1
Exponential manifold by reproducing kernel Hilbert spaces

1.1 Introduction

The purpose of this paper is to propose a method of constructing exponential families of Hilbert manifold, on which estimation theory can be built. Although there have been works on infinite dimensional exponential families of Banach manifolds [14, 10, 13], they are not appropriate for discussing statistical estimation with a finite sample; the likelihood function with a finite sample is not realized as a continuous function on the manifold.

The proposed exponential manifold uses a Reproducing Kernel Hilbert Space (RKHS) as a functional space in the construction. A RKHS is defined as a Hilbert space of functions such that evaluation of a function at an arbitrary point is a continuous functional on the Hilbert space. Since evaluation of the likelihood function is necessary for the estimation theory, it is very natural to use a manifold associated with a RKHS in defining an exponential family. Such a manifold can be either finite or infinite dimensional depending of the choice of RKHS.

This paper focuses on the Maximum Likelihood Estimation (MLE) with the exponential manifold associated with a RKHS. As in many non-parametric estimation methods, straightforward extension of MLE to an infinite dimensional exponential manifold suffers the problem of ill-posedness; the estimator is chosen from the infinite dimensional space, while only a finite number of constraints is given by the sample. To solve this problem, a pseudo-maximum likelihood method is proposed by restricting the infinite dimensional manifold to a series of finite dimensional submanifolds, which enlarge as the sample size increases. Some asymptotic results in the limit of infinite sample are shown, including the consistency of the pseudo-MLE.

This paper is an extended version of the previous conference paper [6].

1.2 Exponential family associated with a reproducing kernel Hilbert space

1.2.1 Reproducing kernel Hilbert space

This subsection provides a brief review of reproducing kernel Hilbert spaces. Only real Hilbert spaces are discussed in this paper, while a RKHS is defined as a complex Hilbert space in general. For the details on RKHS, see [1].

Let Ω be a set, and $\mathscr{H}$ be a Hilbert space included in the set of all real-valued functions on Ω. The inner product of $\mathscr{H}$ is denoted by $\langle\, , \,\rangle_{\mathscr{H}}$. The Hilbert space $\mathscr{H}$ is called a *reproducing kernel Hilbert space* (RKHS) if there is a function

$$k : \Omega \times \Omega \to \mathbb{R}$$

such that (i) $k(\cdot, x) \in \mathscr{H}$ for all $x \in \Omega$, and (ii) for any $f \in \mathscr{H}$ and $x \in \Omega$,

$$\langle f, k(\cdot, x)\rangle_{\mathscr{H}} = f(x)$$

The condition (ii) is called the *reproducing property* and k is called a *reproducing kernel.*

A reproducing kernel is symmetric, because $k(x, y) = \langle k(\cdot, y), k(\cdot, x)\rangle_{\mathscr{H}} = \langle k(\cdot, x), k(\cdot, y)\rangle_{\mathscr{H}} = k(y, x)$. It is easy to see that a reproducing kernel is unique if it exists. The following proposition is a characterization of RKHS.

Proposition 1 *A Hilbert space of functions on Ω is a RKHS if and only if the evaluation mapping $e_x : \mathscr{H} \to \mathbb{R}$, $f \mapsto f(x)$, is a continuous linear functional on $\mathscr{H}$ for any $x \in \Omega$.*

Proof Suppose $k : \Omega \times \Omega \to \mathbb{R}$ is a reproducing kernel of $\mathscr{H}$. For any $x \in \Omega$ and $f \in \mathscr{H}$, we have $|e_x(f)| = |f(x)| = |\langle f, k(\cdot, x)\rangle_{\mathscr{H}}| \leq \|f\|_{\mathscr{H}} \|k(\cdot, x)\|_{\mathscr{H}} = \|f\|_{\mathscr{H}} \sqrt{k(x, x)}$, which shows e_x is bounded. Conversely, if the evaluation mapping e_x is bounded, by Riesz's representation theorem, there exists $\phi_x \in \mathscr{H}$ such that $f(x) = e_x(f) = \langle f, \phi_x\rangle_{\mathscr{H}}$. The function $k(y, x) = \phi_x(y)$ is then a reproducing kernel on $\mathscr{H}$. □

A function $k : \Omega \times \Omega \to \mathbb{R}$ is said to be *positive definite* if it is symmetric, $k(x, y) = k(y, x)$ for any $x, y \in \Omega$, and for any points $x_1, \ldots, x_n \in \Omega$ the symmetric matrix $(k(x_i, x_j))_{i,j}$ is positive semidefinite, i.e., for any real numbers $c_1, \ldots, c_n$ the inequality $\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$ holds.

A RKHS and a positive definite kernel have one-to-one correspondence. If $\mathscr{H}$ is a RKHS on Ω, the reproducing kernel $k(x, y)$ is positive definite, because $\sum_{i,j} c_i c_j k(x_i, x_j) = \|\sum_i c_i k(\cdot, x_i)\|_{\mathscr{H}}^2 \geq 0$. It is also known ([1]) that for a positive definite kernel k on Ω there uniquely exists a RKHS $\mathscr{H}_k$

such that $\mathscr{H}_k$ consists of functions on Ω, the class of functions $\sum_{i=1}^m a_i k(\cdot, x_i)$ ($m \in \mathbb{N}, x_i \in \Omega, a_i \in \mathbb{R}$) is dense in $\mathscr{H}_k$, and $\langle f, k(\cdot, x)\rangle_{\mathscr{H}_k} = f(x)$ holds for any $f \in \mathscr{H}_k$ and $x \in \Omega$. Thus, a Hilbert space $\mathscr{H}$ of functions on Ω is a RKHS if and only if $\mathscr{H} = \mathscr{H}_k$ for some positive definite kernel k. In practice, a RKHS is often given by a positive definite kernel.

Many functions are known to be positive definite. On $\mathbb{R}^n$, the basic examples include the linear kernel $k(x, y) = x^T y$, and more generally, the polynomial kernel $k(x, y) = (x^T y + c)^d$ ($c \geq 0, d \in \mathbb{N}$). The RKHS defined by the linear kernel is isomorphic to the n-dimensional Euclidean space. The RKHS given by $(x^T y + c)^d$ ($c > 0$) is the polynomials of degree d or less as a vector space. It is also known that the shift-invariant kernel $\exp(-|x-y|^p)$ on $\mathbb{R}$ is positive definite for $0 < p \leq 2$. The positive definite kernel $\exp(-\frac{1}{2\sigma^2}|x-y|^2)$ ($\sigma > 0$) is often referred to as Gaussian RBF kernel, and the associated RKHS is infinite dimensional. When $p = 1$, it is known that the RKHS defined by $\exp(-|x-y|)$ is the Sobolev space $H^1(\mathbb{R}) = \{u \in L^2(\mathbb{R}) \mid \text{ there exists } u' \in L^2(\mathbb{R}) \text{ such that } u(x) = \int_{-\infty}^x u'(y)dy\}$. Many examples of positive definite kernels and the associated RKHSs are shown in [3] and [2].

It is also important to note that if a positive definite kernel k on a topological space is continuous, all the functions in $\mathscr{H}_k$ are continuous. This is easily seen from $|f(x) - f(y)| = |\langle f, k(\cdot, x) - k(\cdot, y)\rangle_{\mathscr{H}_k}| \leq \|f\|_{\mathscr{H}_k}(k(x, x) + k(y, y) - 2k(x, y))^{1/2}$.

1.2.2 Exponential manifold associated with a RKHS

Let Ω be a topological space, and μ be a Borel probability measure on Ω. The *support* of μ is defined by the smallest closed set F such that $\mu(\Omega \backslash F) = 0$. Throughout this paper, it is assumed that the support of μ is Ω. The set of positive probability density functions with respect to μ is denoted by

$$\mathscr{M}_\mu = \left\{ f : \Omega \to \mathbb{R} \;\middle|\; f > 0 \text{ almost everywhere-}\mu\text{, and } \int_\Omega f d\mu = 1 \right\}.$$

Hereafter, the probability given by the density $f \in \mathscr{M}_\mu$ is denoted by $f\mu$, and the expectation of a measurable function on Ω with respect to $f\mu$ is denoted by $E_f[u]$ or $E_f[u(X)]$.

Let $k : \Omega \times \Omega \to \mathbb{R}$ be a continuous positive definite kernel on Ω. Define a subclass of $\mathscr{M}_\mu$ by

$$\mathscr{M}_\mu(k) = \left\{ f \in \mathscr{M}_\mu^c \;\middle|\; \text{there exists } \delta > 0 \text{ such that } \int e^{\delta\sqrt{k(x,x)}} f(x) d\mu(x) < \infty \right\}.$$

A positive definite kernel k is bounded if and only if the function $k(x,x)$ on Ω is bounded, since $|k(x,y)| \leq k(x,x)k(y,y)$ by the positive semidefiniteness. For bounded k, we have $\mathscr{M}_\mu(k) = \mathscr{M}_\mu$.

It is also worth noting that $f \in \mathscr{M}_\mu(k)$ if and only if the function $x \mapsto \sqrt{k(x,x)}$ belongs to the Orlicz space $L^{\cosh -1}(f)$, which is used for constructing the Banach manifold by [14]. In fact, $L^{\cosh -1}(f)$ is defined by the class of function u for which there is $\alpha > 0$ such that

$$E_f\Big[\cosh\Big(\frac{u}{\alpha}\Big) - 1\Big] < \infty.$$

Throughout this paper, the following assumption is made unless otherwise mentioned;

(A) The RKHS $\mathscr{H}_k$ contains the constant functions.

This is a mild assumption, because for any RKHS $\mathscr{H}_k$ the direct sum $\mathscr{H}_k + \mathbb{R}$, where $\mathbb{R}$ denotes the RKHS associated with the positive definite kernel 1 on Ω, is again a RKHS with reproducing kernel $k(x,y)+1$ ([1]). This assumption is made so that subtracting a constant may be operated within $\mathscr{H}_k$.

For any $f \in \mathscr{M}_\mu(k)$, the expectation $E_f[\sqrt{k(X,X)}]$ is finite, because $\delta E_f[\sqrt{k(X,X)}] \leq E_f[e^{\delta\sqrt{k(X,X)}}] < \infty$. From $|u(x)| = |\langle u, k(\cdot,x)\rangle_{\mathscr{H}_k}| \leq \sqrt{k(x,x)}\|u\|_{\mathscr{H}_k}$, the mapping $u \mapsto E_f[u(X)]$ is a bounded functional on $\mathscr{H}_k$ for any $f \in \mathscr{M}_\mu(k)$. We define a closed subspace T_f of $\mathscr{H}_k$ by

$$T_f := \{u \in \mathscr{H}_k \mid E_f[u(X)] = 0\},$$

which works as a tangent space at f, as we will see later. Note that, by the assumption (A), $u - E_f[u]$ is included in T_f for any $u \in \mathscr{H}_k$.

For $f \in \mathscr{M}_\mu(k)$, let $\mathscr{W}_f$ be a subset of T_f defined by

$$\mathscr{W}_f = \big\{u \in T_f \,\big|\, \text{there exists } \delta > 0 \text{ such that } E_f[e^{\delta\sqrt{k(X,X)}+u(X)}] < \infty\big\}.$$

The cumulant generating function Ψ_f on $\mathscr{W}_f$ is defined by

$$\Psi_f(u) = \log E_f[e^{u(X)}].$$

Lemma 1 *For any $u \in \mathscr{W}_f$, the probability density function*

$$e^{u-\Psi_f(u)}f$$

belongs to $\mathscr{M}_\mu(k)$.

Proof It is obvious that $\Psi(u)$ is finite for any $u \in \mathscr{W}_f$, so that the above probability density function is well-defined. By the definition of $\mathscr{W}_f$, there is $\delta > 0$ such that $E_f[e^{\delta\sqrt{k(X,X)}+u(X)}] < \infty$, which derives

$$\int e^{\delta\sqrt{k(x,x)}} e^{u(x)-\Psi_f(u)} f(x) d\mu(x) = e^{-\Psi_f(u)} E_f\left[e^{\delta\sqrt{k(X,X)}+u(X)}\right] < \infty.$$

This implies $e^{u-\Psi_f(u)} f \in \mathscr{M}_\mu(k)$. □

From Lemma 1, the mapping

$$\xi_f : \mathscr{W}_f \to \mathscr{M}_\mu(k), \qquad u \mapsto e^{u-\Psi_f(u)} f$$

is well-defined. The map ξ_f is one-to-one, because $\xi_f(u) = \xi_f(v)$ implies $u - v$ is constant, which is necessarily zero from $E_f[u] = E_f[v] = 0$, since u is continuous and the support of $f\mu$ is Ω †.

Let $\mathscr{S}_f = \xi_f(\mathscr{W}_f)$, and φ_f be the inverse of ξ_f, that is,

$$\varphi_f : \mathscr{S}_f \to \mathscr{W}_f, \qquad g \mapsto \log\frac{g}{f} - E_f\left[\log\frac{g}{f}\right].$$

It will be shown that φ_f works as a local coordinate that makes $\mathscr{M}_\mu(k)$ a Hilbert manifold. The following facts are basic;

Lemma 2 *Let f and g be arbitrary elements in $\mathscr{M}_\mu(k)$.*

(i) $\mathscr{W}_f$ is an open subset of T_f.

(ii) $g \in \mathscr{S}_f$ if and only if $\mathscr{S}_g = \mathscr{S}_f$.

Proof (i). For an arbitrary $u \in \mathscr{W}_f$, take $\delta > 0$ so that $E_f[e^{u(X)+\delta\sqrt{k(X,X)}}] < \infty$. Define an open neighborhood V_u of u in T_f by $V_u = \{v \in T_f \mid \|v-u\|_{\mathscr{H}_k} < \delta/2\}$. Then, for any $v \in V_u$,

$$\begin{aligned} E_f\left[e^{(\delta/2)\sqrt{k(X,X)}+v(X)}\right] &= E_f\left[e^{(\delta/2)\sqrt{k(X,X)}+\langle v-u,k(\cdot,X)\rangle_{\mathscr{H}_k}+u(X)}\right] \\ &\le E_f\left[e^{(\delta/2)\sqrt{k(X,X)}+\|v-u\|_{\mathscr{H}_k}\sqrt{k(X,X)}+u(X)}\right] \\ &\le E_f\left[e^{\delta\sqrt{k(X,X)}+u(X)}\right] \quad < \infty, \end{aligned}$$

which implies $\mathscr{W}_f$ is open.

(ii). "If" part is obvious. For the "only if" part, we first prove $\mathscr{S}_g \subset \mathscr{S}_f$ on condition that $g \in \mathscr{S}_f$. Let h be an arbitrary element in $\mathscr{S}_g$, and take

† The continuity assumption on k is made to guarantee the injectiveness of ξ_f. With an almost-everywhere positive density function f, it is obvious that two density functions $e^{u(x)-\Psi_f(u)}f(x)$ and $e^{v(x)-\Psi_f(v)}f(x)$ define the same probability if and only if $u - v$ is constant almost everywhere with respect to $f\mu$. We wish to further guarantee, however, that $u - v$ is exactly constant, because a function is identified as the zero element in a RKHS only if it is exactly zero. We thus assume that the functions in $\mathscr{H}_k$ are continuous and the support of μ is Ω.

$u \in \mathscr{W}_f$ and $v \in \mathscr{W}_g$ such that $g = e^{u-\Psi_f(u)}f$ and $h = e^{v-\Psi_g(v)}g$. From the fact $g \in \mathscr{W}_f$, there is $\delta > 0$ such that $E_g[e^{v(X)+\delta\sqrt{k(X,X)}}] < \infty$. We have $\int e^{v(x)+u(x)+\delta\sqrt{k(x,x)}-\Psi_f(u)} f(x) d\mu(x) < \infty$, which means $v + u - E_f[v] \in \mathscr{W}_f$. From $h = e^{(v+u-E_f[v])-(\Psi_f(u)+\Psi_g(v)-E_f[v])}f$, we have $\Psi_f(v + u - E_f[v]) = \Psi_f(u) + \Psi_g(v) - E_f[v]$ and $h = \xi_f(v + u - E_f[v]) \in \mathscr{S}_f$.

For the opposite inclusion, it suffices to show $f \in \mathscr{S}_g$. Let $\gamma > 0$ be a constant so that $E_f[e^{\gamma\sqrt{k(X,X)}}] < \infty$. From $e^{-u}g = e^{-\Psi_f(u)}f$, we see $\int e^{\gamma\sqrt{k(x,x)}-u(x)} g(x) d\mu(x) < \infty$, which means $-u + E_g[u] \in \mathscr{W}_g$. It follows that $f = e^{-u+\Psi_f(u)}g = e^{(-u+E_g[u])-(-\Psi_f(u)+E_g[u])}g$ means $f = \xi_g(-u + E_g[u]) \in \mathscr{S}_g$. □

The map φ_f defines a structure of Hilbert Manifold on $\mathscr{M}_\mu(k)$, which we call *reproducing kernel exponential manifold.*

Theorem 1 *The system $\{(\mathscr{S}_f, \varphi_f)\}_{f\in\mathscr{M}_\mu(k)}$ is a C^∞-atlas of $\mathscr{M}_\mu(k)$, that is,*

(i) If $\mathscr{S}_f \cap \mathscr{S}_g \neq \emptyset$, then $\varphi_f(\mathscr{S}_f \cap \mathscr{S}_g)$ is an open set in T_f.

(ii) If $\mathscr{S}_f \cap \mathscr{S}_g \neq \emptyset$, then

$$\varphi_g \circ \varphi_f^{-1}|_{\varphi_f(\mathscr{S}_f\cap\mathscr{S}_g)} : \varphi_f(\mathscr{S}_f \cap \mathscr{S}_g) \to \varphi_g(\mathscr{S}_f \cap \mathscr{S}_g)$$

is a C^∞ map.

Thus, $\mathscr{M}_\mu(k)$ admits a structure of C^∞-Hilbert manifold.

Proof The assertion (i) is obvious, because $\mathscr{S}_f \cap \mathscr{S}_g \neq \emptyset$ means $\mathscr{S}_f = \mathscr{S}_g$ from Lemma 2. Suppose $\mathscr{S}_f \cap \mathscr{S}_g \neq \emptyset$, that is, $\mathscr{S}_f = \mathscr{S}_g$. For any $u \in \mathscr{W}_f$,

$$\begin{aligned}\varphi_g \circ \varphi_f^{-1}(u) = \varphi_g\big(e^{u-\Psi_f(u)}f\big) &= \log \frac{e^{u-\Psi_f(u)}f}{g} - E_g\Big[\log \frac{e^{u-\Psi_f(u)}f}{g}\Big] \\ &= u + \log(f/g) - E_g\big[u + \log(f/g)\big],\end{aligned}$$

from which the assertion (ii) is obtained, because $u \mapsto E_g[u]$ is of C^∞ on $\mathscr{W}_f$.

It is known that with the assertions (i) and (ii) a topology is introduced on $\mathscr{M}_\mu(k)$ so that all $\mathscr{S}_f$ are open, and $\mathscr{M}_\mu(k)$ is equipped with the structure of C^∞-Hilbert manifold (see [12]). □

The open set $\mathscr{S}_f$ is regarded as a maximal exponential family in $\mathscr{M}_\mu(k)$. In fact, we have the following

Theorem 2 *For any $f \in \mathscr{M}_\mu(k)$,*

$$\mathscr{S}_f = \{g \in \mathscr{M}_\mu(k) \mid \text{ there exists } u \in T_f \text{ such that } g = e^{u-\Psi_f(u)}f\}.$$

Proof It suffices to show that $g = e^{u-\Psi_f(u)}f$ in the right hand side is included in the left hand side, as the opposite inclusion is obvious. From $g \in \mathscr{M}_\mu(k)$, there is $\delta > 0$ such that $E_g[e^{\delta\sqrt{k(X,X)}}] < \infty$, which means $E_f[e^{\delta\sqrt{k(X,X)}+u(X)}] < \infty$. Therefore, $u \in \mathscr{W}_f$ and $g = \xi_f(u) \in \mathscr{S}_f$. □

From Lemma 2 (ii), we can define an equivalence relation such that f and g are equivalent if and only if they are in the same local maximal exponential family, that is, if and only if $\mathscr{S}_f \cap \mathscr{S}_g \neq \emptyset$. Let $\{\mathscr{S}^{(\lambda)}\}_{\lambda\in\Lambda}$ be the equivalence classes. Then, they are equal to the set of connected components.

Theorem 3 *Let $\{\mathscr{S}^{(\lambda)}\}_{\lambda\in\Lambda}$ be the equivalence classes of the maximum local exponential families described above. Then, $\{\mathscr{S}^{(\lambda)}\}_{\lambda\in\Lambda}$ are the connected components of $\mathscr{M}_\mu(k)$. Moreover, each component $\mathscr{S}^{(\lambda)}$ is simply connected.*

Proof From Lemma 2 and Theorem 1, $\{\mathscr{S}^{(\lambda)}\}_{\lambda\in\Lambda}$ are disjoint open covering of $\mathscr{M}_\mu(k)$. The proof is completed if every $\mathscr{W}_f$ is shown to be convex. Let u_0 and u_1 be arbitrary elements in $\mathscr{W}_f$. Then, there exists $\delta > 0$ such that $E_f\big[e^{\delta\sqrt{k(X,X)}+u_0(X)}\big] < \infty$ and $E_f\big[e^{\delta\sqrt{k(X,X)}+u_1(X)}\big] < \infty$. For $u_t = tu_1 + (1-t)u_0 \in T_f$ $(t \in [0,1])$, we have $e^{u_t(x)} \le te^{u_1(x)} + (1-t)e^{u_0(x)}$ by the convexity of $z \mapsto e^z$. It leads

$$\begin{aligned} &E_f\Big[e^{\delta\sqrt{k(X,X)}+u_t(X)}\Big] \\ &\qquad \le tE_f\Big[e^{\delta\sqrt{k(X,X)}+u_1(X)}\Big] + (1-t)E_f\Big[e^{\delta\sqrt{k(X,X)}+u_0(X)}\Big] < \infty, \end{aligned}$$

which means $u_t \in \mathscr{W}_f$. □

The Hilbert space $\mathscr{H}_k$, which is used for giving the manifold structure to $\mathscr{M}_\mu(k)$, has stronger topology than the Orlicz space $L^{\cosh -1}(f)$. Recall that the norm of $u \in L^{\cosh -1}(f)$ is defined by

$$\|u\|_{L^{\cosh -1}(f)} = \inf\Big\{\alpha > 0 \;\Big|\; E_f\Big[\cosh\Big(\frac{u}{\alpha}\Big) - 1\Big] \le 1\Big\}.$$

Proposition 2 *For any $f \in \mathscr{M}_\mu(k)$, the RKHS $\mathscr{H}_k$ is continuously included in $L^{\cosh -1}(f)$. Moreover, if a positive number A_f is defined by*

$$A_f = \inf\Big\{\alpha > 0 \;\Big|\; \int e^{\frac{\sqrt{k(x,x)}}{\alpha}} f(x)d\mu(x) \le 2\Big\},$$

then for any $u \in \mathscr{H}_k$

$$\|u\|_{L^{\cosh -1}(f)} \leq A_f \|u\|_{\mathscr{H}_k}.$$

Proof From the inequality

$$\begin{aligned} E_f\big[\cosh(u(X)/\alpha) - 1\big] &\leq E_f\big[e^{|u(X)|/\alpha}\big] - 1 \\ &\leq E_f\Big[e^{\frac{1}{\alpha}\|u\|_{\mathscr{H}_k}\sqrt{k(X,X)}}\Big] - 1, \end{aligned}$$

if $\|u\|_{\mathscr{H}_k}/\alpha < 1/A_f$, then $E_f[\cosh(u/\alpha) - 1] \leq 1$. This means $A_f\|u\|_{\mathscr{H}_k} \geq \|u\|_{L^{\cosh -1}(f)}$. □

Proposition 2 tells that the manifold $\mathscr{M}_\mu(k)$ is a subset of the maximum exponential manifold. However, the former is not necessarily a submanifold of the latter, because $\mathscr{H}_k$ is not a closed subspace of $L^{\cosh -1}(f)$ in general. Note also that $L^{\cosh -1}(f)$ is continuously embedded in $L^p(f)$ for all $p \geq 1$. Thus, $E_f|u|^p$ is finite for any $f \in \mathscr{M}_\mu(k)$, $u \in \mathscr{H}_k$, and $p \geq 1$.

The reproducing kernel exponential manifold and its connected components depend on the underlying RKHS. It may be either finite or infinite dimensional. A different choice of the positive definite kernel results in a different exponential manifold. A connected component of $\mathscr{M}_\mu(k)$ in Theorem 3 is in general smaller than the maximal exponential model discussed in [14].

1.2.3 Mean and covariance on reproducing kernel exponential manifolds

As in the case of finite dimensional exponential families and the exponential manifold by [14], the derivatives of the cumulant generating function provide the cumulants or moments of the random variables given by tangent vectors. Let $f \in \mathscr{M}_\mu(k)$ and $v_1, \ldots, v_d \in T_f$. The d-th Fréchet derivative of Ψ_f in the directions $v_1, \ldots, v_d$ at $f_u = e^{u-\Psi_f(u)}f$ is denoted by $D_u^d\Psi_f(v_1, \ldots, v_d)$. From Proposition 2 and the known results on the derivatives for the maximal exponential manifolds [14, 5], the Ψ_f is C^∞-Fréchet differentiable on $\mathscr{M}_\mu(k)$, and in particular, we have

$$D_u\Psi_f(v) = E_{f_u}[v], \qquad D_u^2\Psi_f(v_1, v_2) = \mathrm{Cov}_{f_u}[v_1(X), v_2(X)],$$

where $\mathrm{Cov}_g[v_1, v_2] = E_g[v_1(X)v_2(X)] - E_g[v_1(X)]E_g[v_2(X)]$ is the covariance of v_1 and v_2 under the probability $g\mu$.

The first and second moments are expressed also by an element and an operator of the Hilbert space. Let P be a probability on Ω such that

$E_P[\sqrt{k(X,X)}] < \infty$. Because the functional $\mathscr{H}_k \ni u \mapsto E_P[u(X)]$ is bounded, there exists $m_P \in \mathscr{H}_k$ such that

$$E_P[u(X)] = \langle u, m_P \rangle_{\mathscr{H}_k}$$

for all $u \in \mathscr{H}_k$. We call m_P the *mean element* for P. Noticing that the mapping $\mathscr{H}_k \times \mathscr{H}_k \ni (v_1, v_2) \mapsto \mathrm{Cov}_P[v_1(X), v_2(X)]$ is a bounded bilinear form, we see that there uniquely exists a bounded operator Σ_P on $\mathscr{H}_k$ such that

$$\mathrm{Cov}_P[v_1(X), v_2(X)] = \langle v_1, \Sigma_P v_2 \rangle_{\mathscr{H}_k}$$

holds for all $v_1, v_2 \in \mathscr{H}_k$. The operator Σ_P is called the *covariance operator* for P. For the detail of covariance operators on a RKHS, see [7].

When a local coordinate $(\varphi_{f_0}, \mathscr{S}_{f_0})$ in a reproducing kernel exponential manifold $\mathscr{M}_\mu(k)$ is assumed, the notations m_u and Σ_u are also used for the mean element and covariance operator, respectively, with respect to the probability density $f_u = e^{u - \Psi_{f_0}(u)} f_0$. We have

$$D_u \Psi_f(v) = \langle m_u, v \rangle_{\mathscr{H}_k}, \qquad D_u^2 \Psi_f(v_1, v_2) = \langle v_1, \Sigma_u v_2 \rangle_{\mathscr{H}_k}.$$

The mapping $\mathscr{W}_f \ni u \mapsto m_u \in \mathscr{H}_k$ is locally one-to-one, because the derivative $\Sigma_u|_{T_{f_0}}$ is strictly positive for non-degenerate μ. The element m_u is equal to the *mean parameter* ([13]) for the density f_u by identifying the bounded linear functional $D_u \Psi_f$ with the element of T_f.

The mean element $m_P(y)$ as a function is explicitly expressed by

$$m_P(y) = E_P[k(X, y)]$$

from $m_P(y) = \langle m_P, k(\cdot, y) \rangle_{\mathscr{H}_k} = E_P[k(X, y)]$. The operator Σ_u is an extension of Fisher information matrix.

It is interesting to ask when the mean element specifies a probability.

Definition 1 *Let $(\Omega, \mathscr{B})$ be a measurable space, and k be a measurable positive definite kernel on Ω such that $\int k(x,x) dP(x)$ is finite for any probability P on $(\Omega, \mathscr{B})$. The kernel k is called* characteristic *if the mapping $P \mapsto m_P$ uniquely determines a probability.*

It is known that Gaussian RBF kernels and Laplacian kernels are characteristic on $\mathbb{R}^n$ and any compact subset in $\mathbb{R}^n$ with Borel σ-field ([9, 15]). If $k(x, y) = \exp(-|x - y|)$ is used for defining $\mathscr{M}_k(\mu)$ on the unit interval $[0, 1]$ with the uniform distribution μ, then the mean parameter m_u uniquely determines a probability in $\mathscr{M}_\mu$.

1.2.4 Kullback-Leibler divergence

Let $f_0 \in \mathscr{M}_\mu(k)$ and $u, v \in \mathscr{W}_{f_0}$. With the local coordinate $(\varphi_{f_0}, \mathscr{S}_{f_0})$, it is easy to see that the Kullback-Leibler divergence from $f_u = e^{u-\Psi_{f_0}(u)} f_0$ to $f_v = e^{v-\Psi_{f_0}(v)} f_0$ is given by

$$\mathrm{KL}(f_u || f_v) = \Psi_{f_0}(v) - \Psi_{f_0}(u) - \langle v - u, m_u \rangle_{\mathscr{H}_k}. \qquad (1.1)$$

Let f_u, f_v, and f_w be points in $\mathscr{S}_{f_0}$. It is straightforward to see

$$\mathrm{KL}(f_u || f_w) = \mathrm{KL}(f_u || f_v) + \mathrm{KL}(f_v || f_w) - \langle w - v, m_u - m_v \rangle_{\mathscr{H}_k}. \qquad (1.2)$$

Let U be a closed subspace of T_{f_0}, and $\mathcal{V} = U \cap \mathscr{W}_{f_0}$. The subset $\mathcal{N} = \varphi_{f_0}^{-1}(\mathcal{V})$ is a submanifold of $\mathscr{S}_{f_0}$, which is also an exponential family. Let $f_* = e^{u_* - \Psi_{f_0}(u_*)}$ be a point in $\mathscr{S}_{f_0}$, and consider the minimizer of the KL divergence from f_* to a point in $\mathcal{N}$;

$$u_{opt} = \arg\min_{u \in \mathcal{V}} \mathrm{KL}(f_* || f_u). \qquad (1.3)$$

Theorem 4 *Under the assumption that the minimizer u_{opt} of Eq.(1.3) exists, the orthogonal relation*

$$\langle u - u_{opt}, m_{u_*} - m_{u_{opt}} \rangle_{\mathscr{H}_k} = 0. \qquad (1.4)$$

and the Pythagorean equation

$$\mathrm{KL}(f_* || f_u) = \mathrm{KL}(f_* || f_{u_{opt}}) + \mathrm{KL}(f_{u_{opt}} || f_u) \qquad (1.5)$$

hold for any $u \in \mathcal{V}$.

Proof Since $\mathscr{W}_{f_0}$ is an open convex set, $u_t = t(u - u_{opt}) + u_{opt}$ lies in $\mathscr{W}_{f_0}$ for all $t \in (-\delta, \delta)$ with sufficiently small $\delta > 0$. From Eq. (1.2), $\mathrm{KL}(f_* || f_{u_t})$ is differentiable with respective to t, and $\frac{d}{dt}\mathrm{KL}(f_* || f_{u_t})|_{t=0} = 0$ by the minimality. This derives

$$\langle u - u_{opt}, m_{u_{opt}} \rangle_{\mathscr{H}_k} - \langle u - u_{opt}, m_{u_*} \rangle_{\mathscr{H}_k} = 0,$$

which is the orthogonal relation. Pythagorean relation is obvious from Eqs.(1.2) and (1.4). □

1.3 Pseudo maximum likelihood estimation with $\mathscr{M}_\mu(k)$

In this section, statistical estimation with a reproducing kernel exponential manifold is discussed. Throughout this section, a positive definite kernel k with the assumption (A) and a connected component $\mathscr{S}$ of $\mathscr{M}_\mu(k)$ are fixed.

From Lemma 2 and Theorem 2, for any $f_0 \in \mathscr{S}$ the component $\mathscr{S}$ can be expressed by

$$\mathscr{S} = \{f \in \mathscr{M}_\mu(k) \mid f = e^{u-\Psi_0(u)} f_0 \text{ for some } u \in T_{f_0}\},$$

where Ψ_0 is an abbreviation of Ψ_{f_0}. For notational simplicity, $\mathscr{W}_0 = \mathscr{W}_{f_0}$ and $f_u = e^{u-\Psi_0(u)} f_0$ for $u \in \mathscr{W}_0$ are used.

It is assumed that $(X_1, X_2, \ldots, X_n)$ is an i.i.d. sample with probability $f_* \mu$ with $f_* \in \mathscr{S}$, which is called a true probability density. We discuss the problem of estimating f_* with the statistical model $\mathscr{S}$ given the finite sample.

1.3.1 Likelihood equation on a reproducing kernel exponential manifold

The maximum likelihood estimation (MLE) is the most popular estimation method for finite dimensional exponential families. In the following, we consider the MLE approach with the reproducing kernel exponential manifold $\mathscr{S}$, which may not be finite dimensional. The objective function of MLE with $\mathscr{S}$ is given by

$$\sup_{u \in \mathscr{W}_0} L_n(u), \qquad L_n(u) = \frac{1}{n} \sum_{i=1}^{n} u(X_i) - \Psi_0(u),$$

where $L_n(u)$ is called the log likelihood function. By introducing the empirical mean element

$$\widehat{m}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i),$$

the log likelihood function is rewritten by

$$L_n(u) = \langle \widehat{m}^{(n)}, u \rangle_{\mathscr{H}_k} - \Psi_0(u).$$

Taking the partial derivative of $L_n(u)$, we obtain the likelihood equation,

$$\langle \widehat{m}^{(n)}, v \rangle_{\mathscr{H}_k} = \langle m_u, v \rangle_{\mathscr{H}_k} \qquad (\forall v \in \mathscr{H}_k), \tag{1.6}$$

where m_u is the mean parameter corresponding to the density f_u. Note that the above equation holds not only for $v \in T_{f_0}$ but for all $v \in \mathscr{H}_k$, since $\langle \widehat{m}^{(n)} - m_u, 1 \rangle_{\mathscr{H}_k}$ always vanishes. The log likelihood equation is thus reduced to

$$m_u = \widehat{m}^{(n)}, \tag{1.7}$$

that is, the mean parameter for the maximum likelihood estimator shoudl be the empirical mean element $\widehat{m}^{(n)}$.

If $\mathscr{H}_k$ is finite dimensional and $(\phi_1, \ldots, \phi_d)$ is a basis of T_{f_0}, Eq. (1.7) is equivalent to

$$m_u^j = \frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i) \qquad (j = 1, \ldots, d),$$

where $(m_u^1, \ldots, m_u^d)$ is the component of m_u with respect to the basis $(\phi_1, \ldots, \phi_d)$. If the mapping $u \mapsto m_u$ is invertible, which is often the case with ordinary finite dimensional exponential families, the MLE $\widehat{u}$ is given by the inverse image of $\widehat{m}^{(n)}$.

Unlike the finite dimensional exponential family, the likelihood equation Eq. (1.7) does not necessarily have a solution in the canonical parameter u. As [13] points out for their exponential manifold, the inverse mapping from the mean parameter to the canonical parameter u is not bounded in general. For reproducing kernel exponential manifolds, the unboundedness of the inverse of $u \mapsto m_u$ can been seen by investigating its derivative. The derivative of the map $u \mapsto m_u$ is given by the covariance operator Σ_u, which is known to be of trace class ([7], Section 4.2). In fact, it is easy to see $\mathrm{Tr}[\Sigma_u] = E_{f_u}[\|k(\cdot, X) - m_u\|^2_{\mathscr{H}_k}] = E[k(X, X)] - E[k(X, \tilde{X})]$, where X and $\tilde{X}$ are independent variables with the same distribution $f_u\mu$. If $\mathscr{H}_k$ is infinite dimensional, Σ_u has arbitrary small positive eigenvalues, which implies Σ_u does not have a bounded inverse. Thus, the mean parameter does not give a coordinate system for infinite dimensional manifolds.

If k is characteristic, another explanation by the moment matching is possible to the fact that the likelihood equation does not have a solution. From Eq.**??**1.7), the empirical distribution $\frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ and the probability $e^{u - \Psi_0(u)} f_0\mu$ must have the same mean element. For a characteristic k, however, these two probabilities must be the same; this is impossible if the support of μ is uncountable.

To solve this problem, a method of pseudo maximum likelihood estimation will be proposed in Section 1.3.3, in which asymptotic properties of the mean parameter yet play an important role.

1.3.2 $\sqrt{n}$-consistency of the mean parameter

Although the mean parameter does not give a local coordinate of $\mathscr{M}_\mu(k)$, it is useful to analyze the asymptotic behavior of an estimator based on the likelihood approach. The next theorem establishes $\sqrt{n}$-consistency of the mean parameter in a general form. While this is a known result ([2], Lemma 22, Section 9.1), the proof is shown for completeness.

Theorem 5 *Let $(\Omega, \mathcal{B}, P)$ be a probability space, $k : \Omega \times \Omega \to \mathbb{R}$ be a positive definite kernel so that $E_P[k(X, X)] < \infty$, and $m_P \in \mathscr{H}_k$ be the mean element with respect to P. Suppose $X_1, \dots, X_n$ are i.i.d. sample from P, and define the empirical mean element $\widehat{m}^{(n)}$ by $\widehat{m}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} k(\cdot, X_i)$. Then, we have*

$$\|\widehat{m}^{(n)} - m_P\|_{\mathscr{H}_k} = O_p\bigl(1/\sqrt{n}\bigr) \quad (n \to \infty).$$

Proof Let $E_X[\cdot]$ denote the expectation with respect to the random variable X which follows P. Suppose $X, \tilde{X}, X_1, \dots, X_n$ are i.i.d. We have

$$\begin{aligned}
E\|\widehat{m}^{(n)} - m_P\|^2_{\mathscr{H}_k} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E_{X_i}E_{X_j}[k(X_i, X_j)] \\
&\quad - \frac{2}{n}\sum_{i=1}^{n} E_{X_i}E_X[k(X_i, X)] + E_X E_{\tilde{X}}[k(X, \tilde{X})] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} E[k(X_i, X_j)] + \frac{1}{n}E_X[k(X, X)] - E_X E_{\tilde{X}}[k(X, \tilde{X})] \\
&= \frac{1}{n}\{E_X[k(X, X)] - E_X E_{\tilde{X}}[k(X, \tilde{X})]\} \\
&= O(1/n).
\end{aligned}$$

The assertion is obtained by Chebyshev's inequality. □

It is further known ([2], Section 9.1) that $\sqrt{n}(\widehat{m}^{(n)} - m_P)$ converges in law to a Gaussian distribution on $\mathscr{H}_k$.

1.3.3 Pseudo maximum likelihood estimation

This subsection proposes the pseudo maximum likelihood estimation using a series of finite dimensional subspaces in $\mathscr{H}_k$ to make the inversion from the mean parameter to the canonical parameter possible. With an infinite dimensional reproducing kernel exponential manifold, the estimation of the true density with a finite sample is an ill-posed problem, as discussed in Section 1.3.1. Among many methods of regularization to solve such ill-posed problems, one of the most well-known methods is Tikhonov regularization [11], which adds a regularization term to the objective function for making inversion stable. Canu and Smola [4] have proposed a kernel method for density estimation using an exponential family defined by a positive definite kernel, while they do not formulate it rigorously. They discuss Tikhonov-type regularization for estimation. Another major approach to regularization is to approximate the original infinite dimensional space by finite dimensional subspaces [11]. This paper uses the latter approach, because it matches better with the geometrical apparatus developed in the previous sections.

Let $\{\mathscr{H}^{(\ell)}\}_{\ell=1}^{\infty}$ be a series of finite dimensional subspaces of $\mathscr{H}_k$ such that $\mathscr{H}^{(\ell)} \subset \mathscr{H}^{(\ell+1)}$ for all $\ell \in \mathbb{N}$. For any $f \in \mathscr{M}_\mu(k)$, a subspace $T_f^{(\ell)}$ of T_f is defined by $T_f^{(\ell)} = T_f \cap \mathscr{H}^{(\ell)}$, and an open set $\mathscr{W}_f^{(\ell)}$ of $T_f^{(\ell)}$ is defined by $\mathscr{W}_f^{(\ell)} = \mathscr{W}_f \cap \mathscr{H}^{(\ell)}$. For simplicity, the notations $\mathscr{W}^{(\ell)}$ and $\mathscr{S}^{(\ell)}$ are used for $\mathscr{W}_{f_0}^{(\ell)}$ and $\{f_u \in \mathscr{S} \mid u \in \mathscr{W}^{(\ell)}\}$, respectively.

For each $\ell \in \mathbb{N}$, the pseudo maximum likelihood estimator $\widehat{u}^{(\ell)}$ in $\mathscr{W}^{(\ell)}$ is defined by

$$\widehat{u}^{(\ell)} = \arg\max_{u \in \mathscr{W}^{(\ell)}} \langle \widehat{m}^{(n)}, u \rangle_{\mathscr{H}_k} - \Psi_0(u).$$

In the following discussion, it is assumed that the maximizer $\widehat{u}^{(\ell)}$ exists in $\mathscr{W}^{(\ell)}$, and further the following two assumptions are made;

(A-1) For all $u \in \mathscr{W}_0$, let $u_*^{(\ell)} \in \mathscr{W}^{(\ell)}$ ($\ell \in \mathbb{N}$) be the minimizer of

$$\min_{u^{(\ell)} \in \mathscr{W}^{(\ell)}} \mathrm{KL}(f_u || f_{u^{(\ell)}}).$$

Then

$$\|u - u_*^{(\ell)}\|_{\mathscr{H}_k} \to 0 \qquad (\ell \to \infty).$$

(A-2) For $u \in \mathscr{W}_0$, let $\lambda^{(\ell)}(u)$ be the least eigenvalue of the covariance operator Σ_u restricted on $T_{f_u}^{(\ell)}$, that is,

$$\lambda^{(\ell)}(u) = \inf_{v \in T_{f_u}^{(\ell)},\, \|v\|_{\mathscr{H}_k}=1} \langle v, \Sigma_u v \rangle_{\mathscr{H}_k}.$$

Then, there exists a subsequence $(\ell_n)_{n=1}^{\infty}$ of $\mathbb{N}$ such that for all $u \in \mathscr{W}_0$ we can find $\delta > 0$ for which

$$\tilde{\lambda}_u^{(\ell)} = \inf_{u' \in \mathscr{W}_0,\, \|u'-u\|_{\mathscr{H}_k} \le \delta} \lambda^{(\ell)}(u')$$

satisfies

$$\lim_{n\to\infty} \sqrt{n}\tilde{\lambda}_u^{(\ell_n)} = +\infty.$$

The assumption (A-1) means $\mathscr{S}^{(\ell)}$ can approximate a function in $\mathscr{S}$ at any precision as ℓ goes to infinity. The assumption (A-2) provides a stable

MLE in the submodel $\mathscr{S}^{(\ell)}$ by lower-bounding the least eigenvalue of the derivative of the map $u \mapsto m_u$.

Theorem 6 *Under the assumptions (A-1) and (A-2),*

$$\mathrm{KL}(f_* || f_{\widehat{u}^{(\ell_n)}}) \to 0 \qquad (n \to \infty)$$

in probability.

Moreover, let $u_* \in \mathscr{W}_0$ be the element which gives $f_{u_*} = f_*$, and $u_*^{(\ell)}$ be the element in (A-1) with respect to u_*. If positive constants γ_n and ε_n satisfy

$$\|u_* - u_*^{(\ell_n)}\|_{\mathscr{H}_k} = o(\gamma_n) \qquad (n \to \infty) \tag{1.8}$$

and

$$\frac{1}{\sqrt{n}\tilde{\lambda}_{u_*}^{(\ell_n)}} = o(\varepsilon_n) \qquad (n \to \infty), \tag{1.9}$$

then we have

$$\mathrm{KL}(f_* || f_{\widehat{u}^{(\ell_n)}}) = o_p(\max\{\gamma_n, \varepsilon_n\}) \qquad (n \to \infty).$$

Proof We prove the second assertion of the theorem. The first one is similar. Let m_* and $m_*^{(\ell)}$ be the mean parameters corresponding to u_* and $u_*^{(\ell)}$, respectively. From Eqs. (1.4) and (1.5), we have

$$\langle u - u_*^{(\ell)}, m_*^{(\ell)} \rangle_{\mathscr{H}_k} = \langle u - u_*^{(\ell)}, m_* \rangle_{\mathscr{H}_k} \tag{1.10}$$

for all $u \in \mathscr{W}^{(\ell)}$, and

$$\mathrm{KL}(f_* || f_{\widehat{u}^{(\ell_n)}}) = \mathrm{KL}(f_* || f_{u_*^{(\ell_n)}}) + \mathrm{KL}(f_{u_*^{(\ell_n)}} || f_{\widehat{u}^{(\ell_n)}}).$$

Eqs. (1.1) and (1.8) imply

$$\mathrm{KL}(f_* || f_{u_*^{(\ell_n)}}) = o(\gamma_n) \qquad (n \to \infty).$$

Thus, the proof is done if we show

$$\Pr\big(\|\widehat{u}^{(\ell_n)} - u_*^{(\ell_n)}\|_{\mathscr{H}_k} \geq \varepsilon_n\big) \to 0 \qquad (n \to \infty). \tag{1.11}$$

In fact, since Eqs. (1.1) and (1.10) give

$$\mathrm{KL}(f_{u_*^{(\ell_n)}} || f_{\widehat{u}^{(\ell_n)}}) = \Psi_0(\widehat{u}^{(\ell_n)}) - \Psi_0(u_*^{(\ell_n)}) - \langle m_*, \widehat{u}^{(\ell_n)} - u_*^{(\ell_n)} \rangle_{\mathscr{H}_k},$$

Eq. (1.11) means $\mathrm{KL}(f_{u_*^{(\ell_n)}} || f_{\widehat{u}^{(\ell_n)}}) = o_p(\varepsilon_n)$ $(n \to \infty)$.

Let $\delta > 0$ be the constant in the assumption (A-2) with respect to u_*. If the event of the probability in Eq. (1.11) holds, we have

$$\sup_{\substack{u \in \mathscr{W}^{(\ell_n)} \\ \|u - u_*^{(\ell_n)}\|_{\mathscr{H}_k} \geq \varepsilon_n}} L_n(u) - L_n(u_*^{(\ell_n)}) \geq 0, \tag{1.12}$$

where $L_n(u) = \langle u, \widehat{m}^{(n)} \rangle_{\mathscr{H}_k} - \Psi_0(u)$. On the other hand, it follows from Eq. (1.10) and Taylor expansion that for any $u \in \mathscr{W}^{(\ell_n)}$

$$\begin{aligned} & L_n(u) - L_n(u_*^{(\ell_n)}) \\ & = \langle u - u_*^{(\ell_n)}, \widehat{m}^{(n)} - m_* \rangle_{\mathscr{H}_k} - \left\{ \Psi_0(u) - \Psi_0(u_*^{(\ell_n)}) - \langle u - u_*^{(\ell_n)}, m_*^{(\ell_n)} \rangle_{\mathscr{H}_k} \right\} \\ & = \langle u - u_*^{(\ell_n)}, \widehat{m}^{(n)} - m_* \rangle_{\mathscr{H}_k} - \frac{1}{2} \langle u - u_*^{(\ell_n)}, \Sigma_{\tilde{u}} (u - u_*^{(\ell_n)}) \rangle_{\mathscr{H}_k}, \end{aligned}$$

where $\tilde{u}$ is a point in the line segment between u and $u_*^{(\ell_n)}$. By the definition of $\tilde{\lambda}^{(\ell)}$, for sufficiently large n so that $\|u_*^{(\ell_n)} - u_*\|_{\mathscr{H}_k} \leq \delta$, we obtain

$$\begin{aligned} & \sup_{\substack{u \in \mathscr{W}^{(\ell_n)} \\ \|u - u_*^{(\ell_n)}\|_{\mathscr{H}_k} \geq \varepsilon_n}} L_n(u) - L_n(u_*^{(\ell_n)}) \\ & \leq \sup_{\substack{u \in \mathscr{W}^{(\ell_n)} \\ \|u - u_*^{(\ell_n)}\|_{\mathscr{H}_k} \geq \varepsilon_n}} \|u - u_*^{(\ell_n)}\|_{\mathscr{H}_k} \|\widehat{m}^{(n)} - m_*\|_{\mathscr{H}_k} - \frac{1}{2} \tilde{\lambda}^{(\ell_n)} \|u - u_*^{(\ell_n)}\|_{\mathscr{H}_k}^2 \\ & \leq \sup_{\substack{u \in \mathscr{W}^{(\ell_n)} \\ \|u - u_*^{(\ell_n)}\|_{\mathscr{H}_k} \geq \varepsilon_n}} \|u - u_*^{(\ell_n)}\|_{\mathscr{H}_k} \left\{ \|\widehat{m}^{(n)} - m_*\|_{\mathscr{H}_k} - \frac{1}{2} \tilde{\lambda}^{(\ell_n)} \varepsilon_n \right\}. \end{aligned} \tag{1.13}$$

Eqs. (1.12) and (1.13) show that the probability in Eq. (1.11) is upper bounded by

$$\Pr\left(\|\widehat{m}^{(n)} - m_*\|_{\mathscr{H}_k} \geq \tfrac{1}{2} \tilde{\lambda}^{(\ell_n)} \varepsilon_n \right),$$

which converges to zero by Theorem 5 and Eq. (1.9). □

There is a trade-off between the decay rates of ε_n and γ_n; if the subspace $\mathscr{W}^{(\ell_n)}$ enlarges rapidly, the approximation accuracy γ_n decreases fast, while a small value for $\tilde{\lambda}_{u_*}^{(\ell_n)}$ results in a slow rate of ε_n.

1.4 Concluding Remarks

This paper has proposed a new family of statistical models, reproducing kernel exponential manifold, which includes infinite dimensional exponential families. The most significant property of this exponential manifold is that the empirical mean parameter is included in the Hilbert space. Thus, estimation of the density function with a finite sample can be discussed based on this exponential manifold, while many other formulation of infinite dimensional exponential manifold cannot provide basis for estimation with a finite sample. Using the reproducing kernel exponential manifold, a method of pseudo maximum likelihood estimation has been proposed with a series of finite dimensional submanifolds, and consistency of the estimator has been shown.

Many problems remain unsolved, however. One of them is a practical method for constructing a sequence of subspaces used for the pseudo maximum likelihood estimation. A possible way of defining the sequence is to use the subspace spanned by $k(\cdot, X_1), \ldots, k(\cdot, X_\ell)$. However, with this construction the subspaces are also random depending on the sample, and the results in this paper should be extended to the case of random subspaces to guarantee the consistency. Another practical issue is how to choose the subsequence ℓ_n so that the assumption (A-2) is satisfied. We need to elucidate the properties of least eigenvalue of the covariance operator restricted on finite dimensional subspaces, which is not necessarily obvious. Also, providing examples of the estimator for specific kernels is practically important. Investigation of these problems will be among our future works.

Another important problem, which is not discussed in this paper, is the dual geometry on the infinite dimensional exponential family. Unlike the finite dimensional cases, it is not straightforward to define the dual connections on the tangent bundle of the infinite dimensional exponential manifold ([10]). It will be interesting to consider the dual geometric structure on the reproducing kernel exponential manifolds for various choices of the space and kernel k.

Acknowledgements

The author thanks Professor Govanni Pistone for helpful and insightful discussions on this work. The author thanks also the anonymous reviewer for his helpful comments to improve this article. This work has been partially supported by JSPS KAKENHI 19500249, Japan, and Alexander-von-Humboldt fellowship, Germany.

Bibliography

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68(3): 337–404, 1950.

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics.* Kluwer Academic Publisher, 2004.

C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups.* Springer-Verlag, 1984.

S. Canu, A.J. Smola. Kernel methods and the exponential family. *Neurocomputing*, 69(7-9): 714–720, 2006

A. Cena and G. Pistone. Exponential statistical manifold. *Annals of the Institute of Statistical Mathematics*, 59: 27–56, 2007.

K. Fukumizu. Infinite dimensional exponential families by reproducing kernel Hilbert spaces. *Proceedings of the 2nd International Symposium on Information Geometry and its Applications (IGAIA2005)*, December, 2005, Tokyo. `http://www.stat.t.u-tokyo.ac.jp/~infogeo/abst/KenjiFUKUMIZU.pdf`

K. Fukumizu, F. R. Bach and M. I. Jordan. Kernel dimension reduction in regression. *Technical Report 715, Dept. Statistics, University of California, Berkeley*, 2006.

K. Fukumizu, F. R. Bach and A. Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8: 361–383, 2007.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. *Advances in Neural Information Processing Systems*, 21, 489–496, 2008.

P. Gibilisco and G. Pistone. Connections on non-parametric statistical manifolds by Orlicz space geometry. *Infinite Dimensional Analysis, Quantum Probability and Related Topics*, 1(2): 325–347, 1998.

C.W. Groetsch. *The Theory of Tikhonov regularization for Fredholm equations of the first kind.* Pitman:London. 1984.

S. Lang. *Differential Manifolds.* Springer-Verlag, 1985.

G. Pistone and M.-P. Rogantin. The exponential statistical manifold: Mean parameters, orthogonality, and space transformation. *Bernoulli*, 5: 721–760, 1999.

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

B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet and B. Schölkopf. Injective Hilbert Space Embeddings of Probability Measures. *Proceedings of the*

21st Annual Conference on Learning Theory (COLT 2008), 2008, to appear.