Chapter 2: Kernel Exponential Family and Density Estimation Problem in It

In this chapter, we formally introduce the kernel exponential family and discuss the density estimation problem in it.

2.1 Kernel Exponential Families

2.1.1 A Review of Finite-dimensional Exponential Family

In order to introduce the kernel exponential family, we first review the classic finite-dimensional exponential family.

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the sample space and $\varphi : \mathcal{X} \to \mathbb{R}^m$ be a measurable vector-valued function such that $\varphi(x) = (\varphi_1(x), \cdots, \varphi_m(x))^\top \in \mathbb{R}^m$ for all $x \in \mathcal{X}$. An *m-dimensional exponential family* (Brown, 1986; Barndorff-Nielsen, 2014), denoted by $\mathcal{Q}_{\text{fin}}$, in its natural parametrization form, contains all pdfs of the form

$$\tilde{q}_\theta(x) := \mu(x) \exp\big(\langle \varphi(x), \theta \rangle - B(\theta)\big) \text{ for all } x \in \mathcal{X}, \qquad \theta \in \Theta, \tag{2.1}$$

where $\mu : \mathcal{X} \to [0, \infty)$ is a pdf referred to as the *base density*, $\theta \in \mathbb{R}^m$ is the *natural parameter*, φ is referred to as the *canonical statistic*, $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathbb{R}^m$,

$$B(\theta) := \log\left(\int_{\mathcal{X}} \mu(x) \exp\big(\langle \theta, \varphi(x) \rangle\big) \mathrm{d}x\right) \tag{2.2}$$

is the *log-partition function*, and $\Theta := \left\{\theta \in \mathbb{R}^m \;\middle|\; B(\theta) < \infty\right\}$ is the *natural parameter space*.

The finite-dimensional exponential family was first discovered by Darmois (1935), Koopman (1936), and Pitman (1936) in studying the family of distributions whose sufficient statistics have fixed dimensionality as the sample size increases. It can also be motivated via the principle of maximum entropy: given n i.i.d samples $X_1, \cdots, X_n \in \mathcal{X}$ and m measurable functions $\varphi_j : \mathcal{X} \to \mathbb{R}$, for all $j = 1, \cdots, m$, the unique pdf p over $\mathcal{X}$ that maximizes Shannon's entropy

$$-\int_{\mathcal{X}} p(x) \log p(x) \mathrm{d}x$$

subject to the linear constraints

$$\int_{\mathcal{X}} p(x)\varphi_j(x)\mathrm{d}x = \frac{1}{n}\sum_{i=1}^{n} \varphi_j(X_i), \qquad \text{for all } j = 1, \cdots, m,$$

takes on the form (2.1).

Exponential families are ubiquitous in statistics. Many parametric families are special cases of exponential families, such as the family of Gaussian pdfs and the family of probability mass functions (pmfs) of the binomial distribution with the known number of trials. In addition, they form the basis for the generalized linear models (McCullagh and Nelder, 1989). In recent years, they are also used extensively in the studies of graphical models (Wainwright and Jordan, 2008).

2.1.2 Kernel Exponential Family

Note that in (2.1), the function φ maps to an m-dimensional Euclidean space, which can be limited in some applications. In addition, $\tilde{q}_\theta$ depends on φ only through its inner product with $\theta \in \Theta$. Motivated by these observations, Canu and Smola (2006) proposed to replace the inner product in the Euclidean space in (2.1) by the

one in a RKHS, and introduced the *kernel exponential family*, denoted by $\mathcal{Q}_{\mathrm{ker}}$, that contains all pdfs over $\mathcal{X}$ of the form

$$q_f(x) := \mu(x) \exp\bigl(f(x) - A(f)\bigr) \text{ for all } x \in \mathcal{X}, \qquad f \in \mathcal{F}, \tag{2.3}$$

where

$$A(f) := \log\left(\int_{\mathcal{X}} \mu(x) \exp(f(x)) \mathrm{d}x\right) \tag{2.4}$$

is the *log-partition functional*, and $\mathcal{F} := \left\{ f \in \mathcal{H} \;\middle|\; A(f) < \infty \right\}$ is referred to as the *natural parameter space*.

Due to the reproducing property of k, we have $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$, where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the kernel function associated with the underlying RKHS $\mathcal{H}$, and $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ denotes the inner product in $\mathcal{H}$. Thus, comparing with (2.1), we see that $f \in \mathcal{H}$ plays the role of the natural parameter and $k(x, \cdot) \in \mathcal{H}$ plays the role of the canonical statistic.

The kernel exponential family $\mathcal{Q}_{\mathrm{ker}}$ has been used in various statistical applications, such as the anomaly detection (Canu and Smola, 2006), and the estimation of the conditional independence structure of graphical models (Sun, Kolar, and Xu, 2015).

2.1.3 Properties of $\mathcal{Q}_{\mathrm{ker}}$

The kernel exponential family $\mathcal{Q}_{\mathrm{ker}}$ has many nice properties, some of which are in common with those of $\mathcal{Q}_{\mathrm{fin}}$. In the following subsections, we discuss these properties.

2.1.3.1 Characterization of $\mathcal{F}$ for Bounded Kernels

The first property we discuss here is related to the characterization of $\mathcal{F}$ when a bounded kernel function k is used, where k is said to be *bounded* if $\kappa_1 := \sup_{x \in \mathcal{X}} \sqrt{k(x, x)} < \infty$.

Proposition 2.1. *If the kernel function k is bounded, then we have $\mathcal{F} = \mathcal{H}$.*

The proof of Proposition 2.1 can be found in Subsection 2.3.1.

As a consequence of Proposition 2.1, if $\mathcal{X} \subset \mathbb{R}^d$ is compact and k is continuous over $\mathcal{X}$, we must have $\kappa_1 < \infty$ and $\mathcal{F} = \mathcal{H}$. As another example, if $\mathcal{X} = \mathbb{R}^d$ or an unbounded proper subset of $\mathbb{R}^d$, and k is the Gaussian kernel function, $k(x, y) = \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right)$ for all $x, y \in \mathcal{X}$, or the rational quadratic kernel function, $k(x, y) = \left(1 + \frac{\|x-y\|_2^2}{\sigma^2}\right)^{-1}$ for all $x, y \in \mathcal{X}$, where $\sigma > 0$ is the bandwidth parameter associated with each kernel function, we also have $\kappa_1 < \infty$ and $\mathcal{F} = \mathcal{H}$.

The boundedness of k is *not* a necessary condition for $\mathcal{F} = \mathcal{H}$, though. For instance, if we let $\mathcal{X} = \mathbb{R}$ and $k(x, y) = xy$ for all $x, y \in \mathbb{R}$, the corresponding $\mathcal{H}$ contains all functions f of the form

$$f(x) = \sum_{i=1}^{\infty} y_i x, \qquad \text{for all } x \in \mathcal{X},$$

that satisfies $\|f\|_{\mathcal{H}}^2 = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} y_i y_j < \infty$. It is easy to observe $\sup_{x \in \mathcal{X}} \sqrt{k(x, x)} = \sup_{\mathcal{X}} |x| = \infty$. In addition, choose μ to be $\mu(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ for all $x \in \mathcal{X}$. It follows that $A(f) < \infty$ for all $f \in \mathcal{H}$ and $\mathcal{F} = \mathcal{H}$.

2.1.3.2 Convexity of A

It is a classic result that the log-partition function B defined in (2.2) is a convex function and the natural parameter space Θ is a convex set (Theorem 7.1 in Barndorff-Nielsen, 2014). As the following proposition demonstrates, similar results hold for $\mathcal{Q}_{\text{ker}}$.

Proposition 2.2. *The log-partition functional A defined in (2.4) is convex over $\mathcal{F}$, and is strictly convex over $\mathcal{F}$ if $\mathcal{H}$ does* not *contain constant functions. In addition, $\mathcal{F}$ is convex.*

The proof of Proposition 2.2 can be found in Subsection 2.3.2.

As we will see in Subsection 2.2.1 of this chapter and Chapter 4, the convexity of A plays an important role in density estimation using the (penalized) NLL loss functional, which ensures the convexity of the (penalized) NLL loss functional and directly relates to the existence and the uniqueness of its minimizer.

2.1.3.3 Differential Properties of A

It is well-known that the log-partition function B in $\mathcal{Q}_{\text{fin}}$ is infinitely differentiable and has a close relationship with the cumulant and moment generating functions of the canonical statistic (Theorem 2.2 in Brown, 1986).

In this section, we study differential properties of the log-partition functional A in $\mathcal{Q}_{\text{ker}}$ defined in (2.4). We will show that A is also infinitely differentiable and link its derivatives (suitably defined) to the moments of $k(X, \cdot)$.

To start with, notice that the domain of A is a collection of functions over $\mathcal{X}$, or more precisely, a subset of a Hilbert space. We need a version of differentiability defined for functionals whose domain is a normed space or a subset of it. We choose the *Frechét differentiability*, whose definition, together with those of Fréchet derivative and gradient, is given in Section A.1 in Appendix A.

The following lemma illustrates these definitions and serves as a preparation for the derivation of the Fréchet derivative and gradient of A in Proposition 2.3.

Lemma 2.1. *Suppose* $\sup_{x\in\mathcal{X}} \sqrt{k(x,x)} < \infty$ *so that* $\mathcal{F} = \mathcal{H}$. *Let* $x \in \mathcal{X}$ *be fixed and* $J_x : \mathcal{H} \to \mathbb{R}$ *be the evaluation functional* $J_x(f) = \langle f, k(x, \cdot)\rangle_{\mathcal{H}} = f(x)$ *for all* $f \in \mathcal{H}$. *Then,* J_x *is Frechét differentiable over* $\mathcal{H}$ *with the Frechét derivative at* $f \in \mathcal{H}$ *being*

$$\mathrm{D}J_x(f)(g) = \langle g, k(x, \cdot)\rangle_{\mathcal{H}}, \qquad \text{for all } g \in \mathcal{H}.$$

In addition, the Fréchet gradient is $\nabla J_x : \mathcal{H} \to \mathcal{H}$ *with* $\nabla J_x(f) = k(x, \cdot)$ *for all* $f \in \mathcal{H}$.

The proof of Lemma 2.1 can be found in Section 2.3.3.

With this lemma, we now establish the Fréchet differentiability of A over $\mathcal{H}$ and derive its Fréchet derivative and gradient at $f \in \mathcal{H}$.

Proposition 2.3 (Fréchet differentiability, derivative and gradient of A)**.** *Suppose* $\sup_{x \in \mathcal{X}} \sqrt{k(x,x)} < \infty$ *so that* $\mathcal{F} = \mathcal{H}$. *Then,* $A : \mathcal{H} \to \mathbb{R}$ *is Fréchet differentiable over* $\mathcal{H}$ *and its Fréchet derivative at* $f \in \mathcal{H}$ *is a bounded linear operator from* $\mathcal{H}$ *to* $\mathbb{R}$ *given by*

$$\mathrm{D}A(f)(g) = \left\langle g, \int_{\mathcal{X}} k(x, \cdot) q_f(x) \mathrm{d}x \right\rangle_{\mathcal{H}} = \int_{\mathcal{X}} g(x) q_f(x) \mathrm{d}x, \tag{2.5}$$

for all $g \in \mathcal{H}$. *In addition, the Fréchet gradient of* A *is*

$$\nabla A : \mathcal{H} \to \mathcal{H}, \qquad f \longmapsto \int_{\mathcal{X}} k(x, \cdot) q_f(x) \mathrm{d}x.$$

The proof of Proposition 2.3 can be found in Section 2.3.4.

Remark 2.1. Note that the integrand of $\int_{\mathcal{X}} k(x, \cdot) q_f(x) \mathrm{d}x$ is an element in $\mathcal{H}$ and this integral is a Bochner integral (Diestel and Uhl, 1977; Denkowski, Migórski, and Papageorgiou, 2013); see Section A.2 in Appendix A for its definition and properties. In particular, the second equality in (2.5) follows from Proposition A.4(c) there.

In addition, we can view $\int_{\mathcal{X}} k(x, \cdot) q_f(x) \mathrm{d}x$ as the output of the *kernel mean embedding* (Bertinet and Agnan, 2004; Smola et al., 2007), i.e., we map a distribution F, whose density function is q_f, into $\mathcal{H}$ via the map

$$F \longmapsto \int_{\mathcal{X}} k(x, \cdot) \mathrm{d}F(x) = \int_{\mathcal{X}} k(x, \cdot) q_f(x) \mathrm{d}x.$$

Kernel mean embedding has drawn a lot of attention in recent years. Its statistical properties have been extensively studied by, to name a few, Le (2008), Sriperumbudur et al. (2010), Sriperumbudur, Fukumizu, and others (2011), and Muandet et al. (2016). It can be used in statistical hypothesis testing for independence (Gretton et al., 2005) and for the equality of two sets of random samples (Gretton et al., 2012), statistical clustering (Jegelka et al., 2009), the estimation of graphical models (Song, Gretton, and Guestrin, 2010; Song, Fukumizu, and Gretton, 2013; Song et al., 2014). More details about the kernel mean embedding can be found in the recent comprehensive survey by Muandet et al. (2017). ▶

Proposition 2.3 only considers the first-order Fréchet differentiability of A. By an inductive argument, we can show that A is r-times Fréchet differentiable for all $r \in \mathbb{N}$. In particular, the following proposition shows the result when $r = 2$.

Proposition 2.4 (Second-order Fréchet differentiability, derivative and gradient of A)**.** *Suppose* $\sup_{x \in \mathcal{X}} \sqrt{k(x,x)} < \infty$ *so that* $\mathcal{F} = \mathcal{H}$*. Then, the log-partition functional* A *is twice Fréchet differentiable over* $\mathcal{H}$ *and its second-order Frechét derivative at* $f \in \mathcal{H}$*, denoted by* $\mathrm{D}^2 A(f)$*, is a map from* $\mathcal{H}$ *to the collection of bounded linear operators from* $\mathcal{H}$ *to* $\mathbb{R}$ *given by*

$$\begin{aligned}\mathrm{D}^2 A(f)(g) = &\left[\left(\int_{\mathcal{X}} k(x,\cdot) \otimes k(x,\cdot) q_f(x)\mathrm{d}x\right)\right. \\ &\left. - \left(\int_{\mathcal{X}} k(x,\cdot) q_f(x)\mathrm{d}x\right) \otimes \left(\int_{\mathcal{X}} k(x,\cdot) q_f(x)\mathrm{d}x\right)\right] g \\ = &\left(\int_{\mathcal{X}} k(x,\cdot) g(x) q_f(x)\mathrm{d}x\right) - \left(\int_{\mathcal{X}} q_f(x) g(x)\mathrm{d}x\right)\left(\int_{\mathcal{X}} k(x,\cdot) q_f(x)\mathrm{d}x\right),\end{aligned}$$

for all $g \in \mathcal{H}$*.*

In addition, the second-order Fréchet gradient of A *at* $f \in \mathcal{F}$*, denoted by* $\nabla^2 A(f)$*, is a bounded linear operator from* $\mathcal{H}$ *to itself given by*

$$\nabla^2 A(f) = \int_{\mathcal{X}} k(x,\cdot) \otimes k(x,\cdot) q_f(x)\mathrm{d}x$$

$$- \left(\int_{\mathcal{X}} k(x, \cdot) q_f(x) \mathrm{d}x \right) \otimes \left(\int_{\mathcal{X}} k(x, \cdot) q_f(x) \mathrm{d}x \right)$$

The proof of Proposition 2.4 is similar to that of Proposition 2.3 and is omitted here.

The bounded linear operator $\nabla^2 A(f)$, or more generally, the operator

$$\Sigma_F := \left(\int_{\mathcal{X}} k(x, \cdot) \otimes k(x, \cdot) \mathrm{d}F(x) \right) - \left(\int_{\mathcal{X}} k(x, \cdot) \mathrm{d}F(x) \right) \otimes \left(\int_{\mathcal{X}} k(x, \cdot) \mathrm{d}F(x) \right)$$

where F is a distribution over $\mathcal{X}$, maps from $\mathcal{H}$ to itself, and is referred to as the *covariance operator* in the literature, since, for any $f, g \in \mathcal{H}$, we have

$$\langle f, \Sigma_F g \rangle_{\mathcal{H}} = \mathbb{E}_F[f(X)g(X)] - \mathbb{E}_F[f(X)]\,\mathbb{E}_F[g(X)],$$

which is the covariance between $f(X)$ and $g(X)$ with $X \sim F$. The operator Σ_F is known to be linear, bounded, self-adjoint, and of trace-class (Baker, 1973). The covariance operator has been used in such statistical and machine learning applications as dimensionality reduction (Fukumizu, Bach, and Jordan, 2004; Fukumizu, Bach, and Jordan, 2009), kernel principal component analysis (Schölkopf, Smola, and Müller, 1998), kernel canonical correlation analysis (Fukumizu, Bach, and Gretton, 2007), and independence and conditional independence measures (Gretton et al., 2005; Fukumizu et al., 2007). We will also see the covariance operator appears in the influence function of the ML log-density projection in $\mathcal{Q}_{\text{ker}}$ in Chapter 5. More details on the covariance operator and, more generally, the cross-covariance operator can be found in Section 3.2 and 4.3 in Muandet et al. (2017).

2.1.4 Connection to Finite-dimensional Exponential Families

In this section, we show the connection between $\mathcal{Q}_{\text{fin}}$ and $\mathcal{Q}_{\text{ker}}$. In particular, we show that density functions in $\mathcal{Q}_{\text{fin}}$ can be written in the form of those in $\mathcal{Q}_{\text{ker}}$ by

choosing $\mathcal{H}$ to be a finite-dimensional RKHS. Therefore, $\mathcal{Q}_{\text{fin}}$ is a special case of $\mathcal{Q}_{\text{ker}}$. If we choose $\mathcal{H}$ to be an infinite-dimensional RKHS, $\mathcal{Q}_{\text{ker}}$ is a strict generalization of $\mathcal{Q}_{\text{fin}}$.

We start with $\mathcal{Q}_{\text{fin}}$ that contains all pdfs of the form (2.1). Consider the following collection of functions

$$\mathcal{H}_0 := \left\{ \sum_{j=1}^{m} \alpha_j \varphi_j \;\middle|\; \alpha_1, \cdots, \alpha_m \in \mathbb{R} \right\},$$

which is a vector space with addition and scalar multiplication defined as

$$(f+g)(x) = f(x) + g(x), \qquad \text{and} \qquad (cf)(x) = cf(x), \qquad \text{for all } x \in \mathcal{X},$$

for all $f, g \in \mathcal{H}_0$ and all $c \in \mathbb{R}$.

Now, for any $f = \sum_{j=1}^{m} \alpha_j \varphi_j \in \mathcal{H}_0$ and $g = \sum_{j=1}^{m} \beta_j \varphi_j \in \mathcal{H}_0$, where $\alpha_j, \beta_j \in \mathbb{R}$ for all $j = 1, \cdots, m$, define the inner product between them to be

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{j=1}^{m} \alpha_j \beta_j. \tag{2.6}$$

Then, we have the following proposition.

Proposition 2.5. *The vector space $\mathcal{H}_0$ equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ defined in (2.6), denoted by $(\mathcal{H}_0, \langle \cdot, \cdot \rangle_{\mathcal{H}_0})$, is a RKHS with the kernel function*

$$k(x, y) = \sum_{j=1}^{m} \varphi_j(x) \varphi_j(y), \qquad \text{for all } x, y \in \mathcal{X}. \tag{2.7}$$

The proof of Proposition 2.5 can be found in Subsection 2.3.5.

With the construction above, we have

$$\begin{aligned} \tilde{q}_\theta(x) &= \mu(x) \exp\big(\langle \theta, \varphi(x) \rangle - B(\theta)\big) \\ &= \mu(x) \exp\left(\sum_{j=1}^{m} \theta_j \varphi_j(x) - B(\theta)\right) \\ &= \mu(x) \exp\big(\langle f, k(x, \cdot) \rangle_{\mathcal{H}_0} - A(f)\big), \end{aligned}$$

where $\theta := (\theta_1, \cdots, \theta_m)^\top \in \Theta \subseteq \mathbb{R}^m$, $f := \sum_{j=1}^m \theta_j \varphi_j \in \mathcal{H}_0$, $k(x, \cdot) := \sum_{j=1}^m \varphi_j(x)\varphi_j \in \mathcal{H}_0$, and

$$\begin{aligned} A(f) &= \log\left(\int_{\mathcal{X}} \mu(x) \exp\big(\langle f, k(x, \cdot)\rangle_{\mathcal{H}_0}\big) \mathrm{d}x\right) \\ &= \log\left(\int_{\mathcal{X}} \mu(x) \exp\left(\sum_{j=1}^m \theta_j \varphi_j(x)\right) \mathrm{d}x\right) = B(\theta). \end{aligned}$$

Thus, every $\tilde{q}_\theta \in \mathcal{Q}_{\text{fin}}$ can be written in the form of pdfs in $\mathcal{Q}_{\text{ker}}$, and, with the RKHS being $\mathcal{H}_0$, $\mathcal{Q}_{\text{fin}}$ is a special case of $\mathcal{Q}_{\text{ker}}$.

In the rest of this section, we provide several examples of frequently used finite-dimensional exponential families to illustrate Proposition 2.5. In these examples, we temporarily ignore the fact that μ is a pdf over $\mathcal{X}$, which is assumed in the definitions of $\mathcal{Q}_{\text{fin}}$ and $\mathcal{Q}_{\text{ker}}$.

Example 2.1 (Binomial distribution). Consider the family of pmfs of the binomial distribution with the known number of trials n and the success probability $\eta \in (0, 1)$. The general form of the pmf is

$$p_\eta(x) := \binom{n}{x} \eta^x (1-\eta)^{n-x}, \qquad \text{for all } x \in \mathcal{X} := \big\{0, 1, \cdots, n\big\}.$$

In the natural parametrization, we can rewrite p_η as

$$\tilde{q}_\theta(x) := \binom{n}{x} \exp\big(x\theta - n\log(1 + e^\theta)\big), \qquad \text{for all } x \in \mathcal{X},$$

where the natural parameter is $\theta := \log\big(\frac{\eta}{1-\eta}\big)$ and the natural parameter space is $\mathbb{R}$. We recognize

$$\mu(x) = \binom{n}{x} \quad \text{and} \quad \varphi(x) = x \text{ for all } x \in \mathcal{X}, \quad \text{and} \quad B(\theta) = n\log(1 + e^\theta).$$

Here, the canonical statistic is $\varphi = \mathrm{Id}$, the identity map, with $\varphi(x) = x$ for all $x \in \mathcal{X}$. Then, $\mathcal{H}_0$ in this case contains all functions of the form $f(x) = \theta x$ for all $x \in \mathcal{X}$, where $\theta \in \mathbb{R}$. With $\theta_1, \theta_2 \in \mathbb{R}$, the inner product between $f = \theta_1 \cdot \mathrm{Id} \in \mathcal{H}_0$ and $g = \theta_2 \cdot \mathrm{Id} \in \mathcal{H}_0$

is $\langle f, g\rangle_{\mathcal{H}_0} = \theta_1\theta_2$. The reproducing kernel is $k(x, y) = \langle x \cdot \mathrm{Id}, y \cdot \mathrm{Id}\rangle_{\mathcal{H}_0} = xy$ for all $x, y \in \mathcal{X}$.

Furthermore, with $f = \theta \cdot \mathrm{Id}$ for some $\theta \in \mathbb{R}$, we have

$$A(f) = \log\left(\sum_{x=0}^{n} \binom{n}{x} e^{\theta x}\right) = \log((1 + e^{\theta})^n) = n \log(1 + e^{\eta}) = B(\theta),$$

which is the desired result. Since $A(f) = A(\theta \cdot \mathrm{Id}) = n \log(1+e^{\theta}) < +\infty$ for all $\theta \in \mathbb{R}$, we have $\mathcal{F} = \big\{f \in \mathcal{H}_0 \mid f := \theta \cdot \mathrm{Id}, \theta \in \mathbb{R}\big\}$. ▶

Example 2.2 (Poisson distribution). Consider the family of the pmfs of the Poisson distribution with the mean parameter $\eta > 0$. The general form of the pmf is

$$p_\eta(x) = e^{-\eta}\frac{\eta^x}{x!}, \qquad \text{for all } x \in \mathcal{X} := \{0, 1, 2, \cdots\}.$$

We rewrite p_η in the natural parametrization form as

$$\tilde{q}_\theta(x) = \frac{1}{x!}\exp(x\theta - e^{\theta}), \qquad \text{for all } x \in \mathcal{X},$$

where the natural parameter is $\theta = \log \eta$ and the natural parameter space is $\mathbb{R}$. We recognize

$$\mu(x) = \frac{1}{x!} \quad \text{and} \quad \varphi(x) = x \text{ for all } x \in \mathcal{X}, \qquad \text{and} \qquad B(\theta) = e^{\theta}.$$

Here, similar to the binomial example we have considered earlier, the canonical statistic is $\varphi = \mathrm{Id}$, the identity map. Then, $\mathcal{H}_0$ contains all functions of the form $f(x) = \theta x$ for all $x \in \mathcal{X}$, where $\theta \in \mathbb{R}$. With $\theta_1, \theta_2 \in \mathbb{R}$, the inner product between $f = \theta_1 \cdot \mathrm{Id} \in \mathcal{H}_0$ and $g = \theta_2 \cdot \mathrm{Id} \in \mathcal{H}_0$ is $\langle f, g\rangle_{\mathcal{H}_0} = \theta_1\theta_2$. The reproducing kernel is $k(x, y) = \langle x \cdot \mathrm{Id}, y \cdot \mathrm{Id}\rangle_{\mathcal{H}_0} = xy$ for all $x, y \in \mathcal{X}$.

Furthermore, with $f = \theta \cdot \mathrm{Id}$ for some $\theta \in \mathbb{R}$, we have

$$A(f) = \log\left(\sum_{x=0}^{\infty} \frac{(e^{\theta})^x}{x!}\right) = \log(\exp(e^{\theta})) = e^{\theta} = B(\theta),$$

which is the desired result. Since $A(f) = e^{\theta} < +\infty$ if and only if $\theta \in \mathbb{R}$, the natural parameter space is $\mathcal{F} = \big\{f \in \mathcal{H}_0 \mid f = \theta \cdot \mathrm{Id}, \theta \in \mathbb{R}\big\}$. ▶

Example 2.3 (Exponential distribution). Consider the family of pdfs of the exponential distribution with the scale parameter $\eta > 0$. The general form of the pdf is

$$p_\eta(x) = \frac{1}{\eta}\exp\left(-\frac{x}{\eta}\right), \qquad \text{for all } x \in \mathcal{X} := \big\{x \mid x > 0\big\}.$$

We can rewrite it in the natural parametrization form as

$$\tilde{q}_\theta(x) = \exp\Big(x\theta + \log(-\theta)\Big), \qquad \text{for all } x \in \mathcal{X},$$

where the natural parameter is $\theta := -\frac{1}{\eta}$ and the natural parameter space is $\Theta := (-\infty, 0)$. We recognize that

$$\mu(x) = 1 \quad \text{and} \quad \varphi(x) = x \text{ for all } x \in \mathcal{X}, \qquad \text{and} \qquad B(\theta) = -\log(-\theta).$$

Here, the canonical statistic is $\varphi = \mathrm{Id}$. Then, $\mathcal{H}_0$ contains all functions of the form $f(x) = \theta x$ for all $x \in \mathcal{X}$, for some $\theta \in \Theta$. With $\theta_1, \theta_2 \in \Theta$, the inner product between $f = \theta_1 \cdot \mathrm{Id} \in \mathcal{H}$ and $g = \theta_2 \cdot \mathrm{Id} \in \mathcal{H}$ is $\langle f, g\rangle_{\mathcal{H}_0} = \theta_1\theta_2$. It follows that the reproducing kernel is $k(x, y) = \langle x \cdot \mathrm{Id}, y \cdot \mathrm{Id}\rangle_{\mathcal{H}_0} = xy$, for all $x, y \in \mathcal{X}$.

Finally, with $f = \theta \cdot \mathrm{Id}$ for some $\theta \in \Theta$, we have

$$A(f) = \log\left(\int_{\mathcal{X}} \mu(x)\exp(f(x))\mathrm{d}x\right) = \log\left(\int_0^{+\infty} \exp(\theta x)\mathrm{d}x\right) = \log\left(-\frac{1}{\theta}\right) = B(\theta),$$

which is the desired result. Since $A(f) = A(\theta \cdot \mathrm{Id}) = \log(-\frac{1}{\theta}) < +\infty$ if and only if $\theta < 0$, the natural parameter space is $\mathcal{F} = \big\{f \in \mathcal{H}_0 \mid f = \theta \cdot \mathrm{Id}, \theta \in (-\infty, 0)\big\}$. ▶

Example 2.4 (Univariate normal distribution). Consider the univariate normal distribution with the unknown mean $\eta \in \mathbb{R}$ and the unknown variance $\sigma^2 > 0$. The general form of the pdf is

$$p_{\eta,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\eta)^2\right), \qquad \text{for all } x \in \mathcal{X} := \mathbb{R}.$$

We can rewrite p_{η,σ^2} in the natural parametrization form as

$$\tilde{q}_\theta(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\theta_1 x^2 + \theta_2 x - \left(-\frac{\theta_2^2}{4\theta_1} - \frac{1}{2}\log(-2\theta_1)\right)\right),$$

where

$$\theta := (\theta_1, \theta_2)^\top = \left(-\frac{1}{2\sigma^2}, \frac{\eta}{\sigma^2}\right)^\top, \qquad \Theta := \left\{\theta \;\middle|\; \theta := (\theta_1, \theta_2)^\top, \theta_1 < 0, \theta_2 \in \mathbb{R}\right\}.$$

$$\varphi(x) := (\varphi_1(x), \varphi_2(x))^\top = (x^2, x)^\top \in \mathbb{R}^2, \quad \text{and} \quad \mu(x) = \frac{1}{\sqrt{2\pi}} \text{ for all } x \in \mathcal{X},$$

$$B(\theta) = -\frac{\theta_2^2}{4\theta_1} - \frac{1}{2}\log(-2\theta_1).$$

This is an example of a 2-dimensional exponential family.

In this case, $\mathcal{H}_0$ contains all functions f of the form

$$f(x) = \theta_1 x^2 + \theta_2 x, \qquad \text{for all } x \in \mathbb{R},$$

for some $(\theta_1, \theta_2)^\top \in \Theta$. For any $f, g \in \mathcal{H}_0$ with $f(x) = \theta_{1,1}x^2 + \theta_{1,2}x$ and $g(x) = \theta_{2,1}x^2 + \theta_{2,2}x$ for all $x \in \mathcal{X}$, where $(\theta_{1,1}, \theta_{1,2})^\top \in \Theta$ and $(\theta_{2,1}, \theta_{2,2})^\top \in \Theta$, define the inner product between f and g to be

$$\langle f, g \rangle_{\mathcal{H}_0} = \theta_{1,1}\theta_{2,1} + \theta_{1,2}\theta_{2,2}.$$

Then, $(\mathcal{H}_0, \langle \cdot, \cdot \rangle_{\mathcal{H}_0})$ forms a RKHS with the reproducing kernel

$$k(x, y) = x^2 y^2 + xy, \qquad \text{for all } x, y \in \mathbb{R}.$$

We finally verify $A(f) = B(\theta)$ with $f(x) = \theta_1 x^2 + \theta_2 x$ for all $x \in \mathcal{X}$, where $\theta := (\theta_1, \theta_2)^\top \in \Theta$. Note the following

$$\begin{aligned} A(f) = \log\left(\int_{\mathcal{X}} \mu(x) \exp(f(x)) \mathrm{d}x\right) &= \log\left(\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\bigl(\theta_1 x^2 + \theta_2 x\bigr) \mathrm{d}x\right) \\ &= \log\left(\frac{1}{\sqrt{-2\theta_1}} \exp\left(-\frac{\theta_2^2}{4\theta_1}\right)\right) \\ &= -\frac{\theta_2^2}{4\theta_1} - \frac{1}{2}\log(-2\theta_1) \end{aligned}$$

$$= B(\theta).$$

Since $A(f) < +\infty$ if and only if $\theta_1 < 0$ and $\theta_2 \in \mathbb{R}$, we have $\mathcal{F} = \left\{ f \in \mathcal{H}_0 \mid f(x) = \theta_1 x^2 + \theta_2 x \text{ for all } x \in \mathcal{X}, \theta_1 < 0, \theta_2 \in \mathbb{R} \right\}$. ▶

2.1.5 Assumptions on $\mathcal{H}$ and k and Their Implications

In the rest of this dissertation, we are going to make the following assumptions on $\mathcal{H}$, k and μ, unless explicitly stated otherwise:

(A1) The RKHS $\mathcal{H}$ is infinite-dimensional.

(A2) The kernel function k is continuous and bounded, i.e., $\kappa_1 := \sqrt{k(x,x)} < \infty$.

(A3) The RKHS $\mathcal{H}$ does *not* constant functions.

(A4) The base density μ is a continuously differentiable pdf over $\mathcal{X}$ with the support being $\mathcal{X}$.

If we assume $\mathcal{H}$ to be finite-dimensional, rather than infinite-dimensional as we have done in **(A1)**, we are returning to the case of finite-dimensional exponential families discussed in Section 2.1.1, as we have seen in Section 2.1.4. Classic theories on the estimation in $\mathcal{Q}_{\text{fin}}$ are directly applicable and are not very interesting. Thus, we rule out this case.

The main motivation of **(A2)** is to ensure $\mathcal{F} = \mathcal{H}$, as we have shown in Proposition 2.1. This is going to make all the optimization problems considered in Section 2.2 unconstrained. In addition, since $\mathcal{X} \subseteq \mathbb{R}^d$ and k is continuous, Lemma 4.33 in Steinwart and Christmann (2008) implies that $\mathcal{H}$ is separable and has an orthonormal basis (Theorem 11 in Royden and Fitzpatrick, 2018).

Assumptions **(A3)** and **(A4)** together ensure the identifiability of $\mathcal{Q}_{\text{ker}}$, i.e., $q_{f_1} = q_{f_2}$ if and only if $f_1 = f_2$. One sufficient condition that guarantees $\mathcal{H}$ does not contain

constant functions is that the kernel function k is continuous on $\mathcal{X} \times \mathcal{X}$ and vanishes at infinity (see Remark 3(iii) in Sriperumbudur et al., 2017).

2.2 Nonparametric Density Estimation in $\mathcal{Q}_{\text{ker}}$

We now turn to the nonparametric density estimation problem in $\mathcal{Q}_{\text{ker}}$. Borrowing the framework in Chapter 1, we consider the following minimization problem

$$\underset{q_f \in \mathcal{Q}_{\text{ker}}}{\text{minimize}} \left\{ \widehat{L}(q_f) + \frac{\lambda}{2} \widetilde{P}(f) \right\}. \tag{2.8}$$

The very first question we consider is how rich $\mathcal{Q}_{\text{ker}}$ is as a class of pdfs to estimate p_0. More specifically, we want to understand what class of density functions over $\mathcal{X}$ can be approximated arbitrarily well by those in $\mathcal{Q}_{\text{ker}}$. Proposition 1, Corollary 2 and Proposition 13 in Sriperumbudur et al. (2017) answered this question and state that, under certain regularity conditions, $\mathcal{Q}_{\text{ker}}$ can approximate arbitrarily well any continuous p_0 that vanishes at infinity under the KL-divergence, L^r norm with $r \in [1, \infty]$, the Hellinger distance, and the H-divergence. Hence, $\mathcal{Q}_{\text{ker}}$ is a rather rich class of density functions to estimate p_0.

In the rest of this section, we again consider the two loss functionals, $\widehat{L}_{\text{NLL}}$ and $\widehat{L}_{\text{SM}}$, and give a review of density estimation problem in $\mathcal{Q}_{\text{ker}}$ using them.

2.2.1 Density Estimation in $\mathcal{Q}_{\text{ker}}$ using $\widehat{L}_{\text{NLL}}$

We let $\widehat{L}$ in (2.8) to be the NLL loss functional $\widehat{L}_{\text{NLL}}$. Then, using (2.3), $\widehat{L}_{\text{NLL}}(q_f)$ becomes

$$\widehat{J}_{\text{NLL}}(f) := A(f) - \frac{1}{n} \sum_{i=1}^{n} f(X_i), \qquad \text{for all } f \in \mathcal{F},$$

up to an additive constant.

Since A is convex by Proposition 2.2, and $-\frac{1}{n}\sum_{i=1}^{n} f(X_i) = -\frac{1}{n}\sum_{i=1}^{n} \langle f, k(X_i, \cdot)\rangle_{\mathcal{H}}$ is linear and, hence, convex in f, their sum, $\widehat{J}_{\text{NLL}}$, is convex as well.

We first have a bad news stated in the following proposition.

Proposition 2.6 (Fukumizu (2005)). *Under* ***(A1)*** *and* ***(A2)****, minimizing* $\widehat{J}_{\mathrm{NLL}}$ *over* $\mathcal{H}$ *does* not *have a solution.*

The direct consequence of Proposition 2.6 is that the ML density estimator in $\mathcal{Q}_{\mathrm{ker}}$ does *not* exist.

Remark 2.2. Proposition 2.6 exemplifies a distinct difference between $\mathcal{Q}_{\mathrm{ker}}$ and $\mathcal{Q}_{\mathrm{fin}}$.

Suppose we estimate p_0 using elements in $\mathcal{Q}_{\mathrm{fin}}$ via maximizing the log-likelihood function

$$\left\langle \theta, \frac{1}{n}\sum_{i=1}^{n}\varphi(X_i)\right\rangle - B(\theta), \qquad \text{subject to } \theta \in \Theta. \tag{2.9}$$

Under certain regularity conditions, the maximizer of (2.9), denoted by $\hat{\theta}_{\mathrm{ML}}$, exists and is unique, and must satisfy the equation

$$\nabla B(\hat{\theta}_{\mathrm{ML}}) = \frac{1}{n}\sum_{i=1}^{n}\varphi(X_i),$$

which has the moment matching interpretation that the sample mean of the canonical statistic must match the population mean at the MLE, since $\nabla B(\hat{\theta}_{\mathrm{ML}}) = \int_{\mathcal{X}} \tilde{q}_{\hat{\theta}}(x)\varphi(x)\mathrm{d}x$. In particular, the inverse map of ∇B exists and

$$\hat{\theta}_{\mathrm{ML}} = (\nabla B)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(X_i)\right).$$

▶

With the result in Proposition 2.6, the density estimation problem in $\mathcal{Q}_{\mathrm{ker}}$ via minimizing $\widehat{J}_{\mathrm{NLL}}$ is ill-posed. In order to make the problem well-posed and obtain a solution, one has to impose certain kind of regularization. Serval ideas have been proposed in the literature to tackle this issue.

The first idea is to regularize the function space over which we minimize $\widehat{J}_{\mathrm{NLL}}$. Rather than minimize it over the entire $\mathcal{H}$, Fukumizu (2005) proposed to construct a sequence of nested finite-dimensional subspaces of $\mathcal{H}$ that enlarges with the sample size n, $\{\mathcal{H}^{(m_n)}\}_{m_n \in \mathbb{N}}$ satisfying $\mathcal{H}^{(m_n)} \subset \mathcal{H}^{(m_{n+1})}$ for all n, and minimize $\widehat{J}_{\mathrm{NLL}}$ over $\mathcal{H}^{(m_n)}$. Supposing $\hat{f}_{\mathrm{ML}}^{(m_n)} := \arg\min_{f \in \mathcal{H}^{(m_n)}} \widehat{J}_{\mathrm{NLL}}(f)$ exists for all n, Fukumizu (2005) showed that, if $p_0 \in \mathcal{Q}_{\mathrm{ker}}$, together with some additional assumptions, $q_{\hat{f}_{\mathrm{ML}}^{(m_n)}}$ is consistent for p_0 under the KL-divergence.

Even though this approach is theoretically interesting, it suffers from several theoretical and practical drawbacks. On the theoretical side, the consistency of $q_{\hat{f}_{\mathrm{ML}}^{(m_n)}}$ relies on the decay rate of the smallest eigenvalue of certain covariance operator, which is hard or even impossible to check in practice. On the practical side, Fukumizu (2005) did not elucidate guidelines on which class of RKHS should be used or how to choose the sequence of nested finite-dimensional subspaces. Moreover, even if such guidelines were provided, the minimization problem is nonlinear by its nature and one has to rely on an iterative optimization algorithm to compute $\hat{f}_{\mathrm{ML}}^{(m_n)}$. Then, it is inevitable to deal with A and its derivative, both of which involve integration over a possibly high-dimensional space and are hard to handle in practice. Thus, the density estimator constructed using this approach is not attractive.

A different approach proposed in the literature is to add a nonzero penalty functional $\widetilde{P}$ to $\widehat{J}_{\mathrm{NLL}}$ and minimize the penalized NLL loss functional. One such work was carried out by Gu and Qiu (1993) who chose $\widetilde{P}$ to be a square seminorm of $\mathcal{H}$. They showed the minimizer of $\widehat{J}_{\mathrm{NLL}}(f) + \lambda \widetilde{P}(f)$ with $f \in \mathcal{H}$ exists and is unique under very mild conditions. However, since $\mathcal{H}$ is infinite-dimensional, this minimizer is *not* computable. They proposed to minimize the penalized NLL loss functional over $\mathcal{H}_0 \oplus \widetilde{\mathcal{H}}_n$, where $\mathcal{H}_0 := \left\{ f \in \mathcal{H} \mid \widetilde{P}(f) = 0 \right\}$ is the null space of $\widetilde{P}$ and $\widetilde{\mathcal{H}}_n := \{ f \mid f :=$

$\sum_{i=1}^{n} \alpha_i k(X_i, \cdot), \alpha_1, \cdots, \alpha_n \in \mathbb{R}\}$ is an n-dimensional subspace of $\mathcal{H}$, and established asymptotic properties of $q_{\tilde{f}_{\mathrm{ML}}^{(\lambda)}}$, where $\tilde{f}_{\mathrm{ML}}^{(\lambda)} := \arg\min_{f \in \mathcal{H}_0 \oplus \widetilde{\mathcal{H}}_n} \{\widehat{J}_{\mathrm{NLL}}(f) + \lambda \widetilde{P}(f)\}$. Gu (1993) proposed an iterative algorithm to compute $\tilde{f}_{\mathrm{ML}}^{(\lambda)}$ and used the quadrature rule over a dense mesh to approximate A and its derivatives at each iteration. All numerical examples therein were limited to cases $d \leq 2$. If d is large, the computation would become prohibitively expensive and the approximations via this approach could be very poor.

In order to avoid working with A directly, Dai et al. (2018) proposed a doubly dual embedding approach. Suppose k satisfies $\int_{\mathcal{X}} k(x, x)\mathrm{d}x < \infty$ so that any $f \in \mathcal{H}$ is square-integrable. We then have

$$A(f) = \sup_{p \in \mathcal{P} \cap L^2(\mathcal{X})} \Big\{ \langle p, f \rangle_{L^2(\mathcal{X})} - \mathrm{KL}(p \| \mu) \Big\}, \qquad \text{for all } f \in \mathcal{H}, \tag{2.10}$$

$$\mathrm{KL}(p \| \mu) = \sup_{g \in \mathcal{H}} \Big\{ \langle p, g \rangle_{L^2(\mathcal{X})} - \int_{\mathcal{X}} e^{g(x)} \mu(x)\mathrm{d}x + 1 \Big\}, \quad \text{for all } p \in \mathcal{P} \cap L^2(\mathcal{X}), \tag{2.11}$$

where $\mathcal{P}$ denotes the set of all pdfs over $\mathcal{X}$, $L^2(\mathcal{X}) := \big\{f : \mathcal{X} \to \mathbb{R} \mid \int_{\mathcal{X}} (f(x))^2 \mathrm{d}x < \infty\big\}$, and $\langle f_1, f_2 \rangle_{L^2(\mathcal{X})} := \int_{\mathcal{X}} f_1(x) f_2(x) \mathrm{d}x$ for all $f_1, f_2 \in L^2(\mathcal{X})$. Plugging (2.10) and (2.11) successively into $\widehat{J}_{\mathrm{NLL}}(f) + \lambda \widetilde{P}(f)$, Dai et al. (2018) proposed to solve the following max-min problem

$$\underset{p \in \mathcal{P} \cap L^2(\mathcal{X})}{\text{maximize}} \ \underset{f, g \in \mathcal{H}}{\text{minimize}} \Big\{ -\frac{1}{n} \sum_{i=1}^{n} f(X_i) + \int_{\mathcal{X}} (f(x) - g(x)) p(x) \mathrm{d}x + \int_{\mathcal{X}} e^{g(x)} \mu(x) \mathrm{d}x + \lambda \widetilde{P}(f) \Big\},$$

and proposed a stochastic gradient ascent-descent algorithm to iterate between the inner and outer optimization problems until convergence. While this approach avoids directly working with A, it incurs additional computational burdens as, in order to compute the minimizer with respect to f, one has to compute the optimal solutions with respect to p and g at the same time.

2.2.2 Density estimation in $\mathcal{Q}_{\text{ker}}$ using $\widehat{L}_{\text{SM}}$

With the discussions in the preceding subsection, we see that the main difficulty in the approach via minimizing $\widehat{J}_{\text{NLL}}$ is that one has to deal with A and its derivative, which is computationally intractable in practice. The SM loss functional, as we have discussed in Chapter 1, can help us avoid this difficulty.

Recall that, under certain conditions, the SM loss functional, in its original form, is

$$\widehat{L}_{\text{SM}}(q) := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\big(\partial_u \log q(X_i)\big)^2 + \partial_u^2 \log q(X_i)\right). \tag{2.12}$$

If we let $q = q_f \in \mathcal{Q}_{\text{ker}}$ in (2.12), under certain additional regularity conditions that will be made explicit in Chapter 3, we can rewrite $\widehat{L}_{\text{SM}}$, up to additive constants, as

$$\begin{aligned}\widehat{J}_{\text{SM}}(f) &:= \frac{1}{2}\sum_{i=1}^{n}\sum_{u=1}^{d}(\partial_u f(X_i))^2 + \sum_{i=1}^{n}\sum_{u=1}^{d}\big(\partial_u \log \mu(X_i)\partial_u f(X_i) + \partial_u^2 f(X_i)\big)\\ &\overset{(\star)}{=} \frac{1}{2}\langle f, \widehat{C} f\rangle_{\mathcal{H}} - \langle f, \hat{z}\rangle_{\mathcal{H}},\end{aligned} \tag{2.13}$$

where $\widehat{C} : \mathcal{H} \to \mathcal{H}$ is given by

$$\widehat{C} := \frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\partial_u k(X_i, \cdot) \otimes \partial_u k(X_i, \cdot), \tag{2.14}$$

with $\widehat{C} f = \frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\partial_u f(X_i)\partial_u k(X_i, \cdot)$ for all $f \in \mathcal{H}$, and

$$\hat{z} := -\frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\big(\partial_u \log \mu(X_i)\partial_u k(X_i, \cdot) + \partial_u^2 k(X_i, \cdot)\big) \in \mathcal{H}. \tag{2.15}$$

In (2.14) and (2.15), with a fixed $x \in \mathcal{X}$, $\partial_u^s k(x, y) := \frac{\partial^s}{\partial w_u^s} k(x, y)\big|_{w=x}$, for all $y \in \mathcal{X}$, $s = 1, 2$, and $u = 1, \cdots, d$. In $(\star)$, we use $\partial_u^s k(x, \cdot) \in \mathcal{H}$ and the reproducing property of partial derivatives of k, $\partial_u^s f(x) = \langle f, \partial_u^s k(x, \cdot)\rangle_{\mathcal{H}}$, for $s = 1, 2$ and $u = 1, \cdots, d$ (Zhou, 2008). Details about the partial derivatives of k and their reproducing properties can be found in Section A.3 in Appendix A. In particular, notice that $\widehat{J}_{\text{SM}}$ does *not* involve A.

Now, $\widehat{J}_{\mathrm{SM}}$ is a quadratic functional in f. It is not hard to show that the operator $\widehat{C}$ is linear, self-adjoint, and positive semidefinite. Hence, minimizing $\widehat{J}_{\mathrm{SM}}$ over $\mathcal{H}$ is a convex optimization problem. Suppose that $\hat{f}_{\mathrm{SM}} := \arg\min_{f\in\mathcal{H}} \widehat{J}_{\mathrm{SM}}(f)$ exists. Then, $\hat{f}_{\mathrm{SM}}$ must satisfy the first-order optimality condition $\widehat{C}\hat{f}_{\mathrm{SM}} = \hat{z}$; in other words, minimizing $\widehat{J}_{\mathrm{SM}}$ amounts to solving an infinite-dimensional linear system, and $\hat{f}_{\mathrm{SM}} = \widehat{C}^{-1}\hat{z}$. However, since $\widehat{C}$ has finite rank and must be a compact operator in an infinite-dimensional RKHS $\mathcal{H}$, it is *not* invertible (Section 16.5 in Royden and Fitzpatrick, 2018) and, hence, $\hat{f}_{\mathrm{SM}}$ does *not* exist.

Thus, minimizing $\widehat{J}_{\mathrm{SM}}$ is an ill-posed problem. In order to remedy this, certain kind of regularization has to be imposed. Sriperumbudur et al. (2017) proposed to add a penalty term and minimize

$$\widehat{J}_{\mathrm{SM}}(f) + \frac{\rho}{2}\|f\|_{\mathcal{H}}^2, \qquad \text{subject to } f \in \mathcal{H},$$

where we use $\rho > 0$ to denote the penalty parameter associated with the SM loss functional. This is exactly the Tikhonov regularization. Sriperumbudur et al. (2017) showed this penalized SM loss functional has a unique minimizer given by

$$\hat{f}_{\mathrm{SM}}^{(\rho)} := \arg\min_{f\in\mathcal{H}} \left\{ \widehat{J}_{\mathrm{SM}}(f) + \frac{\rho}{2}\|f\|_{\mathcal{H}}^2 \right\} = (\widehat{C} + \lambda I)^{-1}\hat{z}, \tag{2.16}$$

where $I : \mathcal{H} \to \mathcal{H}$ denotes the identity operator in $\mathcal{H}$. In practice, however, it may not be easy to compute the minimizer in the form of (2.16) as it involves solving an infinite-dimensional linear system. With the help of a general representer theorem (Theorem A.2 in Sriperumbudur et al., 2017), it can be shown that

$$\hat{f}_{\mathrm{SM}}^{(\rho)} = \sum_{i=1}^{n}\sum_{u=1}^{d} \alpha_{(i-1)d+u}^{(\rho)} \partial_u k(X_i, \cdot) + \frac{1}{\rho}\hat{z},$$

where $\boldsymbol{\alpha}^{(\rho)} := (\alpha_1^{(\rho)}, \cdots, \alpha_{nd}^{(\rho)})^\top \in \mathbb{R}^{nd}$ can be obtained by solving the linear system

$$(\mathbf{G} + n\rho\mathbf{I}_{nd})\boldsymbol{\alpha}^{(\rho)} = \frac{1}{\rho}\mathbf{h},$$

the $((i-1)d+u,(j-1)+v)$-entry of $\mathbf{G} \in \mathbb{R}^{nd\times nd}$ is $\langle \partial_u k(X_i,\cdot), \partial_v k(X_j,\cdot)\rangle_{\mathcal{H}} = \partial_i\partial_{i+d}k(X_i,X_j)$, and $((i-1)d+u)$-entry of $\mathbf{h} \in \mathbb{R}^{nd}$ is

$$\langle \hat{z}, \partial_u \partial k(X_i,\cdot)\rangle_{\mathcal{H}} = -\frac{1}{n}\sum_{j=1}^{n}\sum_{v=1}^{d}\big(\partial_u\partial_{v+d}k(X_i,X_j)\partial_v \log\mu(X_j) + \partial_u\partial_{v+d}^2 k(X_i,X_j)\big),$$

and $\mathbf{I}_{nd}$ denotes the $nd\times nd$ identity matrix. Note that the matrix $\mathbf{G}$ is positive semidefinite and $n\rho\mathbf{I}_{nd}$ is positive definite, so their sum must be positive definite and must be invertible, and it follows that $\boldsymbol{\alpha}^{(\rho)} = \frac{1}{\rho}(\mathbf{G}+n\rho\mathbf{I}_{nd})^{-1}\mathbf{h}$.

Theoretical properties of the penalized SM density estimator $q_{\hat{f}_{\mathrm{SM}}^{(\rho)}}$ was studied by Sriperumbudur et al. (2017). If $p_0 \in \mathcal{Q}_{\mathrm{ker}}$, they established the consistency and convergence rate of $q_{\hat{f}_{\mathrm{SM}}^{(\rho)}}$ under the KL-divergence, the H-divergence, the Hellinger distance, and the total variation distance. If $p_0 \notin \mathcal{Q}_{\mathrm{ker}}$, they showed $q_{\hat{f}_{\mathrm{SM}}^{(\rho)}}$ converges to the element in $\mathcal{Q}_{\mathrm{ker}}$ that has the smallest H-divergence to p_0 and established the corresponding convergence rate under the H-divergence.

In their simulation studies, Sriperumbudur et al. (2017) showed their penalized SM density estimator outperforms the kernel density estimator (Section 6.3 in Wasserman, 2006), especially when d is large. However, no comparison with the (penalized) ML density estimator was conducted.

Note that computing $\hat{f}_{\mathrm{SM}}^{(\rho)}$ requires to solve a linear system of nd equations in nd variables, which requires elementary operations of order $\mathcal{O}(n^3d^3)$. This can be computationally expensive when n or d is large. To alleviate the computational cost, Sutherland et al. (2017) adopted the idea of the Nyström approximation (Williams and Seeger, 2001) and proposed to minimize $\widehat{J}_{\mathrm{SM}}(f) + \frac{\rho}{2}\|f\|_{\mathcal{H}}^2$ subject to

$$f \in \mathrm{Span}\Big\{\partial_u k(Y_j,\cdot) \text{ for all } j=1,\cdots,m \text{ and } u=1,\cdots,d\Big\},$$

where $Y_1,\cdots,Y_m$ are m randomly selected observations from $\{X_1,\cdots,X_n\}$ with $m \ll n$. Then, one only needs to work with a linear system of md equations in md variables,

which requires elementary operations of order $\mathcal{O}(m^3 d^3)$. This penalized SM density estimator is more efficient to compuate and empirically performs very close to the one proposed by Sriperumbudur et al. (2017).

2.3 Proofs

2.3.1 Proof of Proposition 2.1

Proof. Let $f \in \mathcal{H}$ be arbitrary. Due to the Cauchy-Schwartz inequality, we have $\left|\langle f, k(x, \cdot)\rangle_{\mathcal{H}}\right| \leq \|f\|_{\mathcal{H}}\|k(x, \cdot)\|_{\mathcal{H}} \leq \kappa_1\|f\|_{\mathcal{H}}$. Then, we can bound $A(f)$ from above by

$$A(f) \leq \log\left(\exp(\kappa_1\|f\|_{\mathcal{H}})\int_{\mathcal{X}}\mu(x)\mathrm{d}x\right) \overset{(\star)}{=} \kappa_1\|f\|_{\mathcal{H}} < \infty,$$

where we use the assumption that μ is a pdf over $\mathcal{X}$ in deriving $(\star)$. Since the choice of $f \in \mathcal{H}$ is arbitrary, we conclude $\mathcal{H} \subseteq \mathcal{F}$. On the other hand, it is obvious that $\mathcal{F} \subseteq \mathcal{H}$. We conclude that $\mathcal{F} = \mathcal{H}$. ∎

2.3.2 Proof of Proposition 2.2

Proof of Proposition 2.2. We first show the convexity of A. That is, we need to show, for any distinct $f, g \in \mathcal{F}$ and $\alpha \in [0, 1]$, the following inequality holds

$$A(\alpha f + (1-\alpha)g) \leq \alpha A(f) + (1-\alpha)A(g).$$

Notice that if $\alpha = 1$ or $\alpha = 0$, the inequality above becomes an equality and the result holds trivially.

Hence, we assume $\alpha \in (0, 1)$. Then,

$$\begin{aligned}
A(\alpha f + (1-\alpha)g) &= \log\left[\int_{\mathcal{X}}\big(\mu(x)\exp(f(x))\big)^{\alpha}\big(\mu(x)\exp(g(x))\big)^{1-\alpha}\mathrm{d}x\right] \\
&\overset{(\star)}{\leq} \log\left[\left(\int_{\mathcal{X}}\big(\mu(x)\exp(f(x))\big)^{\alpha\cdot\frac{1}{\alpha}}dx\right)^{\alpha}\cdot\left(\int_{\mathcal{X}}\big(\mu(x)\exp(g(x))\big)^{(1-\alpha)\cdot\frac{1}{1-\alpha}}\mathrm{d}x\right)^{1-\alpha}\right] \\
&= \alpha\log\left[\int_{\mathcal{X}}\mu(x)\exp(f(x))\mathrm{d}x\right] + (1-\alpha)\log\left[\int_{\mathcal{X}}\mu(x)\exp(g(x))\mathrm{d}x\right] \\
&= \alpha A(f) + (1-\alpha)A(g),
\end{aligned}$$

where $(\star)$ is due to the Hölder's inequality. This established the convexity of A.

The inequality ($\star$) becomes an equality if and only if there exist $\beta_1 > 0$ and $\beta_2 > 0$ such that $\beta_1 \exp(f(x)) = \beta_2 \exp(g(x))$ for almost all $x \in \mathcal{X}$, or, equivalently,

$$f(x) - g(x) = \langle f - g, k(x, \cdot)\rangle_{\mathcal{H}} = \log \beta_2 - \log \beta_1, \qquad \text{for almost all } x \in \mathcal{X}.$$

Now, if $\mathcal{H}$ does *not* contain constant functions, $f - g$ cannot be a constant function, meaning that the equality cannot hold. Thus, A is strictly convex.

Finally, we show the convexity of $\mathcal{F}$. Let $f, g \in \mathcal{F}$ so that $A(f) < \infty$ and $A(g) < \infty$. By the proof above, we have $A(\alpha f + (1-\alpha)g) < \infty$, i.e, $\alpha f + (1-\alpha)g \in \mathcal{F}$. In other words, $\mathcal{F}$ is convex. ∎

2.3.3 Proof of Lemma 2.1

Proof of Lemma 2.1. Let $f \in \mathcal{H}$ be arbitrary and $g \in \mathcal{H}$ with $g \neq 0$. Note that

$$J_x(f + g) - J_x(f) = \langle f + g, k(x, \cdot)\rangle_{\mathcal{H}} - \langle f, k(x, \cdot)\rangle_{\mathcal{H}} = \langle g, k(x, \cdot)\rangle_{\mathcal{H}}.$$

Then,

$$\frac{\left|J_x(f + g) - J_x(f) - \langle g, k(x, \cdot)\rangle_{\mathcal{H}}\right|}{\|g\|_{\mathcal{H}}} = 0,$$

and, by Definition A.1 in Appendix A, we conjecture $\mathrm{D}J_x(f)(g) = \langle g, k(x, \cdot)\rangle_{\mathcal{H}}$, for all $g \in \mathcal{H}$.

We next verify that the map $\mathrm{D}J_x(f)$ is linear and bounded. The linearity part follows from the linearity of inner product and is omitted. To establish the boundedness, we have, for all $g \in \mathcal{H}$,

$$|\mathrm{D}J_x(f)(g)| = |\langle g, k(x, \cdot)\rangle_{\mathcal{H}}| \leq \|g\|_{\mathcal{H}} \|k(x, \cdot)\|_{\mathcal{H}} \leq \kappa_1 \|g\|_{\mathcal{H}},$$

where $\kappa_1 := \sup_{x \in \mathcal{X}} \sqrt{k(x, x)} < \infty$, by the assumption. Hence, $\mathrm{D}J_x(f)$ is a bounded linear operator. We conclude that J_x is Fréchet differentiable at $f \in \mathcal{H}$ with the

Fréchet derivative at $f \in \mathcal{H}$ being $\mathrm{D}J_x(f)(g) = \langle g, k(x, \cdot)\rangle_{\mathcal{H}}$ for all $g \in \mathcal{H}$. Since the choice of $f \in \mathcal{H}$ is arbitrary, we conclude J_x is Fréchet differentiable over $\mathcal{H}$.

Using the definition of Fréchet gradient, it is easy to see $\nabla J_x(f) = k(x, \cdot)$ for all $f \in \mathcal{H}$. ∎

2.3.4 Proof of Proposition 2.3

Proof of Proposition 2.3. Observe that $A(f) = (J_2 \circ J_1)(f)$ for all $f \in \mathcal{H}$, where

$$\begin{aligned} J_2(x) &:= \log x && \text{for all } x > 0, \\ J_1(f) &:= \int_{\mathcal{X}} \mu(x) \exp(f(x)) \mathrm{d}x && \text{for all } f \in \mathcal{H}. \end{aligned}$$

Since $J_1(f) > 0$ for all $f \in \mathcal{H}$, this composition is well-defined.

Let $f \in \mathcal{H}$ be arbitrary. We first show J_1 is Fréchet differentiable at $f \in \mathcal{H}$ with the Fréchet derivative

$$\mathrm{D}J_1(f)(g) = \left\langle g, \int_{\mathcal{X}} \mu(x) \exp(f(x)) k(x, \cdot) \mathrm{d}x \right\rangle_{\mathcal{H}}, \qquad \text{for all } g \in \mathcal{H}.$$

Note that the integrand of $\int_{\mathcal{X}} \mu(x) \exp(f(x)) k(x, \cdot) \mathrm{d}x$ is an element in $\mathcal{H}$, and

$$\begin{aligned} \left\| \int_{\mathcal{X}} \mu(x) \exp(f(x)) k(x, \cdot) \mathrm{d}x \right\|_{\mathcal{H}} &\leq \int_{\mathcal{X}} \mu(x) \exp(f(x)) \|k(x, \cdot)\|_{\mathcal{H}} \mathrm{d}x \\ &\leq \kappa_1 \exp(A(f)) < \infty, \end{aligned}$$

where $\kappa_1 := \sup_{x \in \mathcal{X}} \sqrt{k(x, x)} < \infty$ by the assumption. Hence, $\int_{\mathcal{X}} \mu(x) \exp(f(x)) k(x, \cdot) \mathrm{d}x$ is Bochner integrable with respect to the Lebesgue measure (by Proposition A.3 in Appendix A), and we can interchange the inner product and the integral (by Proposition A.4(c) in Appendix A) and have

$$\mathrm{D}J_1(f)(g) = \int_{\mathcal{X}} \mu(x) \exp(f(x)) g(x) \mathrm{d}x, \qquad \text{for all } g \in \mathcal{F},$$

Now, let $0 \neq g \in \mathcal{H}$. We have

$$J_1(f + g) - J_1(f) - \mathrm{D}J_1(f)(g)$$

$$
\begin{aligned}
&= \int_{\mathcal{X}} \mu(x) \exp\big(f(x)\big) \Big[\exp(g(x)) - 1 - g(x)\Big] \mathrm{d}x \\
&= \int_{\mathcal{X}} \mu(x) \exp(f(x)) \left[\sum_{j=0}^{\infty} \frac{(g(x))^m}{m!} - 1 - g(x)\right] \mathrm{d}x \\
&= \int_{\mathcal{X}} \mu(x) \exp(f(x)) \left[\sum_{m=2}^{\infty} \frac{(g(x))^m}{m!}\right] \mathrm{d}x \\
&\overset{\text{(i)}}{\leq} \int_{\mathcal{X}} \mu(x) \exp\big(\|f\|_{\mathcal{H}} \sqrt{k(x,\,\cdot\,)}\big) \left[\sum_{m=2}^{\infty} \frac{\|g\|_{\mathcal{H}}^m k(x,x)^{m/2}}{m!}\right] \mathrm{d}x \\
&\overset{\text{(ii)}}{\leq} \int_{\mathcal{X}} \mu(x) \exp(\kappa_1 \|f\|_{\mathcal{H}}) \left[\sum_{m=2}^{\infty} \frac{\|g\|_{\mathcal{H}}^m \kappa_1^{m/2}}{m!}\right] \mathrm{d}x \\
&\overset{\text{(iii)}}{=} \exp(\kappa_1 \|f\|_{\mathcal{H}}) \left[\sum_{m=2}^{\infty} \frac{\|g\|_{\mathcal{H}}^m \kappa_1^{m/2}}{m!}\right],
\end{aligned}
$$

where (i) follows from the Cauchy-Schwartz inequality, (ii) is due to $\sqrt{k(x,x)} \leq \kappa_1$, and (iii) is because μ is a density function over $\mathcal{X}$. To proceed, we have

$$
\frac{\big|J_1(f+g) - J_1(f) - \mathrm{D}J_1(f)(g)\big|}{\|g\|_{\mathcal{H}}} \leq \exp(\kappa_1 \|f\|_{\mathcal{H}}) \left[\sum_{m=2}^{\infty} \frac{\|g\|_{\mathcal{H}}^{m-1} \kappa_1^{m/2}}{m!}\right] \to 0,
$$

as $\|g\|_{\mathcal{H}} \to 0$.

Furthermore, we need to show $\mathrm{D}J_1(f)$ is linear and bounded. The linearity follows from that of the inner product and is omitted. To show the boundedness, we let $g \in \mathcal{H}$ be arbitrary and notice

$$
\begin{aligned}
\big|\mathrm{D}J_1(f)(g)\big|_{\mathcal{H}} &\leq \|g\|_{\mathcal{H}} \int_{\mathcal{X}} \mu(x) \exp\big(\|f\|_{\mathcal{H}} \sqrt{k(x,x)}\big) \sqrt{k(x,x)} \mathrm{d}x \\
&\leq \Big[\kappa_1 \exp(\kappa_1 \|f\|_{\mathcal{H}})\Big] \|g\|_{\mathcal{H}},
\end{aligned}
$$

from which we conclude that $\mathrm{D}J_1(f)$ is a bounded operator. Thus, J_1 is Fréchet differentiable at $f \in \mathcal{H}$ with the desired Fréchet derivative.

Since $A = J_2 \circ J_1$, applying Proposition A.2(b) in Appendix A, we obtain, for any $f \in \mathcal{H}$,

$$
\mathrm{D}A(f)(g) = \frac{1}{J_1(f)} \mathrm{D}J_1(f)(g)
$$

$$= \frac{1}{\exp(A(f))} \left\langle g, \int_{\mathcal{X}} \mu(x) \exp(\langle f, k(x, \cdot)\rangle_{\mathcal{H}}) k(x, \cdot) \mathrm{d}x \right\rangle_{\mathcal{H}}$$
$$= \left\langle g, \int_{\mathcal{X}} q_f(x) k(x, \cdot) \mathrm{d}x \right\rangle_{\mathcal{H}}.$$

Since our choice of $f \in \mathcal{H}$ is arbitrary, we conclude A is Fréchet differentiable over $\mathcal{H}$.

Finally, by the definition of Fréchet gradient operator, we conclude that the Fréchet gradient operator is $\int_{\mathcal{X}} q_f(x) k(x, \cdot) \mathrm{d}x$ as claimed. This completes the proof.

■

2.3.5 Proof of Proposition 2.5

Proof of Proposition 2.5. We will show the desired result by the three steps:

(a) show $(\mathcal{H}_0, \langle \cdot, \cdot \rangle_{\mathcal{H}_0})$ is an inner product space,

(b) show $(\mathcal{H}_0, \langle \cdot, \cdot \rangle_{\mathcal{H}_0})$ is a Hilbert space, and

(c) show $(\mathcal{H}_0, \langle \cdot, \cdot \rangle_{\mathcal{H}_0})$ is a RKHS.

(a) To show $(\mathcal{H}_0, \langle \cdot, \cdot \rangle_{\mathcal{H}_0})$ is an inner product space, we verify using the definition of an inner product space. Let $f = \sum_{j=1}^m \alpha_j \varphi_j \in \mathcal{H}_0$, $g = \sum_{j=1}^m \beta_j \varphi_j \in \mathcal{H}_0$, $h = \sum_{j=1}^m \gamma_j \varphi_j \in \mathcal{H}_0$, $\alpha_j, \beta_j, \gamma_j \in \mathbb{R}$ for all $j = 1, \cdots, m$, and $a, b \in \mathbb{R}$ be arbitrary. Then, the operation $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ is

(i) *symmetric*, since $\langle f, g \rangle_{\mathcal{H}_0} = \sum_{j=1}^m \alpha_j \beta_j = \sum_{j=1}^m \beta_j \alpha_j = \langle g, f \rangle_{\mathcal{H}_0}$;

(ii) *linear*, since

$$\begin{aligned}
\langle af + bg, h \rangle_{\mathcal{H}_0} &= \left\langle a \sum_{j=1}^m \alpha_j \varphi_j + b \sum_{j=1}^m \beta_j \varphi_j, \sum_{j=1}^m \gamma_j \varphi_j \right\rangle_{\mathcal{H}_0} \\
&= \left\langle \sum_{j=1}^m (a\alpha_j + b\beta_j) \varphi_j, \sum_{j=1}^m \gamma_j \varphi_j \right\rangle_{\mathcal{H}_0} \\
&= \sum_{j=1}^m (a\alpha_j + b\beta_j) \gamma_j
\end{aligned}$$

$$= a\sum_{j=1}^{m}\alpha_j\gamma_j + b\sum_{j=1}^{m}\beta_j\gamma_j$$
$$= a\langle f, h\rangle_{\mathcal{H}_0} + b\langle g, h\rangle_{\mathcal{H}_0};$$

(iii) *positive definite*, since $\langle f, f\rangle_{\mathcal{H}_0} = \sum_{j=1}^{m}\alpha_j^2 \geq 0$ for all $f \in \mathcal{H}_0$, and $\langle f, f\rangle_{\mathcal{H}_0} = 0$ if and only if $\alpha_j = 0$ for all $j = 1, \cdots, m$, implying $f = 0$.

(b) To show $(\mathcal{H}_0, \langle\,\cdot\,,\,\cdot\,\rangle_{\mathcal{H}_0})$ is a Hilbert space, first note that the inner product space $(\mathcal{H}_0, \langle\,\cdot\,,\,\cdot\,\rangle_{\mathcal{H}_0})$ is m-dimensional. The desired result follows directly from the fact that any finite-dimensional inner product space over $\mathbb{R}$ is complete and the definition that a Hilbert space is a complete inner product space.

(c) Finally, to show $(\mathcal{H}_0, \langle\,\cdot\,,\,\cdot\,\rangle_{\mathcal{H}_0})$ is a RKHS with the reproducing kernel (2.7), we let $x \in \mathcal{X}$ be fixed and $E_x : \mathcal{H}_0 \to \mathbb{R}$ be the evaluation functional, i.e., $E_x(f) = f(x)$ for all $f \in \mathcal{H}_0$, and need to show E_x is a bounded operator.

To this end, let $f, g \in \mathcal{H}_0$ be arbitrary and $a, b \in \mathbb{R}$, and note the following

$$|E_x(f)| = |f(x)| = \left|\sum_{j=1}^{m}\alpha_j\varphi_j(x)\right| \leq \sqrt{\sum_{j=1}^{m}\alpha_j^2}\sqrt{\sum_{j=1}^{m}(\varphi_j(x))^2}$$
$$= \|f\|_{\mathcal{H}_0}\sqrt{\sum_{j=1}^{m}(\varphi_j(x))^2} = M_x\|f\|_{\mathcal{H}_0},$$

where $M_x := \sqrt{\sum_{j=1}^{m}(\varphi_j(x))^2} < \infty$. Hence, the evaluation functional E_x is a bounded operator, from which we conclude $\mathcal{H}_0$ is a RKHS.

Finally, we identify the reproducing kernel associated with $\mathcal{H}_0$. Let $k(x, \cdot\,) = \sum_{j=1}^{m}\varphi_j(x)\varphi_j \in \mathcal{H}_0$, so that $k(x, y)$ is given by (2.7). Letting $f = \sum_{j=1}^{m}\alpha_j\varphi_j \in \mathcal{H}_0$, we have

$$\langle f, k(x, \cdot\,)\rangle_{\mathcal{H}_0} = \left\langle \sum_{j=1}^{m}\alpha_j\varphi_j, \sum_{j=1}^{m}\varphi_j(x)\varphi_j \right\rangle_{\mathcal{H}_0} = \sum_{j=1}^{m}\alpha_j\varphi_j(x) = f(x).$$

Thus, k defined in (2.7) satisfies the reproducing property and is the reproducing kernel of $\mathcal{H}_0$. ∎