Appendix A: Math Background

A.1 Fréchet Differentiability and Derivative

We provide details on the Fréchet differentiability and derivative. Throughout this section, we let $\mathcal{H}$ be a real Hilbert space, $J : \mathcal{H} \to \mathbb{R}$ be a map, $\mathcal{B}(\mathcal{H}, \mathbb{R})$ denote the collection of all bounded linear operators from $\mathcal{H}$ to $\mathbb{R}$, and, similarly, $\mathcal{B}(\mathcal{H}, \mathcal{H})$ denote the collection of all bounded linear operators from $\mathcal{H}$ to itself. All materials of this section come from Section 2.6 in Bauschke and Combettes (2011) and Section 5.1 in Denkowski, Migórski, and Papageorgiou (2013).

Definition A.1 (Fréchet differentiability and derivative)**.** The map J is said to be *(first-order) Fréchet differentiable* at $f \in \mathcal{H}$ if there exists an operator $\mathrm{D}J(f) \in \mathcal{B}(\mathcal{H}, \mathbb{R})$ such that

$$\lim_{\substack{\|g\|_{\mathcal{H}} \to 0 \\ g \neq 0}} \frac{\big|J(f+g) - J(f) - \mathrm{D}J(f)(g)\big|}{\|g\|_{\mathcal{H}}} = 0, \tag{A.1}$$

and the operator $\mathrm{D}J(f)$ is called the *(first-order) Fréchet derivative*. The map J is said to be *(first-order) Fréchet differentiable on* $\mathcal{H}$ if it is Fréchet differentiable at all $f \in \mathcal{H}$.

Proposition A.1. *Suppose J is Fréchet differentiable at $f \in \mathcal{H}$ and the Fréchet derivative $\mathrm{D}J(f)$ exists. Then, $\mathrm{D}J(f)$ is unique.*

Remark A.1. If J is Fréchet differentiable at $f \in \mathcal{H}$, we then can write

$$J(f+g) = J(f) + \mathrm{D}J(f)(g) + o(\|g\|_{\mathcal{H}}), \tag{A.2}$$

for all $g \in \mathcal{H}$ in a small neighborhood of the origin, where $o(\|g\|_{\mathcal{H}})$ denotes $\frac{o(\|g\|_{\mathcal{H}})}{\|g\|_{\mathcal{H}}} \to 0$ as $\|g\|_{\mathcal{H}} \to 0$. Thus, from (A.2), we see $J(f) + \mathrm{D}J(f)(g)$ provides the best linear approximation of J in a small neighborhood of f, which is the similar interpretation of the derivative of a real-valued function of a single variable. ▶

Frechét derivative shares many properties of the derivative of a real-valued function of a single variable. The following proposition lists two properties we use in studying the Frechét differentiability and deriving the Frechét derivative of the log-partition functional A in Chapter 2.

Proposition A.2.

(a) Suppose $J_1, J_2 : \mathcal{H} \to \mathbb{R}$ are Frechét differentiable at $f \in \mathcal{H}$ and $\alpha, \beta \in \mathbb{R}$ are arbitrary. Then, $\alpha J_1 + \beta J_2$ is also Frechét differentiable at $f \in \mathcal{H}$, and

$$\mathrm{D}(\alpha J_1 + \beta J_2)(f) = \alpha \mathrm{D}J_1(f) + \beta \mathrm{D}J_2(f).$$

(b) (Chain rule) Suppose $J_1 : \mathcal{H} \to \mathbb{R}$ is Frechét differentiable at $f \in \mathcal{H}$ and $J_2 : \mathbb{R} \to \mathbb{R}$ is differentiable at $J_1(f)$. Then, $J_2 \circ J_1 : \mathcal{H} \to \mathbb{R}$ is Frechét differentiable at $f \in \mathcal{H}$, and

$$\mathrm{D}(J_2 \circ J_1)(f) = J_2'(J_1(f))\mathrm{D}J_1(f). \tag{A.3}$$

Definition A.2 (Fréchet gradient)**.** Suppose $J : \mathcal{H} \to \mathbb{R}$ is Frechét differentiable at $f \in \mathcal{H}$. Since $\mathrm{D}J(f)$ is a bounded linear map from $\mathcal{H}$ to $\mathbb{R}$, the Riesz-Fréchet representation theorem (Fact 2.24 in Bauschke and Combettes, 2011) guarantees there exists a unique element $\nabla J(f) \in \mathcal{H}$ such that, for any $g \in \mathcal{H}$,

$$\mathrm{D}J(f)(g) = \langle g, \nabla J(f) \rangle_{\mathcal{H}}, \tag{A.4}$$

and $\nabla J(f)$ is called the *Fréchet gradient* of J at f. If J is Fréchet differentiable on $\mathcal{H}$, the *Fréchet gradient operator* is defined to be $\nabla J : \mathcal{H} \to \mathcal{H}, f \mapsto \nabla J(f)$.

Remark A.2. Note that $\mathrm{D}J(f)$ is a bounded linear map from $\mathcal{H}$ to $\mathbb{R}$, and belongs to the dual space of $\mathcal{H}$, denoted by $\mathcal{H}^*$. Since we have $\mathrm{D}J(f)(g) = \langle \nabla J(f), g \rangle_{\mathcal{H}}$, the Riesz-Fréchet representation theorem implies that $\|\mathrm{D}J(f)\|_{\mathcal{H}^*} = \|\nabla J(f)\|_{\mathcal{H}}$, where $\|\cdot\|_{\mathcal{H}^*}$ denotes the norm of the dual space $\mathcal{H}^*$. ▶

We now extend Definition A.1 to higher orders.

Definition A.3 (Higher-order Fréchet differentiability and derivatives)**.** Higher-order Fréchet differentiability and derivatives are defined inductively.

In particular, the map J is said to be *twice Fréchet differentiable at* $f \in \mathcal{H}$ if J itself is Fréchet differentiable at $f \in \mathcal{H}$ and the map $\mathrm{D}J(f) : \mathcal{H} \to \mathbb{R}$ is also Fréchet differentiable at $f \in \mathcal{H}$. The *second Fréchet derivative* of J at $f \in \mathcal{H}$, denoted by $\mathrm{D}^2 J(f)$, is an operator from $\mathcal{H}$ to $\mathcal{B}(\mathcal{H}, \mathbb{R})$, that satisfies

$$\lim_{\substack{\|g\|_{\mathcal{H}} \to 0 \\ g \neq 0}} \frac{\left\| \mathrm{D}J(f+g) - \mathrm{D}J(f) - \mathrm{D}^2 J(f)(g) \right\|_{\mathcal{H}^*}}{\|g\|_{\mathcal{H}}} = 0, \tag{A.5}$$

where $\|\cdot\|_{\mathcal{H}^*}$ denotes the norm of the dual space of $\mathcal{H}$. The map J is said to be *twice Fréchet differentiable on* $\mathcal{H}$ if it is twice Fréchet differentiable at all $f \in \mathcal{H}$.

Suppose J is twice Fréchet differentiable on $\mathcal{H}$. The *second-order Fréchet gradient operator*, denoted by $\nabla^2 J$, is a bounded linear operator that maps from $\mathcal{H}$ to $\mathcal{B}(\mathcal{H}, \mathcal{H})$ and satisfies

$$\mathrm{D}^2 J(f)(g)(h) = \langle h, \nabla^2 J(f)(g) \rangle_{\mathcal{H}}, \qquad \text{for all } f, g, h \in \mathcal{H}.$$

In other words, $\nabla^2 J \in \mathcal{B}(\mathcal{H}, \mathcal{B}(\mathcal{H}, \mathcal{H}))$ and $\nabla^2 J(f) \in \mathcal{B}(\mathcal{H}, \mathcal{H})$ for all $f \in \mathcal{H}$.

A.2 Bochner Integral

In this section, we present the definition of the Bochner integral, which is the extension of the Lebesgue integral of real-valued functions to the integral of functions taking values in a Banach space. We also present some properties of the Bochner integral that we have used in the dissertation (in particular, in Chapter 2 and 3). All materials of this section come from Appendix A.5.3 in Steinwart and Christmann (2008) and Section 3.10 Denkowski, Migórski, and Papageorgiou (2013).

Throughout this section, let $\mathcal{E}$ be a Banach space whose norm is denoted by $\|\cdot\|_{\mathcal{E}}$, and $(\mathcal{X}, \Sigma, \mu)$ be a σ-finite measure space (note that this μ differs from the one in the definition of finite-dimensional and kernel exponential families in Chapter 2). We first define the simple function (Definition A.4) and the measurable function (Definition A.5) in the Banach space setting and then define the Bochner μ-integral (Definition A.6).

Definition A.4 ($\mathcal{E}$-valued simple function)**.** A function $s : \mathcal{X} \to \mathcal{E}$ is said to be an *$\mathcal{E}$-valued simple function* if there exist $e_1, \cdots, e_n \in \mathcal{E}$ and $A_1, \cdots, A_n \in \Sigma$ such that

$$s(x) = \sum_{i=1}^{n} \mathbb{1}_{A_i}(x) e_i, \qquad \text{for all } x \in \mathcal{X},$$

where $\mathbb{1}_A$ is the indicator function of the set A, and is equal to 1 if $x \in A$ and to 0, otherwise.

Definition A.5 ($\mathcal{E}$-valued measurable function)**.** A function $f : \mathcal{X} \to \mathcal{E}$ is said to be an *$\mathcal{E}$-valued measurable function* if there exists a sequence of $\mathcal{E}$-valued simple functions, $\{s_n\}_{n \in \mathbb{N}}$, such that

$$\lim_{n \to \infty} \|f(x) - s_n(x)\|_{\mathcal{E}} = 0 \tag{A.6}$$

holds for all $x \in \mathcal{X}$.

Definition A.6 (Bochner μ-integral)**.** An $\mathcal{E}$-valued measurable function $f : \mathcal{X} \to \mathcal{E}$ is said to be *Bochner μ-integrable* if there exists a sequence of $\mathcal{E}$-valued simple functions, $\{s_n\}_{n\in\mathbb{N}}$, such that

$$\lim_{n\to\infty} \int_{\mathcal{X}} \|s_n(x) - f(x)\|_{\mathcal{E}} \, \mathrm{d}\mu(x) = 0. \tag{A.7}$$

In this case, the limit

$$\int_{\mathcal{X}} f(x)\mathrm{d}\mu(x) := \lim_{n\to\infty} \int_{\mathcal{X}} s_n(x)\mathrm{d}\mu(x)$$

exists and is called the *Bochner μ-integral* of f.

A criterion to check the Bochner μ-integrability is the following.

Proposition A.3. *A measurable function $f : \mathcal{X} \to \mathcal{E}$ is Bochner μ-integrable if and only if $\int_{\mathcal{X}} \|f(x)\|_{\mathcal{E}} \mathrm{d}\mu(x) < \infty$.*

Finally, we look at some properties of Bochner μ-integral we use.

Proposition A.4. *The Bochner μ-integral defined above has the following properties:*

(a) The Bochner μ-integral is linear.

(b) If $f : \mathcal{X} \to \mathcal{E}$ is Bochner μ-integrable, we have

$$\left\| \int_{\mathcal{X}} f(x)\mathrm{d}\mu(x) \right\|_{\mathcal{E}} \leq \int_{\mathcal{X}} \|f(x)\|_{\mathcal{E}} \mathrm{d}\mu(x).$$

(c) Suppose $\mathcal{E}'$ is another Banach space. If $S : \mathcal{E} \to \mathcal{E}'$ is a bounded linear operator and $f : \mathcal{X} \to \mathcal{E}$ is Bochner μ-integrable, then $S \circ f : \mathcal{X} \to \mathcal{E}'$ is also Bochner μ-integrable. In this case, the integral commutes with S, that is,

$$S\left(\int_{\mathcal{X}} f(x)\mathrm{d}\mu(x) \right) = \int_{\mathcal{X}} (S \circ f)(x) \, \mathrm{d}\mu(x).$$

A.3 Partial Derivative of a Kernel Function

In this section, we discuss the partial derivatives of the kernel function associated with a RKHS and their reproducing property. We follow the development in Section 4.3 in Steinwart and Christmann (2008) and the paper by Zhou (2008). Throughout this section, we let $\mathcal{X} \subseteq \mathbb{R}^d$ be an open set and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$.

We first consider a real-valued function $f : \mathcal{X} \to \mathbb{R}$. The function f is said to be *m-times continuously differentiable* if, for all $\alpha := (\alpha_1, \cdots, \alpha_d)^\top \in \mathbb{N}_0^d$ with $|\alpha| := \sum_{i=1}^d \alpha_i \leq m$ and all $x \in \mathcal{X}$,

$$\partial^\alpha f(x) = \frac{\partial^{|\alpha|}}{\partial u_1^{\alpha_1} \cdots \partial u_d^{\alpha_d}} f(u)\bigg|_{u=x},$$

exists, where $u := (u_1, \cdots, u_d)^\top \in \mathcal{X}$.

We then define the m-times continuous differentiability of the kernel function k.

Definition A.7 (m-times continuous differentiability of a kernel function)**.** Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel function and $m \in \mathbb{N}$. We say k is *m-times continuously differentiable* if $\partial^{\alpha,\alpha} k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ exists and is continuous for all $\alpha := (\alpha_1, \cdots, \alpha_d) \in \mathbb{N}_0^d$ with $|\alpha| := \sum_{i=1}^d \alpha_i \leq m$, where

$$\partial^{\alpha,\alpha} k(x, y) = \frac{\partial^{2|\alpha|}}{\partial u_1^{\alpha_1} \cdots \partial u_d^{\alpha_d} \partial v_1^{\alpha_1} \cdots \partial v_d^{\alpha_d}} k(u, v)\bigg|_{u=x, v=y}, \qquad \text{for all } x, y \in \mathcal{X}.$$

The partial derivative of k is an element in $\mathcal{H}$ and has the reproducing property as k does, as the following proposition states.

Proposition A.5 (Partial derivatives of kernels and its reproducing property)**.** *Let $\mathcal{H}$ be a RKHS with the kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and assume k is m-times continuously differentiable on $\mathcal{X}$. Then,*

(a) we have

$$\partial^\alpha k(x, \cdot) = \frac{\partial^{|\alpha|}}{\partial u_1^{\alpha_1} \cdots \partial u_d^{\alpha_d}} k(u, \cdot)\bigg|_{u=x} \in \mathcal{H} \tag{A.8}$$

where $u := (u_1, \cdots, u_d)^\top \in \mathcal{X}$, *and*

(b) every $f \in \mathcal{H}$ *is* m*-times continuously differentiable, and, for all* $\alpha \in \mathbb{N}_0^d$ *with* $|\alpha| \leq m$ *and all* $x \in \mathcal{X}$, *the partial derivative reproducing property holds, i.e.,*

$$\partial^\alpha f(x) = \langle \partial^\alpha k(x, \cdot), f \rangle_{\mathcal{H}}, \qquad \text{for all } x \in \mathcal{X}. \tag{A.9}$$

In particular, we have $\partial^{\alpha,\alpha} k(x, y) = \langle \partial^\alpha k(x, \cdot), \partial^\alpha k(y, \cdot) \rangle_{\mathcal{H}}$ *for all* $x, y \in \mathcal{X}$.

A.4 Some Theories on Bounded Linear Operators

In our development in Chapters 2 and 3, we have used some theories on the bounded linear operators in a Hilbert space. In this section, we aim to give a very brief overview of these theories. More information can be found in, for example, Chapters 13 and 16 in Royden and Fitzpatrick (2018) and Chapter VI in Reed and Simon (2012).

Throughout this section, we let $\mathcal{H}$ be a real Hilbert space with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and the norm $\| \cdot \|_{\mathcal{H}}$, and assume $\mathcal{H}$ is separable, meaning that it contains a dense countable subset. Due to the separability of $\mathcal{H}$, it admits a countable orthonormal basis.

Definition A.8 (Linear operator)**.** An operator $C : \mathcal{H} \to \mathcal{H}$ is said to be *linear* if, for any $f, g \in \mathcal{H}$ and any $\alpha, \beta \in \mathbb{R}$, we have

$$C(\alpha f + \beta g) = \alpha C f + \beta C g.$$

Definition A.9 (Bounded operator)**.** A linear operator $C : \mathcal{H} \to \mathcal{H}$ is said to be *bounded* if there exists a constant $M \geq 0$ such that

$$\|Cf\|_{\mathcal{H}} \leq M \cdot \|f\|_{\mathcal{H}}, \qquad \text{for all } f \in \mathcal{H}.$$

The infimum of all such M is called the *operator norm* of C and is denoted by $\|C\|$.

Definition A.10 (Adjoint and self-adjoint operators)**.** Let $C : \mathcal{H} \to \mathcal{H}$ be a bounded linear operator. Then, the *adjoint* of C is the bounded linear operator $C^* : \mathcal{H} \to \mathcal{H}$ satisfying

$$\langle Cf, g\rangle_{\mathcal{H}} = \langle f, C^* g\rangle_{\mathcal{H}}, \qquad \text{for all } f, g \in \mathcal{H}.$$

The adjoint C^* exists and is unique.

The operator C is said to be *self-adjoint*, if $C = C^*$, that is, $\langle Cf, g\rangle_{\mathcal{H}} = \langle f, Cg\rangle_{\mathcal{H}}$ for all $f, g \in \mathcal{H}$.

Definition A.11 (Positive semidefinite and definite operators)**.** Let $C : \mathcal{H} \to \mathcal{H}$ be a self-adjoint bounded linear operator. Then, C is said to be *positive semidefinite* if, for all $f \in \mathcal{H}$, $\langle f, Cf\rangle_{\mathcal{H}} \geq 0$, and to be *positive definite* if, for all $f \in \mathcal{H}\backslash\{0\}$, $\langle f, Cf\rangle_{\mathcal{H}} > 0$.

Definition A.12 (Compact operator)**.** A bounded linear operator $C : \mathcal{H} \to \mathcal{H}$ is said to be *compact*, if, for every bounded sequence $\{f_n\}_{n\in\mathbb{N}}$ in $\mathcal{H}$, $\{Cf_n\}_{n\in\mathbb{N}}$ has a subsequence that converges in $\mathcal{H}$ (with respect to the norm $\|\cdot\|_{\mathcal{H}}$).

Other equivalent ways of defining a compact operator can be found, for example, in Section 16.5 in Royden and Fitzpatrick (2018) and Section VI.5 in Reed and Simon (2012).

Definition A.13 (Finite rank operator)**.** A bounded linear operator $C : \mathcal{H} \to \mathcal{H}$ is said to be of *finite rank* if its range is finite-dimensional. That is, every element in range(C) can be written as $Cf = \sum_{i=1}^{m} \alpha_i g_i$, for some $f \in \mathcal{H}$, some fixed family $\{g_i\}_{i=1}^{m} \subset \mathcal{H}$, some $\alpha_1, \cdots, \alpha_m \in \mathbb{R}$, and $m < \infty$.

The relationship between the compact operator and the finite rank operator is given in the following proposition.

Proposition A.6. *Every finite rank operator is compact. In addition, every compact operator on $\mathcal{H}$ is the norm limit of a sequence of finite rank operators.*

In addition, we have the following characterization when the identity operator is compact.

Proposition A.7. *The identity operator $I : \mathcal{H} \to \mathcal{H}$ is compact if and only if $\mathcal{H}$ is finite-dimensional.*

We also have the following result characterizing the relationship between the invertibility and the compactness of a bounded linear operator.

Proposition A.8. *Let $\mathcal{H}$ be infinite-dimensional and $C : \mathcal{H} \to \mathcal{H}$ be a bounded linear operator. If C is compact, it is* not *invertible.*

In the development in Chapter 3, we have repeatedly used the following Hilbert-Schmidt theorem, which is an extension of the eigen-decomposition of a real symmetric matrix.

Theorem A.1 (Hilbert-Schmidt theorem)**.** *Let $C : \mathcal{H} \to \mathcal{H}$ be a self-adjoint compact operator. There exists an orthonormal basis $\{\psi_\nu\}_{\nu=1}^R$ for $\overline{\text{range}(C)}$, together with a monotonically non-increasing sequence of nonzero real numbers $\{\xi_\nu\}_{\nu=1}^R$, such that $C\psi_\nu = \xi_\nu\psi_\nu$ for all $\nu = 1, \cdots, R$. In addition, the following identity holds*

$$Cf = \sum_{\nu=1}^{R} \xi_\nu \langle f, \psi_\nu \rangle_{\mathcal{H}} \psi_\nu, \qquad \text{for all } f \in \mathcal{H}.$$

If C is of finite rank, then $R < \infty$ and R is the rank of C; if C is not of finite rank, $R = \infty$ and $\lim_{\nu\to\infty} \xi_\nu = 0$.

In Theorem A.1, the functions $\psi_1, \cdots, \psi_R$ are called the *eigenfunctions* of C, and ξ_ν is called the *eigenvalue* of C associated with the eigenfunction ψ_ν, for all $\nu = 1, \cdots, R$.

Using Theorem A.1, it is easy to see that, if C is positive semidefinite, all its eigenvalues are positive.

We next introduce two special classes of bounded linear operators, the trace class and the Hilbert-Schmidt class, that are of great importance in proving various properties of penalized and early stopping SM density estimators in Chapter 2 and 3.

We first introduce the trace of a bounded linear operator, based on which we define the trace class.

Definition A.14 (Trace)**.** Let $\{\psi_\nu\}_{\nu=1}^{\infty}$ be an orthonormal basis of $\mathcal{H}$. For any positive definite bounded linear operator $C : \mathcal{H} \to \mathcal{H}$, the *trace* of C is defined to be

$$\text{trace}(C) := \sum_{\nu=1}^{\infty} \langle \psi_\nu, C\psi_\nu \rangle_{\mathcal{H}},$$

where the sum is independent of the choice of the orthonormal basis.

Definition A.15 (Trace class)**.** A bounded linear operator $C : \mathcal{H} \to \mathcal{H}$ is said to be of *trace class* if $\text{trace}(|C|) < \infty$.

Then, we have the following properties of operators that are of trace class.

Proposition A.9. *Let $C : \mathcal{H} \to \mathcal{H}$ be of trace class. Then, C is compact, and its operator norm, $\|C\|$, and its trace,* $\text{trace}(C)$*, are related by* $\|C\| \leq \text{trace}(C)$.

We now turn to the Hilbert-Schmidt operator.

Definition A.16 (Hilbert-Schmidt operator)**.** Let $\{\psi_\nu\}_{\nu=1}^{\infty}$ be an orthonormal basis of $\mathcal{H}$. A bounded linear operator $C : \mathcal{H} \to \mathcal{H}$ is said to be *Hilbert-Schmidt (HS)* if

$$\sum_{\nu=1}^{\infty} \|C\psi_\nu\|_{\mathcal{H}}^2 < \infty.$$

Still let $\{\psi_\nu\}_{\nu=1}^{\infty}$ be an orthonormal basis of $\mathcal{H}$. Given two HS operators $C_1, C_2 : \mathcal{H} \to \mathcal{H}$, define the *HS inner product* between them to be

$$\langle C_1, C_2 \rangle_{\mathrm{HS}} = \sum_{\nu=1}^{\infty} \langle C_1\psi_\nu, C_2\psi_\nu \rangle_{\mathcal{H}},$$

and the *HS norm* to be

$$\|C_1\|_{\mathrm{HS}} := \sqrt{\langle C_1, C_1 \rangle_{\mathrm{HS}}} = \sqrt{\mathrm{trace}(C_1^* C_1)} = \left(\sum_{\nu=1}^{\infty} \|C_1\psi_\nu\|_{\mathcal{H}}^2 \right)^{1/2}.$$

We then have the following properties of HS operators.

Proposition A.10 (Properties of HS operators)**.** *The following properties of the HS operators hold:*

(a) Let $C : \mathcal{H} \to \mathcal{H}$ be a HS operator. Then, its operator norm $\|C\|$, its trace $\mathrm{trace}(C)$, and its HS norm $\|C\|_{\mathrm{HS}}$ are related by $\|C\| \leq \|C\|_{\mathrm{HS}} \leq \mathrm{trace}(C)$.

(b) The class of all HS operators with the inner product $\langle \cdot, \cdot \rangle_{\mathrm{HS}}$ forms a Hilbert space.

(c) Every HS operator is compact.