Diffeological Statistical Models, the Fisher Metric and Probabilistic Mappings

Hông Vân Lê

Institute of Mathematics, Czech Academy of Sciences, Zitna 25, 11567 Praha 1, Czech Republic; hvle@math.cas.cz

Abstract: We introduce the notion of a C^k-diffeological statistical model, which allows us to apply the theory of diffeological spaces to (possibly singular) statistical models. In particular, we introduce a class of almost 2-integrable C^k-diffeological statistical models that encompasses all known statistical models for which the Fisher metric is defined. This class contains a statistical model which does not appear in the Ay–Jost–Lê–Schwachhöfer theory of parametrized measure models. Then, we show that, for any positive integer k, the class of almost 2-integrable C^k-diffeological statistical models is preserved under probabilistic mappings. Furthermore, the monotonicity theorem for the Fisher metric also holds for this class. As a consequence, the Fisher metric on an almost 2-integrable C^k-diffeological statistical model $P \subset \mathcal{P}(\mathcal{X})$ is preserved under any probabilistic mapping $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ that is sufficient w.r.t. P. Finally, we extend the Cramér–Rao inequality to the class of 2-integrable C^k-diffeological statistical models.

Keywords: statistical model; diffeology; the Fisher metric; probabilistic mapping; Cramér-Rao inequality

1. Introduction

In mathematical statistics, the notion of *a statistical model* and the notion of *a parameterized statistical model* are of central importance [1]. For a measurable space $\mathcal{X}$, let us denote by $\mathcal{P}(\mathcal{X})$ the space of all probability measures on $\mathcal{X}$. According to currently accepted theories, see e.g., [1] and the references therein, a statistical model is a subset $P_\mathcal{X} \subset \mathcal{P}(\mathcal{X})$ and a parameterized statistical model is a parameter set Θ, together with a mapping $\mathbf{p} : \Theta \to \mathcal{P}(\mathcal{X})$. The image $\mathbf{p}(\Theta) \subset \mathcal{P}(\mathcal{X})$ is a statistical model endowed with the parameterization $\mathbf{p} : \Theta \to \mathbf{p}(\Theta)$. If the parameter set Θ is a smooth manifold, then we can study a statistical model $\mathbf{p}(\Theta)$, endowed with a parameterization $\mathbf{p} : \Theta \to \mathbf{p}(\Theta) \subset \mathcal{P}(\mathcal{X})$, by applying differential geometric techniques to Θ and to smooth the mappings $\mathbf{p} : \Theta \to \mathcal{P}(\mathcal{X})$.

This idea lies in the heart of the field of information geometry, which is in the domain of mathematical statistics, where we study (parameterized) statistical models using techniques of differential geometry [2–5]. In the book "Information Geometry" by Ay, Jost, Lê, and Schwachhöfer, a parameterized statistical model is a triple $(M, \mathcal{X}, \mathbf{p})$ where M is a Banach manifold, $\mathcal{X}$ is a measurable space, and $i \circ \mathbf{p} : M \xrightarrow{\mathbf{p}} \mathcal{P}(\mathcal{X}) \xrightarrow{i} \mathcal{S}(\mathcal{X})$ is a C^1-map. Here $\mathcal{S}(\mathcal{X})$ is the Banach space of all signed finite measures on $\mathcal{X}$ endowed with the total variation norm $\| \cdot \|_{TV}$ and i is the natural inclusion. We would like to emphasize that the concept of a parameterized statistical model introduced in [5–7] encompasses statistical models endowed with the structure of a finite dimensional manifold [2,3,8], or with the structure of an infinite dimensional Banach manifold [9]. The theory of parameterized measure models, moreover, allows us to study *singular statistical models* $P_\mathcal{X}$ using differential geometric techniques, if $P_\mathcal{X}$ is endowed with a parameterization by a Banach manifold.

In this study, inspired by the theory of diffeological spaces founded by Souriau and developed further by many people, we shall generalize the concept of a parameterized statistical model to the

concept of a C^k-*diffeological statistical model* $P \subset \mathcal{P}(\mathcal{X})$, which, by definition, is a subset in $\mathcal{P}(\mathcal{X})$ endowed with a *compatible C^k-diffeology*. We shall show that the concept of a C^k-diffeological statistical model is more flexible than the concept of a parameterized statistical model. In particular, the image $\mathbf{p}(M)$ of any parameterized statistical model $(M, \mathcal{X}, \mathbf{p})$ has a natural compatible C^1-diffeology. Moreover, for any $k \in \mathbb{N}^+ \cup \infty$, any subset in $\mathcal{P}(\mathcal{X})$ can be provided with a compatible C^k-diffeology (and hence it has a structure of a C^k-diffeological statistical model).

Furthermore, not every subset in $\mathcal{P}(\mathcal{X})$ can be written as $\mathbf{p}(M)$ for some parameterized statistical model $(M, \mathcal{X}, \mathbf{p})$. Hence the class of C^1-diffeological statistical models is larger than the class of statistical models parameterized by Banach manifolds as the Ay–Jost–Lê–Schwachhöfer theory. We also extend conceptually many results in the Ay–Jost–Lê–Schwachhöfer theory concerning the differential geometry of parameterized statistical models and their application to statistics and to the class of C^k-diffeological statistical models, using the theory of probabilistic mappings, developed in a recent work by Jost, Lê, Luu and Tran [10].

Our paper is organized as follows. In the second section we introduce the notions of C^k-diffeological statistical models, almost 2-integrable C^k-diffeological statistical models, and 2-integrable C^k-diffeological statistical models. In the third section we recall the notion of probabilistic mappings and related results in [10] and prove that the class of (almost 2-integrable/resp. 2-integrable) C^k-statistical models is preserved under probabilistic mappings (Theorem 1). Then we extend the monotonicity of the Fisher metric on 2-integrable parameterized statistical models to the class of almost 2-integrable C^k-diffeological statistical models (Theorem 2). In the last section, we prove a diffeological version of the Cramér–Rao inequality (Theorem 3) which extends previously known versions of the Cramér–Rao inequality in [5,11]. We conclude our paper with a discussion on some future directions and open questions.

2. Almost 2-Integrable Diffeological Statistical Models

Given a statistical model, $P \subset \mathcal{P}(\mathcal{X})$, which we also denote by $P_{\mathcal{X}}$, it is known that $P_{\mathcal{X}}$ is endowed with a *natural geometric structure* induced from the Banach space $(\mathcal{S}(\mathcal{X}), ||, ||_{TV})$.

Definition 1. *(cf. [5], Definition 3.2, p. 141) (1) Let $(V, \| \cdot \|)$ be a Banach space, $X \overset{i}{\hookrightarrow} V$ be an arbitrary subset, where i denotes the inclusion, and $x_0 \in X$. Then $v \in V$ is called a tangent vector of X at x_0, if there is a C^1-map $c : \mathbb{R} \to X$, i.e., the composition $i \circ c : \mathbb{R} \to V$ is a C^1-map, such that $c(0) = x_0$ and $\dot{c}(0) = v$.*

(2) The tangent (double) cone $C_x X$ at a point $x \in X$ is defined as the subset of the tangent space $T_x V = V$ that consists of tangent vectors of X at x. The tangent space $T_x X$ is the linear hull of the tangent cone $C_x X$.

(3) The tangent cone fibration CX (resp. the tangent fibration TX) is the union $\cup_{x \in X} C_x X$ (resp. $\cup_{x \in X} T_x X$), which is a subset of $V \times V$ and, therefore, it is endowed with the induced topology from $V \times V$.

Remark 1. *(1) The notion of a tangent cone in Definition 1 occurs in a similar fashion in the theory of singular spaces, see e.g., [12], §3, [13], §3, [14], p. 166.*

(2) Definition 1 differs from [5], Definition 3.1, in that, in Definition 1, the domain of a C^1-curve c is $\mathbb{R}$ and in [5] the domain of a C^1-curve c is $(-\varepsilon, \varepsilon)$. Since $(-\varepsilon, \varepsilon)$ is diffeomorphic to $\mathbb{R}$, both the two choices of the domain of c are equivalent.

Example 1. *Let us consider a mixture family $P_{\mathcal{X}}$ of probability measures $p_\eta \mu_0$ on $\mathcal{X}$ that are dominated by $\mu_0 \in \mathcal{P}(\mathcal{X})$, where the density functions, p_η, are of the following form*

$$p_\eta(x) := g^1(x)\eta_1 + g^2(x)\eta_1 + g^3(x)(1 - \eta_1 - \eta_2) \text{ for } x \in \mathcal{X}. \tag{1}$$

Here g^i, for $i = 1, 2, 3$, are nonnegative functions on $\mathcal{X}$, such that $\mathbb{E}_{\mu_0}(g^i) = 1$ and $\eta = (\eta_1, \eta_2) \in D_b \subset \mathbb{R}^2$ is a parameter, which will be specified as follows. Let us divide the square $D = [0,1] \times [0,1] \subset \mathbb{R}^2$ into smaller squares and color them in black and white as with a chessboard. Let D_b be the closure of the subset

of D colored in black. If η is an interior point of D_b, then $C_{p_\eta} P_{\mathcal{X}} = \mathbb{R}^2$. If η is a boundary point of D_b, then $C_{p_\eta} P_{\mathcal{X}} = \mathbb{R}$. If η is a corner point of D_b, then $C_{p_\eta} P_{\mathcal{X}}$ consists of two intersecting lines.

- Let $P_{\mathcal{X}}$ be a statistical model. Then it is known that any $v \in C_\xi P_{\mathcal{X}}$ is dominated by ξ. Hence *the logarithmic representation of v*

$$\log v := dv/d\xi \tag{2}$$

  is an element of $L^1(\mathcal{X}, \xi)$. The set $\{\log v | \ v \in C_\xi P_{\mathcal{X}}\}$ is a subset in $L^1(\mathcal{X}, \xi)$. We denote it by $\log(C_\xi P_{\mathcal{X}})$ and will call it *the logarithmic representation of $C_\xi P_{\mathcal{X}}$*.

- Next we want to put a Riemannian metric on a statistical model $P_{\mathcal{X}}$ i.e., to put a positive quadratic form $\mathfrak{g}$ on each *tangent space* $T_\xi P_{\mathcal{X}} \subset L^1(\mathcal{X}, \xi)$. The space $L^1(\mathcal{X}, \xi)$ does not have a natural metric but its subspace $L^2(\mathcal{X}, \xi)$ is a Hilbert space.

Definition 2. *A statistical model $P_{\mathcal{X}}$ will be called almost 2-integrable, if*

$$\log(C_\xi P_{\mathcal{X}}) \subset L^2(\mathcal{X}, \xi) \tag{3}$$

for all $\xi \in P_{\mathcal{X}}$. In this case we define the Fisher metric $\mathfrak{g}$ on $P_{\mathcal{X}}$ as follows. For each $v, w \in C_\xi P_{\mathcal{X}}$

$$\mathfrak{g}_\xi(v, w) := \langle \log v, \log w \rangle_{L^2(\mathcal{X}, \xi)} = \int_{\mathcal{X}} \log v \cdot \log w \, d\xi. \tag{4}$$

Since $T_\xi P_{\mathcal{X}}$ is the linear hull of $C_\xi P_{\mathcal{X}}$, Formula (4) extends uniquely to a positive quadratic form on $T_\xi P_{\mathcal{X}}$, which is called *the Fisher metric*.

Example 2. *Let us reconsider Example 1. Recall that our statistical model $P_{\mathcal{X}}$ is parameterized by a map*

$$\mathbf{p} : D_b \to \mathcal{S}(\mathcal{X}), \eta \mapsto p_\eta \cdot \mu_0,$$

which is the restriction of the affine map $L : \mathbb{R}^2 \to \mathcal{S}(\mathcal{X})$, defined by the same formula. Hence, any tangent vector $\tilde{v} \in T_\eta P_{\mathcal{X}}$ can be written as $\tilde{v} = d\mathbf{p}(v)$ where $v \in T_\eta D_b$. For $v = (v_1, v_2) \in T_\eta D_b$, we have $d\mathbf{p}(v) = [(g^1 - g^3)v_1 + (g^2 - g^3)v_2]\mu_0$. If $g^i(x) > 0$ for all $x \in \mathcal{X}$ and $i = 1, 2, 3$, then $p_\eta(x) > 0$ for all $x \in \mathcal{X}$ and all $\eta \in D_b$. Therefore

$$\log d\mathbf{p}(v)_{|\mathbf{p}(\eta)} = \frac{d\mathbf{p}(v)}{d(p_\eta \mu_0)} = \frac{(g^1 - g^3)v_1 + (g^2 - g^3)v_2}{p_\eta} \in L^1(\mathcal{X}, \mathbf{p}(\eta)).$$

Hence $P_{\mathcal{X}}$ is almost 2-integrable, if

$$\frac{g^1 - g^3}{\sqrt{p_\eta}}, \frac{g^2 - g^3}{\sqrt{p_\eta}} \in L^2(\mathcal{X}, \mu_0) \ \forall \eta \in D_b.$$

In this case we have

$$\mathfrak{g}_{|\mathbf{p}(\eta)}(d\mathbf{p}(v), d\mathbf{p}(w)) = \langle \log d\mathbf{p}(v), \log d\mathbf{p}(w) \rangle_{L^2(\mathcal{X}, \mathbf{p}(\eta))}. \tag{5}$$

Next we shall introduce the notion of a C^k-diffeological statistical model.

Definition 3. *For $k \in \mathbb{N}^+ \cup \infty$ and a nonempty set X, a C^k-diffeology of X is a set $\mathcal{D}$ of mappings $\mathbf{p} : U \to X$, where U is an open domain in $\mathbb{R}^n$, and n runs over nonnegative integers, such that the three following axioms are satisfied.*

D1. Covering. The set $\mathcal{D}$ contains the constant mappings $\mathbf{x} : r \mapsto x$, defined on $\mathbb{R}^n$, for all $x \in X$ and for all $n \in \mathbb{N}$.

D2. Locality. Let $\mathbf{p} : U \to X$ *be a mapping. If for every point* $r \in U$ *there exists an open neighborhood* V *of* r, *such that* $\mathbf{p}_{|V}$ *belongs to* $\mathcal{D}$ *then the map* $\mathbf{p}$ *belongs to* $\mathcal{D}$.

D3. Smooth compatibility. For every element $\mathbf{p} : U \to X$ *of* $\mathcal{D}$, *for every real domain* V, *for every* $\psi \in C^k(V, U)$, $\mathbf{p} \circ \psi$ *belongs to* $\mathcal{D}$.

A C^k-*diffeological space is a nonempty set equipped with a* C^k-*diffeology* $\mathcal{D}$. *Elements* $\mathbf{p} : U \to X$ *of* $\mathcal{D}$ *will be called* C^k-*maps from* U *to* X.

A statistical model $P_{\mathcal{X}}$ *endowed with a* C^k-*diffeology* $\mathcal{D}_{\mathcal{X}}$ *will be called a* C^k-*diffeological statistical model, if for any map* $\mathbf{p} : U \to P_{\mathcal{X}}$ *in* $\mathcal{D}_{\mathcal{X}}$ *the composition* $i \circ \mathbf{p} : U \to \mathcal{S}(\mathcal{X})$ *is a* C^k-*map.*

Remark 2. *(1) In [14], Iglesias-Zemmour considered only* C^∞-*diffeologies. The notion of a* C^k-*diffeology, as given in Definition 3 is a straightforward adaptation of the concept of a smooth diffeology, as given in [14], §1.5.*

(2) As $(\mathcal{S}(\mathcal{X}), \| \cdot \|_{TV})$ *is a Banach space, by [15], Lemma 3.11, p. 30, a compatible* C^∞-*diffeology on a statistical model* $P_{\mathcal{X}}$ *is defined by smooth maps* $c : \mathbb{R} \to P_{\mathcal{X}}$.

(3) Given a C^k-*diffeological statistical model* $(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ *and* $\xi \in P_{\mathcal{X}}$, *the tangent cone* $C_{\xi}(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ *is the subset of* $C_{\xi} P_{\mathcal{X}}$ *that consists of the tangent vectors* $\dot{c}(0)$ *of* C^k-*curves* $c : \mathbb{R} \to \mathcal{X}$ *in* $\mathcal{D}_{\mathcal{X}}$, *such that* $c(0) = \xi$. *Similarly, the tangent space* $T_{\xi}(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ *is the linear hull of* $C_{\xi}(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$.

(4) Let $(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ *be a* C^k-*diffeological statistical model and* V *a locally convex vector space. A map* $\varphi : P_{\mathcal{X}} \to V$ *is called Gateaux-differentiable on* $(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ *if for any* C^k-*curve* $c : \mathbb{R} \to P_{\mathcal{X}}$ *in* $\mathcal{D}_{\mathcal{X}}$ *the composition* $\varphi \circ c : \mathbb{R} \to V$ *is differentiable. We recommend [15] for differential calculus on locally convex vector spaces.*

Example 3. *(1) Let* $(M, \mathcal{X}, \mathbf{p})$ *be a parametrized statistical model. Then* $(\mathbf{p}(M), \mathcal{D}_{\mathcal{X}})$ *is a* C^1-*diffeological statistical model where* $\mathcal{D}_{\mathcal{X}}$ *consists of all* C^1-*maps* $\mathbf{q} : \mathbb{R}^n \supset U \to \mathbf{p}(M)$, *such that there exists a* C^1-*map* $\psi^M : U \to M$ *and* $\mathbf{q} = \mathbf{p} \circ \psi^M$.

(2) Let $P_{\mathcal{X}}$ *be a statistical model. Then* $P_{\mathcal{X}}$ *can be endowed with a structure of a* C^k-*diffeological statistical model for any* $k \in \mathbb{N}^+ \cup \infty$, *where its diffeology* $\mathcal{D}_{\mathcal{X}}^{(k)}$ *consists of all mappings* $\mathbf{p} : U \to P_{\mathcal{X}}$, *such that the composition* $i \circ \mathbf{p} : U \to \mathcal{S}(\mathcal{X})$ *is of the class* C^k, *where* U *is any open domain in* $\mathbb{R}^n$ *for* $n \in \mathbb{N}$.

(3) Let $\mathcal{X}$ *be the closed interval* $[0, 1]$. *Let* $P_{\mathcal{X}} := f \cdot \mu_0$, *where* $f \in C^\infty(\mathcal{X})$, *such that* $\int_{\mathcal{X}} f d\mu_0 = 1$ *and* $f(x) > 0$ *for all* $x \in \mathcal{X}$. *We claim that, there does not exist a parameterized statistical model* $(M, \mathcal{X}, \mathbf{p})$, *such that* $P_{\mathcal{X}} = \mathbf{p}(M)$. *Assume the opposite, i.e., there is a* C^1-*map* $\mathbf{p} : M \to \mathcal{S}(\mathcal{X})$, *such that* $\mathbf{p}(M) = P_{\mathcal{X}}$. *Then for any* $m \in M$ *we have* $d\mathbf{p}(T_m(M)) = T_{\mathbf{p}(m)} P_{\mathcal{X}} = \{ f \in C^\infty(\mathcal{X}) | \int_{\mathcal{X}} f d\mu_0 = 0 \}$. *However, this is not the case, as it is known that the space* $C^\infty([0, 1])$ *cannot be the image of a linear bounded map from a Banach space* M *to* $L_1([0, 1])$, *see e.g., [16], p. 1434.*

Definition 4. *A* C^k-*diffeological statistical model* $(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ *will be called almost 2-integrable, if* $\log(C_{\xi}(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})) \subset L^2(\mathcal{X}, \xi)$ *for all* $\xi \in P_{\mathcal{X}}$.

An almost 2-integrable C^k-*diffeological statistical model* $(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ *will be called 2-integrable, if for any* C^k-*map* $\mathbf{p} : U \to P_{\mathcal{X}}$ *in* $\mathcal{D}_{\mathcal{X}}$, *the function* $v \mapsto |d\mathbf{p}(v)|_{\mathfrak{g}}$ *is continuous on* TU.

Example 4. *(1) By [5], Theorem 3.2, p. 155, a parameterized statistical model* $(M, \mathcal{X}, \mathbf{p})$ *is 2-integrable, if and only if* $(\mathbf{p}(M), \mathbf{p}_*(\mathcal{D}_M))$ *is a 2-integrable* C^1-*diffeological statistical model.*

(2) The C^1-*diffeological statistical model* $(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}}^{(1)})$ *in Example 3(3) is 2-integrable, though there is no parameterized statistical model* $(M, \mathcal{X}, \mathbf{p})$ *such that* $\mathbf{p}(M) = P_{\mathcal{X}}$.

(3) Let $\mathcal{X}$ *be a measurable space and* λ *be a* σ-*finite measure. In [17], p. 274, Friedrich considered a family* $P(\lambda) := \{ \mu \in \mathcal{P}(\mathcal{X}) | \mu \ll \lambda \}$ *that is endowed with the following diffeology* $\mathcal{D}(\lambda)$. *A curve* $c : \mathbb{R} \to P(\lambda)$ *is a* C^1-*curve, if*

$$\log \dot{c}(t) \in L^2(\mathcal{X}, c(t)).$$

Hence $(P(\lambda), \mathcal{D}(\lambda))$ *is an almost 2-integrable* C^1-*diffeological statistical model.*

Remark 3. *The axiomatics of Espaces différentiels, which became later the diffeological spaces, were introduced by J.-M. Souriau in the beginning of the nineteen-eighties [18]. Diffeology is a variant of the theory of differentiable spaces, introduced and developed a few years before by K.T. Chen [19]. As I have worked with a different theory of smooth structures on singular spaces [12,13], I appreciate the elegance of the theory of diffeology for its consistent and simple treatment of smooth structures on (possibly infinite dimensional) singular spaces. The best source for diffeology is the monograph by P. Iglesias-Zemmour [14].*

3. Probabilistic Mappings

In 1962, Lawvere proposed a categorical approach to probability theory, where morphisms are Markov kernels, and most importantly, he supplied the space $\mathcal{P}(\mathcal{X})$ with a natural σ-algebra Σ_w, making the notion of Markov kernels and hence many constructions in probability theory and mathematical statistics functorial.

Let us recall the definition of Σ_w. Given a measurable space $\mathcal{X}$, let $\mathcal{F}_s(\mathcal{X})$ denote the linear space of simple functions on $\mathcal{X}$. Recall that $\mathcal{S}(\mathcal{X})$ is the space of all signed finite measures on $\mathcal{X}$. There is a natural homomorphism $I : \mathcal{F}_s(\mathcal{X}) \to \mathcal{S}^*(\mathcal{X}) := Hom(S(\mathcal{X}), \mathbb{R})$, $f \mapsto I_f$, defined by integration: $I_f(\mu) := \int_{\mathcal{X}} f d\mu$ for $f \in \mathcal{F}_s(\mathcal{X})$ and $\mu \in \mathcal{S}(\mathcal{X})$. Following Lawvere [20], we define Σ_w to be the smallest σ-algebra on $\mathcal{S}(\mathcal{X})$, such that I_f is measurable for all $f \in \mathcal{F}_s(\mathcal{X})$. Let $\mathcal{M}(\mathcal{X})$ denote the space of all finite nonnegative measures on $\mathcal{X}$. We also denote by Σ_w , the restriction of Σ_w to $\mathcal{M}(\mathcal{X})$, $\mathcal{M}^*(\mathcal{X}) := \mathcal{M}(\mathcal{X}) \setminus \{0\}$, and $\mathcal{P}(\mathcal{X})$.

- For a topological space $\mathcal{X}$ we shall consider the natural Borel σ-algebra $\mathcal{B}(\mathcal{X})$. Then, every continuous function is measurable w.r.t. $\mathcal{B}(\mathcal{X})$. If $\mathcal{X}$ is, moreover, a metric space, then $\mathcal{B}(\mathcal{X})$ is the smallest algebra making any continuous function measurable ([21], Lemma 2.13).
- Let $C_b(\mathcal{X})$ be the space of *bounded continuous functions* on a topological space $\mathcal{X}$. We denote by τ_v, the smallest topology on $\mathcal{S}(\mathcal{X})$, such that for any $f \in C_b(\mathcal{X})$ the map $I_f : (\mathcal{S}(\mathcal{X}), \tau_v) \to \mathbb{R}$ is continuous. We also denote by τ_v, the restriction of τ_v to $\mathcal{M}(\mathcal{X})$ and $\mathcal{P}(\mathcal{X})$, which is also called the *weak topology* that generates the weak convergence of probability measures. It is known that $(\mathcal{P}(\mathcal{X}), \tau_v)$ is separable, and metrizable if, and only if, $\mathcal{X}$ is [21], Theorem 3.1.4, p. 104. If $\mathcal{X}$ is separable and metrizable then the Borel σ-algebra on $\mathcal{P}(\mathcal{X})$ generated by τ_v coincides with Σ_w.

Definition 5. *([10], Definition 2.4) A probabilistic mapping (or an arrow) from a measurable space $\mathcal{X}$ to a measurable space $\mathcal{Y}$ is a measurable mapping from $\mathcal{X}$ to $(\mathcal{P}(\mathcal{Y}), \Sigma_w)$.*

We shall denote by $\overline{T} : \mathcal{X} \to (\mathcal{P}(\mathcal{Y}), \Sigma_w)$ the measurable mapping defining/generating a probabilistic mapping $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$. Similarly, for a measurable mapping $\mathbf{p} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ we shall denote by $\underline{\mathbf{p}} : \mathcal{X} \rightsquigarrow \mathcal{Y}$ the generated probabilistic mapping. Note that a probabilistic mapping is denoted by a curved arrow and a measurable mapping by a straight arrow.

Example 5. *([10], Example 2.6) (1) Assume that $\mathcal{X}$ is separable and metrizable. Then the identity mapping $Id_\mathcal{P} : (\mathcal{P}(\mathcal{X}), \tau_v) \to (\mathcal{P}(\mathcal{X}), \tau_v)$ is continuous, and hence measurable w.r.t. the Borel σ-algebra $\Sigma_w = \mathcal{B}(\tau_v)$. Consequently, $Id_\mathcal{P}$ generates a probabilistic mapping $ev : (\mathcal{P}(\mathcal{X}), \mathcal{B}(\tau_v)) \rightsquigarrow (\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and we write $\overline{ev} = Id_\mathcal{P}$. Similarly, for any measurable space $\mathcal{X}$, we also have an arrow (a probabilistic mapping) $ev : (\mathcal{P}(\mathcal{X}), \Sigma_w) \rightsquigarrow \mathcal{X}$ generated by the measurable mapping $\overline{ev} = Id_\mathcal{P}$.*

(2) Let δ_x denote the Dirac measure concentrated at x. It is known that the map $\delta : \mathcal{X} \to (\mathcal{P}(\mathcal{X}), \Sigma_w)$, $x \mapsto \delta(x) := \delta_x$, is measurable [22]. If $\mathcal{X}$ is a topological space, then the map $\delta : \mathcal{X} \to (\mathcal{P}(\mathcal{X}), \tau_v)$ is continuous, as the composition $I_f \circ \delta : \mathcal{X} \to \mathbb{R}$ is continuous for any $f \in C_b(\mathcal{X})$. Hence, if $\kappa : \mathcal{X} \to \mathcal{Y}$ is a measurable mapping between measurable spaces (resp. a continuous mapping between separable metrizable spaces), then the map $\overline{\kappa} : \mathcal{X} \overset{\delta \circ \kappa}{\to} \mathcal{P}(\mathcal{Y})$ is a measurable mapping (resp. a continuous mapping). We regard κ as a probabilistic mapping defined by $\delta \circ \kappa : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$. In particular, the identity mapping $Id : \mathcal{X} \to \mathcal{X}$ of a measurable space $\mathcal{X}$ is a probabilistic mapping generated by $\delta : \mathcal{X} \to \mathcal{P}(\mathcal{X})$. Graphically

speaking, any straight arrow (a measurable mapping) $\kappa : \mathcal{X} \to \mathcal{Y}$ between measurable spaces can be seen as a curved arrow (a probabilistic mapping).

Given a probabilistic mapping $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$, we define a linear map $S_*(T) : \mathcal{S}(\mathcal{X}) \to \mathcal{S}(\mathcal{Y})$, called *Markov morphism*, as follows [2], Lemma 5.9, p. 72,

$$S_*(T)(\mu)(B) := \int_{\mathcal{X}} \overline{T}(x)(B) d\mu(x) \tag{6}$$

for any $\mu \in \mathcal{S}(\mathcal{X})$ and $B \in \Sigma_{\mathcal{Y}}$.

Proposition 1. *Assume that $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ is a probabilistic mapping.*

(1) Then T induces a linear bounded map $S_*(T) : \mathcal{S}(\mathcal{X}) \to \mathcal{S}(\mathcal{Y})$ w.r.t. the total variation norm $|| \cdot ||_{TV}$. The restriction $M_*(T)$ of $S_*(T)$ to $\mathcal{M}(\mathcal{X})$ (resp. $P_*(T)$ of $S_*(T)$ to $\mathcal{P}(\mathcal{X})$) maps $\mathcal{M}(\mathcal{X})$ to $\mathcal{M}(\mathcal{Y})$ (resp. $\mathcal{P}(\mathcal{X})$ to $\mathcal{P}(\mathcal{Y})$).

(2) Probabilistic mappings are morphisms in the category of measurable spaces; i.e., for any probabilistic mappings $T_1 : \mathcal{X} \rightsquigarrow \mathcal{Y}$ and $T_2 : \mathcal{Y} \rightsquigarrow \mathcal{Z}$, we have

$$M_*(T_2 \circ T_1) = M_*(T_2) \circ M_*(T_1), \; P_*(T_2 \circ T_1) = P_*(T_2) \circ P_*(T_1). \tag{7}$$

(3) M_* and P_* are faithful functors.
(4) If $\nu \ll \mu \in \mathcal{M}^*(\mathcal{X})$ then $M_*(T)(\nu) \ll M_*(T)(\mu)$.

Remark 4. *The first assertion of Proposition 1 is due to Chentsov [2], Lemma 5.9, p. 72. The second assertion has been proven in [10], Theorem 2.14 (1), extending Giry's result in [22]. The third assertion has been proven in [10]. The last assertion of Proposition 1 is due to Morse–Sacksteder [23], Proposition 5.1.*

We also denote by T_* the map $S_*(T)$, if no confusion can arise.

Given a probabilistic mapping $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ and a C^k-diffeological statistical model $(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$, we define a C^k-diffeological space $(T_*(P_{\mathcal{X}}), T_*(\mathcal{D}_{\mathcal{X}}))$ as the image of $\mathcal{D}$ by T [14], §1.43, p. 24. In other words, a mapping $\mathbf{p} : U \to T_*(P_{\mathcal{X}})$ belongs to $T_*(\mathcal{D}_{\mathcal{X}})$ if and only if it satisfies the following condition. For every $r \in U$ there exists an open neighborhood $V \subset U$ of r, such that either $\mathbf{p}_{|V}$ is a constant mapping, or there exists a mapping $\mathbf{q} : U \to P_{\mathcal{X}}$ in $\mathcal{D}_{\mathcal{X}}$, such that $\mathbf{p}_{|V} = T_* \circ \mathbf{q}$.

Theorem 1. *Let $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ be a probabilistic mapping and $(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ is a C^k-diffeological statistical model.*

(1) Then $(T_*(P_{\mathcal{X}}), T_*(\mathcal{D}_{\mathcal{X}}))$ is a C^k-diffeological statistical model.

(2) If $(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ is an almost 2-integrable C^k-diffeological statistical model, then $(T_*(P_{\mathcal{X}}), T_*(\mathcal{D}_{\mathcal{X}}))$ is also an almost 2-integrable C^k-diffeological statistical model.

(3) If $(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ is a 2-integrable C^k-diffeological statistical model, then $(T_*(P_{\mathcal{X}}), T_*(\mathcal{D}_{\mathcal{X}}))$ is also a 2-integrable C^k-diffeological statistical model.

Proof. (1) The first assertion is straightforward, since $T_* : \mathcal{S}(\mathcal{X}) \to \mathcal{S}(\mathcal{Y})$ is a linear bounded map by Proposition 1(1).

(2) Assume that $(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ is an almost 2-integrable C^k-statistical model and $v \in C_\xi(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$. Then there exits a C^k-map $c : \mathbb{R} \to P_{\mathcal{X}}$ in $\mathcal{D}_{\mathcal{X}}$, such that $\frac{d}{dt}_{|t=0} c(\xi) = v$. Since $T_* : \mathcal{S}(\mathcal{X}) \to \mathcal{S}(\mathcal{Y})$ is a bounded linear map,

$$\frac{d}{dt}_{|t=0} T_* \circ c = T_*(v).$$

By the monotonicity theorem [5], Corollary 5.1, p. 260, we have

$$\|\frac{dT_* v}{dT_* \xi}\|_{L^2(\mathcal{Y}, T_* \xi)} \leq \|v\|_{L^2(\mathcal{X}, \xi)}. \tag{8}$$

This proves that $(T_*(P_\mathcal{X}), T_*(\mathcal{D}_\mathcal{X}))$ is almost 2-integrable.

(3) Assume that $(P_\mathcal{X}, \mathcal{D}_\mathcal{X})$ is a C^k-diffeological statistical model. Let $c : \mathbb{R} \to T_*(P_\mathcal{X})$ be an element in $T_*(\mathcal{D}_\mathcal{X})$. Then $c = T_* \circ c'$, where $c : \mathbb{R} \to P_\mathcal{X}$ is an element of $\mathcal{D}_\mathcal{X}$, i.e., $i \circ c : \mathbb{R} \to \mathcal{S}(\mathcal{X})$ is of class C^k and $(\mathbb{R}, \mathcal{X}, c)$ is a parameterized 2-integrable statistical model. By [5], Theorem 5.4, p. 264, $(\mathbb{R}, \mathcal{Y}, T_* \circ c)$ is a 2-integrable parameterized statistical model. Combined with the first assertion of Theorem 1 this proves the last assertion of Theorem 1. □

Denote by $L(\mathcal{X})$, the space of bounded measurable functions on a measurable space $\mathcal{X}$. Given a probabilistic mapping $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$, we define a linear map $T^* : L(\mathcal{Y}) \to L(\mathcal{X})$, as follows [10], (2.2),

$$T^*(f)(x) := I_f(\overline{T}(x)) = \int_\mathcal{Y} f d\overline{T}(x), \tag{9}$$

which coincides with the classical formula (5.1) in [2], p. 66, for the transformation of a bounded measurable f under a Markov morphism (i.e., a probabilistic mapping) T. In particular, if $\kappa : \mathcal{X} \to \mathcal{Y}$ is a measurable mapping, then we have $\kappa^*(f)(x) = f(\kappa(x))$, since $\overline{\kappa} = \delta \circ \kappa$.

Definition 6. *([10], Definition 2.22, cf. [23]) Let $P_\mathcal{X} \subset \mathcal{P}(\mathcal{X})$ and $P_\mathcal{Y} \subset \mathcal{P}(\mathcal{Y})$. A probabilistic mapping $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ will be called sufficient for $P_\mathcal{X}$ if there exists a probabilistic mapping $\underline{\mathbf{p}} : \mathcal{Y} \rightsquigarrow \mathcal{X}$, such that for all $\mu \in P_\mathcal{X}$ and $h \in L(\mathcal{X})$ we have*

$$T_*(h\mu) = \underline{\mathbf{p}}^*(h) T_*(\mu), \text{ i.e., } \underline{\mathbf{p}}^*(h) = \frac{dT_*(h\mu)}{dT_*(\mu)} \in L^1(\mathcal{Y}, T_*(\mu)). \tag{10}$$

In this case we shall call the measurable mapping $\mathbf{p} : \mathcal{Y} \to \mathcal{P}(\mathcal{X})$ defining the probabilistic mapping $\underline{\mathbf{p}} : \mathcal{Y} \rightsquigarrow \mathcal{X}$ a conditional mapping for T.

Example 6. *Assume that $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$ is a measurable mapping (i.e., a statistic) which is a probabilistic mapping sufficient for $P_\mathcal{X} \subset \mathcal{P}(\mathcal{X})$. Let $\mathbf{p} : \mathcal{Y} \to \mathcal{P}(\mathcal{X})$, $y \mapsto \mathbf{p}_y$, be a conditional mapping for κ. By (9), $\underline{\mathbf{p}}^*(1_A)(y) = \mathbf{p}_y(A)$, and we rewrite (10) as follows*

$$\mathbf{p}_y(A) = \frac{d\kappa_*(1_A\mu)}{d\kappa_*\mu} \in L^1(\mathcal{Y}, \kappa_*(\mu)). \tag{11}$$

The RHS of (11) is the conditional measure of μ applied to A w.r.t. the measurable mapping κ. The equality (11) implies that this conditional measure is regular and independent of μ. Thus the notion of sufficiency of a measurable mapping κ for $P_\mathcal{X}$ coincides with the classical notion of sufficiency of κ for $P_\mathcal{X}$, see e.g., [2], p. 28, [24], Definition 2.8, p. 85. We also note that the equality in (11) is understood as equivalence class in $L^1(\mathcal{Y}, \kappa_*(\mu))$ and hence every statistic κ' that coincides with a sufficient statistic κ except on a zero μ-measure set, for all $\mu \in P_\mathcal{X}$, is also a sufficient statistic for $P_\mathcal{X}$.

Example 7. *(cf. [2], Lemma 2.8, p. 28) Assume that $\mu \in \mathcal{P}(\mathcal{X})$ has a regular conditional distribution w.r.t. to a statistic $\kappa : \mathcal{X} \to \mathcal{Y}$; i.e., there exists a measurable mapping $\mathbf{p} : \mathcal{Y} \to \mathcal{P}(\mathcal{X})$, $y \mapsto \mathbf{p}_y$, such that*

$$\mathbb{E}_\mu^{\sigma(\kappa)}(1_A|y) = \mathbf{p}_y(A) \tag{12}$$

for any $A \in \Sigma_\mathcal{X}$ and $y \in \mathcal{Y}$. Let Θ be a set and $P := \{\nu_\theta \in \mathcal{P}(\mathcal{X}) | \theta \in \Theta\}$ be a parameterized family of probability measures dominated by μ. If there exists a function $h : \mathcal{Y} \times \Theta \to \mathbb{R}$ such that for all $\theta \in \Theta$, and we have

$$\nu_\theta = h(\kappa(x))\mu, \tag{13}$$

then κ is sufficient for P, since, for any $\theta \in \Theta$,

$$\mathbf{p}^*(1_A) = \frac{d\kappa_*(1_A\nu_\theta)}{d\kappa_*\nu_\theta}$$

does not depend on θ. Condition (13) is the Fisher–Neymann sufficiency condition for a family of dominated measures.

Example 8. *Let $\kappa : \mathcal{X} \to \mathcal{Y}$ be a measurable 1-1 mapping. Then for any statistical model $P_{\mathcal{X}} \subset \mathcal{P}(\mathcal{X})$, the statistic κ is sufficient w.r.t. $\mathcal{P}_{\mathcal{X}}$, since, for any $A \in \Sigma_{\mathcal{X}}$ and any $\mu \in \mathcal{P}_{\mathcal{X}}$, we have*

$$\frac{d\kappa_*(1_A\mu)}{d\kappa_*\mu} = (\kappa^{-1})^*(1_A) \in L^1(\mathcal{Y}, \kappa_*(\mu)).$$

Next, we shall show that probabilistic mappings do not increase the Fisher metrics on almost 2-integrable C^k-diffeological statistical models. Thus the Fisher metric serves as a "information quantity" of almost 2-integrable C^k-diffeological statistical models.

Theorem 2. *Let $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ be a probabilistic mapping and $(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ an almost 2-integrable C^k-diffeological statistical model. Then for any $\mu \in P_{\mathcal{X}}$ and any $v \in T_\mu(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$, we have*

$$\mathfrak{g}_\mu(v, v) \geq \mathfrak{g}_{T_*\mu}(T_*v, T_*v)$$

with the equality, if T is sufficient w.r.t. $P_{\mathcal{X}}$.

Proof. The monotonicity assertion of Theorem 2 follows from (8). The second assertion of Theorem 2 follows from the first assertion, taking into account Theorem 2.8.2 in [10], which states the existence of a probabilistic mapping $\mathbf{p} : \mathcal{Y} \rightsquigarrow \mathcal{X}$, such that $\mathbf{p}_*(T_*(P_{\mathcal{X}})) = P_{\mathcal{X}}$, and therefore $\mathbf{p}_*(T_*(\mathcal{D}_{\mathcal{X}})) = \mathcal{D}_{\mathcal{X}}$. □

Let us apply Theorem 2 to Example 4 (3), originally from [17]. In [17], Satz 1, p.274, Friedrich considered the group $\mathcal{G}(\mathcal{X}, \Sigma_{\mathcal{X}}, \lambda)$ of all measurable 1-1 mappings $\Phi : \mathcal{X} \to \mathcal{X}$, such that $\Phi_*(\lambda) \ll \lambda$. Clearly $\Phi_*(P(\lambda)) \subset P(\lambda)$. Example 8 says that Φ is a sufficient statistic w.r.t. $P(\lambda)$. Hence Theorem 2 implies the following

Corollary 1. *([17], Satz 1) The group $\mathcal{G}(\mathcal{X}, \Sigma_{\mathcal{X}}, \lambda)$ acts isometrically on $P(\lambda)$.*

Remark 5. *Theorem 2 extends the Monotonicity Theorem [5], Theorem 5.5, p. 265, for 2-integrable parameterized statistical models. (As we remarked in Section 5, Theorem 2 can be easily extended to the case of almost l-integrable C^k-diffeological measure models.)*

4. The Cramér–Rao Inequality for 2-Integrable Diffeological Statistical Models

In this section we shall prove a version of the Cramér–Rao inequality for estimators with values in a 2-integrable C^k-diffeological statistical model.

Definition 7. *Let $P_{\mathcal{X}} \subset \mathcal{P}(\mathcal{X})$ be a statistical model. An estimator is a map $\hat{\sigma} : \mathcal{X} \to P_{\mathcal{X}}$.*

Assume that V is a locally convex topological vector space. Then we denote, by $Map(P_{\mathcal{X}}, V)$, the space of all mappings $\varphi : P_{\mathcal{X}} \to V$ and by V', the topological dual of V. It is usually easier to estimate only a "coordinate" $\varphi(\xi)$ of a probability measure $\xi \in P_{\mathcal{X}}$, which determines ξ uniquely, if φ is embedded.

Definition 8. *Let $P_{\mathcal{X}}$ be a statistical model and $\varphi \in Map(P_{\mathcal{X}}, V)$. A φ-estimator $\hat{\sigma}_\varphi$ is a composition $\varphi \circ \hat{\sigma} : \mathcal{X} \xrightarrow{\hat{\sigma}} P_{\mathcal{X}} \xrightarrow{\varphi} V$.*

Example 9. *Assume that $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric and positive definite kernel function and let V be the associated RKHS. For any $x \in \mathcal{X}$, we denote by k_x, the function on $\mathcal{X}$ defined by $k_x(y) := k(x, y)$,*

for any $y \in \mathcal{X}$. Then k_x is an element of V. Let $P_{\mathcal{X}} = \mathcal{P}(\mathcal{X})$. Then we define the kernel mean embedding $\varphi : \mathcal{P}(\mathcal{X}) \to V$ as follows [25]

$$\varphi(\xi) := \int_{\mathcal{X}} k_x d\xi(x),$$

where the integral should be understood as a Bochner integral.

Remark 6. *(1) In classical statistics (see e.g., [26], §13, p. 51, [27], p. 4, [8], §4, p. 82, [5], Definition 5.1, p. 277) one considers only the parameter estimations for parameterized statistical models. In this case, an estimator is a map from $\mathcal{X}$ to the parameter set Θ of a statistical model $\mathbf{p}(\Theta) \subset \mathcal{P}(\mathcal{X})$. Usually one assumes that the parameterization $\mathbf{p} : \Theta \to \mathbf{p}(\Theta)$ is 1-1, hence, a parameter estimation is equivalent to a nonparametric estimation in the sense of Definition 7. Note that the ultimate aim of a statistical experiment is to estimate the probability measure generating the observable of the experiment. In general, we can only assume that the unknown generating probability measure belongs to a statistical model $P_{\mathcal{X}} \subset \mathcal{P}(\mathcal{X})$. In this case, we need to use non-parametric estimation; see e.g., [28], p. 1. Note that, by Example 3, $P_{\mathcal{X}}$ has a natural structure of a C^1-diffeological statistical model.*

(2) The notion of a φ-estimation occurs in classical statistics in similar fashion; see e.g., [26], p. 52, where the author called similar estimators substitution estimators, and in [29], Definition 1.2, p. 4, where the authors consider estimands, which are versions of φ-estimators for a parameter estimation problem, see [5], p. 279.

For $\varphi \in Map(P_{\mathcal{X}}, V)$ and $l \in V'$ we denote by φ^l the composition $l \circ \varphi$. Then we set

$$L^2_\varphi(\mathcal{X}, P_{\mathcal{X}}) := \{\hat{\sigma} : \mathcal{X} \to P_{\mathcal{X}} | \, \varphi^l \circ \hat{\sigma} \in L^2_\xi(\mathcal{X}) \text{ for all } \xi \in P_{\mathcal{X}} \text{ and } l \in V'\}.$$

For $\hat{\sigma} \in L^2_\varphi(\mathcal{X}, P_{\mathcal{X}})$ we define the φ-mean value of $\hat{\sigma}$, denoted by $\varphi_{\hat{\sigma}} : P_{\mathcal{X}} \to V''$, as follows (cf. [5], (5.54), p. 279)

$$\varphi_{\hat{\sigma}}(\xi)(l) := \mathbb{E}_\xi(\varphi^l \circ \hat{\sigma}) \text{ for } \xi \in P_{\mathcal{X}} \text{ and } l \in V'.$$

Let us identify V with a subspace in V'' via the canonical pairing.

The difference $b^\varphi_{\hat{\sigma}} := \varphi_{\hat{\sigma}} - \varphi \in Map(\mathcal{P}_{\mathcal{X}}, V'')$ will be called *the bias* of the φ-estimator $\hat{\sigma}_\varphi$.

For all $\xi \in P_{\mathcal{X}}$ we define a quadratic function $MSE^\varphi_\xi[\hat{\sigma}]$ on V', which is called the *mean square error quadratic function* at ξ, by setting for $l, h \in V'$ (cf. [5], (5.56), p. 279)

$$MSE^\varphi_\xi[\hat{\sigma}](l, h) := \mathbb{E}_\xi[(\varphi^l \circ \hat{\sigma}(x) - \varphi^l(\xi)) \cdot (\varphi^h \circ \hat{\sigma}(x) - \varphi^h(\xi))]. \tag{14}$$

Similarly we define the *variance quadratic function* of the φ-estimator $\varphi \circ \hat{\sigma}$ at $\xi \in \mathcal{P}_{\mathcal{X}}$ is the quadratic form $V^\varphi_\xi[\hat{\sigma}]$ on V', such that, for all $l, h \in V'$ we have (cf. [5], (5.57), p. 279)

$$V^\varphi_\xi[\hat{\sigma}](l, h) = \mathbb{E}_\xi[\varphi^l \circ \hat{\sigma}(x) - \mathbb{E}_\xi(\varphi^l \circ \hat{\sigma}(x)) \cdot \varphi^h \circ \hat{\sigma}(x) - \mathbb{E}_\xi(\varphi^h \circ \hat{\sigma}(x))].$$

Then it is known that [5], (5.58), p. 279,

$$MSE^\varphi_\xi[\hat{\sigma}](l, h) = V^\varphi_\xi[\hat{\sigma}](l, h) + \langle b^\varphi_{\hat{\sigma}}(\xi), l \rangle \cdot \langle b^\varphi_{\hat{\sigma}}(\xi), h \rangle. \tag{15}$$

Remark 7. *Assume that V is a real Hilbert space with a scalar product $\langle \cdot, \cdot \rangle$ and the associated norm $\| \cdot \|$. Then the scalar product defines a canonical isomorphism $V = V'$, $v(w) := \langle v, w \rangle$, for all $v, w \in V$. For $\hat{\sigma} \in L^2_\varphi(\mathcal{X}, P_{\mathcal{X}})$, the mean square error $MSE^\varphi_\xi(\hat{\sigma})$ of the φ-estimator $\varphi \circ \hat{\sigma}$ is defined by*

$$MSE^\varphi_\xi(\hat{\sigma}) := \mathbb{E}_\xi(\|\varphi \circ \hat{\sigma} - \varphi(\xi)\|^2). \tag{16}$$

The RHS of (16) is well-defined, since $\hat{\sigma} \in L^2_\varphi(\mathcal{X}, P_{\mathcal{X}})$, and therefore

$$\langle \varphi \circ \hat{\sigma}(x), \varphi \circ \hat{\sigma}(x) \rangle \in L^1(\mathcal{X}, \xi) \text{ and } \langle \varphi \circ \hat{\sigma}(x), \varphi(\xi) \rangle \in L^2(\mathcal{X}, \xi).$$

Similarly, we define the variance of a φ-estimator $\varphi \circ \hat{\sigma}$ at ξ as follows

$$V_\xi^\varphi(\hat{\sigma}) := \mathbb{E}_\xi(\|\varphi \circ \hat{\sigma} - \mathbb{E}_\xi(\varphi \circ \hat{\sigma})\|^2).$$

If V has a countable basis of orthonormal vectors $v_1, \cdots, v_\infty$, then we have

$$MSE_\xi^\varphi(\hat{\sigma}) = \sum_{i=1}^\infty MSE_\xi^\varphi[\hat{\sigma}](v_i, v_i), \tag{17}$$

$$V_\xi^\varphi(\hat{\sigma}) = \sum_{i=1}^\infty V_\xi^\varphi[\hat{\sigma}](v_i, v_i). \tag{18}$$

Now, we assume that $(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ is an almost 2-integrable C^k-diffeological statistical model. For any $\xi \in P_{\mathcal{X}}$, let $T_\xi^{\mathfrak{g}}(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ be the completion of $T_\xi(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ w.r.t. the Fisher metric $\mathfrak{g}$. Since $T_\xi^{\mathfrak{g}}(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ is a Hilbert space, the map

$$L_{\mathfrak{g}} : T_\xi^{\mathfrak{g}}(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}}) \to (T_\xi^{\mathfrak{g}}(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}}))', \ L_{\mathfrak{g}}(v)(w) := \langle v, w \rangle_{\mathfrak{g}},$$

is an isomorphism. Then we define the inverse $\mathfrak{g}^{-1}$ of the Fisher metric $\mathfrak{g}$ on $(T_\xi^{\mathfrak{g}}(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}}))'$ as follows

$$\langle L_g v, L_g w \rangle_{\mathfrak{g}^{-1}} := \langle v, w \rangle_{\mathfrak{g}}. \tag{19}$$

Definition 9. *(cf. [5], Definition 5.18, p. 281) Assume that $\hat{\sigma} \in L_\varphi^2(\mathcal{X}, P_{\mathcal{X}})$. We shall call $\hat{\sigma}$ a φ-regular estimator, if for all $l \in V'$ the function $\xi \mapsto \|\varphi^l \circ \hat{\sigma}\|_{L^2(\mathcal{X}, \xi)}$ is locally bounded, i.e., for all $\xi_0 \in P_{\mathcal{X}}$*

$$\lim_{\xi \to \xi_0} \sup \|\varphi^l \circ \hat{\sigma}\|_{L^2(\mathcal{X}, \xi)} < \infty.$$

Proposition 2. *Assume that $(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ is a 2-integrable C^k-diffeological statistical model, V is a topological vector space, $\varphi \in Map(P_{\mathcal{X}}, V)$ and $\hat{\sigma} : \mathcal{X} \to P_{\mathcal{X}}$ is a φ-regular estimator. Then the V''-valued function $\varphi_{\hat{\sigma}}$ is Gateaux-differentiable on $(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$. Furthermore, for any $l' \in V'$, the differential $d\varphi_{\hat{\sigma}}^l(\xi)$ extends to an element in $(T_\xi^{\mathfrak{g}}(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}}))'$ for all $\xi \in P_{\mathcal{X}}$.*

Proof. Assume that a map $c : \mathbb{R} \to P_{\mathcal{X}}$ belongs to $\mathcal{D}_{\mathcal{X}}$. Then $(\mathbb{R}, \mathcal{X}, c)$ is a 2-integrable parametrized statistical model. By Lemma 5.2 in [5], p. 282, the composition $\varphi_{\hat{\sigma}} \circ c$ is differentiable. This proves the first assertion of Proposition 2.

Next, we shall show that $d\varphi_{\hat{\sigma}}(\xi)$ extends to an element in $(T_\xi^{\mathfrak{g}}(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}}))'$ for all $\xi \in P_{\mathcal{X}}$. Let $X \in C_\xi(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}})$ and $c : \mathbb{R} \to P_{\mathcal{X}}$ be a C^k-curve, such that $c(0) = \xi$ and $\dot{c}(0) = X$. By Lemma 5.3 [5], p. 284, we have

$$\partial_X(\varphi_{\hat{\sigma}}^l) = \int_{\mathcal{X}} (\varphi^{\mathfrak{l}} \circ \hat{\sigma}(x) - \mathbb{E}_\xi(\varphi^l \circ \hat{\sigma}) \cdot \log X \, d\xi(x), \tag{20}$$

where $\varphi^{\mathfrak{l}} \circ \hat{\sigma}(x) - \mathbb{E}_\xi(\varphi^l \circ \hat{\sigma}) \in L^2(\mathcal{X}, \xi)$. Denote by $\Pi_\xi : L^2(\mathcal{X}, \xi) \cdot \xi \to T_\xi^{\mathfrak{g}} P_{\mathcal{X}}$, the orthogonal projection. Set

$$\mathrm{grad}_g(\varphi_{\hat{\sigma}}^l) := \Pi_\xi[(\varphi^{\mathfrak{l}} \circ \hat{\sigma}(x) - \mathbb{E}_\xi(\varphi^l \circ \hat{\sigma})) \cdot \xi] \in T_\xi^{\mathfrak{g}} P_{\mathcal{X}}. \tag{21}$$

Then we rewrite (20), as follows

$$\partial_X(\varphi^l) = \langle \mathrm{grad}_g(\varphi_{\hat{\sigma}}^l), X \rangle_{\mathfrak{g}}.$$

Hence $d\varphi_{\hat{\sigma}}^l$ is the restriction of $L_{\mathfrak{g}}(\mathrm{grad}_{\mathfrak{g}}(\varphi_{\hat{\sigma}}^l)) \in (T_\xi^{\mathfrak{g}}(P_{\mathcal{X}}, \mathcal{D}_{\mathcal{X}}))'$. This completes the proof of Proposition 2. $\square$

For any $\xi \in \mathcal{P}_\mathcal{X}$, we denote $(\mathfrak{g}_{\hat\sigma}^\varphi)^{-1}(\xi)$ to be the following quadratic form on V':

$$(\mathfrak{g}_{\hat\sigma}^\varphi)^{-1}(\xi)(l,k) := \langle d\varphi_{\hat\sigma}^l, d\varphi_{\hat\sigma}^k \rangle_{\mathfrak{g}^{-1}}(\xi) := \langle \mathrm{grad}_{\mathfrak{g}}(\varphi_{\hat\sigma}^l), \mathrm{grad}_{\mathfrak{g}}(\varphi_{\hat\sigma}^k) \rangle. \tag{22}$$

Theorem 3 (Diffeological Cramér–Rao inequality)**.** *Let* $(P_\mathcal{X}, \mathcal{D}_\mathcal{X})$ *be a 2-integrable* C^k*-diffeological statistical model,* φ*, a* V*-valued function on* $P_\mathcal{X}$ *and* $\hat\sigma \in L_\varphi^2(\mathcal{X}, P_\mathcal{X})$*, a* φ*-regular estimator. Then the difference* $\mathcal{V}_\xi^\varphi[\hat\sigma] - (\hat{\mathfrak{g}}_{\hat\sigma}^\varphi)^{-1}(\xi)$ *is a positive semi-definite quadratic form on* V' *for any* $\xi \in P_\mathcal{X}$*.*

Proof. To prove Theorem 3 it suffices to show that for any $l \in V'$ we have

$$\mathbb{E}_\xi(\varphi^l \circ \hat\sigma - \mathbb{E}_\xi(\varphi^l \circ \hat\sigma))^2 \geq \|\mathrm{grad}_{\mathfrak{g}}(\varphi_{\hat\sigma}^l)\|_{\mathfrak{g}}^2. \tag{23}$$

Clearly (23) follows from (21). This completes the proof of Theorem 3. $\quad\square$

Theorem 3 is an extension of the general Cramér–Rao inequality [11], Theorem 2, see also [5], Theorem 5.7, p. 286.

5. Discussion

The extension of the notion of a *k*-integrable parametrized measure model (as introduced in [6,7], see also [5]) to the notion of an almost *k*-integrable diffeological measure model can be done.

(1) There are two main differences between parameterized statistical models and C^k-diffeological statistical models. First, the parameter space of a parameterized statistical model is a *single smooth Banach manifold*, and *parameter spaces* for a C^k-diffeological statistical model can be *different but compatible*. Secondly, parameter spaces for a C^k-diffeological statistical model are *finite dimensional*. If $k = \infty$, this assumption is well-motivated [14], see also Remark 2 (2).

(2) It would be interesting to apply the theory of C^k-statistical models to stochastic processes. It is known that Banach manifolds are not suitable for many questions of global analysis, see e.g., [15], p. 1, and therefore, the theory of parameterized measure models might have limited applications to stochastic processes. On the other hand, there are many open questions in the theory of C^∞-diffeological spaces, e.g., we do not know under which conditions we can define the Levi–Civita connection on a Riemannian C^∞-diffeological space. Furthermore, the theory of C^k-diffeological spaces has not been considered before, with $k \neq \infty$.

(3) The variational calculus founded by Leibniz and Newton is a cornerstone of differential geometry and modern analysis. In our opinion, it is best expressed in the language of diffeological spaces that declare which mappings into a diffeological space are smooth. This language is a counterpart to the language of ringed spaces in algebraic geometry that declares which functions are algebraic.

Conflicts of Interest: The author declares no conflict of interest.

References

1. McCullagh, P. What is a statistical model. *Ann. Stat.* **2002**, *30*, 1225–1310. [CrossRef]
2. Chentsov, N. *Statistical Decision Rules and Optimal Inference*; Nauka: Moscow, Russia, 1972; English translation in: Translation of Math. Monograph vol. 53, Amer. Math. Soc.: Providence, RI, USA, 1982.
3. Amari, S. *Differential-Geometric Methods in Statistics*; Lecture Notes in Statistics 28; Springer: Heidelberg, Germany, 1985.
4. Amari, S. *Information Geometry and Its Applications*; Applied Mathematical Sciences; Springer: Berlin, Germany, 2016; Volume 194.
5. Ay, N.; Jost, J.; Lê, H.V.; Schwachhöfer, L. *Information Geometry*; Springer Nature: Cham, Switzerland, 2017.
6. Ay, N.; Jost, J.; Lê, H.V.; Schwachhöfer, L. Information geometry and sufficient statistics. *Probab. Theory Relat. Fields* **2015**, *162*, 327–364. [CrossRef]
7. Ay, N.; Jost J.; Lê, H.V.; Schwachhöfer, L. Parameterized measure models. *Bernoulli* **2018**, *24*, 1692–1725. [CrossRef]
8. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; Translations of Mathematical Monographs 191; Amer. Math. Soc.: Providence, RI, USA, 2000.
9. Pistone, G.; Sempi, C. An infinite-dimensional structure on the space of all the probability measures equivalent to a given one. *Ann. Stat.* **1995**, *23*, 1543–1561. [CrossRef]
10. Jost, J.; Lê, H.V.; Luu, D.H.; Tran, T.D. Probabilistic mappings and Bayesian nonparametrics. *arXiv* **2019**, arXiv:1905.11448.
11. Lê, H.V.; Jost, J.; Schwachhöfer, L. The Cramér-Rao Inequality on Singular Statistical Models. In Proceedings of the Conference "Geometric Science of Information", GSI 2017, Paris, France, 7–9 November 2017; LNCS; Springer Nature: Cham, Switzerland, 2017; Volume 10589, pp. 552–560.
12. Lê, H.V.; Somberg, P.; Vanžura, J. Smooth structures on pseudomanifolds with isolated conical singularities. *Acta Math. Vietnam.* **2013**, *38*, 33–54. [CrossRef]
13. Lê, H.V.; Somberg, P.; Vanžura, J. Poisson smooth structures on stratified symplectic spaces. In *The Springer Proceedings in Mathematics & Statistics "Mathematics in the 21st Century, 6th World Conference", Lahore, March 2013*; Springer: Basel, Switzerland, 2015; Volume 98, Chapter 7, pp. 181–204.
14. Iglesias-Zemmour, P. *Diffeology*; Amer. Math. Soc.: Providence, RI, USA, 2013.
15. Kriegl, A.; Michor, P.W. *The Convenient Setting of Global Analysis*; Amer. Math. Soc.: Providence, RI, USA, 1997.
16. Grabiner, S. Range of products of operators. *Can. J. Math.* **1974**, *XXVI*, 1430–1441. [CrossRef]
17. Friedrich, T. Die Fisher-Information und symplektische Strukturen. *Math. Nachr.* **1991**, *153*, 273–296. [CrossRef]
18. Souriau, J.-M. Groupes différentiels. In *Lecture Notes in Mathematics, Vol. 836*; Springer: Berlin, Germany, 1980; pp. 91–128.
19. Chen, K.T. Iterated path integrals. *Bull. Am. Math. Soc.* **1977**, *83*, 831–879. [CrossRef]
20. Lawvere, W.F. The Category of Probabilistic Mappings. 1962. Unpublished. Available online: https://ncatlab.org/nlab/files/lawvereprobability1962.pdf (accessed on 19 December 2019).
21. Bogachev, V.I. *Weak Convergence of Measures*; Mathematical Surveys and Monographs; Amer. Math. Soc.: Providence, RI, USA, 2018; Volume 234.
22. Giry, M. A categorical approach to probability theory. In *Categorical Aspects of Topology and Analysis*; Banaschewski, B., Ed.; Lecture Notes in Mathematics; Springer: Berlin/Heidelberg, Germany, 1982; Volume 915, pp. 68–85.
23. Morse, N.; Sacksteder, R. Statistical isomorphism. *Ann. Math. Stat.* **1966**, *37*, 203–214. [CrossRef]
24. Schervish, M.J. *Theory of Statistics*, 2nd ed.; Springer: New York, NY, USA, 1997.
25. Muandet, K.; Fukumizu, K.; Sriperumbudur, B.; Schölkopf, B. Kernel Mean Embedding of Distributions: A Review and Beyonds. *Found. Trends Mach. Learn.* **2017**, *10*, 1–141. [CrossRef]
26. Borovkov, A.A. *Mathematical Statistics*; Gordon and Breach Science Publishers: Amsterdam, The Netherlands, 1998.
27. Ibragimov, I.A.; Has'minskii, R.Z. *Statistical Estimation: Asymptotic Theory*; Springer: New York, NY, USA, 1981.

28.　Tsybakov, A.B. *Introduction to Nonparametric Estimation*; Springer Science+Business Media: New York, NY, USA, 2009.

29.　Lehmann, E.L.; Casella, G. *Theory of Point Estimation*, 2nd ed.; Springer: New York, NY, USA, 1998.