Appendix 1: Math Background

Contents

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.

Definition 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

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

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 every $f \in \mathcal{H}$.

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

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

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 (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 2. *(a) Suppose $J_1, J_2 : \mathcal{H} \to \mathbb{R}$ are Frechét differentiable at $f \in \mathcal{H}$ and $\alpha_1, \alpha_2 \in \mathbb{R}$. Then, $\alpha_1 J_1 + \alpha_2 J_2$ is also Frechét differentiable at $f \in \mathcal{H}$, and*

$$\mathrm{D}(\alpha_1 J_1 + \alpha_2 J_2)(f) = \alpha_1 \mathrm{D} J_1(f) + \alpha_2 \mathrm{D} J_2(f).$$

(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

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

Definition 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) implies there exists a unique element $\nabla J(f) \in \mathcal{H}$ such that, for any $g \in \mathcal{H}$,

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

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 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 1 to higher orders.

Definition 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

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

where $\|\cdot\|_{\mathcal{H}^*}$ denotes the norm of the dual space of $\mathcal{H}$.

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

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

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

Remark 3. By Proposition 5.1.17 in Denkowski, Migórski, and Papageorgiou (2013), there exists an isometric isomorphism between $\mathcal{B}(\mathcal{H}, \mathcal{B}(\mathcal{H}, \mathcal{H}))$ and $\mathcal{B}(\mathcal{H} \times \mathcal{H}, \mathcal{H})$. Let Φ denote this isometric isomorphism. Then, $\Phi(\nabla^2 J)$ is a map from $\mathcal{H} \times \mathcal{H}$ to $\mathcal{H}$ such that $\Phi(\nabla^2 J)(f,\, g) = \nabla^2 J(f)(g)$, for all $f, g \in \mathcal{H}$.

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 4) and the measurable function (Definition 5) in the Banach space setting and then define the Bochner μ-integral (Definition 6).

Definition 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 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, $\{s_m\}_{m \in \mathbb{N}}$, such that

$$\lim_{m \to \infty} \|f(x) - s_m(x)\|_{\mathcal{E}} = 0 \tag{6}$$

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

Definition 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, $\{s_m\}_{m \in \mathbb{N}}$, such that

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

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 integral* of f.

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

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

(a) The Bochner 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,

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

3 Partial Derivative of a Kernel Function

In this section, we discuss the partial derivatives of a kernel function of a RKHS and its 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 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 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

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

for all $u := (u_1, \cdots, u_d)^{\top} \in \mathcal{X}$, and *(b) every $f \in \mathcal{H}$ is m-times continuously differentiable, and, for all $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \le m$ and all $x \in \mathcal{X}$, the partial derivative reproducing property holds, i.e.,*

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

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 $x, y \in \mathcal{X}$.

References

Bauschke, Heinz H. and Patrick L. Combettes (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. 1st. Springer Publishing Company. ISBN: 9781441994660.

Denkowski, Zdzislaw, Stanislaw Migórski, and Nikolaos S Papageorgiou (Dec. 2013). *An Introduction to Nonlinear Analysis: Theory*. en. Springer Science & Business Media.

Steinwart, Ingo and Andreas Christmann (Sept. 2008). *Support Vector Machines*. en. Springer Science & Business Media.

Zhou, Ding-Xuan (Oct. 2008). “Derivative reproducing properties for kernel methods in learning theory”. In: *J. Comput. Appl. Math.* 220.1, pp. 456–463.