breve{f}_{\lambda,n}$ requires the knowledge of M, however, which we assume is known *a priori*. Using the representer theorem in RKHS, it can be shown (see Section B.2.1) that

$$\breve{f}_{\lambda,n} = \breve{\delta}\hat{\xi} + \sum_{b=1}^{n}\sum_{j=1}^{d}\breve{\beta}_{(b-1)d+j}\partial_j k(X_b,\cdot)$$

where $\breve{\delta}$ and $\breve{\boldsymbol{\beta}}$ are obtained by solving the following quadratically constrained quadratic program (QCQP),

$$(\breve{\boldsymbol{\beta}},\breve{\delta}) =: \breve{\Theta} = \arg\min_{\Theta\in\mathbb{R}^{nd+1}} \frac{1}{2}\Theta^T\boldsymbol{H}\Theta + \Theta^T\Delta \;\text{ subject to }\; \Theta^T\boldsymbol{K}\Theta \le M^2,$$

with $\Delta := (\boldsymbol{h}, \|\hat{\xi}\|^2_{\mathcal{H}})$, $\Theta := (\boldsymbol{\beta},\delta)$ and $\boldsymbol{K}$, $\boldsymbol{H}$ being defined in the proof of Theorem 5 and the remark following it. The following result investigates the consistency and convergence rates for $p_{\breve{f}_{\lambda,n}}$.

Theorem B.2 (Consistency and rates for $p_{\breve{f}_{\lambda,n}}$) *Let $M \ge \|f_0\|_{\mathcal{H}}$ be a fixed constant, and $\breve{f}_{n,\lambda}$ be a clipped estimator given by (B.3). Suppose* **(A)**–**(D)** *with $\varepsilon = 2$ hold. Let* $\text{supp}(q_0) = \Omega$ *and $\int_\Omega e^{M\sqrt{k(x,x)}}q_0(x)\,dx < \infty$. Define $\eta(x) = \sqrt{k(x,x)}e^{M\sqrt{k(x,x)}}$. Then, as $\lambda\sqrt{n}\to\infty$, $\lambda\to 0$ and $n\to\infty$,*

(i) $\|p_{\breve{f}_{\lambda,n}} - p_0\|_{L^1(\Omega)} \to 0$, $KL(p_0\|p_{\breve{f}_{\lambda,n}}) \to 0$ *if* $\eta \in L^1(\Omega, q_0)$;

(ii) for $1 < r \le \infty$, $\|p_{\breve{f}_{\lambda,n}} - p_0\|_{L^r(\Omega)} \to 0$ *if* $\eta q_0 \in L^1(\Omega)\cap L^r(\Omega)$ *and* $e^{M\sqrt{k(\cdot,\cdot)}}q_0 \in L^r(\Omega)$;

(iii) $h(p_{\breve{f}_{\lambda,n}}, p_0) \to 0$ *if* $\sqrt{k(\cdot,\cdot)}\eta \in L^1(\Omega, q_0)$*;*

(iv) $J(p_0\|p_{\breve{f}_{\lambda,n}}) \to 0$.

In addition, if $f_0 \in \mathcal{R}(C^\beta)$ *for some* $\beta > 0$*, then* $\|p_{\breve{f}_{\lambda,n}} - p_0\|_{L^r(\Omega)} = O_{p_0}(\theta_n)$*,* $h(p_0, p_{\breve{f}_{\lambda,n}}) = O_{p_0}(\theta_n)$*,* $KL(p_0\|p_{\breve{f}_{\lambda,n}}) = O_{p_0}(\theta_n)$ *and* $J(p_0\|p_{\breve{f}_{\lambda,n}}) = O_{p_0}(\theta_n^2)$ *where* $\theta_n := n^{-\min\left\{\frac{1}{4},\frac{\beta}{2(\beta+1)}\right\}}$ *with* $\lambda = n^{-\max\left\{\frac{1}{4},\frac{1}{2(\beta+1)}\right\}}$ *assuming the respective conditions in (i)-(iii) above hold.*

Proof For any $x \in \Omega$, since $|f_0(x)| \le \|f_0\|_{\mathcal{H}}\sqrt{k(x,x)} \le M\sqrt{k(x,x)}$ and $|\breve{f}_{\lambda,n}(x)| \le M\sqrt{k(x,x)}$, we have

$$\left|e^{\breve{f}_{\lambda,n}(x)} - e^{f_0(x)}\right| \le e^{M\sqrt{k(x,x)}}\left|\breve{f}_{\lambda,n}(x) - f_0(x)\right| \le \eta(x)\left\|\breve{f}_{\lambda,n} - f_0\right\|_{\mathcal{H}}, \tag{B.4}$$

where we used the fact that $|e^x - e^y| \le e^a|x-y|$ for $x, y \in [-a, a]$ and $\eta(x) := \sqrt{k(x,x)}e^{M\sqrt{k(x,x)}}$. In the following, we obtain bounds for $\left\|p_{\breve{f}_{\lambda,n}} - p_0\right\|_{L^r(\Omega)}$ for any $1 \le r \le \infty$, $h(p_{\breve{f}_{\lambda,n}}, p_0)$ and $KL(p_0\|p_{\breve{f}_{\lambda,n}})$ in terms of $\|\breve{f}_{\lambda,n} - f_0\|_{\mathcal{H}}$. Define $B(f) := \int_\Omega e^f q_0\,dx$. Since k satisfies $\int_\Omega e^{M\sqrt{k(x,x)}}q_0(x)\,dx < \infty$, then it is clear that $\breve{f}_{\lambda,n} \in \mathcal{F}$ as $B(\breve{f}_{\lambda,n}) < \infty$ since

$$\int_\Omega e^{\breve{f}_{\lambda,n}(x)}q_0(x)\,dx \le \int_\Omega e^{\|\breve{f}_{\lambda,n}\|_{\mathcal{H}}\sqrt{k(x,x)}}q_0(x)\,dx \le \int_\Omega e^{M\sqrt{k(x,x)}}q_0(x)\,dx < \infty.$$

Similarly, it is easy to verify that $B(f_0) < \infty$.
(i) Recalling (A.1), we have

$$\left\|p_{\breve{f}_{\lambda,n}} - p_0\right\|_{L^r(\Omega)} \le \frac{|B(\breve{f}_{\lambda,n}) - B(f_0)|\|e^{f_0}q_0\|_{L^r(\Omega)}}{B(\breve{f}_{\lambda,n})B(f_0)} + \frac{\|(e^{f_0} - e^{\breve{f}_{\lambda,n}})q_0\|_{L^r(\Omega)}}{B(\breve{f}_{\lambda,n})}.$$

If $r = 1$, we obtain

$$\left\|p_{\breve{f}_{\lambda,n}} - p_0\right\|_{L^1(\Omega)} \le \frac{|B(\breve{f}_{\lambda,n}) - B(f_0)|}{B(\breve{f}_{\lambda,n})} + \frac{\|(e^{f_0} - e^{\breve{f}_{\lambda,n}})q_0\|_{L^1(\Omega)}}{B(\breve{f}_{\lambda,n})}.$$

Using (B.4), we bound $|B(\breve{f}_{\lambda,n}) - B(f_0)|$ as

$$|B(\breve{f}_{\lambda,n}) - B(f_0)| \le \int_\Omega\left|e^{\breve{f}_{\lambda,n}(x)} - e^{f_0(x)}\right|q_0(x)\,dx \le \|\eta\|_{L^1(\Omega,q_0)}\left\|\breve{f}_{\lambda,n} - f_0\right\|_{\mathcal{H}}.$$

Also for any $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \le M$, we have $B(f) \ge \int_\Omega e^{-M\sqrt{k(x,x)}}q_0(x)\,dx =: \theta$, where $\theta > 0$. Again using (B.4), we have

$$\|(e^{f_0} - e^{\breve{f}_{\lambda,n}})q_0\|_{L^r(\Omega)} \le \|\eta q_0\|_{L^r(\Omega)}\|\breve{f}_{\lambda,n} - f_0\|_{\mathcal{H}}$$

and $\|e^{f_0}q_0\|_{L^r(\Omega)} \le \|e^{M\sqrt{k(x,x)}}q_0\|_{L^r(\Omega)}$. Therefore,

$$\left\|p_{\breve{f}_{\lambda,n}} - p_0\right\|_{L^r(\Omega)} \le \frac{\|\eta\|_{L^1(\Omega,q_0)}\|e^{M\sqrt{k(x,x)}}q_0\|_{L^r(\Omega)}\left\|\breve{f}_{\lambda,n} - f_0\right\|_{\mathcal{H}}}{\theta^2}$$

$$+\frac{\|\eta q_0\|_{L^r(\Omega)}\|\breve{f}_{\lambda,n}-f_0\|_{\mathcal{H}}}{\theta}$$

and for $r=1$,

$$\left\|p_{\breve{f}_{\lambda,n}}-p_0\right\|_{L^1(\Omega)}\le\frac{2\,\|\eta\|_{L^1(\Omega,q_0)}\left\|\breve{f}_{\lambda,n}-f_0\right\|_{\mathcal{H}}}{\theta}.$$

(ii) Also

$$\begin{aligned}KL(p_0\|p_{\breve{f}_{\lambda,n}})&=\int_\Omega p_0\log\frac{p_0}{p_{\breve{f}_{\lambda,n}}}\,dx=\int_\Omega\log\left(e^{f_0-\breve{f}_{\lambda,n}}\frac{B(\breve{f}_{\lambda,n})}{B(f_0)}\right)p_0(x)\,dx\\&=\int_\Omega\left(f_0-\breve{f}_{\lambda,n}+\log\frac{B(\breve{f}_{\lambda,n})}{B(f_0)}\right)p_0(x)\,dx\\&\le\frac{|B(\breve{f}_{\lambda,n})-B(f_0)|}{B(f_0)}+\|\breve{f}_{\lambda,n}-f_0\|_{L^1(\Omega,p_0)}\le\frac{2\,\|\eta q_0\|_{L^1(\Omega)}}{\theta}\|\breve{f}_{\lambda,n}-f_0\|_{\mathcal{H}}.\end{aligned}$$

(iii) It is easy to verify that

$$\begin{aligned}h(p_{\breve{f}_{\lambda,n}},p_0)&=\left\|\frac{e^{\breve{f}_{\lambda,n}/2}}{\|e^{\breve{f}_{\lambda,n}/2}\|_{L^2(\Omega,q_0)}}-\frac{e^{f_0/2}}{\left\|e^{f_0/2}\right\|_{L^2(\Omega,q_0)}}\right\|_{L^2(\Omega,q_0)}\\&\le\frac{2\|e^{\breve{f}_{\lambda,n}/2}-e^{f_0/2}\|_{L^2(\Omega,q_0)}}{\left\|e^{f_0/2}\right\|_{L^2(\Omega,q_0)}}\end{aligned}$$

where the above inequality is obtained by carrying out and simplifying the decomposition as in (A.1). Using (B.4), we therefore have

$$h(p_{\breve{f}_{\lambda,n}},p_0)\le\sqrt{\frac{\int_\Omega k(x,x)e^{M\sqrt{k(x,x)}}q_0\,dx}{\theta}}\|\breve{f}_{\lambda,n}-f_0\|_{\mathcal{H}}.$$

(iv) As f_0, $\breve{f}_{\lambda,n}\in\mathcal{F}$, by Theorem 4, we obtain $J(p_0\|p_{\breve{f}_{\lambda,n}})=\frac{1}{2}\|\sqrt{C}(\breve{f}_{\lambda,n}-f_0)\|^2_{\mathcal{H}}\le\frac{1}{2}\|\sqrt{C}\|^2\|\breve{f}_{\lambda,n}-f_0\|^2_{\mathcal{H}}$.

Note that we have bounded the various distances between $p_{\breve{f}_{\lambda,n}}$ and p_0 in terms of $\|\breve{f}_{\lambda,n}-f_0\|_{\mathcal{H}}$. Since $\breve{f}_{\lambda,n}=f_{\lambda,n}$ with probability converging to 1, the assertions on consistency are proved by Theorem 6(i) in combination with Lemma 14—as we did not explicitly assume $f_0\in\overline{\mathcal{R}(C)}$—and the rates follow from Theorem 6(iii). ■

Remark The following observations can be made while comparing the scenarios of using bounded vs. unbounded kernels in the problem of estimating p_0 through Theorems 7 and B.2. First, the consistency results in L^r, Hellinger and KL distances are the same but for additional integrability conditions on k and q_0. The additional integrability conditions are not too difficult to hold in practice as they involve k and q_0 which can be chosen appropriately. However, the unbounded situation in Theorem B.2 requires the knowledge of M which is usually not known. On the other hand, the consistency result in J in Theorem B.2 is slightly weaker than in Theorem 7. This may be an artifact of our analysis as we are not able to adapt the bounding technique used in the proof of Theorem 7 to bound $J(p_0\|p_{\breve{f}_{\lambda,n}}) = \frac{1}{2}\|\sqrt{C}(\breve{f}_{\lambda,n} - f_0)\|^2_{\mathcal{H}}$ as it critically depends on the boundedness of k. Therefore, we used a trivial bound of $J(p_0\|p_{\breve{f}_{\lambda,n}}) = \frac{1}{2}\|\sqrt{C}(\breve{f}_{\lambda,n} - f_0)\|^2_{\mathcal{H}} \le \frac{1}{2}\|\sqrt{C}\|^2\|\breve{f}_{\lambda,n} - f_0\|^2_{\mathcal{H}}$, which yields the result through Theorem 6(i). Due to the same reason, we also obtain a slower rate of convergence in J. Second, the rate of convergence in KL is slower than in Theorem B.2, which again may be an artifact of our analysis. The convergence rate for KL in Theorem 7 is based on the application of Theorem 6(ii) in Lemma A.1, where the bound on KL in Lemma A.1 critically uses the boundedness to upper bound KL in terms of squared Hellinger distance.

B.2.1 Derivation of $\breve{f}_{\lambda,n}$

Any $f \in \mathcal{H}$ can be decomposed as $f = f_{\|} + f_{\perp}$ where

$$f_{\|} \in \text{span}\left\{\hat{\xi}, (\partial_j k(X_b, \cdot))_{b,j}\right\} =: \mathcal{H}_{\|},$$

which is a closed subset of $\mathcal{H}$ and $f_{\perp} \in \mathcal{H}_{\|}^{\perp} := \left\{g \in \mathcal{H} : \langle g, h\rangle_{\mathcal{H}} = 0, \forall\, h \in \mathcal{H}_{\|}\right\}$ so that $\mathcal{H} = \mathcal{H}_{\|} \oplus \mathcal{H}_{\|}^{\perp}$. Since the objective function in (B.3) matches with the one in Theorem 5, using the above decomposition in (B.3), it is easy to verify that $\hat{J}$ depends only on $f_{\|} \in \mathcal{H}_{\|}$ so that (B.3) reduces to

$$(\breve{f}^{\|}_{\lambda,n}, \breve{f}^{\perp}_{\lambda,n}) = \arg \inf_{\substack{f_{\|} \in \mathcal{H}_{\|}, f_{\perp} \in \mathcal{H}^{\perp} \\ \|f_{\|}\|^2_{\mathcal{H}} + \|f_{\perp}\|^2_{\mathcal{H}} \le M^2}} \hat{J}_{\lambda}(f_{\|}) + \frac{\lambda}{2}\|f_{\|}\|^2_{\mathcal{H}} + \frac{\lambda}{2}\|f_{\perp}\|^2_{\mathcal{H}} \tag{B.5}$$

and $\breve{f}_{\lambda,n} = \breve{f}^{\|}_{\lambda,n} + \breve{f}^{\perp}_{\lambda,n}$. Since $f_{\|}$ is of the form in (14), using it in (B.5), it is easy to show that $\hat{J}_{\lambda}(f_{\|}) + \frac{\lambda}{2}\|f_{\|}\|^2_{\mathcal{H}} = \frac{1}{2}\Theta^T \boldsymbol{H}\Theta + \Theta^T\Delta$. Similarly, it can be shown that $\|f_{\|}\|^2_{\mathcal{H}} = \Theta^T\boldsymbol{K}\Theta$. Since $f_{\perp}$ appears in (B.5) only through $\|f_{\perp}\|^2_{\mathcal{H}}$, (B.5) reduces to

$$(\Theta_{\|}, c_{\perp}) = \arg \inf_{\substack{\Theta \in \mathbb{R}^{nd+1}, c \ge 0 \\ \Theta^T\boldsymbol{K}\Theta + c \le M^2}} \frac{1}{2}\Theta^T\boldsymbol{H}\Theta + \Theta^T\Delta + \frac{\lambda}{2}c, \tag{B.6}$$

where $\breve{f}^{\|}_{\lambda,n}$ is constructed as in (14) using $\Theta_{\|}$ and $\breve{f}^{\perp}_{\lambda,n}$ is such that $\|\breve{f}^{\perp}_{\lambda.n}\|^2_{\mathcal{H}} = c_{\perp}$. The necessary and sufficient conditions for the optimality of $(\Theta_{\|}, c_{\perp})$ is given by the following Karush-Kuhn-Tucker conditions,

$$(\boldsymbol{H} + 2\tau\boldsymbol{K})\Theta_{\|} + \Delta = 0,\ \frac{\lambda}{2} + \eta - \tau = 0 \qquad \text{(Stationarity)}$$

$$\Theta_{\|}^T\boldsymbol{K}\Theta_{\|} + c_{\perp} \le M^2,\ c_{\perp} \ge 0 \qquad \text{(Primal feasibility)}$$

$$\eta \ge 0,\ \tau \ge 0 \qquad \text{(Dual feasibility)}$$

$$\tau c_{\perp} = 0,\ \eta(\Theta_{\|}^T\boldsymbol{K}\Theta_{\|} + c_{\perp} - M^2) = 0 \qquad \text{(Complementary slackness)}$$

Combining the dual feasibility and stationary conditions, we have $\eta = \tau - \frac{\lambda}{2} \ge 0$, i.e., $\tau \ge \frac{\lambda}{2}$. Using this in the complementary slackness involving τ and $c_{\perp}$, it follows that $c_{\perp} = 0$. Since $\|\breve{f}^{\perp}_{\lambda,n}\|^2 = c_{\perp}$, we have $\breve{f}^{\perp}_{\lambda,n} = 0$, i.e., $\breve{f}_{\lambda,n}$ is completely determined by $\breve{f}^{\|}_{\lambda,n}$. Therefore $\breve{f}^{\|}_{\lambda,n}$ is of the form in (14) and (B.6) reduces to a quadratically constrained quadratic program.

B.3 $\mathcal{R}(C^\beta)$ and Interpolation Spaces

Proposition B.3 presents an interpretation for $\mathcal{R}(C^\beta)$ ($\beta > 0$ and $\beta \notin \mathbb{N}$) as interpolation spaces between $\mathcal{R}(C^{\lceil \beta \rceil})$ and $\mathcal{R}(C^{\lfloor \beta \rfloor})$ where $\mathcal{R}(C^0) := \mathcal{H}$. An inspection of its proof shows that Proposition B.3 holds for any self-adjoint, bounded, compact operator defined on a separable Hilbert space.

Proposition B.3 *Suppose* **(B)** *and* **(D)** *hold with $\varepsilon = 1$. Then for all $\beta > 0$ and $\beta \notin \mathbb{N}$*

$$\mathcal{R}(C^\beta) = \left[\mathcal{R}(C^{\lfloor \beta \rfloor}), \mathcal{R}(C^{\lceil \beta \rceil})\right]_{\beta - \lfloor \beta \rfloor, 2}$$

where $\mathcal{R}(C^0) := \mathcal{H}$, and the spaces $\mathcal{R}(C^\beta)$ and $\left[\mathcal{R}(C^{\lfloor \beta \rfloor}), \mathcal{R}(C^{\lceil \beta \rceil})\right]_{\beta - \lfloor \beta \rfloor, 2}$ have equivalent norms.

To prove Proposition B.3, we need the following result which we quote from Steinwart and Scovel (2012, Lemma 6.3) (also see Tartar, 2007, Lemma 23.1) that interpolates L^2-spaces whose underlying measures are absolutely continuous with respect to a measure ν.

Lemma B.4 *Let ν be a measure on a measurable space Θ and $w_0 : \Theta \to [0, \infty)$ and $w_1 : \Theta \to [0, \infty)$ be measurable functions. For $0 < \beta < 1$, define $w_\beta := w_0^{1-\beta} w_1^\beta$. Then we have*

$$[L^2(w_0 \, d\nu), L^2(w_1 \, d\nu)]_{\beta,2} = L^2(w_\beta \, d\nu)$$

and the norms on these two spaces are equivalent. Moreover, this result still holds for weights $w_0 : \Theta \to (0, \infty)$ and $w_1 : \Theta \to [0, \infty]$, if one uses the convention $0 \cdot \infty := 0$ in the definition of the weighted spaces.

Proof of Proposition B.3. The proof is based on the ideas used in the proof of Theorem 4.6 in Steinwart and Scovel (2012). Recall that by the Hilbert-Schmidt theorem, C has the following representation,

$$C = \sum_{i \in I} \alpha_i \phi_i \langle \phi_i, \cdot \rangle_{\mathcal{H}}$$

where $(\alpha_i)_{i \in I}$ are the positive eigenvalues of C, $(\phi_i)_{i \in I}$ are the corresponding unit eigenvectors that form an ONB for $\mathcal{R}(C)$ and I is an index set which is either finite (if $\mathcal{H}$ is finite-dimensional) or $I = \mathbb{N}$ with $\lim_{i \to \infty} \alpha_i = 0$ (if $\mathcal{H}$ is infinite dimensional). Let $(\psi_i)_{i \in J}$ be an ONB for $\mathcal{N}(C)$ where J is some index set so that any $f \in \mathcal{H}$ can be written as

$$f = \sum_{i \in I} \langle f, \phi_i \rangle_{\mathcal{H}} \phi_i + \sum_{i \in J} \langle f, \psi_i \rangle_{\mathcal{H}} \psi_i =: \sum_{i \in I \cup J} a_i \theta_i$$

where $\theta_i := \phi_i$ if $i \in I$ and $\theta_i := \psi_i$ if $i \in J$ with $a_i := \langle f, \theta_i \rangle_{\mathcal{H}}$. Let $\beta > 0$. By definition, $g \in \mathcal{R}(C^\beta)$ is equivalent to $\exists h \in \mathcal{H}$ such that $g = C^\beta h$, i.e.,

$$g = \sum_{i \in I} \alpha_i^\beta \langle h, \phi_i \rangle_{\mathcal{H}} \phi_i =: \sum_{i \in I} b_i \alpha_i^\beta \phi_i$$

where $b_i := \langle h, \phi_i \rangle_{\mathcal{H}}$. Clearly $\sum_{i\in I} b_i^2 = \sum_{i\in I} \langle h, \phi_i \rangle_{\mathcal{H}}^2 \le \|h\|_{\mathcal{H}}^2 < \infty$, i.e., $(b_i) \in \ell_2(I)$. Therefore

$$\mathcal{R}(C^\beta) = \left\{ \sum_{i\in I} b_i \alpha_i^\beta \phi_i \,:\, (b_i) \in \ell_2(I) \right\} = \left\{ \sum_{i\in I} c_i \phi_i \,:\, (c_i) \in \ell_2(I, \alpha^{-2\beta}) \right\}$$

where $\alpha := (\alpha_i)_{i\in I}$. Let us equip this space with the bilinear form

$$\left\langle \sum_{i\in I} c_i \phi_i, \sum_{i\in I} d_i \phi_i \right\rangle_{\mathcal{R}(C^\beta)} := \langle (c_i), (d_i) \rangle_{\ell_2(I,\alpha^{-2\beta})}$$

so that it induces the norm

$$\left\| \sum_{i\in I} c_i \phi_i \right\|_{\mathcal{R}(C^\beta)} := \|(c_i)\|_{\ell_2(I,\alpha^{-2\beta})}.$$

It is easy to verify that $(\alpha_i^\beta \phi_i)_{i\in I}$ is an ONB of $\mathcal{R}(C^\beta)$. Also since $\mathcal{R}(C^{\beta_1}) \subset \mathcal{R}(C^{\beta_2})$ for $0 < \beta_2 < \beta_1 < \infty$ and $\mathrm{id} : \mathcal{R}(C^{\beta_1}) \to \mathcal{R}(C^{\beta_2})$ is continuous, i.e., for any $g \in \mathcal{R}(C^{\beta_1})$,

$$\begin{aligned}\|g\|_{\mathcal{R}(C^{\beta_2})} = \|(c_i)\|_{\ell_2(I,\alpha^{-2\beta_2})} = \sqrt{\sum_{i\in I} \frac{c_i^2}{\alpha_i^{2\beta_2}}} &\le \sup_{i\in I} |\alpha_i|^{\beta_1-\beta_2} \|(c_i)\|_{\ell_2(I,\alpha^{-2\beta_1})} \\ &= \|C\|^{\beta_1-\beta_2} \|g\|_{\mathcal{R}(C^{\beta_1})} < \infty\end{aligned}$$

and so $\mathcal{R}(C^{\beta_1}) \hookrightarrow \mathcal{R}(C^{\beta_2})$. Similarly, we can show that $\mathcal{R}(C) \hookrightarrow \mathcal{H}$. In the following, we first prove the result for $0 < \beta < 1$ and then for $\beta > 1$.

(a) $\underline{0 < \beta < 1}$: For any $f \in \mathcal{H}$ and $g \in \mathcal{R}(C)$, we have

$$\|f - g\|_{\mathcal{H}}^2 = \left\| \sum_{i\in I\cup J} a_i \theta_i - \sum_{i\in I} c_i \phi_i \right\|_{\mathcal{H}}^2 = \left\| \sum_{i\in I\cup J} (a_i - c_i) \theta_i \right\|_{\mathcal{H}}^2 = \|(a_i - c_i)\|_{\ell_2(I\cup J)}^2$$

where we define $c_i := 0$ for $i \in J$. For $t > 0$, we find

$$\begin{aligned}K(f, t, \mathcal{H}, \mathcal{R}(C)) &= \inf_{g\in\mathcal{R}(C)} \|f - g\|_{\mathcal{H}} + t\|g\|_{\mathcal{R}(C)} \\ &= \inf_{(c_i)\in\ell_2(I,\alpha^{-2})} \|(a_i - c_i)\|_{\ell_2(I\cup J)} + t\|(c_i)\|_{\ell_2(I,\alpha^{-2})} \\ &= K(a, t, \ell_2(I\cup J), \ell_2(I, \alpha^{-2})).\end{aligned}$$

From this we immediately obtain the equivalence

$$f \in [\mathcal{H}, \mathcal{R}(C)]_{\beta,2} \iff (a_i) \in [\ell_2(I\cup J), \ell_2(I, \alpha^{-2})]_{\beta,2}$$

where $0 < \beta < 1$. Applying the second part of Lemma B.4 to the counting measure on $I \cup J$ yields

$$[\ell_2(I\cup J), \ell_2(I, \alpha^{-2})]_{\beta,2} = \ell_2(I, \alpha^{-2\beta}).$$

Since $\mathcal{R}(C^\beta)$ and $\ell_2(I, \alpha^{-2\beta})$ are isometrically isomorphic, we obtain $\mathcal{R}(C^\beta) = [\mathcal{H}, \mathcal{R}(C)]_{\beta,2}$.

(b) $\underline{\beta > 1 \text{ and } \beta \notin \mathbb{N}}$: Define $\gamma := \lfloor \beta \rfloor$. Let $f \in \mathcal{R}(C^\gamma)$ and $g \in \mathcal{R}(C^{\gamma+1})$, i.e., $\exists\, (c_i) \in \ell_2(I, \alpha^{-2\gamma})$ and $(d_i) \in \ell_2(I, \alpha^{-2\gamma-2})$ such that $f = \sum_{i \in I} c_i \phi_i$ and $g = \sum_{i \in I} d_i \phi_i$. Since

$$\|f - g\|^2_{\mathcal{R}(C^\gamma)} = \|(c_i - d_i)\|^2_{\ell_2(I, \alpha^{-2\gamma})},$$

for $t > 0$, we have

$$\begin{aligned} K(f, t, \mathcal{R}(C^\gamma), \mathcal{R}(C^{\gamma+1})) &= \inf_{g \in \mathcal{R}(C^{\gamma+1})} \|f - g\|_{\mathcal{R}(C^\gamma)} + t\|g\|_{\mathcal{R}(C^{\gamma+1})} \\ &= \inf_{(d_i) \in \ell_2(I, \alpha^{-2\gamma-2})} \|(c_i - d_i)\|_{\ell_2(I,\alpha^{-2\gamma})} + t\|(d_i)\|_{\ell_2(I,\alpha^{-2\gamma-2})} \\ &= K(c, t, \ell_2(I, \alpha^{-2\gamma}), \ell_2(I, \alpha^{-2\gamma-2})), \end{aligned}$$

from which we obtain the following equivalence

$$f \in [\mathcal{R}(C^\gamma), \mathcal{R}(C^{\gamma+1})]_{\beta-\gamma,2} \iff (c_i) \in [\ell_2(I, \alpha^{-2\gamma}), \ell_2(I, \alpha^{-2\gamma-2})]_{\beta-\gamma,2} \overset{(*)}{=} \ell_2(I, \alpha^{-2\beta}),$$

where $(*)$ follows from Lemma B.4 and the result is obtained by noting that $\ell_2(I, \alpha^{-2\beta})$ and $\mathcal{R}(C^\beta)$ are isometrically isomorphic. ■

B.4 Denseness of $I_k\mathcal{H}$ in $W_2(\mathbb{R}^d, p)$

In this section, we discuss the denseness of $I_k\mathcal{H}$ in $W_2(\mathbb{R}^d, p)$ for a given $p \in \mathcal{P}_{\text{FD}}$, where $\mathcal{P}_{\text{FD}}$ is defined in Theorem 13, which is equivalent to the injectivity of S_k (see Rudin, 1991, Theorem 4.12). To this end, in the following result we show that under certain conditions on a bounded continuous translation invariant kernel on $\mathbb{R}^d$, the restriction of S_k to $\mathcal{W}_2^\sim(\mathbb{R}^d, p)$ is injective when $d = 1$, while the result for any general $d > 1$ is open. However, even for $d = 1$, this does not guarantee the injectivity of S_k (which is defined on $W_2(\mathbb{R}^d, p)$). Therefore, the question of characterizing the injectivity of S_k (or equivalently the denseness of $I_k\mathcal{H}$ in $W_2(\mathbb{R}^d, p)$) is open.

Proposition B.5 *Suppose $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$ where $\psi \in C_b(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$, $\int \|\omega\|_2 \psi^\wedge(\omega)\, d\omega < \infty$ and $\text{supp}(\psi^\wedge) = \mathbb{R}^d$. If $d = 1$, then the restriction of S_k to $\mathcal{W}_2^\sim(\mathbb{R}^d, p)$ is injective for any $p \in \mathcal{P}_{FD}$.*

Proof Fix any $p \in \mathcal{P}_{\text{FD}}$. We need to show that for $[f]_\sim \in \mathcal{W}_2^\sim(\mathbb{R}^d, p)$, $S_k[f]_\sim = 0$ implies $[f]_\sim = 0$. From Proposition 11, we have

$$\begin{aligned} S_k[f]_\sim &= \int_{\mathbb{R}^d} \sum_{j=1}^d \partial_j k(x, \cdot) \partial_j f(x)\, p(x)\, dx = \int_{\mathbb{R}^d} \sum_{j=1}^d \partial_j \psi(x - \cdot) \partial_j f(x)\, p(x)\, dx \\ &= \int_{\mathbb{R}^d} \sum_{j=1}^d \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} i\omega_j \psi^\wedge(\omega) e^{i\langle \omega, \cdot - x\rangle}\, d\omega\, \partial_j f(x)\, p(x)\, dx \\ &= \int_{\mathbb{R}^d} \sum_{j=1}^d \phi_j(\omega) \psi^\wedge(\omega) e^{i\langle \omega, \cdot\rangle}\, d\omega \end{aligned}$$

where

$$\phi_j(\omega) := \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} (i\omega_j) e^{-i\langle \omega, x\rangle} \partial_j f(x)\, p(x)\, dx.$$

$S_k f = 0$ implies $\sum_{j=1}^d \phi_j(\omega)\psi^\wedge(\omega) = 0$ for all $\omega \in \mathbb{R}^d$. Since $\mathrm{supp}(\psi^\wedge) = \mathbb{R}^d$, we have $\sum_{j=1}^d \phi_j(\omega) = 0$ a.e., i.e., for ω-a.e.,

$$0 = \sum_{j=1}^d \int_{\mathbb{R}^d} (i\omega_j)\partial_j f(x)\, p(x)\, e^{-i\langle \omega, x\rangle}\, dx = \sum_{j=1}^d (i\omega_j)(p\partial_j f)^\wedge(\omega).$$

For $d = 1$, this implies $(\partial_j f)p = 0$ a.e. and so $\|f\|_{W_2} = 0$. ■

Examples of kernels that satisfy the conditions in Proposition B.5 include the Gaussian, Matérn (with $\beta > 1$) and inverse multiquadrics on $\mathbb{R}$.

References

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

A. Barron and C-H. Sheu. Approximation of density functions by sequences of exponential families. *Annals of Statistics*, 19(3):1347–1369, 1991.

F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007.

C. Bennett and R. Sharpley. *Interpolation of Operators*. Academic Press, Boston, 1988.

L. D. Brown. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. IMS, Hayward, CA, 1986.

S. Canu and A. J. Smola. Kernel methods and the exponential family. *Neurocomputing*, 69: 714–720, 2005.

A. Caponnetto and E. De Vito. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, 1991.

E. De Vito, L. Rosasco, and A. Toigo. Learning sets with separating kernels. *http://arxiv.org/abs/1204.3573*, April 2012.

R. DeVore, G. Kerkyacharian, D. Picard, and V. Temlyakov. Mathematical methods for supervised learning. IMI Preprints, University of South Carolina, 2004.

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

A. Doucet, N. de Freitas, and N. J. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York, 2001.

J. J. Duistermaat and J. A. C. Kolk. *Multidimensional Real Analysis II: Integration*. Cambridge University Press, Cambridge, UK, 2004.

H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley-Interscience, New York, 1999.

K. Fukumizu. Exponential manifold by reproducing kernel Hilbert spaces. In P. Gibilisco, E. Riccomagno, M.-P. Rogantin, and H. Winn, editors, *Algebraic and Geometric mothods in Statistics*, pages 291–306. Cambridge University Press, 2009.

K. Fukumizu, F. Bach, and M. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 489–496, Cambridge, MA, 2008. MIT Press.

K. Fukumizu, F. Bach, and M. Jordan. Kernel dimension reduction in regression. *Annals of Statistics*, 37(4):1871–1905, 2009.

K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14:3753–3783, 2013.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

C. Gu and C. Qiu. Smoothing spline density estimation: Theory. *Annals of Statistics*, 21 (1):217–234, 1993.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

A. Hyvärinen. Some extensions of score matching. *Computational Statistics & Data Analysis*, 51:2499–2512, 2007.

O. Johnson. *Information Theory and The Central Limit Theorem*. Imperial College Press, London, UK, 2004.

G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33:82–95, 1971.

C. Ley and Y. Swan. Stein's density approach and information inequalities. *Electronic Communications in Probability*, 18:1–14, 2013.

S. Lyu. Interpretation and generalization of score matching. In *Proc. 25^{th} conference on Uncertainty in Artificial Intelligence*, Montreal, Canada, 2009.

G. Pistone and C. Sempi. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Annals of Statistics*, 23(5):1543–1561, 1995.

M. Reed and B. Simon. *Functional Analysis.* Academic Press, New York, 1972.

L. Rosasco, E. De Vito, and A. Verri. Spectral methods for regularization in learning theory. Technical Report DISI-TR-05-18, DISI, Universitá degli Studi di Genova, 2005.

W. Rudin. *Functional Analysis.* McGraw-Hill, USA, 1991.

H. Sasaki, A. Hyvärinen, and M. Sugiyama. Clustering via mode seeking by direct estimation of the gradient of a log-density. In *Proc. of European Conference on Machine Learning*, pages 19–34, 2014.

B. Schölkopf, R. Herbrich, and A. Smola. A generalized representer theorem. In *Proceedings of the 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory*, pages 416–426, London, UK, 2001. Springer-Verlag.

S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153–172, 2007.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12: 2389–2410, 2011.

I. Steinwart and A. Christmann. *Support Vector Machines.* Springer, New York, 2008.

I. Steinwart and C. Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels and RKHSs. *Constructive Approximation*, 35:363–417, 2012.

H. Strathmann, D. Sejdinovic, S. Livingstone, Z. Szabó, and A. Gretton. Gradient-free Hamiltonian Monte Carlo with efficient kernel exponential families. In C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 955–963. Curran Associates, Inc., 2015.

S. Sun, M. Kolar, and J. Xu. Learning structured densities via infinite dimensional exponential families. In C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2278–2286. Curran Associates, Inc., 2015.

L. Tartar. *An Introduction to Sobolev Spaces and Interpolation Spaces.* Springer-Verlag, Berlin, 2007.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge, UK, 2000.

A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Annals of Statistics*, 36(3):1435–1463, 2008.

L. Wasserman. *All of Nonparametric Statistics*. Springer, 2006.

H. Wendland. *Scattered Data Approximation*. Cambridge University Press, Cambridge, UK, 2005.

D.-X. Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220:456–463, 2008.