Chapter 4
The Intrinsic Geometry of Statistical Models

4.1 Extrinsic Versus Intrinsic Geometric Structures

In geometry, an extrinsic and intrinsic perspective can be distinguished. Differential geometry started as the geometry of curves and surfaces in three-dimensional Euclidean space. Of course, this can be generalized to higher dimensions and codimensions, but Gauss had a deeper insight [102]. His *Theorema Egregium* says that the curvature of a surface, now referred to as Gauss curvature, only depends on intrinsic measurements within the surface and does not refer to the particular way the surface is embedded in the ambient three-dimensional Euclidean space. Although this theorem explicitly refers to the curvature of surfaces, it highlights a more general paradigm of geometry. In fact, Riemann developed a systematic approach to geometry, now called Riemannian geometry, that treats geometric quantities intrinsically, as metric structures on manifolds without referring to any embedding into some Euclidean space [225]. Nevertheless, geometry can also be developed extrinsically, because Nash's embedding theorem [196] tells us that any Riemannian manifold can be isometrically embedded into some Euclidean space, that is, it can be realized as a submanifold of some Euclidean space, and its intrinsic Riemannian metric then coincides with the restriction of the extrinsic Euclidean metric to that submanifold. It is therefore a matter of convenience whether differential geometry is developed intrinsically or extrinsically. In most cases, after all, the intrinsic view is the more convenient and transparent one. (In algebraic geometry, the situation is somewhat similar, although more complicated. A projective variety is a subvariety of some complex projective space, and inherits the latter's structures; in particular, we can restrict the Fubini–Study metric to a projective variety, and when the latter is smooth, it thus becomes a Riemannian manifold. On the other hand, there is the abstract notion of an algebraic variety. In contrast to the differential geometric situation, however, not every abstract algebraic variety can be embedded into some complex projective space, that is, realized as a projective variety. Nevertheless, many algebraic varieties can be so embedded, and for those thus also both an intrinsic and an extrinsic perspective are possible.)

N. Ay et al., *Information Geometry*, Ergebnisse der Mathematik und ihrer Grenzgebiete.
3. Folge / A Series of Modern Surveys in Mathematics 64,
DOI 10.1007/978-3-319-56478-4_4

Nevertheless, the two approaches are not completely equivalent. This stems from the fact that the isometric embedding produced by Nash's theorem is in general not unique. Submanifolds of a Euclidean space need not be rigid, that is, there may exist different isometric embeddings of the same manifold into the same ambient Euclidean space that cannot be transformed into each other by a Euclidean isometry of that ambient space. In many cases, submanifolds can even be continuously deformed in the ambient space while keeping their intrinsic geometry. In general, the rigidity or non-rigidity of submanifolds is a difficult topic that depends not only on the codimension—the higher the codimension, the more room we have for deformations—, but also on intrinsic properties. For instance, closed surfaces of positive curvature in 3-space are rigid, but there exist other closed surfaces that are not rigid. We do not want to enter into this topic here, but simply point out that the non-uniqueness of isometric embeddings means that such an embedding is an additional datum that is not determined by the intrinsic geometry.

Thus, on the one hand, the intrinsic geometry by its very definition does not depend on an isometric embedding. Distances and notions derived from them, like that of a geodesic curve or a totally geodesic submanifold, are intrinsically defined. On the other hand, the shape of the embedded object in the ambient space influences constructions like projections from the exterior onto that object.

Also, the intrinsic and the extrinsic geometry are not equivalent, in the sense that distances measured intrinsically are generally larger than those measured extrinsically, in the ambient space. The only exception occurs for convex subsets of affine linear subspaces of Euclidean spaces, but compact manifolds cannot be isometrically embedded in such a manner. We should point out, however, that there is another, stronger, notion of isometric embedding where intrinsic and extrinsic distances are the same. For that, the target space can no longer be chosen as a finite-dimensional Euclidean space, but it has to be some Banach space. More precisely, every bounded metric space (X, d) can be isometrically embedded, in this stronger sense, into an L^∞-space. We simply associate to every point $x \in X$ the distance function $d(x, \cdot)$,

$$\begin{aligned} X &\to L^\infty(X) \\ x &\mapsto d(x, \cdot). \end{aligned} \tag{4.1}$$

Since by the triangle inequality

$$\big\|d(x_1, \cdot) - d(x_2, \cdot)\big\|_{L^\infty} = \sup_{y \in X} \big|d(x_1, y) - d(x_2, y)\big| = d(x_1, x_2),$$

this embedding does indeed preserve distances.

In information geometry, the situation is somewhat analogous, as we shall now explain. In Chap. 2, we have developed an extrinsic approach, embedding our parameter spaces into spaces of measures and defining the Fisher metric in terms of such embeddings, like the restriction of the Euclidean metric to a submanifold. The same holds for Chap. 3, where, however, we have mainly addressed the complications arising from the fact that our space of measures was of infinite dimension. In

our formal Definition 3.4 of statistical models, we explicitly include the embedding $p : M \to \mathcal{P}_+(\Omega)$ and thereby interpret the points of M as being elements of the ambient set of strictly positive probability measures.[1] This way we can pull back the natural geometric structures on $\mathcal{P}_+(\Omega)$, not only the Fisher metric $\mathfrak{g}$, but also the Amari–Chentsov tensor $\mathbf{T}$, and consider them as natural structures defined on M. The question then arises whether information geometry can also alternatively be developed in an intrinsic manner, analogous to Riemannian geometry. After all, some important aspects of information geometry are completely captured in terms of the structures defined on M. For instance, we have shown that the m-connection and the e-connection on $\mathcal{P}_+(\Omega)$ are dual with respect to the Fisher metric. If we equip M with the pullback g of the Fisher metric and the pullbacks ∇ and ∇^* of the m-and the e-connection, respectively, then it is easy to see that the duality of ∇ and ∇^* with respect to g, which is an intrinsic property, is inherited from the duality of the corresponding objects on the bigger space. This duality already captures important aspects of information geometry, which led Amari and Nagaoka to the definition of a *dualistic structure* (g, ∇, ∇^*) on a manifold M. It turns out that much of the theory presented so far can indeed be derived based on a dualistic structure, without assuming that the metric and the dual connections are distinguished as natural objects such as the Fisher metric and the m- and e-connections. Alternatively, Lauritzen [160] proposed to consider as the basic structure of information geometry a manifold M together with a Riemannian metric g and a 3-symmetric tensor T, where g corresponds to the Fisher metric and T corresponds to the Amari–Chentsov tensor. He referred to such a triple (M, g, T) as a *statistical manifold*, which, compared to a statistical model, ignores the way M is embedded in $\mathcal{P}_+(\Omega)$ and therefore only refers to intrinsic aspects captured by g and T. Note that any torsion-free dualistic structure in the sense of Amari and Nagaoka defines a statistical manifold in terms of $T(A, B, C) := g(\nabla^*_A B - \nabla_A B, C)$. Importantly, we can also go back from this intrinsic approach to an extrinsic one, analogously to Nash's theorem. In Sect. 4.5 we shall derive Lê's theorem which says that any statistical manifold that is compact, possibly with boundary, can be embedded, in a structure preserving way, in some $\mathcal{P}_+(\Omega)$ so that it can be interpreted as a statistical model, even with a finite set Ω of elementary events. Although we cannot expect such an embedding to be naturally distinguished, there is a great advantage of having an intrinsically defined structure similar to the dualistic structure of Amari and Nagaoka or the statistical manifold of Lauritzen. Such a structure is general enough to be applicable to other contexts, not restricted to our context of measures. For instance, quantum information geometry, which is not a subject of this book, can be treated in terms of a dualistic structure. It turns out that the dualistic structure also captures essential information-geometric aspects in many other fields of application [16].

As in the Nash case, the embedding produced by Lê's theorem is not unique in general. Therefore, we also need to consider such a statistical embedding as an additional datum that is not determined by the intrinsic geometry. Therefore, our

[1] Here, for simplicity of the discussion, we restrict attention to strictly positive probability measures, instead of finite signed measures.

notion of a parametrized measure model includes some structure not contained in the notion of a statistical manifold. The question then is to what extent this is relevant for statistics. Clearly, different embeddings, that is, different parametrized measure models based on the same intrinsic statistical manifold, constitute different families of probability measures in parametric statistics. Nevertheless, certain aspects are intrinsic. For instance, whether a submanifold is autoparallel w.r.t. a connection ∇ depends only on that connection, but not on any embedding. Intrinsically, the two connections ∇ and ∇^* play equivalent roles, but extrinsically, of course, exponential and mixture families play different roles. By the choice of our embedding, we can therefore let either ∇ or ∇^* become the exponential connection. When, however, one of them, say ∇, is so chosen then it becomes an intrinsic notion of what an exponential subfamily is. Therefore, even if we embed the same statistical manifold differently into a space of probability measures, that is, let it represent different parametrized families in the sense of parametric statistics, the corresponding notions of exponential subfamily coincide.

We should also point out that the embedding produced in Lê's theorem is different from that on which Definition 3.4 of a signed parametrized measure model depends, because for the latter the target space $\mathcal{S}(\Omega)$ is infinite-dimensional (unless Ω is finite), like the L^∞-space of (4.1). The strength of Lê's theorem derives precisely from the fact that the target space is finite-dimensional, if the statistical manifold is compact (possibly with boundary). In any case, in Lê's theorem, the structure of a statistical manifold is given and the embedding is constructed. In contrast, in Definition 3.4, the space $\mathcal{P}(\Omega)$ is used to impose the structure of a statistical model onto M. The latter situation is also different from that of (4.1), because the embedding into $\mathcal{P}(\Omega)$ is not extrinsically isometric in general, but rather induces a metric (and a pair of connections) on M in the same manner as a submanifold of a Euclidean space inherits a metric from the latter.

In this chapter, we first approach the structures developed in our previous chapters from a more differential-geometric perspective. In fact, with concepts and tools from finite-dimensional Riemannian geometry, presented in Appendix B, we can identify and describe relations among these structures in a very efficient and transparent way. In this regard, the concept of duality plays a dominant role. This is the perspective developed by Amari and Nagaoka [16] which we shall explore in Sects. 4.2 and 4.3. The intrinsic description of information geometry will naturally lead us to the definition of a dualistic structure. This allows us to derive results solely based on that structure. In fact, the situation that will be analyzed in depth is when we not only have dual connections, but when we also have two such distinguished connections that are flat. Manifolds that possess two such dually flat connections have been called affine Kähler by Cheng–Yau [59] and Hessian manifolds by Shima [236]. They are real analogues of Kähler manifolds, and in particular, locally the entire structure is encoded by some strictly convex function. The second derivatives of that function yield the metric. In particular, that function then is determined up to some affine linear term. The third derivatives yield the Christoffel symbols of the metric as well as those of the two dually flat connections. In fact, for one of them, the Christoffel symbols vanish, and it is affine in the coordinates w.r.t. which we

have defined and computed the convex function. By a Legendre transform, we can pass to dual coordinates and a dual convex functions. With respect to those dual coordinates, the Christoffel symbols, now derived from the dual convex function, of the other flat connection vanish, that is, they are affine coordinates for the second connection. Also, the second derivatives of the dual convex function yield the inverse of the metric. This structure can also be seen as a generalization of the basic situation of statistical mechanics where one of our convex functions becomes the free energy and the other the (negative of the) entropy, see [38]. We'll briefly return to that aspect in Sect. 4.5 below.

Finally, we return to the general case of a dualistic structure that is not necessarily flat and address the question to what extent such a dualistic structure, more precisely a statistical manifold, is different from our previous notion of a statistical model. This question will be answered by Lê's embedding theorem, whose proof we shall present.

4.2 Connections and the Amari–Chentsov Structure

Let now $\mathbf{p}(\xi)$ be a d-dimensional smooth family of probability measures depending on the parameter ξ. The base measure will not play an important role, and so we shall simply write integration as $\int dx$ in this chapter. The family $\mathbf{p}(\xi)$ then has to define a statistical model in the sense of Definition 3.4 with a logarithmic derivative in the sense of Definition 3.6, and it has to be 2-integrable when we compute the Fisher metric and 3-integrable for the Amari–Chentsov tensor, in the sense of Definition 3.7. The latter simply means that the integrals underlying the expectation values in (4.2) and (4.3) below exist; see, for instance, (4.6). We shall henceforth assume that.

For the Fisher metric (3.41), we then have

$$\begin{aligned} g_{ij}(\xi) &= \mathbb{E}_{\mathbf{p}(\xi)}\left(\frac{\partial}{\partial\xi^i}\log p(\cdot;\xi)\frac{\partial}{\partial\xi^j}\log p(\cdot;\xi)\right) \\ &= \int_\Omega \frac{\partial}{\partial\xi^i}\log p(x;\xi)\frac{\partial}{\partial\xi^j}\log p(x;\xi)\; p(x;\xi)\,dx, \end{aligned} \tag{4.2}$$

and so

$$\begin{aligned} \frac{\partial}{\partial\xi^k}g_{ij}(\xi) &= \mathbb{E}_{\mathbf{p}}\left(\frac{\partial}{\partial\xi^k}\frac{\partial}{\partial\xi^i}\log p\,\frac{\partial}{\partial\xi^j}\log p\right) \\ &\quad + \mathbb{E}_{\mathbf{p}}\left(\frac{\partial}{\partial\xi^i}\log p\,\frac{\partial}{\partial\xi^k}\frac{\partial}{\partial\xi^j}\log p\right) \\ &\quad + \mathbb{E}_{\mathbf{p}}\left(\frac{\partial}{\partial\xi^i}\log p\,\frac{\partial}{\partial\xi^j}\log p\,\frac{\partial}{\partial\xi^k}\log p\right). \end{aligned} \tag{4.3}$$

Therefore, by (B.45),

$$\Gamma^{(0)}_{ijk} = \mathbb{E}_{\mathbf{p}}\left(\frac{\partial^2}{\partial\xi^i\partial\xi^j}\log p\,\frac{\partial}{\partial\xi^k}\log p + \frac{1}{2}\frac{\partial}{\partial\xi^i}\log p\,\frac{\partial}{\partial\xi^j}\log p\,\frac{\partial}{\partial\xi^k}\log p\right) \tag{4.4}$$

yields the Levi-Civita connection $\nabla^{(0)}$ for the Fisher metric. More generally, we can define a family $\nabla^{(\alpha)}$, $-1 \le \alpha \le 1$, of connections via

$$\begin{aligned}\Gamma^{(\alpha)}_{ijk} &= \mathbb{E}_{\mathbf{p}}\left(\frac{\partial^2}{\partial\xi^i\partial\xi^j}\log p\,\frac{\partial}{\partial\xi^k}\log p + \frac{1-\alpha}{2}\frac{\partial}{\partial\xi^i}\log p\,\frac{\partial}{\partial\xi^j}\log p\,\frac{\partial}{\partial\xi^k}\log p\right)\\ &= \Gamma^{(0)}_{ijk} - \frac{\alpha}{2}\mathbb{E}_{\mathbf{p}}\left(\frac{\partial}{\partial\xi^i}\log p\,\frac{\partial}{\partial\xi^j}\log p\,\frac{\partial}{\partial\xi^k}\log p\right).\end{aligned} \tag{4.5}$$

We also recall the *Amari–Chentsov tensor* (3.42)

$$\begin{aligned}&\mathbb{E}_{\mathbf{p}}\left(\frac{\partial}{\partial\xi^i}\log p\,\frac{\partial}{\partial\xi^j}\log p\,\frac{\partial}{\partial\xi^k}\log p\right)\\ &\quad= \int_\Omega \frac{\partial}{\partial\xi^i}\log p(x;\xi)\frac{\partial}{\partial\xi^j}\log p(x;\xi)\frac{\partial}{\partial\xi^k}\log p(x;\xi)\;p(x;\xi)\,dx.\end{aligned} \tag{4.6}$$

We note the analogy between the Fisher metric tensor (4.2) and the Amari–Chentsov tensor (4.6). The family $\nabla^{(\alpha)}$ of connections thus is determined by a combination of first derivatives of the Fisher tensor and the Amari–Chentsov tensor.

Lemma 4.1 *All the connections $\nabla^{(\alpha)}$ are torsion-free.*

Proof A connection is torsion-free iff its Christoffel symbols Γ_{ijk} are symmetric in the indices i and j, see (B.32). Equation (4.5) exhibits that symmetry. □

Lemma 4.2 *The connections $\nabla^{(-\alpha)}$ and $\nabla^{(\alpha)}$ are dual to each other.*

Proof

$$\Gamma^{(-\alpha)}_{ijk} + \Gamma^{(\alpha)}_{ijk} = 2\Gamma^{(0)}_{ijk}$$

yields $\frac{1}{2}(\nabla^{(-\alpha)} + \nabla^{(\alpha)}) = \nabla^{(0)}$. As observed in (B.50), this implies that the two connections are dual to each other. □

The preceding can also be developed in abstract terms. We shall proceed in several steps in which we shall successively narrow down the setting. We start with a Riemannian metric g. During our subsequent steps, that metric will emerge as the abstract version of the Fisher metric. g determines a unique torsion-free connection that respects the metric, the Levi-Civita connection $\nabla^{(0)}$. The steps then consist in the following:

1. We consider two further connections ∇, ∇^* that are dual w.r.t. g. In particular, $\nabla^{(0)} = \frac{1}{2}(\nabla + \nabla^*)$.

2. We assume that both ∇ and ∇^* are torsion-free. It turns out that the tensor T defined by $T(X,Y,Z) = g(\nabla^*_X Y - \nabla_X Y, Z)$ is symmetric in all three entries. (While connections themselves do not define tensors, the difference of connections does, see Lemma B.1.) The structure then is compactly encoded by the symmetric 2-tensor g and the symmetric 3-tensor T. T will emerge as an abstract version of the Amari–Chentsov tensor. Alternatively, the structure can be derived from a divergence, as we shall see in Sect. 4.4. Moreover, as we shall see in Sect. 4.5, any such structure can be isostatistically immersed into a standard structure as defined by the Fisher metric and the Amari–Chentsov tensor.
3. We assume furthermore that ∇ and ∇^* are flat, that is, their curvatures vanish. In that case, locally, there exists a convex function ψ whose second derivatives yield g and whose third derivatives yield T. Moreover, passing from ∇ to ∇^* then is achieved by the Legendre transform of ψ. The primary examples of this latter structure are exponential and mixture families. Their geometries will be explored in Sect. 4.3.

The preceding structures have been given various names in the literature, and we shall try to record them during the course of our mathematical analysis.

We shall now implement those steps. First, following Amari and Nagaoka [16], we formulate

Definition 4.1 A triple (g, ∇, ∇^*) on a differentiable manifold M consisting of a Riemannian metric g and two connections ∇, ∇^* that are dual to each other with respect to g in the sense of Lemma 4.2 is called a *dualistic structure* on M.

(We might then call the quadruple (M, g, ∇, ∇^*) a dualistic space or dualistic manifold, but usually, the underlying manifold M is fixed and therefore need not be referred to in our terminology.)

Of particular importance are dualistic structures with two torsion-free dual connections. According to the preceding, when g is the Fisher metric, then for any $-1 \leq \alpha \leq 1$, $(g, \nabla^{(\alpha)}, \nabla^{(-\alpha)})$ is such a torsion-free dualistic structure. As we shall see, any torsion-free dualistic structure is equivalently encoded by a Riemannian metric g and a 3-symmetric tensor T, leading to the following notion, introduced by Lauritzen [160] and generalizing the pair consisting of the Fisher metric and the Amari–Chentsov tensor.

Definition 4.2 A *statistical structure* on a manifold M consists of a Riemannian metric g and a 3-tensor T that is symmetric in all arguments. A *statistical manifold* is a manifold M equipped with a statistical structure.

We shall now develop the relation between the preceding notions.

Definition 4.3 Let ∇, ∇^* be torsion-free connections that are dual w.r.t. the Riemannian metric g. Then the 3-tensor

$$T = \nabla^* - \nabla \tag{4.7}$$

is called the *3-symmetric tensor* of the triple (g, ∇, ∇^*).

Remark 4.1 The tensor T has been called the skewness tensor by Lauritzen [160].

When the Christoffel symbols of ∇ and ∇^* are Γ_{ijk} and Γ^*_{ijk}, then T has components

$$T_{ijk} = \Gamma^*_{ijk} - \Gamma_{ijk}. \tag{4.8}$$

By Lemma B.1, T is indeed a tensor, in contrast to the Christoffel symbols, because the non-tensorial term in (B.29) drops out when we take the difference of two Christoffel symbols. We now justify the name (see [8] and [11, Thm. 6.1]).

Theorem 4.1 *The* 3*-symmetric tensor* T_{ijk} *is symmetric in all three indices.*

Proof First, T_{ijk} is symmetric in the indices i, j, since Γ_{ijk} and Γ^*_{ijk} are, because they are torsion-free.

To show symmetry w.r.t. j, k, we compare (B.43), that is,

$$Zg(V, W) = g(\nabla_Z V, W) + g\big(V, \nabla^*_Z W\big) \tag{4.9}$$

which expresses the duality of ∇ and ∇^*, with

$$Zg(V, W) = (\nabla_Z g)(V, W) + g(\nabla_Z V, W) + g(V, \nabla_Z W) \tag{4.10}$$

which follows from the product rule for the connection ∇. This yields

$$(\nabla_Z g)(V, W) = g\big(V, \big(\nabla^*_Z - \nabla_Z\big)W\big), \tag{4.11}$$

and since the LHS is symmetric in V and W, so then is the RHS. Writing this out in indices yields the required symmetry of T. Indeed, (4.11) yields

$$\begin{aligned}(\nabla_{\frac{\partial}{\partial x^i}} g)_{jk} &= g\left(\frac{\partial}{\partial x^j}, \big(\nabla^* - \nabla\big)_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^k}\right) \\ &=: g\left(\frac{\partial}{\partial x^j}, T^\ell_{ij} \frac{\partial}{\partial x^\ell}\right),\end{aligned}$$

and hence

$$T_{ijk} = g_{k\ell} T^\ell_{ij} \tag{4.12}$$

is symmetric w.r.t. j, k. □

Conversely, we have the result of Lauritzen [160].

Theorem 4.2 *A metric* g *and a symmetric* 3*-tensor* T *yield a dualistic structure with torsion-free connections.*

Proof Let $\nabla^{(0)}$ be the Levi-Civita connection for g. We define the connection ∇ by

$$g(\nabla_Z V, W) := g\big(\nabla^{(0)}_Z V, W\big) - \frac{1}{2}T(Z, V, W), \tag{4.13}$$

or equivalently, with $\mathcal{T}$ defined by

$$g\big(\mathcal{T}(Z, V), W\big) = T(Z, V, W), \tag{4.14}$$

$$\nabla_Z V = \nabla^{(0)}_Z V - \frac{1}{2}\mathcal{T}(Z, V). \tag{4.15}$$

Then ∇ is linear in Z and V and satisfies the product rule (B.26), because $\nabla^{(0)}$ does and T is a tensor. It is torsion free, because $\mathcal{T}$ is symmetric. Indeed,

$$\nabla_Z V - \nabla_V Z - [Z, V] = \nabla_Z V - \nabla_V Z - [Z, V] - \frac{1}{2}\big(\mathcal{T}(Z, V) - \mathcal{T}(V, Z)\big) = 0.$$

Moreover,

$$\nabla^*_Z V = \nabla^{(0)}_Z V + \frac{1}{2}\mathcal{T}(Z, V) \tag{4.16}$$

is a torsion free connection that is dual to ∇ w.r.t. g:

$$\begin{aligned}
&g(\nabla_Z V, W) + g\big(V, \nabla^*_Z W\big) \\
&\quad = g\big(\nabla^{(0)}_Z V, W\big) + g\big(V, \nabla^{(0)}_Z W\big) - \frac{1}{2}T(Z, V, W) + \frac{1}{2}T(Z, W, V) \\
&\quad = g\big(\nabla^{(0)}_Z V, W\big) + g\big(V, \nabla^{(0)}_Z W\big) \quad \text{by the symmetry of } T \\
&\quad = Z g(V, W)
\end{aligned}$$

since $\nabla^{(0)}$ is the Levi-Civita connection, see (B.46). □

The pair (g, T) represents the structure more compactly than the triple (g, ∇, ∇^*), because in contrast to the connections, T transforms as a tensor by Lemma B.1. In fact, from such a pair (g, T), we can generate an entire family of torsion free connections

$$\nabla^{(\alpha)}_Z V = \nabla^{(0)}_Z V - \frac{\alpha}{2}\mathcal{T}(Z, V) \quad \text{for } -1 \le \alpha \le 1, \tag{4.17}$$

or written with indices

$$\Gamma^{(\alpha)}_{ijk} = \Gamma^{(0)}_{ijk} - \frac{\alpha}{2}T_{ijk}. \tag{4.18}$$

The connections $\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$ are then dual to each other. And

$$\nabla^{(0)} = \frac{1}{2}\big(\nabla^{(\alpha)} + \nabla^{(-\alpha)}\big) \tag{4.19}$$

then is the Levi-Civita connection, because it is metric and torsion free.

Such statistical structures will be further studied in Sects. 4.3 and 4.5.

We now return to an extrinsic setting and consider families of probability distributions, equipped with their Fisher metric. We shall then see that for $\alpha = \pm 1$, we obtain additional properties. We begin with an exponential family (3.31),

$$p(x;\vartheta) = \exp\big(\gamma(x) + f_i(x)\vartheta^i - \psi(\vartheta)\big). \tag{4.20}$$

So, here we require that the function $\exp(\gamma(x) + f_i(x)\vartheta^i - \psi(\vartheta))$ be integrable for all parameter values ϑ under consideration, and likewise that expressions like $f_j(x)\exp(\gamma(x) + f_i(x)\vartheta^i - \psi(\vartheta))$ or $f_j(x)f_k(x)\exp(\gamma(x) + f_i(x)\vartheta^i - \psi(\vartheta))$ be integrable as well. In particular, as analyzed in detail in Sects. 3.2 and 3.3, this is a more stringent requirement than the $f_i(x)$ simply being L^1-functions. But if the model provides a family of finite measures, i.e., if it is a parametrized measure model, then it is always ∞-integrable, so that all canonical tensors and, in particular, the Fisher metric and the Amari–Chentsov tensor are well-defined; cf. Example 3.3. We observe that we may allow for points x with $\gamma(x) = -\infty$, as long as those integrability conditions are not affected. The reason is that $\exp(\gamma(x))$ can be incorporated into the base measure.

When we have such integrability conditions, then differentiability w.r.t. the parameters ϑ^j holds.

We compute

$$\frac{\partial}{\partial\vartheta^j}\log p(x;\vartheta) = \big(f_j(x) - \mathbb{E}_{\mathbf{p}}(f_j)\big)p(x;\vartheta). \tag{4.21}$$

In particular,

$$\mathbb{E}_{\mathbf{p}}\left(\frac{\partial}{\partial\vartheta^k}\log p\right) = 0, \tag{4.22}$$

because $\mathbb{E}_{\mathbf{p}}(f_j(x) - \mathbb{E}_{\mathbf{p}}(f_j)) = 0$ or because of the normalization $\int p\,dx = 1$. We then get

$$\begin{aligned}\Gamma^{(1)}_{ijk} &= \mathbb{E}_{\mathbf{p}}\left(\frac{\partial^2}{\partial\vartheta^i\partial\vartheta^j}\log p\,\frac{\partial}{\partial\xi^k}\log p\right)\\ &= -\frac{\partial^2}{\partial\vartheta^i\partial\vartheta^j}\psi(\vartheta)\,\mathbb{E}_{\mathbf{p}}\left(\frac{\partial}{\partial\vartheta^k}\log p\right)\\ &= 0.\end{aligned}$$

Thus, ϑ yields an affine coordinate system for the connection $\nabla^{(1)}$, and we have

Lemma 4.3 *The connection $\nabla^{(1)}$ is flat.*

Proof See (B.34) in the Appendix. □

Definition 4.4 The connection $\nabla^{(1)}$ is called the *exponential connection*, abbreviated as *e-connection*.

We now consider a family

$$p(x;\eta) = c(x) + \sum_{i=1}^{d} g^i(x)\eta_i, \tag{4.23}$$

an affine family of probability measures, a so-called mixture family. (We might wish to add a term $\nu(\eta)$ in order to achieve the normalization $\int p(x;\eta)\,dx = 1$, but as one readily computes that this ν is given by $\nu(\eta) = 1 - \int (c(x) + \sum g^i(x)\eta_i)\,dx$, which is linear in η, it can simply be incorporated by a redefinition of the functions $c(x)$ and $g^i(x)$. Here, we require that the functions $c(x)$ and $g^i(x)$ be integrable, and the expressions (4.25) and (4.26) below as well. Again, differentiability w.r.t. the parameters η^j is then obvious.

We then have

$$\int c(x)\,dx = 1, \qquad \int g^i(x)\,dx = 0 \quad \text{for all } i.) \tag{4.24}$$

And then,

$$\frac{\partial}{\partial\eta_i}\log p(x;\eta) = \frac{g^i(x)}{p(x;\eta)}, \tag{4.25}$$

and

$$\frac{\partial^2}{\partial\eta_i\partial\eta_j}\log p(x;\eta) = -\frac{g^i(x)g^j(x)}{p(x;\eta)^2}. \tag{4.26}$$

Consequently,

$$\frac{\partial^2}{\partial\eta_i\partial\eta_j}\log p + \frac{\partial}{\partial\eta_i}\log p\,\frac{\partial}{\partial\eta_j}\log p = 0.$$

This implies

$$\Gamma^{(-1)}_{ijk} = 0. \tag{4.27}$$

In other words, now η is an affine coordinate system for the connection $\nabla^{(-1)}$, and analogously to Lemma 4.3, (4.27) implies

Lemma 4.4 *The connection $\nabla^{(-1)}$ is flat.* □

Definition 4.5 $\nabla^{(-1)}$ is called the *mixture* or *m-connection*.

These connections have already been described in more concrete terms in Sect. 2.4. Amari and Nagaoka [16] call a triple (g, ∇, ∇^*) consisting of a Riemannian metric g and two connections that are dual to each other and both flat a *dually flat structure*.

We shall now develop an intrinsic perspective, that is, no longer speak about families of probability distributions, but simply consider a triple consisting of a

Riemannian metric and two torsion-free flat connections that are dual to each other. Let ∇ and ∇^* be dually flat connections. We choose affine coordinates $\vartheta^1, \ldots, \vartheta^d$, for ∇; the vector fields $\partial_i := \frac{\partial}{\partial \vartheta^i}$ are then parallel. We define vector fields ∂^j via

$$\langle \partial_i, \partial^j \rangle = \delta_i^j \quad \left(= \begin{cases} 1 & \text{for } i = j \\ 0 & \text{else} \end{cases} \right). \tag{4.28}$$

We have for any vector V

$$0 = V\langle \partial_i, \partial^j \rangle = \langle \nabla_V \partial_i, \partial^j \rangle + \langle \partial_i, \nabla_V^* \partial^j \rangle,$$

and since ∂_i is parallel for ∇, we conclude that ∂^j is parallel for ∇^*. Since ∇^* is torsion-free, then also $[\partial^j, \partial^k] = 0$ for all j and k, and so, we may find ∇^*-affine coordinates η_j with $\partial^j = \frac{\partial}{\partial \eta_j}$. Here and in the following, the position of the indices, i.e., whether we have upper or lower indices, is important because it indicates the transformation behavior under coordinate changes. For example, if when changing the ϑ-coordinates ∂_i transforms as a vector (contravariantly), then ∂^j transforms as a 1-form (covariantly). For changes of the η-coordinates, the rules are reversed. In particular, we have

$$\partial^j = \big(\partial^j \vartheta^i\big)\partial_i \quad \text{and} \quad \partial_i = (\partial_i \eta_j)\partial^j \tag{4.29}$$

as the transition rules between the ϑ- and η-coordinates. Writing then the metric tensor in terms of the ϑ- and η-coordinates, resp., as

$$g_{ij} := \langle \partial_i, \partial_j \rangle, \qquad g^{ij} := \big\langle \partial^i, \partial^j \big\rangle, \tag{4.30}$$

we obtain from $\langle \partial_i, \partial^j \rangle = \delta_i^j$

$$\frac{\partial \eta_j}{\partial \vartheta^i} = g_{ij}, \qquad \frac{\partial \vartheta^i}{\partial \eta_j} = g^{ij}. \tag{4.31}$$

Theorem 4.3 *There exist strictly convex potential functions $\varphi(\eta)$ and $\psi(\vartheta)$ satisfying*

$$\eta_i = \partial_i \psi(\vartheta), \qquad \vartheta^i = \partial^i \varphi(\eta), \tag{4.32}$$

as well as

$$g_{ij} = \partial_i \partial_j \psi, \tag{4.33}$$

$$g^{ij} = \partial^i \partial^j \varphi. \tag{4.34}$$

Proof The first equation of (4.32) can be solved locally iff

$$\partial_i \eta_j = \partial_j \eta_i. \tag{4.35}$$

From the preceding equation, this is nothing but the symmetry

$$g_{ij} = g_{ji}, \tag{4.36}$$

and so, local solvability holds; moreover, we obtain

$$g_{ij} = \partial_i \partial_j \psi. \tag{4.37}$$

Thus, ψ is strictly convex. φ can be found by the same reasoning or, more elegantly, by duality; in fact, we simply put

$$\varphi := \vartheta^i \eta_i - \psi \tag{4.38}$$

from which

$$\partial^i \varphi = \vartheta^i + \frac{\partial \vartheta^j}{\partial \eta_i} \eta_j - \frac{\partial \vartheta^j}{\partial \eta_i} \frac{\partial}{\partial \vartheta^j} \psi = \vartheta^i. \qquad \square$$

Since ψ and φ are strictly convex, the relation

$$\varphi(\eta) + \psi(\vartheta) = \vartheta^i \eta_i \tag{4.39}$$

means that they are related by Legendre transformations,

$$\varphi(\eta) = \max_{\vartheta} \big(\vartheta^i \eta_i - \psi(\vartheta)\big), \tag{4.40}$$

$$\psi(\vartheta) = \max_{\eta} \big(\vartheta^i \eta_i - \phi(\eta)\big). \tag{4.41}$$

Of course, all these formulae are valid locally, i.e., where ψ and φ are defined. In fact, the construction can be reversed, and all that is needed locally is a convex function $\psi(\vartheta)$ of some local coordinates.

Remark 4.2

(1) Cheng–Yau [59] call an affine structure that is obtained from such local convex functions a Kähler affine structure and consider it a real analogue of Kähler geometry. The analogy consists in the fact that the Kähler form of a Kähler manifold can be locally obtained as the complex Hessian of some function, in the same manner that here, the metric is locally obtained from the real Hessian. In fact, concepts that have been developed in the context of Kähler geometry, like Chern classes, can be transferred to this affine context and can then be used to derive restrictions for a manifold to carry such a dually flat structure. Shima [236] speaks of a Hessian structure instead. For instance, Amari and Armstrong in [13] show that the Pontryagin forms vanish on Hessian manifolds. This is a strong local constraint.

(2) In the context of decision theory, this duality was worked out by Dawid and Lauritzen [81]. This includes concepts like the Bregman divergences.

(3) In statistics, so-called curved exponential families also play a role; see, for instance, [16]. A curved exponential family is a submanifold of some exponential family. That is, we have some mapping $M' \to M, \xi \mapsto \vartheta(\xi)$ from some parameter space M' into the parameter space M of an exponential family as in (4.20) and consider a family of the form

$$p(x;\xi) = \exp\big(\gamma(x) + f_i(x)\vartheta^i(\xi) - \psi\big(\vartheta(\xi)\big)\big) \tag{4.42}$$

parametrized by $\xi \in M'$. The family is called curved because M' does not need to carry an affine structure here, and even if it does, the mapping $\xi \mapsto \vartheta(\xi)$ does not need to be affine.

In order to see that everything can be derived from a strictly convex function $\psi(\vartheta)$, we define the metric

$$g_{ij} = \partial_i \partial_j \psi$$

and the α-connection through

$$\Gamma_{ijk}^{(\alpha)} = \Gamma_{ijk}^{(0)} - \frac{\alpha}{2}\partial_i \partial_j \partial_k \psi$$

where $\Gamma_{ijk}^{(0)}$ is the Levi-Civita connection for g_{ij}. Since

$$\Gamma_{ijk}^{(0)} = \frac{1}{2}(g_{ik,j} + g_{jk,i} - g_{ij,k}) = \frac{1}{2}\partial_i \partial_j \partial_k \psi, \tag{4.43}$$

we have

$$\Gamma_{ijk}^{(\alpha)} = \frac{1}{2}(1-\alpha)\partial_i \partial_j \partial_k \psi, \tag{4.44}$$

and since this is symmetric in i and j, $\nabla^{(\alpha)}$ is torsion free. Since $\Gamma_{ijk}^{(\alpha)} + \Gamma_{ijk}^{(-\alpha)} = 2\Gamma_{ijk}^{(0)}$, $\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$ are dual to each other. Recalling (4.18),

$$T_{ijk} = \partial_i \partial_j \partial_k \psi \tag{4.45}$$

is the 3-symmetric tensor.

In particular, $\Gamma_{ijk}^{(1)} = 0$, and so $\nabla^{(1)}$ defines a flat structure, and the coordinates ϑ are affine coordinates for $\nabla^{(1)}$.

Remark 4.3 As an aside, we observe that in the ϑ-coordinates, the curvature of the Levi-Civita connection becomes

$$\begin{aligned} R_{lij}^k &= \frac{1}{2}g^{kn}g^{mr}\Big(\partial_j\partial_n\partial_r\psi\ \partial_i\partial_l\partial_m\psi - \partial_j\partial_l\partial_m\psi\ \partial_i\partial_n\partial_r\psi \\ &\qquad + \frac{1}{2}\partial_j\partial_l\partial_r\psi\ \partial_i\partial_m\partial_n\psi - \frac{1}{2}\partial_j\partial_m\partial_n\psi\ \partial_i\partial_l\partial_r\psi\Big) \\ &= \frac{1}{4}\big(T_{jr}^k T_{\ell i}^r - T_{ir}^k T_{\ell j}^r\big) \end{aligned}$$

when writing it in terms of the 3-symmetric tensor. Remarkably, the curvature tensor can be computed from the second and third derivatives of ψ; no fourth derivatives are involved. The reason is that the derivatives of the Christoffel symbols in (B.37) drop out by symmetry because the Christoffel symbols in turn can be computed from derivatives of ψ, see (4.43), and those commute. The curvature tensor thus becomes a quadratic expression of coefficients of the 3-symmetric tensor.

In particular, if we subject the ϑ-coordinates to a linear transformation so that at the point under consideration,

$$g^{ij} = \delta^{ij},$$

we get

$$\begin{aligned} R^k_{lij} &= \frac{1}{4}(\partial_j\partial_m\partial_k\psi\ \partial_i\partial_m\partial_l\psi - \partial_j\partial_m\partial_l\psi\ \partial_i\partial_m\partial_k\psi) \\ &= \frac{1}{4}(T_{jkm}T_{i\ell m} - T_{j\ell m}T_{ikm}) \end{aligned} \tag{4.46}$$

(where we sum over the index m on the right). In a terminology developed in the context of mirror symmetry, we thus have an affine structure that in general is not a Frobenius manifold (see [86] for this notion) because the latter condition would require that the Witten–Dijkgraaf–Verlinde–Verlinde equations hold, which are equivalent to the vanishing of the curvature. Here, we have found a nice representation of these equations in terms of the 3-symmetric tensor. While this vanishing may well happen for the potential functions ψ for certain dually flat structures, it does not hold for the Fisher metric as we have seen already that it has constant positive sectional curvature.

The dual connection then is $\nabla^{(-1)}$, with Christoffel symbols

$$\Gamma^{(-1)}_{ijk} = \partial_i\partial_j\partial_k\psi \tag{4.47}$$

with respect to the ϑ-coordinates. The dually affine coordinates η can be obtained as before:

$$\eta_j = \partial_j\psi, \tag{4.48}$$

and so also

$$g_{ij} = \partial_i\eta_j. \tag{4.49}$$

The corresponding potential is again obtained by a Legendre transformation

$$\varphi(\eta) = \max_{\vartheta}\big(\vartheta^i\eta_i - \psi(\vartheta)\big), \qquad \psi(\vartheta) + \varphi(\eta) - \vartheta\cdot\eta = 0, \tag{4.50}$$

and

$$\vartheta^j = \partial^j\varphi(\eta), \qquad g^{ij} = \frac{\partial\vartheta^j}{\partial\eta_i} = \partial^i\partial^j\varphi(\eta). \tag{4.51}$$

We also remark that the Christoffel symbols for the Levi-Civita connection for the metric g^{ij} with respect to the ϑ-coordinates are given by

$$\tilde{\Gamma}^{ijk} = -\Gamma_{ijk} = -\frac{1}{2}\partial_i\partial_j\partial_k\psi, \tag{4.52}$$

and so

$$\tilde{\Gamma}^{(\alpha)ijk} = \tilde{\Gamma}^{ijk} - \frac{\alpha}{2}\partial_i\partial_j\partial_k\psi = -\Gamma^{(-\alpha)}_{ijk},$$

and so, with respect to the dual metric g^{ij}, α and $-\alpha$ reverse their rules. So, $\tilde{\Gamma}^{(1)} = -\Gamma^{(-1)}$ vanishes in the η-coordinates.

In conclusion, we have shown

Theorem 4.4 *A dually flat structure, i.e., a Riemannian metric g together with two flat connections ∇ and ∇^* that are dual with respect to g is locally equivalent to the datum of a single convex function ψ, where convexity here refers to local coordinates ϑ and not to any metric.*

As observed at the end of Appendix B, it suffices that the connections ∇ and ∇^* are dual with respect to g and torsion-free and that the curvature of one of them vanishes, because this then implies that the other curvature also vanishes, see Lemma B.2.

One should point out, however, in order to get a global structure from such local data, compatibility conditions under coordinate changes need to be satisfied. In general, this is a fundamental point in geometry, but for our present purposes, this is not so relevant as the manifolds of probability distributions do not exhibit nontrivial global phenomena. For example, in the case of a finite underlying space, the probability distributions are represented by a simplex or a spherical sector, as we have seen above, and so, the topology is trivial.

For a dually flat structure, completeness of one of the connections does not imply completeness of the other one. In fact, in information geometry, the exponential connection is complete (under appropriate assumptions) while the mixture connection is not. (In this regard, see also the example constructed from (B.41) in Appendix B of an incomplete connection on a complete manifold.)

Finally, we observe that in a dually flat structure, the assumption that a submanifold be autoparallel implies the seemingly stronger property that it is itself dually flat w.r.t. the induced structure (see [16]).

Lemma 4.5 *Let (g, ∇, ∇^*) be a dually flat structure on the manifold M, and let S be a submanifold of M. Then, if S is autoparallel for ∇ or ∇^*, then it is dually flat w.r.t. the metrics and connections induced from (g, ∇, ∇^*).*

Proof That S is autoparallel w.r.t., say, ∇ means that the restriction of ∇ to vector fields tangent to S is a connection of S, as observed after (B.42), which then is also

flat. And by (B.43), the connection $\nabla^{S,*}$ induced by ∇^* on S satisfies

$$Z\langle V, W\rangle = \langle \nabla_Z V, W\rangle + \langle V, \nabla^*_Z W\rangle \tag{4.53}$$

for all tangent vectors Z of S and vector fields V, W that are tangent to S. Since ∇^* is torsion-free, so then is $\nabla^{S,*}$, and since the curvature of $\nabla^S = \nabla$ vanishes, so then does the curvature of the dual connection $\nabla^{S,*}$ by Lemma B.2. Thus, both ∇^S and $\nabla^{S,*}$ are flat, and S is dually flat w.r.t. the induced metric and connections. □

4.3 The Duality Between Exponential and Mixture Families

In this section, we shall assume that the parameter space M is a (finite-dimensional) differentiable manifold. Instead of considering φ and ψ as functions of the coordinates η and ϑ, resp., we may consider them as functions on our manifold M, i.e., for $p \in M$, instead of

$$\psi\big(\vartheta(p)\big),$$

we simply write

$$\psi(p)$$

by abuse of notation.

We now discuss another important concept of Amari and Nagaoka, see Sect. 3.4 in [16].

Definition 4.6 For $p, q \in M$, we define the *canonical divergence*

$$D(p\|q) := \psi(p) + \varphi(q) - \vartheta^i(p)\eta_i(q). \tag{4.54}$$

(Note the contrast to (4.39) where all expressions were evaluated at the same point.) From (4.40) or (4.41), we have

$$D(p\|q) \geq 0, \tag{4.55}$$

and

$$D(p\|q) = 0 \quad \Longleftrightarrow \quad p = q \tag{4.56}$$

by (4.39) and strict convexity. In general, however, $D(p\|q)$ is not symmetric in p and q. Thus, while $D(p\|q)$ behaves like the square of a distance function in a certain sense, it is not a true squared distance function. Also, for a derivative with

respect to the first argument[2]

$$\begin{aligned} D\big((\partial_i)p\,\big\|\,q\big) &:= \partial_i D(p\|q) \\ &= \partial_i \psi(p) - \eta_i(q) \\ &= \eta_i(p) - \eta_i(q) \quad \text{by (4.32).} \end{aligned} \tag{4.57}$$

This vanishes for all i precisely at $p = q$. Again, this is the same behavior as shown locally by the square of a distance function on a Riemannian manifold which has a global minimum at $p = q$. Likewise for a derivative with respect to the second argument

$$D\big(p\,\big\|\,(\partial^j)q\big) = \partial^j D(p\|q) = \partial^j \varphi(q) - \vartheta^j(p). \tag{4.58}$$

Again, this vanishes for all j iff $p = q$.

For the second derivatives

$$D\big((\partial_i\partial_j)p\,\big\|\,q\big)_{|q=p} = \partial_i\partial_j D(p\|q)_{|q=p} = g_{ij}(p) \quad \text{by (4.37)} \tag{4.59}$$

and

$$D\big(p\,\big\|\,(\partial^i\partial^j)q\big)_{|p=q} = \partial^i\partial^j D(p\|q)_{|p=q} = g^{ij}(q). \tag{4.60}$$

Thus, the metric is reproduced from the distance-like 2-point function $D(p\|q)$.

Theorem 4.5 *The divergence is characterized by the relation*

$$D(p\|q) + D(q\|r) - D(p\|r) = \big(\vartheta^i(p) - \vartheta^i(q)\big)\big(\eta_i(r) - \eta_i(q)\big), \tag{4.61}$$

using (4.39), *i.e.*, $\psi(q) + \varphi(q) = \vartheta^i(q)\eta_i(q)$.

This is the product of the two tangent vectors at q. Equation (4.61) can be seen as a generalization of the cosine formula in Hilbert spaces,

$$\frac{1}{2}\|p-q\|^2 + \frac{1}{2}\|q-r\|^2 - \frac{1}{2}\|p-r\|^2 = \langle p-q, r-q\rangle.$$

Proof To obtain (4.61), we have from (4.54)

$$\begin{aligned} &D(p\|q) + D(q\|r) - D(p\|r) \\ &\quad = \psi(p) + \varphi(q) - \vartheta^i(p)\eta_i(q) + \psi(q) + \varphi(r) - \vartheta^i(q)\eta_i(r) \\ &\qquad - \psi(p) - \varphi(r) + \vartheta^i(p)\eta_i(r) \end{aligned}$$

[2]The notation employed is based on the convention that a derivative involving ϑ-variables operates only on those expressions that are naturally expressed in those variables, and not on those expressed in the η-variables. Thus, it operates, for example, on ψ, but not on φ.

using (4.39), i.e., $\psi(q) + \varphi(q) = \vartheta^i(q)\eta_i(q)$. Conversely, if (4.61) holds, then $D(p\|p) = 0$ and, by differentiating (4.61) w.r.t. p and then setting $r = p$, $D((\partial_i)p\|q) = \eta_i(p) - \eta_i(q)$. Since a solution of this differential equation is unique, D has to be the divergence. Thus, the divergence is indeed characterized by (4.61). □

The following conclusions from (4.61), Corollaries 4.1, 4.2, and 4.3, can be seen as abstract instances of Theorem 2.8 (note, however, that Theorem 2.8 also handles situations where the image of the projection is in the boundary of the simplex; this is not covered by the present result).

Corollary 4.1 *The ∇-geodesic from q to p is given by $t\vartheta^i(p) + (1-t)\vartheta^i(q)$, since ϑ^i are affine coordinates for ∇, and likewise, the ∇^*-geodesic from q to r is $t\eta_i(r) + (1-t)\eta_i(q)$. Thus, if those two geodesics are orthogonal at q, we obtain the Pythagoras relation*

$$D(p\|r) = D(p\|q) + D(q\|r). \tag{4.62}$$

□

In particular, such a q where these two geodesics meet orthogonally is the point closest to p on that ∇^*-geodesic. We may therefore consider such a q as the projection of p on that latter geodesic. By the same token, we can characterize the projection of $p \in M$ onto an autoparallel submanifold N of M by such a relation. That is, the unique point q on N minimizing the canonical divergence $D(p\|r)$ among all $r \in N$ is characterized by the fact that the ∇-geodesic from p to N meets N orthogonally. This observation admits the following generalization:

Corollary 4.2 *Let N be a differentiable submanifold of M. Then $q \in N$ is a stationary point of the function $D(p\|\cdot) : N \to \mathbb{R}, r \mapsto D(p\|r)$ iff the ∇-geodesic from p to q meets N orthogonally.*

Proof We differentiate (4.61) w.r.t. $r \in N$ and then put $q = r$. Recalling that (4.58) vanishes when the two points coincide, we obtain

$$-D\big(p\big\|\partial^j r\big) = \big(\vartheta^i(p) - \vartheta^i(q)\big)\partial^j\eta_i(r).$$

Thus, when we consider a curve $r(t), t \in [0,1]$ with $r(0) = q$ in N, we have

$$-\frac{d}{dt}D\big(p\big\|r(t)\big) = \big(\vartheta^i(p) - \vartheta^i(q)\big)\frac{\partial\eta_i(r)}{\partial\eta_j}\frac{d\eta_j(t)}{dt}.$$

For $t = 0$, this is the product between the tangent vector of the geodesic from p to q and the tangent vector of the curve $r(t)$ at $q = r(0)$, as in (4.61). This yields the claim. □

We now turn to a situation where we can even find and characterize minima for the projection onto a submanifold.

Corollary 4.3 *Let N be a submanifold of M that is autoparallel for the connection ∇^*. Let $p \in M$. Then $q \in N$ satisfies*

$$q = \operatorname*{argmin}_{r \in N} D(p\|r) \tag{4.63}$$

precisely if the ∇-geodesic from p to q is orthogonal to N at q.

Proof This follows directly from Corollary 4.1, see (4.61). □

Corollary 4.4 *Let $N_1 \subseteq N_2$ be differentiable submanifolds of M which are autoparallel w.r.t. to ∇^*, and assume that N_2 is complete w.r.t. ∇.*[3] *Let q_i be the projection in the above sense of some distribution p onto N_i, $i = 1, 2$. Then q_1 is also the projection of q_2 onto N_1, and we have*

$$D(p\|q_1) = D(p\|q_2) + D(q_2\|q_1). \tag{4.64}$$

Proof Since $q_1 \in N_1 \subseteq N_2$, we may apply Corollary 4.1 to get the Pythagoras relation (4.64). By Lemma 4.5, both N_2 and N_1 are dually flat (although here we actually need this property only for N_2). Therefore, the minimizing ∇-geodesic from q_2 to N_1 (which exists as N_2 is assumed to be ∇-complete) stays inside N_2, and it is orthogonal to N_1 at its endpoint q_1^* by Corollary 4.3, and by Corollary 4.1 again, we get the Pythagoras relation

$$D\big(p\big\|q_1^*\big) = D(p\|q_2) + D\big(q_2\big\|q_1^*\big). \tag{4.65}$$

Since $D(p\|q_1^*) \geq D(p\|q_1)$ and $D(q_2\|q_1^*) \leq D(q_2\|q_1)$ by the respective minimizing properties, comparing (4.64) and (4.65) shows that we must have equality in both cases. Likewise, we may apply the Pythagoras relation in N_1 to get $D(q_2\|q_1) = D(q_2\|q_1^*) + D(q_1^*|q_1)$ to then infer $D(q_1^*\|q_1) = 0$ which by the properties of the divergence D (see (4.56)) implies $q_1^* = q_1$. This concludes the proof. □

We now return to families of probability distributions and consider the prime example to which we want to apply the preceding theory, the exponential family (4.20)

$$p(x;\vartheta) = \exp\big(\gamma(x) + f_i(x)\vartheta^i - \psi(\vartheta)\big), \tag{4.66}$$

with

$$\psi(\vartheta) = \log \int \exp\big(\gamma(x) + f_i(x)\vartheta^i\big)\,dx, \tag{4.67}$$

that is,

$$p(x;\vartheta) = \frac{1}{Z(\vartheta)} \exp\big(\gamma(x) + f_i(x)\vartheta^i\big) \tag{4.68}$$

[3] We shall see in the proof why this assumption is meaningful.

with the expression

$$Z(\vartheta) := \int \exp\bigl(\gamma(x) + f_i(x)\vartheta^i\bigr)\,dx = e^{\psi(\vartheta)}, \tag{4.69}$$

which is called the *zustandssumme* or *partition function* in statistical mechanics. According to the theory developed in Sect. 4.2, such an exponential family carries a dually flat structure with the Fisher metric and the exponential and the mixture connection. We can therefore explore the implications of Theorem 4.4. Also, exponential subfamilies, that is, when we restrict the ϑ-coordinates to some *linear* subspace, inherit such a dually flat structure, according to Lemma 4.5.

There are two special cases of (4.66) where the subsequent formulae will simplify somewhat. The first case occurs when $\gamma(x) = 0$. In fact, the general case can be reduced to this one, because $\exp(\gamma(x))$ can be incorporated in the base measure $\mu(x)$, compare (3.12). Nevertheless, the function γ will play a role in Sect. 6.2.3. We could also introduce $f_0(x) = \gamma(x)$ and put the corresponding coefficient $\vartheta^0 = 1$.

The other simple case occurs when there are no f_i. Anticipating some of the discussion in Sect. 6.4, we call

$$\Gamma\bigl(p(\cdot;\vartheta)\bigr) := -\int p(x;\vartheta)\gamma(x)\,dx \tag{4.70}$$

the *potential energy* and obtain the relation

$$\begin{aligned}
\psi &= \log Z(\vartheta) \\
&= \int \exp\bigl(\gamma(x) - \psi(\vartheta)\bigr)\psi(\vartheta)\,dx \quad \text{since } \int p(x;\vartheta)\,dx = 1 \\
&= -\int p(x;\vartheta)\log p(x;\vartheta)\,dx + \int p(x;\vartheta)\gamma(x)\,dx \\
&= -\varphi - \Gamma,
\end{aligned} \tag{4.71}$$

where $-\varphi$ is the *entropy*. Thus, the *free energy*, defined as $-\psi$, is the difference between the potential energy and the entropy.[4]

We return to the general case (4.66). We have the simple identity

$$\frac{\partial^k Z(\vartheta)}{\partial\vartheta^{i_1}\dots\partial\vartheta^{i_k}} = \int f_{i_1}\cdots f_{i_k}\exp\bigl(\gamma(x) + f_i(x)\vartheta^i\bigr)\,dx, \tag{4.72}$$

and hence

$$\mathbb{E}_p(f_{i_1}\cdots f_{i_k}) = \frac{1}{Z(\vartheta)}\frac{\partial^k Z(\vartheta)}{\partial\vartheta^{i_1}\cdots\partial\vartheta^{i_k}}, \tag{4.73}$$

an identity that will appear in various guises in the rest of this section. Of course, by (4.69), we can and shall also express this identity in terms of $\psi(\vartheta)$. Thus, when we

[4] The various minus signs here come from the conventions of statistical physics, see Sect. 6.4.

know Z, or equivalently ψ, we can compute all expectation values of the "observables" f_i.

First, we put

$$\eta_i(\vartheta) := \int f_i(x)\, p(x;\vartheta)\, dx = \mathbb{E}_p(f_i),$$

$$\text{the expectation of the coefficient of } \vartheta^i \text{ w.r.t.}$$

$$\text{the probability measure } p(\cdot;\vartheta), \tag{4.74}$$

and have from (4.73)

$$\eta_i = \partial_i \psi, \tag{4.75}$$

as well as, recalling (3.34),

$$g_{ij} = \partial_i \partial_j \psi \quad \text{for the Fisher information metric,} \tag{4.76}$$

as computed above. For the dual potential, we find

$$\begin{aligned} \varphi(\eta) &= \vartheta^i \eta_i - \psi(\vartheta) \\ &= \int \big(\log p(x;\vartheta) - \gamma(x)\big)\, p(x;\vartheta)\, dx, \end{aligned} \tag{4.77}$$

involving the entropy $-\int \log p(x;\vartheta)\ p(x;\vartheta)\, dx$; this generalizes (4.71). For the divergence, we get

$$\begin{aligned} D(p\|q) &= \psi(\vartheta) - \vartheta^i \int f_i(x) q(x;\eta)\, dx + \int \big(\log q(x;\eta) - \gamma(x)\big) q(x;\eta)\, dx \\ &= \psi(\vartheta) - \int \big(\log p(x;\vartheta) - \gamma(x) + \psi(\vartheta)\big) q(x;\eta)\, dx \\ &\quad + \int \big(\log q(x;\eta) - \gamma(x)\big) q(x;\eta)\, dx \\ &= \int \big(\log q(x) - \log p(x)\big) q(x)\, dx, \end{aligned} \tag{4.78}$$

the dual of the *Kullback–Leibler divergence* D_{KL} introduced in (2.154).

Equation (4.74) is a linear relationship between η and p, and so, by inverting it, we can express the $p(\cdot;\vartheta)$ as linear combinations of the η_i. (We assume here, as always, that the f_i are independent, i.e., that the parametrization of the family $p(x;\vartheta)$ is non-redundant.) In other words, when replacing the coordinates ϑ by η, we obtain a mixture family

$$p(x;\eta) = c(x) + g^i(x)\eta_i.$$

(Obviously, we are abusing the notation here, because the functional dependence of p on η will be different from the one on ϑ.)

We check the consistency

$$\eta_i(\vartheta) = \int f_i(x)\, p(x;\vartheta)\, dx = \int f_i(x)\big(c(x) + g^j(x)\eta_j\big)\, dx$$

from (4.74), from which we obtain

$$\int f_i(x)c(x)\, dx = 0, \qquad \int f_i(x)g^j(x)\, dx = \delta_i^j. \tag{4.79}$$

If we consider the potential $\varphi(\eta)$ given by (4.77), for computing the inverse of the Fisher metric through its second derivatives as in (4.51), we may suppress the term $-\int \gamma(x)\ p(x;\eta)dx$ as this is linear in η. Then the potential is the negative of the entropy, and

$$\begin{aligned}
&\frac{\partial^2}{\partial\eta_i\partial\eta_j}\int \log p(x;\eta)\ p(x;\eta)\, dx \\
&\quad = \int \frac{\partial}{\partial\eta_i}\log p(x;\eta)\frac{\partial}{\partial\eta_j}\log p(x;\eta)\ p(x;\eta)\, dx \\
&\quad = \int g^i(x)g^j(x)\frac{1}{c(x) + g^k(x)\eta_k}\, dx
\end{aligned}$$

is the inverse of the Fisher metric.

Thus, we have

Theorem 4.6 *With respect to the mixture coordinates, (the negative of) the entropy is a potential for the Fisher metric.* □

It is also instructive to revert the construction and go from the η_i back to the ϑ^i; namely, we have

$$\begin{aligned}
\vartheta^j &= \frac{\partial}{\partial\eta_j}\int \big(c(x) + g^i(x)\eta_i\big)\big(\log\big(c(x) + g^i(x)\eta_i\big) - \gamma(x)\big)\, dx \\
&= \int g^j(x)\big(\log\big(c(x) + g^i(x)\eta_i\big) - \gamma(x) + 1\big)\, dx \\
&= \int g^j(x)\big(\log p(x;\eta) - \gamma(x) + 1\big)\, dx \\
&= \int g^j(x)\big(\log p(x;\eta) - \gamma(x)\big)\, dx
\end{aligned}$$

because $\int g^j(x)\, dx = 0$. This is a linear relationship between ϑ and $\log p(x;\eta) - \gamma(x)$, and so, we can invert it and write

$$\log p(x;\vartheta) = \gamma(x) + f_i(x)\vartheta^i \tag{4.80}$$

to express $\log p$ as a function of ϑ, except that then the normalization $\int p(x;\vartheta)\,dx = 1$ does not necessarily hold, and so, we need to subtract a term $\psi(\vartheta)$, i.e.,

$$\log p(x;\vartheta) = \gamma(x) + f_i(x)\vartheta^i - \psi(\vartheta). \tag{4.81}$$

The reason that $\psi(\vartheta)$ is undetermined comes from $\int g^j(x)\,dx = 0$; namely, we must have the consistency

$$\vartheta^j = \int g^j(x)\big(f_i(x)\vartheta^i - \psi(\vartheta) + 1\big)\,dx$$

from the above, and this holds because of $\int g^j(x)\,dx = 0$ and $\int g^j(x) f_i(x)\,dx = \delta_i^j$.

For our exponential family

$$p(x;\vartheta) = \exp\big(\gamma(x) + f_i(x)\vartheta^i - \psi(\vartheta)\big), \tag{4.82}$$

with $\psi(\vartheta) = \log\int \exp(\gamma(x) + f_i(x)\vartheta^i)dx$, we also obtain a relationship between the expectation values η_i for the functions f_i and the expectation values η_{ij} of the products $f_i f_j$ (this is a special case of the general identity (4.73)):

$$\begin{aligned}
\eta_{ij} &= \int f_i(x) f_j(x) \exp\big(\gamma(x) + f_k(x)\vartheta^k - \psi(\vartheta)\big)\,dx \\
&= \exp\big(-\psi(\vartheta)\big)\frac{\partial^2}{\partial\vartheta^i\partial\vartheta^j}\int \exp\big(\gamma(x) + f_k(x)\vartheta^k\big)\,dx \\
&= \exp\big(-\psi(\vartheta)\big)\frac{\partial}{\partial\vartheta^i}\int f_j(x)\exp\big(\gamma(x) + f_k(x)\vartheta^k\big)\,dx \\
&= \exp\big(-\psi(\vartheta)\big)\frac{\partial}{\partial\vartheta^i}\big(\exp\big(\psi(\vartheta)\big)\eta_j\big) \\
&= \exp\big(-\psi(\vartheta)\big)\eta_j\frac{\partial}{\partial\vartheta^i}\int \exp\big(\gamma(x) + f_k(x)\vartheta^k\big)\,dx + \frac{\partial\eta_j}{\partial\vartheta^i} \\
&= \eta_i\eta_j + g_{ij},
\end{aligned} \tag{4.83}$$

see (4.49). We thus have the important result

Theorem 4.7

$$g_{ij} = \eta_{ij} - \eta_i\eta_j. \tag{4.84}$$

Thus, when we consider our coordinates ϑ^i as the weights given to the observables $f_i(x)$ on the basis of $\gamma(x)$, our metric $g_{ij}(\vartheta)$ is then simply the covariance matrix of those observables at the given weights or coordinates. □

To put this another way, if our observations yield not only the average values η_i of the functions f_i, but also the averages η_{ij} of the products $f_i f_j$—which in general are different from the products $\eta_i\eta_j$ of the average values of the f_i—, and we wish

to represent those in our probability distribution, we need to introduce additional parameters ϑ^{ij} and construct

$$p\big(x;\vartheta^i,\vartheta^{ij}\big)=\exp\big(\gamma(x)+f_i(x)\vartheta^i+f_i(x)f_j(x)\vartheta^{ij}-\psi(\vartheta)\big) \tag{4.85}$$

with

$$\vartheta^{ij}=\frac{\partial}{\partial\eta_{ij}}\varphi(\eta_i,\eta_{ij}), \tag{4.86}$$

$$\varphi(\eta_i,\eta_{ij})=\int\big(\log p(x;\eta_i,\eta_{ij})-\gamma(x)\big)p(x;\eta_i,\eta_{ij})\,dx \tag{4.87}$$

analogously to the above considerations.

Our definition (4.54) can also be interpreted in the sense that for fixed φ, $D(p\|q)$ can be taken as our potential $\psi(\vartheta)$ as $\varphi(\eta)$ is independent of ϑ and $\vartheta\cdot\eta$ is linear in ϑ so that neither of them enters into the second derivatives. Of course, this is (4.59). Here, $p=q$ then corresponds to $\eta=0$. From (4.57) and Theorem 4.3 (see (4.32) or (4.75)), we obtain the negative gradient flow for this potential $\psi(p)=D(p\|q)$,

$$\dot\vartheta^i=-g^{ij}\partial_j\psi(\vartheta)=-g^{ij}\eta_j, \tag{4.88}$$

or, since

$$\dot\eta_j=g_{ji}\dot\vartheta^i, \tag{4.89}$$

$$\dot\eta_j=-\eta_j. \tag{4.90}$$

This is a linear equation, and the solution moves along a straight line in the η-coordinates, i.e., along a geodesic for the dual connection ∇^* (see Proposition 2.5), towards the point $\eta=0$, i.e., $p=q$. In particular, the gradient flow for the Kullback–Leibler divergence D_{KL} moves on a straight line in the mixture coordinates.

A special case of an exponential family is a Gaussian one. Let $A=(A_{ij})_{i,j=1,\dots,n}$ be a symmetric, positive definite $n\times n$-matrix and let $\vartheta\in\mathbb{R}^n$ be a vector. The observables here are the components $x^1,\dots,x^n$ of $x\in\mathbb{R}^n$. We shall use the notation $\vartheta^t x=\sum_{i=1}^n\vartheta^i x^i$ and so on.

The Gaussian integral is

$$\begin{aligned}I(A,\vartheta):&=\int dx^1\cdots dx^n\exp\left(-\frac{1}{2}x^tAx+\vartheta^tx\right)\\&=\exp\left(\frac{1}{2}\vartheta^tA^{-1}\vartheta\right)\int\exp\left(-\frac{1}{2}y^tAy\right)dy^1\cdots dy^n\\&=\exp\left(\frac{1}{2}\vartheta^tA^{-1}\vartheta\right)\left(\frac{(2\pi)^n}{\det A}\right)^{\frac{1}{2}}\end{aligned} \tag{4.91}$$

(with the substitution $x = A^{-1}\vartheta + y$; note that the integral exists because A is positive definite). Thus, with

$$\psi(\vartheta) := \frac{1}{2}\vartheta^t A^{-1}\vartheta + \frac{1}{2}\log\frac{(2\pi)^n}{\det A}, \tag{4.92}$$

we have our exponential family

$$p(x;\vartheta) = \exp\left(-\frac{1}{2}x^t Ax + \vartheta^t x - \psi(\vartheta)\right). \tag{4.93}$$

Since ψ is a quadratic function of ϑ, all higher derivatives vanish, and in particular the connection $\Gamma^{(-1)} = 0$ and also the curvature tensor R vanishes, see (4.47), (4.46).

By (3.34), the metric is given by

$$g_{ij} = \frac{\partial^2}{\partial\vartheta^i\partial\vartheta^j}\psi = \left(A^{-1}\right)_{ij} \tag{4.94}$$

and is thus independent of ϑ. This should be compared with the results around (3.35). In contrast to those computations, here A is fixed, and not variable as σ^2. Equation (4.93) is equivalent to (3.35), noting that $\mu = \vartheta^1\sigma^2$ there. It can also be expressed in terms of moments

$$\begin{aligned}\left\langle x^{i_1}\cdots x^{i_m}\right\rangle &:= \mathbb{E}_p\left(x^{i_1}\cdots x^{i_m}\right)\\ &= \frac{\int x^{i_1}\cdots x^{i_m}\,\exp(-\frac{1}{2}x^tAx + \vartheta^t x)\,dx^1\cdots dx^n}{\int \exp(-\frac{1}{2}x^tAx + \vartheta^t x)\,dx^1\cdots dx^n}\\ &= \frac{1}{I(A,\vartheta)}\,\frac{\partial}{\partial\vartheta^{i_1}}\cdots\frac{\partial}{\partial\vartheta^{i_m}}I(A,\vartheta).\end{aligned} \tag{4.95}$$

In fact, we have

$$\left\langle x^i x^j\right\rangle - \left\langle x^i\right\rangle\left\langle x^j\right\rangle = \left(A^{-1}\right)_{ij} \tag{4.96}$$

by (4.91), in agreement with the general result of (4.84). (In the language of statistical physics, the second-order moment $\langle x^i x^j\rangle$ is also called a propagator.) For $\vartheta = 0$, the first-order moments $\langle x^i\rangle$ vanish because the exponential is then quadratic and therefore even.

4.4 Canonical Divergences

4.4.1 Dual Structures via Divergences

In this chapter we have studied the intrinsic geometry of statistical models, leading to the notion of a dualistic structure (g, ∇, ∇^*) on M. Of particular interest are

dualistic structures with torsion-free dual connections ∇ and ∇^*. As we have proved in Sect. 4.2, see Theorems 4.1 and 4.2, such a structure is equivalently given by (g, T), where g is a Riemannian metric and T a 3-symmetric tensor. This is an abstract and intrinsically defined version of the pair consisting of the Fisher metric and the Amari–Chentsov tensor. A natural approach to such a structure has been proposed by Eguchi [93] based on divergences.

Definition 4.7 Let M be a differentiable manifold. A *divergence* or *contrast function* on M is a real-valued smooth function $D : M \times M \to \mathbb{R}$, $(p, q) \mapsto D(p\|q)$, satisfying

$$D(p\|q) \geq 0, \qquad D(p\|q) = 0 \Leftrightarrow p = q, \tag{4.97}$$

and moreover,

$$V_p V_q D(p\|q)|_{p=q} > 0 \tag{4.98}$$

for any smooth vector field V on M that is non-zero at p. Given a divergence D, its *dual*

$$D^* : M \times M \to \mathbb{R}, \qquad D^*(p\|q) := D(q\|p) \tag{4.99}$$

also satisfies the conditions (4.97) and (4.98) and is therefore a divergence (contrast function) on M.

In Sect. 2.7, we have introduced and discussed various divergences defined on $\mathcal{M}_+(I)$ and $\mathcal{P}_+(I)$, respectively, including the relative entropy and the α-divergence. These divergences were tightly coupled with the Fisher metric and the α-connection. Furthermore, we have seen the tight coupling of a dually flat structure with the canonical divergence of Definition 4.6.

In what follows, we elaborate on how divergences induce dualistic structures. We use the following notation for a function on M defined by the value of the partial derivative in $M \times M$ with respect to the smooth vector fields $V_1, \ldots, V_n$, and $W_1, \ldots, W_m$ on M:

$$D(V_1 \cdots V_n \| W_1 \cdots W_m)(p) := (V_1)_p \cdots (V_n)_p (W_1)_q \cdots (W_m)_q D(p\|q)|_{p=q}.$$

We note that by (4.97), cf. (4.57)

$$V_q D(p\|q)|_{p=q} = V_p D(p\|q)|_{p=q} = 0. \tag{4.100}$$

By (4.98) the tensor $g^{(D)}$

$$g^{(D)}(V, W) := -D(V\|W) \tag{4.101}$$

is a Riemannian metric on M (compare with (4.59)). (Note that this expression is symmetric, that is, $D(V\|W) = D(W\|V)$, in contrast to $D(p\|q)$ which in general

is not symmetric.) In addition to this metric, the divergence induces a connection $\nabla^{(D)}$, given by

$$g^{(D)}\big(\nabla^{(D)}_X Y, Z\big) := -D(XY\|Z). \tag{4.102}$$

Applying (4.102) to the dual divergence D^* and noting that $g^{(D)} = g^{(D^*)}$, we obtain a connection $\nabla^{(D^*)}$ that satisfies

$$g^{(D)}\big(\nabla^{(D^*)}_X Y, Z\big) = -D^*(XY\|Z) = -D(Z\|XY) \tag{4.103}$$

for all smooth vector fields Z on M.

Theorem 4.8 *The two connections $\nabla^{(D)}$ and $\nabla^{(D^*)}$ are torsion-free and dual with respect to $g^{(D)} = g^{(D^*)}$.*

Proof We shall appeal to Theorem 4.2 and show that the tensor

$$T^{(D)}(X, Y, Z) := g^{(D)}\big(\nabla^{(D^*)}_X Y - \nabla^{(D)}_X Y, Z\big) \tag{4.104}$$

is a symmetric 3-tensor.

We first observe that $T^{(D)}$ is a tensor. It is linear in all its arguments, and for any $f \in C^\infty(M)$ we have

$$T^{(D)}(fX, Y, Z) = f\, T^{(D)}(X, Y, Z),$$

because, by Lemma B.1, the difference of two connections is a tensor. Moreover, by the symmetry of $D(X\|Y)$, we have

$$\begin{aligned} T^{(D)}(X, Y, Z) - T^{(D)}(Y, X, Z) &= D\big([X, Y]\big\|Z\big) - D\big(Z\big\|[X, Y]\big) \\ &= 0, \\ -T^{(D)}(X, Z, Y) + T^{(D)}(X, Y, Z) &= D(XY\|Z) + D(Y\|XZ) \\ &\quad - D(XZ\|Y) - D(Z\|XY) \\ &= X\big(D(Y\|Z) - D(Z\|Y)\big) \\ &= 0, \end{aligned}$$

and hence $T^{(D)}$ is symmetric. □

We say that a torsion-free dualistic structure (g, ∇, ∇^*) is *induced* by a divergence D if

$$g = g^{(D)}, \quad \nabla = \nabla^{(D)}, \quad \text{and} \quad \nabla^* = \nabla^{(D^*)}. \tag{4.105}$$

4.4.2 A General Canonical Divergence

We have the following inverse problem:

Given a torsion-free dualistic structure (g, ∇, ∇^*)*, is there always a corresponding divergence* D *that induces that structure?*

This question has been positively answered by Matumoto [173]. His result also follows from Lê's embedding Theorem 4.10 of Sect. 4.5 (see Corollary 4.5). On the one hand, it is quite satisfying to know that any torsion-free dualistic structure can be encoded by a divergence in terms of (4.105). For instance, in Sect. 2.7 we have shown that the Fisher metric $\mathfrak{g}$ together with the m- and e-connections can be encoded in terms of the relative entropy. More generally, $\mathfrak{g}$ together with the $\pm\alpha$-connections can be encoded by the α-divergence (see Proposition 2.13). Clearly, these divergences are special and have very particular meanings within information theory and statistical physics. The relative entropy generalizes Shannon information as reduction of uncertainty and the α-divergence is closely related to the Rényi entropy [223] and the Tsallis entropy [248, 249]. Indeed, these quantities are coupled with the underlying dualistic structure, or equivalently with $\mathfrak{g}$ and $\mathbf{T}$, in terms of partial derivatives, as formulated in Proposition 2.13. However, the relative entropy and the α-divergence are more strongly coupled with $\mathfrak{g}$ and $\mathbf{T}$ than expressed by this proposition. In general, we have many possible divergences that induce a given torsion-free dualistic structure, and there is no way to recover the relative entropy and the α-divergence without making stronger requirements than (4.105). On the other hand, in the dually flat case there is a distinguished divergence, the canonical divergence as introduced in Definition 4.6, which induces the underlying dualistic structure. This canonical divergence represents a natural choice among the many possible divergences that satisfy (4.105). It turns out that the canonical divergence recovers the Kullback–Leibler divergence in the case of a dually flat statistical model (see Eq. (4.78)), which highlights the importance of a canonical divergence. But which divergence should we choose, if the manifold is not dually flat? For instance in the general Riemannian case, where ∇ and ∇^* both coincide with the Levi-Civita connection of g, we do not necessarily have dual flatness. The need for a general canonical divergence in such cases has been highlighted in [35]. We can reformulate the above inverse problem as follows:

Given a torsion-free dualistic structure (g, ∇, ∇^*)*, is there always a corresponding "canonical" divergence* D *that induces that structure?*

We use quotation marks because it is not fully clear what we should mean by "canonical." Clearly, in addition to the basic requirement (4.105), such a divergence should coincide with those divergences that we had already identified as "canonical" in the basic cases of Sect. 2.7 and Definition 4.6. We therefore impose the following two

Requirements:

1. In the self-dual case where $\nabla = \nabla^*$ coincides with the Levi-Civita connection of g, the canonical divergence should simply be $D(p \parallel q) = \frac{1}{2}d^2(p, q)$.

2. We already have a canonical divergence in the dually flat case. Therefore, a generalized notion of a canonical divergence, which applies to any dualistic structure, should recover the canonical divergence of Definition 4.6 if applied to a dually flat structure.

In [23], Ay and Amari propose a canonical divergence that satisfies these requirements, following the gradient-based approach of Sect. 2.7.1 (see also the related work [14]). Assume that we have a manifold M equipped with a Riemannian metric g and an affine connection ∇. Here, we are only concerned with the local construction of a canonical divergence and assume that for each pair of points $q, p \in M$, there is a unique ∇-geodesic $\gamma_{q,p} : [0,1] \to M$ satisfying $\gamma_{q,p}(0) = q$ and $\gamma_{q,p}(1) = p$. This is equivalent to saying that for each pair of points q and p there is a unique vector $X(q,p) \in T_q M$ satisfying $\exp_q(X(q,p)) = p$, where exp denotes the exponential map associated with ∇ (see (B.39) and (B.40) in Appendix B). Given a point p, this allows us to consider the vector field $q \mapsto X(q,p)$, which we interpreted as difference field in Sect. 2.7.1 (see Fig. 2.5). Now, if this vector field is the (negative) gradient field of a function D_p, in the sense of Eq. (2.95), then $D_p(q) = D(p \,\|\, q)$ can be written as an integral along any path from q to p (see Eq. (2.96)). Choosing the ∇-geodesic $\gamma_{q,p}$ as a particular path, we obtain

$$D(p \,\|\, q) = \int_0^1 \left\langle X\big(\gamma_{q,p}(t), p\big), \dot{\gamma}_{q,p}(t) \right\rangle dt. \tag{4.106}$$

Since the geodesic connecting $\gamma_{q,p}(t)$ and p is a part of the geodesic connecting q and p, corresponding to the interval $[t,1]$, we have

$$X\big(\gamma_{q,p}(t), p\big) = (1-t)\,\dot{\gamma}_{q,p}(t). \tag{4.107}$$

Using also the reversed geodesic $\gamma_{p,q}(t) = \gamma_{q,p}(1-t)$, this leads to the following representations of the integral (4.106), which we use as a

Definition 4.8 We define the *canonical divergence* associated with a Riemannian metric g and an affine connection ∇ locally by

$$\begin{aligned} D(p \,\|\, q) &:= D^{\nabla}(p \,\|\, q) \\ &:= \int_0^1 (1-t) \left\| \dot{\gamma}_{q,p}(t) \right\|^2 dt \qquad (4.108) \\ &= \int_0^1 t \left\| \dot{\gamma}_{p,q}(t) \right\|^2 dt. \qquad (4.109) \end{aligned}$$

It is obvious from this definition that $D(p \,\|\, q) \geq 0$, and $D(p \,\|\, q) = 0$ if and only if $p = q$, which implies that D is actually a divergence. In the self-dual case, $\nabla = \nabla^*$ is the Levi-Civita connection with respect to g. In that case, the velocity field $\dot{\gamma}_{p,q}$ is parallel along the geodesic $\gamma_{p,q}$, and therefore

$$\left\| \dot{\gamma}_{p,q}(t) \right\|_{\gamma(t)} = \left\| \dot{\gamma}_{p,q}(0) \right\|_p = \left\| X(p,q) \right\|_p = d(p,q),$$

where $d(p,q)$ denotes the Riemannian distance between p and q. This implies that the canonical divergence has the following natural form:

$$D(p \parallel q) = \frac{1}{2} d^2(p,q)\,. \tag{4.110}$$

This shows that the canonical divergence satisfies Requirement 1 above. In the general case, where ∇ is not necessarily the Levi-Civita connection, we obtain the energy of the geodesic $\gamma_{p,q}$ as the symmetrized version of the canonical divergence:

$$\frac{1}{2}\big(D(p \parallel q) + D(q \parallel p)\big) = \frac{1}{2}\int_0^1 \big\|\dot{\gamma}_{p,q}(t)\big\|^2\, dt. \tag{4.111}$$

Remark 4.4

(1) We have defined the canonical divergence based on the affine connection ∇ of a given dualistic structure (g, ∇, ∇^*). We can apply the same definition to the dual connection ∇^* instead, leading to a canonical divergence $D^{(\nabla^*)}$. Note that, in general we do not have $D^{(\nabla^*)}(p \parallel q) = D^{(\nabla)}(q \parallel p)$, a property that is satisfied for the relative entropy and the α-divergence introduced in Sect. 2.7 (see Eq. (2.111)). In Sect. 4.4.3, we will see that this relation generally holds for dually flat structures. On the other hand, the mean

$$\bar{D}^{(\nabla)}(p \parallel q) := \frac{1}{2}\big(D^{(\nabla)}(p \parallel q) + D^{(\nabla^*)}(q \parallel p)\big)$$

always satisfies $\bar{D}^{(\nabla^*)}(p \parallel q) = \bar{D}^{(\nabla)}(q \parallel p)$, suggesting yet another definition of a canonical divergence [11, 23]. For a comparison of various notions of divergence duality, see also the work of Zhang [260].

(2) Motivated by Hooke's law, Henmi and Kobayashi [119] propose a canonical divergence that is similar to that of Definition 4.109.

(3) In the context of a dually flat structure (g, ∇, ∇^*) and its canonical divergence (4.54) of Definition 4.6, Fujiwara and Amari [100] studied gradient fields that are closely related to those given by (2.95).

In what follows, we are going to prove that the canonical divergence also satisfies Requirement 2.

4.4.3 Recovering the Canonical Divergence of a Dually Flat Structure

In the case of a dually flat structure (g, ∇, ∇^*), a canonical divergence is well-known, which is given by (4.54) in Definition 4.6. This is a distinguished divergence with many natural properties, which we have elaborated on in Sect. 4.3. We are now going to show that the divergence given by (4.109) coincides with the canonical

divergence of Definition 4.6 in the dually flat case. In order to do so, we consider ∇-affine coordinates $\vartheta = (\vartheta^1, \ldots, \vartheta^d)$ and ∇^*-affine coordinates $\eta = (\eta_1, \ldots, \eta_d)$. In the ϑ-coordinates, the ∇-geodesic connecting p with q has the form

$$\vartheta(t) := \vartheta(p) + t\big(\vartheta(q) - \vartheta(p)\big). \tag{4.112}$$

Hence, the velocity is constant

$$\dot{\vartheta}(t) = \vartheta(q) - \vartheta(p) =: z. \tag{4.113}$$

The canonical divergence of ∇ is given by

$$D^{(\nabla)}(p \,\|\, q) = \int_0^1 t\, g_{ij}\big(\vartheta(t)\big) z^i z^j \, dt. \tag{4.114}$$

Since $g_{ij}(\vartheta) = \partial_i \partial_j \psi(\vartheta)$ according to (4.33), where ψ is a strictly convex potential function, we have

$$\begin{aligned}
D^{(\nabla)}(p \,\|\, q) &= \int_0^1 t\, \partial_i \partial_j \psi\big(\vartheta(p) + t\, z\big) z^i z^j \, dt \\
&= \int_0^1 t\, \frac{d^2}{dt^2} \psi\big(\vartheta(t)\big)\, dt \\
&= -\int_0^1 \frac{d}{dt} \psi\big(\vartheta(t)\big)\, dt + \left[t\, \frac{d}{dt} \psi\big(\vartheta(t)\big) \right]_0^1 \\
&= \psi\big(\vartheta(p)\big) - \psi\big(\vartheta(q)\big) + \partial_i \psi\big(\vartheta(q)\big)\big(\vartheta^i(q) - \vartheta^i(p)\big) \\
&\overset{(4.48)}{=} \psi\big(\vartheta(p)\big) - \psi\big(\vartheta(q)\big) + \eta_i(q)\big(\vartheta^i(q) - \vartheta^i(p)\big) \\
&\overset{(4.39)}{=} \psi\big(\vartheta(p)\big) + \varphi\big(\eta(q)\big) - \vartheta^i(p)\, \eta_i(q).
\end{aligned} \tag{4.115}$$

Thus, we obtain exactly the definition (4.54) where $\psi(\vartheta(p))$ is abbreviated by $\psi(p)$ and $\varphi(\eta(q))$ by $\varphi(q)$.

We derived the canonical divergence D for the affine connection ∇ based on (4.109). We now use the same definition, in order to derive the canonical divergence $D^{(\nabla^*)}$ of the dual connection ∇^*. The ∇^*-geodesic connecting p with q has the following form in the η-coordinates:

$$\eta(t) = \eta(p) + t\big(\eta(q) - \eta(p)\big). \tag{4.116}$$

Hence, the velocity is constant

$$\dot{\eta}(t) = \eta(q) - \eta(p) =: z^*. \tag{4.117}$$

The divergence $D^{(\nabla^*)}$ is given by

$$D^{(\nabla^*)}(p \,\|\, q) = \int_0^1 t\, g^{ij}\big(\eta(t)\big) z_i^* z_j^* \, dt. \tag{4.118}$$

Since $g^{ij}(\boldsymbol{\eta}) = \partial^i \partial^j \varphi(\boldsymbol{\eta})$, we have

$$\begin{aligned} D^{(\nabla^*)}(p \,\|\, q) &= \int_0^1 t\, \partial^i \partial^j \varphi\big(\eta(p) + t z^*\big) z_i^* z_j^*\, dt \\ &= \int_0^1 t\, \frac{d^2}{dt^2} \varphi\big(\eta(t)\big)\, dt \\ &= -\int_0^1 \frac{d}{dt} \varphi\big(\eta(t)\big)\, dt + \left[t\, \frac{d}{dt} \varphi\big(\eta(t)\big) \right]_0^1 \\ &= \varphi\big(\eta(p)\big) - \varphi\big(\eta(q)\big) + \partial^i \varphi\big(\eta(q)\big)\big(\eta_i(q) - \eta_i(p)\big) \\ &\overset{(4.51)}{=} \varphi\big(\eta(p)\big) - \varphi\big(\eta(q)\big) + \vartheta^i(q)\big(\eta_i(q) - \eta_i(p)\big) \\ &\overset{(4.39)}{=} \varphi\big(\eta(p)\big) + \psi\big(\vartheta(q)\big) - \vartheta^i(q)\, \eta_i(p). \end{aligned}$$

A comparison with (4.115) shows

$$D^{(\nabla^*)}(p \,\|\, q) = D^{(\nabla)}(q \,\|\, p). \tag{4.119}$$

This proves that ∇ and ∇^* give the same canonical divergence except that p and q are interchanged because of the duality. Instances of this general relation in the dually flat case are given by the α-divergences, see (2.111).

4.4.4 Consistency with the Underlying Dualistic Structure

We have defined our canonical divergence D based on a metric g and an affine connection ∇ (see (4.109)). It is natural to require that the corresponding dualistic structure (g, ∇, ∇^*) is encoded by this divergence in terms of (4.105), or, in local coordinates $\xi = (\xi^1, \ldots, \xi^n)$, by

$$g_{ij} = -D(\partial_i \| \partial_j), \qquad \Gamma_{ijk} = -D(\partial_i \partial_j \| \partial_k), \qquad \Gamma^*_{ijk} = -D(\partial_k \| \partial_i \partial_j). \tag{4.120}$$

Since the geometry is determined by the derivatives of $D(p \,\|\, q)$ at $p = q$, we consider the case where p and q are close to each other, that is

$$z^i = \xi^i(q) - \xi^i(p) \tag{4.121}$$

is small for all i. We evaluate the divergence by Taylor expansion up to $O(\|z\|^3)$. Note that $X(p, q)$ is of order $\|z\|$.

Proposition 4.1 *When $\|z\| = \|\boldsymbol{\xi}(q) - \boldsymbol{\xi}(p)\|$ is small, the canonical divergence* (4.109) *is expanded as*

$$D(p \,\|\, q) = \frac{1}{2}\, g_{ij}(p)\, z^i z^j + \frac{1}{6}\, \Lambda_{ijk}(p)\, z^i z^j z^k + O\big(\|z\|^4\big) \tag{4.122}$$

where

$$\Lambda_{ijk} = 2\,\partial_i g_{jk} - \Gamma_{ijk}. \tag{4.123}$$

Proof The Taylor series expansion of the local coordinates $\boldsymbol{\xi}(t)$ of the geodesic $\gamma_{p,q}(t)$ is given by

$$\xi^i(t) = \xi^i(p) + t\,X^i - \frac{t^2}{2}\,\Gamma^i_{jk}\,X^j X^k + O\big(t^3\|X\|^3\big), \tag{4.124}$$

where $X(p,q) = X^i\partial_i$. This follows from $\gamma_{p,q}(0) = p$, $\dot\gamma_{p,q}(0) = X(p,q)$, and $\nabla_{\dot\gamma_{p,q}}\dot\gamma_{p,q} = 0$ (see the geodesic equation (B.38) in Appendix B). When $\|z\|$ is small, X is of order $O(\|z\|)$. Hence, we regard (4.124) as a Taylor expansion with respect to X and $t \in [0,1]$ when z is small. When $t = 1$, we have

$$z^i = X^i - \frac{1}{2}\,\Gamma^i_{jk}X^j\,X^k + O\big(\|X\|^3\big). \tag{4.125}$$

This in turn gives

$$X^i = z^i + \frac{1}{2}\,\Gamma^i_{jk}\,z^j z^k + O\big(\|z\|^3\big). \tag{4.126}$$

We calculate $D(p\,\|\,q)$ by using the representation (4.109). The velocity at t is given as

$$\dot\xi^i(t) = X^i - t\,\Gamma^i_{jk}\,X^j X^k + O\big(t^2\|X\|^3\big) \tag{4.127}$$

$$= z^i + \frac{1}{2}(1-2t)\,\Gamma^i_{jk}\,z^j z^k + O\big(t^2\|z\|^3\big). \tag{4.128}$$

We also use

$$g_{ij}\big(\gamma_{p,q}(t)\big) = g_{ij}(p) + t\,\partial_k g_{ij}(p)\,z^k + O\big(t^2\|z\|^2\big). \tag{4.129}$$

Collecting these terms, we obtain

$$\begin{aligned} &t\,g_{ij}\big(\gamma_{p,q}(t)\big)\,\dot\xi^i(t)\,\dot\xi^j(t) \\ &\quad = t\,g_{ij}(p)\,z^i z^j + \big\{t^2\,\partial_i g_{jk}(p) + (-2t^2+t)\Gamma_{ijk}(p)\big\}\,z^i z^j z^k + O\big(t^3\|z\|^4\big). \end{aligned}$$

By integration, we have

$$D(p\,\|\,q) = \int_0^1 t\,g_{ij}\big(\gamma_{p,q}(t)\big)\,\dot\xi^i(t)\,\dot\xi^j(t)\,dt \tag{4.130}$$

$$= \frac{1}{2}\,g_{ij}(p)\,z^i z^j + \frac{1}{6}\,\Lambda_{ijk}(p)\,z^i z^j z^k + O\big(\|z\|^4\big), \tag{4.131}$$

where indices of Λ_{ijk} are symmetrized by means of multiplication with $z^i z^j z^k$. □

Theorem 4.9 ([23, Theorem 1]) *Let* (g, ∇, ∇^*) *be a torsion-free dualistic structure. Then the canonical divergence* $D^{(\nabla)}(p \,\|\, q)$ *of Definition* 4.8 *is consistent with* (g, ∇, ∇^*) *in the sense of* (4.105).

Proof Without loss of generality, we restrict attention to the connection ∇ and consider only the canonical divergence $D = D^{(\nabla)}$. By differentiating equation (4.122) with respect to $\boldsymbol{\xi}(p)$, we obtain

$$\begin{aligned}&\partial_i D(p \,\|\, q)\\ &\quad = \frac{1}{2}\partial_i g_{jk}(p)\, z^j z^k - g_{ij}(p)\, z^j - \frac{1}{2}\Lambda_{ijk}(p)\, z^j z^k + O\big(\|z\|^3\big), \qquad (4.132)\\ &\partial_i \partial_j D(p \,\|\, q)\\ &\quad = \frac{1}{2}\partial_i \partial_j g_{kl}(p)\, z^k z^l - 2\,\partial_i g_{jk}(p)\, z^k + g_{ij} + \Lambda_{ijk}(p)\, z^k + O\big(\|z\|^2\big). \qquad (4.133)\end{aligned}$$

We need to symmetrize the indexed quantities of the RHS with respect to i, j. By evaluating $\partial_i \partial_j D(p \,\|\, q)$ at $p = q$, i.e., $z = 0$, we have

$$g_{ij}^{(D)} = -D(\partial_i \,\|\, \partial_j) = D(\partial_i \partial_j \,\|\, \cdot) = g_{ij}, \qquad (4.134)$$

proving that the Riemannian metric derived from D is the same as the original one. We further differentiate (4.133) with respect to $\boldsymbol{\xi}(q)$ and evaluate it at $p = q$. This yields

$$\begin{aligned}\Gamma_{ijk}^{(D)} &= -D(\partial_i \partial_j \,\|\, \partial_k) = 2\,\partial_i g_{jk} - \Lambda_{ijk}\\ &= \Gamma_{ijk}. \qquad (4.135)\end{aligned}$$

Hence, the affine connection $\nabla^{(D)}$ derived from D coincides with the original affine connection ∇, given that ∇ is assumed to be torsion-free. □

4.5 Statistical Manifolds and Statistical Models

In Sect. 4.2 we have analyzed the notion of a statistical manifold (Definition 4.2), introduced by Lauritzen [160] as a formalization of the notion of a statistical model and showed the equivalence between a statistical manifold and a manifold provided with a torsion-free dualistic structure. In this section, we shall analyze the relation between the Lauritzen question, whether any statistical manifold is induced by a statistical model, and the existence of an isostatistical immersion (Definition 4.9, Lemma 4.8). The main theorem of this section asserts that any statistical manifold is induced by a statistical model (Theorem 4.10). The proof will occupy Sects. 4.5.3, 4.5.4. In Sect. 4.5.2 we shall study some simple obstructions to the existence of isostatistical immersions, which are helpful for understanding the strategy of the proof of Theorem 4.10. Finally, in Sect. 4.5.5 we shall strengthen our

immersion theorem by showing that we can embed any compact statistical manifold (possibly with boundary) into $(\mathcal{P}_+([N]), \mathfrak{g}, \mathbf{T})$ for some finite N (Theorem 4.11). An analogous (but weaker) embedding statement for non-compact statistical manifolds will also follow.

All manifolds in this section are assumed to have finite dimension.

4.5.1 Statistical Manifolds and Isostatistical Immersions

In Definition 4.2, we have introduced Lauritzen's notions of a statistical manifold, that is, a manifold M equipped with a Riemannian metric g and a 3-symmetric tensor T.

Remark 4.5 As in the Riemannian case [195], we call a smooth manifold M provided with a statistical structure (g, T) that are C^k-differentiable a *C^k-statistical manifold*. Occasionally, we shall drop "C^k" before "statistical manifold" if there is no danger of confusion.

The Riemannian metric g generalizes the notion of the Fisher metric and the 3-symmetric tensor T generalizes the notion of the Amari–Chentsov tensor. Statistical manifolds also encompass the class of differentiable manifolds supplied with a divergence as we have introduced in Definition 4.7.

As follows from Theorem 4.1, a torsion-free dualistic structure (g, ∇, ∇^*) defines a statistical structure. Conversely, by Theorem 4.2 any statistical structure (g, T) on M defines a torsion-free dualistic structure (g, ∇, ∇^*) by (4.15) and the duality condition $\nabla_A B + \nabla^*_A B = 2\nabla^{(0)}_A B$, where $\nabla^{(0)}$ is the Levi-Civita connection, see (4.16). As Lauritzen remarked, the representation (M, g, T) is practical for mathematical purposes, because as a symmetric 3-tensor, T has simpler transformational properties than ∇ [160].

Lauritzen raised in [160, §4, p. 179] the question of whether any statistical manifold is induced by a statistical model. More precisely, he wrote after giving the definition of a statistical manifold: "The above defined notion could seem a bit more general than necessary, in the sense that some Riemannian manifolds with a symmetric trivalent tensor T might not correspond to a particular statistical model." Turning this positively, the question is whether for a given (C^k)-statistical manifold (M, g, T) we can find a sample space Ω and a $(C^{k+1}$-)smooth family of probability distributions $p(x; \xi)$ on Ω with parameter $\xi \in M$ such that (cf. (4.2), (4.6))

$$g(\xi; V_1, V_2) = \mathbb{E}_{p(\cdot;\xi)}\left(\frac{\partial}{\partial V_1} \log p(\cdot; \xi) \frac{\partial}{\partial V_2} \log p(\cdot; \xi)\right), \tag{4.136}$$

$$T(\xi; V_1, V_2, V_3) = \mathbb{E}_{p(\cdot;\xi)}\left(\frac{\partial}{\partial V_1} \log p(\cdot; \xi) \frac{\partial}{\partial V_2} \log p(\cdot; \xi) \frac{\partial}{\partial V_3} \log p(\cdot; \xi)\right). \tag{4.137}$$

If (4.136) and (4.137) hold, we shall call the function $p(x;\xi)$ *a probability density* for g and T, see also Remark 3.7. We regard the Lauritzen question as the existence question of a probability density for the tensors g and T on a statistical manifold (M, g, T).

Our approach in solving the Lauritzen question is to reduce the existence problem of probability densities on a statistical manifold to an immersion problem of statistical manifolds.

Definition 4.9 A smooth (resp. C^1) map h from a smooth (resp. C^0) statistical manifold (M_1, g_1, T_1) to a statistical manifold (M_2, g_2, T_2) will be called *an isostatistical immersion* if h is an immersion of M_1 into M_2 such that $g_1 = h^*(g_2)$, $T_1 = h^*(T_2)$.

Of course, the notion of an isostatistical immersion is an intrinsic counterpart of that of a sufficient statistic. In fact, a sufficient statistic as defined in (5.1) or in Definition 5.8 is characterized by the fact that it preserves the Fisher metric and the Amari–Chentsov tensor, see Theorems 5.5 and 5.6.

Lemma 4.6 *Assume that $h : (M_1, g_1, T_1) \to (M_2, g_2, T_2)$ is an isostatistical immersion. If there exist a measure space Ω and a function $p(x;\xi_2) : \Omega \times M_2 \to \mathbb{R}$ such that p is a probability density for the tensors g_2 and T_2 then $h^*(p)(x;\xi_1) := p(x;h(\xi_1))$ is a probability density for g_1 and T_1.*

Proof Since h is an isostatistical immersion, we have

$$\begin{aligned} g_1(\xi; V_1, V_2) &= g_2\big(h(\xi); h_*(V_1), h_*(V_2)\big) \\ &= \int_\Omega \frac{\partial}{\partial h_*(V_1)} \log p\big(x; h(\xi)\big) \frac{\partial}{\partial h_*(V_2)} \log p\big(x; h(\xi)\big)\; p\big(x; h(\xi)\big)\, dx \\ &= \mathbb{E}_{h^*(p)}\left(\frac{\partial}{\partial V_1} \log h^*(p)(\cdot;\xi) \frac{\partial}{\partial V_2} \log h^*(p)(\cdot;\xi) \right). \end{aligned}$$

Thus $h^*(p)$ is a probability density for g_1. In the same way, $h^*(p)$ is a probability density for T_1. This completes the proof of Lemma 4.6. □

Example 4.1

(1) The statistical manifold[5] $(\mathcal{P}_+([n]), \mathfrak{g}, \mathbf{T})$ has a natural probability density $p \in C^\infty([n] \times \mathcal{P}_+([n]))$ defined by $p(x;\xi) := \xi(x)$.
(2) Let g_0 denote the Euclidean metric on $\mathbb{R}^n$ as well as its restriction to the positive sector $\mathbb{R}^n_+$. Let $\{e_i\}$ be an orthonormal basis of $\mathbb{R}^n$. Denote by $\{x_i\}$ the dual basis

[5] $\mathcal{P}_+([n])$ is the interior of the probability simplex Σ^{n-1}.

of $(\mathbb{R}^n)^*$. Set

$$T^* := \sum_{i=1}^{n} \frac{2dx_i^3}{x_i}.$$

Then $(\mathbb{R}^n_+, g_0, T^*)$ is a statistical manifold. By Proposition 2.1 the embedding

$$\pi^{1/2} : \mathcal{P}_+\big([n]\big) \to \mathbb{R}^n_+, \qquad \xi = \sum_{i=1}^{n} p(i;\xi)\,\delta^i \mapsto 2\sum_{i=1}^{n} \sqrt{p(i;\xi)}\,e_i,$$

where δ^i is the Dirac measure concentrated at $i \in [n]$, is an isometric embedding of the Riemannian manifold $(\mathcal{P}_+([n]), \mathfrak{g})$ into the Riemannian manifold $(\mathbb{R}^n_+, g_0)$.

Now let us compute $(\pi^{1/2})^*(T^*)$. Since $x_i(\pi^{1/2}(\xi)) = 2\sqrt{p(i;\xi)}$, we obtain

$$\begin{aligned}\big(\pi^{1/2}\big)^*\big(T^*\big)(\xi; V_1, V_1, V_1) &= \sum_{i=1}^{n} 2\frac{(\partial_V(2\sqrt{p(i;\xi)}))^3}{2\sqrt{p(i;\xi)}} \\ &= \sum_{i=1}^{n} \frac{(\partial_V p(i;\xi))^3}{p(i;\xi)^2} \\ &= \sum_{i=1}^{n} \big(\partial_V \log p(i;\xi)\big)^3 p(i;\xi) = \mathbf{T}(\xi; V_1, V_1, V_1).\end{aligned}$$

This shows that $\pi^{1/2}$ is an isostatistical immersion of the statistical manifold $(\mathcal{P}_+([n]), \mathfrak{g}, \mathbf{T})$ into the statistical manifold $(\mathbb{R}^n_+, g_0, T^*)$.

Now we formulate our answer to Lauritzen's question.

Theorem 4.10 (Existence of isostatistical immersions (cf. [162])) *Any smooth* (*resp.* C^0) *compact statistical manifold* (M, g, T) (*possibly with boundary*) *admits an isostatistical immersion into the statistical manifold* $(\mathcal{P}_+([N]), \mathfrak{g}, \mathbf{T})$ *for some finite number* N. *Any non-compact statistical manifold* (M, g, T) *admits an immersion* I *into the space* $\mathcal{P}_+(\mathbb{N})$ *of all positive probability measures on the set* $\mathbb{N}$ *of all natural numbers such that* g *is equal to the Fisher metric defined on* $I(M)$ *and* T *is equal to the Amari–Chentsov tensor defined on* $I(M)$. *Hence any statistical manifold is a statistical model.*

This theorem then links the abstract differential geometry developed in this chapter with the functional analysis established in Chap. 3.

This result will be proved in Sects. 4.5.3, 4.5.4. In the remainder of this section, compact manifolds may have non-empty boundary.

Since the statistical structure $(\mathfrak{g}, \mathbf{T})$ on $\mathcal{P}_+([N])$ is defined by the canonical divergence [16, Theorem 3.13], see also Proposition 2.13, we obtain from Theorem 4.10 the following result due to Matumoto.

Corollary 4.5 (Cf. [173, Theorem 1]) *For any statistical manifold* (M, g, T) *we can find a divergence* D *of* M *which defines* g *and* T *by the formulas* (4.101), (4.103).

Proof Assume that a statistical manifold (M, g, T) is compact. By Theorem 4.10 (M, g, T) admits an isostatistical immersion I into a statistical manifold $(\mathcal{P}_+([N]), \mathfrak{g}, \mathbf{T})$ for some finite number N. This statistical manifold is compatible with the KL-divergence (compare with Proposition 2.13 for $\alpha = -1$), which implies that the contrast function $M \times M \to \mathbb{R}$, $(p, q) \mapsto D_{KL}(I(p) \,\|\, I(q))$, is compatible with (M, g, T).

Now assume that (M, g, T) is a non-compact manifold. It is known that (M, g, T) admits a countable locally finite open cover $\{U_i\}$ such that each U_i is a subset of a compact subset in M. By the argument above, each statistical manifold $(U_i, g_{|U_i}, T_{|U_i})$ admits a compatible divergence D_i. With a partition of unity we can glue the divergence functions D_i and define a smoothly extended contrast function on $M \times M$, thereby following Matumoto's final step of his construction [173]. □

4.5.2 Monotone Invariants of Statistical Manifolds

Before going to develop a strategy for a proof of Theorem 4.10 we need to understand what could be an obstruction for the existence of an isostatistical immersion between statistical manifolds.

Definition 4.10 Let $K(M, e)$ denote the category of statistical manifolds M with morphisms being embeddings. A functor of this category is called *a monotone invariant* of statistical manifolds.

Remark 4.6 Since any isomorphism between statistical manifolds defines an invertible isostatistical immersion, any monotone invariant is an invariant of statistical manifolds.

In this subsection we study some simple monotone invariants of statistical manifolds and refer the reader to [163] for more sophisticate monotone invariants.

Let $f : (M_1, g_1, T_1) \to (M_2, g_2, T_2)$ be a statistical immersion. Then for any $x \in M_1$ the differential $df : (T_x M_1, g_1(x), T_1(x)) \to (T_{f(x)} M_2, g_2(f(x)), T_2(f(x)))$ defines an isostatistical immersion of the statistical manifold $(T_x M_1, g_1(x), T_1(x))$ into the statistical manifold $(T_{f(x)} M_2, g_2(f(x)), T_2(f(x)))$.

A statistical manifold $(\mathbb{R}^m, g, T)$ is called a *linear statistical manifold* if g and T are constant tensors.

Thus we start our study by investigating functors of the subcategory $K_l(M, e)$ of linear statistical manifolds $M = (\mathbb{R}^n, g, T)$. Such a functor will be called a *linear monotone invariant*.

Given a linear statistical manifold $M = (\mathbb{R}^n, g, T)$ we set

$$\mathcal{M}^3(T) := \max_{|x|=1,|y|=1,|z|=1} T(x, y, z),$$

$$\mathcal{M}^2(T) := \max_{|x|=1,|y|=1} T(x, y, y),$$

$$\mathcal{M}^1(T) := \max_{|x|=1} T(x, x, x).$$

Clearly, we have

$$0 \leq \mathcal{M}^1(T) \leq \mathcal{M}^2(T) \leq \mathcal{M}^3(T).$$

Proposition 4.2 *The comasses $\mathcal{M}^i$, $i \in [1, 3]$, are non-negative linear monotone invariants, which vanish if and only if $T = 0$.*

Proof Clearly $\mathcal{M}^i(T) \geq 0$ for $i = 1, 2, 3$. Now we are going to show that $\mathcal{M}^1$ vanishes at T only if $T = 0$. Observe that $\mathcal{M}^1 = 0$ if and only if $T(x, x, x) = 0$ for all $x \in \mathbb{R}^n$. Writing T in coordinate expression $T(x, y, z) = \sum a_{ijk} x^i y^j z^k$, we note that $T(x, x, x) = 0$ if and only if $T = 0$, since T is symmetric.

Next we shall show that $\mathcal{M}^i(T)$ is a linear monotone invariant for $i = 1, 2, 3$. Assume that e is a linear embedding $(\mathbb{R}^n, g, T)$ into $(\mathbb{R}^m, \bar{g}, \bar{T})$. Then T is a restriction of the 3-symmetric tensor $\bar{T}$. Hence we have

$$\mathcal{M}^i(T) \leq \mathcal{M}^i(\bar{T}), \quad \text{for } i = 1, 2, 3.$$

This implies that $\mathcal{M}^i$ are linear monotone invariants. □

Thus $\mathcal{M}^i$ is a functor from the category $K_l(M, e)$ of linear statistical manifolds to the category $(\mathbb{R}, \leq)$ of real numbers with morphism being the relation "$\leq$".

Using the linear statistical monotone invariant $\mathcal{M}^1$, we define for any statistical manifold (M, g, T) the following number

$$\mathcal{M}_0^1(T) := \sup_{x \in M} \mathcal{M}^1\big(T(x)\big).$$

Since the restriction of $\mathcal{M}_0^1$ to the subcategory $K_l(M, e)$ is equal to $\mathcal{M}^1$, we shall abbreviate $\mathcal{M}_0^1$ as $\mathcal{M}^1$, if there is no danger of confusion.

By Proposition 4.2 we obtain immediately

Proposition 4.3 *The comass $\mathcal{M}^1$ is a non-negative monotone invariant, which vanishes if and only if $T = 0$.*

Thus $\mathcal{M}^1$ is a functor from the category $K(M, e)$ of statistical manifolds to the category $(\mathbb{R}, \leq)$ of real numbers with morphism being the relation "$\leq$".

In what follows we shall show two applications of the monotone invariant $\mathcal{M}^1$. Proposition 4.4 below will guide our strategy of the proof of Theorem 4.10 in the

later part of this section. The equality (4.138) below suggests that the statistical manifold $(\mathcal{P}_+([N]), \mathfrak{g}, \mathbf{T})$ might be a good candidate for a target of isostatistical embeddings of statistical manifolds.

Proposition 4.4 *A statistical line* $(\mathbb{R}, g_0, T)$ *can be embedded into a linear statistical manifold* $(\mathbb{R}^N, g_0, T')$ *if and only if* $\mathcal{M}^1(T) \leq \mathcal{M}^1(T')$.

Proof The "only" assertion of Proposition 4.4 is obvious. Now we shall show that we can embed $(\mathbb{R}, g_0, T)$ into $(\mathbb{R}^N, g_0, T')$ if we have $\mathcal{M}^1(T) \leq \mathcal{M}^1(T')$. We note that $T'(v, v, v)$ defines an anti-symmetric function on the sphere $S^{N-1}(|v| = 1) \subseteq \mathbb{R}^N$. Thus there is a point $v \in S^{N-1}$ such that $T'(v, v, v) = \mathcal{M}^1(T)$. Clearly, the line $\{t \cdot v | t \in \mathbb{R}\}$ defines the required embedding. □

The example of the family of normal distributions treated on page 132 yields the normal Gaussian statistical manifold $(\Gamma^2, \mathfrak{g}, \mathbf{T})$. Recall that Γ^2 is the upper half of the plane $\mathbb{R}^2(\mu, \sigma)$, $\mathfrak{g}$ is the Fisher metric and $\mathbf{T}$ is the Amari–Chentsov tensor associated to the probability density

$$p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right),$$

where $x \in \mathbb{R}$.

Proposition 4.5 *The statistical manifold* $(\mathcal{P}_+([N]), \mathfrak{g}, \mathbf{T})$ *cannot be embedded into the Cartesian product of m copies of the normal Gaussian statistical manifold* $(\Gamma^2, \mathfrak{g}, \mathbf{T})$ *for any* $N \geq 4$ *and finite m*.

Proof By Lemma 4.9 proved below, we obtain for $N \geq 4$

$$\mathcal{M}^1\big(\mathcal{P}_+\big([N]\big), \mathfrak{g}, \mathbf{T}\big) = \infty. \tag{4.138}$$

(In chronological order Lemma 4.9 has been invented for proving (4.138). We decide to move Lemma 4.9 to a later subsection for a better understanding of the proof of Theorem 4.10.)

The tensor $\mathfrak{g}$ of the manifold $(\Gamma^2, \mathfrak{g}, \mathbf{T})$ has been computed on p. 132. $\mathbf{T}$ can be computed analogously. (The formulas are due to [160].)

$$\mathfrak{g}\left(\frac{\partial}{\partial\mu}, \frac{\partial}{\partial\mu}\right) = \frac{1}{\sigma^2},$$

$$\mathfrak{g}\left(\frac{\partial}{\partial\mu}, \frac{\partial}{\partial\sigma}\right) = 0,$$

$$\mathfrak{g}\left(\frac{\partial}{\partial\sigma}, \frac{\partial}{\partial\sigma}\right) = \frac{2}{\sigma^2},$$

$$\mathbf{T}\left(\frac{\partial}{\partial\mu}, \frac{\partial}{\partial\mu}, \frac{\partial}{\partial\mu}\right) = 0 = \mathbf{T}\left(\frac{\partial}{\partial\mu}, \frac{\partial}{\partial\sigma}, \frac{\partial}{\partial\sigma}\right),$$

$$\mathbf{T}\left(\frac{\partial}{\partial\mu},\frac{\partial}{\partial\mu},\frac{\partial}{\partial\sigma}\right)=\frac{2}{\sigma^3},\qquad \mathbf{T}\left(\frac{\partial}{\partial\sigma},\frac{\partial}{\partial\sigma},\frac{\partial}{\partial\sigma}\right)=\frac{8}{\sigma^3}.$$

$\mathcal{M}^1(\mathbb{R}^2(\mu,\sigma))<\infty$. It follows that the norm $\mathcal{M}^1$ of a direct product of finite copies of $\mathbb{R}^2(\mu,\sigma)$ is also finite. Since $\mathcal{M}^1$ is a monotone invariant, the space $\mathcal{P}_+([N])$ cannot be embedded into the direct product of m copies of the normal Gaussian statistical manifold for any $N\geq 4$ and any $m<\infty$. □

After investigating obstructions to the existence of isostatistical immersions, we now return to the proof of Theorem 4.10.

4.5.3 *Immersion of Compact Statistical Manifolds into Linear Statistical Manifolds*

We denote by T_0 the "standard" 3-tensor on $\mathbb{R}^m$:

$$T_0=\sum_{i=1}^{m}dx_i^3.$$

One important class of linear statistical manifolds are those of the form $(\mathbb{R}^m, g_0, A\cdot T_0)$ where $A\in\mathbb{R}$.

In the first step of our proof of Theorem 4.10 we need the following

Proposition 4.6 *Let (M^m,g,T) be a compact smooth (resp. C^0) statistical manifold. Then there exist numbers $N\in\mathbb{N}^+$ and $A>0$ as well as a smooth (resp. C^1) immersion $f:(M^m,g,T)\to(\mathbb{R}^N,g_0,A\cdot T_0)$ such that $f^*(g_0)=g$ and $f^*(A\cdot T_0)=T$.*

The constant A enters into Proposition 4.6 to ensure that the monotone invariants $\mathcal{M}^i(\mathbb{R}^N,g_0,A\cdot T_0)$ can be sufficiently large.

Our proof of Proposition 4.6 uses the Nash immersion theorem, the Gromov immersion theorem and an algebraic trick. We also note that, unlike the Riemannian case, Proposition 4.6 does not hold for non-compact statistical manifolds. For example, for any $n\geq 4$, the statistical manifold $(\mathcal{P}_+([n]),\mathfrak{g},\mathbf{T})$ does not admit an isostatistical immersion to any linear statistical manifold. This follows from the theory of monotone invariants of statistical manifolds developed in [163], where we showed that the monotone invariant $\mathcal{M}^1$ of $(\mathcal{P}_+([n]),\mathfrak{g},\mathbf{T})$ is infinite, and the monotone invariant $\mathcal{M}^1$ of any linear statistical manifold is finite [163, §3.6, Proposition 4.2], or see the proof of Proposition 4.5 above.

Nash's embedding theorem ([195, 196]) *Any smooth (resp. C^0) Riemannian manifold (M^n,g) can be isometrically embedded into $(\mathbb{R}^{N(n)},g_0)$ for some $N(n)$ depending on n and on the compactness property of M^n.*

Remark 4.7 One important part in the proof of the Nash embedding theorem for C^0-Riemannian manifolds (M^n, g) is his immersion theorem ([195, Theorems 2, 3, 4, p. 395]), where the dimension $N(n)$ of the target Euclidean space depends on the dimension $W(n)$ of Whitney's immersion [256] of M^n into $\mathbb{R}^{W(n)}$ and hence $N(n)$ can be chosen not greater than $\min(W(n)+2, 2n-1)$. If M^n is compact (resp. non-compact), then Nash proved that (M^n, g) can be C^1-isometrically embedded in $(\mathbb{R}^{2n}, g_0)$ (resp. $(\mathbb{R}^{2n+1}, g_0)$). Nash's results on C^1-embedding have been sharpened by Kuiper in 1955 by weakening the dependency of $N(n)$ on the dimension of the Whitney embedding of M^n into $\mathbb{R}^{W'(n)}$ [154].

Nash proved his isometric embedding theorem for smooth (actually C^k, $k \geq 3$) Riemannian manifolds (M^n, g) in 1956 for $N(n) = (n/2)(3n+11)$ if M^n is compact [196, Theorem 2, p. 59], and for $N(n) = \frac{3}{2}n^3 + 7n^2 + \frac{5}{2}n$ if M^n is non-compact [196, Theorem 3, p. 63]. The best upper bound estimate $N(n) \leq \max\{n(n+5)/2, n(n+3)/2+5\}$ for the smooth (compact or non-compact) case has been obtained by Günther in 1989, using a different proof strategy [114, pp. 1141, 1142]. Whether the isometric immersion theorem is also true for C^2-Riemannian manifolds remains unknown. The problem is that the C^0-case and the case of $C^{2+\alpha}$, $\alpha > 0$ (including the smooth case) are proved by different methods. The C^0-case is proved by a limit process, where we have control on the first derivatives but no control on higher derivatives of an immersion. So the limit immersion may not be smooth though the original immersion is smooth. On the other hand, the $C^{2+\alpha}$-case is proved by the famous Nash implicit function theorem, which was developed later by Gromov in [112, 113].

Gromov's immersion theorem ([113, 2.4.9.3' (p. 205), 3.1.4 (p. 234)]) *Suppose that M^m is given with a smooth (resp. C^0) symmetric 3-form T. Then there exists a smooth immersion $f : M^m \to \mathbb{R}^{N_1(m)}$ with $N_1(m) = 3(n + \binom{n+1}{2} + \binom{n+2}{3})$ (resp. a C^1-immersion f with $N_1(m) = (m+1)(m+2)/2 + m$) such that $f^*(T_0) = T$.*

Proof of Proposition 4.6 First we choose an immersion $f_1 : (M^m, g, T) \to (\mathbb{R}^{N_1(m)}, g_0, T_0)$ such that

$$f_1^*(T_0) = T. \tag{4.139}$$

The existence of f_1 follows from the Gromov immersion theorem.

Then we choose a positive (large) number A such that

$$g - A^{-1} \cdot f_1^*(g_0) = g_1 \tag{4.140}$$

is a Riemannian metric on M, i.e., g_1 is a positive symmetric bilinear form. Such a number A exists, since M is compact.

Next we choose an isometric immersion

$$f_2 : \big(M^m, g_1\big) \to \big(\mathbb{R}^{N(m)}, g_0\big).$$

The existence of f_2 follows from the Nash isometric immersion theorem.

Lemma 4.7 *For all N there is a linear isometric embedding $L_N : (\mathbb{R}^N, g_0) \to (\mathbb{R}^{2N}, g_0)$ such that $L_N^*(T_0) = 0$.*

Proof For $x = (x_1, \ldots, x_N) \in \mathbb{R}^N$ we set

$$L_N(x_1, \ldots, x_N) := \big(f^1(x_1), \ldots, f^N(x_N)\big)$$

where f^i embeds the statistical line $(\mathbb{R}(x_i), (dx_i)^2, 0)$ into the statistical plane $(\mathbb{R}^2(x_{2i-1}, x_{2i}), (dx_{2i-1})^2 + (dx_{2i})^2, (dx_{2i-1})^3 + (dx_{2i})^3)$ as follows:

$$f^i(x_i) := \frac{1}{\sqrt{2}}(x_{2i-1} - x_{2i}).$$

Since f_i is an isometric embedding of $(\mathbb{R}(x_i), (dx_i)^2)$ into $(\mathbb{R}^2(x_{2i-1}, x_{2i}), (dx_{2i-1})^2 + (dx_{2i})^2)$, L_N is an isometric embedding of $(\mathbb{R}^N, g_0)$ into $(\mathbb{R}^{2N}, g_0)$. Set $T_0^{2i} := (dx_{2i-1})^3 + (dx_{2i})^3$. Clearly, $(f^i)^* T_0^{2i} = 0$. Since $T_0 = \sum_{i=1}^N T_0^{2i}$, it follows that $L_N^*(T_0) = 0$. This completes the proof of Lemma 4.7. □

Completion of the proof of Proposition 4.6 We choose an immersion

$$f_3 : M^m \to \mathbb{R}^{N_1(m)} \oplus \mathbb{R}^{2N(m)}$$

as follows

$$f_3(x) := A^{-1} \cdot f_1(x) \oplus (L_{N(m)} \circ f_2)(x).$$

Using (4.140) and the isometry property of f_2, we obtain

$$(f_3)^*(g_0) = A^{-1} \cdot f_1^*(g_0|_{\mathbb{R}^{N_1(m)}}) + f_2^*(g_0|_{\mathbb{R}^{2N(m)}}) = (g - g_1) + g_1 = g,$$

which implies that f_3 is an isometric embedding. Using (4.139) and Lemma 4.7, we obtain

$$(f_3)^*(A \cdot T_0) = A^{-1} \cdot f_1^*(A \cdot T_0|_{\mathbb{R}^{N_1(m)}}) = f_1^*(T_0) = T.$$

This implies that the immersion f_3 satisfies the condition of Proposition 4.6. □

4.5.4 Proof of the Existence of Isostatistical Immersions

Proposition 4.6 plays an important role in the proof of Theorem 4.10. Using it, we deduce Theorem 4.10 from the following

Proposition 4.7 *For any linear statistical manifold $(\mathbb{R}^n, g_0, A \cdot T_0)$ there exists an isostatistical immersion of $(\mathbb{R}^n, g_0, A \cdot T_0)$ into $(\mathcal{P}_+([4n]), \mathfrak{g}, \mathbf{T})$.*

Proof We shall choose a very large positive number

$$\bar{A} = \bar{A}(n, A), \tag{4.141}$$

which will be specified later in the proof of Lemma 4.8. First, $\bar{A}$ in (4.141) is required to be so large that there exists a number $1 < \lambda = \lambda(\bar{A}) < 2$ satisfying the following equation:

$$\lambda^2 + \frac{3n}{(2\bar{A})^2} = 4. \tag{4.142}$$

Equation (4.142) implies that $(\lambda, (2\bar{A})^{-1}, (2\bar{A})^{-1}, (2\bar{A})^{-1}) \in \mathbb{R}^4$ is a point in the positive sector $S^3_{2/\sqrt{n},+}$.

Hence there exists a positive number $r(\bar{A})$ such that for all $0 < r \leq r(\bar{A})$ the ball $U(\bar{A}, r)$ of radius r in the sphere $S^3_{2/\sqrt{n}}$ centered at the point

$$\left(\lambda, (2\bar{A})^{-1}, (2\bar{A})^{-1}, (2\bar{A})^{-1}\right)$$

also belongs to the positive sector $S^3_{2/\sqrt{n},+}$. For such r the Cartesian product $\bigtimes_{n \text{ times}} U(\bar{A}, r)$ is a subset in $S^{4n-1}_{2,+} \subseteq \mathbb{R}^{4n}$. This geometric observation helps us to reduce the proof of Proposition 4.7 to the proof of the following simpler statement.

Lemma 4.8 *For given positive number $A > 0$ there exist a positive number $\bar{A}$, satisfying (4.142) and depending only on n and A, a positive number $r < r(\bar{A})$ and an isostatistical immersion h from $(\mathbb{R}^n, g_0, A \cdot T_0)$ into $(\mathcal{P}_+([4n]), \mathfrak{g}, \mathbf{T})$ such that $h(\mathbb{R}^n, g_0, A \cdot T_0) \subseteq \bigtimes_{n \text{ times}} U(\bar{A}, r)$.*

Proof Since $(U(\bar{A}, r), g_0|_{U(\bar{A},r)}, T^*|_{U(\bar{A},r)})$ is a statistical submanifold of $(\mathbb{R}^4_+, g_0, T^*)$, the Cartesian product

$$\left(\bigtimes_{n \text{ times}} U(\bar{A}, r), \bigoplus_{i=1}^{n} (g_0)|_{U(\bar{A},r)}, \bigoplus_{i=1}^{n} T^*\big|_{U(\bar{A},r)}\right)$$

is a statistical submanifold of the statistical manifold $(\mathbb{R}^{4n}_+, g_0, T^*)$. Taking into account Example 4.1.2, we conclude that

$$\left(\bigtimes_{n \text{ times}} U(\bar{A}, r), \bigoplus_{i=1}^{n} (g_0)|_{U(\bar{A},r)}, \bigoplus_{i=1}^{n} T^*\big|_{U(\bar{A},r)}\right)$$

is a statistical submanifold of $(\mathcal{P}_+([4n]), \mathfrak{g}, \mathbf{T})$. Hence, to prove Lemma 4.8, it suffices to show that there are positive numbers $\bar{A} = \bar{A}(n, A), r < r(\bar{A})$ and an isostatistical immersion $f : ([0, R], dx^2, A \cdot dx^3) \to (U(\bar{A}, r), (g_0)|_{U(\bar{A},r)}, T^*|_{U(\bar{A},r)})$. On

$U(\bar{A}, r)$, for any given $\rho > 0$ we consider the distribution $D(\rho) \subseteq TU(\bar{A}, r)$ defined by

$$D_x(\rho) := \left\{ v \in T_x U(\bar{A}, r) : |v|_{g_0} = 1, T^*(v, v, v) = \rho \right\}.$$

Clearly, the existence of an isostatistical immersion $f : (\mathbb{R}, dx^2, A \cdot dx^3) \to (U(\bar{A}, r), (g_0)|_{U(\bar{A},r)}, T^*|_{U(A,r)})$ is equivalent to the existence of an integral curve of the distribution $D(A)$ on $U(\bar{A}, r)$. Intuitively, $\bar{A}$ should be as large as possible to ensure that the monotone invariant $\mathcal{M}^1(U)$ is as large as possible for a small neighborhood $U \ni x$ in $(\mathcal{P}_+([4n]), \mathfrak{g}, \mathbf{T})$, see Corollary 4.6 below.

We shall search for the required integral curve using the following geometric lemma.

Lemma 4.9 *There exist a positive number $\bar{A} = \bar{A}(n, A)$ and an embedded torus T^2 in $U(\bar{A}, r)$ which is provided with a unit vector field V on T^2 such that $T^*(V, V, V) = A$.*

Proof of Lemma 4.9 Set

$$x_0 = x_0(\bar{A}) := \left(\lambda, (2\bar{A})^{-1}, (2\bar{A})^{-1}, (2\bar{A})^{-1}\right) \in S^3_{2/\sqrt{n},+},$$

where $\lambda = \lambda(\bar{A})$ is defined by (4.142). The following lemma is a key step in the proof of Lemma 4.9.

Lemma 4.10 *There exists a positive number $\bar{A} = \bar{A}(n, A)$ such that the following assertion holds. Let H be any 2-dimensional subspace in $T_{x_0}U(\bar{A}, r) \subseteq \mathbb{R}^4$. Then there exists a unit vector $w \in H$ such that $T^*(w, w, w) \geq \sqrt{2}A$.*

Proof of Lemma 4.10 Denote by $\vec{x}_0$ the vector in $\mathbb{R}^4$ with the same coordinates as those of the point x_0. For any given H as in Lemma 4.10 there exists a unit vector $\vec{h}$ in $\mathbb{R}^4$, which is not co-linear with $\vec{x}_0$ and which is orthogonal to H, such that a vector $w \in \mathbb{R}^4$ belongs to H if and only if w is a solution to the following two linear equations:

$$\langle w, \vec{x}_0 \rangle = 0, \tag{4.143}$$

$$\langle w, \vec{h} \rangle = 0. \tag{4.144}$$

Adding a multiple of $\vec{x}_0$ to $\vec{h}$ if necessary, and taking the normalization, we can assume that

$$\vec{h} = (0 = h_1, h_2, h_3, h_4) \quad \text{and} \quad \sum_i h_i^2 = 1.$$

Case 1. Suppose that not all the coordinates h_i of $\vec{h}$ are of the same sign. Since the statistical manifold $(\mathbb{R}^n, g_0, T^*)$ as well as the positive sector $S^3_{2/\sqrt{n},+}$ are invariant under the permutation of coordinates (x_2, x_3, x_4), observing that the last three

coordinates of x_0 are equal, w.l.o.g. we assume that $h_2 \leq 0, h_3 > 0$. We put

$$k_2 := \frac{-h_2}{\sqrt{(h_2)^2 + (h_3)^2}}, \qquad k_3 := \frac{h_3}{\sqrt{(h_2)^2 + (h_3)^2}}.$$

We shall search for the required vector w for Lemma 4.10 in the following form:

$$w := \big(w_1, w_2 = (1-\varepsilon_2)k_3, w_3 = (1-\varepsilon_2)k_2, 0 = w_4\big) \in \mathbb{R}^4. \tag{4.145}$$

Recall that w must satisfy (4.144) and (4.143). We observe that for any choice of w_1 and ε_2 Eq. (4.144) for w is satisfied. Now we need to find the parameters (w_1, ε_2) of w in (4.145) such that w satisfies (4.143). For this purpose we choose (w_1, ε_2) to be a solution of the following system of equations

$$\lambda \cdot w_1 + (1-\varepsilon_2) \cdot (2\bar{A})^{-1} \cdot (k_2 + k_3) = 0, \tag{4.146}$$

$$w_1^2 = \big(2\varepsilon_2 - \varepsilon_2^2\big). \tag{4.147}$$

Note that (4.146) is equivalent to (4.143), and (4.147) normalizes w so that $|w|^2 = 1$.

From (4.146) we express w_1 in terms of ε_2 as follows:

$$w_1 = -\frac{(1-\varepsilon_2)(k_2+k_3)}{\lambda \cdot 2\bar{A}}. \tag{4.148}$$

Substituting the value of w_1 from (4.148) into (4.147), we get the following equation for ε_2:

$$\left(\frac{(k_2+k_3)^2}{(\lambda \cdot 2\bar{A})^2} + 1\right)\varepsilon_2^2 - \left(2 + \frac{2(k_2+k_3)^2}{(\lambda \cdot 2\bar{A})^2}\right)\varepsilon_2 + \left(\frac{k_2+k_3}{\lambda \cdot 2\bar{A}}\right)^2 = 0,$$

which we simplify as follows:

$$\varepsilon_2^2 - 2\varepsilon_2 + \frac{(k_2+k_3)^2}{(k_2+k_3)^2 + 4\lambda^2\bar{A}^2} = 0. \tag{4.149}$$

Clearly, the following choice of ε_2 is a solution to (4.149):

$$\varepsilon_2 = 1 - \frac{2\lambda\bar{A}}{\sqrt{(k_2+k_3)^2 + 4\lambda^2\bar{A}^2}}. \tag{4.150}$$

By our assumption on h_2 and h_3, we have $0 \leq k_2, k_3 \leq 1$. Since $1 < \lambda < 2$ by (4.142), we conclude that when $\bar{A}$ goes to infinity, the value ε_2 goes to zero. Hence there exists a number $N_1 > 0$ such that if $\bar{A} > N_1$ then

$$\varepsilon_2 > 0 \quad \text{and} \quad (1-\varepsilon_2)^2 \geq \frac{3}{4}. \tag{4.151}$$

We shall show that for ε_2 in (4.150) that also satisfies (4.151) if $\bar{A}$ is sufficiently large, and for w_1 defined by (4.148), the vector w defined by (4.145) satisfies the required condition of Lemma 4.10. Since $x_0 = (\lambda, (2\bar{A})^{-1}, (2\bar{A})^{-1}, (2\bar{A})^{-1})$ we have

$$T^*_{x_0}(w, w, w) = \frac{2w_1^3}{\lambda} + (4\bar{A})\big(w_2^3 + w_3^3\big). \tag{4.152}$$

Now assume that $\bar{A} > N_1$. Noting that ε_2 is positive and close to zero, and using $k_2 \geq 0$, $k_3 \geq 0$, we obtain from (4.145)

$$w_2 \geq 0,\ w_3 \geq 0. \tag{4.153}$$

Since $0 < \varepsilon < 1$, $0 < k_2 + k_3 < 2$, and λ, $\bar{A}$ are positive, we obtain from (4.148)

$$w_1 < 0 \quad \text{and} \quad |w_1| < \frac{1}{\lambda \bar{A}}. \tag{4.154}$$

Taking into account (4.145) and (4.151), we obtain

$$w_2^2 + w_3^2 = (1 - \varepsilon_2)^2 \geq \frac{3}{4}. \tag{4.155}$$

Using (4.154), we obtain from (4.152)

$$T^*_{x_0}(w, w, w) \geq \frac{-2}{\lambda^4 \bar{A}^3} + (4\bar{A}) \cdot \big(w_2^3 + w_3^3\big). \tag{4.156}$$

Observing that the function $x^{3/2} + (c - x)^{3/2}$ is convex on the interval $[0, c]$ for any $c > 0$, and therefore $(w_2^3 + w_3^3)$ reaches the minimum under the constraints (4.153) and (4.155) at $w_2 = w_3 = \sqrt{3}/\sqrt{2}$, we obtain from (4.156)

$$T^*_{x_0}(w, w, w) \geq \frac{-2}{\lambda^4 \bar{A}^3} + (4\bar{A}) \cdot 2\left(\frac{\sqrt{3}}{\sqrt{2}}\right)^3 = \frac{-2}{\lambda^4 \bar{A}^3} + 8\left(\sqrt{\frac{3}{2}}\right)^3 \bar{A}. \tag{4.157}$$

Increasing $\bar{A}$ if necessary, noting that $1 < \lambda = \lambda(A)$, Eq. (4.157) implies that there exists a large positive number $\bar{A}(n, A)$ depending only on n and A such that any subspace H defined by Eqs. (4.143) and (4.144), where h is in Case 1, contains a unit vector w that satisfies the condition in Lemma 4.10, i.e., the RHS of (4.157) is larger than $\sqrt{2}A$.

Case 2. Without loss of generality we assume that $h_2 \geq h_3 \geq h_4 > 0$ and therefore we have

$$\alpha := \frac{h_2 + h_3}{h_4} \geq 2. \tag{4.158}$$

We shall search for the required vector w for Lemma 4.10 in the following form:

$$w := \big(w_1, w_2 = -(1 - \varepsilon_2), w_3 = -(1 - \varepsilon_2), w_4 = \alpha(1 - \varepsilon_2)\big). \tag{4.159}$$

Equations (4.159) and (4.158) ensure that $\langle w, \vec{h} \rangle = 0$ for any choice of parameters (w_1, ε_2) of w in (4.159). Next we require that the parameters (w_1, ε_2) of w satisfy the following two equations:

$$\lambda \cdot w_1 + \frac{(1-\varepsilon_2)(\alpha-2)}{2\bar{A}} = 0, \tag{4.160}$$

$$w_1^2 + (1-\varepsilon_2)^2\big(2+\alpha^2\big) = 1. \tag{4.161}$$

Note that (4.160) is equivalent to (4.143) and (4.161) normalizes w. From (4.160) we express w_1 in terms of ε_2 as follows:

$$w_1 = -\frac{(1-\varepsilon_2)(\alpha-2)}{\lambda 2\bar{A}}. \tag{4.162}$$

Set

$$B := \big(2+\alpha^2\big) + \frac{(\alpha-2)^2}{4\lambda^2\bar{A}^2}. \tag{4.163}$$

Plugging (4.162) into (4.161) and using (4.163), we obtain the following equation for ε_2:

$$(1-\varepsilon_2)^2 B - 1 = 0,$$

which is equivalent to the following equation:

$$(1-\varepsilon_2)^2 = \frac{1}{B}. \tag{4.164}$$

Since $\alpha \geq 2$ by (4.158), from (4.163) we have $B > 0$. Clearly,

$$\varepsilon_2 := 1 - \frac{1}{\sqrt{B}} \tag{4.165}$$

is a solution to (4.164).

Since $\alpha \geq 2$ and $\varepsilon_2 \leq 1$ by (4.165), we obtain from (4.162) that $w_1 \leq 0$. Taking into account $1 < \lambda$, $\bar{A} > 0$, we derive from (4.162) and (4.165) the following estimates:

$$\begin{aligned}
T^*_{x_0}(w,w,w) &= \frac{2w_1^3}{\lambda} + (4\bar{A})(1-\varepsilon_2)^3\big(\alpha^3-2\big) \\
&> 2w_1^3 + (4\bar{A})\big(\alpha^3-2\big)(1-\varepsilon_2)^3 \\
&= -\frac{(\alpha-2)^3}{4\bar{A}^3(\sqrt{B})^3} + 4\bar{A}\frac{(\alpha^3-2)}{(\sqrt{B})^3} \\
&\geq -\frac{\alpha^3-2}{4\bar{A}^3(\sqrt{B})^3} + 4\bar{A}\frac{(\alpha^3-2)}{(\sqrt{B})^3} \text{ (since } \alpha \geq 2) \\
&= \frac{\alpha^3-2}{(\sqrt{B})^3}\left(-\frac{1}{4\bar{A}^3} + 4\bar{A}\right).
\end{aligned} \tag{4.166}$$

Lemma 4.11 *There exists a large number $\bar{A} = \bar{A}(n, A)$ depending only on n such that for all choices of $\alpha \geq 2$ we have*

$$\frac{(\alpha^3 - 2)}{(\sqrt{B})^3} \geq \frac{1}{10^2}.$$

Proof To prove Lemma 4.11, it suffices to show that for $\alpha \geq 2$ we have

$$10^4\big(\alpha^3 - 2\big)^2 \geq B^3. \tag{4.167}$$

Clearly, there exists a positive number N_2 such that if $\bar{A} > N_2$, then by (4.163), we have

$$B < \frac{3}{2}\big(2 + \alpha^2\big) \tag{4.168}$$

for any $\alpha \geq 2$. Hence (4.167) is a consequence of the following relation:

$$10^4\big(\alpha^3 - 2\big)^2 \geq \left[\frac{3}{2}\big(2 + \alpha^2\big)\right]^3, \tag{4.169}$$

which we shall establish now. To prove (4.169), it suffices to show that

$$10^3\big(\alpha^3 - 2\big)^2 \geq \big(2 + \alpha^2\big)^3. \tag{4.170}$$

The inequality (4.170) is equivalent to the following:

$$999\alpha^6 - 6\alpha^4 - 4000\alpha^3 - 12\alpha^2 + 3992 \geq 0. \tag{4.171}$$

Since $\alpha \geq 2$, it follows that $\alpha^3 \geq 8$ and hence

$$999\alpha^6 - 4000\alpha^3 = 499\alpha^6 + 500\alpha^3\big(\alpha^3 - 8\big) \geq 499\alpha^6. \tag{4.172}$$

Using $2\alpha^6 \geq 6\alpha^4$, we obtain

$$499\alpha^6 - 6\alpha^4 \geq 497\alpha^6. \tag{4.173}$$

Using $a^4 \geq 16$, we obtain

$$497\alpha^6 - 12\alpha^2 = 496\alpha^6 + \alpha^2\big(\alpha^4 - 12\big) > 496\alpha^6 > 496\alpha^6. \tag{4.174}$$

From (4.172), (4.173) and (4.174), we obtain

$$999\alpha^6 - 6\alpha^4 - 4000\alpha^3 - 12\alpha^2 + 3992 \geq 496\alpha^6 + 3992 > 0. \tag{4.175}$$

This proves (4.170) and hence completes the proof of Lemma 4.11. □

Lemma 4.11 implies that when $\bar{A} = \bar{A}(A, n)$ is sufficiently large, the RHS of (4.166) is larger than $\sqrt{2}A$. This proves the existence of $\bar{A}$, which depends only on n and A, for Case 2.

This completes the proof of Lemma 4.10. □

From Lemma 4.10 we immediately obtain the following.

Corollary 4.6 *There exists a small neighborhood $U_1 \ni x_0$ in $\bar{U}(\bar{A}, r)$ such that the following statement holds. For any $x \in U_1$ and any two-dimensional subspace $H \subseteq T_x U_1$, we have*

$$\max\left\{T^*(v, v, v) \;\middle|\; v \in H \text{ and } |v|_{g_0} = 1\right\} \geq \frac{5}{4}A.$$

Completion of the proof of Lemma 4.9 Let $\bar{A} = \bar{A}(n, A)$ satisfy the condition of Lemma 4.10. Now we choose a small embedded torus T^2 in $U_1 \subseteq U(\bar{A}, r)$. By Corollary 4.6, for all $x \in T^2$ we have

$$\max\left\{T^*(v, v, v) \;\middle|\; v \in T_x T^2 \text{ and } |v|_{g_0} = 1\right\} \geq \frac{5}{4}A. \tag{4.176}$$

Denote by $T_1 T^2$ the bundle of the unit tangent vectors of T^2. Since $T^2 = \mathbb{R}^2/\mathbb{Z}^2$ is parallelizable, we have $T_1 T^2 = T^2 \times S^1$. Thus the existence of a vector field V required in Lemma 4.9 is equivalent to the existence of a function $T^2 \to S^1$ satisfying the condition of Lemma 4.9. Next we claim that there exists a unit vector field W on T^2 such that $T^*(W, W, W) = 0$. First we choose some orientation for T^2, that induces an orientation on $T_1 T^2$ and hence on the circle S^1. Take an arbitrary unit vector field W' on T^2, equivalently we pick a function $W' : T^2 \to S^1$. Now we consider the fiber bundle F over T^2 whose fiber over $x \in T^2$ consists of the interval $[W', -W']$ defined by the chosen orientation on the circle of unit vectors in $T_x S^2$. Since $T^*(W', W', W') = -T^*(W, W, W)$, for each $x \in T^2$ there exists a value W on $F(x)$ such that $T^*(W, W, W) = 0$ and W is closest to W'. Using W we identify the circle S^1 with the interval $[0, 1)$. The existence of W implies that the existence of a function $V : T^2 \to [0, 1)$, regarded as a unit vector field V on T^2, that satisfies the condition of Lemma 4.9 is equivalent to the existence of a function $f : T^2 \to [0, 1)$ satisfying the same condition. Now let $V(x)$ be the smallest value of unit vector $V(x) \in [0, 1) \subseteq S^1(T_x T^2)$ such that

$$T^*\big(V(x), V(x), V(x)\big) = A$$

for each $x \in T^2$. The existence of $V(x)$ follows from (4.176). This completes the proof of Lemma 4.9. □

As we have noted, Lemma 4.9 implies Lemma 4.8. □

This finishes the proof of Proposition 4.7. □

Proof of Theorem 4.10 Case I. M is a compact manifold. In this case, the existence of an isostatistical immersion of a statistical manifold (M, g, T) into $(\mathcal{P}_+([N]), \mathfrak{g}, \mathbf{T})$ for some finite N follows from Proposition 4.6 and Proposition 4.7.

Case II. M is a non-compact manifold. We shall reduce the existence of an immersion of (M^m, g, T) into $\mathcal{P}_+(\mathbb{N})$ satisfying the condition of Theorem 4.10 to Case I, using a partition of unity and Nash's trick. (The Nash trick is a bit more complicated and can be used to embed a non-compact Riemannian manifold into a finite-dimensional Euclidean manifold.) Since (M^m, g, T) is finite-dimensional, there exists a countable locally finite open bounded cover U_i, $i = \overline{1, \infty}$, of M^m. We can then find compact submanifolds with boundary $A_i \subseteq U_i$ whose union also covers M^m.

Let $\{v_i\}$ be a partition of unity subjected to the cover $\{U_i\}$ such that v_i is strictly positive on A_i. Let S_i be a sphere of dimension m.

The following lemma is based on a trick that is similar to Nash's trick in [196, part D, pp. 61–62].

Lemma 4.12 *For each i there exists a smooth map $\phi^i : A_i \to S_i$ with the following properties*:

(i) *ϕ^i can be extended smoothly to the whole M^m.*
(ii) *For each S_i there exists a statistical structure (g_i, T_i) on S_i such that*

$$g = \sum_i \left(\phi^i\right)^*(g_i), \tag{4.177}$$

$$T = \sum_i \left(\phi^i\right)^*(T_i). \tag{4.178}$$

Proof Let ϕ^i map the boundary of U_i into the north point of the sphere S_i. Furthermore, we can assume that this map ϕ^i is injective in A_i. Clearly, ϕ^i can be extended smoothly to the whole M^n. This proves assertion (i) of Lemma 4.12.

(ii) The existence of a Riemannian metric g_i on S_i that satisfies (4.177)

$$g = \sum_i \left(\phi^i\right)^*(g_i)$$

has been proved in [196, part D, pp. 61–62]. For the reader's convenience, and for the proof of the last assertion of Lemma 4.12, we shall repeat Nash's proof.

Let γ_i be a Riemannian metric on S_i. Set

$$g_0 = \sum_i \left(\phi^i\right)^*(\gamma_i), \tag{4.179}$$

where by Lemma 4.12 ϕ^i is a smooth map from M^m to S_i. This is a well-defined metric, since the covering $\{U_i\}$ is locally finite. By rescaling the metric γ_i, we can assume that $g - g_0$ is a positive metric. Now we set

$$g_i := \left(\phi^i\right)^*(\gamma_i) + v_i \cdot (g - g_0). \tag{4.180}$$

We claim that there is a Riemannian metric $\tilde{\gamma}_i$ on S_i such that

$$\left(\phi^i\right)^*(\tilde{\gamma}_i) = g_i. \tag{4.181}$$

Note $g_i - (\phi^i)^*(\gamma_i)$ has a support on U_i since $\mathrm{supp}(v_i) \subseteq U_i$. Since ϕ^i is injective in A_i, $((\phi^i)^{-1})^*(g_i - (\phi^i)^*(\gamma_i))$ is a non-negative quadratic form on S_i. Hence

$$\tilde{\gamma}_i := \left(\left(\phi^i\right)^{-1}\right)^*\left(g_i - \left(\phi^i\right)^*(\gamma_i)\right) + \gamma_i$$

is a Riemannian metric on S_i that satisfies (4.181).

Now we compute

$$\begin{aligned}\sum_i \left(\phi^i\right)^*(\tilde{\gamma}_i) &= \sum_i g_i \quad \text{(by (4.181))} \\ &= \sum_i \left(\phi^i\right)^*(\gamma_i) + v_i \cdot (g - g_0) \quad \text{(by (4.180))} \\ &= g_0 + (g - g_0) = g \quad \text{(by (4.179))}.\end{aligned}$$

This proves (4.177).

The proof of the existence of T_i that satisfies (4.178) follows the same scheme, as for the proof of the existence of g_i; it is even easier, since we do not have the issue of positivity of T_i. So we leave it to the reader as an exercise. □

Continuation of the proof of Theorem 4.10 Let $a_1, \ldots, a_\infty$ be a sequence of positive numbers with

$$\sum_{i=1}^{\infty} a_i^2 = 4.$$

By the proof of Theorem 4.10, there exist a large number $l(m)$ depending only on m and an isostatistical immersion

$$\psi^i : (S_i, g_i, T_i) \to \left(S^{4l(m)-1}_{a_i,+}, g_0, T^*\right)$$

for any $i \in \mathbb{N}$. Here the sphere $S^{4l(m)-1}_{a_i}$ has radius smaller than 2, so we have to adjust the number $\bar{A}$ in the proof of Case 1. The main point is that the value $\lambda = \lambda(\bar{A})$ defined by a modified Eq. (4.142), where the RHS is replaced by a_i^2, is bounded from below and from above by a number that depends only on $l(m)$ and the radius a_i. Thus the RHS of (4.157) goes to infinity, when $\bar{A}$ goes to infinity. Similarly, the RHS of (4.166) goes to infinity when $\bar{A}$ goes to infinity.

Set

$$I^i := \psi^i \circ \phi^i : M^m \to \left(S^{4l(m)-1}_{a_i,+}, g_0, T^*\right) \subseteq \left(\mathbb{R}^{4l(m)}, g_0, T^*\right).$$

Clearly, the map $I := (I^1 \times \cdots \times I^\infty)$ maps M^m into the Cartesian product of the positive sectors $S^{4l(m)-1}_{a_i,+}$, that is, a subset of the positive sectors $S^\infty_{\sqrt{2},+}$ of all positive

probability measures on $\mathbb{N}$. Since ψ^i are isostatistical immersions, by (4.177) and (4.178) the map I satisfies the required condition of Theorem 4.10. □

4.5.5 Existence of Statistical Embeddings

Theorem 4.11 *Any smooth (resp. C^0) compact statistical manifold (M^n, g, T) admits an isostatistical embedding into the statistical manifold $(\mathcal{P}_+([N]), \mathfrak{g}, \mathbf{T})$ for some finite number N. Any smooth (resp. C^0) non-compact statistical manifold (M^n, g, T) admits an embedding I into the space $\mathcal{P}_+(\mathbb{N})$ of all probability measures on $\mathbb{N}$ such that g and T coincide with the Fisher metric and the Amari–Chentsov tensor on $I(M^n)$, respectively.*

Proof To prove Theorem 4.11, we repeat the proof of Theorem 4.10, replacing the Nash immersion theorem by the Nash embedding theorem. First we observe that our immersion f_3 constructed in Sect. 4.5.3 is an embedding, if f_2 is an isometric embedding. The existence of an isometric embedding f_2 is ensured by the Nash theorem. Hence, if M^n is compact, to prove the existence of an isostatistical embedding of (M^n, g, T) into $(\mathcal{P}_+([N]), \mathfrak{g}, \mathbf{T})$, it suffices to prove the strengthened version of Proposition 4.7, where the existence of an isostatistical immersion is replaced by the existence of an isostatistical embedding, but we need only to embed a bounded domain D in a statistical manifold $(\mathbb{R}^n, g_0, A \cdot T_0)$ into $(\mathcal{P}_+([N]), \mathfrak{g}, \mathbf{T})$.

As in the proof of Proposition 4.7, the proof of the new strengthened version of Proposition 4.7 is reduced to the proof of the existence of an isostatistical immersion of a bounded statistical interval $([0, R], dt^2, A \cdot dt^3)$ into a torus T^2 of a small domain in $(S^7_{2/\sqrt{n},+}, \mathfrak{g}, T^*) \subseteq (\mathbb{R}^8, g_0, T^*)$, see the proof of Lemma 4.8.

The statistical immersion produced with the help of Lemma 4.9 will be an embedding if not all the integral curves of the distribution $D(A)$ on the torus T^2 are closed curves. Now we shall search for an isostatistical embedding of $([0, R], dt^2, A \cdot dt^3)$ into a torus $T^2 \times T^2$ of a small domain in $(S^3_{1/\sqrt{n},+}, g_0, T^*) \times (S^3_{1/\sqrt{n},+}, g_0, T^*) \subseteq (S^7_{2/\sqrt{n},+}, \mathfrak{g}, T^*) \subseteq (\mathbb{R}^8, g_0, T^*)$. Since T^4 is parallelizable, repeating the argument at the end of the proof of Lemma 4.8, we choose a distribution $D(A) \subseteq TT^4$ such that $D(A) = T^4 \times S^2$ and

$$D_x A = \left\{ v \in T_x T^4 \;\middle|\; |v|_{g_0} = 1, \text{ and } T^*(v, v, v) = A \right\}.$$

Now assume that the integral curves of $D(A)$ that lie on the first factor $T^2 \times y$ for all $y \in S^3_{1/\sqrt{n},+}$ are closed. Since T^2 is compact, there is a positive number p_1 such that the periods of these integral curves are at least p_1.

Now let us consider the following integral curve $\gamma(t)$ of $D(A)$ on T^4. The curve $\gamma(t)$ begins at a point $(0, 0, 0, 0) \in T^4$. Here we identify T^1 with $[0, 1]/(0 = 1)$. The integral curve lies on $T^2 \times (0, 0)$ until it approaches $(0, 0, 0, 0)$ again. Since $D_x(A) = S^2$, we can slightly modify the direction of $\gamma(t)$ and let it leave the torus

$T^2 \times (0,0)$ and after a very short time $\gamma(t)$ must stay on the torus $T^2 \times (\varepsilon, \varepsilon)$ where ε is sufficiently small. Without loss of generality we assume that the period of any closed curve of the distribution $D(A) \cap T(T^2 \times (\varepsilon, \varepsilon))$ is at least p_1. Repeating this procedure, since R and p_1 are finite, we produce an embedding of $([0, R], dt^2, A \cdot dt^3)$ into $T^4 \subseteq (S^3_{1/\sqrt{n},+}, g_0, T^*) \times (S^3_{1/\sqrt{n},+}, g_0, T^*)$.

This completes the proof of Theorem 4.11 in the case when M^n is compact.

It remains to consider the case where M^n is non-compact. Using the existence of isostatistical embeddings for the compact case we can assume that ψ^i is an embedding for all i. Now we shall show that the map I is an embedding. Assume that $x \in M$. By the assumption, there exists an A_i such that x is an interior point of A_i. Then for any $y \in M$, $I^i(x) \neq I^i(y)$, since ψ^i is an embedding, ϕ^i is injective in the interior of A_i and maps the boundary of A_i to the north pole of S_i. □

Remark 4.8 There are many open questions concerning immersions of C^k-statistical manifolds. One important problem is to find a class of statistical manifolds $(M, \mathfrak{g}, \mathbf{T})$ of exponential type (i.e., M are exponential families as in (3.31)) that admit an isostatistical embedding into linear statistical manifolds $(\mathbb{R}^n, g_0, A \cdot T_0)$ or into statistical manifolds $(\mathcal{P}_+([N]), \mathfrak{g}, \mathbf{T})$. This is a difficult problem, if $\dim M \geq 2$. A version of this problem is to find an explicit embedding of the Riemannian manifold $(M, \mathfrak{g})$ underlying a statistical manifold $(M, \mathfrak{g}, \mathbf{T})$ of exponential type into Euclidean spaces. The problem of embedding of hyperbolic Riemannian spaces into Euclidean spaces has been considered by many geometers. We refer the reader to [52] for a survey.