

# Parametrized measure models

NIHAT AY<sup>1,\*</sup>, JÜRGEN JOST<sup>1,\*\*</sup>,  
HÔNG VÂN LÊ<sup>2,†</sup> and LORENZ SCHWACHHÖFER<sup>3,‡</sup>

<sup>1</sup>Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany.

E-mail: \*[nay@mis.mpg.de](mailto:nay@mis.mpg.de); \*\*[jjost@mis.mpg.de](mailto:jjost@mis.mpg.de)

<sup>2</sup>Academy of Sciences of the Czech Republic, Prague. E-mail: †[hvle@math.cas.cz](mailto:hvle@math.cas.cz)

<sup>3</sup>TU Dortmund University, Dortmund, Germany. E-mail: ‡[lschwach@math.tu-dortmund.de](mailto:lschwach@math.tu-dortmund.de)

We develop a new and general notion of parametric measure models and statistical models on an arbitrary sample space  $\Omega$  which does not assume that all measures of the model have the same null sets. This is given by a differentiable map from the parameter manifold  $M$  into the set of finite measures or probability measures on  $\Omega$ , respectively, which is differentiable when regarded as a map into the Banach space of all signed measures on  $\Omega$ . Furthermore, we also give a rigorous definition of roots of measures and give a natural characterization of the Fisher metric and the Amari–Chentsov tensor as the pullback of tensors defined on the space of roots of measures. We show that many features such as the preservation of this tensor under sufficient statistics and the monotonicity formula hold even in this very general set-up.

*Keywords:* Amari–Chentsov tensor; Fisher quadratic form; monotonicity; sufficient statistic

## 1. Introduction

Information geometry is concerned with the use of differential geometric methods in probability theory. An important object of investigation are families of probability measures or, more generally, of finite measures on a given sample space  $\Omega$  which depend differentiably on a finite number of parameters. Associated to such a family there are two symmetric tensors on the parameter space  $M$ . The first is a quadratic form (i.e., a Riemannian metric), called the *Fisher metric*  $g^F$ , and the second is a 3-tensor, called the *Amari–Chentsov tensor*  $\mathbf{T}^{\text{AC}}$ . The Fisher metric was first suggested by Rao [27], followed by Jeffreys [16], Efron [13] and then systematically developed by Chentsov and Morozova [10,12] and [20]; the Amari–Chentsov tensor and its significance was discovered by Amari [1,2] and Chentsov [11].

These tensors are of interest from the differential geometric point of view as they do not depend on the particular choice of parametrization of the family, but they are also natural objects from the point of view of statistics, as they are unchanged under sufficient statistics and are in fact characterized by this property; this was shown in the case of finite sample spaces by Chentsov in [12] and more recently for general sample spaces in [5]. In fact, Chentsov not only showed the invariance of these tensors under sufficient statistics, but also under what he called *congruent embeddings* of probability measures. These are Markov kernels between finite sample spaces which are right inverses of a statistic. We use this property to give a definition of congruent embeddings between arbitrary sample spaces (cf. Definition 3.1). As it turns out, every Markov kernel induces a congruent embedding in this sense, but there are congruent embeddings which are *not* induced by Markov kernels, cf. Theorem 3.1.

The main conceptual difficulty in the investigation of families of probability measures is the lack of a canonical manifold structure on the spaces  $\mathcal{M}(\Omega)$  and  $\mathcal{P}(\Omega)$  of finite measures and probability measures on  $\Omega$ . If  $\Omega$  is finite, then a measure is given by finitely many nonnegative parameters, allowing to identify  $\mathcal{M}(\Omega)$  with the closure of the positive orthant in  $\mathbb{R}^{|\Omega|}$  and  $\mathcal{P}(\Omega)$  with the intersection of this closure with an affine hyperplane in  $\mathbb{R}^{|\Omega|}$ , so that both are (finite dimensional) manifolds with corners. If one does not assume that all elements of  $\Omega$  have positive mass for all measures in the family, that is, allowing the model to contain elements of the boundary of  $\mathcal{M}(\Omega)$  or  $\mathcal{P}(\Omega)$ , then technical difficulties arise for example, when describing the Fisher metric and the Amari–Chentsov tensor. If  $\Omega$  is infinite, then there is a priori not even a differentiable structure on  $\mathcal{M}(\Omega)$  and  $\mathcal{P}(\Omega)$ .

Attempts have been made to provide  $\mathcal{P}(\Omega)$  and  $\mathcal{M}(\Omega)$  with a Banach manifold structure. For instance, Pistone and Sempi [26] equipped these spaces with a topology, the so-called *e-topology*. With this,  $\mathcal{P}(\Omega)$  and  $\mathcal{M}(\Omega)$  become Banach manifolds and have many remarkable features, see, for example, [9,25]. On the other hand, the *e-topology* is very strong in the sense that many families of measures on  $\Omega$  fail to be continuous w.r.t. the *e-topology*, so it cannot be applied as widely as one would wish.

Another approach was recently pursued by Bauer, Bruveris and Michor [7] under the assumption that  $\Omega$  is a manifold. In this case, the space of smooth densities also carries a natural topology, and they were able to show that the invariance under diffeomorphisms already suffices to characterize the Fisher metric of a family of such densities.

In [5], the authors of the present article proposed to define parametrized measure models as a family  $(\mathbf{p}(\xi))_{\xi \in M}$  of finite measures on  $\Omega$ , labelled by elements  $\xi$  of a finite dimensional manifold  $M$ , such that for a measurable set  $A \subset \Omega$

$$\mathbf{p}(\xi)(A) = \int_A d\mathbf{p}(\xi) = \int_A p(\omega; \xi) d\mu(\omega) \tag{1.1}$$

for some reference measure  $\mu$  and a positive function  $p$  on  $\Omega \times M$  which is differentiable in  $\xi \in M$ . This closely follows the notion of Amari [1]. That is, for fixed  $\xi \in M$ , the function  $p(\cdot; \xi)$  on  $\Omega$  is the Radon–Nikodym derivative of the measure  $\mathbf{p}(\xi)$  w.r.t.  $\mu$ , whence by a slight abuse of notation (e.g., [6], Definition 17.2) we abbreviate (1.1) as

$$\mathbf{p}(\xi) = p(\cdot; \xi)\mu. \tag{1.2}$$

While this notion embraces many interesting families of measures, it is still restricted as it requires the existence of a reference measure  $\mu$  dominating all measures  $\mathbf{p}(\xi)$ , and on the other hand, the positivity of the density function implies that all measures  $\mathbf{p}(\xi)$  on  $\Omega$  are equivalent, that is, have the same null sets. While the existence of a measure  $\mu$  dominating all measures  $\mathbf{p}(\xi)$  is satisfied for example, if  $M$  is a finite dimensional manifold, the condition that all measures  $\mathbf{p}(\xi)$  have the same null sets is a more severe restriction of the admissible families.

It is the aim of the present article to provide a yet more general definition of parametrized measure models which embraces all of the aforementioned definitions, but is more general and more natural than these at the same time. Namely, in this article we define *parametrized measure models* and *statistical models*, respectively, as families  $(\mathbf{p}(\xi))_{\xi \in M}$  which are given by a map  $\mathbf{p}$  from  $M$  to  $\mathcal{M}(\Omega)$  and  $\mathcal{P}(\Omega)$ , respectively, which is differentiable when regarded as a map between the

(finite or infinite dimensional) manifold  $M$  and the Banach space  $\mathcal{S}(\Omega)$  of finite signed measures on  $\Omega$ , since evidently  $\mathcal{P}(\Omega)$  and  $\mathcal{M}(\Omega)$  are subsets of  $\mathcal{S}(\Omega)$ . That is, the geometric structure on  $\mathcal{M}(\Omega)$  and  $\mathcal{P}(\Omega)$  is given by the inclusions  $\mathcal{P}(\Omega) \hookrightarrow \mathcal{M}(\Omega) \hookrightarrow \mathcal{S}(\Omega)$ .

For the models defined in [5], the function  $p : \Omega \times M \rightarrow \mathbb{R}$  in (1.2) is differentiable into the  $\xi$ -direction, and such a  $p$  is called a *regular density function*. Even if a parametrized measure model in the sense of the present paper has a dominating measure  $\mu$  and hence is given by (1.2), the density function  $p$  is not necessarily regular, cf. Remark 4.2 and Example 4.2(2) below, and  $p$  is not required to be positive  $\mu$ -a.e., making this notion more general than that in [5]. We shall show that most of the statements shown in [5] for parametrized measure models or statistical models with a positive regular density function also hold in this more general setup.

The Fisher metric  $\mathfrak{g}^F$  and the Amari–Chentsov tensor  $\mathbf{T}^{\text{AC}}$  associated to a parametrized measure model are the two symmetric tensors given by

$$\begin{aligned} \mathfrak{g}^F(V, W) &:= \int_{\Omega} \partial_V \log p(\omega; \xi) \partial_W \log p(\omega; \xi) d\mathbf{p}(\xi), \\ \mathbf{T}^{\text{AC}}(V, W, U) &:= \int_{\Omega} \partial_V \log p(\omega; \xi) \partial_W \log p(\omega; \xi) \partial_U \log p(\omega; \xi) d\mathbf{p}(\xi). \end{aligned}$$

The crucial observation is that even though the function  $\log p(\omega; \xi)$  is not defined everywhere if we drop the assumption that the density function  $p$  is positive, the partial derivatives  $\partial_V \log p(\omega; \xi)$  still may be given sense for an arbitrary parametrized measure model. Thus, the notion of  $k$ -integrability from [5] requiring that  $\partial_V \log p(\omega; \xi) \in L^k(\Omega, p(\xi))$  for all  $V \in T_{\xi}M$  generalizes to parametrized measure models.

We also introduce the Banach space  $\mathcal{S}^r(\Omega)$  of  $r$ th powers of measures on  $\Omega$  for  $r \in (0, 1]$ , which has been discussed in [23], Ex. IV.1.4, for general  $\Omega$  and generalizes the concept of half densities on a manifold  $\Omega$  in [22], Section 6.9.1. The elements of  $\mathcal{S}^r(\Omega)$  can be raised to the  $1/r$ th power to become finite signed measures, and for each measure  $\mu \in \mathcal{M}(\Omega) \subset \mathcal{S}(\Omega)$  there is a well defined power  $\mu^r \in \mathcal{S}^r(\Omega)$ . Thus, for a parametrized measure model  $\mathbf{p} : M \rightarrow \mathcal{M}(\Omega)$  the  $r$ th power defines a map  $\mathbf{p}^r : M \rightarrow \mathcal{S}^r(\Omega)$ , and if the model is  $k$ -integrable for  $k = 1/r \geq 1$ , then  $\mathbf{p}^r$  is differentiable, and for  $k = 2$  or  $k = 3$ ,  $\mathfrak{g}^F$  and  $\mathbf{T}^{\text{AC}}$  are pull-backs of canonical tensors on  $\mathcal{S}^{1/2}(\Omega)$  under  $\mathbf{p}^{1/2}$  and  $\mathcal{S}^{1/3}(\Omega)$  under  $\mathbf{p}^{1/3}$ , respectively.

We also discuss the behavior of the Fisher metric under statistics, i.e., under measurable maps  $\kappa : \Omega \rightarrow \Omega'$  or, more general, under Markov kernels  $K : \Omega \rightarrow \mathcal{P}(\Omega')$ . These transitions can be interpreted as data processing in statistical decision theory, which can be deterministic (given by a measurable map, i.e., a statistic) or randomized (i.e., given by a Markov kernel). The earliest occurrence of this point of view appears to be [11].

Given a parametrized measure model  $\mathbf{p} : M \rightarrow \mathcal{M}(\Omega)$ , it induces a map  $\mathbf{p}'(\xi) := \kappa_*\mathbf{p}(\xi)$  or  $\mathbf{p}'(\xi) := K_*\mathbf{p}(\xi)$ , respectively. We show that this process preserves  $k$ -integrability, i.e., if  $\mathbf{p}$  is  $k$ -integrable, then so is  $\mathbf{p}'$  (cf. Theorem 5.1). Moreover, in Theorem 5.2 we show in this general setup the estimate

$$\|\partial_V \log p(\cdot; \xi)\|_k \geq \|\partial_V \log p'(\cdot; \xi)\|_k, \quad \text{whence } \mathfrak{g}^F(V, V) \geq \mathfrak{g}'^F(V, V), \quad (1.3)$$

where the second estimate is called the *monotonicity formula* and follows from the first for  $k = 2$ . The difference  $\|\partial_V \log p(\cdot; \xi)\|_k^k - \|\partial_V \log p'(\cdot; \xi)\|_k^k \geq 0$  is called the  *$k$ th order information loss*

under  $\kappa$  (or  $K$ ) in direction  $V$ . If the information loss in any direction vanishes, then we call the statistic *sufficient* for the model.

There is a remarkable difference between parametrized measure models with *positive* regular density functions, that is, those considered in [5], and the more general notion establishes in this paper. Namely, in case of a positive regular density function the vanishing of the information loss for a statistic  $\kappa : \Omega \rightarrow \Omega'$  implies that the statistic admits a Fisher-Neyman factorization, cf. Proposition 5.1. Remarkably, this is no longer true in our setting. That is, if we admit parametrized measure models with inequivalent measures, then there are statistics which have vanishing information loss, but do not admit a Fisher-Neyman factorization, cf. Example 5.2.

This paper is structured as follows. In Section 2, we give the formal definition of the spaces of  $r$ th powers of measures. In Section 3, we provide a precise definition of congruent embeddings for arbitrary sample spaces  $\Omega$  and discuss their relations with Markov kernels and the existence of transverse measures. In the following Section 4, we establish the notion of  $k$ -integrability, which is applied in the final Section 5 to the discussion of sufficient statistics and the proof of the monotonicity formula.

## 2. The spaces of measures and their powers

### 2.1. The space of (signed) finite measures

Let  $(\Omega, \Sigma)$  be a measurable space, that is an arbitrary set  $\Omega$  together with a sigma algebra  $\Sigma$  of subsets of  $\Omega$ . Regarding the sigma algebra  $\Sigma$  on  $\Omega$  as fixed, we let

$$\begin{aligned} \mathcal{P}(\Omega) &:= \{\mu : \mu \text{ a probability measure on } \Omega\}, \\ \mathcal{M}(\Omega) &:= \{\mu : \mu \text{ a finite measure on } \Omega\}, \\ \mathcal{S}(\Omega) &:= \{\mu : \mu \text{ a signed finite measure on } \Omega\}, \\ \mathcal{S}_0(\Omega) &:= \left\{ \mu \in \mathcal{S}(\Omega) : \int_{\Omega} d\mu = 0 \right\}. \end{aligned}$$

Clearly,  $\mathcal{P}(\Omega) \subset \mathcal{M}(\Omega) \subset \mathcal{S}(\Omega)$ , and  $\mathcal{S}_0(\Omega), \mathcal{S}(\Omega)$  are real vector spaces. In fact, both  $\mathcal{S}_0(\Omega)$  and  $\mathcal{S}(\Omega)$  are Banach spaces whose norm is given by the total variation of a signed measure, defined as

$$\|\mu\|_{\text{TV}} := \sup \sum_{i=1}^n |\mu(A_i)|,$$

where the supremum is taken over all finite partitions  $\Omega = A_1 \dot{\cup} \dots \dot{\cup} A_n$  with disjoint sets  $A_i \in \Sigma$ . Here, the symbol  $\dot{\cup}$  stands for the disjoint union of sets.

For a measurable function  $\phi : \Omega \rightarrow [-\infty, \infty]$ , we define  $\phi_+ := \max(\phi, 0)$  and  $\phi_- := \max(-\phi, 0)$ , so that  $\phi_{\pm} \geq 0$  are measurable with disjoint support, and

$$\phi = \phi_+ - \phi_-, \quad |\phi| = \phi_+ + \phi_-. \tag{2.1}$$

By the *Jordan decomposition theorem*, each measure  $\mu \in \mathcal{S}(\Omega)$  can be decomposed uniquely as

$$\mu = \mu_+ - \mu_- \quad \text{with } \mu_{\pm} \in \mathcal{M}(\Omega), \mu_+ \perp \mu_- \tag{2.2}$$

That is, there is a decomposition  $\Omega = P \dot{\cup} N$  with  $\mu_+(N) = \mu_-(P) = 0$ . Thus, if we define

$$|\mu| := \mu_+ + \mu_- \in \mathcal{M}(\Omega),$$

then (2.2) implies

$$|\mu(A)| \leq |\mu|(A) \quad \text{for all } \mu \in \mathcal{S}(\Omega) \text{ and } A \in \Sigma, \tag{2.3}$$

so that

$$\|\mu\|_{\text{TV}} = \||\mu|\|_{\text{TV}} = |\mu|(\Omega).$$

In particular,

$$\mathcal{P}(\Omega) = \{\mu \in \mathcal{M}(\Omega) : \|\mu\|_{\text{TV}} = 1\}.$$

Moreover, fixing a measure  $\mu_0 \in \mathcal{M}(\Omega)$ , we let

$$\begin{aligned} \mathcal{P}(\Omega, \mu_0) &:= \{\mu \in \mathcal{P}(\Omega) : \mu \text{ is dominated by } \mu_0\}, \\ \mathcal{M}(\Omega, \mu_0) &:= \{\mu \in \mathcal{M}(\Omega) : \mu \text{ is dominated by } \mu_0\}, \\ \mathcal{P}_+(\Omega, \mu_0) &:= \{\mu \in \mathcal{P}(\Omega, \mu_0) : \mu \text{ is equivalent to } \mu_0\}, \\ \mathcal{M}_+(\Omega, \mu_0) &:= \{\mu \in \mathcal{M}(\Omega, \mu_0) : \mu \text{ is equivalent to } \mu_0\}, \\ \mathcal{S}(\Omega, \mu_0) &:= \{\mu \in \mathcal{S}(\Omega) : \mu \text{ is dominated by } \mu_0\}, \\ \mathcal{S}_0(\Omega, \mu_0) &:= \mathcal{S}(\Omega, \mu_0) \cap \mathcal{S}_0(\Omega), \end{aligned} \tag{2.4}$$

where we say that  $\mu_0$  dominates  $\mu$  if every  $\mu_0$ -null set is also a  $|\mu|$ -null set and where we call two measures equivalent if they dominate each other and hence have the same null sets. The spaces in (2.4) do not change when replacing  $\mu_0$  by an equivalent measure.

We may canonically identify  $\mathcal{S}(\Omega, \mu_0)$  with  $L^1(\Omega, \mu_0)$  by the correspondence

$$\iota_{\text{can}} : L^1(\Omega, \mu_0) \longrightarrow \mathcal{S}(\Omega, \mu_0), \quad \phi \longmapsto \phi \mu_0.$$

By the Radon–Nikodym theorem, this is an isomorphism whose inverse is given by the Radon–Nikodym derivative  $\mu \mapsto \frac{d\mu}{d\mu_0}$ . With this,  $\mathcal{M}(\Omega, \mu_0) = \{\phi \mu_0 : \phi \geq 0\}$  and  $\mathcal{M}_+(\Omega, \mu_0) = \{\phi \mu_0 : \phi > 0\}$  and the corresponding descriptions apply to  $\mathcal{P}(\Omega, \mu_0)$  and  $\mathcal{P}_+(\Omega, \mu_0)$ , respectively. Observe that  $\iota_{\text{can}}$  is an isomorphism of Banach spaces, since evidently

$$\|\phi\|_{L^1(\Omega, \mu_0)} = \int_{\Omega} |\phi| d\mu_0 = \|\phi \mu_0\|_{\text{TV}}.$$

## 2.2. Differential maps between Banach manifolds and tangent double cone fibrations

In this section, we shall recall some basic notions of maps between Banach manifolds. For simplicity, we shall restrict ourselves to maps between open subsets of Banach spaces, even though this notion can be generalized to general Banach manifolds, see, for example, [17].

Let  $V$  and  $W$  be Banach spaces and  $U \subset V$  an open subset. A map  $\phi : U \rightarrow W$  is called *differentiable at  $x \in U$* , if there is a bounded linear operator  $d_x\phi \in \text{Lin}(V, W)$  such that

$$\lim_{h \rightarrow 0} \frac{\|\phi(x+h) - \phi(x) - d_x\phi(h)\|_W}{\|h\|_V} = 0. \tag{2.5}$$

In this case,  $d_x\phi$  is called the *(total) differential of  $\phi$  at  $x$* . Moreover,  $\phi$  is called *continuously differentiable* or shortly a  $C^1$ -map, if it is differentiable at every  $x \in U$ , and the map  $d\phi : U \rightarrow \text{Lin}(V, W)$ ,  $x \mapsto d_x\phi$  is continuous. Furthermore, a differentiable map  $c : (-\varepsilon, \varepsilon) \rightarrow W$  is called a *curve in  $W$* .

**Definition 2.1.** Let  $X \subset V$  be an arbitrary subset and let  $x_0 \in X$ . Then  $v \in V$  is called a *tangent vector of  $X$  at  $x_0$* , if there is a curve  $c : (-\varepsilon, \varepsilon) \rightarrow X \subset V$  such that  $c(0) = x_0$  and  $\dot{c}(0) := d_0c(1) = v$ .

The set of all tangent vectors at  $x_0$  is called the *tangent double cone of  $X$  at  $x_0$*  and is denoted by  $T_{x_0}X$ .

Since reparametrization of the curve  $c$  easily implies that  $T_{x_0}X$  is invariant under multiplication by positive or negative scalars, it is a double cone in  $V$ . However, for general subsets  $X \subset V$ ,  $T_{x_0}X$  may fail to be a vector subspace, and for  $x_0 \neq x_1$ , the tangent cones  $T_{x_0}X$  and  $T_{x_1}X$  need not be homeomorphic. We also let

$$TX := \dot{\bigcup}_{x_0 \in X} T_{x_0}X \subset X \times V \subset V \times V,$$

equipped with the induced topology. Again,  $\dot{\bigcup}$  stands for the disjoint union of sets. Then  $TX$  together with the map  $TX \rightarrow X$  mapping  $T_{x_0}X$  to  $x_0$  is a topological fibration, called the *tangent double cone fibration of  $X$* . Since this is a rather bulky terminology, we shall simply refer to  $TX \rightarrow X$  as the *tangent fibration*, but the reader should be aware that, unlike in some texts, this is not the a synonym for the tangent bundle, as  $X$  needs not be a manifold in general.

If  $U \subset V$  is open and  $\phi : U \rightarrow W$  is a  $C^1$ -map whose image is contained in  $X \subset W$ , then  $d_{x_0}\phi(V) \subset T_{\phi(x_0)}X$ , whence  $\phi$  induces a continuous map

$$d\phi : TU = U \times V \longrightarrow TX, \quad (u, v) \longmapsto d_u\phi(v).$$

**Theorem 2.1.** Let  $\mathcal{S}(\Omega)$  be the Banach space of signed finite measures on  $\Omega$ . Then the tangent cones of  $\mathcal{M}(\Omega)$  and  $\mathcal{P}(\Omega)$  at  $\mu$  are  $T_\mu\mathcal{M}(\Omega) = \mathcal{S}(\Omega, \mu)$  and  $T_\mu\mathcal{P}(\Omega) = \mathcal{S}_0(\Omega, \mu)$ , respectively,

so that the tangent fibrations are given as

$$T\mathcal{M}(\Omega) = \dot{\bigcup}_{\mu \in \mathcal{M}(\Omega)} \mathcal{S}(\Omega, \mu) \subset \mathcal{M}(\Omega) \times \mathcal{S}(\Omega)$$

and

$$T\mathcal{P}(\Omega) = \dot{\bigcup}_{\mu \in \mathcal{P}(\Omega)} \mathcal{S}_0(\Omega, \mu) \subset \mathcal{P}(\Omega) \times \mathcal{S}(\Omega).$$

**Proof.** Let  $v \in T_{\mu_0}\mathcal{M}(\Omega)$  and let  $(\mu_t)_{t \in (-\varepsilon, \varepsilon)}$  be a curve in  $\mathcal{M}(\Omega)$  with  $\dot{\mu}_0 = v$ . Let  $A \subset \Omega$  be such that  $\mu_0(A) = 0$ . Then as  $\mu_t(A) \geq 0$ , the function  $t \mapsto \mu_t(A)$  has a minimum at  $t_0 = 0$ , whence

$$0 = \left. \frac{d}{dt} \right|_{t=0} \mu_t(A) = \dot{\mu}_0(A) = v(A),$$

where the second equation is evident from (2.5). Thus,  $v(A) = 0$  whenever  $\mu_0(A) = 0$ , i.e.,  $\mu_0$  dominates  $v$ , so that  $v \in \mathcal{S}(\Omega, \mu_0)$ . Thus,  $T_{\mu_0}\mathcal{M}(\Omega) \subset \mathcal{S}(\Omega, \mu_0)$ .

Conversely, given  $v = \phi\mu_0 \in \mathcal{S}(\Omega, \mu_0)$ , define  $\mu_t := p(\omega; t)\mu_0$  where

$$p(\omega; t) := \begin{cases} 1 + t\phi(\omega) & \text{if } t\phi(\omega) \geq 0, \\ \exp(t\phi(\omega)) & \text{if } t\phi(\omega) < 0. \end{cases}$$

As  $p(\omega; t) \leq \max(1 + t\phi(\omega), 1)$ , it follows that  $\mu_t \in \mathcal{M}(\Omega)$ , and as  $|\partial_t p(\omega; t)| \leq |\phi(\omega)| \in L^1(\Omega, \mu_0)$  for all  $t$ , it follows that  $t \mapsto \mu_t$  is a  $C^1$ -curve in  $\mathcal{M}(\Omega)$  with  $\dot{\mu}_0 = \phi\mu_0 = v$ , whence  $v \in T_{\mu_0}\mathcal{M}(\Omega)$  as claimed.

To show the statement for  $\mathcal{P}(\Omega)$ , let  $(\mu_t)_{t \in (-\varepsilon, \varepsilon)}$  be a curve in  $\mathcal{P}(\Omega)$  with  $\dot{\mu}_0 = v$ . Then as  $\mu_t$  is a probability measure for all  $t$ , we conclude

$$\left| \int_{\Omega} dv \right| = \left| \int_{\Omega} \frac{1}{t} d(\mu_t - \mu_0 - tv) \right| \leq \frac{\|\mu_t - \mu_0 - tv\|_{\text{TV}}}{|t|} \xrightarrow{t \rightarrow 0} 0,$$

so that  $v \in \mathcal{S}_0(\Omega)$ . Since  $\mathcal{P}(\Omega) \subset \mathcal{M}(\Omega)$ , it follows that  $T_{\mu_0}\mathcal{P}(\Omega) \subset T_{\mu_0}\mathcal{M}(\Omega) \cap \mathcal{S}_0(\Omega) = \mathcal{S}_0(\Omega, \mu_0)$  for all  $\mu_0 \in \mathcal{P}(\Omega)$ .

Conversely, given  $v = \phi\mu_0 \in \mathcal{S}_0(\Omega, \mu_0)$ , define the curve  $\lambda_t := \mu_t \|\mu_t\|_{\text{TV}}^{-1} \in \mathcal{P}(\Omega)$  with  $\mu_t$  from above, which is a  $C^1$ -curve in  $\mathcal{P}(\Omega)$  as  $\|\mu_t\|_{\text{TV}} > 0$ , and it is straightforward that  $\lambda_0 = \mu_0$  and  $\dot{\lambda}_0 = \phi\mu_0 = v$ . □

**Remark 2.1.**

- (1) Observe that the curves  $\mu_t$  and  $\lambda_t$  in the proof of Theorem 2.1 are contained in  $\mathcal{M}_+(\Omega, \mu_0)$  and  $\mathcal{P}_+(\Omega, \mu_0)$ , respectively, whence

$$T_{\mu}\mathcal{M}_+(\Omega, \mu_0) = \mathcal{S}(\Omega, \mu) \quad \text{and} \quad T_{\mu}\mathcal{P}_+(\Omega, \mu) = \mathcal{S}_0(\Omega, \mu).$$

But if  $\mu \in \mathcal{M}_+(\Omega, \mu_0)$ , the  $\mu$  and  $\mu_0$  are equivalent measures so that  $\mathcal{S}(\Omega, \mu) = \mathcal{S}(\Omega, \mu_0) =: V$  and  $\mathcal{S}_0(\Omega, \mu) = \mathcal{S}_0(\Omega, \mu_0) =: V_0$ . Thus, the tangent space is the same at all points.

That is,  $\mathcal{M}_+(\Omega, \mu_0) \subset V$  has the property that  $T_\mu \mathcal{M}_+(\Omega, \mu_0) = V$  for all  $\mu$ , but  $\mathcal{M}_+(\Omega, \mu_0) \subset V$  is not an open subset if  $\Omega$  is infinite, and the corresponding statement holds for  $\mathcal{P}(\Omega, \mu_0) \subset \mu_0 + V_0$ . This is a rather unusual phenomenon.

- (2) The sets  $\mathcal{P}(\Omega)$  and  $\mathcal{M}(\Omega)$  are not Banach submanifolds of  $\mathcal{S}(\Omega)$ , and the tangent fibrations  $T\mathcal{P}(\Omega) \rightarrow \mathcal{P}(\Omega)$  and  $T\mathcal{M}(\Omega) \rightarrow \mathcal{M}(\Omega)$  are not vector bundles, even though the fibers at each point are closed subspaces. This even fails in the case where  $\Omega = \{\omega_1, \dots, \omega_k\}$  is finite. In this case, we may identify  $\mathcal{S}(\Omega)$  with  $\mathbb{R}^k$  by the map  $\sum_{i=1}^k x_i \delta^{\omega_i} \cong (x_1, \dots, x_k)$ , and with this,

$$T\mathcal{M}(\Omega) \cong \left\{ (x_1, \dots, x_k; y_1, \dots, y_k) \in \mathbb{R}^k \times \mathbb{R}^k : \begin{array}{l} x_i \geq 0, \\ x_i = 0 \Rightarrow y_i = 0 \end{array} \right\} \subset \mathbb{R}^{2k},$$

and this is evidently not a submanifold of  $\mathbb{R}^{2k}$ . Indeed, in this case the dimension of  $T_\mu \mathcal{M}(\Omega) = \mathcal{S}(\Omega, \mu)$  equals  $|\{\omega \in \Omega \mid \mu(\omega) > 0\}|$ , which varies with  $\mu$ .

### 2.3. Powers of measures

Let us now give the formal definition of roots of measures. On the set  $\mathcal{M}(\Omega)$  we define the preordering  $\mu_1 \leq \mu_2$  if  $\mu_2$  dominates  $\mu_1$ . Then  $(\mathcal{M}(\Omega), \leq)$  is a directed set, meaning that for any pair  $\mu_1, \mu_2 \in \mathcal{M}(\Omega)$  there is a  $\mu_0 \in \mathcal{M}(\Omega)$  dominating both of them (use e.g.  $\mu_0 := \mu_1 + \mu_2$ ).

For fixed  $r \in (0, 1]$  and measures  $\mu_1 \leq \mu_2$  on  $\Omega$  we define the linear embedding

$$i_{\mu_2}^{\mu_1} : L^{1/r}(\Omega, \mu_1) \longrightarrow L^{1/r}(\Omega, \mu_2), \quad \phi \longmapsto \phi \left( \frac{d\mu_1}{d\mu_2} \right)^r.$$

Observe that

$$\begin{aligned} \|i_{\mu_2}^{\mu_1}(\phi)\|_{L^{1/r}(\Omega, \mu_2)} &= \left| \int_{\Omega} |i_{\mu_2}^{\mu_1}(\phi)|^{1/r} d\mu_2 \right|^r \\ &= \left| \int_{\Omega} |\phi|^{1/r} \frac{d\mu_1}{d\mu_2} d\mu_2 \right|^r \\ &= \left| \int_{\Omega} |\phi|^{1/r} d\mu_1 \right|^r \\ &= \|\phi\|_{L^{1/r}(\Omega, \mu_1)}, \end{aligned} \tag{2.6}$$

so that  $i_{\mu_2}^{\mu_1}$  is an isometry. Evidently  $i_{\mu_2}^{\mu_1} i_{\mu_3}^{\mu_2} = i_{\mu_3}^{\mu_1}$  whenever  $\mu_1 \leq \mu_2 \leq \mu_3$ . Then we define the space of  $r$ th powers of measures on  $\Omega$  to be the directed limit over the directed set  $(\mathcal{M}(\Omega), \leq)$

$$\mathcal{S}^r(\Omega) := \lim_{\longrightarrow} L^{1/r}(\Omega, \mu). \tag{2.7}$$



Let us give a more concrete definition of  $\mathcal{S}^r(\Omega)$ . On the disjoint union of the spaces  $L^{1/r}(\Omega, \mu)$  for  $\mu \in \mathcal{M}(\Omega)$  we define the equivalence relation

$$\begin{aligned} L^{1/r}(\Omega, \mu_1) \ni \phi \sim \psi \in L^{1/r}(\Omega, \mu_2) &\iff \iota_{\mu_0}^{\mu_1}(\phi) = \iota_{\mu_0}^{\mu_2}(\psi) \\ &\iff \phi \left(\frac{d\mu_1}{d\mu_0}\right)^r = \psi \left(\frac{d\mu_2}{d\mu_0}\right)^r \end{aligned}$$

for some  $\mu_0 \geq \mu_1, \mu_2$ . Then  $\mathcal{S}^r(\Omega)$  is the set of all equivalence classes of this relation.

Denote the equivalence class of  $\phi \in L^{1/r}(\Omega, \mu)$  by  $\phi\mu^r$ , so that  $\mu^r \in \mathcal{S}^r(\Omega)$  is the equivalence class represented by  $1 \in L^{1/r}(\Omega, \mu)$ . Then the equivalence relation yields

$$\mu_1^r = \left(\frac{d\mu_1}{d\mu_2}\right)^r \mu_2^r \quad \text{as elements of } \mathcal{S}^r(\Omega) \tag{2.8}$$

whenever  $\mu_1 \leq \mu_2$ , justifying this notation. In fact, from this description in the case  $r = 1$  we see that

$$\mathcal{S}^1(\Omega) = \mathcal{S}(\Omega).$$

Observe that by (2.6)  $\|\phi\|_{L^{1/r}(\Omega, \mu)}$  is constant on equivalence classes, whence there is a norm on  $\mathcal{S}^r(\Omega)$ , denoted by  $\|\cdot\|_{1/r}$ , for which the inclusions

$$L^{1/r}(\Omega, \mu) \longrightarrow \mathcal{S}^r(\Omega), \quad \phi \longmapsto \phi\mu^r$$

are isometries. For  $r = 1$ , we have  $\|\cdot\|_1 = \|\cdot\|_{\text{TV}}$ . Thus,

$$\|\phi\mu^r\|_{1/r} = \|\phi\|_{L^{1/r}(\Omega, \mu)} = \left| \int_{\Omega} |\phi|^{1/r} d\mu \right|^r \quad \text{for } 0 < r \leq 1. \tag{2.9}$$

Note that the equivalence relation also preserves nonnegativity of functions, whence we may define the subsets

$$\begin{aligned} \mathcal{M}^r(\Omega) &:= \{\phi\mu^r : \mu \in \mathcal{M}(\Omega), \phi \geq 0\}, \\ \mathcal{P}^r(\Omega) &:= \{\phi\mu^r : \mu \in \mathcal{P}(\Omega), \phi \geq 0, \|\phi\mu^r\|_{1/r} = 1\}. \end{aligned} \tag{2.10}$$

In analogy to (2.4) we define for a fixed measure  $\mu_0 \in \mathcal{M}(\Omega)$  and  $r \in (0, 1]$  the spaces

$$\begin{aligned} \mathcal{S}^r(\Omega, \mu_0) &:= \{\phi\mu_0^r : \phi \in L^{1/r}(\Omega, \mu_0)\}, \\ \mathcal{M}^r(\Omega, \mu_0) &:= \{\phi\mu_0^r : \phi \in L^{1/r}(\Omega, \mu_0), \phi \geq 0\}, \\ \mathcal{P}^r(\Omega, \mu_0) &:= \{\phi\mu_0^r : \phi \in L^{1/r}(\Omega, \mu_0), \phi \geq 0, \|\phi\mu_0^r\|_{1/r} = 1\}, \\ \mathcal{S}_0^r(\Omega, \mu_0) &:= \left\{ \phi\mu_0^r : \phi \in L^{1/r}(\Omega, \mu_0), \int_{\Omega} \phi d\mu = 0 \right\}. \end{aligned}$$

The elements of  $\mathcal{P}^r(\Omega, \mu_0)$ ,  $\mathcal{M}^r(\Omega, \mu_0)$ ,  $\mathcal{S}^r(\Omega, \mu_0)$  are said to be *dominated by*  $\mu_0^r$ .

If  $\{\mu_n \in \mathcal{S}(\Omega) : n \in \mathbb{N}\}$  is a countable family of (signed) finite measures, then they are dominated by the finite measure  $\mu_0 := \sum_n 2^{-n} \|\nu_n\|_{\mathbb{T}V}^{-1} |\nu_n|$  (cf. e.g., [23], Ex. IV.1.3). Therefore, any Cauchy sequence  $(\mu_{r;n})_{n \in \mathbb{N}} \in \mathcal{S}^r(\Omega)$  is contained in  $\mathcal{S}^r(\Omega, \mu_0)$  for some  $\mu_0$ . As the embedding  $\mathcal{S}^r(\Omega, \mu_0) \hookrightarrow \mathcal{S}^r(\Omega)$  is an isometry,  $(\mu_{r;n})_{n \in \mathbb{N}} \in \mathcal{S}^r(\Omega, \mu_0) \cong L^{1/r}(\Omega, \mu_0)$  is also a Cauchy sequence and hence convergent. Thus,  $(\mathcal{S}^r(\Omega), \|\cdot\|_{1/r})$  is a Banach space.

**Remark 2.2.** The concept of  $r$ th root of measures has been indicated in [23], Ex. IV.1.4. Moreover, if  $\Omega$  is a manifold and  $r = 1/2$ , then  $\mathcal{S}^{1/2}(\Omega)$  is even a Hilbert space which has been considered in [22], Section 6.9.1. In this case, the diffeomorphism group of  $\Omega$  acts by isometries on  $\mathcal{S}^{1/2}(\Omega)$  [15].

The product of powers of measures can now be defined for all  $r, s \in (0, 1)$  with  $r + s \leq 1$  and for measures  $\phi \mu^r \in \mathcal{S}^r(\Omega, \mu)$  and  $\psi \mu^s \in \mathcal{S}^s(\Omega, \mu)$ :

$$(\phi \mu^r) \cdot (\psi \mu^s) := \phi \psi \mu^{r+s}.$$

By definition  $\phi \in L^{1/r}(\Omega, \mu)$  and  $\psi \in L^{1/s}(\Omega, \mu)$ , whence Hölder's inequality implies that  $\|\phi \psi\|_{1/(r+s)} \leq \|\phi\|_{1/r} \|\psi\|_{1/s} < \infty$ , so that  $\phi \psi \in L^{1/(r+s)}(\Omega, \mu)$  and hence,  $\phi \psi \mu^{r+s} \in \mathcal{S}^{r+s}(\Omega, \mu)$ . Since by (2.8) this definition of the product is independent of the choice of representative  $\mu$ , it follows that it induces a bilinear product

$$\cdot : \mathcal{S}^r(\Omega) \times \mathcal{S}^s(\Omega) \longrightarrow \mathcal{S}^{r+s}(\Omega), \quad \text{where } r, s, r + s \in (0, 1], \quad (2.11)$$

satisfying the Hölder inequality

$$\|\nu_r \cdot \nu_s\|_{1/(r+s)} \leq \|\nu_r\|_{1/r} \|\nu_s\|_{1/s},$$

so that the product in (2.11) is a bounded bilinear map.

In analogy to Theorem 2.1, we can also determine the tangent fibrations of the subsets  $\mathcal{P}^r(\Omega) \subset \mathcal{M}^r(\Omega) \subset \mathcal{S}^r(\Omega)$ .

**Proposition 2.1.** *For each  $\mu \in \mathcal{M}(\Omega)$  ( $\mu \in \mathcal{P}(\Omega)$ , respectively), the tangent cones of  $\mathcal{P}^r(\Omega) \subset \mathcal{M}^r(\Omega) \subset \mathcal{S}^r(\Omega)$  at  $\mu^r$  are  $T_{\mu^r} \mathcal{M}^r(\Omega) = \mathcal{S}^r(\Omega, \mu)$  and  $T_{\mu^r} \mathcal{P}^r(\Omega) = \mathcal{S}_0^r(\Omega, \mu)$ , respectively, so that the tangent fibrations are given as*

$$T\mathcal{M}^r(\Omega) = \dot{\bigcup}_{\mu_r \in \mathcal{M}^r(\Omega)} \mathcal{S}^r(\Omega, \mu) \subset \mathcal{M}^r(\Omega) \times \mathcal{S}^r(\Omega)$$

and

$$T\mathcal{P}^r(\Omega) = \dot{\bigcup}_{\mu_r \in \mathcal{P}^r(\Omega)} \mathcal{S}_0^r(\Omega, \mu) \subset \mathcal{P}^r(\Omega) \times \mathcal{S}^r(\Omega).$$

**Proof.** We have to adapt the proof of Theorem 2.1. The proof of the statements  $\mathcal{S}^r(\Omega, \mu) \subset T_{\mu^r} \mathcal{M}^r(\Omega)$  and  $\mathcal{S}_0^r(\Omega, \mu) \subset T_{\mu^r} \mathcal{P}^r(\Omega)$  is identical to that of the corresponding statement in Theorem 2.1; just as in that case, one shows that for  $\phi \in L^{1/r}(\Omega, \mu_0)$  the curves  $\mu_t^r := p(\omega; t) \mu_0^r$

with  $p(\omega; t) := 1 + t\phi(\omega)$  if  $t\phi(\omega) \geq 0$  and  $p(\omega; \xi) = \exp(t\phi(\omega))$  if  $t\phi(\omega) < 0$  is a differentiable curve in  $\mathcal{M}^r(\Omega)$ , and  $\lambda_t^r := \mu_t^r / \|\mu_t^r\|_{1/r}$  is a differentiable curve in  $\mathcal{P}^r(\Omega)$ , and their derivative is  $\phi\mu_0^r$  at  $t = 0$ .

In order to show the other direction, let  $(\mu_t^r)_{t \in (-\varepsilon, \varepsilon)}$  be a curve in  $\mathcal{M}^r(\Omega)$ . Since there is a measure  $\hat{\mu}$  dominating the countable family  $(\mu_t^r)_{t \in \mathbb{Q} \cap (-\varepsilon, \varepsilon)}$  and since  $\mathcal{S}^r(\Omega, \hat{\mu}) \subset \mathcal{S}^r(\Omega)$  is closed, it follows that  $\mu_t^r \in \mathcal{M}(\Omega, \hat{\mu})$  for all  $t$ . Now we can apply the argument from the proof of Theorem 2.1 to the curve  $t \mapsto (\mu_t^r \cdot \hat{\mu}^{1-r})(A)$  for  $A \subset \Omega$ . □

Besides multiplication of roots of measures, we also wish to take their powers. Here, we have two possibilities to deal with signs. For  $0 < k \leq r^{-1}$  and  $v_r = \phi\mu^r \in \mathcal{S}^r(\Omega)$  we define

$$|v_r|^k := |\phi|^k \mu^{rk} \quad \text{and} \quad \tilde{v}_r^k := \text{sign}(\phi)|\phi|^k \mu^{rk}.$$

Since  $\phi \in L^{1/r}(\Omega, \mu)$ , it follows that  $|\phi|^k \in L^{1/rk}(\Omega, \mu)$ , so that  $|v_r|^k, \tilde{v}_r^k \in \mathcal{S}^{rk}(\Omega)$ . By (2.8) these powers are well defined, independent of the choice of the measure  $\mu$ , and, moreover,

$$\| |v_r|^k \|_{1/(rk)} = \| \tilde{v}_r^k \|_{1/(rk)} = \| v_r \|_{1/r}^k. \tag{2.12}$$

**Proposition 2.2.** *Let  $r \in (0, 1]$  and  $0 < k \leq 1/r$ , and consider the maps*

$$\pi^k, \tilde{\pi}^k : \mathcal{S}^r(\Omega) \longrightarrow \mathcal{S}^{rk}(\Omega), \quad \begin{aligned} \pi^k(v) &:= |v|^k, \\ \tilde{\pi}^k(v) &:= \tilde{v}^k. \end{aligned}$$

*Then  $\pi^k, \tilde{\pi}^k$  are continuous maps. Moreover, for  $1 < k \leq 1/r$  they are  $C^1$ -maps between Banach spaces, and their derivatives are given as*

$$d_{v_r} \tilde{\pi}^k(\rho_r) = k|v_r|^{k-1} \cdot \rho_r \quad \text{and} \quad d_{v_r} \pi^k(\rho_r) = k\tilde{v}_r^{k-1} \cdot \rho_r. \tag{2.13}$$

Observe that for  $k = 1$ ,  $\pi^1(v_r) = |v_r|$  fails to be  $C^1$ , whereas  $\tilde{\pi}^1(v_r) = v_r$ , so that  $\tilde{\pi}^1$  is the identity and hence a  $C^1$ -map.

**Proof of Proposition 2.2.** Let us first assume that  $0 < k \leq 1$ . We assert that in this case, there are constants  $C_k, \tilde{C}_k > 0$  such that for all  $x, y \in \mathbb{R}$

$$\begin{aligned} ||x + y|^k - |x|^k| &\leq C_k |y|^k \quad \text{and} \\ |\text{sign}(x + y)|x + y|^k - \text{sign}(x)|x|^k| &\leq \tilde{C}_k |y|^k. \end{aligned} \tag{2.14}$$

Namely, by homogeneity it suffices to show this for  $y = 1$ , and since the functions

$$x \mapsto |x + 1|^k - |x|^k \quad \text{and} \quad x \mapsto \text{sign}(x + 1)|x + 1|^k - \text{sign}(x)|x|^k$$

are continuous and have finite limits for  $x \rightarrow \pm\infty$ , it follows that they are bounded, showing (2.14).

Let  $v_1, v_2 \in \mathcal{S}^r(\Omega)$ , and choose  $\mu_0 \in \mathcal{M}(\Omega)$  such that  $v_1, v_2 \in \mathcal{S}^r(\Omega, \mu_0)$ , i.e.,  $v_i = \phi_i \mu_0^r$  with  $\phi_i \in L^{1/r}(\Omega, \mu_0)$ . Then

$$\begin{aligned} \|\pi^k(v_1 + v_2) - \pi^k(v_1)\|_{1/(rk)} &= \||\phi_1 + \phi_2|^k - |\phi_1|^k\|_{1/(rk)} \\ &\leq C_k \|\phi_2\|_{1/rk}^k \quad \text{by (2.14)} \\ &= C_k \|v_2\|_{1/r}^k \quad \text{by (2.12),} \end{aligned}$$

so that  $\lim_{\|v_2\|_{1/r} \rightarrow 0} \|\pi^k(v_1 + v_2) - \pi^k(v_1)\|_{1/(rk)} = 0$ , showing the continuity of  $\pi^k$  for  $0 < k \leq 1$ . The continuity of  $\tilde{\pi}^k$  follows analogously.

Now let us assume that  $1 < k \leq 1/r$ . In this case, the functions

$$x \mapsto |x|^k \quad \text{and} \quad x \mapsto \text{sign}(x)|x|^k$$

with  $x \in \mathbb{R}$  are  $C^1$ -maps with respective derivatives

$$x \mapsto k \text{sign}(x)|x|^{k-1} \quad \text{and} \quad x \mapsto k|x|^{k-1}.$$

Thus, if we pick  $v_i = \phi_i \mu_0^r$  as above, then by the mean value theorem we have

$$\begin{aligned} \pi^k(v_1 + v_2) - \pi^k(v_1) &= (|\phi_1 + \phi_2|^k - |\phi_1|^k) \mu_0^{rk} \\ &= k \text{sign}(\phi_1 + \eta\phi_2) |\phi_1 + \eta\phi_2|^{k-1} \phi_2 \mu_0^{rk} \\ &= k \text{sign}(\phi_1 + \eta\phi_2) |\phi_1 + \eta\phi_2|^{k-1} \mu_0^{r(k-1)} \cdot v_2 \end{aligned}$$

for some function  $\eta : \Omega \rightarrow (0, 1)$ . If we let  $v_\eta := \eta\phi_2 \mu_0^r$ , then  $\|v_\eta\|_{1/r} \leq \|v_2\|_{1/r}$ , and we get

$$\pi^k(v_1 + v_2) - \pi^k(v_1) = k \tilde{\pi}^{k-1}(v_1 + v_\eta) \cdot v_2.$$

With the definition of  $d_{v_1} \tilde{\pi}^k$  from (2.13) we have

$$\begin{aligned} &\|\pi^k(v_1 + v_2) - \pi^k(v_1) - d_{v_1} \pi^k(v_2)\|_{1/(rk)} \\ &= \|k(\tilde{\pi}^{k-1}(v_1 + v_\eta) - \tilde{\pi}^{k-1}(v_1)) \cdot v_2\|_{1/(rk)} \\ &\leq k \|\tilde{\pi}^{k-1}(v_1 + v_\eta) - \tilde{\pi}^{k-1}(v_1)\|_{1/(r(k-1))} \|v_2\|_{1/r} \end{aligned}$$

and hence,

$$\frac{\|\pi^k(v_1 + v_2) - \pi^k(v_1) - d_{v_1} \pi^k(v_2)\|_{\frac{1}{rk}}}{\|v_2\|_{\frac{1}{r}}} \leq k \|\tilde{\pi}^{k-1}(v_1 + v_\eta) - \tilde{\pi}^{k-1}(v_1)\|_{\frac{1}{r(k-1)}}.$$

Thus, the differentiability of  $\pi^k$  will follow if

$$\|\tilde{\pi}^{k-1}(v_1 + v_\eta) - \tilde{\pi}^{k-1}(v_1)\|_{1/(r(k-1))} \xrightarrow{\|v_2\|_{1/r} \rightarrow 0} 0,$$

and because of  $\|v_\eta\|_{1/r} \leq \|v_2\|_{1/r}$ , this is the case if  $\tilde{\pi}^{k-1}$  is continuous.

Analogously, one shows that  $\tilde{\pi}^k$  is differentiable if  $\pi^{k-1}$  is continuous.

Since we already know continuity of  $\pi^k$  and  $\tilde{\pi}^k$  for  $0 < k \leq 1$ , and since  $C^1$ -maps are continuous, the claim now follows by induction on  $[k]$ . □

Thus, (2.13) implies that the differentials of  $\pi^k$  and  $\tilde{\pi}^k$  (which coincide on  $\mathcal{P}^r(\Omega)$  and  $\mathcal{M}^r(\Omega)$ ) yield continuous maps

$$d\pi^k = d\tilde{\pi}^k : \begin{matrix} T\mathcal{P}^r(\Omega) \longrightarrow T\mathcal{P}^{rk}(\Omega), \\ T\mathcal{M}^r(\Omega) \longrightarrow T\mathcal{M}^{rk}(\Omega), \end{matrix} \quad (\mu, \rho) \longmapsto k\mu^{rk-r} \cdot \rho.$$

### 3. Congruent embeddings

#### 3.1. Statistics and congruent embeddings

Given two measurable sets  $\Omega$  and  $\Omega'$ , a measurable map

$$\kappa : \Omega \longrightarrow \Omega'$$

will be called a *statistic*. Any (signed) measure  $\mu$  on  $\Omega$ , induces a (signed) measure  $\kappa_*\mu$  on  $\Omega'$ , via

$$\kappa_*\mu(A) := \mu(\kappa^{-1}A), \tag{3.1}$$

which is called the *push-forward of  $\mu$  by  $\kappa$* . Note that

$$\kappa_* : \mathcal{S}(\Omega) \longrightarrow \mathcal{S}(\Omega') \tag{3.2}$$

is a bounded linear map which is monotone, that is, it maps nonnegative measures to nonnegative measures. When using the Jordan decomposition (2.2), we obtain

$$\|\kappa_*\mu\|_{\text{TV}} = |\kappa_*\mu_+ - \kappa_*\mu_-|(\Omega') \leq \kappa_*\mu_+(\Omega') + \kappa_*\mu_-(\Omega') = |\mu|(\Omega) = \|\mu\|_{\text{TV}}.$$

Thus,

$$\|\kappa_*\mu\|_{\text{TV}} \leq \|\mu\|_{\text{TV}} \quad \text{with equality iff} \quad \kappa_*\mu_+ \perp \kappa_*\mu_-. \tag{3.3}$$

In particular,  $\kappa_*$  preserves the total variation of nonnegative measures, and whence maps probability measures to probability measures, i.e.

$$\kappa_*(\mathcal{P}(\Omega)) \subset \mathcal{P}(\Omega').$$

Furthermore, if  $\mu_1$  dominates  $\mu_2$ , then  $\kappa_*\mu_1$  dominates  $\kappa_*\mu_2$  by (3.1), whence  $\kappa_*$  yields bounded linear maps

$$\kappa_* : \mathcal{S}(\Omega, \mu) \longrightarrow \mathcal{S}(\Omega', \kappa_*\mu), \tag{3.4}$$

and if we write

$$\kappa_*(\phi\mu) = \phi' \kappa_*\mu, \tag{3.5}$$

then  $\phi' \in L^1(\Omega', \kappa_*\mu)$  is called the *conditional expectation* of  $\phi \in L^1(\Omega, \mu)$  given  $\kappa$ . This yields a bounded linear map

$$\kappa_*^\mu : L^1(\Omega, \mu) \longrightarrow L^1(\Omega', \mu'), \quad \phi \longmapsto \phi' \tag{3.6}$$

with  $\phi'$  from (3.5).

We also define the pull-back of a measurable function  $\phi' : \Omega' \rightarrow \mathbb{R}$  as

$$\kappa^*\phi' := \phi' \circ \kappa.$$

If  $A' \subset \Omega'$  and  $A := \kappa^{-1}(A')$  we have  $\chi_A = \kappa^*\chi_{A'}$ , and thus, (3.1) is equivalent to  $\chi_{A'}\kappa_*\mu = \kappa_*(\chi_A\mu) = \kappa_*(\kappa^*\chi_{A'}\mu)$ , and by linearity and the density of step functions in  $L^1(\Omega', \kappa_*\mu)$  this implies for  $\phi' \in L^1(\Omega', \kappa_*\mu)$

$$\kappa_*(\kappa^*\phi'\mu) = \phi'\kappa_*\mu \quad \text{or, equivalently,} \quad \kappa_*^\mu(\kappa^*\phi') = \phi'. \tag{3.7}$$

Recall that  $\mathcal{M}(\Omega)$  and  $\mathcal{S}(\Omega)$  denote the spaces of *all* (signed) measures on  $\Omega$ , whereas  $\mathcal{M}(\Omega, \mu)$  and  $\mathcal{S}(\Omega, \mu)$  denote the subspaces of the (signed) measures on  $\Omega$  which are dominated by  $\mu$ .

**Definition 3.1 (Congruent embedding).** Let  $\kappa : \Omega \rightarrow \Omega'$  be a statistic and  $\mu' \in \mathcal{M}(\Omega')$ . A  $\kappa$ -congruent embedding is a bounded linear map  $K_* : \mathcal{S}(\Omega', \mu') \rightarrow \mathcal{S}(\Omega)$  such that:

- (1)  $K_*$  is monotone, i.e., it maps nonnegative measures to nonnegative measures, or shortly:  $K_*(\mathcal{M}(\Omega', \mu')) \subset \mathcal{M}(\Omega)$ .
- (2)  $\kappa_*K_*(\nu') = \nu'$  for all  $\nu' \in \mathcal{S}(\Omega', \mu')$ .

Furthermore, the image of a  $\kappa$ -congruent embedding  $K_*$  in  $\mathcal{S}(\Omega)$  is called a  $\kappa$ -congruent subspace of  $\mathcal{S}(\Omega)$ .

**Example 3.1.** Let  $\kappa : \Omega \rightarrow \Omega'$  be a statistic, let  $\mu \in \mathcal{M}(\Omega)$  and  $\mu' := \kappa_*\mu \in \mathcal{M}(\Omega')$ . Then the map

$$K_\mu : \mathcal{S}(\Omega', \mu') \longrightarrow \mathcal{S}(\Omega, \mu) \subset \mathcal{S}(\Omega), \quad \phi'\mu' \longmapsto \kappa^*\phi'\mu \tag{3.8}$$

for all  $\phi' \in L^1(\Omega', \mu')$  is a  $\kappa$ -congruent embedding, since

$$\kappa_*(K_\mu(\phi'\mu')) = \kappa_*(\kappa^*\phi'\mu) \stackrel{(3.7)}{=} \phi'\kappa_*\mu = \phi'\mu'.$$

We shall now see that the above example exhausts *all* possibilities of congruent embeddings.

**Proposition 3.1.** Let  $\kappa : \Omega \rightarrow \Omega'$  be a statistic, let  $K_* : \mathcal{S}(\Omega', \mu') \rightarrow \mathcal{S}(\Omega)$  for some  $\mu' \in \mathcal{M}(\Omega')$  be a  $\kappa$ -congruent embedding, and let  $\mu := K_*\mu' \in \mathcal{M}(\Omega)$ .

Then  $K_* = K_\mu$  with the map  $K_\mu$  given in (3.8).

**Proof.** We have to show that  $K_*(\phi'\mu') = \kappa^*\phi'\mu$  for all  $\phi' \in L^1(\Omega', \mu')$ . By continuity, it suffices to show this for step functions, as these are dense in  $L^1(\Omega', \mu')$ , whence by linearity, we have to show that for all  $A' \subset \Omega'$ ,  $A := \kappa^{-1}(A') \subset \Omega$

$$K_*(\chi_{A'}\mu') = \chi_A\mu. \tag{3.9}$$

Let  $A'_1 := A'$  and  $A'_2 = \Omega' \setminus A'$ , and let  $A_i := \kappa^{-1}(A'_i)$ . We define the measures  $\mu'_i := \chi_{A'_i}\mu' \in \mathcal{M}(\Omega')$ , and  $\mu_i := K_*\mu'_i \in \mathcal{M}(\Omega)$ . Observe that the monotonicity of  $K_*$  implies that  $\mu_i$  are indeed (nonnegative) measures. Since  $\mu'_1 + \mu'_2 = \mu'$ , it follows that  $\mu_1 + \mu_2 = \mu$  by the linearity of  $K_*$ .

Taking indices mod 2, and using  $\kappa_*\mu_i = \kappa_*K_*\mu'_i = \mu'_i$  by the  $\kappa$ -congruency of  $K_*$ , note that

$$\mu_i(A_{i+1}) = \mu_i(\kappa^{-1}(A'_{i+1})) = \kappa_*\mu_i(A'_{i+1}) = \mu'_i(A'_{i+1}) = 0.$$

Thus, for any measurable  $B \subset \Omega$  we have

$$\begin{aligned} \mu_1(B) &= \mu_1(B \cap A_1) && \text{since } \mu_1(B \cap A_2) \leq \mu_1(A_2) = 0 \\ &= \mu_1(B \cap A_1) + \mu_2(B \cap A_1) && \text{since } \mu_2(B \cap A_1) \leq \mu_2(A_1) = 0 \\ &= \mu(B \cap A_1) && \text{since } \mu = \mu_1 + \mu_2 \\ &= (\chi_A\mu)(B) && \text{since } A_1 = A. \end{aligned}$$

That is,  $\chi_A\mu = \mu_1 = K_*\mu'_1 = K_*(\chi_{A'}\mu')$ , so that (3.9) follows. □

### 3.2. Markov kernels and Markov morphisms

**Definition 3.2 (Markov kernel and Markov morphism, cf. [5,10,21]).** A *Markov kernel* between two measurable spaces  $\Omega$  and  $\Omega'$  is a map  $K : \Omega \rightarrow \mathcal{P}(\Omega')$ , associating to each  $\omega \in \Omega$  a probability measure on  $\Omega'$  such that for each fixed measurable  $A' \subset \Omega'$  the map

$$\Omega \longrightarrow [0, 1], \quad \omega \longmapsto K(\omega; A') := K(\omega)(A')$$

is measurable. The *Markov morphism induced by  $K$*  is the linear map

$$K_* : \mathcal{S}(\Omega) \longrightarrow \mathcal{S}(\Omega'), \quad K_*\mu(A') := \int_{\Omega} K(\omega; A') d\mu(\omega). \tag{3.10}$$

We shall use the notation  $K_*(\mu; A') := K_*\mu(A')$ . Since  $K(\omega) \in \mathcal{P}(\Omega')$ , it follows that  $K(\omega; \Omega') = 1$  and hence (3.10) implies that  $K_*\mu(\Omega') = \mu(\Omega)$ . Thus,

$$\|K_*\mu\|_{\text{TV}} = \|\mu\|_{\text{TV}} \quad \text{for all } \mu \in \mathcal{M}(\Omega). \tag{3.11}$$

In particular, a Markov morphism maps probability measures to probability measures. For a general measure  $\mu \in \mathcal{S}(\Omega)$ , (2.3) implies that  $|K_*(\mu; A')| \leq K_*(|\mu|; A')$  for all  $A'$  and hence,

$$\|K_*\mu\|_{\text{TV}} \leq \|K_*|\mu|\|_{\text{TV}} = \|\mu\|_{\text{TV}} \quad \text{for all } \mu \in \mathcal{S}(\Omega),$$

so that  $K_* : \mathcal{S}(\Omega) \rightarrow \mathcal{S}(\Omega')$  is a bounded linear map.

Observe that we can recover the Markov kernel  $K$  from  $K_*$  using the relation

$$K(\omega) = K_*\delta^\omega \quad \text{for all } \omega \in \Omega,$$

where  $\delta^\omega$  denotes the Dirac measure supported at  $\omega \in \Omega$ .

**Remark 3.1.** From (3.10) it is immediate that  $K_*$  preserves dominance of measures, i.e., if  $\mu$  dominates  $\tilde{\mu}$ , then  $K_*\mu$  dominates  $K_*\tilde{\mu}$ . Thus, for each  $\mu \in \mathcal{M}(\Omega)$  there is a restriction

$$K_* : \mathcal{S}(\Omega, \mu) \longrightarrow \mathcal{S}(\Omega', \mu'),$$

where  $\mu' := K_*\mu$ . This again induces a bounded linear map

$$K_*^\mu : L^1(\Omega, \mu) \longrightarrow L^1(\Omega', \mu'), \quad \phi \longmapsto \phi', \tag{3.12}$$

where  $\phi'$  is given by

$$K_*(\phi\mu) = \phi'\mu', \tag{3.13}$$

and as for statistics,  $\phi'$  is called the *conditional expectation of  $\phi$  given  $K$* , cf. (3.5).

**Definition 3.3 (Composition of Markov kernels).** Let  $\Omega_i$ ,  $i = 1, 2, 3$  be measurable spaces, and let  $K_i : \Omega_i \rightarrow \mathcal{P}(\Omega_{i+1})$  for  $i = 1, 2$  be Markov kernels. The *composition of  $K_1$  and  $K_2$*  is the Markov kernel

$$K_2K_1 : \Omega_1 \longrightarrow \mathcal{P}(\Omega_3), \quad \omega \longmapsto (K_2)_*(K_1(\omega)).$$

Since  $\|(K_2)_*(K_1(\omega))\|_{\text{TV}} = \|K_1(\omega)\|_{\text{TV}} = 1$  by (3.11),  $(K_2)_*(K_1(\omega))$  is a probability measure, hence this composition yields indeed a Markov kernel. Moreover, it is straightforward to verify that this composition is associative, and for the induced Markov morphism we have

$$(K_2K_1)_* = (K_2)_*(K_1)_*. \tag{3.14}$$

Markov kernels are generalizations of statistics. In fact, a statistic  $\kappa : \Omega \rightarrow \Omega'$  induces a Markov kernel by

$$K^\kappa(\omega) := \delta^{\kappa(\omega)}, \quad \text{so that } K^\kappa(\omega; A') := \chi_{\kappa^{-1}(A')}(\omega).$$

In this case, the Markov morphism induced by  $K^\kappa$  is the map  $\kappa_* : \mathcal{S}(\Omega) \rightarrow \mathcal{S}(\Omega')$  from (3.2). We shall write the Markov kernel  $K^\kappa$  also as  $\kappa$  if there is no danger of confusion.

**Definition 3.4 (Congruent Markov kernels).** A Markov kernel  $K : \Omega' \rightarrow \mathcal{P}(\Omega)$  is called  *$\kappa$ -congruent* for a statistic  $\kappa : \Omega \rightarrow \Omega'$  if

$$\kappa_*K(\omega') = \delta^{\omega'} \quad \text{for all } \omega' \in \Omega', \tag{3.15}$$

or, equivalently,

$$(K^\kappa K)_* = \text{Id}_{\mathcal{S}(\Omega')} : \mathcal{S}(\Omega') \longrightarrow \mathcal{S}(\Omega').$$

In this case, we also call the induced Markov morphism  $K_* : \mathcal{S}(\Omega') \rightarrow \mathcal{S}(\Omega)$   $\kappa$ -congruent.



In order to relate the notions of  $\kappa$ -congruent Markov morphism and  $\kappa$ -congruent embeddings from Definition 3.1, we need the notion of  $\kappa$ -transverse measures.

**Definition 3.5 (Transverse measures).** Let  $\kappa : \Omega \rightarrow \Omega'$  be a statistic. A measure  $\mu \in \mathcal{M}(\Omega)$  is said to admit  $\kappa$ -transverse measures if there are measures  $\mu_{\omega'}^\perp$  on  $\kappa^{-1}(\omega')$  such that for all  $\phi \in L^1(\Omega, \mu)$

$$\int_{\Omega} \phi \, d\mu = \int_{\Omega'} \left( \int_{\kappa^{-1}(\omega')} \phi \, d\mu_{\omega'}^\perp \right) d\mu'(\omega'), \tag{3.16}$$

where  $\mu' := \kappa_*\mu$ . In particular, the function

$$\Omega' \longrightarrow \hat{\mathbb{R}}, \quad \omega' \longmapsto \int_{\kappa^{-1}(\omega')} \phi \, d\mu_{\omega'}^\perp$$

is measurable for all  $\phi \in L^1(\Omega, \mu)$ .

Observe that the choice of  $\kappa$ -transverse measures  $\mu_{\omega'}^\perp$  is not unique, but rather, one can change these measures for all  $\omega'$  in a  $\mu'$ -null set.

**Proposition 3.2.** Let  $\kappa : \Omega \rightarrow \Omega'$  be a statistic and  $\mu \in \mathcal{M}(\Omega)$  a measure which admits  $\kappa$ -transverse measures  $\{\mu_{\omega'}^\perp : \omega' \in \Omega'\}$ . Then  $\mu_{\omega'}^\perp$  is a probability measure for almost every  $\omega' \in \Omega'$  and hence, we may assume w.l.o.g. that  $\mu_{\omega'}^\perp \in \mathcal{P}(\kappa^{-1}(\omega'))$  for all  $\omega' \in \Omega'$ .

**Proof.** Given  $\varepsilon > 0$ , define  $A'_\varepsilon := \{\omega' \in \Omega' : \mu_{\omega'}^\perp(\kappa^{-1}(\omega')) \geq 1 + \varepsilon\}$ . Then for  $\phi := \chi_{\kappa^{-1}(A'_\varepsilon)}$  the two sides of equation (3.16) read

$$\begin{aligned} \int_{\Omega} \chi_{\kappa^{-1}(A'_\varepsilon)} \, d\mu &= \mu(\kappa^{-1}(A'_\varepsilon)) = \mu'(A'_\varepsilon) \\ \int_{\Omega'} \left( \int_{\kappa^{-1}(\omega')} \chi_{\kappa^{-1}(A'_\varepsilon)} \, d\mu_{\omega'}^\perp \right) d\mu'(\omega') &= \int_{A'_\varepsilon} \left( \int_{\kappa^{-1}(\omega')} d\mu_{\omega'}^\perp \right) d\mu'(\omega') \\ &= \int_{A'_\varepsilon} \mu_{\omega'}^\perp(\kappa^{-1}(\omega')) \, d\mu'(\omega') \\ &\geq (1 + \varepsilon)\mu'(A'_\varepsilon). \end{aligned}$$

Thus, (3.16) implies

$$\mu'(A'_\varepsilon) \geq (1 + \varepsilon)\mu'(A'_\varepsilon),$$

and hence,  $\mu'(A'_\varepsilon) = 0$  for all  $\varepsilon > 0$ . Thus,

$$\mu'(\{\omega' \in \Omega' : \mu_{\omega'}^\perp(\kappa^{-1}(\omega')) > 1\}) = \mu' \left( \bigcup_{n=1}^{\infty} A'_{1/n} \right) \leq \sum_{n=1}^{\infty} \mu'(A'_{1/n}) = 0,$$

whence  $\{\omega' \in \Omega' : \mu_{\omega'}^\perp(\kappa^{-1}(\omega')) > 1\}$  is a  $\mu'$ -null set. Analogously,  $\{\omega' \in \Omega' : \mu_{\omega'}^\perp(\kappa^{-1}(\omega')) < 1\}$  is a  $\mu'$ -null set, that is,  $\mu_{\omega'}^\perp \in \mathcal{P}(\kappa^{-1}(\omega'))$  and hence  $\|\mu_{\omega'}^\perp\|_{TV} = 1$  for  $\mu'$ -a.e.  $\omega' \in \Omega'$ .

Thus, if we replace  $\mu_{\omega'}^\perp$  by  $\tilde{\mu}_{\omega'}^\perp := \|\mu_{\omega'}^\perp\|_{TV}^{-1} \mu_{\omega'}^\perp$ , then  $\tilde{\mu}_{\omega'}^\perp \in \mathcal{P}(\kappa^{-1}(\omega'))$  for all  $\omega' \in \Omega'$ , and since  $\tilde{\mu}_{\omega'}^\perp = \mu_{\omega'}^\perp$  for  $\mu'$ -a.e.  $\omega' \in \Omega'$ , it follows that (3.16) holds when replacing  $\mu_{\omega'}^\perp$  by  $\tilde{\mu}_{\omega'}^\perp$ .  $\square$

We are now ready to relate the notions of  $\kappa$ -congruent embeddings and  $\kappa$ -congruent Markov kernels.

**Theorem 3.1.** *Let  $\kappa : \Omega \rightarrow \Omega'$  be a statistic and  $\mu' \in \mathcal{M}(\Omega')$  be a measure.*

- (1) *If  $K : \Omega' \rightarrow \mathcal{P}(\Omega)$  is a  $\kappa$ -congruent Markov kernel, then the restriction of  $K_*$  to  $\mathcal{S}(\Omega', \mu') \subset \mathcal{S}(\Omega')$  is a  $\kappa$ -congruent embedding and hence, for  $\phi' \in L^1(\Omega', \mu')$  we have*

$$K_*(\phi' \mu') = \kappa^* \phi' K_* \mu', \quad \text{or, equivalently,} \quad K_*^{\mu'}(\phi') = \kappa^* \phi'.$$

- (2) *Conversely, if  $K_* : \mathcal{S}(\Omega', \mu') \rightarrow \mathcal{S}(\Omega)$  is a  $\kappa$ -congruent embedding, then the following are equivalent.*
  - (a)  *$K_*$  is the restriction of a  $\kappa$ -congruent Markov morphism to  $\mathcal{S}(\Omega', \mu') \subset \mathcal{S}(\Omega')$ .*
  - (b)  *$\mu := K_* \mu' \in \mathcal{S}(\Omega)$  admits  $\kappa$ -transverse measures.*

Theorem 3.1 implies that the two notions of congruency, that is, congruent embeddings and congruent Markov morphisms, are equivalent for large classes of statistics  $\kappa$ , since the existence of transversal measures is guaranteed under rather mild hypotheses, for example, if one of  $\Omega, \Omega'$  is a finite set, or if  $\Omega, \Omega'$  are differentiable manifolds equipped with a Borel measure  $\mu$  and  $\kappa$  is a differentiable map.

However, there are examples of statistics and measures which do not admit  $\kappa$ -transverse measures, cf. Example 3.2 below.

**Proof of Theorem 3.1.** The first statement follows directly from  $(K^\kappa K)_* = (K^\kappa)_* K_* = \kappa_* K_*$  by (3.14) and Proposition 3.1.

For the second, suppose that  $K_* : \mathcal{S}(\Omega', \mu') \rightarrow \mathcal{S}(\Omega)$  is a  $\kappa$ -congruent embedding. Then  $K_* = K_\mu$  given in (3.8) for the measure  $\mu := K_* \mu'$  by Proposition 3.1.

If we assume that  $K_*$  is the restriction of a  $\kappa$ -congruent Markov morphism induced by the  $\kappa$ -congruent Markov kernel  $K : \Omega' \rightarrow \mathcal{P}(\Omega)$ , then we define the measures

$$\mu_{\omega'}^\perp := K(\omega')|_{\kappa^{-1}(\omega)} \in \mathcal{M}(\kappa^{-1}(\omega')).$$

Note that for  $\omega' \in \Omega'$

$$\begin{aligned} K(\omega'; \Omega \setminus \kappa^{-1}(\omega')) &= \int_{\Omega \setminus \kappa^{-1}(\omega')} dK(\omega') = \int_{\Omega' \setminus \omega'} d\kappa_* K(\omega') \\ &\stackrel{(3.15)}{=} \int_{\Omega' \setminus \omega'} d\delta^{\omega'} = 0. \end{aligned}$$

That is,  $K(\omega')$  is supported on  $\kappa^{-1}(\omega')$  and hence, for an arbitrary set  $A \subset \Omega$  we have

$$K(\omega'; A) = K(\omega'; A \cap \kappa^{-1}(\omega')) = \mu_{\omega'}^\perp(A \cap \kappa^{-1}(\omega')) = \int_{\kappa^{-1}(\omega')} \chi_A d\mu_{\omega'}^\perp.$$

Substituting this into the definition of  $K_*$  we obtain for a subset  $A \subset \Omega$

$$\begin{aligned} \int_{\Omega} \chi_A d\mu &= \mu(A) = K_*(\mu'; A) \stackrel{(3.10)}{=} \int_{\Omega'} K(\omega'; A) d\mu'(\omega') \\ &= \int_{\Omega'} \left( \int_{\kappa^{-1}(\omega')} \chi_A d\mu_{\omega'}^\perp \right) d\mu'(\omega'), \end{aligned}$$

showing that (3.16) holds for  $\phi = \chi_A$ . But then, by linearity (3.16) holds for any step function  $\phi$ , and since these are dense in  $L^1(\Omega, \mu)$ , it follows that (3.16) holds for all  $\phi$ , so that the measures  $\mu_{\omega'}^\perp$  defined above yield indeed  $\kappa$ -transverse measures of  $\mu$ .

Conversely, suppose that  $\mu := K_*\mu'$  admits  $\kappa$ -transverse measures  $\mu_{\omega'}^\perp$ , and by Proposition 3.2 we may assume w.l.o.g. that  $\mu_{\omega'}^\perp \in \mathcal{P}(\kappa^{-1}(\omega'))$ . Then we define the map

$$K : \Omega' \longrightarrow \mathcal{P}(\Omega), \quad K(\omega'; A) := \mu_{\omega'}^\perp(A \cap \kappa^{-1}(\omega')) = \int_{\kappa^{-1}(\omega')} \chi_A d\mu_{\omega'}^\perp.$$

Since for fixed  $A \subset \Omega$  the map  $\omega' \mapsto \int_{\kappa^{-1}(\omega')} \chi_A d\mu_{\omega'}^\perp$  is measurable by the definition of transversal measures,  $K$  is indeed a Markov kernel. Moreover, for  $A' \subset \Omega'$

$$\kappa_* K(\omega')(A') = K(\omega'; \kappa^{-1}(A')) = \mu_{\omega'}^\perp(\kappa^{-1}(A') \cap \kappa^{-1}(\omega')) = \delta^{\omega'}(A'),$$

so that  $\kappa_* K(\omega') = \delta^{\omega'}$  for all  $\omega' \in \Omega'$ , whence  $K$  is  $\kappa$ -congruent. Moreover, for any  $\phi' \in L^1(\Omega', \mu')$  and  $A \subset \Omega$  we have

$$\begin{aligned} K_\mu(\phi'\mu')(A) &\stackrel{(3.8)}{=} \kappa^*\phi'\mu(A) = \int_{\Omega} \chi_A \kappa^*\phi' d\mu \\ &\stackrel{(3.16)}{=} \int_{\Omega'} \left( \int_{\kappa^{-1}(\omega')} \chi_A \kappa^*\phi' d\mu_{\omega'}^\perp \right) d\mu'(\omega') \\ &= \int_{\Omega'} \left( \int_{\kappa^{-1}(\omega')} \chi_A d\mu_{\omega'}^\perp \right) \phi'(\omega') d\mu'(\omega') \\ &= \int_{\Omega'} K(\omega'; A) d(\phi'\mu')(\omega') \stackrel{(3.10)}{=} K_*(\phi'\mu')(A). \end{aligned}$$

Thus,  $K_\mu(\phi'\mu') = K_*(\phi'\mu')$  for all  $\phi' \in L^1(\Omega', \mu')$  and hence,  $K_\mu(v) = K_*v$  for all  $v \in \mathcal{S}(\Omega', \mu')$ . That is, the given congruent embedding  $K_\mu$  coincides with the Markov morphism  $K_*$  induced by  $K$ , and this completes the proof.  $\square$

Now we give an example of a statistic which does not admit  $\kappa$ -transverse measures.

**Example 3.2.** Let  $\Omega := S^1$  be the unit circle group in the complex plain with the 1-dimensional Borel algebra  $\mathfrak{B}$ . Let  $\Gamma := \exp(2\pi\sqrt{-1}\mathbb{Q}) \subset S^1$  be the subgroup of rational rotations, and let  $\Omega' := S^1/\Gamma$  be the quotient space with the canonical projection  $\kappa : \Omega \rightarrow \Omega'$ . Let  $\mathfrak{B}' := \{A' \subset \Omega' : \kappa^{-1}(A') \in \mathfrak{B}\}$ , so that  $\kappa : \Omega \rightarrow \Omega'$  is measurable. For  $\gamma \in \Gamma$ , we let  $m_\gamma : S^1 \rightarrow S^1$  denote the multiplication by  $\gamma$ .

Let  $\lambda$  be the 1-dimensional Lebesgue measure on  $\Omega$  and  $\lambda' := \kappa_*\lambda$  be the induced measure on  $\Omega'$ . Suppose that  $\lambda$  admits  $\kappa$ -transverse measures  $(\lambda_{\omega'}^\perp)_{\omega' \in \Omega'}$ . Then for each  $A \in \mathfrak{B}$  we have

$$\lambda(A) = \int_{\Omega'} \left( \int_{A \cap \kappa^{-1}(\omega')} d\lambda_{\omega'}^\perp \right) d\lambda'(\omega'). \quad (3.17)$$

Since  $\lambda$  is invariant under rotations, we have on the other hand for  $\gamma \in \Gamma$

$$\begin{aligned} \lambda(A) &= \lambda(m_\gamma^{-1}A) = \int_{\Omega'} \left( \int_{(m_\gamma^{-1}A) \cap \kappa^{-1}(\omega')} d\lambda_{\omega'}^\perp \right) d\lambda'(\omega') \\ &= \int_{\Omega'} \left( \int_{A \cap \kappa^{-1}(\omega')} d((m_\gamma)_*\lambda_{\omega'}^\perp) \right) d\lambda'(\omega'). \end{aligned} \quad (3.18)$$

Comparing (3.17) and (3.18) implies that  $((m_\gamma)_*\lambda_{\omega'}^\perp)_{\omega' \in \Omega'}$  is another family of  $\kappa$ -transverse measures of  $\lambda$  which implies that  $(m_\gamma)_*\lambda_{\omega'}^\perp = \lambda_{\omega'}^\perp$  for  $\lambda'$ -a.e.  $\omega' \in \Omega'$ , and as  $\Gamma$  is countable, it follows that

$$(m_\gamma)_*\lambda_{\omega'}^\perp = \lambda_{\omega'}^\perp \quad \text{for all } \gamma \in \Gamma \text{ and } \lambda'\text{-a.e. } \omega' \in \Omega'.$$

Thus, for a.e.  $\omega' \in \Omega'$  we have  $\lambda_{\omega'}^\perp(\{\gamma \cdot x\}) = \lambda_{\omega'}^\perp(\{x\})$ , and since  $\Gamma$  acts transitively on  $\kappa^{-1}(\omega')$ , it follows that singleton subsets have equal measure, i.e., there is a constant  $c_{\omega'}$  with

$$\lambda_{\omega'}^\perp(A') = c_{\omega'} |A'|$$

for all  $A' \subset \kappa^{-1}(\omega')$ . As  $\kappa^{-1}(\omega')$  is countable infinite, this implies that  $\lambda_{\omega'}^\perp = 0$  if  $c_{\omega'} = 0$ , and  $\lambda_{\omega'}^\perp(\kappa^{-1}(\omega')) = \infty$  if  $c_{\omega'} > 0$ . Thus,  $\lambda_{\omega'}^\perp$  is not a probability measure for a.e.  $\omega' \in \Omega'$ , contradicting Proposition 3.2. This shows that  $\lambda$  does not admit  $\kappa$ -transverse measures.

We conclude this section by the following result (cf. [5], Theorem 4.10).

**Theorem 3.2.** Any Markov kernel  $K = \Omega \rightarrow \mathcal{P}(\Omega')$  can be decomposed into a statistic and a congruent Markov kernel. That is, there is a Markov kernel  $K^{\text{cong}} : \Omega \rightarrow \mathcal{P}(\hat{\Omega})$  which is congruent w.r.t. some statistic  $\kappa_1 : \hat{\Omega} \rightarrow \Omega$ , and a statistic  $\kappa_2 : \hat{\Omega} \rightarrow \Omega'$  such that

$$K = K^{\kappa_2} K^{\text{cong}}.$$

**Proof.** Let  $\hat{\Omega} := \Omega \times \Omega'$  and let  $\kappa_1 : \hat{\Omega} \rightarrow \Omega$  and  $\kappa_2 : \hat{\Omega} \rightarrow \Omega'$  be the canonical projections. We define the Markov kernel

$$K^{\text{cong}} : \Omega \longrightarrow \mathcal{P}(\hat{\Omega}), \quad K^{\text{cong}}(\omega) := \delta^\omega \times K(\omega),$$

i.e.,  $K^{\text{cong}}(\omega; \hat{A}) := K(\omega; \kappa_2(\hat{A} \cap (\{\omega\} \times \Omega')))$  for  $\hat{A} \subset \hat{\Omega}$ . Then evidently,  $(\kappa_1)_*(K^{\text{cong}}(\omega)) = \delta^\omega$ , so that  $K^{\text{cong}}$  is  $\kappa_1$ -congruent, and  $(\kappa_2)_*K^{\text{cong}}(\omega) = K(\omega)$ , so the claim follows.  $\square$

### 3.3. Powers of densities and congruent embeddings

As we saw in the preceding section, a Markov kernel  $K : \Omega \rightarrow \mathcal{P}(\Omega')$  (e.g., a statistic  $\kappa : \Omega \rightarrow \Omega'$ ), induces the monotone bounded linear map  $K_* : \mathcal{S}(\Omega) \rightarrow \mathcal{S}(\Omega')$  from (3.10) and for each  $\mu \in \mathcal{M}(\Omega)$  the restriction yields a bounded linear map  $K_* : \mathcal{S}(\Omega, \mu) \rightarrow \mathcal{S}(\Omega', \mu')$ , where  $\mu' := K_*\mu \in \mathcal{M}(\Omega')$ . This induces the bounded linear map  $K_*^\mu : L^1(\Omega, \mu) \rightarrow L^1(\Omega', \mu')$  from (3.12) (or in case of a statistic, the map  $\kappa_*^\mu : L^1(\Omega, \mu) \rightarrow L^1(\Omega', \mu')$  from (3.6), respectively).

We wish to show that when restricting this map to  $L^k(\Omega, \mu) \subset L^1(\Omega', \mu')$ , the  $k$ -regularity is preserved by  $\kappa_*^\mu$  and  $K_*^\mu$ , respectively, cf. Theorem 3.3 below. The first step towards this is to consider congruent Markov kernels.

**Proposition 3.3.** *Let  $K : \Omega_1 \rightarrow \mathcal{P}(\Omega_2)$  be a Markov kernel which is congruent w.r.t. some statistic  $\kappa : \Omega_2 \rightarrow \Omega_1$ . Let  $\mu_1 \in \mathcal{M}(\Omega_1)$  and  $\mu_2 := K_*\mu_1 \in \mathcal{M}(\Omega_2)$ , and consider the map  $K_*^{\mu_1} : L^1(\Omega_1, \mu_1) \rightarrow L^1(\Omega_2, \mu_2)$ .*

*Then for all  $\phi \in L^k(\Omega_1, \mu_1)$  with  $1 \leq k \leq \infty$  we have  $\phi' := K_*^{\mu_1}(\phi) \in L^k(\Omega_2, \mu_2)$ , and*

$$\|\phi'\|_k = \|\phi\|_k.$$

**Proof.** Since  $K$  is  $\kappa$ -congruent, by Theorem 3.1 we have  $\phi' := K_*^{\mu_1}(\phi) = \kappa^*\phi$ . Thus, for  $1 \leq k < \infty$ ,

$$\|\phi'\|_k^k = \int_{\Omega_2} |\phi'|^k d\mu_2 = \int_{\Omega_2} \kappa^*|\phi|^k d\kappa_*\mu_1 = \int_{\Omega_1} |\phi|^k d\mu_1 = \|\phi\|_k^k, \tag{3.19}$$

showing the assertion. For  $k = \infty$ ,  $\|\phi'\|_\infty = \|\kappa^*\phi\|_\infty = \|\phi\|_\infty$  is obvious.  $\square$

Next, we shall deal with statistics  $\kappa : \Omega \rightarrow \Omega'$ .

**Proposition 3.4.** *Let  $\kappa : \Omega \rightarrow \Omega'$  be a statistic and  $\mu \in \mathcal{M}(\Omega)$ ,  $\mu' := \kappa_*\mu \in \mathcal{M}(\Omega')$ , and let  $\kappa_*^\mu : L^1(\Omega; \mu) \rightarrow L^1(\Omega', \mu')$  be the map from (3.6). Then the following hold.*

(1) *If  $\phi \in L^k(\Omega, \mu)$  for  $1 \leq k \leq \infty$ , then  $\phi' := \kappa_*^\mu(\phi) \in L^k(\Omega', \mu')$ , and*

$$\|\phi'\|_k \leq \|\phi\|_k. \tag{3.20}$$

(2) *For  $1 < k < \infty$ , equality in (3.20) holds iff  $\phi = \kappa^*\phi'$ .*

**Remark 3.2.** The estimate (3.20) in Proposition 3.4 also follows from [23], Proposition IV.3.1.

**Proof of Proposition 3.4.** We decompose  $\phi = \phi_+ - \phi_-$  as in (2.1). Then  $\|\phi'\|_1 = \|\kappa_*(\phi\mu)\|_{\text{TV}}$  and  $\|\phi\|_1 = \|\phi\mu\|_{\text{TV}}$ , so that (3.3) implies (3.20) for  $k = 1$ .

If  $\phi \in L^\infty(\Omega, \mu)$ , then  $|\phi\mu| \leq \|\phi\|_\infty\mu$ , and by monotonicity of  $\kappa_*$  it follows that

$$|\phi'\mu'| = |\kappa_*(\phi\mu)| \leq \kappa_*|\phi\mu| \leq \|\phi\|_\infty\mu',$$

whence  $\|\phi'\|_\infty \leq \|\phi\|_\infty$ , so that (3.20) holds for  $k = \infty$ .

For  $\phi' \in L^k(\Omega', \mu')$  (3.19) implies that  $\kappa^*\phi' \in L^k(\Omega, \mu)$  and

$$\|\kappa^*\phi'\|_k = \|\phi'\|_k \quad \text{for all } \phi' \in L^k(\Omega', \mu'), 1 \leq k \leq \infty. \tag{3.21}$$

Suppose now that  $\phi \in L^k(\Omega, \mu)$  with  $1 < k < \infty$  is such that  $\phi' \in L^k(\Omega', \mu')$ , and assume that  $\phi \geq 0$  and hence,  $\phi' \geq 0$ . Then

$$\begin{aligned} \|\phi'\|_k^k &= \int_{\Omega'} \phi'^k d\mu' = \int_{\Omega'} \phi'^{k-1} d\kappa_*(\phi\mu) = \int_{\Omega} \kappa^*(\phi'^{k-1})\phi d\mu \\ &\stackrel{(*)}{\leq} \|\kappa^*(\phi'^{k-1})\|_{k/(k-1)} \|\phi\|_k \\ &\stackrel{(**)}{=} \|\phi'^{k-1}\|_{k/(k-1)} \|\phi\|_k = \|\phi'\|_k^{k-1} \|\phi\|_k. \end{aligned}$$

From this, (3.20) follows. Here we used Hölder's inequality at (\*), and (3.21) applied to  $\phi'^{k-1} \in L^{k/(k-1)}(\Omega', \mu')$  at (\*\*). Moreover, equality at (\*) holds iff  $\phi = c\kappa^*\phi'$  for some  $c \in \mathbb{R}$ , and the fact that  $\kappa_*(\phi\mu) = \phi'\mu'$  easily implies that  $c = 1$ , i.e., equality in (3.20) occurs iff  $\phi = \kappa^*\phi'$ .

If we drop the assumption that  $\phi \geq 0$ , we decompose  $\phi = \phi_+ - \phi_-$  as in (2.1) and let  $\phi'_\pm := \kappa_*^\mu(\phi_\pm) \geq 0$ . Although in general,  $\phi'_+$  and  $\phi'_-$  do not have disjoint support, the linearity of  $\kappa_*$  still implies that  $\phi' = \phi'_+ - \phi'_-$ . Let us assume that  $\phi'_\pm \in L^k(\Omega', \mu')$ . Then

$$\|\phi'\|_k = \|\phi'_+ - \phi'_-\|_k \leq \|\phi'_+\|_k + \|\phi'_-\|_k \leq \|\phi_+\|_k + \|\phi_-\|_k = \|\phi\|_k,$$

using (3.20) applied to  $\phi_\pm \geq 0$  in the second estimate. Equality in the second estimate holds iff  $\phi_\pm = \kappa^*\phi'_\pm$ , and thus,  $\phi = \phi_+ - \phi_- = \kappa^*(\phi'_+ - \phi'_-) = \kappa^*\phi'$ .

Thus, it remains to show that  $\phi' \in L^k(\Omega', \mu')$  whenever  $\phi \in L^k(\Omega, \mu)$ . For this, let  $\phi \in L^k(\Omega, \mu)$ ,  $(\phi_n)_{n \in \mathbb{N}}$  be a sequence in  $L^\infty(\Omega, \mu)$  converging to  $\phi$  in  $L^k(\Omega, \mu)$ , and let  $\phi'_n := \kappa_*^\mu(\phi_n) \in L^\infty(\Omega', \mu') \subset L^k(\Omega', \mu')$ . As  $(\phi'_n - \phi'_m)_\pm \in L^\infty(\Omega', \mu') \subset L^k(\Omega', \mu')$ , (3.20) holds for  $\phi_n - \phi_m$  by the previous argument, that is,

$$\|\phi'_n - \phi'_m\|_k \leq \|\phi_n - \phi_m\|_k,$$

which tends to 0 for  $n, m \rightarrow \infty$ , as  $(\phi)_n$  is convergent and hence a Cauchy sequence in  $L^k(\Omega, \mu)$ . Thus,  $(\phi'_n)$  is also a Cauchy sequence, whence it converges to some  $\tilde{\phi}' \in L^k(\Omega', \mu')$ . It follows that  $\phi_n - \kappa^*\phi'_n$  converges in  $L^k(\Omega, \mu)$  to  $\phi - \kappa^*\tilde{\phi}'$ , and as  $\kappa_*((\phi_n - \kappa^*\phi'_n)\mu) = 0$  for all  $n$ , we have

$$0 = \kappa_*((\phi - \kappa^*\tilde{\phi}')\mu) = \phi'\mu' - \tilde{\phi}'\mu',$$

whence  $\phi' = \tilde{\phi}' \in L^k(\Omega', \mu')$ . □

Putting the last two results together, we obtain the following theorem.

**Theorem 3.3.** *Let  $K : \Omega \rightarrow \mathcal{P}(\Omega')$  be a Markov kernel,  $\mu \in \mathcal{M}(\Omega)$  and  $\mu' := K_*\mu \in \mathcal{M}(\Omega')$ . Then for  $\phi \in L^k(\Omega, \mu)$  with  $1 \leq k \leq \infty$  we have  $K_*^\mu(\phi) \in L^k(\Omega', \mu')$ , and*

$$\|K_*^\mu(\phi)\|_k \leq \|\phi\|_k.$$

**Proof.** By Theorem 3.2, we can decompose  $K = \kappa_* K^{\text{cong}}$ , where  $K^{\text{cong}} : \Omega \rightarrow \mathcal{P}(\hat{\Omega})$  is congruent w.r.t. some statistic  $\hat{\kappa} : \hat{\Omega} \rightarrow \Omega$ , and with a statistic  $\kappa : \hat{\Omega} \rightarrow \Omega'$ . Then it follows that  $K_* = \kappa_* K_*^{\text{cong}}$ , and whence,

$$K_*^\mu = \kappa_*^{\hat{\mu}} K_*^{\text{cong}\mu},$$

where  $\hat{\mu} := K_*^{\text{cong}}(\mu) \in \mathcal{M}(\hat{\Omega})$ . Given  $\phi \in L^k(\Omega, \mu)$ , then by Theorem 3.1,  $\hat{\phi} := K_*^{\text{cong}\mu}(\phi) = \hat{\kappa}^*\phi$ , whence  $\phi' := K_*^\mu(\phi) = \kappa_*^{\hat{\mu}}(\hat{\phi})$ . Thus,

$$\|K_*^\mu(\phi)\|_k = \|\kappa_*^{\hat{\mu}}(\hat{\phi})\|_k \leq \|\hat{\phi}\|_k = \|\phi\|_k,$$

where the first estimate follows from Proposition 3.4, whereas the second equation follows from Proposition 3.3. □

**Remark 3.3.** Theorem 3.3 can be interpreted in a different way. Namely, given a Markov kernel  $K : \Omega \rightarrow \mathcal{P}(\Omega')$  and  $r \in (0, 1]$ , one can define the map  $K_*^r : \mathcal{S}^r(\Omega) \rightarrow \mathcal{S}^r(\Omega')$  by

$$K_*^r(\tilde{\mu}^r) := \tilde{\pi}^r(K_*\mu) \quad \text{for } \mu \in \mathcal{S}(\Omega), \tag{3.22}$$

with the signed  $r$ th power  $\tilde{\mu}^r$  defined before. Since  $\tilde{\pi}^r$  and  $\tilde{\pi}^{1/r}$  are both continuous by Proposition 2.2, the map  $K_*^r$  is continuous, but it fails to be  $C^1$  for  $r < 1$ , even for finite  $\Omega$ .

Let  $\mu \in \mathcal{M}(\Omega)$  and  $\mu' := K_*\mu \in \mathcal{M}(\Omega')$ , so that  $K_*^r(\mu^r) = \mu'^r$ . If there was a derivative of  $K_*^r$  at  $\mu^r$ , then it would have to be a map between the tangent spaces  $T_{\mu^r}\mathcal{M}(\Omega)$  and  $T_{\mu'^r}\mathcal{M}(\Omega')$ , i.e., according to Proposition 2.1 between  $\mathcal{S}^r(\Omega, \mu)$  and  $\mathcal{S}^r(\Omega', \mu')$ . Let  $k := 1/r > 1$ ,  $\phi \in L^k(\Omega, \mu) \subset L^1(\Omega, \mu)$ , so that  $\phi' := K_*^\mu(\phi) \in L^k(\Omega', \mu')$  by Theorem 3.3. Then by Proposition 2.2 and the chain rule we obtain

$$\begin{aligned} d(\tilde{\pi}^k K_*^r)_{\mu^r}(\phi\mu^r) &= k\mu'^{1-r} \cdot d(K_*^r)_{\mu^r}(\phi\mu^r), \\ d(K_*\tilde{\pi}^k)_{\mu'^r}(\phi\mu'^r) &= kK_*(\phi\mu) = k\phi'\mu', \end{aligned}$$

and these should coincide as  $\tilde{\pi}^k K_*^r = K_*\tilde{\pi}^k$  by (3.22). Since  $d(K_*^r)_{\mu^r}(\phi\mu^r) \in \mathcal{S}^r(\Omega', \mu')$ , we thus must have

$$d(K_*^r)_{\mu^r}(\phi\mu^r) = \phi'\mu'^r, \quad \text{where } \phi' = K_*^\mu(\phi). \tag{3.23}$$

Thus, Theorem 3.3 states that this map is a well defined linear operator with operator norm  $\leq 1$ . The map  $d(K_*^r)_{\mu^r} : \mathcal{S}^r(\Omega, \mu) \rightarrow \mathcal{S}^r(\Omega', \mu')$  from (3.23) is called the *formal derivative of  $K_*^r$  at  $\mu^r$* .

## 4. Parametrized measure models and $k$ -integrability

In this section, we shall now present our notion of a parametrized measure model.

**Definition 4.1 (Parametrized measure model).** Let  $\Omega$  be a measure space.

- (1) A *parametrized measure model* is a triple  $(M, \Omega, \mathbf{p})$  where  $M$  is a (finite or infinite dimensional) Banach manifold and  $\mathbf{p} : M \rightarrow \mathcal{M}(\Omega) \subset \mathcal{S}(\Omega)$  is a  $C^1$ -map in the sense explained in Section 2.2.
- (2) The triple  $(M, \Omega, \mathbf{p})$  is called a *statistical model* if it consists only of probability measures, that is, such that the image of  $\mathbf{p}$  is contained in  $\mathcal{P}(\Omega)$ .
- (3) We call such a model *dominated by  $\mu_0$*  if the image of  $\mathbf{p}$  is contained in  $\mathcal{S}(\Omega, \mu_0)$ . In this case, we use the notation  $(M, \Omega, \mu_0, \mathbf{p})$  for this model.

**Remark 4.1.** Evidently, for the applications we have in mind, we are interested mainly in statistical models. However, we can take the point of view that  $\mathcal{P}(\Omega)$  is the projectivization of  $\mathcal{S}(\Omega) = \mathbb{P}(\mathcal{M}(\Omega) \setminus 0)$  via rescaling. Thus, given a parametrized measure model  $(M, \Omega, \mathbf{p})$ , normalization yields a statistical model  $(M, \Omega, \mu_0, \mathbf{p}_0)$  defined by

$$\mathbf{p}_0(\xi) := \frac{\mathbf{p}(\xi)}{\|\mathbf{p}(\xi)\|_{\text{TV}}},$$

which is again a  $C^1$ -map. Indeed, the map  $\mu \mapsto \|\mu\|_{\text{TV}}$  on  $\mathcal{M}(\Omega)$  is a  $C^1$ -map, being the restriction of the linear (and hence differentiable) map  $\mu \mapsto \int_{\Omega} d\mu$  on  $\mathcal{S}(\Omega)$ .

Observe that while  $\mathcal{S}(\Omega)$  is a Banach space, the subsets  $\mathcal{M}(\Omega)$  and  $\mathcal{P}(\Omega)$  do not carry a canonical manifold structure.

If a parametrized measure model  $(M, \Omega, \mu_0, \mathbf{p})$  is dominated by  $\mu_0$ , then there is a *density function*  $p : \Omega \times M \rightarrow \mathbb{R}$  such that

$$\mathbf{p}(\xi) = p(\cdot; \xi)\mu_0. \tag{4.1}$$

Evidently, we must have  $p(\cdot; \xi) \in L^1(\Omega, \mu_0)$  for all  $\xi$ . In particular, for fixed  $\xi$ ,  $p(\cdot; \xi)$  is defined only up to changes on a  $\mu_0$ -null set.

**Definition 4.2 (Regular density function).**

Let  $(M, \Omega, \mu_0, \mathbf{p})$  be a parametrized measure model dominated by  $\mu_0$ . We say that this model has a *regular density function* if the density function  $p : \Omega \times M \rightarrow \mathbb{R}$  satisfying (4.1) can be chosen such that for all  $V \in T_{\xi}M$  the partial derivative  $\partial_V p(\cdot; \xi)$  exists and lies in  $L^1(\Omega, \mu_0)$ .

**Remark 4.2.** The standard notion of a statistical model always assumes that it is dominated by some measure and has a positive regular density function (e.g., [4], Section 2, p. 25, [3], Section 2.1, [27], [5], Definition 2.4). In fact, the definition of a parametrized measure model or statistical model in [5], Definition 2.4, is equivalent to a parametrized measure model or statistical model with a positive regular density function in the sense of Definition 4.2.



Let us point out why the present notion is indeed more general. The formal definition of differentiability of  $\mathbf{p}$  implies that for each  $C^1$ -path  $\xi(t) \in M$  with  $\xi(0) = \xi, \dot{\xi}(0) =: V \in T_\xi M$ , the curve  $t \mapsto p(\cdot; \xi(t)) \in L^1(\Omega, \mu_0)$  is differentiable. This implies that there is a  $d_\xi \mathbf{p}(V) \in L^1(\Omega, \mu_0)$  such that

$$\left\| \frac{p(\cdot; \xi(t)) - p(\cdot; \xi)}{t} - d_\xi \mathbf{p}(V)(\cdot) \right\|_1 \xrightarrow{t \rightarrow 0} 0.$$

If this is a *pointwise* convergence, then  $d_\xi \mathbf{p}(V) = \partial_V p(\cdot; \xi)$  is the partial derivative and whence,  $\partial_V p(\cdot; \xi)$  lies in  $L^1(\Omega, \mu_0)$ , so that the density function is regular.

However, in general convergence in  $L^1(\Omega, \mu_0)$  does *not* imply pointwise convergence, whence there are parametrized measure models in the sense of Definition 4.1 without a regular density function, cf. Example 4.1 below. Nevertheless, for simplicity we shall frequently use the notation  $\partial_V p(\cdot; \xi)$  instead of  $d_\xi \mathbf{p}(V)(\cdot)$ , even if the density function is *not* regular.

By this convention, for a parametrized measure model  $(M, \Omega, \mu_0, \mathbf{p})$  we can describe its derivative in the direction of  $V \in T_\xi M$  as

$$d_\xi \mathbf{p}(V) = \partial_V p(\cdot; \xi) \mu_0.$$

**Example 4.1.** To see that there are parametrized measure models without a regular density function, consider the family of measures on  $\Omega = (0, \pi)$  and  $\xi \in (-1, \infty)$

$$\mathbf{p}(\xi) := p(t; \xi) dt \quad \text{with } p(t; \xi) = \begin{cases} (1 + \xi (\sin^2(t - 1/\xi))^{1/\xi^2}) dt & \text{for } \xi \neq 0, \\ 1 & \text{for } \xi = 0. \end{cases}$$

This model is dominated by the Lebesgue measure  $dt$  with density function  $p$ , and the partial derivative  $\partial_\xi p$  does not exist at  $\xi = 0$ , whence the density function is not regular.

On the other hand,  $\mathbf{p} : \mathbb{R} \rightarrow \mathcal{M}(\Omega, dt)$  is differentiable in the above sense at  $\xi = 0$  with  $d_0 \mathbf{p}(\partial_\xi) = 0$ , so that  $(M, \Omega, \mathbf{p})$  is a parametrized measure model in the sense of Definition 4.1. To see this, we calculate

$$\begin{aligned} \left\| \frac{\mathbf{p}(\xi) - \mathbf{p}(0)}{\xi} \right\|_1 &= \left\| (\sin^2(t - 1/\xi))^{1/\xi^2} dt \right\|_1 \\ &= \int_0^\pi (\sin^2(t - 1/\xi))^{1/\xi^2} dt \\ &= \int_0^\pi (\sin^2 t)^{1/\xi^2} dt \xrightarrow{\xi \rightarrow 0} 0, \end{aligned}$$

which shows the claim. Here, we used the  $\pi$ -periodicity of the integrand for fixed  $\xi$  and dominated convergence in the last step.

Since for a parametrized measure model  $(M, \Omega, \mathbf{p})$  the map  $\mathbf{p}$  is  $C^1$ , it follows that its derivative yields a continuous map between the tangent fibrations

$$d\mathbf{p} : TM \longrightarrow T\mathcal{M}(\Omega) = \bigcup_{\mu \in \mathcal{M}(\Omega)} \mathcal{S}(\Omega, \mu).$$

That is, for each tangent vector  $V \in T_\xi M$ , its differential  $d_\xi \mathbf{p}(V)$  is contained in  $\mathcal{S}(\Omega, \mathbf{p}(\xi))$  and hence dominated by  $\mathbf{p}(\xi)$ .

**Definition 4.3.** Let  $(M, \Omega, \mathbf{p})$  be a parametrized measure model. Then for each tangent vector  $V \in T_\xi M$  of  $M$ , we define

$$\partial_V \log \mathbf{p}(\xi) := \frac{d\{d_\xi \mathbf{p}(V)\}}{d\mathbf{p}(\xi)} \in L^1(\Omega, \mathbf{p}(\xi)) \tag{4.2}$$

and call this the *logarithmic derivative of  $\mathbf{p}$  at  $\xi$  in direction  $V$* .

If such a model is dominated by  $\mu_0$  and has a regular density function  $p$  for which (4.1) holds, then we can calculate the Radon–Nikodym derivative as

$$\begin{aligned} \frac{d\{d_\xi \mathbf{p}(V)\}}{d\mathbf{p}(\xi)} &= \frac{d\{d_\xi \mathbf{p}(V)\}}{d\mu_0} \cdot \left(\frac{d\mathbf{p}(\xi)}{d\mu_0}\right)^{-1} \\ &= \partial_V p(\cdot; \xi) (p(\cdot; \xi))^{-1} = \partial_V \log p(\cdot; \xi), \end{aligned}$$

where we use the convention  $\log 0 = 0$ . This justifies the notation in (4.2) even for models without a regular density function.

For a parametrized measure model  $(M, \Omega, \mathbf{p})$  and  $k > 1$  we consider the map

$$\mathbf{p}^{1/k} := \pi^{1/k} \circ \mathbf{p} : M \longrightarrow \mathcal{S}^{1/k}(\Omega), \quad \xi \longmapsto \mathbf{p}(\xi)^{1/k}.$$

Since  $\pi^{1/k}$  is continuous by Proposition 2.2, it follows that  $\mathbf{p}^{1/k}$  is continuous as well. Let us pretend for the moment that  $\mathbf{p}^{1/k}$  is a  $C^1$ -map, so that  $d_\xi \mathbf{p}^{1/k}(V) \in T_{\mathbf{p}(\xi)^{1/k}} \mathcal{M}^{1/k}(\Omega) = \mathcal{S}^{1/k}(\Omega, \mathbf{p}(\xi))$ . In this case, because of  $\pi^k \circ \pi^{1/k} = \text{Id}$ , we have

$$\mathbf{p} = \pi^k \circ \mathbf{p}^{1/k},$$

whence by the chain rule and (2.13) we have for  $\xi \in M$  and  $V \in T_\xi M$

$$d_\xi \mathbf{p}(V) = k\mathbf{p}(\xi)^{1-1/k} \cdot (d_\xi \mathbf{p}^{1/k}(V)).$$

Thus with (4.2) this implies

$$d_\xi \mathbf{p}^{1/k}(V) = \frac{1}{k} \partial_V \log \mathbf{p}(\xi) \mathbf{p}^{1/k}(\xi) \in \mathcal{S}^{1/k}(\Omega, \mathbf{p}(\xi)) \tag{4.3}$$

and hence, in particular,  $\partial_V \log \mathbf{p}(\xi) \in L^k(\Omega, \mathbf{p}(\xi))$ , and depends continuously on  $V \in TM$ . This motivates the following definition.

**Definition 4.4** (*k-integrable parametrized measure model*).

A parametrized measure model  $(M, \Omega, \mathbf{p})$  is called *k-integrable* for  $k \geq 1$  if for all  $\xi \in M$  and  $V \in T_\xi M$  we have

$$\partial_V \log \mathbf{p}(\xi) = \frac{d\{d_\xi \mathbf{p}(V)\}}{d\mathbf{p}(\xi)} \in L^k(\Omega, \mathbf{p}(\xi)),$$

and moreover, the map

$$d\mathbf{p}^{1/k} : TM \longrightarrow T\mathcal{S}^{1/k}(\Omega)$$

given in (4.3) is continuous. Furthermore, we call the model *∞-integrable* if it is *k-integrable* for all  $k \geq 1$ .

Since  $\mathbf{p}(\xi)$  is a finite measure, we have  $L^k(\Omega, \mathbf{p}(\xi)) \subset L^l(\Omega, \mathbf{p}(\xi))$  for all  $1 \leq l \leq k$ . Thus, *k-integrability* implies *l-integrability* for all such *l*.

**Remark 4.3.** The reader who is familiar with the definition of *k-integrability* in [5], Definition 2.4, might notice that there only the continuity of the norm  $\|d_\xi \mathbf{p}^{1/k}\|$  on  $TM$  is required. Also, by our previous discussion, a parametrized measure model  $(M, \Omega, \mathbf{p})$  for which  $\mathbf{p}^{1/k}$  is a  $C^1$ -map is always *k-integrable*.

As it turns out, these three notions of *k-integrability* are equivalent. That is,  $\|d_\xi \mathbf{p}^{1/k}\|$  is continuous if and only if  $d_\xi \mathbf{p}^{1/k}$  is continuous if and only if  $\mathbf{p}^{1/k}$  is a  $C^1$ -map.

**Definition 4.5 (Canonical n-tensor).** For  $n \in \mathbb{N}$ , the *canonical n-tensor* is the covariant *n-tensor* on  $\mathcal{S}^{1/n}(\Omega)$ , given by

$$L_\Omega^n(v_1, \dots, v_n) = n^n \int_\Omega d(v_1 \cdots v_n), \quad \text{where } v_i \in \mathcal{S}^{1/n}(\Omega). \tag{4.4}$$

The main purpose of defining the notion of *k-integrability* is that for a *k-integrable* model, we can for any  $n \leq k$  define the pullback

$$\begin{aligned} \tau_{(M, \Omega, \mathbf{p})}^n &:= (\mathbf{p}^{1/n})^* L_\Omega^n, \\ \tau_{(M, \Omega, \mathbf{p})}^n(V_1, \dots, V_n) &= L_\Omega^n(d_\xi \mathbf{p}^{1/n}(V_1), \dots, d_\xi \mathbf{p}^{1/n}(V_n)) \\ &= \int_\Omega \partial_{V_1} \log \mathbf{p}(\xi) \cdots \partial_{V_n} \log \mathbf{p}(\xi) d\mathbf{p}(\xi), \end{aligned} \tag{4.5}$$

where the second line follows immediately from (4.3) and (4.4). This is well defined as  $\mathbf{p}^{1/n} : M \rightarrow \mathcal{S}^{1/n}(\Omega)$  is differentiable by Remark 4.3

**Example 4.2.**

(1) For  $n = 1$ , the canonical 1-form is given as

$$\tau_{(M, \Omega, \mathbf{p})}^1(V) := \int_\Omega \partial_V \log \mathbf{p}(\xi) d\mathbf{p}(\xi) = \partial_V \|\mathbf{p}(\xi)\|.$$

Thus, it vanishes if and only if  $\|\mathbf{p}(\xi)\|$  is locally constant, e.g., if  $(M, \Omega, \mathbf{p})$  is a *statistical model*.

(2) For  $n = 2$ ,  $\tau_{(M, \Omega, \mathbf{p})}^2$  coincides with the *Fisher metric*

$$\mathbf{g}^F(V, W)_\xi := \int_{\Omega} \partial_V \log \mathbf{p}(\xi) \partial_W \log \mathbf{p}(\xi) d\mathbf{p}(\xi) \tag{4.6}$$

(3) For  $n = 3$ ,  $\tau_{(M, \Omega, \mathbf{p})}^3$  coincides with the *Amari–Chentsov 3-symmetric tensor*

$$T^{\text{AC}}(V, W, X)_\xi := \int_{\Omega} \partial_V \log \mathbf{p}(\xi) \partial_W \log \mathbf{p}(\xi) \partial_X \log \mathbf{p}(\xi) d\mathbf{p}(\xi).$$

Observe that the Fisher metric  $\mathbf{g}^F$  is a Riemannian metric on  $M$  iff  $\mathbf{p}$  is an immersion, i.e., if  $\ker d_\xi \mathbf{p} = 0$ .

**Remark 4.4.** While the Fisher metric and the Amari–Chentsov tensor give an interpretation of  $\tau_{(M, \Omega, \mathbf{p})}^n$  for  $n = 2, 3$ , we do not know of any statistical significance of  $\tau_{(M, \Omega, \mathbf{p})}^n$  for  $n \geq 4$ . However, we shall show later that  $\tau_M^{2n}$  can be used to measure the information loss of statistics and Markov kernels, cf. Theorem 5.2. Moreover, in [18], p. 212, the question is posed if there are other significant tensors on statistical manifolds, and the canonical  $n$ -tensors may be considered as natural candidates.

## 5. Parametrized measure models and sufficient statistics

Given a parametrized measure model (statistical model, respectively)  $(M, \Omega, \mathbf{p})$  and a Markov kernel  $K : \Omega \rightarrow \mathcal{P}(\Omega')$  which induces the Markov morphism  $K_* : \mathcal{M}(\Omega) \rightarrow \mathcal{M}(\Omega')$  as in (3.10), we obtain another parametrized measure model (statistical model, respectively)  $(M, \Omega', \mathbf{p}')$  by defining  $\mathbf{p}'(\xi) := K_* \mathbf{p}(\xi)$ . These transitions can be interpreted as data processing in statistical decision theory, which can be deterministic (i.e., given by a statistic) or randomized (i.e., given by a Markov kernel). We refer to for example, [11] where this is elaborated in detail.

It is the purpose of this section to investigate the relation between these two models in more detail.

**Theorem 5.1.** *Let  $(M, \Omega, \mathbf{p})$ ,  $K : \Omega \rightarrow \mathcal{P}(\Omega')$  and  $(M, \Omega', \mathbf{p}')$  be as above, and suppose that  $(M, \Omega, \mathbf{p})$  is  $k$ -integrable for some  $k \geq 1$ . Then  $(M, \Omega', \mathbf{p}')$  is also  $k$ -integrable, and*

$$\|\partial_V \log \mathbf{p}'(\xi)\|_k \leq \|\partial_V \log \mathbf{p}(\xi)\|_k \quad \text{for all } V \in T_\xi M, \tag{5.1}$$

where the norms are taken in  $L^k(\Omega, \mathbf{p}(\xi))$  and  $L^k(\Omega', \mathbf{p}'(\xi))$ , respectively. If  $K$  is congruent, then equality in (5.1) holds for all  $V$ .

Moreover, if  $K$  is given by a statistic  $\kappa : \Omega \rightarrow \Omega'$  and  $k > 1$ , then equality in (5.1) holds iff  $\partial_V \log \mathbf{p}(\xi) = \kappa^*(\partial_V \log \mathbf{p}'(\xi))$ . In particular, equality in (5.1) either holds for all  $k > 1$  for which the model is  $k$ -integrable or for no such  $k > 1$ .

**Proof.** Since  $K_*$  is the restriction of a bounded linear map, it is obvious that  $\mathbf{p}' : M \rightarrow \mathcal{M}(\Omega')$  is again differentiable, and in fact,

$$d_\xi \mathbf{p}'(V) = K_*(d_\xi \mathbf{p}(V)) \tag{5.2}$$

for all  $V \in T_\xi M$ ,  $\xi \in M$ .

Let  $\mu := \mathbf{p}(\xi)$  and  $\mu' := \mathbf{p}'(\xi) = K_*\mu$ , and let  $\phi := \partial_V \log \mathbf{p}(\xi)$  and  $\phi' := \partial_V \log \mathbf{p}'(\xi)$ , so that  $d_\xi \mathbf{p}(V) = \phi\mu$  and  $d_\xi \mathbf{p}'(V) = \phi'\mu'$ . By (5.2) we thus have

$$K_*(\phi\mu) = \phi'\mu',$$

so that  $\phi' = K_*^\mu(\phi)$  is the expectation value of  $\phi$  given  $K$ . If  $\mathbf{p}$  is  $k$ -integrable, then  $\phi = \partial_V \log \mathbf{p}(\xi) \in L^k(\Omega, \mu)$ , whence  $\phi' \in L^k(\Omega', \mu')$ , and  $\|\phi'\|_k \leq \|\phi\|_k$ , by Theorem 3.3. That is,  $\mathbf{p}'$  is  $k$ -integrable as well and (5.1) holds.

If  $K$  is congruent, then  $\|\phi'\|_k = \|\phi\|_k$  by Proposition 3.3.

If  $k > 1$  and  $K$  is given by a statistic  $\kappa$ , then equality in (5.1) occurs iff  $\phi = \kappa^*\phi'$  by Proposition 3.4. □

Since the Fisher metrics  $\mathfrak{g}^F$  of  $(M, \Omega, \mathbf{p})$  and  $\mathfrak{g}'^F$  of  $(M, \Omega', \mathbf{p}')$  are defined as

$$\mathfrak{g}(V, V) = \|\partial_V \log \mathbf{p}(\xi)\|_2^2 \quad \text{and} \quad \mathfrak{g}'(V, V) = \|\partial_V \log \mathbf{p}'(\xi)\|_2^2$$

by (4.6), Theorem 5.1 immediately implies the following.

**Theorem 5.2 (Monotonicity theorem, cf. [3,5,28]).** *Let  $(M, \Omega, \mathbf{p})$  be a  $k$ -integrable parametrized measure model for  $k \geq 2$ , let  $K : \Omega \rightarrow \mathcal{P}(\Omega')$  be a Markov kernel, and let  $(M, \Omega', \mathbf{p}')$  be given by  $\mathbf{p}'(\xi) = K_*\mathbf{p}(\xi)$ . Then*

$$\mathfrak{g}(V, V) \geq \mathfrak{g}'(V, V) \quad \text{for all } V \in T_\xi M \text{ and } \xi \in M. \tag{5.3}$$

**Remark 5.1.** Note that our approach allows to prove the Monotonicity Theorem 5.2 with no further assumption on the model  $(M, \Omega, \mathbf{p})$ . In order for (5.3) to hold we can work with arbitrary Markov kernels, not just statistics  $\kappa$ . Even if  $K$  is given by a statistic  $\kappa$ , we do not need to assume that  $\Omega$  is a topological space with its Borel  $\sigma$ -algebra as in [19], Theorem 1.2, nor do we need to assume the existence of transversal measures of the map  $\kappa$  (e.g. [3], Theorem 2.1), nor do we need to assume that all measures  $\mathbf{p}(\xi)$  have the same null sets ([5], Theorem 3.11). In this sense, our statement generalizes these versions of the monotonicity theorem, as it even covers a rather peculiar statistic as in Example 3.2.

In [3], p. 98, the difference

$$\mathfrak{g}(V, V) - \mathfrak{g}'(V, V) \geq 0 \tag{5.4}$$

is called the *information loss* of the model under the statistic  $\kappa$ , a notion which is highly relevant for statistical inference. This motivates the following definition.

**Definition 5.1.** Let  $(M, \Omega, \mathbf{p})$  be  $k$ -integrable for some  $k > 1$ , let  $K : \Omega \rightarrow \mathcal{P}(\Omega')$  and  $(M, \Omega', \mathbf{p}')$  be as above, so that  $(M, \Omega', \mathbf{p}')$  is  $k$ -integrable as well. Then for each  $V \in T_\xi M$  we define the  $k$ th order information loss under  $K$  in direction  $V$  as

$$\|\partial_V \log \mathbf{p}(\xi)\|_k^k - \|\partial_V \log \mathbf{p}'(\xi)\|_k^k \geq 0,$$

where the norms are taken in  $L^k(\Omega, \mathbf{p}(\xi))$  and  $L^k(\Omega', \mathbf{p}'(\xi))$ , respectively.

That is, the information loss in (5.4) is simply the special case  $k = 2$  in Definition 5.1. Observe that due to Theorem 5.1 the vanishing of the information loss for *some*  $k > 1$  implies the vanishing for *all*  $k > 1$  for which this norm is defined. That is, the  $k$ th order information loss measures the same quantity by different means.

For instance, if  $(M, \Omega, \mathbf{p})$  is  $k$ -integrable for  $1 < k < 2$ , but not 2-integrable, then the Fisher metric and hence the classical information loss in (5.4) is not defined. Nevertheless, we still can quantify the  $k$ th order information loss of a statistic of this model.

Observe that for  $k = 2n$  an even integer  $\tau_{(M, \Omega, \mathbf{p})}^{2n}(V, \dots, V) = \|\partial_V \log \mathbf{p}(\xi)\|_{2n}^{2n}$ , whence the difference

$$\tau_{(M, \Omega, \mathbf{p})}^{2n}(V, \dots, V) - \tau_{(M, \Omega', \mathbf{p}')}^{2n}(V, \dots, V) \geq 0$$

represents the  $2n$ th order information loss of  $\kappa$  in direction  $V$ . This gives an interpretation of the canonical  $2n$ -tensors  $\tau_{(M, \Omega, \mathbf{p})}^{2n}$ .

It is a natural problem to characterize statistics of a model which do not produce any information loss. Fisher [14] called such a statistic *sufficient* writing that "... the criterion of sufficiency, which latter requires that the whole of the relevant information supplied by a sample shall be contained in the statistics calculated" [14], p. 367. This motivates the following definition.

**Definition 5.2 (Sufficient statistic).** Let  $(M, \Omega, \mathbf{p})$  be a parametrized measure model which is  $k$ -integrable for some  $k > 1$ . Then a statistic  $\kappa : \Omega \rightarrow \Omega'$  or, more general, a Markov kernel  $K : \Omega \rightarrow \mathcal{P}(\Omega')$  is called a *sufficient for the model* if the  $k$ th order information loss vanishes for all tangent vectors  $V$ , that is, if

$$\|\partial_V \log \mathbf{p}'(\xi)\|_k = \|\partial_V \log \mathbf{p}(\xi)\|_k \quad \text{for all } V \in T_\xi M,$$

where  $\mathbf{p}'(\xi) = \kappa_* \mathbf{p}(\xi)$  or  $\mathbf{p}'(\xi) = K_* \mathbf{p}(\xi)$ , respectively.

Again, in this definition it is irrelevant which  $k > 1$  is used, as long as  $k$ -integrability of the model is satisfied.

**Example 5.1 (Fisher–Neyman [24]).** Let  $(M, \Omega', \mu'_0, \mathbf{p}')$  be a parametrized measure model dominated by  $\mu'_0$ , given by

$$\mathbf{p}'(\xi) = \phi'(\cdot; \xi) \mu'_0, \quad \phi' : \Omega \times M \longrightarrow [0, \infty].$$

Moreover, let  $\kappa : \Omega \rightarrow \Omega'$  be a statistic and  $\mu_0 \in \mathcal{M}(\Omega)$  such that  $\kappa_*(\mu_0) = \mu'_0$ . Define the parametrized measure model  $(M, \Omega, \mu_0, \mathbf{p})$  as

$$\mathbf{p}(\xi) := \phi'(\kappa(\cdot), \xi) \mu_0. \tag{5.5}$$

Then  $\kappa$  is a sufficient statistic for  $(M, \Omega, \mu_0, \mathbf{p})$ . Indeed,  $\kappa_*(\mathbf{p})(\xi) = \mathbf{p}'(\xi)$  by (3.7), and  $d_\xi \mathbf{p}(V) = \kappa^*(d\mathbf{p}'_\xi(V))$  for all  $V \in T_\xi M$ , so that  $\partial_V \log \mathbf{p}(\xi) = \kappa^*(\partial_V \log \mathbf{p}'(V))$ . By Theorem 5.1 it follows that equality holds in (5.1), so that  $\kappa$  is a sufficient statistic for  $(M, \Omega, \mu_0, \mathbf{p})$ .

Under some further assumptions, the statistics given in Example 5.1 exhaust all sufficient statistics. More precisely, the following is known as the Fisher–Neyman factorization.

**Proposition 5.1 ([24]).** *Let  $(M, \Omega, \mu, \mathbf{p})$  be a parametrized measure model with a positive regular density function  $p : \Omega \times M \rightarrow (0, \infty)$ , and let  $\kappa : \Omega \rightarrow \Omega'$  be a sufficient statistic of the model.*

*Then  $(M, \Omega, \mu, \mathbf{p})$  admits a Fisher–Neyman factorization, that is, it is of the form (5.5) in Example 5.1 for some measure  $\mu_0$  on  $\Omega$ ,  $\mu'_0 := \kappa_*(\mu_0)$  and some function  $\phi' : \Omega' \times M \rightarrow (0, \infty)$ .*

**Proof.** If  $\mathbf{p}(\xi) = p(\cdot; \xi)\mu$  where  $p$  is positive and differentiable in the  $\xi$ -variable, then  $\log p(\cdot; \xi)$  and  $\log p'(\cdot; \xi)$  are well defined functions on  $\Omega \times M$  and  $\Omega' \times M$ , respectively and differentiable in  $\xi$ . In particular,  $\kappa^*(\partial_V \log p'(\cdot; \xi)) = \partial_V (\log p'(\kappa(\cdot); \xi))$ , so that by Theorem 5.1 equality in (5.1) holds for  $k > 1$  iff

$$\partial_V \log \frac{p(\cdot; \xi)}{p'(\kappa(\cdot); \xi)} = \partial_V (\log p(\cdot; \xi) - (\log p'(\kappa(\cdot); \xi))) = 0.$$

If  $M$  is connected, then this is the case for all  $V \in TM$  iff the positive function  $h(\cdot) := \frac{p(\cdot; \xi)}{p'(\kappa(\cdot); \xi)}$  does not depend on  $\xi \in M$ . Thus, setting  $\mu_0 := h\mu$  implies (5.5), showing the assertion.  $\square$

Observe that the proof uses the positivity of the density function  $p$  in a crucial way. In fact, without this assumption the conclusion is false, as the following example shows.

**Example 5.2.** Let  $\Omega := (-1, 1) \times (0, 1)$ ,  $\Omega' := (-1, 1)$  and  $\kappa : \Omega \rightarrow \Omega'$  be the projection onto the first component. For  $\xi \in \mathbb{R}$  we define the statistical model  $\mathbf{p}$  on  $\Omega$  as  $\mathbf{p}(\xi) := p(s, t; \xi) ds dt$ , where

$$p(s, t; \xi) := \begin{cases} h(\xi) & \text{for } \xi \geq 0 \text{ and } s \geq 0, \\ 2h(\xi)t & \text{for } \xi < 0 \text{ and } s \geq 0, \\ 1 - h(\xi) & \text{for } s < 0, \end{cases}$$

with  $h(\xi) := \exp(-|\xi|^{-1})$  for  $\xi \neq 0$  and  $h(0) := 0$ . Then  $\mathbf{p}(\xi)$  is a probability measure, and

$$\mathbf{p}'(\xi) := \kappa_* \mathbf{p}(\xi) = p'(s; \xi) ds \quad \text{with } p'(s; \xi) := (1 - h(\xi))\chi_{(-1,0)}(s) + h(\xi)\chi_{[0,1)}(s),$$

and thus,

$$\|\partial_\xi \log p(s, t; \xi)\|_k = \|\partial_\xi \log p'(s; \xi)\|_k = k \left( \left| \frac{d}{d\xi} h(\xi)^{1/k} \right|^k + \left| \frac{d}{d\xi} (1 - h(\xi))^{1/k} \right|^k \right)^{1/k},$$

where the norm is taken in  $L^k(\Omega, \mathbf{p}(\xi))$  and  $L^k(\Omega', \mathbf{p}'(\xi))$ , respectively. Since this expression is continuous in  $\xi$  for all  $k$ , the models  $(\mathbb{R}, \Omega, \mathbf{p})$  and  $(\mathbb{R}, \Omega', \mathbf{p}')$  are  $\infty$ -integrable, and there is no information loss of  $k$ th order for any  $k \geq 1$ , so that  $\kappa$  is a sufficient statistic of the model in the sense of Definition 5.2. Thus,  $\kappa$  is a sufficient statistic for the model.

Indeed, this model admits a Fisher–Neyman factorization when restricted to  $\xi \geq 0$  and to  $\xi \leq 0$ , respectively; in these cases, we have

$$\mathbf{p}(\xi) = p'(s; \xi)\mu_{\pm},$$

with the measures  $\mu_+ := ds dt$  for  $\xi \geq 0$  and  $\mu_- := (\chi_{(-1,0)}(s) + 2t\chi_{[0,1)}(s)) ds dt$  for  $\xi \leq 0$ , respectively.

However, since  $\mu_+ \neq \mu_-$ ,  $\kappa$  is *not* of the form (5.5) and hence *not among the sufficient statistics given in Example 5.1* when defining it for all  $\xi \in \mathbb{R}$ . This does not contradict Proposition 5.1 since  $p(s, t; \xi)$  is not positive a.e. for  $\xi = 0$ .

The reader might be aware that most texts use the description (5.5) in Example 5.1 as a definition for a sufficient statistic, for example, [14], [5], Definition 3.1, [3], (2.17), [8], Theorem 1, p. 117. In the light of the Fisher–Neyman factorization in Proposition 5.1, this is equivalent to our Definition 5.2 under the assumption that the model is given by a regular positive density function, an assumption that has been made in all these references.

However, the significance of Example 5.2 is that the two notions of sufficiency are no longer equivalent if the assumption of positivity of the density function is dropped. But since in this example, all statistical information of  $(M, \Omega, \mathbf{p})$  can be recovered from  $(M, \Omega', \mathbf{p}')$ , it seems natural to define sufficiency of a statistic in such a way that this example is subsumed, that is, as in Definition 5.2.

## Acknowledgements

This work was mainly carried out at the Max Planck Institute for Mathematics in the Sciences in Leipzig, and we are grateful for the excellent working conditions provided at that institution. H.V. Lê is partially supported by Grant RVO: 67985840. J. Jost acknowledges support from the ERC Advanced Grant FP7-267087. We also thank the referees for numerous helpful comments and suggestions which helped to significantly improve the manuscript.

## References

- [1] Amari, S. (1980). Theory of information space: A differential-geometrical foundation of statistics. *RAAG Res. Notes* **106** 1–54. [MR0567621](#)
- [2] Amari, S. (1982). Differential geometry of curved exponential families – curvatures and information loss. *Ann. Statist.* **10** 357–385. [MR0653513](#)
- [3] Amari, S. and Nagaoka, H. (2000). *Methods of Information Geometry. Translations of Mathematical Monographs* **191**. Providence, RI: Amer. Math. Soc.; Oxford: Oxford Univ. Press. Translated from the 1993 Japanese original by Daishi Harada. [MR1800071](#)



- [4] Amari, S.-I., Barndorff-Nielsen, O.E., Kass, R.E., Lauritzen, S.L. and Rao, C.R. (1987). *Differential Geometry in Statistical Inference. Institute of Mathematical Statistics Lecture Notes – Monograph Series* **10**. Hayward, CA: IMS. [MR0932246](#)
- [5] Ay, N., Jost, J., Lê, H.V. and Schwachhöfer, L. (2015). Information geometry and sufficient statistics. *Probab. Theory Related Fields* **162** 327–364. [MR3350047](#)
- [6] Bauer, H. (2001). *Measure and Integration Theory. De Gruyter Studies in Mathematics* **26**. Berlin: de Gruyter. Translated from the German by Robert B. Burckel. [MR1897176](#)
- [7] Bauer, M., Bruveris, M. and Michor, P.W. (2016). Uniqueness of the Fisher–Rao metric on the space of smooth densities. *Bull. Lond. Math. Soc.* **48** 499–506. [MR3509909](#)
- [8] Borovkov, A.A. (1998). *Mathematical Statistics*. Amsterdam: Gordon and Breach Science Publishers. Translated from the Russian by A. Moullagaliev and revised by the author. [MR1712750](#)
- [9] Cena, A. and Pistone, G. (2007). Exponential statistical model. *Ann. Inst. Statist. Math.* **59** 27–56.
- [10] Chentsov, N. (1965). Category of mathematical statistics. *Dokl. Acad. Nauk USSR* **164** 511–514.
- [11] Chentsov, N. (1972). *Statistical Decision Rules and Optimal Inference*. Moscow: Nauka (in Russian). English translation in: Translation of Math. Monograph 53, AMS, Providence, 1982.
- [12] Chentsov, N. (1978). Algebraic foundation of mathematical statistics. *Math. Operforsch. Stat., Ser. Stat.* **9** 267–276. [MR0512264](#)
- [13] Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* **3** 1189–1242. With a discussion by by C.R. Rao, Don A. Pierce, D.R. Cox, D.V. Lindley, Lucien LeCam, J.K. Ghosh, J. Pfanzagl, Niels Keiding, A.P. Dawid, Jim Reeds and with a reply by the author.
- [14] Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. Lond. Ser. A* **222** 309–368.
- [15] Guillemin, V. and Sternberg, S. (1977). *Geometric Asymptotics. Mathematical Surveys* **14**. Providence, RI: Amer. Math. Soc. [MR0516965](#)
- [16] Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. Ser. A* **186** 453–461. [MR0017504](#)
- [17] Lang, S. (2002). *Introduction to Differentiable Manifolds*, 2nd ed. *Universitext*. New York: Springer. [MR1931083](#)
- [18] Lauritzen, S. (1987). Statistical manifolds. In *Differential Geometry in Statistical Inference. Institute of Mathematical Statistics, Lecture Note – Monograph Series* **10**. Hayward, CA: IMS.
- [19] Lê, H.V. (2017). The uniqueness of the Fisher metric as information metric. *Ann. Inst. Statist. Math.* **69** 879–896. [MR3671642](#)
- [20] Morozova, E. and Chentsov, N. (1991). Natural geometry on families of probability laws. In *Probability theory, 8. Itogi Nauki i Techniki* **83** 133–265. Moscow: VINITI.
- [21] Morse, N. and Sacksteder, R. (1966). Statistical isomorphism. *Ann. Math. Stat.* **37** 203–214.
- [22] Murray, M.K. and Rice, J.W. (1993). *Differential Geometry and Statistics. Monographs on Statistics and Applied Probability* **48**. London: Chapman & Hall. [MR1293124](#)
- [23] Neveu, J. (1965). *Mathematical Foundations of the Calculus of Probability. Translated by Amiel Feinstein*. San Francisco, CA–London–Amsterdam: Holden–Day, Inc. [MR0198505](#)
- [24] Neyman, J. (1935). Sur un teorema concernente le cosiddette statistiche sufficienti. *G. Ist. Ital. Attuari* **6** 320–334.
- [25] Pistone, G. (2013). Nonparametric information geometry. In *Geometric Science of Information. Lecture Notes in Computer Science* **8085** 5–36. Heidelberg: Springer. [MR3126029](#)
- [26] Pistone, G. and Sempi, C. (1995). An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Ann. Statist.* **23** 1543–1561. [MR1370295](#)

- [27] Rao, C.R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **37** 81–91. [MR0015748](#)
- [28] Schwachhöfer, L., Ay, N., Jost, J. and Lê, H.V. (2015). Invariant geometric structures on statistical models. In *Geometric Science of Information. Lecture Notes in Computer Science* **9389** 150–158. Cham: Springer. [MR3442195](#)

*Received December 2015 and revised September 2016*