Chapter 2
Finite Information Geometry

The considerations of this chapter are restricted to the situation of probability distributions on a finite number of symbols, and are hence of a more elementary nature. We pay particular attention to this case for two reasons. On the one hand, many applications of information geometry are based on statistical models associated with finite sets, and, on the other hand, the finite case will guide our intuition within the study of the infinite-dimensional setting. Some of the definitions in this chapter can and will be directly extended to more general settings.

2.1 Manifolds of Finite Measures

Basic Geometric Objects We consider a non-empty and finite set I.[1] The real algebra of functions $I \to \mathbb{R}$ is denoted by $\mathcal{F}(I)$, and its unity $\mathbb{1}_I$ or simply $\mathbb{1}$ is given by $\mathbb{1}(i) = 1$, $i \in I$. This vector spans the space $\mathbb{R} \cdot \mathbb{1} := \{c \cdot \mathbb{1} \in \mathcal{F}(I) : c \in \mathbb{R}\}$ of constant functions which we also abbreviate by $\mathbb{R}$. Given a function $g : \mathbb{R} \to \mathbb{R}$ and an $f \in \mathcal{F}(I)$, by $g(f)$ we denote the composition $i \mapsto g(f)(i) := g(f(i))$.

The space $\mathcal{F}(I)$ has the canonical basis e_i, $i \in I$, with

$$e_i(j) = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases}$$

and every function $f \in \mathcal{F}(I)$ can be written as

$$f = \sum_{i \in I} f^i \, e_i,$$

where the coordinates f^i are given by the corresponding values $f(i)$. We naturally interpret linear forms $\sigma : \mathcal{F}(I) \to \mathbb{R}$ as signed measures on I and denote the

[1] This set I will play the role of the no longer necessarily finite space Ω in Chap. 3.

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_2

corresponding dual space $\mathcal{F}(I)^*$, the space of $\mathbb{R}$-valued linear forms on $\mathcal{F}(I)$, by $\mathcal{S}(I)$. In a more general context, this interpretation is justified by the Riesz representation theorem. Here, it allows us to highlight a particular geometric perspective, which makes it easier to introduce natural information-geometric objects. Later in the book, we will treat general signed measures, and thereby have to carefully distinguish between various function spaces and their dual spaces.

The space $\mathcal{S}(I)$ has the dual basis δ^i, $i \in I$, defined by

$$\delta^i(e_j) := \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases}$$

Each element δ^i of the dual basis corresponds, interpreted as a measure, to the Dirac measure concentrated in i. A linear form $\mu \in \mathcal{S}(I)$, with $\mu_i := \mu(e_i)$, then has the representation

$$\mu = \sum_{i \in I} \mu_i \, \delta^i$$

with respect to the dual basis. This representation highlights the fact that μ can be interpreted as a signed measure, given by a linear combination of Dirac measures. The basis e_i, $i \in I$, allows us to consider the natural isomorphism between $\mathcal{F}(I)$ and $\mathcal{S}(I)$ defined by $e_i \mapsto \delta^i$. Note that this isomorphism is based on the existence of a distinguished basis of $\mathcal{F}(I)$. Information geometry, on the other hand, aims at identifying structures that are independent of such a particular choice of a basis. Therefore, the canonical basis will be used only for convenience, and all relevant information-geometric structures will be independent of this choice.

In what follows, we introduce several submanifolds of $\mathcal{S}(I)$ which play an important role in this chapter and which will be generalized and studied later in the book:

$$\begin{aligned} \mathcal{S}_a(I) &:= \left\{ \sum_{i \in I} \mu_i \, \delta^i : \sum_{i \in I} \mu_i = a \right\} \quad (\text{for } a \in \mathbb{R}), \\ \mathcal{M}(I) &:= \left\{ \mu \in \mathcal{S}(I) : \mu_i \geq 0 \text{ for all } i \in I \right\}, \\ \mathcal{M}_+(I) &:= \left\{ \mu \in \mathcal{M}(I) : \mu_i > 0 \text{ for all } i \in I \right\}, \\ \mathcal{P}(I) &:= \left\{ \mu \in \mathcal{M}(I) : \mu_i \geq 0 \text{ for all } i \in I, \text{ and } \sum_{i \in I} \mu_i = 1 \right\}, \\ \mathcal{P}_+(I) &:= \left\{ \mu \in \mathcal{P}(I) : \mu_i > 0 \text{ for all } i \in I, \text{ and } \sum_{i \in I} \mu_i = 1 \right\}. \end{aligned} \tag{2.1}$$

Obviously, we have the following inclusion chain of submanifolds of $\mathcal{S}(I)$:

$$\mathcal{P}_+(I) \subseteq \mathcal{M}_+(I) \subseteq \mathcal{S}(I).$$

In Sect. 3.1, we shall also alternatively interpret $\mathcal{P}_+(I)$ as the set of measures in $\mathcal{M}_+(I)$ that are defined up to a scaling factor, that is, as the projective space associated with $\mathcal{M}_+(I)$. From that point of view, $\mathcal{P}_+(I)$ is a positive spherical sector rather than a simplex.

Tangent and Cotangent Bundles We start with the vector space $\mathcal{S}(I)$. Given a point $\mu \in \mathcal{S}(I)$, clearly the tangent space is given by

$$T_\mu \mathcal{S}(I) = \{\mu\} \times \mathcal{S}(I). \tag{2.2}$$

Consider the natural identification of $\mathcal{S}(I)^* = \mathcal{F}(I)^{**}$ with $\mathcal{F}(I)$:

$$\mathcal{F}(I) \longrightarrow \mathcal{S}(I)^*, \qquad f \longmapsto \big(\mathcal{S}(I) \to \mathbb{R}, \mu \mapsto \mu(f)\big). \tag{2.3}$$

With this identification, the cotangent space of $\mathcal{S}(I)$ in μ is given by

$$T^*_\mu \mathcal{S}(I) = \{\mu\} \times \mathcal{F}(I). \tag{2.4}$$

As an open submanifold of $\mathcal{S}(I)$, $\mathcal{M}_+(I)$ has the same tangent and cotangent space at a point $\mu \in \mathcal{M}_+(I)$:

$$T_\mu \mathcal{M}_+(I) = \{\mu\} \times \mathcal{S}(I), \qquad T^*_\mu \mathcal{M}_+(I) = \{\mu\} \times \mathcal{F}(I). \tag{2.5}$$

Finally, we consider the manifold $\mathcal{P}_+(I)$. Obviously, for $\mu \in \mathcal{P}_+(I)$ we have

$$T_\mu \mathcal{P}_+(I) = \{\mu\} \times \mathcal{S}_0(I). \tag{2.6}$$

In order to specify the cotangent space, we consider the natural identification map (2.3). In terms of this identification, each $f \in \mathcal{F}(I)$ defines a linear form on $\mathcal{S}(I)$, which we now restrict to $\mathcal{S}_0(I)$. We obtain the map $\rho : \mathcal{F}(I) \to \mathcal{S}_0(I)^*$ that assigns to each f the linear form $\mathcal{S}_0(I) \to \mathbb{R}$, $\mu \mapsto \mu(f)$. Obviously, the kernel of ρ consists of the space of constant functions, and we obtain the natural isomorphism $\overline{\rho} : \mathcal{F}(I)/\mathbb{R} \to \mathcal{S}_0(I)^*$, $f + \mathbb{R} \mapsto \overline{\rho}(f + \mathbb{R}) := \rho(f)$. It will be useful to express the inverse $\overline{\rho}^{-1}$ in terms of the basis δ^i, $i \in I$, of $\mathcal{S}(I)$. In order to do so, assume $f \in \mathcal{S}_0(I)^*$, and consider an extension $\widetilde{f} \in \mathcal{S}(I)^*$. One can easily see that, with $f^i := \widetilde{f}(\delta^i)$, $i \in I$, the following holds:

$$\overline{\rho}^{-1}(f) = \left(\sum_{i \in I} f^i \, e_i \right) + \mathbb{R}. \tag{2.7}$$

Summarizing, we obtain

$$T^*_\mu \mathcal{P}_+(I) = \{\mu\} \times \big(\mathcal{F}(I)/\mathbb{R}\big) \tag{2.8}$$

as the cotangent space of $\mathcal{P}_+(I)$ at μ.

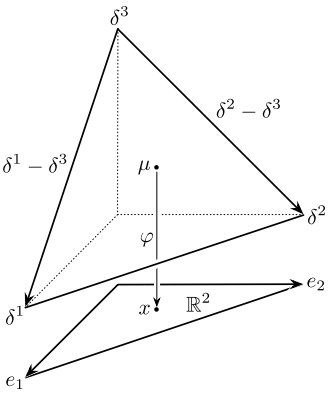

Fig. 2.1 Illustration of the chart φ for $n = 2$, with the two coordinate vectors $\delta^1 - \delta^3$ and $\delta^2 - \delta^3$

After having specified tangent and cotangent spaces at individual points μ, we finally list the corresponding tangent and cotangent bundles:

$$\begin{aligned} T\mathcal{S}(I) &= \mathcal{S}(I) \times \mathcal{S}(I), & T^*\mathcal{S}(I) &= \mathcal{S}(I) \times \mathcal{F}(I), \\ T\mathcal{M}_+(I) &= \mathcal{M}_+(I) \times \mathcal{S}(I), & T^*\mathcal{M}_+(I) &= \mathcal{M}_+(I) \times \mathcal{F}(I), \\ T\mathcal{P}_+(I) &= \mathcal{P}_+(I) \times \mathcal{S}_0(I), & T^*\mathcal{P}_+(I) &= \mathcal{P}_+(I) \times \big(\mathcal{F}(I)/\mathbb{R}\big). \end{aligned} \tag{2.9}$$

Example 2.1 (Local coordinates) In this example we consider a natural coordinate system of $\mathcal{P}_+(I)$ which is quite useful (see Fig. 2.1). We assume $I = [n+1] = \{1, \ldots, n, n+1\}$. With the open set

$$U := \left\{ x = (x_1, \ldots, x_n) \in \mathbb{R}^n : x_i > 0 \text{ for all } i, \text{ and } \sum_{i=1}^n x_i < 1 \right\},$$

we consider the map

$$\varphi : \mathcal{P}_+(I) \to U, \qquad \mu = \sum_{i=1}^{n+1} \mu_i \, \delta^i \mapsto \varphi(\mu) := (\mu_1, \ldots, \mu_n)$$

and its inverse

$$\varphi^{-1} : U \to \mathcal{P}_+(I), \qquad (x_1, \ldots, x_n) \mapsto \sum_{i=1}^n x_i \, \delta^i + \left(1 - \sum_{i=1}^n x_i\right) \delta^{n+1}.$$

We have the coordinate vectors

$$\frac{\partial}{\partial x_i}\bigg|_\mu = \frac{\partial \varphi^{-1}}{\partial x_i}\bigg|_{\varphi(\mu)} = \delta^i - \delta^{n+1}, \quad i = 1, \ldots, n,$$

which form a basis of $\mathcal{S}_0(I)$. Formula (2.7) allows us to identify its dual basis with the following basis of $\mathcal{F}(I)/\mathbb{R}$:

$$dx_i := e_i + \mathbb{R}, \quad i = 1, \ldots, n.$$

Each vector $f + \mathbb{R}$ in $\mathcal{F}(I)/\mathbb{R}$ can be expressed with respect to dx_i, $i = 1, \ldots, n$:

$$\begin{aligned} f + \mathbb{R} &= \left(\sum_{i=1}^{n+1} f^i \, e_i \right) + \mathbb{R} \\ &= \sum_{i=1}^{n+1} f^i \, (e_i + \mathbb{R}) \\ &= \sum_{i=1}^{n} f^i \, (e_i + \mathbb{R}) + f^{n+1}(e_{n+1} + \mathbb{R}) \\ &= \sum_{i=1}^{n} f^i \, (e_i + \mathbb{R}) + f^{n+1}\left(\left(\mathbb{1} - \sum_{i=1}^{n} e_i \right) + \mathbb{R} \right) \\ &= \sum_{i=1}^{n} f^i (e_i + \mathbb{R}) - \sum_{i=1}^{n} f^{n+1}(e_i + \mathbb{R}) \\ &= \sum_{i=1}^{n} \big(f^i - f^{n+1} \big)(e_i + \mathbb{R}). \end{aligned}$$

The coordinate system of this example will be useful for explicit calculations later on.

2.2 The Fisher Metric

The Definition Given a measure $\mu \in \mathcal{M}_+(I)$, we have the following natural L^2-product on $\mathcal{F}(I)$:

$$\langle f, g \rangle_\mu = \mu(f \cdot g), \quad f, g \in \mathcal{F}(I). \tag{2.10}$$

This product allows us to consider the vector space isomorphism

$$\mathcal{F}(I) \longrightarrow \mathcal{S}(I), \qquad f \longmapsto f\mu := \langle f, \cdot \rangle_\mu. \tag{2.11}$$

The notation $f\mu$ emphasizes that, via this isomorphism, functions define linear forms on $\mathcal{F}(I)$ in terms of densities with respect to μ. The inverse, which we denote by $\widetilde{\phi}_\mu$, maps linear forms to functions and represents a simple version of the

Radon–Nikodym derivative with respect to μ:

$$\widetilde{\phi}_\mu : \mathcal{S}(I) \longrightarrow \mathcal{F}(I) = \mathcal{S}(I)^*, \qquad a = \sum_i a_i\, \delta^i \longmapsto \frac{da}{d\mu} := \sum_i \frac{a_i}{\mu_i}\, e_i. \tag{2.12}$$

This gives rise to the formulation of (2.10) on the dual space of $\mathcal{F}(I)$:

$$\langle a, b\rangle_\mu = \mu\left(\frac{da}{d\mu} \cdot \frac{db}{d\mu}\right) = \sum_i \frac{1}{\mu_i}\, a_i b_i, \quad a, b \in \mathcal{S}(I). \tag{2.13}$$

With this metric, the orthogonal complement of $\mathcal{S}_0(I)$ in $\mathcal{S}(I)$ is given by $\mathbb{R} \cdot \mu = \{\lambda \cdot \mu : \lambda \in \mathbb{R}\}$, and we have the orthogonal decomposition $a = \Pi_\mu^\top a + \Pi_\mu^\perp a$ of vectors $a \in \mathcal{S}(I)$, where

$$\Pi_\mu^\top(a) = \sum_{i\in I}\left(a_i - \mu_i \sum_{j\in I} a_j\right)\delta^i, \qquad \Pi_\mu^\perp(a) = \sum_{i\in I}\left(\mu_i \sum_{j\in I} a_j\right)\delta^i. \tag{2.14}$$

If we restrict this metric to $\mathcal{S}_0(I) \subseteq \mathcal{S}(I)$, then we obtain the following identification of $\mathcal{S}_0(I)$ with the dual space:

$$\phi_\mu : \mathcal{S}_0(I) \longrightarrow \mathcal{F}(I)/\mathbb{R} = \mathcal{S}_0(I)^*, \qquad a \longmapsto \frac{da}{d\mu} + \mathbb{R}. \tag{2.15}$$

With the natural inclusion map $\imath : \mathcal{S}_0(I) \to \mathcal{S}(I)$, and $\imath_\mu := \widetilde{\phi}_\mu \circ \imath \circ {\phi_\mu}^{-1}$, the following diagram commutes:

$$\begin{array}{ccc} \mathcal{S}(I) & \xrightarrow{\widetilde{\phi}_\mu} & \mathcal{S}(I)^* \\ {\scriptstyle\imath}\uparrow & & \uparrow{\scriptstyle\imath_\mu} \\ \mathcal{S}_0(I) & \xrightarrow[\phi_\mu]{} & \mathcal{S}_0(I)^* \end{array} \tag{2.16}$$

This diagram defines linear maps between tangent and cotangent spaces in the individual points of $\mathcal{M}_+(I)$ and $\mathcal{P}_+(I)$. Collecting all these maps to corresponding bundle maps, we obtain a commutative diagram between the tangent and cotangent bundles:

$$\begin{array}{ccc} T\mathcal{M}_+(I) & \xrightarrow{\widetilde{\phi}} & T^*\mathcal{M}_+(I) \\ {\scriptstyle\imath}\uparrow & & \uparrow{\scriptstyle\imath} \\ T\mathcal{P}_+(I) & \xrightarrow[\phi]{} & T^*\mathcal{P}_+(I) \end{array} \tag{2.17}$$

The inner product (2.13) defines a Riemannian metric on $\mathcal{M}_+(I)$, on which the maps $\widetilde{\phi}$ and ϕ are based.

Definition 2.1 (Fisher metric) Given two vectors $A = (\mu, a)$ and $B = (\mu, b)$ of the tangent space $T_\mu \mathcal{M}_+(I)$, we consider

$$\mathfrak{g}_\mu(A, B) := \langle a, b \rangle_\mu. \tag{2.18}$$

This Riemannian metric $\mathfrak{g}$ on $\mathcal{M}_+(I)$ is called the *Fisher metric*.

The Fisher metric was introduced as a Riemannian metric by Rao [219]. It is relevant for estimation theory within statistics and also appears in mathematical population genetics where it is known as the *Shahshahani metric* [123, 124]. We shall outline the biological perspective of this metric in Sect. 6.2.

We now express the Fisher metric with respect to the coordinates of Example 2.1, where we concentrate on the restriction of the Fisher metric to $\mathcal{P}_+(I)$. With respect to the chart φ of Example 2.1, the first fundamental form of the Fisher metric is given as

$$g_{ij}(\mu) = \sum_{k=1}^{n} \frac{1}{\mu_k}\, \delta_{ki}\, \delta_{kj} + \frac{1}{\mu_{n+1}} = \begin{cases} \frac{1}{\mu_i} + \frac{1}{\mu_{n+1}}, & \text{if } i = j, \\ \frac{1}{\mu_{n+1}}, & \text{otherwise.} \end{cases} \tag{2.19}$$

The inverse of this matrix is given as

$$g^{ij}(\mu) = \begin{cases} \mu_i\,(1 - \mu_i), & \text{if } i = j, \\ -\mu_i\, \mu_j, & \text{otherwise.} \end{cases} \tag{2.20}$$

Written as matrices, we have

$$G(\mu) := (g_{ij})(\mu) = \frac{1}{\mu_{n+1}} \begin{pmatrix} \frac{\mu_{n+1}}{\mu_1} + 1 & 1 & \cdots & 1 \\ 1 & \frac{\mu_{n+1}}{\mu_2} + 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & \frac{\mu_{n+1}}{\mu_n} + 1 \end{pmatrix}, \tag{2.21}$$

$$G^{-1}(\mu) = \left(g^{ij}\right)(\mu) = \begin{pmatrix} \mu_1\,(1 - \mu_1) & -\mu_1\, \mu_2 & \cdots & -\mu_1\, \mu_n \\ -\mu_2\, \mu_1 & \mu_2\,(1 - \mu_2) & \cdots & -\mu_2\, \mu_n \\ \vdots & \vdots & \ddots & \vdots \\ -\mu_n\, \mu_1 & -\mu_n\, \mu_2 & \cdots & \mu_n\,(1 - \mu_n) \end{pmatrix}. \tag{2.22}$$

This is nothing but the covariance matrix of the probability distribution μ, in the following sense. We draw the element $i \in \{1, \ldots, n\}$ with probability μ_i, and we put $X_i = 1$ and $X_j = 0$ for $j \neq i$ when i happens to be drawn. We then have the expectation values

$$\mu_i = \mathbb{E}(X_i) = \mathbb{E}\left(X_i^2\right), \tag{2.23}$$

and hence, the variances and covariances are

$$\mathrm{Var}(X_i) = \mu_i(1-\mu_i), \qquad \mathrm{Cov}(X_i X_j) = -\mu_i\mu_j \quad \text{for } j \neq i, \tag{2.24}$$

that is, (2.22). In fact, this is the statistical origin of the Fisher metric as a covariance matrix [219].

The Fisher metric is an example of a covariant 2-tensor on M, in the sense of the following definition (see also (B.16) and (B.17) in Appendix B).

Definition 2.2 A *covariant n-tensor Θ on a manifold M* is a collection of n-multilinear maps

$$\Theta_\xi : \bigtimes^n T_\xi M \longrightarrow \mathbb{R}, \qquad (V_1, \ldots, V_n) \longmapsto \Theta_\xi(V_1, \ldots, V_n)$$

which is continuous in the sense that for continuous vector fields V_i the function $\xi \mapsto \Theta_\xi(V_1, \ldots, V_n)$ is continuous.

If $f : M_1 \to M_2$ is a differentiable map between the manifolds M_1 and M_2, then it can be used to pull back covariant n-tensors. That is, if Θ is a covariant n-tensor on M_2, then its pullback to M_1 by f is defined to be the tensor $f^*(\Theta)$ on M_1 given as

$$f^*(\Theta)_\xi(V_1, \ldots, V_n) := \Theta_{f(\xi)}\left(\frac{\partial f}{\partial V_1}, \ldots, \frac{\partial f}{\partial V_n}\right). \tag{2.25}$$

Information geometry deals with statistical models, that is, submanifolds of $\mathcal{P}_+(I)$. Instead of considering submanifolds directly, we take a slightly different perspective here. We consider statistical models as a manifold together with an embedding into $\mathcal{P}_+(I)$, or more generally, into $\mathcal{M}_+(I)$. To be more precise, let be M an n-dimensional (differentiable) manifold and $M \hookrightarrow \mathcal{M}_+(I)$, $\xi \mapsto p(\xi) = \sum_{i\in I} p_i(\xi)\,\delta^i$, an embedding. The pullback (2.25) of the Fisher metric $\mathfrak{g}$ defines a metric on M. More precisely, for $A, B \in T_\xi M$ we set

$$\begin{aligned} g_\xi(A,B) &:= p^*(\mathfrak{g})_\xi(A,B) \\ &\overset{(2.25)}{=} \mathfrak{g}_{p(\xi)}\left(\frac{\partial p}{\partial A}, \frac{\partial p}{\partial B}\right) \\ &= \sum_i \frac{1}{p_i(\xi)} \frac{\partial\, p_i}{\partial A}(\xi) \frac{\partial\, p_i}{\partial B}(\xi) \\ &= \sum_i p_i(\xi) \frac{\partial \log p_i}{\partial A}(\xi) \frac{\partial \log p_i}{\partial B}(\xi). \end{aligned} \tag{2.26}$$

This representation of the Fisher metric is more familiar within the standard information-geometric treatment.

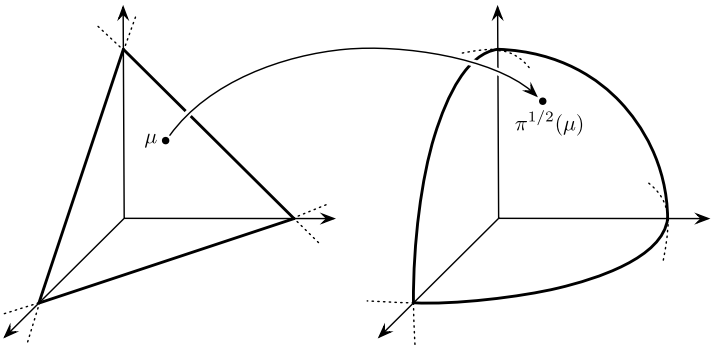

Fig. 2.2 Mapping from simplex to sphere

Extending the Fisher Metric to the Boundary As is obvious from (2.13) and also from the first fundamental form (2.19), the Fisher metric is not defined at the boundary of the simplex. It is, however, possible to extend the Fisher metric to the boundary by identifying the simplex with a sector of a sphere in $\mathbb{R}^I = \mathcal{F}(I)$. In order to be more precise, we consider the following sector of the sphere with radius 2 (or, equivalently (up to the factor 2, of course), the positive part of the projective space, according to the interpretation of the set of probability measures as a projective version of the space of positive measures alluded to above and taken up in Sect. 3.1):

$$S_{2,+}(I) := \left\{ f \in \mathcal{F}(I) : f(i) > 0 \text{ for all } i \in I, \text{ and } \sum_i f^2(i) = 4 \right\}.$$

As a submanifold of $\mathcal{F}(I)$ it carries the induced standard metric $\langle \cdot, \cdot \rangle$ of $\mathcal{F}(I)$. We consider the following diffeomorphism (see Fig. 2.2):

$$\pi^{1/2} : \mathcal{P}_+(I) \to S_{2,+}(I), \qquad \mu = \sum_i \mu_i \, \delta^i \mapsto 2 \sum_i \sqrt{\mu_i} e_i .$$

Note that $\|\pi^{1/2}(\mu)\| = \sqrt{\sum_i (2\sqrt{\mu_i})^2} = 2\sqrt{\sum_i \mu_i} = 2$.

Proposition 2.1 *The map $\pi^{1/2}$ is an isometry between $\mathcal{P}_+(I)$ with the Fisher metric $\mathfrak{g}$ and $S_{2,+}(I)$ with the induced canonical scalar product of $\mathcal{F}(I)$:*

$$\left\langle \frac{\partial \pi^{1/2}}{\partial A}(\mu), \frac{\partial \pi^{1/2}}{\partial B}(\mu) \right\rangle = \mathfrak{g}_\mu(A, B), \quad A, B \in T_\mu \mathcal{P}_+(I).$$

That is, the Fisher metric coincides with the pullback of the standard metric on $\mathcal{F}(I)$ by the map $\pi^{1/2}$. In particular, the extension of the standard metric on $S_{2,+}(I)$ to the boundary can be considered as an extension of the Fisher metric.

Proof With $a, b \in \mathcal{S}_0(I)$ such that $A = (\mu, a)$ and $B = (\mu, b)$, we have

$$\begin{aligned}\left\langle \frac{\partial \pi^{1/2}}{\partial A}(\mu), \frac{\partial \pi^{1/2}}{\partial B}(\mu) \right\rangle &= \left\langle \frac{d}{dt}\pi^{1/2}(\mu + ta)\Big|_{t=0}, \frac{d}{dt}\pi^{1/2}(\mu + tb)\Big|_{t=0} \right\rangle \\ &= \sum_i \frac{1}{\sqrt{\mu_i}} a_i \cdot \frac{1}{\sqrt{\mu_i}} b_i \\ &= \mathfrak{g}_\mu(A, B). \end{aligned}$$

□

Fisher and Hellinger Distance Proposition 2.1 allows us to give an explicit formula for the Riemannian distance between two points $\mu, \nu \in \mathcal{P}_+(I)$ which is defined as

$$d^F(\mu, \nu) := \inf_{\substack{\gamma:[r,s]\to\mathcal{P}_+(I) \\ \gamma(r)=\mu,\, \gamma(s)=\nu}} L(\gamma),$$

where $L(\gamma)$ denotes the length of a curve $\gamma : [r, s] \to \mathcal{P}_+(I)$ given by

$$L(\gamma) = \int_r^s \left\|\dot\gamma(t)\right\|_{\gamma(t)} dt = \int_r^s \sqrt{\mathfrak{g}_{\gamma(t)}\big(\dot\gamma(t), \dot\gamma(t)\big)}\, dt.$$

We refer to d^F as the *Fisher distance*. With Proposition 2.1 we directly obtain the following corollary.

Corollary 2.1 *Let $d : S_{2,+}(I) \times S_{2,+}(I) \to \mathbb{R}$ denote the metric that is induced from the standard metric on $\mathcal{F}(I)$. Then*

$$d^F(\mu, \nu) = d\big(\pi^{1/2}(\mu), \pi^{1/2}(\nu)\big) = 2\arccos\left(\sum_i \sqrt{\mu_i\, \nu_i}\right). \tag{2.27}$$

Proof We have

$$\cos\alpha = \frac{\langle \pi^{1/2}(\mu), \pi^{1/2}(\nu)\rangle}{\|\pi^{1/2}(\mu)\| \cdot \|\pi^{1/2}(\nu)\|} = \frac{\sum_i (2\sqrt{\mu_i})(2\sqrt{\nu_i})}{2 \cdot 2} = \sum_i \sqrt{\mu_i\, \nu_i}.$$

This implies

$$\frac{d^F(\mu, \nu)}{2} = \alpha = \arccos\left(\sum_i \sqrt{\mu_i\, \nu_i}\right).$$

□

A distance measure that is closely related to the Fisher distance is the *Hellinger distance*. It is not induced from $\mathcal{F}(I)$ onto $S_{2,+}(I)$ but restricted to $S_{2,+}(I)$:

$$d^H(\mu, \nu) := \sqrt{\sum_i (\sqrt{\mu_i} - \sqrt{\nu_i})^2}. \tag{2.28}$$

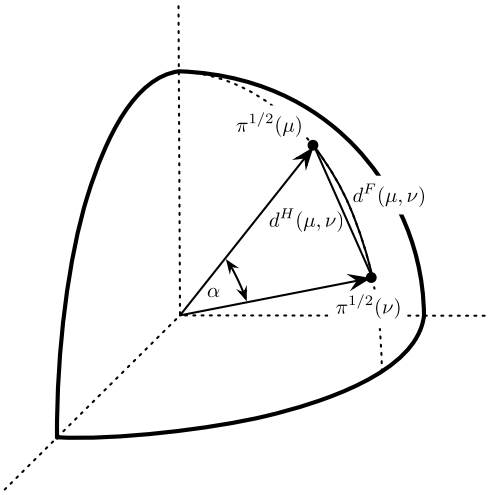

Fig. 2.3 Illustration of the relation between the Fisher distance $d^F(\mu,\nu)$ and the Hellinger distance $d^H(\mu,\nu)$ of two probability measures μ and ν, see Eq. (2.29)

We have the following relation between d^F and d^H (see Fig. 2.3):

$$\begin{aligned} d^H(\mu,\nu) &= \sqrt{\sum_i (\sqrt{\mu_i} - \sqrt{\nu_i})^2} \\ &= \sqrt{\sum_i (\mu_i - 2\sqrt{\mu_i\,\nu_i} + \nu_i)} \\ &= \sqrt{2\Big(1 - \sum_i \sqrt{\mu_i\,\nu_i}\Big)} \\ &= \sqrt{2\Big(1 - \cos\Big(\frac{1}{2}\,d^F(\mu,\nu)\Big)\Big)}. \end{aligned} \tag{2.29}$$

Chentsov's Characterization of the Fisher Metric In what follows, we present a classical characterization of the Fisher metric through invariance properties. This is due to Chentsov [64].

Consider two non-empty and finite sets I and I'. A *Markov kernel* is a map

$$K : I \to \mathcal{P}(I'), \qquad i \mapsto K^i := \sum_{i' \in I'} K^i_{i'}\,\delta^{i'}. \tag{2.30}$$

Particular examples of Markov kernels are given in terms of (deterministic) maps $f : I \to I'$. Given such a map, we simply define K^f by $i \mapsto \delta^{f(i)}$. Each Markov kernel K induces a corresponding map between probability distributions:

$$K_* : \mathcal{P}(I) \to \mathcal{P}(I'), \qquad \mu = \sum_{i \in I} \mu_i\,\delta^i \mapsto \sum_{i \in I} \mu_i\,K^i. \tag{2.31}$$

The map K_* is called the *Markov morphism induced by* K. Note that K_* may also be regarded as a linear map $K_* : \mathcal{S}(I) \to \mathcal{S}(I')$. Given a map $f : I \to I'$, we use $f_* := (K^f)_*$ as short-hand notation.

Now assume that $|I| \leq |I'|$. We call a Markov kernel K *congruent* if there is a partition A_i, $i \in I$, of I', such that the following condition holds:

$$K^i_{i'} > 0 \quad \Leftrightarrow \quad i' \in A_i. \tag{2.32}$$

If K is congruent and $\mu \in \mathcal{P}_+(I)$ then $K_*(\mu) \in \mathcal{P}_+(I')$. This implies a differentiable map

$$K_* : \mathcal{P}_+(I) \to \mathcal{P}_+\big(I'\big),$$

and the differential in μ is given by

$$d_\mu K_* : T_\mu \mathcal{P}_+(I) \to T_{K_*(\mu)} \mathcal{P}_+\big(I'\big), \qquad (\mu, \nu - \mu) \mapsto \big(K_*(\mu), K_*(\nu) - K_*(\mu)\big).$$

The following theorem has been proven by Chentsov.

Theorem 2.1 (Cf. [65, Theorem 11.1]) *We assign to each non-empty and finite set I a metric h^I on $\mathcal{P}_+(I)$. If for each congruent Markov kernel $K : I \to \mathcal{P}(I')$ we have invariance in the sense*

$$h^I_p(A, B) = h^{I'}_{K_*(p)}\big(d_p K_*(A), d_p K_*(B)\big),$$

or for short $(K_*)^*(h^{I'}) = h^I$ in the notation of (2.25)*, then there is a constant $\alpha > 0$, such that $h^I = \alpha\, \mathfrak{g}^I$ for all I, where the latter is the Fisher metric on $\mathcal{P}_+(I)$.*

Proof **Step 1:** First we consider permutations $\pi : I \to I$. The center $c_I := \frac{1}{|I|} \sum_{i \in I} \delta^i$ is left-invariant, that is, $\pi_*(c_I) = c_I$. With $E_i := (c_I, \delta^i - c_I) \in T_{c_I} \mathcal{P}_+(I)$, we also have

$$d_{c_I} \pi_*(E_i) = E_{\pi(i)}, \quad i \in I.$$

For each $i, j \in I$, $i \neq j$, choose a transposition π of i and j, that is, $\pi(i) = j$, $\pi(j) = i$, and $\pi(k) = k$ if $k \notin \{i, j\}$. This implies

$$\begin{aligned} h^I_{ii}(c_I) &= h^I_{c_I}(E_i, E_i) = h^I_{\pi_*(c_I)}\big(d_{c_I}\pi_*(E_i), d_{c_I}\pi_*(E_i)\big) = h^I_{c_I}(E_{\pi(i)}, E_{\pi(i)}) \\ &= h^I_{c_I}(E_j, E_j) = h^I_{jj}(c_I) =: f_1(n), \end{aligned}$$

where we set $n := |I|$. In a similar way, we obtain that all $h^I_{ij}(c_I)$ with $i \neq j$ coincide. We denote them by $f_2(n)$. The functions $f_1(n)$ and $f_2(n)$ have to satisfy the following:

$$\begin{aligned} f_1(n) + (n-1) f_2(n) &= \sum_{j \in I} h^I_{ij}(c_I) = \sum_{j \in I} h^I_{c_I}(E_i, E_j) \\ &= h^I_{c_I}\Big(E_i, \sum_{j \in I} E_j\Big) = h^I_{c_I}(E_i, 0) = 0. \end{aligned}$$

Consider two vectors

$$a = \sum_{i \in I} a_i \, \delta^i, \qquad b = \sum_{i \in I} b_i \, \delta^i.$$

Assuming $a, b \in \mathcal{S}_0(I)$, we have $\sum_{i \in I} a_i = 0$ and $\sum_{i \in I} b_i = 0$ and therefore

$$a = \sum_{i \in I} a_i \left(\delta^i - c_I\right), \qquad b = \sum_{i \in I} b_i \left(\delta^i - c_I\right).$$

This implies for $A = (c_I, a)$ and $B = (c_I, b)$

$$\begin{aligned}
h^I_{c_I}(A, B) &= \sum_{i,j \in I} a_i \, b_j \, h^I_{ij}(c_I) = \sum_{i \in I} a_i \, b_i \, h^I_{ii}(c_I) + \sum_{\substack{i,j \in I \\ i \neq j}} a_i \, b_j \, h^I_{ij}(c_I) \\
&= f_1(n) \sum_{i \in I} a_i \, b_i + f_2(n) \sum_{\substack{i,j \in I \\ i \neq j}} a_i \, b_j \\
&= -(n-1) \, f_2(n) \sum_{i \in I} a_i \, b_i + f_2(n) \sum_{\substack{i,j \in I \\ i \neq j}} a_i \, b_j \\
&= f_2(n) \left\{ -n \sum_{i \in I} a_i \, b_i + \sum_{i,j \in I} a_i \, b_j \right\} \\
&= -f_2(n) \sum_{i \in I} \frac{1}{\frac{1}{n}} \, a_i \, b_i \\
&= -f_2(n) \, \mathfrak{g}^I_{c_I}(A, B).
\end{aligned}$$

Step 2: In this step, we show that the function $f(n)$ is actually independent of n and therefore a constant. In order to do so, we divide each element $i \in I$ into k elements. More precisely, we set $I' := I \times \{1, \ldots, k\}$. With the partition $A_i := \{(i, j) : 1 \leq j \leq k\}$, $i \in I$, we define the Markov kernel K by

$$i \mapsto K^i = \frac{1}{k} \sum_{j=1}^{k} \delta^{(i,j)}.$$

This kernel satisfies $K_*(c_I) = c_{I'}$, and we have

$$\begin{aligned}
d_{c_I} K_*(E_i) &= d_{c_I} K_*\left(c_I, \delta^i - c_I\right) \\
&= \left(c_{I'}, \frac{1}{k} \sum_{j=1}^{k} \left(\delta^{(i,j)} - \frac{1}{n} \sum_{i' \in I} \delta^{(i',j)} \right) \right)
\end{aligned}$$

$$= \left(c_{I'}, \frac{1}{k} \sum_{j=1}^{k} \left(\delta^{(i,j)} - \frac{1}{n\,k} \sum_{i' \in I} \sum_{j'=1}^{k} \delta^{(i',j')} \right) \right)$$

$$= \frac{1}{k} \sum_{j=1}^{k} E'_{(i,j)}.$$

With $r, s \in I$, $r \neq s$, this implies

$$f_2(n) = h^I_{c_I}(E_r, E_s) = h^{I'}_{c_{I'}} \left(\frac{1}{k} \sum_{j=1}^{k} E'_{r,j}, \frac{1}{k} \sum_{j=1}^{k} E'_{s,j} \right)$$

$$= \frac{1}{k^2} \sum_{j_1, j_2 = 1}^{k} h^{I'}_{c_{I'}} \left(E'_{r,j_1}, E'_{s,j_2} \right)$$

$$= \frac{1}{k^2}\, k^2\, f_2(n \cdot k) = f_2(n \cdot k).$$

Exchanging the role of n and k, we obtain $f_2(k) = f_2(k \cdot n) = f_2(n)$ and therefore $-f_2(n)$ is a positive constant in n, which we denote by α. In the center c_I, we have shown that

$$h^I_{c_I}(A, B) = \alpha\, \mathfrak{g}^I_{c_I}(A, B) = 0, \quad A, B \in T_{c_I} \mathcal{P}_+(I).$$

It remains to show that this equality also holds for all other points. This is our next step.

Step 3: First consider a point $\mu \in \mathcal{P}_+(I)$ that has rational coordinates, that is,

$$\mu = \sum_{i \in I} \frac{k_i}{n}\, \delta^i,$$

with $\sum_i k_i = n$. We now define a set I' and a congruent Markov kernel $K : I \to \mathcal{P}(I')$ so that $K_*(\mu) = c_{I'}$. With

$$I' := \biguplus_{i \in I} \left(\{i\} \times \{1, \ldots, k_i\} \right),$$

("$\biguplus$" denotes the disjoint union) we consider the Markov kernel

$$K : i \mapsto \frac{1}{k_i} \sum_{j=1}^{k_i} \delta^{(i,j)}.$$

Obviously, we have

$$d_\mu K_* : A = \left(\mu, \sum_{i \in I} a_i\, \delta^i \right) \mapsto d_\mu K_*(A) = \left(c_{I'}, \sum_{i \in I} \sum_{j=1}^{k_i} \frac{a_i}{k_i}\, \delta^{(i,j)} \right).$$

This implies

$$\begin{aligned} h^I_\mu(A,B) &= h^{I'}_{K_*(\mu)}\big(d_\mu K_*(A), d_\mu K_*(B)\big) = \alpha\, g^{I'}_{c_{I'}}\big(d_\mu K_*(A), d_\mu K_*(B)\big) \\ &= \alpha \sum_{i\in I} \sum_{j=1}^{k_i} \frac{1}{\frac{1}{n}} \frac{a_i}{k_i} \frac{b_i}{k_i} = \alpha \sum_{i\in I} k_i \frac{1}{\frac{1}{n}} \frac{a_i}{k_i} \frac{b_i}{k_i} = \alpha \sum_{i\in I} \frac{1}{\mu_i} a_i\, b_i \\ &= \alpha\, \mathfrak{g}^I_\mu(A,B). \end{aligned}$$

We have this equality for all probability vectors μ with rational coordinates. As we assume continuity with respect to the base point μ, the equality of $h^I_\mu(A,B) = \alpha\, \mathfrak{g}^{I'}_\mu(A,B)$ holds for all $\mu \in \mathcal{P}_+(I)$. □

2.3 Gradient Fields

In this section, we are going to study vector and covector fields on $\mathcal{M}_+(I)$ and $\mathcal{P}_+(I)$. We begin with the first case, which is the simpler one, and consider covector fields given by a differentiable function $V : \mathcal{M}_+(I) \to \mathbb{R}$. The differential in μ is defined as the linear form

$$d_\mu V : \mathcal{S}(I) \to \mathbb{R}, \qquad a \mapsto d_\mu V(a) = \frac{\partial V}{\partial a}(\mu).$$

In terms of the canonical basis, we have

$$d_\mu V = \sum_i \partial_i V(\mu)\, e_i \in \mathcal{F}(I), \tag{2.33}$$

where $\partial_i V(\mu) := \frac{\partial V}{\partial \mu_i}(\mu) := \frac{\partial V}{\partial \delta^i}(\mu)$. This defines the covector field

$$dV : \mathcal{M}_+(I) \to \mathcal{F}(I), \qquad \mu \mapsto d_\mu V.$$

The Fisher metric $\mathfrak{g}$ allows us to identify $d_\mu V$ with an element of $T_\mu \mathcal{M}_+(I)$ in terms of the map $\widetilde{\phi}_\mu$, the *gradient* of V in μ:

$$\operatorname{grad}_\mu V := \widetilde{\phi}_\mu^{-1}(d_\mu V) = \sum_i \mu_i\, \partial_i V(\mu)\, \delta^i. \tag{2.34}$$

Given a function $f : \mathcal{M}_+(I) \to \mathcal{F}(I)$, $\mu \mapsto f(\mu) = \sum_{i\in I} f^i(\mu)\, e_i$, we can ask whether there exists a differentiable function V such that $f(\mu) = d_\mu V$. In this case, f is called exact. It is easy to see that f is exact if and only if the condition

$$\frac{\partial f^i}{\partial \mu_j} = \frac{\partial f^j}{\partial \mu_i} \tag{2.35}$$

holds on $\mathcal{M}_+(I)$ for all $i, j \in I$.

Now we come to vector and covector fields on $\mathcal{P}_+(I)$. The commutative diagram (2.17) allows us to relate sections to each other. Of particular interest are sections in $T^*\mathcal{P}_+(I) = \mathcal{P}_+(I) \times (\mathcal{F}(I)/\mathbb{R})$ (*covector fields*) as well as sections in $T\mathcal{P}_+(I)$ (*vector fields*). As all bundles are of product form $\mathcal{P}_+(I) \times \mathcal{V}$, sections are given by functions $f : \mathcal{P}_+(I) \to \mathcal{V}$. We assume that f is a C^∞ function. We will also use C^∞ extensions $\widetilde{f} : \mathcal{U} \to \mathcal{V}$, where $\mathcal{U}$ is an open subset of $\mathcal{S}(I)$ containing $\mathcal{P}_+(I)$, and $\widetilde{f}|_{\mathcal{U}} = f$. To simplify the notation, we will also use the same symbol f for the extension $\widetilde{f}$. Given a section $f : \mathcal{P}_+(I) \to \mathcal{F}(I)$, we assign various other sections to it:

$$\begin{aligned}
\overline{f} &: \mathcal{P}_+(I) \to \mathbb{R}, && \mu \mapsto \overline{f}(\mu) := \mu(f(\mu)) = \textstyle\sum_i \mu_i\, f^i(\mu),\\
[f] &: \mathcal{P}_+(I) \to (\mathcal{F}(I)/\mathbb{R}), && \mu \mapsto f(\mu) + \mathbb{R},\\
\widetilde{\phi}(f) &: \mathcal{P}_+(I) \to \mathcal{S}(I), && \mu \mapsto f(\mu)\mu = \textstyle\sum_i \mu_i\, f^i(\mu)\, \delta^i,\\
\widehat{f} &: \mathcal{P}_+(I) \to \mathcal{S}_0(I), && \mu \mapsto (f(\mu) - \overline{f}(\mu))\mu = \textstyle\sum_i \mu_i (f^i(\mu) - \overline{f}(\mu))\delta^i.
\end{aligned}$$

In what follows, we consider covector fields given by a differentiable function $V : \mathcal{P}_+(I) \to \mathbb{R}$. The differential in μ is defined as the linear form

$$d_\mu V : T_\mu \mathcal{P}_+(I) \to \mathbb{R}, \qquad a \mapsto d_\mu V(a) = \frac{\partial V}{\partial a}(\mu),$$

which defines a covector field $dV : \mu \mapsto d_\mu V \in T^*_\mu \mathcal{P}_+(I)$. In order to interpret it as a vector in $\mathcal{F}(I)/\mathbb{R}$, consider an extension $\widetilde{V} : \mathcal{U} \to \mathbb{R}$ of V to an open neighborhood of $\mathcal{P}_+(I)$. This yields a corresponding extension $d_\mu \widetilde{V} : \mathcal{S}(I) \to \mathbb{R}$, and according to (2.7) we have

$$d_\mu V = \sum_i \partial_i \widetilde{V}(\mu)\, e_i + \mathbb{R}, \tag{2.36}$$

where $\partial_i \widetilde{V}(\mu) = \frac{\partial \widetilde{V}}{\partial \delta^i}(\mu)$. The Fisher metric $\mathfrak{g}$ allows us to identify $d_\mu V$ with an element of $T_\mu \mathcal{P}_+(I)$ via the map ϕ_μ, the *gradient* of V in μ:

$$\operatorname{grad}_\mu V := \phi_\mu^{-1}(d_\mu V). \tag{2.37}$$

(See (B.22) in Appendix B for the general construction.)

Proposition 2.2 *Let $V : \mathcal{P}_+(I) \to \mathbb{R}$ be a differentiable function, $\mathcal{U}$ an open subset of $\mathcal{S}(I)$ that contains $\mathcal{P}_+(I)$, and $\widetilde{V} : \mathcal{U} \to \mathbb{R}$ a differentiable continuation of V, that is, $\widetilde{V}|_{\mathcal{P}_+(I)} = V$. Then the coordinates of $\operatorname{grad}_\mu V$ with respect to δ^i are given as*

$$(\operatorname{grad}_\mu V)_i = \mu_i \left(\partial_i \widetilde{V}(\mu) - \sum_j \mu_j\, \partial_j \widetilde{V}(\mu) \right), \quad i \in I.$$

Proof This follows from (2.36), (2.37), and the definition of ϕ_μ. Alternatively, one can show this directly: We have to verify

$$\begin{aligned}
&\mathfrak{g}_\mu(\operatorname{grad}_\mu V, a) = d_\mu V(a), \quad a \in T_\mu \mathcal{P}_+(I).\\
&\mathfrak{g}_\mu(\operatorname{grad}_\mu V, a) = \sum_i \frac{1}{\mu_i}\left(\mu_i\left(\partial_i \widetilde{V}(\mu) - \sum_j \mu_j \partial_j \widetilde{V}(\mu)\right)\right) a_i\\
&\qquad = \sum_i a_i\, \partial_i \widetilde{V}(\mu) - \underbrace{\sum_i a_i \sum_j \mu_j \partial_j \widetilde{V}(\mu)}_{=0}\\
&\qquad = \frac{\partial \widetilde{V}}{\partial a}(\mu)\\
&\qquad = \lim_{t\to 0} \frac{\widetilde{V}(\mu + ta) - \widetilde{V}(\mu)}{t}\\
&\qquad = \lim_{t\to 0} \frac{V(\mu + ta) - V(\mu)}{t}\\
&\qquad = \frac{\partial V}{\partial a}(\mu)\\
&\qquad = d_\mu V(a).
\end{aligned}$$

□

Proposition 2.3 *Consider a map* $f : \mathcal{U} \to \mathcal{F}(I)$, $\mu \mapsto f(\mu) = \sum_{i\in I} f^i(\mu)\, e_i$, *defined on a neighborhood of* $\mathcal{P}_+(I)$. *Then the following statements are equivalent*:

(1) *The vector field* $\widehat{f}$ *is a Fisher gradient field on* $\mathcal{P}_+(I)$.
(2) *The covector field* $[f] : \mathcal{P}_+(I) \to \mathcal{F}(I)/\mathbb{R}$, $\mu \mapsto [f](\mu) := f(\mu) + \mathbb{R}$, *is exact, that is, there exists a function* $V : \mathcal{P}_+(I) \to \mathbb{R}$ *satisfying* $d_\mu V = [f](\mu)$.
(3) *The relation*

$$\frac{\partial f^i}{\partial \mu_j} + \frac{\partial f^j}{\partial \mu_k} + \frac{\partial f^k}{\partial \mu_i} = \frac{\partial f^i}{\partial \mu_k} + \frac{\partial f^k}{\partial \mu_j} + \frac{\partial f^j}{\partial \mu_i} \tag{2.38}$$

holds on $\mathcal{P}_+(I)$ *for all* $i, j, k \in I$.

Proof (1) $\Leftrightarrow$ (2) This is clear.

(2) $\Leftrightarrow$ (3) The covector field $f + \mathbb{R}$ is exact if and only if it is closed. The latter property is expressed in local coordinates. Without restriction of generality we assume $I = \{1, \ldots, n, n+1\}$ and choose the coordinate system of Example 2.1.

$$\left.\frac{\partial \varphi^{-1}}{\partial x_i}\right|_{\varphi(p)} = \delta^i - \delta^{n+1}, \quad i = 1, \ldots, n.$$

This family is a basis of $\mathcal{S}_0(I)$. The dual basis in $\mathcal{F}(I)/\mathbb{R}$ is given as

$$e_i + \mathbb{R}, \quad i = 1, \ldots, n.$$

We now express the covector field $[f]$ in these coordinates:

$$f(\mu)+\mathbb{R}=\left(\sum_{i=1}^{n+1} f^i(p)\,e_i\right)+\mathbb{R}$$
$$=\sum_{i=1}^{n}\big(f^i(\mu)-f^{n+1}(\mu)\big)(e_i+\mathbb{R}).$$

The covector field $f+\mathbb{R}$ is closed, if the coefficients f^i-f^{n+1} satisfy the following integrability condition:

$$\frac{\partial(f^i-f^{n+1})}{\partial(\delta^j-\delta^{n+1})}(\mu)=\frac{\partial(f^j-f^{n+1})}{\partial(\delta^i-\delta^{n+1})}(\mu),\quad i,j=1,\ldots,n.$$

This is equivalent to

$$\frac{\partial f^i}{\partial\delta^j}+\frac{\partial f^j}{\partial\delta^{n+1}}+\frac{\partial f^{n+1}}{\partial\delta^i}=\frac{\partial f^j}{\partial\delta^i}+\frac{\partial f^i}{\partial\delta^{n+1}}+\frac{\partial f^{n+1}}{\partial\delta^j}.$$

Replacing $n+1$ by k yields the integrability condition (2.38). □

2.4 The *m*- and *e*-Connections

The tangent bundle $T\mathcal{M}_+(I)$ and the cotangent bundle $T^*\mathcal{M}_+(I)$ are of product structure. Given two points μ and ν in $\mathcal{M}_+(I)$, this allows for the following natural identification of $T_\mu\mathcal{M}_+(I)$ with $T_\nu\mathcal{M}_+(I)$ and $T^*_\mu\mathcal{M}_+(I)$ with $T^*_\nu\mathcal{M}_+(I)$:

$$\widetilde{\Pi}^{(m)}_{\mu,\nu}: T_\mu\mathcal{M}_+(I)\longrightarrow T_\nu\mathcal{M}_+(I),\qquad (\mu,a)\longmapsto(\nu,a),\tag{2.39}$$
$$\widetilde{\Pi}^{(e)}_{\mu,\nu}: T^*_\mu\mathcal{M}_+(I)\longrightarrow T^*_\nu\mathcal{M}_+(I),\qquad (\mu,f)\longmapsto(\nu,f).\tag{2.40}$$

Note that these identifications of fibers is not a consequence of the triviality of the vector bundles only. In general, a trivial vector bundle has no distinguished trivialization. However, in our case the bundles have a natural product structure.

With the bundle isomorphism $\widetilde{\phi}$ (see diagram (2.17)) one can interpret $\widetilde{\Pi}^{(e)}_{\mu,\nu}$ as a parallel transport in $T\mathcal{M}_+(I)$, given by

$$\widetilde{\Pi}^{(e)}_{\mu,\nu}: T_\mu\mathcal{M}_+(I)\longrightarrow T_\nu\mathcal{M}_+(I),\qquad (\mu,a)\longmapsto\big(\nu,(\widetilde{\phi}_\nu^{-1}\circ\widetilde{\phi}_\mu)(a)\big).$$

Here, one has

$$(\widetilde{\phi}_\nu^{-1}\circ\widetilde{\phi}_\mu)(a)=\frac{da}{d\mu}\,\nu=\sum_i \nu_i\,\frac{a_i}{\mu_i}\,\delta^i.$$

One immediately observes the following duality of the two parallel transports with respect to the Fisher metric. With $A = (\mu, a)$ and $B = (\nu, b)$:

$$\mathfrak{g}_\nu\big(\widetilde{\Pi}^{(e)}_{\mu,\nu} A, \widetilde{\Pi}^{(m)}_{\mu,\nu} B\big) = \sum_i \frac{1}{\nu_i}\left(\nu_i \frac{a_i}{\mu_i}\right) b_i = \sum_i \frac{1}{\mu_i} a_i b_i = \mathfrak{g}_\mu(A, B). \tag{2.41}$$

The correspondence of tangent spaces can be encoded more effectively in terms of an affine connection, which is a differential version of the parallel transport that specifies the directional derivative of a vector field in the direction of another vector field. To be more precise, let A and B be two vector fields $\mathcal{M}_+(I) \to T\mathcal{M}_+(I)$. There exist maps $a, b : \mathcal{M}_+(I) \to \mathcal{S}(I)$ satisfying $B_\mu = (\mu, b_\mu)$ and $A_\mu = (\mu, a_\mu)$. With a curve $\gamma : (-\epsilon, \epsilon) \to \mathcal{M}_+(I)$, $\gamma(0) = \mu$ and $\dot\gamma(0) = A_\mu$ the *covariant derivative of B in the direction of A* can be obtained from the parallel transports as follows (see Eq. (B.33) in Appendix B):

$$\widetilde{\nabla}^{(m,e)}_A B\big|_\mu := \lim_{t\to 0} \frac{1}{t}\big(\widetilde{\Pi}^{(m,e)}_{\gamma(t),\mu}(B_{\gamma(t)}) - B_\mu\big) \in T_\mu\mathcal{M}_+(I). \tag{2.42}$$

The pair (2.42) of affine connections $\widetilde{\nabla}^{(m)}$ and $\widetilde{\nabla}^{(e)}$ corresponds to two kinds of straight line, the so-called geodesic, and exponential maps which specify a natural way of locally identifying the tangent space in μ with a neighborhood of μ (in $\mathcal{M}_+(I)$).

Proposition 2.4

(1) *The affine connections $\widetilde{\nabla}^{(m)}$ and $\widetilde{\nabla}^{(e)}$, defined by (2.42), are given by*

$$\widetilde{\nabla}^{(m)}_A B\big|_\mu = \left(\mu, \frac{\partial b}{\partial a_\mu}(\mu)\right),$$

$$\widetilde{\nabla}^{(e)}_A B\big|_\mu = \left(\mu, \frac{\partial b}{\partial a_\mu}(\mu) - \left(\frac{da_\mu}{d\mu}\cdot\frac{db_\mu}{d\mu}\right)\mu\right).$$

(2) *As corresponding (maximal) m- and e-geodesic with initial point $\mu \in \mathcal{M}_+(I)$ and initial velocity $a \in T_\mu\mathcal{M}_+(I)$ we have*

$$\gamma^{(m)} : \left]t^-, t^+\right[\to \mathcal{M}_+(I), \quad t \mapsto \mu + ta,$$

with

$$t^- := -\min\left\{\frac{\mu_i}{a_i} : i \in I, a_i > 0\right\}, \qquad t^+ := \min\left\{\frac{\mu_i}{|a_i|} : i \in I, a_i < 0\right\}$$

(*we use the convention* $\min\emptyset = \infty$), *and*

$$\gamma^{(e)} : \mathbb{R} \to \mathcal{M}_+(I), \qquad t \mapsto \exp\left(t\,\frac{da}{d\mu}\right)\mu.$$

(3) *As corresponding exponential maps* $\widetilde{\exp}^{(m)}$ *and* $\widetilde{\exp}^{(e)}$, *we obtain*

$$\widetilde{\exp}^{(m)} : T \to \mathcal{M}_+(I), \qquad (\mu, a) \mapsto \mu + a, \tag{2.43}$$

with $T := \{(\mu, \nu - \mu) \in T\mathcal{M}_+(I) : \mu, \nu \in \mathcal{M}_+(I)\}$, *and*

$$\widetilde{\exp}^{(e)} : T\mathcal{M}_+(I) \to \mathcal{M}_+(I), \qquad (\mu, a) \mapsto \exp\left(\frac{da}{d\mu}\right)\mu. \tag{2.44}$$

Proof **The *m*-connection:**

$$\begin{aligned}
\widetilde{\nabla}^{(m)}_A B\big|_\mu &= \lim_{t\to 0} \frac{1}{t}\big(\widetilde{\Pi}^{(m)}_{\gamma(t),\mu}(B_{\gamma(t)}) - B_\mu\big) \\
&= \left(\mu, \lim_{t\to 0} \frac{1}{t}(b_{\gamma(t)} - b_\mu)\right) \\
&= \left(\mu, \frac{\partial b}{\partial a_\mu}(\mu)\right).
\end{aligned}$$

In order to get the geodesic of the m-connection we consider the corresponding equation:

$$\ddot{\gamma} = 0 \quad \text{with } \gamma(0) = \mu,\ \dot{\gamma}(0) = a.$$

Its solution is given by

$$t \mapsto \mu + t\,a$$

which is defined for the maximal time interval $]t^-, t^+[$. Setting $t = 1$ gives us the corresponding exponential map $\widetilde{\exp}^{(m)}$.

The *e*-connection: Now we consider the covariant derivative induced by the exponential parallel transport $\widetilde{\Pi}^{(e)}$:

$$\begin{aligned}
\widetilde{\nabla}^{(e)}_A B\big|_\mu &:= \lim_{t\to 0} \frac{1}{t}\big(\widetilde{\Pi}^{(e)}_{\gamma(t),\mu}(B_{\gamma(t)}) - B_\mu\big) \\
&= \left(\mu, \lim_{t\to 0} \frac{1}{t} \sum_i \left(\mu_i \frac{b_{\gamma(t),i}}{\gamma_i(t)} - b_{\mu,i}\right)\delta^i\right) \\
&= \left(\mu, \sum_i \frac{d}{dt}\left\{\mu_i \frac{b_{\gamma(t),i}}{\gamma_i(t)}\right\}\Big|_{t=0} \delta^i\right) \\
&= \left(\mu, \sum_i \left(\frac{\partial b_i}{\partial a_\mu}(\mu) - \frac{1}{\mu_i} a_{\mu,i} b_{\mu,i}\right)\delta^i\right).
\end{aligned}$$

The equation for the corresponding e-geodesic is given as

$$\ddot{\gamma} - \frac{\dot{\gamma}^2}{\gamma} = 0 \quad \text{with } \gamma(0) = \mu,\ \dot{\gamma}(0) = a. \tag{2.45}$$

One can easily verify that the solution of (2.45) is given by the following curve γ:

$$t \mapsto \sum_i \mu_i \, e^{t \frac{a_i}{\mu_i}} \, \delta^i. \tag{2.46}$$

Setting $t = 1$ in (2.46), we obtain the corresponding exponential map $\widetilde{\exp}^{(e)}$ which is defined on the whole tangent bundle $T\mathcal{M}_+(I)$:

$$(\mu, a) \mapsto \exp\left(\frac{da}{d\mu}\right)\mu = \sum_i \mu_i \, e^{\frac{a_i}{\mu_i}} \, \delta^i. \qquad \square$$

In what follows, we restrict the m- and e-connections to the simplex $\mathcal{P}_+(I)$. First consider the m-connection. Given a point $\mu \in \mathcal{P}_+(I)$ and two vector fields $A, B : \mathcal{P}_+(I) \to T\mathcal{P}_+(I)$, we observe that the covariant derivative in μ is already in the tangent space of $\mathcal{P}_+(I)$ in μ, that is, $\widetilde{\nabla}^{(m)}_A B|_\mu \in T_\mu \mathcal{P}_+(I)$. This allows us to define the m-connection on $\mathcal{P}_+(I)$ simply by

$$\nabla^{(m)}_A B\big|_\mu := \widetilde{\nabla}^{(m)}_A B\big|_\mu. \tag{2.47}$$

The situation is different for the e-connection. There, we have in general $\widetilde{\nabla}^{(e)}_A B|_\mu \notin T_\mu \mathcal{P}_+(I)$. In order to obtain an e-connection on the simplex, we have to project $\widetilde{\nabla}^{(e)}_A B|_\mu$ onto $T_\mu \mathcal{P}_+(I)$ with respect to the Fisher metric $\mathfrak{g}_\mu$ in μ, which leads to the following covariant derivative on the simplex (see (2.14)):

$$\nabla^{(e)}_A B\big|_\mu = \left(\mu, \frac{\partial b}{\partial a_\mu}(\mu) - \left(\frac{da_\mu}{d\mu} \cdot \frac{db_\mu}{d\mu}\right)\mu + \mathfrak{g}_\mu(A_\mu, B_\mu)\,\mu\right). \tag{2.48}$$

Proposition 2.5 *Consider the affine connections $\nabla^{(m)}$ and $\nabla^{(e)}$ defined by (2.47) and (2.48), respectively. Then the following holds*:

(1) *The corresponding (maximal) m- and e-geodesic with initial point $\mu \in \mathcal{P}_+(I)$ and initial velocity $a \in T_\mu \mathcal{P}_+(I)$ are given by*

$$\gamma^{(m)} : \left]t^-, t^+\right[\to \mathcal{P}_+(I), \quad t \mapsto \mu + ta,$$

with

$$t^- := -\min\left\{\frac{\mu_i}{a_i} : i \in I, a_i > 0\right\}, \qquad t^+ := \min\left\{\frac{\mu_i}{|a_i|} : i \in I, a_i < 0\right\},$$

and

$$\gamma^{(e)} : \mathbb{R} \to \mathcal{P}_+(I), \qquad t \mapsto \frac{\exp(t\,\frac{da}{d\mu})}{\mu(\exp(t\,\frac{da}{d\mu}))}\,\mu.$$

(2) *As corresponding exponential maps* $\exp^{(m)}$ *and* $\exp^{(e)}$ *we have*

$$\exp^{(m)} : T \to \mathcal{P}_+(I), \qquad (\mu, a) \mapsto \mu + a,$$

with $T := \{(\mu, \nu - \mu) \in T\mathcal{P}_+(I) : \mu, \nu \in \mathcal{P}_+(I)\}$, *and*

$$\exp^{(e)} : T\mathcal{P}_+(I) \to \mathcal{P}_+(I), \qquad (\mu, a) \mapsto \frac{\exp(\frac{da}{d\mu})}{\mu(\exp(\frac{da}{d\mu}))}\,\mu.$$

Proof Clearly, we only have to prove the statements for the e-connection. From the definition (2.48), we obtain the equation for the corresponding e-geodesic:

$$\ddot{\gamma} - \frac{\dot{\gamma}^2}{\gamma} + \gamma \sum_i \frac{\dot{\gamma}_i^2}{\gamma_i} = 0 \quad \text{with } \gamma(0) = \mu,\ \dot{\gamma}(0) = a. \tag{2.49}$$

The solution of (2.49) is given by the following curve γ:

$$t \mapsto \sum_i \frac{\mu_i\, e^{t\,\frac{a_i}{\mu_i}}}{\sum_j \mu_j\, e^{t\,\frac{a_j}{\mu_j}}}\,\delta^i. \tag{2.50}$$

We now verify this: Obviously, we have $\gamma(0) = \mu$. Furthermore, a straightforward calculation gives us

$$\dot{\gamma}_i(t) = \gamma_i(t)\left(\frac{a_i}{\mu_i} - \sum_j \gamma_j(t)\,\frac{a_j}{\mu_j}\right)$$

and

$$\begin{aligned}\ddot{\gamma}_i(t) &= \dot{\gamma}_i(t)\left(\frac{a_i}{\mu_i} - \sum_j \gamma_j(t)\,\frac{a_j}{\mu_j}\right) - \gamma_i(t)\sum_j \dot{\gamma}_j(t)\,\frac{a_j}{\mu_j} \\ &= \gamma_i(t)\left(\frac{a_i}{\mu_i} - \sum_j \gamma_j(t)\,\frac{a_j}{\mu_j}\right)^2 - \gamma_i(t)\sum_j \dot{\gamma}_j(t)\,\frac{a_j}{\mu_j}.\end{aligned}$$

This implies $\dot{\gamma}(0) = a$ and

$$\begin{aligned}&\ddot{\gamma}_i(t) - \frac{\dot{\gamma}_i(t)^2}{\gamma_i} - \gamma_i(t)\sum_j\left(\ddot{\gamma}_j(t) - \frac{\dot{\gamma}_j(t)^2}{\gamma_j(t)}\right) \\ &\quad = -\gamma_i(t)\sum_j \dot{\gamma}_j(t)\,\frac{a_j}{\mu_j} + \gamma_i(t)\sum_j \gamma_j(t)\sum_k \dot{\gamma}_k(t)\,\frac{a_k}{\mu_k} \\ &\quad = 0,\end{aligned}$$

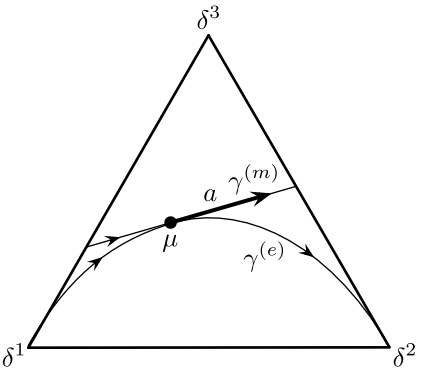

Fig. 2.4 m- and e-geodesic in $\mathcal{P}_+(\{1,2,3\})$ with initial point μ and velocity a

which proves that all conditions (2.49) are satisfied. Setting $t = 1$ in (2.50), we obtain the corresponding exponential map $\exp^{(e)}$ which is defined on the whole tangent bundle $T\mathcal{P}_+(I)$:

$$(\mu, a) \mapsto \frac{\exp(\frac{da}{d\mu})}{\mu(\exp(\frac{da}{d\mu}))}\,\mu = \sum_i \frac{\mu_i e^{\frac{a_i}{\mu_i}}}{\sum_j \mu_j e^{\frac{a_j}{\mu_j}}}\,\delta^i. \qquad \square$$

As an illustration of the m- and e-geodesic of Proposition 2.5(1), see Fig. 2.4.

2.5 The Amari–Chentsov Tensor and the α-Connections

2.5.1 The Amari–Chentsov Tensor

We consider a covariant 3-tensor using the affine connections $\widetilde{\nabla}^{(m)}$ and $\widetilde{\nabla}^{(e)}$: For three vector fields $A : \mu \mapsto A_\mu = (\mu, a_\mu)$, $B : \mu \mapsto B_\mu = (\mu, b_\mu)$, and $C : \mu \mapsto C_\mu = (\mu, c_\mu)$ on $\mathcal{M}_+(I)$, we define

$$\begin{aligned}\mathbf{T}_\mu(A_\mu, B_\mu, C_\mu) &:= \mathfrak{g}_\mu\big(\widetilde{\nabla}^{(m)}_A B\big|_\mu - \widetilde{\nabla}^{(e)}_A B\big|_\mu, C_\mu\big) \\ &= \sum_{i\in I} \mu_i \frac{a_{\mu,i}}{\mu_i}\frac{b_{\mu,i}}{\mu_i}\frac{c_{\mu,i}}{\mu_i}. \end{aligned} \tag{2.51}$$

We refer to this tensor as the *Amari–Chentsov tensor*. Note that for vector fields A, B, C on $\mathcal{P}_+(I)$ and $\mu \in \mathcal{P}_+(I)$ we have

$$\mathbf{T}_\mu(A_\mu, B_\mu, C_\mu) = \mathfrak{g}_\mu\big(\nabla^{(m)}_A B\big|_\mu - \nabla^{(e)}_A B\big|_\mu, C_\mu\big).$$

We have seen that the Fisher metric $\mathfrak{g}$ on $\mathcal{P}_+(I)$ is uniquely characterized in terms of invariance (see Theorem 2.1). Following Chentsov, the same uniqueness property also holds for the tensor $\mathbf{T}$ on $\mathcal{P}_+(I)$, which is the content of the following theorem.

Theorem 2.2 *We assign to each non-empty and finite set I a (non-trivial) covariant 3-tensor S^I on $\mathcal{P}_+(I)$. If for each congruent Markov kernel $K : I \to \mathcal{P}(I')$ we have invariance in the sense that*

$$S^I_\mu(A, B, C) = S^{I'}_{K_*(\mu)}\big(d_\mu K_*(A), d_\mu K_*(B), d_\mu K_*(C)\big)$$

then there is a constant $\alpha > 0$ such that $S^I = \alpha\, \mathbf{T}^I$ for all I, where $\mathbf{T}^I$ denotes the Amari–Chentsov tensor on $\mathcal{P}_+(I)$.[2]

One can prove this theorem by following the same steps as in the proof of Theorem 2.1. Alternatively, it immediately follows from the more general result stated in Theorem 2.3.

By analogy, we can extend the definition (2.51) to a covariant n-tensor for all $n \geq 1$:

$$\begin{aligned}\tau^n_\mu\big(V^{(1)}, V^{(2)}, \ldots, V^{(n)}\big) &:= \sum_{i\in I} \mu_i \frac{v^{(1)}_{\mu,i}}{\mu_i} \frac{v^{(2)}_{\mu,i}}{\mu_i} \cdots \frac{v^{(n)}_{\mu,i}}{\mu_i} \\ &= \sum_{i\in I} \frac{1}{{\mu_i}^{n-1}}\, v^{(1)}_{\mu,i}\, v^{(2)}_{\mu,i} \cdots v^{(n)}_{\mu,i}. \end{aligned} \tag{2.52}$$

Obviously, we have

$$\tau^2 = \mathfrak{g}, \quad \text{and} \quad \tau^3 = \mathbf{T}.$$

It is easy to extend the representation (2.26) of the Fisher metric $\mathfrak{g}$ to the covariant n-tensor τ^n. Given a differentiable manifold M and an embedding $p : M \hookrightarrow \mathcal{M}_+(I)$, one obtains as pullback of τ^n the following covariant n-tensor, defined on M:

$$\tau^n_\xi(V_1, \ldots, V_n) := \sum_i p_i(\xi) \frac{\partial \log p_i}{\partial V_1}(\xi) \cdots \frac{\partial \log p_i}{\partial V_n}(\xi).$$

As suggested by (2.52), the tensor τ^n is closely related to the following multilinear form:

$$\begin{aligned} L^n_I &: \underbrace{\mathcal{F}(I) \times \cdots \times \mathcal{F}(I)}_{n \text{ times}} \to \mathbb{R}, \\ (f_1, \ldots, f_n) &\mapsto L^n_I(f_1, \ldots, f_n) := \sum_i {f_1}^i \cdots {f_n}^i. \end{aligned} \tag{2.53}$$

In order to see this, consider the map

$$\pi^{1/n} : \mathcal{M}_+(I) \to \mathcal{F}(I), \qquad \mu = \sum_i \mu_i\, \delta^i \mapsto \pi^{1/n}(\mu) := n \sum_i {\mu_i}^{\frac{1}{n}}\, e_i.$$

[2] Note that we use the abbreviation $\mathbf{T}$ if corresponding statements are clear without reference to the set I, which is usually the case throughout this book.

This implies

$$L_I^n\left(\frac{\partial \pi^{1/n}}{\partial v^{(1)}}(\mu),\dots,\frac{\partial \pi^{1/n}}{\partial v^{(n)}}(\mu)\right)=\sum_i\left(\mu_i^{-\frac{n-1}{n}}v_i^{(1)}\right)\cdots\left(\mu_i^{-\frac{n-1}{n}}v_i^{(n)}\right)$$
$$=\tau_\mu^n\left(V^{(1)},\dots,V^{(n)}\right).$$

This proves that the tensor τ^n is nothing but the $\pi^{1/n}$-pullback of the multi-linear form L_I^n. In this sense, it is a very natural tensor. Furthermore, for $n=2$ and $n=3$, we have seen that the restrictions of $\mathfrak{g}$ and $\mathbf{T}$ to the simplex $\mathcal{P}_+(I)$ are naturally characterized in terms of their invariance with respect to congruent Markov embeddings (see Theorem 2.1 and Theorem 2.2). This raises the question whether the tensors τ^n on $\mathcal{M}_+(I)$, or their restrictions to $\mathcal{P}_+(I)$, are also characterized by invariance properties. It is easy to see that for all n, τ^n is indeed invariant. However, τ^n are not the only invariant tensors. In fact, Chentsov's results treat the only non-trivial uniqueness cases. Already for $n=2$, Campbell has shown that the metric $\mathfrak{g}$ is not the only one that is invariant if we consider tensors on $\mathcal{M}_+(I)$ rather than on $\mathcal{P}_+(I)$ [57]. Furthermore, for higher n, there are other possible invariant tensors, even when restricting to $\mathcal{P}_+(I)$. For instance, for $n=4$ we can consider the tensors

$$\tau^{\{1,2\},\{3,4\}}(V_1,V_2,V_3,V_4):=\tau^2(V_1,V_2)\,\tau^2(V_3,V_4)=\mathfrak{g}(V_1,V_2)\,\mathfrak{g}(V_3,V_4),$$
$$\tau^{\{1,3\},\{2,4\}}(V_1,V_2,V_3,V_4):=\tau^2(V_1,V_3)\,\tau^2(V_2,V_4)=\mathfrak{g}(V_1,V_3)\,\mathfrak{g}(V_2,V_4),$$
$$\tau^{\{1,4\},\{2,3\}}(V_1,V_2,V_3,V_4):=\tau^2(V_1,V_4)\,\tau^2(V_2,V_3)=\mathfrak{g}(V_1,V_4)\,\mathfrak{g}(V_2,V_3).$$

It is obvious that all of these invariant tensors are mutually different and also different from τ^4. Similarly, for $n=5$ we have, for example,

$$\tau^{\{1,2\},\{3,4,5\}}(V_1,V_2,V_3,V_4,V_5):=\tau^2(V_1,V_2)\,\tau^3(V_3,V_4,V_5)$$
$$=\mathfrak{g}(V_1,V_2)\,\mathbf{T}(V_3,V_4,V_5),$$
$$\tau^{\{1,4\},\{2,3,5\}}(V_1,V_2,V_3,V_4,V_5):=\tau^2(V_1,V_4)\,\tau^3(V_2,V_3,V_5)$$
$$=\mathfrak{g}(V_1,V_4)\,\mathbf{T}(V_2,V_3,V_5).$$

From these examples it becomes evident that for each partition

$$\mathbf{P}=\left\{\left\{i_1^1,\dots,i_1^{n_1}\right\},\dots,\left\{i_l^1,\dots,i_l^{n_l}\right\}\right\}$$

of the set $\{1,\dots,n\}$ with $n=n_1+\cdots+n_l$ one can define an invariant n-tensor $\tau^{\mathbf{P}}(V_1,\dots,V_n)$ in a corresponding fashion, see Definition 2.6 below. Our generalization of Chentsov's uniqueness results, Theorem 2.3, will state that any invariant n-tensor will be a linear combination of these, i.e., the dimension of the space of invariant n-tensors depends on the number of partitions of the set $\{1,\dots,n\}$. In fact, this result will even hold if we consider arbitrary (infinite) measure spaces (see Theorem 5.6).

2.5.2 The α-Connections

The Amari–Chentsov tensor $\mathbf{T}$ is closely related to a family of affine connections, defined as a convex combination of the m- and the e-connections. As in our previous derivations we first consider the affine connections on $\mathcal{M}_+(I)$ and then restrict them to $\mathcal{P}_+(I)$. Given $\alpha \in [-1, 1]$, we define the following convex combination, the *α-connection*:

$$\widetilde{\nabla}^{(\alpha)} := \frac{1-\alpha}{2}\widetilde{\nabla}^{(m)} + \frac{1+\alpha}{2}\widetilde{\nabla}^{(e)} = \widetilde{\nabla}^{(m)} + \frac{1+\alpha}{2}\big(\widetilde{\nabla}^{(e)} - \widetilde{\nabla}^{(m)}\big). \tag{2.54}$$

Obviously, for vector fields A, B, and C we have

$$\mathfrak{g}\big(\widetilde{\nabla}^{(\alpha)}_A B, C\big) = \mathfrak{g}\big(\widetilde{\nabla}^{(m)}_A B, C\big) - \frac{1+\alpha}{2}\,\mathbf{T}(A, B, C).$$

More explicitly, we have

$$\widetilde{\nabla}^{(\alpha)}_A B\big|_\mu = \left(\mu, \sum_i \left(\frac{\partial b_i}{\partial a_\mu}(\mu) - \frac{1+\alpha}{2}\,\frac{a_{\mu,i} b_{\mu,i}}{\mu_i}\right)\delta^i\right). \tag{2.55}$$

This allows us to determine the geodesic and the exponential map of the α-connection. The differential equation for the α-geodesic with initial point μ and initial velocity a follows from (2.55):

$$\ddot{\gamma} - \frac{1+\alpha}{2}\,\frac{\dot{\gamma}^2}{\gamma} = 0, \quad \gamma(0) = \mu,\ \dot{\gamma}(0) = a. \tag{2.56}$$

It is easy to verify that the following curve satisfies this equation:

$$\gamma^{(\alpha)}(t) = \left(\mu^{\frac{1-\alpha}{2}} + t\,\frac{1-\alpha}{2}\,\mu^{-\frac{1+\alpha}{2}} a\right)^{\frac{2}{1-\alpha}}. \tag{2.57}$$

By setting $t = 1$, we can define the corresponding exponential map:

$$\widetilde{\exp}^{(\alpha)} : (\mu, a) \mapsto \left(\mu^{\frac{1-\alpha}{2}} + \frac{1-\alpha}{2}\,\mu^{-\frac{1+\alpha}{2}} a\right)^{\frac{2}{1-\alpha}}. \tag{2.58}$$

Finally, the α-geodesic with initial point μ and endpoint ν has the following more symmetric structure:

$$\gamma^{(\alpha)}(t) = \big((1-t)\,\mu^{\frac{1-\alpha}{2}} + t\,\nu^{\frac{1-\alpha}{2}}\big)^{\frac{2}{1-\alpha}}. \tag{2.59}$$

Now we come to the α-connection $\nabla^{(\alpha)}$ on $\mathcal{P}_+(I)$ by projection of $\widetilde{\nabla}^{(\alpha)}$ (see (2.14)). For $\mu \in \mathcal{P}_+(I)$ and two vector fields A and B that are tangential to $\mathcal{P}_+(I)$, we obtain

as projection

$$\nabla_A^{(\alpha)} B\big|_\mu = \left(\mu, \sum_i \left(\frac{\partial b_i}{\partial a_\mu}(\mu) - \frac{1+\alpha}{2} \left\{ \frac{a_{\mu,i} b_{\mu,i}}{\mu_i} - \mu_i \sum_j \frac{a_{\mu,j} b_{\mu,j}}{\mu_j} \right\} \right) \delta^i \right). \tag{2.60}$$

This implies the following corresponding geodesic equation:

$$\ddot{\gamma} - \frac{1+\alpha}{2} \left\{ \frac{\dot{\gamma}^2}{\gamma} - \gamma \sum_j \frac{\dot{\gamma}_j^2}{\gamma_j} \right\} = 0, \quad \gamma(0) = \mu,\ \dot{\gamma}(0) = a. \tag{2.61}$$

It is reasonable to make a solution ansatz by normalization of the unconstrained geodesic (2.57) and (2.59). However, it turns out that, in order to solve the geodesic Eq. (2.61), both normalized curves have to be reparametrized. More precisely, it has been shown in [187] (Theorems 14.1 and 15.1) that, with appropriate reparametrizations $\tau_{\mu,a}$ and $\tau_{\mu,\nu}$, we have the following forms of the α-geodesic in the simplex $\mathcal{P}_+(I)$:

$$\gamma^{(\alpha)}(t) = \sum_{i \in I} \frac{\mu_i (1 + \tau_{\mu,a}(t) \frac{1-\alpha}{2} \frac{a_i}{\mu_i})^{\frac{2}{1-\alpha}}}{\sum_{j \in I} \mu_j (1 + \tau_{\mu,a}(t) \frac{1-\alpha}{2} \frac{a_j}{\mu_j})^{\frac{2}{1-\alpha}}} \delta^i \tag{2.62}$$

and

$$\gamma^{(\alpha)}(t) = \sum_{i \in I} \frac{(\mu_i^{\frac{1-\alpha}{2}} + \tau_{\mu,\nu}(t)(\nu_i^{\frac{1-\alpha}{2}} - \mu_i^{\frac{1-\alpha}{2}}))^{\frac{2}{1-\alpha}}}{\sum_{j \in I} (\mu_j^{\frac{1-\alpha}{2}} + \tau_{\mu,\nu}(t)(\nu_i^{\frac{1-\alpha}{2}} - \mu_i^{\frac{1-\alpha}{2}}))^{\frac{2}{1-\alpha}}} \delta^i. \tag{2.63}$$

An explicit expression for the reparametrizations $\tau_{\mu,a}$ and $\tau_{\mu,\nu}$ is unknown. In general, we have the following implications:

$$\gamma^{(\alpha)}(0) = \mu, \qquad \frac{d\gamma^{(\alpha)}}{dt}(0) = \dot{\tau}_{\mu,a}(0)\, a = a \quad \Rightarrow \quad \tau_{\mu,a}(0) = 0, \qquad \dot{\tau}_{\mu,a}(0) = 1,$$

and

$$\gamma^{(\alpha)}(0) = \mu, \qquad \gamma^{(\alpha)}_{\mu,\nu}(1) = \nu \quad \Rightarrow \quad \tau_{\mu,\nu}(0) = 0, \qquad \tau_{\mu,\nu}(1) = 1.$$

As the two expressions (2.62) and (2.63) of the geodesic $\gamma^{(\alpha)}$ yield the same velocity a at $t = 0$, we obtain, with $\sum_{i \in I} a_i = 0$,

$$a = \frac{1}{\tau_{\mu,a}(1)} \frac{2}{1-\alpha} \sum_{i \in I} \mu_i \left(\frac{(\frac{\nu_i}{\mu_i})^{\frac{1-\alpha}{2}}}{\sum_{j=1}^n \mu_j (\frac{\nu_j}{\mu_j})^{\frac{1-\alpha}{2}}} - 1 \right) \delta^i \tag{2.64}$$

and

$$a = \dot{\tau}_{\mu,\nu}(0) \frac{2}{1-\alpha} \sum_{i \in I} \mu_i \left(\left(\frac{\nu_i}{\mu_i} \right)^{\frac{1-\alpha}{2}} - \sum_{j \in I} \mu_j \left(\frac{\nu_j}{\mu_j} \right)^{\frac{1-\alpha}{2}} \right) \delta^i. \tag{2.65}$$

A comparison of (2.64) and (2.65) yields

$$\dot{\tau}_{\mu,v}(0)\sum_{j\in I}\mu_j\left(\frac{v_j}{\mu_j}\right)^{\frac{1-\alpha}{2}}=\frac{1}{\tau_{\mu,a}(1)}. \tag{2.66}$$

2.6 Congruent Families of Tensors

In Theorem 2.1 we showed that the Fisher metric $\mathfrak{g}$ on $\mathcal{P}_+(I)$ is characterized by the property that it is invariant under congruent Markov kernels. In this section, we shall generalize this result and give a complete description of all families of covariant n-tensors on $\mathcal{P}_+(I)$ or, more general, on $\mathcal{M}_+(I)$ with this property.

Before doing this, we need to introduce some more notation. Recall that for a non-empty finite set I we defined $\mathcal{S}(I)$ as the *vector space of signed measures on* I on page 26, that is,

$$\mathcal{S}(I)=\left\{\sum_{i\in I}a_i\delta^i : a_i\in\mathbb{R}\right\},$$

where δ^i is the Dirac measure supported at $i\in I$. On this space, we define the norm

$$\|\mu\|_1:=|\mu|(I)=\sum_{i\in I}|a_i|,\quad \text{where } \mu=\sum_{i\in I}a_i\delta^i. \tag{2.67}$$

Remark 2.1 The norm defined in (2.67) is the norm of *total variation*, which we shall define for general measure spaces in (3.1).

Moreover, recall the subsets $\mathcal{P}(I)\subseteq\mathcal{M}(I)\subseteq\mathcal{S}(I)$ introduced in (2.1), where $\mathcal{P}(I)$ denotes the set of probability measures and $\mathcal{M}(I)$ the set of finite measures on I, respectively. By (2.67), we can also write

$$\mathcal{P}(I)=\left\{m\in\mathcal{M}(I):\|m\|_1=1\right\}=\mathcal{M}(I)\cap\mathcal{S}_1(I).$$

By (2.1), $\mathcal{P}_+(I)\subseteq\mathcal{S}_1(I)$ and $\mathcal{M}_+(I)\subseteq\mathcal{S}(I)$ are open subsets, where $\mathcal{S}_1(I)\subseteq\mathcal{S}(I)$ is an affine subspace with underlying vector spaces $\mathcal{S}_0(I)$. Thus, the tangent bundles of these spaces can be naturally given as in (2.9).

For each $\mu\in\mathcal{M}_+(I)$, there is a decomposition of the tangent space

$$T_\mu\mathcal{M}_+(I)=T_\mu\mathcal{P}_+(I)\oplus\mathbb{R}\mu=\mathcal{S}_0(I)\oplus\mathbb{R}\mu. \tag{2.68}$$

Indeed, $T_\mu\mathcal{P}_+(I)=\mathcal{S}_0(I)$ has codimension one in $T_\mu\mathcal{M}_+(I)=\mathcal{S}(I)$, and $\mathbb{R}\mu\cap\mathcal{S}_0(I)=0$.

We also define the projection

$$\pi_I:\mathcal{M}_+(I)\longrightarrow\mathcal{P}_+(I),\qquad \pi_I(\mu)=\frac{1}{\|\mu\|_1}\mu,$$

which rescales an arbitrary finite measure on I to become a probability measure. Obviously, $\pi_I(\mu) = \mu$ if and only if $\mu \in \mathcal{P}_+(I)$. Clearly, π_I is differentiable. To calculate its differential, we let $V \in T_\mu \mathcal{M}_+(I) = \mathcal{S}$, and use

$$d_\mu \pi_I(V) = \frac{d}{dt}\Big|_{t=0} \pi_I(\mu + tV) = \frac{d}{dt}\Big|_{t=0} \frac{1}{\|\mu + tV\|_1}(\mu + tV).$$

If $V \in \mathcal{S}_0(I) \subseteq \mathcal{S}$, then $\|\mu + tV\|_1 = \|\mu\|_1$ by the definition of $\mathcal{S}_0(I)$. On the other hand, if $V = c_0\mu$, then $\pi_I(\mu + tV) = \pi_I(1 + tc_0)\mu = \pi_I(\mu)$ is constant, whence for the differential we obtain

$$d_\mu \pi_I(V) = \begin{cases} \frac{1}{\|\mu\|_1} V & \text{for } V \in T_\mu \mathcal{P}_+(I) = \mathcal{S}_0(I), \\ 0 & \text{for } V \in \mathbb{R}\mu. \end{cases} \tag{2.69}$$

Definition 2.3 (Covariant n-tensors on $\mathcal{M}_+(I)$ and $\mathcal{P}_+(I)$)

(1) A *covariant n-tensor on $\mathcal{P}_+(I)$* is a continuous map

$$\Theta_I^n : \mathcal{P}_+(I) \times \overset{n}{\bigtimes} \mathcal{S}_0(I) \longrightarrow \mathbb{R}, \qquad (\mu; V_1, \ldots, V_n) \longmapsto \big(\Theta_I^n\big)_\mu (V_1, \ldots, V_n)$$

such that $(\Theta_I^n)_\mu$ is n-linear on $\bigtimes^n \mathcal{S}_0(I)$ for fixed $\mu \in \mathcal{P}_+(I)$.

(2) A *covariant n-tensor on $\mathcal{M}_+(I)$* is a continuous map

$$\tilde{\Theta}_I^n : \mathcal{M}_+(I) \times \overset{n}{\bigtimes} \mathcal{S} \longrightarrow \mathbb{R}, \qquad (\mu; V_1, \ldots, V_n) \longmapsto \big(\tilde{\Theta}_I^n\big)_\mu (V_1, \ldots, V_n)$$

such that $(\tilde{\Theta}_I^n)_\mu$ is n-linear on $\bigtimes^n \mathcal{S}$ for fixed $\mu \in \mathcal{M}_+(I)$.

(3) Given a covariant n-tensor Θ_I^n on $\mathcal{P}_+(I)$, we define the *extension of Θ_I^n to $\mathcal{M}_+(I)$* to be the covariant n-tensor

$$\big(\tilde{\Theta}_I^n\big)_\mu (V_1, \ldots, V_n) := \big(\Theta_I^n\big)_{\pi_I(\mu)} \big(d_\mu \pi_I(V_1), \ldots, d_\mu \pi_I(V_n)\big).$$

(4) Given a covariant n-tensor $\tilde{\Theta}_I^n$ on $\mathcal{M}_+(I)$, we define its *restriction to $\mathcal{P}_+(I)$* to be the tensor

$$\big(\tilde{\Theta}_I^n\big)_\mu (V_1, \ldots, V_n) := \big(\tilde{\Theta}_I^n\big)_\mu (V_1, \ldots, V_n).$$

Remark 2.2 By convention, a covariant 0-tensor on $\mathcal{P}_+(I)$ and $\mathcal{M}_+(I)$ is simply a continuous function $\Theta_I^0 : \mathcal{P}_+(I) \to \mathbb{R}$ and $\tilde{\Theta}_I^0 : \mathcal{M}_+(I) \to \mathbb{R}$, respectively.

The extension of Θ_I^n is merely the pull-back of Θ_I^n under the map $\pi_I : \mathcal{M}_+(I) \to \mathcal{P}_+(I)$; the restriction of $\tilde{\Theta}_I^n$ is the pull-back of the inclusion $\mathcal{P}_+(I) \hookrightarrow \mathcal{M}_+(I)$ as defined in (2.25).

In general, in order to describe a covariant n-tensor $\tilde{\Theta}_I^n$ on $\mathcal{M}_+(I)$, we define for a multiindex $\vec{i} := (i_1, \ldots, i_n) \in I^n$

$$\theta_{I;\mu}^{\vec{i}} := \big(\tilde{\Theta}_I^n\big)_\mu \big(\delta^{i_1}, \ldots, \delta^{i_n}\big) =: \big(\tilde{\Theta}_I^n\big)_\mu \big(\delta^{\vec{i}}\big). \tag{2.70}$$

Clearly, these functions are continuous in $\mu \in \mathcal{M}_+(I)$, and they uniquely determine $\tilde{\Theta}^n_I$, since for arbitrary vectors $V_k = \sum_{i\in I} v_{k;i}\delta^i \in \mathcal{S}$ the multilinearity implies

$$\big(\tilde{\Theta}^n_I\big)_\mu(V_1,\dots,V_n) = \sum_{\vec{i}=(i_1,\dots,i_n)\in I^n} \theta^{\vec{i}}_{I;\mu} v_{1,i_1}\cdots v_{n,i_n}. \tag{2.71}$$

Let $K : I \to \mathcal{P}(I')$ be a Markov kernel between the finite sets I and I', as defined in (2.30). Such a map induces a corresponding map between probability distributions as defined in (2.31), and as was mentioned there, this formula also yields a linear map

$$K_* : \mathcal{S}(I) \longrightarrow \mathcal{S}\big(I'\big), \qquad \mu = \sum_{i\in I}\mu_i\delta^i \longmapsto \sum_{i\in I}\mu_i K^i,$$

where

$$K^i := K(i) = \sum_{i'\in I'} K^i_{i'}\delta^{i'}.$$

Clearly, K_* is a linear map between $\mathcal{S}(I)$ and $\mathcal{S}(I')$, and $K^i_{i'} \geq 0$, implies $K_*\mu \in \mathcal{M}(I')$ for all $\mu \in \mathcal{M}(I)$. Moreover, $\sum_{i'\in I'} K^i_{i'} = 1$ implies that for all $\mu \in \mathcal{M}(I)$,

$$\|K_*\mu\|_1 = \Big\| \sum_{i\in I, i'\in I'} \mu_i K^i_{i'}\delta^{i'} \Big\|_1 = \sum_{i\in I, i'\in I'} \mu_i K^i_{i'} = \sum_{i\in I}\mu_i = \|\mu\|_1.$$

That is,

$$\|K_*\mu\|_1 = \|\mu\|_1 \quad \text{for all } \mu \in \mathcal{M}(I). \tag{2.72}$$

This also implies that the image of $\mathcal{P}(I)$ under K_* is contained in $\mathcal{P}(I')$. In particular, it follows that for $\mu \in \mathcal{M}(I)$,

$$K_*(\pi_I\mu) = K_*\left(\frac{1}{\|\mu\|_1}\mu\right) = \frac{1}{\|K_*\mu\|_1}K_*\mu = \pi_{I'}(K_*\mu),$$

i.e.,

$$K_*(\pi_I\mu) = \pi_{I'}(K_*\mu) \quad \text{for all } \mu \in \mathcal{M}(I). \tag{2.73}$$

Definition 2.4 (Tensors invariant under congruent embeddings) A *congruent family of covariant n-tensors* is a collection $\{\tilde{\Theta}^n_I : I \text{ finite}\}$, where $\tilde{\Theta}^n_I$ is a covariant n-tensor on $\mathcal{M}_+(I)$, which is invariant under congruent Markov kernels in the sense that

$$K_*^*\tilde{\Theta}^n_{I'} = \tilde{\Theta}^n_I \tag{2.74}$$

for any congruent Markov kernel $K : I \to \mathcal{P}(I')$ with the definition of the pull-back in (2.25).

A *restricted congruent family of covariant n-tensors* is a collection $\{\Theta_I^n : I \text{ finite}\}$, where Θ_I^n is a covariant n-tensor on $\mathcal{P}_+(I)$, which is invariant under congruent Markov kernels in the sense that (2.74) holds when replacing $\tilde{\Theta}_I^n$ and $\tilde{\Theta}_{I'}^n$ by Θ_I^n and $\Theta_{I'}^n$, respectively.

Proposition 2.6 *There is a correspondence between congruent families of covariant n-tensors and restricted congruent families of covariant n-tensors in the following sense*:

(1) *Let* $\{\tilde{\Theta}_I^n : I \text{ finite}\}$ *be a congruent family of covariant n-tensors, and let* Θ_I^n *be the restriction of* $\tilde{\Theta}_I^n$ *to* $\mathcal{P}_+(I)$.
Then $\{\Theta_I^n : I \text{ finite}\}$ *is a restricted congruent family of covariant n-tensors. Moreover, any restricted congruent family of covariant n-tensors can be described in this way.*

(2) *Let* $\{\Theta_I^n : I \text{ finite}\}$ *be a restricted congruent family of covariant n-tensors, and let* $\tilde{\Theta}_I^n$ *be the extension of* Θ_I^n *to* $\mathcal{M}_+(I)$.
Then $\{\tilde{\Theta}_I^n : I \text{ finite}\}$ *is a congruent family of covariant n-tensors.*

Proof This follows from unwinding the definitions. Namely, if $\{\tilde{\Theta}_I^n : I \text{ finite}\}$ is a congruent family of covariant n-tensors, then the restriction is given as $\Theta_I^n := \tilde{\Theta}_I^n|_{\mathcal{P}(I)}$. Now if $K : I \to \mathcal{P}(I')$ is a congruent Markov kernel, then because of (2.72), K_* maps $\mathcal{P}(I)$ to $\mathcal{P}(I')$, whence if (2.74) holds, it also holds for the restriction of both sides to $\mathcal{P}(I)$ and $\mathcal{P}(I')$, respectively, showing that $\{\Theta_I^n : I \text{ finite}\}$ is a restricted congruent family of covariant n-tensors.

For the second assertion, let $\{\Theta_I^n : I \text{ finite}\}$ be a restricted congruent family of covariant n-tensors. Then the extension of Θ_I^n is given as $\tilde{\Theta}_I^n := \pi_I^* \Theta_I^n$, whence for a congruent Markov kernel $K : I \to \mathcal{P}(I')$ we get from (2.73)

$$K_*^* \tilde{\Theta}_{I'}^n = K_*^* \pi_{I'}^* \Theta_{I'}^n = (\pi_{I'} K_*)^* \Theta_{I'}^n = (K_* \pi_I)^* \Theta_{I'}^n = \pi_I^* K_*^* \Theta_{I'}^n = \pi_I^* \Theta_I^n = \tilde{\Theta}_I^n,$$

so that (2.74) holds. □

In the following, we shall therefore mainly deal with the description of congruent families of covariant n-tensors, since by virtue of Proposition 2.6 this immediately yields a description of all restricted congruent families as well.

Example 2.2 (Congruent families of covariant 0-tensors) Let $\{\tilde{\Theta}_I^0 : \mathcal{M}_+(I) \to \mathbb{R} : I \text{ finite}\}$ be a congruent family of covariant 0-tensors, i.e., of continuous functions $\tilde{\Theta}_I^0 : \mathcal{M}_+(I) \to \mathbb{R}$ (cf. Remark 2.2). Let $\mu \in \mathcal{M}_+(I)$ and $\rho := \pi_I(\mu) \in \mathcal{P}_+(I)$ be the normalization of μ, so that $\mu = \|\mu\|_1 \rho$. Define the congruent embedding determined by

$$K : \mathcal{S}\big(\{0\}\big) \longmapsto \mathcal{S}(I), \qquad \delta^0 \longmapsto \rho.$$

Then by the congruence condition,

$$\tilde{\Theta}_I^0(\mu) = \tilde{\Theta}_I^0\big(\|\mu\|_1 \rho\big) = \tilde{\Theta}_I^0\big(K_*\big(\|\mu\|_1 \delta^0\big)\big) = \tilde{\Theta}_{\{0\}}^0\big(\|\mu\|_1 \delta^0\big) =: a\big(\|\mu\|_1\big).$$

That is, a congruent family of covariant 0-tensors is given as

$$\tilde{\Theta}^0_I(\mu) = a\big(\|\mu\|_1\big) \tag{2.75}$$

for some continuous function $a : (0, \infty) \to \mathbb{R}$. Conversely, every family given as in (2.75) is congruent, since Markov morphisms preserve the total mass by (2.72).

In particular, a restricted congruent family of covariant 0-tensors is given by a constant.

Definition 2.5

(1) Let $\tilde{\Theta}^n_I$ be a covariant n-tensor on $\mathcal{M}_+(I)$ and let σ be a permutation of $\{1, \dots, n\}$. Then the *permutation of* $\tilde{\Theta}^n_I$ *by* σ is defined by

$$\big(\tilde{\Theta}^n_I\big)^\sigma(V_1, \dots, V_n) := \Theta^n_I(V_{\sigma_1}, \dots, V_{\sigma_n}).$$

(2) Let $\tilde{\Theta}^n_I$ and $\tilde{\Psi}^m_I$ be covariant n- and m-tensors on $\mathcal{M}_+(I)$, respectively. Then the *tensor product of* $\tilde{\Theta}^n_I$ *and* $\tilde{\Psi}^m_I$ is the covariant $(n+m)$-tensor on $\mathcal{M}_+(I)$ defined by

$$\big(\tilde{\Theta}^n_I \otimes \tilde{\Psi}^m_I\big)(V_1, \dots, V_{n+m}) := \tilde{\Theta}^n_I(V_1, \dots, V_n) \cdot \tilde{\Psi}^m_I(V_{n+1}, \dots, V_{n+m}).$$

The permutation by σ and the tensor product of covariant tensors on $\mathcal{P}_+(I)$ is defined analogously.

Observe that the tensor product includes multiplication by a continuous function, which is regarded as a covariant 0-tensor.

By the definition of the pull-back K^*_* in (2.25) it follows immediately that

$$\begin{aligned} K^*_*\big(c_1\tilde{\Theta}^n_I + c_2\tilde{\Psi}^n_I\big) &= c_1 K^*_*\big(\tilde{\Theta}^n_I\big) + c_2 K^*_*\big(\tilde{\Psi}^n_I\big), \\ K^*_*\big(\big(\tilde{\Theta}^n_I\big)^\sigma\big) &= \big(K^*_*\big(\tilde{\Theta}^n_I\big)\big)^\sigma, \\ K^*_*\big(\tilde{\Theta}^n_I \otimes \tilde{\Psi}^m_I\big) &= K^*_*\big(\tilde{\Theta}^n_I\big) \otimes K^*_*\big(\tilde{\Psi}^m_I\big). \end{aligned}$$

This implies the following statement.

Proposition 2.7

(1) *Let* $\{\tilde{\Theta}^n_I : I \text{ finite}\}$ *and* $\{\tilde{\Psi}^n_I : I \text{ finite}\}$ *be two congruent families of covariant* n*-tensors. Then any linear combination* $\{c_1\tilde{\Theta}^n_I + c_2\tilde{\Psi}^n_I : I \text{ finite}\}$ *is also a congruent family of covariant* n*-tensors.*

(2) *Let* $\{\tilde{\Theta}^n_I : I \text{ finite}\}$ *be a congruent family of covariant* n*-tensors. Then for any permutation* σ *of* $\{1, \dots, n\}$, $\{(\tilde{\Theta}^n_I)^\sigma : I \text{ finite}\}$ *is a congruent family of covariant* n*-tensors.*

(3) *Let* $\{\tilde{\Theta}^n_I : I \text{ finite}\}$ *and* $\{\tilde{\Psi}^m_I : I \text{ finite}\}$ *be two congruent families of covariant* n*- and* m*-tensors, respectively. Then the tensor product* $\{\tilde{\Theta}^n_I \otimes \tilde{\Psi}^m_I : I \text{ finite}\}$ *is also a congruent family of covariant* $(n+m)$*-tensors.*

The analogous statements hold for restricted congruent family of covariant n-tensors.

The following introduces an important class of congruent families of covariant n-tensors.

Proposition 2.8 *For a finite set I define the* canonical n-tensor τ_I^n *as*

$$\left(\tau_I^n\right)_\mu(V_1,\dots,V_n) := \sum_{i\in I} \frac{1}{m_i^{n-1}} v_{1;i}\cdots v_{n;i}, \tag{2.76}$$

where $V_k = \sum_{i\in I} v_{k;i}\delta^i \in \mathcal{S}(I)$ and $\mu = \sum_{i\in I} m_i\delta^i \in \mathcal{M}_+(I)$. Then $\{\tau_I^n : I \text{ finite}\}$ is a congruent family of covariant n-tensors.

The component functions of this tensor from (2.70) are therefore given as

$$\theta_{I;\mu}^{i_1,\dots,i_n} = \begin{cases} \frac{1}{m_i^{n-1}} & \text{if } i_1 = \cdots = i_n =: i, \\ 0 & \text{otherwise.} \end{cases} \tag{2.77}$$

This is well defined since $m_i > 0$ for all i as $\mu \in \mathcal{M}_+(I)$. Observe that the restriction of τ_I^n to $\mathcal{P}_+(I)$ coincides with the definition in (2.52), so that, in particular, the restriction of τ_I^1 to $\mathcal{P}_+(I)$ vanishes, while τ_I^2 and τ_I^3 are the Fisher metric and the Amari–Chentsov tensor on $\mathcal{P}_+(I)$, respectively.

Proof Let $K : I \to \mathcal{P}(I')$ be a congruent Markov kernel with the partition $(A_i)_{i\in I}$ of I' as defined in (2.32). That is, $K(i) := K_{i'}^i\delta^{i'}$ with $K_{i'}^i = 0$ if $i' \notin A_i$. If $\mu = \sum_{i\in I} m_i\delta^i$, then

$$\mu' := K_*\mu = \sum_{i\in I, i'\in A_i} m_i K_{i'}^i \delta^{i'} =: \sum_{i'\in I'} m'_{i'}\delta^{i'}.$$

Thus,

$$m'_{i'} = m_i K_{i'}^i \quad \text{for the (unique) } i \in I \text{ with } i' \in A_i. \tag{2.78}$$

Then with the notation from before

$$\begin{aligned}\left(\tau_{I'}^n\right)_{\mu'}\left(K_*\delta^{i_1},\dots,K_*\delta^{i_n}\right) &= \left(\tau_{I'}^n\right)_{\mu'}\left(\sum_{i'_1\in A_{i_1}} K_{i'_1}^{i_1}\delta^{i'_1},\dots,\sum_{i'_n\in A_{i_n}} K_{i'_n}^{i_n}\delta^{i'_n}\right) \\ &= \sum_{i'_k\in A_{i_k}} K_{i'_1}^{i_1}\cdots K_{i'_n}^{i_n}\theta_{I';\mu'}^{i'_1,\dots,i'_n}.\end{aligned}$$

By (2.77), the only summands with $\theta_{I';\mu'}^{i'_1,\dots,i'_n} \neq 0$ are those where $i'_1 = \cdots = i'_n =: i'$. If we let $i \in I$ be the index with $i' \in A_i$, then, as K is a congruent Markov morphism, we have $K_{i'}^{i_k} = 0$ unless $i_k = i$.

That is, $K^{i_1}_{i'_1} \cdots K^{i_n}_{i'_n} \theta^{i'_1,\ldots,i'_n}_{I';\mu'} \neq 0$ only if $i'_1 = \cdots = i'_n =: i'$ and $i_1 = \cdots = i_n =: i$ with $i' \in A_i$. In particular, if *not* all of $i_1, \ldots, i_n$ are equal (2.74) holds for $V_k = \delta^{i_k}$, since in this case,

$$\big(\tau^n_{I'}\big)_{\mu'}\big(K_*\delta^{i_1}, \ldots, K_*\delta^{i_n}\big) = 0 = \big(\tau^n_I\big)_{\mu}\big(\delta^{i_1}, \ldots, \delta^{i_n}\big).$$

On the other hand, if $i_1 = \cdots = i_n =: i$, then the above sum reads

$$\begin{aligned}
\big(\tau^n_{I'}\big)_{\mu'}\big(K_*\delta^{i}, \ldots, K_*\delta^{i}\big) &= \sum_{i' \in A_i} K^i_{i'} \cdots K^i_{i'} \theta^{i',\ldots,i'}_{I';\mu'} \\
&\overset{(2.77)}{=} \sum_{i' \in A_i} \big(K^i_{i'}\big)^n \frac{1}{(m'_{i'})^{n-1}} \\
&\overset{(2.78)}{=} \sum_{i' \in A_i} \big(K^i_{i'}\big)^n \frac{1}{(m_i K^i_{i'})^{n-1}} \\
&= \frac{1}{m_i^{n-1}} \sum_{i' \in A_i} K^i_{i'} = \frac{1}{m_i^{n-1}} \underbrace{\sum_{i' \in I'} K^i_{i'}}_{=1} \\
&= \frac{1}{m_i^{n-1}} = \big(\tau^n_I\big)_{\mu}\big(\delta^{i}, \ldots, \delta^{i}\big),
\end{aligned}$$

so that (2.74) holds for $V_1 = \cdots = V_n = \delta^i$ as well. Thus, the n-linearity of the tensors shows that (2.74) always holds, which shows the claim. □

By Propositions 2.7 and 2.8, we can therefore construct further congruent families which we shall now describe in more detail.

For $n \in \mathbb{N}$, we denote by $\mathbf{Part}(n)$ the collection of partitions $\mathbf{P} = \{P_1, \ldots, P_r\}$ of $\{1, \ldots, n\}$, that is, $\bigcup_k P_k = \{1, \ldots, n\}$, and these sets are pairwise disjoint. We denote the number r of sets in the partition by $|\mathbf{P}|$.

Given a partition $\mathbf{P} = \{P_1, \ldots, P_r\} \in \mathbf{Part}(n)$, we associate to it a bijective map

$$\pi_{\mathbf{P}} : \biguplus_{i \in \{1,\ldots,r\}} \big(\{i\} \times \{1, \ldots, n_i\}\big) \longrightarrow \{1, \ldots, n\}, \tag{2.79}$$

where $n_i := |P_i|$, such that $\pi_{\mathbf{P}}(\{i\} \times \{1, \ldots, n_i\}) = P_i$. This map is well defined, up to permutation of the elements in P_i.

$\mathbf{Part}(n)$ is partially ordered by the relation $\mathbf{P} \leq \mathbf{P}'$ if $\mathbf{P}$ is a subdivision of $\mathbf{P}'$. This ordering has the partition $\{\{1\}, \ldots, \{n\}\}$ into singleton sets as its minimum and $\{\{1, \ldots, n\}\}$ as its maximum.

Definition 2.6 (Canonical tensor of a partition) Let $\mathbf{P} \in \mathbf{Part}(n)$ be a partition, and let $\pi_{\mathbf{P}}$ be the bijective map from (2.79). For each finite set I, the *canonical n-tensor*

of $\mathbf{P}$ is the covariant n-tensor defined by

$$\left(\tau_I^{\mathbf{P}}\right)_\mu(V_1,\dots,V_n) := \prod_{i=1}^{r}\left(\tau_I^{n_i}\right)_\mu(V_{\pi_{\mathbf{P}}(i,1)},\dots,V_{\pi_{\mathbf{P}}(i,n_i)}) \tag{2.80}$$

with the canonical tensor $\tau_I^{n_i}$ from (2.76).

Observe that this definition is independent of the choice of the bijection $\pi_{\mathbf{P}}$, since $\tau_I^{k_i}$ is symmetric.

Example 2.3

(1) If $\mathbf{P} = \{\{1,\dots,n\}\}$ is the trivial partition, then

$$\tau_I^{\mathbf{P}} = \tau_I^n.$$

(2) If $\mathbf{P} = \{\{1\},\dots,\{n\}\}$ is the partition into singletons, then

$$\tau_I^{\mathbf{P}}(V_1,\dots,V_n) = \tau_I^1(V_1)\cdots\tau_I^1(V_n).$$

(3) To give a concrete example, let $n = 5$ and $\mathbf{P} = \{\{1,3\},\{2,5\},\{4\}\}$. Then

$$\tau_I^{\mathbf{P}}(V_1,\dots,V_5) = \tau_I^2(V_1,V_3)\cdot\tau_I^2(V_2,V_5)\cdot\tau_I^1(V_4).$$

Observe that the restriction of $\tau^{\mathbf{P}}$ to $\mathcal{P}_+(I)$ vanishes if $\mathbf{P}$ contains a singleton set, since τ_I^1 vanishes on $\mathcal{P}_+(I)$ by (2.52). Thus, the restriction of the last two examples to $\mathcal{P}_+(I)$ vanishes.

Proposition 2.9

(1) *Every family of covariant n-tensors given by*

$$\left(\tilde{\Theta}_I^n\right)_\mu = \sum_{\mathbf{P}\in\mathbf{Part}(n)} a_{\mathbf{P}}\left(\|\mu\|_1\right)\left(\tau_I^{\mathbf{P}}\right)_\mu \tag{2.81}$$

with continuous functions $a_{\mathbf{P}} : (0,\infty) \to \mathbb{R}$ is congruent. Likewise, every family of restricted covariant n-tensors given by

$$\Theta_I^n = \sum_{\substack{\mathbf{P}\in\mathbf{Part}(n)\\ |P_i|\geq 2}} c_{\mathbf{P}}\tau_I^{\mathbf{P}} \tag{2.82}$$

with $c_{\mathbf{P}} \in \mathbb{R}$ is congruent.

(2) *The class of congruent families of (restricted) covariant tensors in (2.81) and (2.82), respectively, is the smallest such class which is closed under taking linear combinations, permutations, and tensor products as described in Proposition 2.7, and which contains the canonical n-tensors $\{\tau_I^n\}$ and the covariant 0-tensors from (2.75).*

(3) *For any congruent family of this class, the functions* $a_{\mathbf{P}}$ *and the constants* $c_{\mathbf{P}}$ *in* (2.81) *and* (2.82)*, respectively, are uniquely determined.*

Proof Evidently, the class of families of (restricted) covariant tensors in (2.81) and (2.82), respectively, is closed under taking linear combinations and permutations. To see that it is closed under taking tensor products, note that

$$\tau_I^{\mathbf{P}} \otimes \tau_I^{\mathbf{P}'} = \tau_I^{\mathbf{P}\cup\mathbf{P}'},$$

where $\mathbf{P}\cup\mathbf{P}' \in \mathbf{Part}(n+m)$ is the partition of $\{1,\dots,n+m\}$ obtained by regarding $\mathbf{P} \in \mathbf{Part}(n)$ and $\mathbf{P}' \in \mathbf{Part}(m)$ as partitions of $\{1,\dots,n\}$ and $\{n+1,\dots,n+m\}$, respectively.

Moreover, if $\mathbf{P} = \{P_1,\dots,P_r\} \in \mathbf{Part}(n)$, we may—after applying a permutation of $\{1,\dots,n\}$—assume that

$$P_1 = \{1,\dots,k_1\},\ P_2 = \{k_1+1,\dots,k_1+k_2\},\ \dots,\ P_r = \{n-k_r+1,\dots,n\},$$

with $k_i = |P_i|$, and in this case, (2.80) and Definition 2.5 imply that

$$\tau_I^{\mathbf{P}} = \big(\tau_I^{k_1}\big) \otimes \big(\tau_I^{k_2}\big) \otimes \cdots \otimes \big(\tau_I^{k_r}\big).$$

Therefore, all (restricted) families given in (2.81) and (2.82), respectively, are congruent by Proposition 2.7, and any class containing the canonical n-tensors and congruent 0-tensors which is closed under linear combinations, permutations and tensor products must contain all families of the form (2.81) and (2.82), respectively. This proves the first two statements.

In order to prove the third part, suppose that

$$\sum_{\mathbf{P}\in\mathbf{Part}(n)} a_{\mathbf{P}}\big(\|\mu\|_1\big)\big(\tau_I^{\mathbf{P}}\big)_\mu = 0 \tag{2.83}$$

for all finite I and $\mu \in \mathcal{M}_+(I)$, but there is a partition $\mathbf{P}_0$ with $a_{\mathbf{P}_0} \not\equiv 0$. In fact, we pick $\mathbf{P}_0$ to be minimal with this property, and choose a multiindex $\vec{i} \in I^n$ with $\mathbf{P}(\vec{i}) = \mathbf{P}_0$. Then

$$\begin{aligned} 0 &= \sum_{\mathbf{P}\in\mathbf{Part}(n)} a_{\mathbf{P}}\big(\|\mu\|_1\big)\big(\tau_I^{\mathbf{P}}\big)_\mu\big(\delta^{\vec{i}}\big) = \sum_{\mathbf{P}\le\mathbf{P}_0} a_{\mathbf{P}}\big(\|\mu\|_1\big)\big(\tau_I^{\mathbf{P}}\big)_\mu\big(\delta^{\vec{i}}\big) \\ &= a_{\mathbf{P}_0}\big(\|\mu\|_1\big)\big(\tau_I^{\mathbf{P}_0}\big)_\mu\big(\delta^{\vec{i}}\big). \end{aligned}$$

The first equation follows since $(\tau_I^{\mathbf{P}})_\mu(\delta^{\vec{i}}) \neq 0$ only if $\mathbf{P} \le \mathbf{P}(\vec{i}) = \mathbf{P}_0$ by Lemma 2.1, whereas the second follows since $a_{\mathbf{P}} \equiv 0$ for $\mathbf{P} < \mathbf{P}_0$ by the minimality assumption on $\mathbf{P}_0$.

But $(\tau_I^{\mathbf{P}_0})_\mu(\delta^{\vec{i}}) \neq 0$ again by Lemma 2.1, since $\mathbf{P}(\vec{i}) = \mathbf{P}_0$, so that $a_{\mathbf{P}_0}(\|\mu\|_1) = 0$ for all μ, contradicting $a_{\mathbf{P}_0} \not\equiv 0$.

Thus, (2.83) occurs only if $a_{\mathbf{P}} \equiv 0$ for all $\mathbf{P}$, showing the uniqueness of the functions $a_{\mathbf{P}}$ in (2.81).

The uniqueness of the constants $c_{\mathbf{P}}$ in (2.82) follows similarly, but we have to account for the fact that $\delta^i \notin \mathcal{S}_0(I) = T_\mu \mathcal{P}_+(I)$. In order to get around this, let I be a finite set and $J := \{0,1,2\} \times I$. For $i \in I$, we define

$$V_i := 2\delta^{(0,i)} - \delta^{(1,i)} - \delta^{(2,i)} \in \mathcal{S}_0(J),$$

and for a multiindex $\vec{i} = (i_1, \dots, i_n) \in I^n$ we let

$$\big(\tau_J^{\mathbf{P}}\big)_\mu\big(V^{\vec{i}}\big) := \big(\tau_J^{\mathbf{P}}\big)_\mu(V_{i_1}, \dots, V_{i_n}).$$

Multiplying this term out, we see that $(\tau_J^{\mathbf{P}})_\mu(V^{\vec{i}})$ is a linear combination of terms of the form $(\tau_J^{\mathbf{P}})_\mu(\delta^{(a_1,i_1)}, \dots, \delta^{(a_n,i_n)})$, where $a_i \in \{0,1,2\}$. Thus, from Lemma 2.1 we conclude that

$$\big(\tau_J^{\mathbf{P}}\big)_\mu\big(V^{\vec{i}}\big) \neq 0 \quad \text{only if } \mathbf{P} \leq \mathbf{P}(\vec{i}). \tag{2.84}$$

Moreover, if $\mathbf{P}(\vec{i}) = \{P_1, \dots, P_r\}$ with $|P_i| = k_i$, then by Definition 2.6 we have

$$\begin{aligned}\big(\tau_J^{\mathbf{P}(\vec{i})}\big)_{c_J}\big(V^{\vec{i}}\big) &= \prod_{i=1}^{r} \big(\tau_J^{k_i}\big)_{c_J}(V_i, \dots, V_i) \\ &= \prod_{i=1}^{r} \big(2^{k_i} + 2(-1)^{k_i}\big)|J|^{k_i - 1} = |J|^{n - |\mathbf{P}(\vec{i})|} \prod_{i=1}^{r} \big(2^{k_i} + 2(-1)^{k_i}\big).\end{aligned}$$

In particular, since $2^{k_i} + 2(-1)^{k_i} > 0$ for all $k_i \geq 2$ we conclude that

$$\big(\tau_J^{\mathbf{P}(\vec{i})}\big)_{c_J}\big(V^{\vec{i}}\big) \neq 0, \tag{2.85}$$

as long as $\mathbf{P}(\vec{i})$ does not contain singleton set.

With this, we can now proceed as in the previous case: assume that

$$\sum_{\mathbf{P} \in \mathbf{Part}(n), |P_i| \geq 2} c_{\mathbf{P}} \tau_I^{\mathbf{P}} = 0 \quad \text{when restricted to } \bigtimes^n \mathcal{S}_0(I), \tag{2.86}$$

for constants $c_{\mathbf{P}}$ which do not all vanish, and we let $\mathbf{P}_0$ be minimal with $c_{\mathbf{P}_0} \neq 0$. Let $\vec{i} = (i_1, \dots, i_n) \in I^n$ be a multiindex with $\mathbf{P}(\vec{i}) = \mathbf{P}_0$, and let $J := \{0,1,2\} \times I$ be as above. Then

$$\begin{aligned}0 &= \sum_{\mathbf{P} \in \mathbf{Part}(n), |P_i| \geq 2} c_{\mathbf{P}} \big(\tau_J^{\mathbf{P}}\big)_\mu\big(V^{\vec{i}}\big) \overset{(2.84)}{=} \sum_{\mathbf{P} \leq \mathbf{P}_0, |P_i| \geq 2} c_{\mathbf{P}} \big(\tau_J^{\mathbf{P}}\big)_\mu\big(V^{\vec{i}}\big) \\ &= c_{\mathbf{P}_0} \big(\tau_J^{\mathbf{P}_0}\big)_\mu\big(V^{\vec{i}}\big),\end{aligned}$$

where the last equality follows by the assumption that $\mathbf{P}_0$ is minimal. But $(\tau_J^{\mathbf{P}_0})_\mu(\vec{V^i}) \neq 0$ by (2.85), whence $c_{\mathbf{P}_0} = 0$, contradicting the choice of $\mathbf{P}_0$.

This shows that (2.86) can happen only if all $c_{\mathbf{P}} = 0$, and this completes the proof. □

In the light of Proposition 2.9, it is thus reasonable to use the following terminology.

Definition 2.7 The class of covariant tensors given in (2.81) and (2.82), respectively, is called *the class of congruent families of (restricted) covariant tensors which is algebraically generated by the canonical n-tensors* $\{\tau_I^n\}$.

We are now ready to state the main result of this section.

Theorem 2.3 (Classification of congruent families of covariant n-tensors) *The class of congruent families of covariant n-tensors on finite sets is the class algebraically generated by the canonical n-tensors* $\{\tau_I^n\}$. *That is, any (restricted) congruent family of covariant n-tensors is of the form* (2.81) *and* (2.82), *respectively.*

For $n = 2$, there are only two partitions, $\{\{1\}, \{2\}\}$ and $\{\{1, 2\}\}$. Thus, in this case the theorem states that each (restricted) congruent family of invariant 2-tensors must be of the form

$$\begin{aligned}\big(\tilde{\Theta}_I^2\big)_\mu(V_1, V_2) &= a\big(\|\mu\|_1\big)\mathfrak{g}(V_1, V_2) + b\big(\|\mu\|_1\big)\tau^1(V_1)\tau^1(V_2),\\ \big(\Theta_I^2\big)_\mu(V_1, V_2) &= c\mathfrak{g}(V_1, V_2).\end{aligned}$$

Therefore, we recover the theorems of Chentsov (cf. Theorem 2.1) and Campbell (cf. [57] or [25]).

In the case $n = 3$, there is no partition with $|P_i| \geq 2$ other than $\{\{1, 2, 3\}\}$, whence it follows that the only restricted congruent family of covariant 3-tensors is—up to multiplication by a constant—the canonical tensor τ_I^3, which coincides with the Amari–Chentsov tensor $\mathbf{T}$ (cf. Theorem 2.2, see also [25] for the non-restricted case).

On the other hand, for $n = 4$, there are several partitions with $|P_i| \geq 2$, hence a restricted congruent family of covariant 4-forms is of the form

$$\begin{aligned}\Theta_I^4(V_1, \ldots, V_4) = c_0\tau^4(V_1, \ldots, V_4) &+ c_1\mathfrak{g}(V_1, V_2)\mathfrak{g}(V_3, V_4)\\ &+ c_2\mathfrak{g}(V_1, V_3)\mathfrak{g}(V_2, V_4) + c_3\mathfrak{g}(V_1, V_4)\mathfrak{g}(V_2, V_3)\end{aligned}$$

for constants $c_0, \ldots, c_3$, where $\mathfrak{g} = \tau_I^2$ is the Fisher metric. Thus, in this case there are more invariant families of such tensors. Evidently, for increasing n, the dimension of the space of invariant families increases rapidly.

The rest of this section will be devoted to the proof of Theorem 2.3 and will be split up into several lemmas.

A multiindex $\vec{i} = (i_1, \dots, i_n) \in I^n$ induces a partition $\mathbf{P}(\vec{i})$ of the set $\{1, \dots, n\}$ into the equivalence classes of the relation $k \sim l \Leftrightarrow i_k = i_l$. For instance, for $n = 6$ and pairwise distinct elements $i, j, k \in I$, the partition induced by $\vec{i} := (j, i, i, k, j, i)$ is

$$\mathbf{P}(\vec{i}) = \big\{\{1, 5\}, \{2, 3, 6\}, \{4\}\big\}.$$

Lemma 2.1 *Let $\tau_I^{\mathbf{P}}$ be the canonical n-tensor from Definition 2.6, and define the center*

$$c_I := \frac{1}{|I|} \sum_{i \in I} \delta^i \in \mathcal{P}_+(I), \tag{2.87}$$

as in the proof of Theorem 2.1. Then for any $\lambda > 0$ we have

$$\big(\tau_I^{\mathbf{P}}\big)_{\lambda c_I}\big(\delta^{\vec{i}}\big) = \begin{cases} (\frac{|I|}{\lambda})^{n-|\mathbf{P}|} & \text{if } \mathbf{P} \le \mathbf{P}(\vec{i}), \\ 0 & \text{otherwise.} \end{cases} \tag{2.88}$$

Proof Let $\mathbf{P} = \{P_1, \dots, P_r\}$ with $|P_i| = k_i$, and let $\pi_{\mathbf{P}}$ be the map from (2.79). Then by (2.80) we have

$$\big(\tau_I^{\mathbf{P}}\big)_\mu\big(\delta^{\vec{i}}\big) = \prod_{i=1}^r \big(\tau_I^{k_i}\big)_\mu\big(\delta^{i_{\pi_{\mathbf{P}}(i,1)}}, \dots, \delta^{i_{\pi_{\mathbf{P}}(i,k_i)}}\big) = \prod_{i=1}^r \theta_{I;\mu}^{i_{\pi_{\mathbf{P}}(i,1)}, \dots, i_{\pi_{\mathbf{P}}(i,k_i)}}.$$

Thus, $(\tau_I^{\mathbf{P}})_\mu(\delta^{\vec{i}}) \neq 0$ if and only if $\theta_{I;\mu}^{i_{\pi_{\mathbf{P}}(i,1)}, \dots, i_{\pi_{\mathbf{P}}(i,k_i)}} \neq 0$ for all i, and by (2.77) this is the case if and only if $i_{\pi_{\mathbf{P}}(i,1)} = \cdots = i_{\pi_{\mathbf{P}}(i,k_i)}$ for all i. But this is equivalent to saying that $\mathbf{P} \le \mathbf{P}(\vec{i})$, showing that $(\tau_I^{\mathbf{P}})_{\lambda c_I}(\delta^{\vec{i}}) = 0$ if $\mathbf{P} \not\le \mathbf{P}(\vec{i})$.

For $\mu = \lambda c_I$, the components m_i of μ all equal $m_i = \lambda/|I|$, whence in this case we have for all multiindices $\vec{i}$ with $\mathbf{P} \le \mathbf{P}(\vec{i})$,

$$\begin{aligned} \big(\tau_I^{\mathbf{P}}\big)_{\lambda c_I}\big(\delta^{\vec{i}}\big) = \prod_{i=1}^r \theta_{I;\lambda c_I}^{i,\dots,i} \overset{(2.77)}{=} \prod_{i=1}^r \left(\frac{|I|}{\lambda}\right)^{k_i - 1} &= \left(\frac{|I|}{\lambda}\right)^{k_1 + \cdots + k_r - r} \\ &= \left(\frac{|I|}{\lambda}\right)^{n-|\mathbf{P}|}, \end{aligned}$$

showing (2.88). □

Now let us suppose that $\{\tilde{\Theta}_I^n : I \text{finite}\}$ is a congruent family of covariant n-tensors, and define $\theta_{I,\mu}^{\vec{i}}$ as in (2.70) and $c_I \in \mathcal{P}_+(I)$ as in (2.87). The following lemma generalizes Step 1 in the proof of Theorem 2.1.

Lemma 2.2 *Let $\{\tilde{\Theta}_I^n : I \text{ finite}\}$ and $\theta_{I,\mu}^{\vec{i}}$ be as before, and let $\lambda > 0$. If $\vec{i}, \vec{j} \in I^n$ are multiindices with $\mathbf{P}(\vec{i}) = \mathbf{P}(\vec{j})$, then*

$$\theta_{I,\lambda c_I}^{\vec{i}} = \theta_{I,\lambda c_I}^{\vec{j}}.$$

Proof If $\mathbf{P}(\vec{i}) = \mathbf{P}(\vec{j})$, then there is a permutation $\sigma : I \to I$ such that $\sigma(i_k) = j_k$ for $k = 1, \ldots, n$. We define the congruent Markov kernel $K : I \to \mathcal{P}(I)$ by $K^i := \delta^{\sigma(i)}$. Then evidently, $K_* c_I = c_I$, and (2.74) implies

$$\begin{aligned}\theta^{\vec{i}}_{I,\lambda c_I} &= \big(\tilde{\Theta}^n_I\big)_{\lambda c_I}\big(\delta^{i_1}, \ldots, \delta^{i_n}\big) \\ &= \big(\tilde{\Theta}^n_I\big)_{K_*(\lambda c_I)}\big(K_*\delta^{i_1}, \ldots, K_*\delta^{i_n}\big) \\ &= \big(\tilde{\Theta}^n_I\big)_{\lambda c_I}\big(\delta^{j_1}, \ldots, \delta^{j_n}\big) = \theta^{\vec{j}}_{I,\lambda c_I},\end{aligned}$$

which shows the claim. □

By virtue of this lemma, we may define

$$\theta^{\mathbf{P}}_{I,\lambda c_I} := \theta^{\vec{i}}_{I,\lambda c_I}, \quad \text{where } \vec{i} \in I^n \text{ is a multiindex with } \mathbf{P}(\vec{i}) = \mathbf{P}.$$

The following two lemmas generalize Step 2 in the proof of Theorem 2.1.

Lemma 2.3 *Let* $\{\tilde{\Theta}^n_I : I \text{ finite}\}$ *and* $\theta^{\mathbf{P}}_{I,\lambda c_I}$ *be as before, and suppose that* $\mathbf{P}_0 \in \mathbf{Part}(n)$ *is a partition such that*

$$\theta^{\mathbf{P}}_{I,\lambda c_I} = 0 \quad \textit{for all } \mathbf{P} < \mathbf{P}_0,\ \lambda > 0 \textit{ and } I. \tag{2.89}$$

Then there is a continuous function $f_{\mathbf{P}_0} : (0, \infty) \to \mathbb{R}$ *such that*

$$\theta^{\mathbf{P}_0}_{I,\lambda c_I} = f_{\mathbf{P}_0}(\lambda)|I|^{n-|\mathbf{P}_0|}. \tag{2.90}$$

Proof Let I, J be finite sets, and let $I' := I \times J$. We define the congruent Markov kernel

$$K : I \longrightarrow \mathcal{P}\big(I'\big), \qquad i \longmapsto \frac{1}{|J|}\sum_{j \in J} \delta^{(i,j)}$$

with the partition $(\{i\} \times J)_{i \in I}$ of I'. Then $K_* c_I = c_{I'}$ is easily verified. Moreover, if $\vec{i} = (i_1, \ldots, i_n) \in I^n$ is a multiindex with $\mathbf{P}(\vec{i}) = \mathbf{P}_0$, then

$$\begin{aligned}\theta^{\mathbf{P}_0}_{I,\lambda c_I} &= \big(\tilde{\Theta}^n_I\big)_{\lambda c_I}\big(\delta^{i_1}, \ldots, \delta^{i_n}\big) \\ &\overset{(2.74)}{=} \big(\tilde{\Theta}^n_{I'}\big)_{K_*(\lambda c_I)}\big(K_*\delta^{i_1}, \ldots, K_*\delta^{i_n}\big) \\ &= \big(\tilde{\Theta}^n_{I'}\big)_{\lambda c_{I'}}\left(\frac{1}{|J|}\sum_{j_1 \in J}\delta^{(i_1,j_1)}, \ldots, \frac{1}{|J|}\sum_{j_n \in J}\delta^{(i_n,j_n)}\right) \\ &= \frac{1}{|J|^n}\sum_{(j_1,\ldots,j_n) \in J^n} \theta^{\mathbf{P}((i_1,j_1),\ldots,(i_n,j_n))}_{I',\lambda c_{I'}}.\end{aligned}$$

Observe that $\mathbf{P}((i_1, j_1), \ldots, (i_n, j_n)) \leq \mathbf{P}(\vec{i}) = \mathbf{P}_0$. If $\mathbf{P}((i_1, j_1), \ldots, (i_n, j_n)) < \mathbf{P}_0$, then $\theta^{\mathbf{P}((i_1,j_1),\ldots,(i_n,j_n))}_{I',\lambda c_{I'}} = 0$ by (2.89).

Moreover, there are $|J|^{|\mathbf{P}_0|}$ multiindices $(j_1, \ldots, j_n) \in J^n$ for which $\mathbf{P}((i_1, j_1), \ldots, (i_n, j_n)) = \mathbf{P}_0$, and since for all of these $\theta^{\mathbf{P}((i_1,j_1),\ldots,(i_n,j_n))}_{I',\lambda c_{I'}} = \theta^{\mathbf{P}_0}_{I',\lambda c_{I'}}$, we obtain

$$\theta^{\mathbf{P}_0}_{I,\lambda c_I} = \frac{1}{|J|^n} \sum_{(j_1,\ldots,j_n)\in J^n} \theta^{\mathbf{P}((i_1,j_1),\ldots,(i_n,j_n))}_{I',\lambda c_{I'}} = \frac{|J|^{|\mathbf{P}_0|}}{|J|^n} \theta^{\mathbf{P}_0}_{I',\lambda c_{I'}},$$

and since $|I'| = |I||J|$, it follows that

$$\frac{1}{|I|^{n-|\mathbf{P}_0|}} \theta^{\mathbf{P}_0}_{I,\lambda c_I} = \frac{1}{|I|^{n-|\mathbf{P}_0|}} \left(\frac{1}{|J|^{n-|\mathbf{P}_0|}} \theta^{\mathbf{P}_0}_{I',\lambda c_{I'}} \right) = \frac{1}{|I'|^{n-|\mathbf{P}_0|}} \theta^{\mathbf{P}_0}_{I',\lambda c_{I'}}.$$

Interchanging the roles of I and J in the previous arguments, we also get

$$\frac{1}{|J|^{n-|\mathbf{P}_0|}} \theta^{\mathbf{P}_0}_{J,\lambda c_J} = \frac{1}{|I'|^{n-|\mathbf{P}_0|}} \theta^{\mathbf{P}_0}_{I',\lambda c_{I'}} = \frac{1}{|I|^{n-|\mathbf{P}_0|}} \theta^{\mathbf{P}_0}_{I,\lambda c_I},$$

whence $f_{\mathbf{P}_0}(\lambda) := \frac{1}{|I|^{n-|\mathbf{P}_0|}} \theta^{\mathbf{P}_0}_{I,\lambda c_I}$ is indeed independent of the choice of the finite set I. □

Lemma 2.4 *Let $\{\tilde{\Theta}^n_I : I \text{ finite}\}$ and $\lambda > 0$ be as before. Then there is a congruent family $\{\tilde{\Psi}^n_I : I \text{ finite}\}$ of the form (2.81) such that*

$$\left(\tilde{\Theta}^n_I - \tilde{\Psi}^n_I\right)_{\lambda c_I} = 0 \quad \text{for all finite sets } I \text{ and all } \lambda > 0.$$

Proof For a congruent family of covariant n-tensors $\{\tilde{\Theta}^n_I : I \text{ finite}\}$, we define

$$N\left(\{\tilde{\Theta}^n_I\}\right) := \left\{\mathbf{P} \in \mathbf{Part}(n) : \left(\tilde{\Theta}^n_I\right)_{\lambda c_I}\left(\delta^{\vec{i}}\right) = 0 \text{ whenever } \mathbf{P}(\vec{i}) \leq \mathbf{P}\right\}.$$

If $N(\{\tilde{\Theta}^n_I\}) \subsetneq \mathbf{Part}(n)$, then let

$$\mathbf{P}_0 = \{P_1, \ldots, P_r\} \in \mathbf{Part}(n) \backslash N\left(\{\tilde{\Theta}^n_I\}\right)$$

be a minimal element, i.e., such that $\mathbf{P} \in N(\{\tilde{\Theta}^n_I\})$ for all $\mathbf{P} < \mathbf{P}_0$. In particular, for this partition (2.89) and hence (2.90) holds. Let

$$\left(\tilde{\Theta}'^n_I\right)_\mu := \left(\tilde{\Theta}^n_I\right)_\mu - \|\mu\|_1^{n-|\mathbf{P}_0|} f_{\mathbf{P}_0}(\|\mu\|_1)\left(\tau^{\mathbf{P}_0}_I\right)_\mu \tag{2.91}$$

with the function $f_{\mathbf{P}_0}$ from (2.90). Then $\{\tilde{\Theta}'^n_I : I \text{ finite}\}$ is again a covariant family of covariant n-tensors.

Let $\mathbf{P} \in N(\{\tilde{\Theta}^n_I\})$ and $\vec{i}$ be a multiindex with $\mathbf{P}(\vec{i}) \leq \mathbf{P}$. If $(\tau^{\mathbf{P}_0}_I)_{\lambda c_I}(\delta^{\vec{i}}) \neq 0$, then by Lemma 2.1 we would have $\mathbf{P}_0 \leq \mathbf{P}(\vec{i}) \leq \mathbf{P} \in N(\{\tilde{\Theta}^n_I\})$ which would imply that $\mathbf{P}_0 \in N(\{\tilde{\Theta}^n_I\})$, contradicting the choice of $\mathbf{P}_0$.

Thus, $(\tau_I^{\mathbf{P}_0})_{\lambda c_I}(\delta^{\vec{i}}) = 0$ and hence $(\tilde{\Theta}'^n_I)_{\lambda c_I}(\delta^{\vec{i}}) = 0$ whenever $\mathbf{P}(\vec{i}) \le \mathbf{P}$, showing that $\mathbf{P} \in N(\{\tilde{\Theta}'^n_I\})$.

Thus, what we have shown is that $N(\{\tilde{\Theta}^n_I\}) \subseteq N(\{\tilde{\Theta}'^n_I\})$. On the other hand, if $\mathbf{P}(\vec{i}) = \mathbf{P}_0$, then again by Lemma 2.1

$$\left(\tau_I^{\mathbf{P}_0}\right)_{\lambda c_I}\left(\delta^{\vec{i}}\right) = \left(\frac{|I|}{\lambda}\right)^{n-|\mathbf{P}_0|},$$

and since $\|\lambda c_I\|_1 = \lambda$, it follows that

$$\begin{aligned}\left(\tilde{\Theta}'^n_I\right)_{\lambda c_I}\left(\delta^{\vec{i}}\right) &\overset{(2.91)}{=} \left(\tilde{\Theta}^n_I\right)_{\lambda c_I}\left(\delta^{\vec{i}}\right) - \lambda^{n-|\mathbf{P}_0|} f_{\mathbf{P}_0}(\lambda)\left(\tau_I^{\mathbf{P}_0}\right)_{\lambda c_I}\left(\delta^{\vec{i}}\right) \\ &= \theta^{\mathbf{P}_0}_{I,\lambda c_I} - \lambda^{n-|\mathbf{P}_0|} f_{\mathbf{P}_0}(\lambda)\left(\frac{|I|}{\lambda}\right)^{n-|\mathbf{P}_0|} \\ &= \theta^{\mathbf{P}_0}_{I,\lambda c_I} - f_{\mathbf{P}_0}(\lambda)|I|^{n-|\mathbf{P}_0|} \overset{(2.90)}{=} 0.\end{aligned}$$

That is, $(\tilde{\Theta}'^n_I)_{\lambda c_I}(\delta^{\vec{i}}) = 0$ whenever $\mathbf{P}(\vec{i}) = \mathbf{P}_0$. If $\vec{i}$ is a multiindex with $\mathbf{P}(\vec{i}) < \mathbf{P}_0$, then $\mathbf{P}(\vec{i}) \in N(\{\tilde{\Theta}'^n_I\})$ by the minimality of $\mathbf{P}_0$, so that $\tilde{\Theta}^n_I(\delta^{\vec{i}}) = 0$. Moreover, $(\tau_I^{\mathbf{P}_0})_{\lambda c_I}(\delta^{\vec{i}}) = 0$ by Lemma 2.1, whence

$$\left(\tilde{\Theta}'^n_I\right)_{\lambda c_I}\left(\delta^{\vec{i}}\right) = 0 \quad \text{whenever } \mathbf{P}(\vec{i}) \le \mathbf{P}_0,$$

showing that $\mathbf{P}_0 \in N(\{\tilde{\Theta}'^n_I\})$. Therefore,

$$N\left(\left\{\tilde{\Theta}^n_I\right\}\right) \subsetneq N\left(\left\{\tilde{\Theta}'^n_I\right\}\right).$$

What we have shown is that given a congruent family of covariant n-tensors $\{\tilde{\Theta}^n_I\}$ with $N(\{\tilde{\Theta}^n_I\}) \subsetneq \mathbf{Part}(n)$, we can enlarge $N(\{\tilde{\Theta}^n_I\})$ by subtracting a multiple of the canonical tensor of some partition. Repeating this finitely many times, we conclude that for some congruent family $\{\tilde{\Psi}^n_I\}$ of the form (2.81)

$$N\left(\left\{\tilde{\Theta}^n_I - \tilde{\Psi}^n_I\right\}\right) = \mathbf{Part}(n),$$

and this implies by definition that $(\tilde{\Theta}^n_I - \tilde{\Psi}^n_I)_{\lambda c_I} = 0$ for all I and all $\lambda > 0$. □

Finally, the next lemma generalizes Step 3 in the proof of Theorem 2.1.

Lemma 2.5 *Let $\{\tilde{\Theta}^n_I : I \text{ finite}\}$ be a congruent family of covariant n-tensors such that $(\tilde{\Theta}^n_I)_{\lambda c_I} = 0$ for all I and $\lambda > 0$. Then $\tilde{\Theta}^n_I = 0$ for all I.*

Proof The proof of Step 3 in Theorem 2.1 carries over almost literally. Namely, consider $\mu \in \mathcal{M}_+(I)$ such that $\pi_I(\mu) = \mu/\|\mu\|_1 \in \mathcal{P}_+(I)$ has rational coefficients,

i.e.,

$$\mu = \|\mu\|_1 \sum_i \frac{k_i}{n} \delta^i$$

for some $k_i, n \in \mathbb{N}$ and $\sum_{i \in I} k_i = n$. Let

$$I' := \biguplus_{i \in I} \big(\{i\} \times \{1, \dots, k_i\}\big),$$

so that $|I'| = n$, and consider the congruent Markov kernel

$$K : i \longmapsto \frac{1}{k_i} \sum_{j=1}^{k_i} \delta^{(i,j)}.$$

Then

$$K_* \mu = \|\mu\|_1 \sum_i \frac{k_i}{n} \left(\frac{1}{k_i} \sum_{j=1}^{k_i} \delta^{(i,j)} \right) = \|\mu\|_1 \frac{1}{n} \sum_i \sum_{j=1}^{k_i} \delta^{(i,j)} = \|\mu\|_1 c_{I'}.$$

Thus, (2.74) implies

$$\big(\tilde{\Theta}^n_I\big)_\mu (V_1, \dots, V_n) = \underbrace{\big(\tilde{\Theta}^n_{I'}\big)_{\|\mu\|_1 c_{I'}}}_{=0} (K_* V_1, \dots, K_* V_n) = 0,$$

so that $(\tilde{\Theta}^n_I)_\mu = 0$ whenever $\pi_I(\mu)$ has rational coefficients. But these μ form a dense subset of $\mathcal{M}_+(I)$, whence $(\tilde{\Theta}^n_I)_\mu = 0$ for *all* $\mu \in \mathcal{M}_+(I)$, which completes the proof. □

Proof of Theorem 2.3 Let $\{\tilde{\Theta}^n_I : I \text{ finite}\}$ be a congruent family of covariant n-tensors. By Lemma 2.4 there is a congruent family $\{\tilde{\Psi}^n_I : I \text{ finite}\}$ of the form (2.81) such that $(\tilde{\Theta}^n_I - \tilde{\Psi}^n_I)_{\lambda c_I} = 0$ for all finite I and all $\lambda > 0$.

Since $\{\tilde{\Theta}^n_I - \tilde{\Psi}^n_I : I \text{ finite}\}$ is again a congruent family, Lemma 2.5 implies that $\tilde{\Theta}^n_I - \tilde{\Psi}^n_I = 0$ and hence $\tilde{\Theta}^n_I = \tilde{\Psi}^n_I$ is of the form (2.81), showing the first part of Theorem 2.3.

For the second part, observe that by Proposition 2.6 any restricted congruent family of covariant n-tensors is the restriction of a congruent family of n-tensors, that is, by the first part of the theorem, the restriction of a family of the form (2.81). This restriction takes the form (2.82) with $c_{\mathbf{P}} := a_{\mathbf{P}}(1)$, observing that the restriction of $\tau^{\mathbf{P}}_I$ with a partition containing a singleton set vanishes as τ^1_I vanishes when restricted to $\mathcal{P}_+(I)$. □

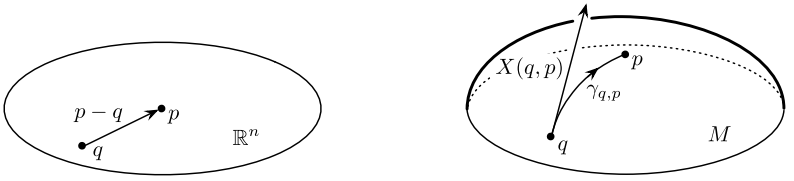

Fig. 2.5 Illustration of (**A**) the difference vector $p - q$ in $\mathbb{R}^n$ pointing from q to p; and (**B**) the difference vector $X(q, p) = \dot{\gamma}_{q,p}(0)$ as the inverse of the exponential map in q

2.7 Divergences

In this section, we derive distance-like functions, so-called divergences, that are naturally associated with a manifold, equipped with a Riemannian metric g and an affine connection ∇, possibly different from the Levi-Civita connection of g. In our context, this will lead to the relative entropy and its extensions, the α-divergences, on $\mathcal{M}_+(I)$. These divergences are special cases of canonical divergences which will be defined in Sect. 4.3.

2.7.1 Gradient-Based Approach

We begin our motivation of a divergence in terms of a simple example where the manifold is $\mathbb{R}^n$, equipped with the standard Euclidean metric and its corresponding connection, the Levi-Civita connection. Consider a point $p \in \mathbb{R}^n$, and the vector field pointing to p (see Fig. 2.5(A)):

$$\mathbb{R}^n \to \mathbb{R}^n, \qquad q \mapsto p - q. \tag{2.92}$$

Obviously, the difference field (2.92) can be seen as the negative gradient of the squared distance

$$D_p : \mathbb{R}^n \to \mathbb{R}, \qquad q \mapsto D_p(q) := D(p \,\|\, q) := \frac{1}{2}\|p - q\|^2 = \frac{1}{2}\sum_{i=1}^{n}(p_i - q_i)^2,$$

that is,

$$p - q = -\operatorname{grad}_q D_p. \tag{2.93}$$

Here, the gradient grad_q is taken with respect to the canonical inner product on $\mathbb{R}^n$.

We shall now generalize the relation (2.93) between the squared distance D_p and the difference of two points p and q to the more general setting of a differentiable manifold M. Given a point $p \in M$, we want to define a vector field $q \mapsto X(q, p)$, at least in a neighborhood of p, that corresponds to the difference vector field (2.92).

Obviously, the problem is that the difference $p - q$ is not naturally defined for a general manifold M. We need an affine connection ∇ in order to have a notion of a difference. Given such a connection ∇, for each point $q \in M$ and each direction $X \in T_q M$ we consider the geodesic $\gamma_{q,X}$, with the initial point q and the initial velocity X, that is, $\gamma_{q,X}(0) = q$ and $\dot{\gamma}_{q,X}(0) = X$ (see (B.38) in Appendix B). If $\gamma_{q,X}(t)$ is defined for all $0 \leq t \leq 1$, the endpoint $p = \gamma_{q,X}(1)$ is interpreted as the result of a translation of the point q along a straight line in the direction of the vector X. The collection of all these translations is summarized in terms of the exponential map

$$\exp_q : U_q \to M, \qquad X \mapsto \gamma_{q,X}(1), \tag{2.94}$$

where $U_q \subseteq T_q M$ denotes the set of tangent vectors X, for which the domain of $\gamma_{q,X}$ contains the unit interval [0, 1] (see (B.39) and (B.40) in Appendix B).

Given two points p and q, one can interpret any X with $\exp_q(X) = p$ as a difference vector X that translates q to p. For simplicity, we assume the existence and uniqueness of such a difference vector, denoted by $X(q, p)$ (see Fig. 2.5(B)).

This is a strong assumption, which is, however, always locally satisfied. On the other hand, although being quite restrictive in general, this property will be satisfied in our information-geometric context, where g is given by the Fisher metric and ∇ is given by the m- and e-connections and their convex combinations, the α-connections. We shall consider these special but important cases in Sects. 2.7.2 and 2.7.3.

If we attach to each point $q \in M$ the difference vector $X(q, p)$, we obtain a vector field that corresponds to the vector field (2.92) in $\mathbb{R}^n$. In order to interpret the vector field $q \mapsto X(q, p)$ as a negative gradient field of a (squared) distance function, and thereby generalize (2.93), we need a Riemannian metric g on M. We search for a function D_p satisfying

$$X(q, p) = -\operatorname{grad}_q D_p, \tag{2.95}$$

where the Riemannian gradient is taken with respect to g (see Appendix B). Obviously, we may set $D_p(p) = 0$. In order to recover D_p from (2.95), we consider any curve $\gamma : [0, 1] \to M$ that connects q with p, that is, $\gamma(0) = q$ and $\gamma(1) = p$, and integrate the inner product of the curve velocity $\dot{\gamma}(t)$ with the vector $X(\gamma(t), p)$ along the curve:

$$\begin{aligned}
\int_0^1 \big\langle X\big(\gamma(t), p\big), \dot{\gamma}(t)\big\rangle dt &= -\int_0^1 \big\langle \operatorname{grad}_{\gamma(t)} D_p, \dot{\gamma}(t)\big\rangle dt \\
&= -\int_0^1 \big(d_{\gamma(t)} D_p\big)\big(\dot{\gamma}(t)\big)\, dt \\
&= -\int_0^1 \frac{d\, D_p \circ \gamma}{d\, t}(t)\, dt \\
&= D_p\big(\gamma(0)\big) - D_p\big(\gamma(1)\big) \\
&= D_p(q) - D_p(p) = D_p(q).
\end{aligned} \tag{2.96}$$

This defines, at least locally, a function D_p that is assigned to the Riemannian metric g and the connection ∇. In what follows, we shall mainly use the standard notation $D(p \,\|\, q) = D_p(q)$ of a divergence as a function D of two arguments.

2.7.2 The Relative Entropy

Now we apply the idea of Sect. 2.7.1 in order to define divergences for the m- and e-connections on the cone $\mathcal{M}_+(I)$ of positive measures. We consider a measure $\mu \in \mathcal{M}_+(I)$ and define two vector fields on $\mathcal{M}_+(I)$ as the inverse of the exponential maps given by (2.43) and (2.44):

$$\begin{aligned} \nu &\mapsto \widetilde{X}^{(m)}(\nu, \mu) := \sum_{i \in I} \nu_i \left(\frac{\mu_i}{\nu_i} - 1 \right) \delta^i, \\ \nu &\mapsto \widetilde{X}^{(e)}(\nu, \mu) := \sum_{i \in I} \nu_i \log \frac{\mu_i}{\nu_i} \, \delta^i. \end{aligned} \tag{2.97}$$

We can easily verify that these vector fields are gradient fields: The functions

$$f^i(\nu) := \frac{\mu_i}{\nu_i} - 1 \quad \text{and} \quad g^i(\nu) := \log \frac{\mu_i}{\nu_i}$$

trivially satisfy the integrability condition (2.35), that is, $\frac{\partial f^i}{\partial \nu_j} = \frac{\partial f^j}{\partial \nu_i}$ and $\frac{\partial g^i}{\partial \nu_j} = \frac{\partial g^j}{\partial \nu_i}$ for all i, j. Therefore, for both connections there are corresponding divergences that satisfy Eq. (2.95).

We derive the divergence of the m-connection first, which we denote by $D^{(m)}$. We consider a curve $\gamma : [0, 1] \to \mathcal{M}_+(I)$ connecting ν with μ, that is, $\gamma(0) = \nu$ and $\gamma(1) = \mu$. This implies

$$\left\langle \widetilde{X}^{(m)}\big(\gamma(t), \mu\big), \dot{\gamma}(t) \right\rangle = \sum_{i \in I} \frac{1}{\gamma_i(t)} \big(\mu_i - \gamma_i(t)\big) \, \dot{\gamma}_i(t) \tag{2.98}$$

and

$$\begin{aligned} D^{(m)}(\mu \,\|\, \nu) &= \int_0^1 \left\langle \widetilde{X}^{(m)}\big(\gamma(t), \mu\big), \dot{\gamma}(t) \right\rangle dt \\ &= \sum_{i \in I} \int_0^1 \frac{1}{\gamma_i(t)} \big(\mu_i - \gamma_i(t)\big) \, \dot{\gamma}_i(t) \, dt \\ &= \sum_{i \in I} \big[\mu_i \log \gamma_i(t) - \gamma_i(t) \big]_0^1 \\ &= \sum_{i \in I} (\mu_i \log \mu_i - \mu_i - \mu_i \log \nu_i + \nu_i) \\ &= \sum_{i \in I} \left(\nu_i - \mu_i + \mu_i \log \frac{\mu_i}{\nu_i} \right). \end{aligned}$$

With the same calculation for the e-connection, we obtain the corresponding divergence, which we denote by $D^{(e)}$. Again, we consider a curve γ connecting ν with μ. This implies

$$\left\langle \widetilde{X}^{(e)}\big(\gamma(t),\mu\big),\dot{\gamma}(t)\right\rangle = \sum_{i\in I}\dot{\gamma}_i(t)\,\log\frac{\mu_i}{\gamma_i(t)} \tag{2.99}$$

and

$$\begin{aligned} D^{(e)}(\mu\,\|\,\nu) &= \int_0^1 \left\langle \widetilde{X}^{(e)}\big(\gamma(t),\mu\big),\dot{\gamma}(t)\right\rangle dt \\ &= \sum_{i\in I}\int_0^1 \dot{\gamma}_i(t)\,\log\frac{\mu_i}{\gamma_i(t)}\,dt \\ &= \sum_{i\in I}\left[\gamma_i(t)\left(1+\log\frac{\mu_i}{\gamma_i(t)}\right)\right]_0^1 \\ &= \sum_{i\in I}\left(\mu_i-\nu_i\left(1+\log\frac{\mu_i}{\nu_i}\right)\right) \\ &= \sum_{i\in I}\left(\mu_i-\nu_i+\nu_i\,\log\frac{\nu_i}{\mu_i}\right) \\ &= D^{(m)}(\nu\,\|\,\mu). \end{aligned}$$

These calculations give rise to the following definition:

Definition 2.8 (Kullback–Leibler divergence ([155, 156])) The function $D_{KL}:\mathcal{M}_+(I)\times\mathcal{M}_+(I)\to\mathbb{R}$ defined by

$$D_{KL}(\mu\,\|\,\nu) := \sum_{i\in I}\nu_i - \sum_{i\in I}\mu_i + \sum_{i\in I}\mu_i\log\frac{\mu_i}{\nu_i} \tag{2.100}$$

is called the *relative entropy*, *information divergence*, or *Kullback–Leibler divergence* (*KL-divergence*). Its restriction to the set of probability distributions is given by

$$D_{KL}(\mu\,\|\,\nu) = \sum_{i\in I}\mu_i\log\frac{\mu_i}{\nu_i}. \tag{2.101}$$

Proposition 2.10 *The following holds*:

$$\begin{aligned} \widetilde{X}^{(m)}(\nu,\mu) &= -\operatorname{grad}_\nu\, D_{KL}(\mu\,\|\,\cdot), \\ \widetilde{X}^{(e)}(\nu,\mu) &= -\operatorname{grad}_\nu\, D_{KL}(\cdot\,\|\,\mu), \end{aligned} \tag{2.102}$$

where D_{KL} is given by (2.100) in Definition 2.8. Furthermore, D_{KL} is the only function on $\mathcal{M}_+(I) \times \mathcal{M}_+(I)$ that satisfies the conditions (2.102) and $D_{KL}(\mu \,\|\, \mu) = 0$ for all μ.

Proof The statement is obvious from the way we introduced D_{KL} as a potential function of the gradient field $\nu \mapsto \widetilde{X}^{(m)}(\nu, \mu)$ and $\nu \mapsto \widetilde{X}^{(e)}(\nu, \mu)$, respectively. The following is an alternative direct verification. We first compute the partial derivatives:

$$\frac{\partial D_{KL}(\mu \,\|\, \cdot)}{\partial \nu_i}(\nu) = -\frac{\mu_i}{\nu_i} + 1, \qquad \frac{\partial D_{KL}(\cdot \,\|\, \mu)}{\partial \nu_i}(\nu) = -\log \frac{\mu_i}{\nu_i}.$$

With the formula (2.34), we obtain

$$\begin{aligned} \big(\operatorname{grad}_\nu D_{KL}(\mu \,\|\, \cdot)\big)_i &= \nu_i \left(-\frac{\mu_i}{\nu_i} + 1 \right) = -\mu_i + \nu_i, \\ \big(\operatorname{grad}_\nu D_{KL}(\cdot \,\|\, \mu)\big)_i &= -\nu_i \log \frac{\mu_i}{\nu_i}. \end{aligned}$$

A comparison with (2.97) verifies (2.102) which uniquely characterize $D_{KL}(\mu \,\|\, \cdot)$ and $D_{KL}(\cdot \,\|\, \mu)$, up to a constant depending on μ. With the additional assumption $D_{KL}(\mu \,\|\, \mu) = 0$ for all μ, this constant is fixed. □

We now ask whether the restriction (2.101) of the Kullback–Leibler divergence to the manifold $\mathcal{P}_+(I)$ is the right divergence function in the sense that (2.102) also holds for the exponential maps of the restricted m- and e-connections. It is easy to verify that this is indeed the case. In order to elaborate on the geometric reason for this, we consider a general Riemannian manifold M and a submanifold N. Given an affine connection $\widetilde{\nabla}$ on M, we can define its restriction ∇ to N. More precisely, denoting the projection of a vector Z in T_pM onto T_pN by $\Pi_p^\top(Z)$, we define $\nabla_X Y|_p := \Pi_p^\top(\widetilde{\nabla}_X Y|_p)$, where X and Y are vector fields on N. Furthermore, we denote the exponential map of ∇ by $\exp_p$ and its inverse by $X(p, q)$.

Now, given $p \in N$, we consider a function $\widetilde{D}_p$ on M that satisfies Eq. (2.95). With the restriction D_p of $\widetilde{D}_p$ to the submanifold N, this directly implies

$$\Pi_q^\top\big(\widetilde{X}(q, p)\big) = -\Pi_q^\top(\operatorname{grad}_q \widetilde{D}_p) = -\operatorname{grad}_q D_p.$$

However, in order to have $X(q, p) = -\operatorname{grad}_q D_p$, which corresponds to Eq. (2.95) on the submanifold N, the following equality is required:

$$X(q, p) = \Pi_q^\top\big(\widetilde{X}(q, p)\big). \tag{2.103}$$

We now verify this condition for the m- and e-connections on $\mathcal{M}_+(I)$ and its submanifold $\mathcal{P}_+(I)$. One can easily show that the vector fields

$$\begin{aligned} \nu \mapsto X^{(m)}(\nu,\mu) &:= \sum_{i\in I} \nu_i \left(\frac{\mu_i}{\nu_i} - 1\right)\delta^i = \widetilde{X}^{(m)}(\nu,\mu), \\ \nu \mapsto X^{(e)}(\nu,\mu) &:= \sum_{i\in I} \nu_i \left(\log\frac{\mu_i}{\nu_i} - \sum_{j\in I} \nu_j \log\frac{\mu_j}{\nu_j}\right)\delta^i \end{aligned} \tag{2.104}$$

satisfy

$$\exp^{(m)}\big(\nu, X^{(m)}(\nu,\mu)\big) = \mu \quad \text{and} \quad \exp^{(e)}\big(\nu, X^{(e)}(\nu,\mu)\big) = \mu, \tag{2.105}$$

respectively, where the exponential maps are given in Proposition 2.5. On the other hand, if we project the vectors $\widetilde{X}^{(m)}(\nu,\mu)$ and $\widetilde{X}^{(e)}(\nu,\mu)$ onto $\mathcal{S}_0(I) \cong T_\nu\mathcal{P}_+(I)$ by using (2.14), we obtain

$$X^{(m)}(\nu,\mu) = \Pi_\nu^\top\big(\widetilde{X}^{(m)}(\nu,\mu)\big) \tag{2.106}$$

and

$$X^{(e)}(\nu,\mu) = \Pi_\nu^\top\big(\widetilde{X}^{(e)}(\nu,\mu)\big). \tag{2.107}$$

This proves that the condition (2.103) is satisfied, which implies

Proposition 2.11 *The following holds*:

$$\begin{aligned} X^{(m)}(\nu,\mu) &= -\operatorname{grad}_\nu D_{KL}(\mu \,\|\, \cdot), \\ X^{(e)}(\nu,\mu) &= -\operatorname{grad}_\nu D_{KL}(\cdot \,\|\, \mu), \end{aligned} \tag{2.108}$$

where D_{KL} is given by (2.101) in Definition 2.8. Furthermore, D_{KL} is the only function on $\mathcal{P}_+(I) \times \mathcal{P}_+(I)$ that satisfies the conditions (2.108) and $D_{KL}(\mu \,\|\, \mu) = 0$ for all μ.

2.7.3 The α-Divergence

We now extend the derivations of Sect. 2.7.2 to the α-connections, leading to a generalization of the relative entropy, the so-called α-divergence (Definition 2.9 below). In order to do, so we define the following vector field as the inverse of the α-exponential map on the manifold $\mathcal{M}_+(I)$ given by (2.58):

$$\nu \mapsto \widetilde{X}^{(\alpha)}(\nu,\mu) := \frac{2}{1-\alpha}\sum_{i\in I} \nu_i \left(\left(\frac{\mu_i}{\nu_i}\right)^{\frac{1-\alpha}{2}} - 1\right)\delta^i. \tag{2.109}$$

Again, we can easily verify that the vector field $\nu \mapsto \widetilde{X}^{(\alpha)}(\nu, \mu)$ is a gradient field by observing that the integrability condition (2.35) is trivially satisfied, that is, $\frac{\partial f^i}{\partial \nu_j} = \frac{\partial f^j}{\partial \nu_i}$ for all i, j, where

$$f^i(\nu) := \left(\frac{\mu_i}{\nu_i}\right)^{\frac{1-\alpha}{2}} - 1.$$

In order to derive the divergence $D^{(\alpha)}$ of the α-connection, we consider a curve $\gamma : [0, 1] \to \mathcal{M}_+(I)$ connecting ν with μ. We obtain

$$\left\langle \widetilde{X}^{(\alpha)}\big(\gamma(t), \mu\big), \dot{\gamma}(t) \right\rangle = \frac{2}{1-\alpha} \sum_{i \in I} \dot{\gamma}_i(t) \left(\left(\frac{\mu_i}{\gamma_i(t)}\right)^{\frac{1-\alpha}{2}} - 1 \right) \tag{2.110}$$

and

$$\begin{aligned}
D^{(\alpha)}(\mu \,\|\, \nu) &= \int_0^1 \left\langle \widetilde{X}^{(\alpha)}\big(\gamma(t), \mu\big), \dot{\gamma}(t) \right\rangle dt \\
&= \sum_{i \in I} \int_0^1 \frac{2}{1-\alpha} \dot{\gamma}_i(t) \left(\left(\frac{\mu_i}{\gamma_i(t)}\right)^{\frac{1-\alpha}{2}} - 1 \right) dt \\
&= \sum_{i \in I} \left[\frac{4}{1-\alpha^2} \gamma_i(t)^{\frac{1+\alpha}{2}} \mu_i^{\frac{1-\alpha}{2}} - \frac{2}{1-\alpha} \gamma_i(t) \right]_0^1 \\
&= \sum_{i \in I} \left(\frac{2}{1+\alpha} \mu_i - \left(\frac{4}{1-\alpha^2} \nu_i^{\frac{1+\alpha}{2}} \mu_i^{\frac{1-\alpha}{2}} - \frac{2}{1-\alpha} \nu_i \right) \right) \\
&= \sum_{i \in I} \left(\frac{2}{1-\alpha} \nu_i + \frac{2}{1+\alpha} \mu_i - \frac{4}{1-\alpha^2} \nu_i^{\frac{1+\alpha}{2}} \mu_i^{\frac{1-\alpha}{2}} \right).
\end{aligned}$$

Obviously, we have

$$D^{(-\alpha)}(\mu \,\|\, \nu) = D^{(\alpha)}(\nu \,\|\, \mu). \tag{2.111}$$

These calculations give rise to the following definition:

Definition 2.9 (α-Divergence) The function $D^{(\alpha)} : \mathcal{M}_+(I) \times \mathcal{M}_+(I) \to \mathbb{R}$ defined by

$$D^{(\alpha)}(\mu \,\|\, \nu) := \frac{2}{1-\alpha} \sum_{i \in I} \nu_i + \frac{2}{1+\alpha} \sum_{i \in I} \mu_i - \frac{4}{1-\alpha^2} \sum_{i \in I} \nu_i^{\frac{1+\alpha}{2}} \mu_i^{\frac{1-\alpha}{2}} \tag{2.112}$$

is called the *α-divergence*. Its restriction to probability measures is given by

$$D^{(\alpha)}(\mu \,\|\, \nu) = \frac{4}{1-\alpha^2} \left(1 - \sum_{i \in I} \nu_i^{\frac{1+\alpha}{2}} \mu_i^{\frac{1-\alpha}{2}} \right). \tag{2.113}$$

Proposition 2.12 *The following holds*:

$$\widetilde{X}^{(\alpha)}(\nu,\mu) = -\operatorname{grad}_{\nu} D^{(\alpha)}(\mu \parallel \cdot), \tag{2.114}$$

where $D^{(\alpha)}$ *is given by* (2.112) *in Definition* 2.9. *Furthermore*, $D^{(\alpha)}$ *is the only function on* $\mathcal{M}_+(I) \times \mathcal{M}_+(I)$ *that satisfies the conditions* (2.114) *and* $D^{(\alpha)}(\mu \parallel \mu) = 0$ *for all* μ.

Proof The statement is obvious from the way we introduced $D^{(\alpha)}$ as a potential function of the gradient field $\nu \mapsto \widetilde{X}^{(\alpha)}(\nu,\mu)$. The following is an alternative direct verification. We compute the partial derivative:

$$\frac{\partial D^{(\alpha)}(\mu \parallel \cdot)}{\partial \nu_i}(\nu) = \frac{2}{1-\alpha}\Big(1 - \nu_i^{\frac{1+\alpha}{2}-1}\mu_i^{\frac{1-\alpha}{2}}\Big).$$

With the formula (2.34), we obtain

$$\begin{aligned}\Big(\operatorname{grad}_{\nu} D^{(\alpha)}(\mu \parallel \cdot)\Big)_i &= \nu_i \cdot \frac{2}{1-\alpha}\Big(1 - \nu_i^{\frac{1+\alpha}{2}-1}\mu_i^{\frac{1-\alpha}{2}}\Big) \\ &= \frac{2}{1-\alpha}\Big(\nu_i - \nu_i^{\frac{1+\alpha}{2}}\mu_i^{\frac{1-\alpha}{2}}\Big).\end{aligned}$$

A comparison with (2.109) verifies (2.114) which uniquely characterizes the function $D^{(\alpha)}(\mu \parallel \cdot)$, up to a constant depending on μ. With the additional assumption $D^{(\alpha)}(\mu \parallel \mu) = 0$ for all μ, this constant is fixed. □

With L'Hopitâl's rule, one can easily verify

$$\lim_{\alpha \to -1} D^{(\alpha)}(\mu \parallel \nu) = D^{(m)}(\mu \parallel \nu) = D_{KL}(\mu \parallel \nu) \tag{2.115}$$

and

$$\lim_{\alpha \to 1} D^{(\alpha)}(\mu \parallel \nu) = D^{(e)}(\mu \parallel \nu) = D_{KL}(\nu \parallel \mu), \tag{2.116}$$

where D_{KL} is relative entropy defined by (2.100).

In what follows, we use the notation $D^{(\alpha)}$ also for $\alpha \in \{-1, 1\}$ by setting $D^{(-1)}(\mu \parallel \nu) := D_{KL}(\mu \parallel \nu)$ and $D^{(1)}(\mu \parallel \nu) := D_{KL}(\nu \parallel \mu)$. This is consistent with the definition of the α-connection, given by (2.54), where we have the m-connection for $\alpha = -1$ and the e-connection for $\alpha = 1$. Note that $D^{(0)}$ is closely related to the Hellinger distance (2.28):

$$D^{(0)}(\mu \parallel \nu) = 2\big(d^H(\mu,\nu)\big)^2. \tag{2.117}$$

We would like to point out that the α-divergence on the simplex $\mathcal{P}_+(I)$, $-1 < \alpha < 1$, based on (2.95) does not coincide with the restriction of the α-divergence on $\mathcal{M}_+(I)$. To be more precise, we have seen that the restriction of the relative entropy, defined on $\mathcal{M}_+(I)$, to the submanifold $\mathcal{P}_+(I)$ is already the right divergence

for the projected m- and e-connections (see Proposition 2.11). The situation turns out to be more complicated for general α. From Eq. (2.65) we obtain

$$X^{(\alpha)}(\nu,\mu) = \dot{\tau}_{\nu,\mu}(0)\, \Pi_\nu^\top \big(\widetilde{X}^{(\alpha)}(\nu,\mu)\big).$$

This equality deviates from the condition (2.103) by the factor $\dot{\tau}_{\nu,\mu}(0)$, which proves that the restriction of the α-divergence, which is defined on $\mathcal{M}_+(I)$, to the submanifold $\mathcal{P}_+(I)$ does not coincide with the α-divergence on $\mathcal{P}_+(I)$. As an example, we consider the case $\alpha = 0$, where the α-connection is the Levi-Civita connection of the Fisher metric. In that case, the canonical divergence equals $\frac{1}{2}(d^F(\mu,\nu))^2$, where d^F denotes the Fisher distance (2.27). Obviously, this divergence is different from the divergence $D^{(0)}$, given by (2.117), which is based on the distance in the ambient space $\mathcal{M}_+(I)$, the Hellinger distance. On the other hand, $\frac{1}{2}(d^F)^2$ can be written as a monotonically increasing function of $D^{(0)}$:

$$\frac{1}{2}\big(d^F(\mu,\nu)\big)^2 = 2\arccos^2\!\left(1 - \frac{1}{4}\, D^{(0)}(\mu \,\|\, \nu)\right). \tag{2.118}$$

2.7.4 The f-Divergence

Our derivation of $D^{(\alpha)}$ was based on the idea of a squared distance function associated with the α-connections in terms of the general Eq. (2.95). However, it turns out that, although being naturally motivated, the functions $D^{(\alpha)}$ do not share all properties of the square of a distance, except for $\alpha = 0$. The symmetry is obviously not satisfied. On the other hand, we have $D^{(\alpha)}(\mu \,\|\, \nu) \geq 0$, and $D^{(\alpha)}(\mu \,\|\, \nu) = 0$ if and only if $\mu = \nu$. One can verify this by considering $D^{(\alpha)}$ as being a function of a more general structure, which we are now going to introduce. Given a strictly convex function $f : \mathbb{R}_+ \to \mathbb{R}$, we define

$$D_f(\mu \,\|\, \nu) := \sum_{i\in I} \mu_i\, f\!\left(\frac{\nu_i}{\mu_i}\right). \tag{2.119}$$

This function is known as the *f-divergence*, and it was introduced and studied by Csiszár [70–73]. Jensen's inequality immediately implies

$$\begin{aligned} D_f(\mu \,\|\, \nu) &= \left(\sum_{j\in I} \mu_j\right) \sum_{i\in I} \frac{\mu_i}{\sum_{j\in I}\mu_j}\, f\!\left(\frac{\nu_i}{\mu_i}\right) \\ &\geq \left(\sum_{j\in I} \mu_j\right) f\!\left(\sum_{i\in I} \frac{\mu_i}{\sum_{j\in I}\mu_j}\, \frac{\nu_i}{\mu_i}\right) \\ &= \left(\sum_{j\in I} \mu_j\right) f\!\left(\frac{\sum_{i\in I}\nu_i}{\sum_{j\in I}\mu_j}\right), \end{aligned}$$

where the equality holds if and only if $\mu = \nu$. If $f(x)$ is non-negative for all x, and $f(1) = 0$, then we obtain

$$D_f(\mu \parallel \nu) \geq 0, \quad \text{and} \quad D_f(\mu \parallel \nu) = 0 \text{ if and only if } \mu = \nu. \tag{2.120}$$

In order to reformulate $D^{(\alpha)}$ as such a function D_f, we define

$$f^{(\alpha)}(x) := \begin{cases} \frac{4}{1-\alpha^2}\left(\frac{1-\alpha}{2} + \frac{1+\alpha}{2}x - x^{\frac{1+\alpha}{2}}\right), & \text{if } \alpha \notin \{-1, 1\}, \\ x - 1 - \log x, & \text{if } \alpha = -1, \\ 1 - x + x\log x, & \text{if } \alpha = 1. \end{cases} \tag{2.121}$$

With this definition, we have $D^{(\alpha)} = D_{f^{(\alpha)}}$. Furthermore, it is easy to verify that for each $\alpha \in [-1, 1]$, the function $f^{(\alpha)}$ is non-negative and vanishes if and only if its argument is equal to one. This proves that the functions $D^{(\alpha)}$ satisfy (2.120), which is a property of a metric. In conclusion, we have seen that, although $D^{(\alpha)}$ is not symmetric and does not satisfy the triangle inequality, it still has some important properties of a squared distance. In this sense, we have obtained a distance-like function that is associated with the α-connection and the Fisher metric on $\mathcal{M}_+(I)$, coupled through Eq. (2.95). The following proposition suggests a way to recover these two objects from the α-divergence.

Proposition 2.13 *The following holds*:

$$\mathfrak{g}_\mu(X, Y) = \frac{\partial^2 D^{(\alpha)}(\mu \parallel \cdot)}{\partial Y \partial X}(\nu)\bigg|_{\nu=\mu}, \tag{2.122}$$

$$\mathbf{T}_\mu(X, Y, Z) = -\frac{2}{3-\alpha}\frac{\partial^3 D^{(\alpha)}(\mu \parallel \cdot)}{\partial Z\, \partial Y\, \partial X}(\nu)\bigg|_{\nu=\mu}. \tag{2.123}$$

The proof of this proposition is by simple calculation. This implies that the Fisher metric can be recovered through the partial derivatives of second order (see (2.122)). In particular, this determines the Levi-Civita connection $\widetilde{\nabla}^{(0)}$ of $\mathfrak{g}$, and we can use $\mathbf{T}$ to derive the α-connection based on the definition (2.54):

$$\mathfrak{g}\big(\widetilde{\nabla}^{(\alpha)}_X Y, Z\big) = \mathfrak{g}\big(\widetilde{\nabla}^{(0)}_X Y, Z\big) - \frac{\alpha}{2}\mathbf{T}(X, Y, Z). \tag{2.124}$$

Thus, we can recover the Fisher metric and the α-connection from the partial derivatives of the α-divergence up to the third order. We obtain the following expansion of the α-divergence:

$$\begin{aligned} D^{(\alpha)}(\mu \parallel \nu) = {} & \frac{1}{2}\sum_{i,j}\mathfrak{g}_\mu(\delta^i, \delta^j)(\mu)\,(\nu_i - \mu_i)(\nu_j - \mu_j) \\ & + \frac{1}{6}\frac{\alpha - 3}{2}\sum_{i,j,k}\mathbf{T}_\mu(\delta^i, \delta^j, \delta^k)(\nu_i - \mu_i)(\nu_j - \mu_j)(\nu_k - \mu_k) \\ & + O\big(\|\nu - \mu\|^4\big). \end{aligned} \tag{2.125}$$

Any function D that satisfies the positivity (2.120) and for which the bilinear form

$$g_\mu(X,Y) := \frac{\partial^2 D(\mu \,\|\, \cdot)}{\partial Y \partial X}(\nu)\bigg|_{\nu=\mu}$$

is positive definite, is called a *divergence* or *contrast function* (see [93, 173]). In Chap. 4, we will revisit divergence functions and related expressions between tensors and affine connections in terms of partial derivatives of potential functions from a more general perspective. We highlight a few important facts already in this section. The coupling between the divergence function $D^{(\alpha)}$ and the tensors $\mathfrak{g}$ and $\mathbf{T}$ through the above expansion (2.125) is clearly not one-to-one, as the derivatives of order greater than three are not fixed. For instance, one could simply neglect the higher-order terms in order to obtain a divergence function that has the same expansion up to order three. A more interesting divergence for $\mathfrak{g}$ and $\mathbf{T}$ is given in terms of the f-divergence. One can easily prove that

$$\begin{aligned} D_f(\mu \,\|\, \nu) = {} & \frac{1}{2} f''(1) \sum_{i,j} \mathfrak{g}_\mu\big(\delta^i, \delta^j\big)(\mu)\,(\nu_i - \mu_i)(\nu_j - \mu_j) \\ & + \frac{1}{6} f'''(1) \sum_{i,j,k} \mathbf{T}_\mu\big(\delta^i, \delta^j, \delta^k\big)(\nu_i - \mu_i)(\nu_j - \mu_j)(\nu_k - \mu_k) \\ & + O\big(\|\nu - \mu\|^4\big). \end{aligned} \tag{2.126}$$

If we choose a function f that satisfies $f''(1) = 1$ and $f'''(1) = \frac{\alpha-3}{2}$, then this expansion coincides with (2.125) up to the third order. Clearly, $f^{(\alpha)}$, as defined in (2.121), satisfies these two conditions. However, this is only one of infinitely many possible choices. This shows that the coupling between a divergence function and an affine connection through (2.95), which uniquely characterizes the relative entropy and the α-divergence on $\mathcal{M}_+(I)$, is stronger than the coupling through the specification of the partial derivatives up to the third order.

2.7.5 The q-Generalization of the Relative Entropy

There is a different way of relating the α-divergence to the relative entropy. Instead of verifying the consistency of D_{KL} and $D^{(\alpha)}$ in terms of (2.115) and (2.116), one can rewrite $D^{(\alpha)}$ so that it resembles the structure of the relative entropy (2.100). This approach is based on Tsallis' so-called q-generalization of the entropy and the relative entropy [248, 249]. Here, q is a parameter with values in the unit interval $]0,1[$ which directly corresponds to $\alpha = 1 - 2q \in]-1,+1[$. With this reparametrization, the α-divergence (2.112) becomes

$$D^{(1-2q)}(\mu \,\|\, \nu) = \frac{1}{q} \sum_i \nu_i + \frac{1}{1-q} \sum_i \mu_i - \frac{1}{q(1-q)} \sum_i \nu_i^{1-q} \mu_i^q. \tag{2.127}$$

This can be rewritten in a way that resembles the structure of the relative entropy D_{KL}. In order to do so, we define the q-exponential function and its inverse, the q-logarithmic function:

$$\exp_q(x) := \big(1 + (1-q)x\big)^{\frac{1}{1-q}}, \qquad \log_q(x) := \frac{1}{1-q}\big(x^{1-q} - 1\big). \tag{2.128}$$

For $q \to 1$, these definitions converge to the ordinary definitions. Now we can rewrite (2.127) as follows:

$$D^{(1-2q)}(\mu \parallel \nu) = \frac{1}{q}\left(\sum_i \nu_i - \sum_i \mu_i - \sum_i \mu_i \log_q\left(\frac{\nu_i}{\mu_i}\right)\right). \tag{2.129}$$

This resembles, up to the factor $\frac{1}{q}$, the Kullback–Leibler divergence (2.100). In this sense, the α-divergence can be considered as a q-generalization of the Kullback–Leibler divergence. These generalizations turn out to be relevant in physics, leading to the field of nonextensive statistical mechanics as a generalization of Boltzmann–Gibbs statistical mechanics [198, 249]. Information geometry contributes to a better geometric understanding of this new field of research [197, 204, 205]. (For a detailed overview of related information-geometric works, see [11].)

2.8 Exponential Families

2.8.1 Exponential Families as Affine Spaces

In Sects. 2.4 and 2.5 we introduced the m- and e-connections and their convex combinations, the α-connections. In general, the notion of an affine connection extends the notion of an affine action of a vector space V on an affine space E. Given a point $p \in E$ and a vector $v \in V$, such an action translates p along v into a new point $p + v$. On the other hand, for each pair of points $p, q \in E$, there is a vector v, called the difference vector between p and q, which translates p into q, that is, $q = p + v$. The affine space E can naturally be interpreted as a manifold with tangent bundle $E \times V$. The translation map $E \times V \to E$ is then nothing but the exponential map exp of the affine connection given by the natural parallel transport $\Pi_{p,q} : (p, v) \mapsto (q, v)$. Clearly, this is a very special parallel transport. It is, however, closely related to the transport maps (2.39) and (2.40), which define the m- and e-connections. Therefore, we can ask the question whether the exponential maps of the m- and e-connections define affine actions on $\mathcal{M}_+(I)$ and $\mathcal{P}_+(I)$. This affine space perspective of $\mathcal{M}_+(I)$ and $\mathcal{P}_+(I)$, and their respective extensions to spaces of σ-finite measures (see Remark 3.8(1)), has been particularly highlighted in [192].

Obviously, $\mathcal{M}_+(I)$, equipped with the m-connection, is not an affine space, simply because the corresponding exponential map is not complete. For each $\mu \in \mathcal{M}_+(I)$ there is a vector $v \in \mathcal{S}(I)$ so that $\mu + v \notin \mathcal{M}_+(I)$. Now, let us come

to the e-connection. The exponential map $\widetilde{\exp}^{(e)}$ is defined on the tangent bundle $T\mathcal{M}_+(I)$ and translates each point μ along a vector $V_\mu \in T_\mu\mathcal{M}_+(I)$. In order to interpret it as an affine action, we identify the tangent spaces $T_\mu\mathcal{M}_+(I)$ and $T_\nu\mathcal{M}_+(I)$ in two points μ and ν in terms of the corresponding parallel transport. More precisely, we introduce an equivalence relation $\sim$ by which we identify two vectors $A_\mu = (\mu, a) \in T_\mu\mathcal{M}_+(I)$ and $B_\nu = (\nu, b) \in T_\nu\mathcal{M}_+(I)$ if B_ν is the parallel transport of A_μ from μ to ν, that is, $B_\nu = \widetilde{\Pi}^{(e)}_{\mu,\nu} A_\mu$. Also, the equivalence class of a vector (μ, a) can be identified with the density $\frac{da}{d\mu}$, which is an element of $\mathcal{F}(I)$. Obviously,

$$A_\mu \sim B_\nu \quad \Leftrightarrow \quad b = \frac{da}{d\mu} \quad \Leftrightarrow \quad \frac{db}{d\nu} = \frac{da}{d\mu}.$$

Now we show that $\mathcal{F}(I)$ acts affinely on $\mathcal{M}_+(I)$. Consider a point μ and a vector f, interpreted as translation vector. This defines a vector $(\mu, f\mu) \in T_\mu\mathcal{M}_+(I)$, which is mapped via the exponential map to

$$\widetilde{\exp}^{(e)}_\mu(f\mu) = e^f\mu.$$

Altogether we have defined the map

$$\mathcal{M}_+(I) \times \mathcal{F}(I) \to \mathcal{M}_+(I), \qquad (\mu, f) \mapsto \mu + f := e^f\mu, \tag{2.130}$$

which satisfies

$$(\mu + f) + g = e^f\mu + g = e^g\left(e^f\mu\right) = e^{f+g}\mu = \mu + (f + g).$$

Furthermore, with the vector $\mathrm{vec}(\mu, \nu) := \log(\frac{d\nu}{d\mu})$ we obviously have $\mu + \mathrm{vec}(\mu, \nu) = \nu$, and this is the only vector that translates μ to ν. This verifies that $+$ is an affine action of $\mathcal{F}(I)$ on $\mathcal{M}_+(I)$.

We apply the same derivation in order to define an affine structure on $\mathcal{P}_+(I)$. This leads to the following version of the map (2.130) defined for the simplex $\mathcal{P}_+(I)$ (see Fig. 2.6):

$$\mathcal{P}_+(I) \times \left(\mathcal{F}(I)/\mathbb{R}\right) \to \mathcal{P}_+(I), \qquad (\mu, f + \mathbb{R}) \mapsto \mu + (f + \mathbb{R}) := \frac{e^f}{\mu(e^f)}\mu. \tag{2.131}$$

This is an affine action of the vector space $\mathcal{F}(I)/\mathbb{R}$ on $\mathcal{P}_+(I)$ with difference vector

$$\mathrm{vec} : \mathcal{P}_+(I) \times \mathcal{P}_+(I) \to \mathcal{F}(I)/\mathbb{R}, \qquad (\mu, \nu) \mapsto \mathrm{vec}(\mu, \nu) = \log\left(\frac{d\nu}{d\mu}\right) + \mathbb{R}.$$

Therefore, $\mathcal{P}_+(I)$ is an affine space over $\mathcal{F}(I)/\mathbb{R}$.

The corresponding affine subspaces play a central role within information geometry.

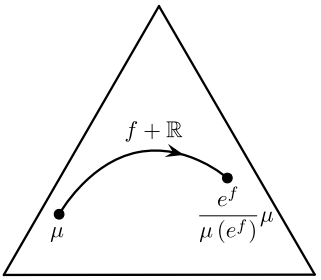

Fig. 2.6 The affine action of $\mathcal{F}(I)/\mathbb{R}$ on the simplex $\mathcal{P}_+(I)$

Definition 2.10 (Exponential family) An affine subspace $\mathcal{E}$ of $\mathcal{P}_+(I)$ with respect to $+$ is called an *exponential family*. Given a measure $\mu_0 \in \mathcal{M}_+(I)$ and a linear subspace $\mathcal{L}$ of $\mathcal{F}(I)$, the following submanifold of $\mathcal{P}_+(I)$ is an exponential family:

$$\mathcal{E}(\mu_0, \mathcal{L}) := \left\{ \frac{e^f}{\mu_0(e^f)} \mu_0 : f \in \mathcal{L} \right\}. \tag{2.132}$$

To simplify the notation, in the case where μ_0 is the counting measure, that is, $\mu_0 = \sum_{i \in I} \delta^i$, we simply write $\mathcal{E}(\mathcal{L})$ instead of $\mathcal{E}(\mu_0, \mathcal{L})$.

Clearly, all exponential families have the structure (2.132). We always assume $\mathbb{1} \in \mathcal{L}$ and thereby ensure uniqueness of $\mathcal{L}$. Furthermore, with this assumption we have $\dim(\mathcal{E}) = \dim(\mathcal{L}) - 1$.

Given two points $\mu, \nu \in \mathcal{P}_+(I)$, the m- and e-connections provide two kinds of straight lines connecting them:

$$\gamma^{(m)}_{\mu,\nu} : [0,1] \to \mathcal{P}_+(I), \qquad t \mapsto (1-t)\,\mu + t\,\nu,$$

$$\gamma^{(e)}_{\mu,\nu} : [0,1] \to \mathcal{P}_+(I), \qquad t \mapsto \frac{(\frac{d\nu}{d\mu})^t}{\mu((\frac{d\nu}{d\mu})^t)}\,\mu.$$

This allows us to consider two kinds of geodesically convex sets. A set $\mathcal{S}$ is said to be *m-convex* if

$$\mu, \nu \in \mathcal{S} \quad \Rightarrow \quad \gamma^{(m)}_{\mu,\nu}(t) \in \mathcal{S} \quad \text{for all } t \in [0,1],$$

and *e-convex* if

$$\mu, \nu \in \mathcal{S} \quad \Rightarrow \quad \gamma^{(e)}_{\mu,\nu}(t) \in \mathcal{S} \quad \text{for all } t \in [0,1].$$

Exponential families are clearly e-convex. On the other hand, they can also be m-convex. Given a partition $\mathfrak{S}$ of I and probability measures μ_A with support A, $A \in \mathfrak{S}$, the following set is an m-convex exponential family:

$$\mathcal{M} := \mathcal{M}(\mu_A : A \in \mathfrak{S}) := \left\{ \sum_{A \in \mathfrak{S}} \eta_A\, \mu_A : \eta_A > 0, \sum_{A \in \mathfrak{S}} \eta_A = 1 \right\}. \tag{2.133}$$

To see this, define the base measure as $\mu_0 := \sum_{A\in\mathfrak{S}} \mu_A$ and $\mathcal{L}$ as the linear hull of the vectors $\mathbb{1}_A$, $A \in \mathfrak{S}$. Then the elements of $\mathcal{M}(\mu_A : A \in \mathfrak{S})$ are precisely the elements of the exponential family $\mathcal{E}(\mu_0, \mathcal{L})$, via the correspondence

$$\begin{aligned}
\sum_{A\in\mathfrak{S}} \eta_A \mu_A &= \sum_{A\in\mathfrak{S}} \frac{\eta_A}{\sum_{B\in\mathfrak{S}} \eta_B} \mu_A \\
&= \sum_{A\in\mathfrak{S}} \frac{e^{\log \eta_A}}{\sum_{B\in\mathfrak{S}} e^{\log \eta_B}} \mu_A \\
&= \sum_{A\in\mathfrak{S}} \frac{e^{\lambda_A}}{\sum_{B\in\mathfrak{S}} e^{\lambda_B}} \mu_A \\
&= \sum_{A\in\mathfrak{S}} \sum_{i\in A} \frac{e^{\lambda_A}}{\sum_{B\in\mathfrak{S}} e^{\lambda_B}} \mu_i \, \delta^i \\
&= \sum_{i\in I} \frac{e^{\sum_{A\in\mathfrak{S}} \lambda_A \mathbb{1}_A(i)}}{\sum_{j\in I} e^{\sum_{A\in\mathfrak{S}} \lambda_A \mathbb{1}_A(j)} \mu_j} \mu_i \, \delta^i \\
&= \frac{e^{\sum_{A\in\mathfrak{S}} \lambda_A \mathbb{1}_A}}{\mu_0(e^{\sum_{A\in\mathfrak{S}} \lambda_A \mathbb{1}_A})} \mu_0, \qquad (2.134)
\end{aligned}$$

where $\lambda_A = \log(\eta_A)$, $A \in \mathfrak{S}$. It turns out that $\mathcal{M}(\mu_A : A \in \mathfrak{S})$ is not just one instance of an m-convex exponential family. In fact, as we shall see in the following theorem, which together with its proof is based on [179], (2.133) describes the general structure of such exponential families. Note that for any set $\mathfrak{S}$ of subsets A of I and corresponding distributions μ_A with support A, the set (2.133) will be m-convex. However, when the sets $A \in \mathfrak{S}$ form a partition of I, this set will also be an exponential family.

Theorem 2.4 *Let $\mathcal{E} = \mathcal{E}(\mu_0, \mathcal{L})$ be an exponential family in $\mathcal{P}_+(I)$. Then the following statements are equivalent*:

(1) *The exponential family $\mathcal{E}$ is m-convex.*
(2) *There exists a partition $\mathfrak{S} \subseteq 2^I$ of I and elements $\mu_A \in \mathcal{P}_+(A)$, $A \in \mathfrak{S}$, such that*

$$\mathcal{E} = \mathcal{M}(\mu_A : A \in \mathfrak{S}), \qquad (2.135)$$

where the RHS of this equation is defined by (2.133).
(3) *The linear space $\mathcal{L}$ is a subalgebra of $\mathcal{F}(I)$, i.e., closed under (pointwise) multiplication.*

The proof of this theorem is based on the following lemma.

Lemma 2.6 *The smallest convex exponential family containing two probability measures $\mu = \sum_{i\in I} \mu_i \, \delta^i$ and $\nu = \sum_{i\in I} \nu_i \, \delta^i$ with the supports equal to I coin-*

cides with $\mathcal{M}(\mu_A : A \in \mathfrak{S}_{\mu,\nu})$ *where* $\mathfrak{S}_{\mu,\nu}$ *is the partition of* I *having* $i, j \in I$ *in the same block if and only if* $\mu_i \nu_j = \mu_j \nu_i$ *and* μ_A *equals the conditioning of* μ *to* A*, that is,*

$$\mu_A = \sum_{i \in I} \mu_{A,i}\, \delta^i, \quad \text{with} \quad \mu_{A,i} := \begin{cases} \frac{\mu_i}{\sum_{j \in A} \mu_j}, & \text{if } i \in A, \\ 0 & \text{otherwise.} \end{cases}$$

Proof Let $\mathfrak{S}_{\mu,\nu}$ have n blocks and an element i_A of A be fixed for each $A \in \mathfrak{S}_{\mu,\nu}$. The numbers ${\mu_{i_A}}^k {\nu_{i_A}}^{-k}$, $A \in \mathfrak{S}_{\mu,\nu}$, $0 \le k < n$, are elements of a Vandermonde matrix which has nonzero determinant because μ_{i_A}/ν_{i_A}, $A \in \mathfrak{S}_{\mu,\nu}$, are pairwise different. Therefore, for $0 \le k < n$ the vectors $({\mu_{i_A}}^k {\nu_{i_A}}^{-k})_{A \in \mathfrak{S}_{\mu,\nu}}$ are linearly independent, and so are the vectors $({\mu_i}^k {\nu_i}^{-k})_{i \in I}$. Then the probability measures proportional to $({\mu_i}^{k+1} {\nu_i}^{-k})_{i \in I}$ are independent. These probability measures belong to any exponential family containing μ and ν and, in turn, their convex hull is contained in any convex exponential family containing μ and ν. In particular, it is contained in $\mathcal{M} = \mathcal{M}(\mu_A : A \in \mathfrak{S}_{\mu,\nu})$ because μ and ν, the latter being equal to $\sum_{A \in \mathfrak{S}} (\sum_{j \in A} \nu_j)\, \mu_A$, belong to $\mathcal{M}$ by construction. Since the convex hull has the same dimension as $\mathcal{M}$, any m-convex exponential family containing μ and ν includes $\mathcal{M}$. □

Proof of Theorem 2.4 (1) ⇒ (2) Let $\mathfrak{S}$ be a partition of I with the maximal number of blocks such that $\mathcal{E} = \mathcal{E}(\mu_0, \mathcal{L})$ contains $\mathcal{M}(\mu_A : A \in \mathfrak{S})$ for some probability measures μ_A. For any probability measure μ with the support equal to I and $i \in A$, $j \in B$, belonging to different blocks A, B of $\mathfrak{S}$, denote by $H_{\mu,i,j}$ the hyperplane of vectors $(t_C)_{C \in \mathfrak{S}}$ satisfying

$$t_A \cdot \mu_i\, \mu_{A,j} - t_B \cdot \mu_j\, \mu_{B,i} = 0.$$

Since no such $H_{\mu,i,j}$ contains the hyperplane given by $\sum_{A \in \mathfrak{S}} t_A = 1$, a probability measure $\nu = \sum_{A \in \mathfrak{S}} t_A\, \mu_A$ in $\mathcal{M}$ exists such that all equations $\mu_i \nu_j = \mu_j \nu_i$ with i, j in different blocks of $\mathfrak{S}$ are simultaneously violated. This implies that each block of $\mathfrak{S}$ is a union of blocks of $\mathfrak{S}_{\mu,\nu}$. If, additionally, $\mu \in \mathcal{E}$ then $\mathcal{M}(\mu_A : A \in \mathfrak{S}_{\mu,\nu})$ is contained in $\mathcal{E}$ on account of Lemma 2.6. By maximality of the number of blocks, $\mathfrak{S}_{\mu,\nu} = \mathfrak{S}$. Hence, $\mu = \sum_{A \in \mathfrak{S}} (\sum_{j \in A} \mu_j) \mu_A$ belongs to $\mathcal{M}$, and thus $\mathcal{E} = \mathcal{M}$.

(2) ⇒ (3) Given the equality (2.135), we can represent $\mathcal{E}$ in terms of (2.134). This implies that $\mathcal{L}$ is spanned by the vectors $\mathbb{1}_A$, $A \in \mathfrak{S}$. The linear space $\mathcal{L}$ obviously forms a subalgebra of $\mathcal{F}(I)$. This is because the multiplication of two indicator functions $\mathbb{1}_A$ and $\mathbb{1}_B$, where $A, B \in \mathfrak{S}$, equals $\mathbb{1}_A$ if $A = B$, and equals the zero function otherwise.

(3) ⇒ (1) Assume $\mu, \nu \in \mathcal{E}(\mu_0, \mathcal{L})$. This means that there exist functions $f, g \in \mathcal{L}$ with $\mu = \frac{e^f}{\mu_0(e^f)}\, \mu_0$ and $\nu = \frac{e^g}{\mu_0(e^g)}\, \mu_0$. Now consider a convex combination $(1-t)\, \mu + t\, \nu$, $0 \le t \le 1$. With f and g, the function $h := \log((1-t)\, \frac{e^f}{\mu_0(e^f)} +$

$t\,\frac{e^g}{\mu_0(e^g)}$) is also an element of the algebra $\mathcal{L}$. Therefore

$$(1-t)\,\mu + t\,\nu = (1-t)\,\frac{e^f}{\mu_0(e^f)}\,\mu_0 + t\,\frac{e^g}{\mu_0(e^g)}\,\mu_0 = \frac{e^h}{\mu_0(e^h)}\,\mu_0 \in \mathcal{E}(\mu_0, \mathcal{L}). \quad \square$$

2.8.2 Implicit Description of Exponential Families

We have introduced exponential families as affine subspaces with respect to the translation (2.131). In many applications, however, it is important to consider not only strictly positive probability measures but also limit points of a given exponential family. In order to incorporate such distributions, we devote this section to the study of closures of exponential families, which turns out to be particularly convenient in terms of implicit equations. Classical work on various extensions of exponential families is due to Chentsov [65], Barndorff-Nielsen [39], Lauritzen [159], and Brown [55]. The theory has been considerably further developed more recently by Csiszár and F. Matúš [76, 77], going far beyond our context of finite state spaces.

Implicit equations play an important role within graphical model theory, where they are related to conditional independence statements and the Hammersley–Clifford Theorem 2.9 (see [161]). We will address graphical models and their generalizations, hierarchical models, in Sect. 2.9. The following material is based on [222] and touches upon the seminal work of Geiger, Meek, and Sturmfels [103].

Let us start with the exponential family itself, without the boundary points. To this end, consider a reference measure $\mu_0 = \sum_{i\in I} \mu_{0,i}\,\delta^i$ and a subspace $\mathcal{L}$ of $\mathcal{F}(I)$ with $\mathbb{1} \in \mathcal{L}$. Throughout this section, we fix a basis $f_0 := \mathbb{1}, f_1, \ldots, f_d$, of $\mathcal{L}$, where d is the dimension of $\mathcal{E}(\mu_0, \mathcal{L})$. Obviously, a probability measure μ is an element of $\mathcal{E}(\mu_0, \mathcal{L})$ if and only if

$$\mathrm{vec}(\mu_0, \mu) \in \mathcal{L}/\mathbb{R},$$

which is equivalent to

$$\log\left(\frac{d\mu}{d\mu_0}\right) \in \mathcal{L}, \tag{2.136}$$

or

$$\left\langle \log\left(\frac{d\mu}{d\mu_0}\right), n \right\rangle = 0 \quad \text{for all } n \in \mathcal{L}^\perp, \tag{2.137}$$

where $\mathcal{L}^\perp$ is the orthogonal complement of $\mathcal{L}$ with respect to the canonical scalar product $\langle \cdot, \cdot \rangle$ on $\mathcal{F}(I)$.

Exponentiating both sides of (2.137) yields

$$\prod_i \left(\frac{\mu_i}{\mu_{0,i}}\right)^{n(i)} = 1 \quad \text{for all } n \in \mathcal{L}^\perp.$$

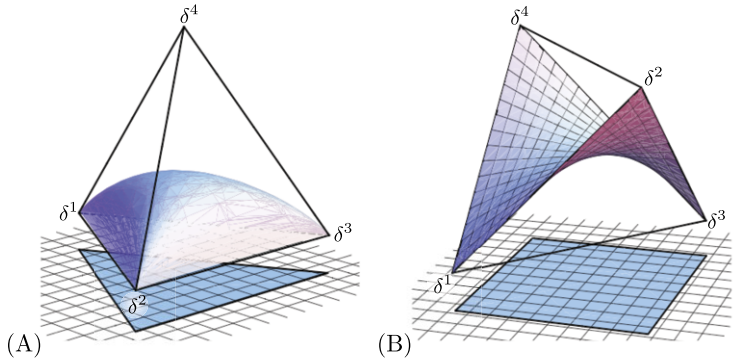

Fig. 2.7 Two examples of exponential families. Reproduced from [S. Weis, A. Knauf (2012) *Entropy distance: New quantum phenomena*, Journal of Mathematical Physics 53(10) 102206], with the permission of AIP Publishing

This is equivalent to

$$\prod_i {\mu_i}^{n(i)} = \prod_i {\mu_{0,i}}^{n(i)} \quad \text{for all } n \in \mathcal{L}^\perp.$$

We define $n^+ := \max\{n(i), 0\}$ and $n^- := \max\{-n(i), 0\}$ and reformulate this condition by

$$\prod_i {\mu_i}^{n^+(i)} \prod_i {\mu_{0,i}}^{n^-(i)} = \prod_i {\mu_i}^{n^-(i)} \prod_i {\mu_{0,i}}^{n^+(i)} \quad \text{for all } n \in \mathcal{L}^\perp. \tag{2.138}$$

With the abbreviation $\mu^n := \prod_i {\mu_i}^{n(i)}$, (2.138) can be written as

$$\mu^{n^+} {\mu_0}^{n^-} = \mu^{n^-} {\mu_0}^{n^+} \quad \text{for all } n \in \mathcal{L}^\perp. \tag{2.139}$$

This proves that $\mu \in \mathcal{E}(\mu_0, \mathcal{L})$ if and only if (2.139) is satisfied. Theorem 2.5 below states that the same criterion holds also for all elements μ in the closure of $\mathcal{E}(\mu_0, \mathcal{L})$. Before we come to this result, we first have to introduce the notion of a facial set.

Non-empty facial sets are the possible support sets that distributions of a given exponential family can have. There is an instructive way to characterize them. Given the basis $f_0 := \mathbb{1}, f_1, \ldots, f_d$ of $\mathcal{L}$, we consider the affine map

$$\mathbb{E} : \mathcal{P}(I) \to \mathbb{R}^{d+1}, \qquad \mu \mapsto \big(1, \mu(f_1), \ldots, \mu(f_d)\big). \tag{2.140}$$

(Here, $\mu(f_k) = \mathbb{E}_\mu(f_k)$ denotes the expectation value of f_k with respect to μ.) Obviously, the image of this map is a polytope, the *convex support* $\mathrm{cs}(\overline{\mathcal{E}})$ of $\mathcal{E}$, which is the convex hull of the images of the Dirac measures δ^i in $\mathcal{P}(I)$, that is, $\mathbb{E}(\delta^i) = (\delta^i(f_0), \delta^i(f_1), \ldots, \delta^i(f_d)) = (1, f_1(i), \ldots, f_d(i))$, $i \in I$. The situation is illustrated in Fig. 2.7 for two examples of two-dimensional exponential families. In each case, the image of the simplex under the map $\mathbb{E}$, the convex support of $\mathcal{E}$, is shown as a "shadow" in the horizontal plane, a triangle in one case and a square in the other case. We observe that the convex supports can be interpreted as "flattened"

versions of the individual closures $\overline{\mathcal{E}}$, where each face of $\mathrm{cs}(\overline{\mathcal{E}})$ corresponds to the intersection of $\overline{\mathcal{E}}$ with a face of the simplex $\mathcal{P}(I)$. Therefore, the faces of the convex support determine the possible support sets, which we call *facial* sets. In order to motivate their definition below, note that a set C is a face of a polytope P in $\mathbb{R}^n$ if either $C = P$ or C is the intersection of P with an affine hyperplane H such that all $x \in P$, $x \notin C$, lie on one side of the hyperplane. Non-trivial faces of maximal dimension are called *facets*. It is a fundamental result that every polytope can equivalently be described as the convex hull of a finite set or as a finite intersection of closed linear half-spaces (corresponding to its facets) (see [261]).

In particular we are interested in the face structure of $\mathrm{cs}(\overline{\mathcal{E}})$. Since we assumed that $\mathbb{1} \in \mathcal{L}$, the image of $\mathbb{E}$ is contained in the affine hyperplane $x_1 = 1$, and we can replace every affine hyperplane H by an equivalent central hyperplane (which passes through the origin). For the convex support $\mathrm{cs}(\overline{\mathcal{E}})$, we want to know which points from $\mathbb{E}(\delta^i)$, $i \in I$, lie on each face. This motivates the following definition.

Definition 2.11 A set $F \subseteq I$ is called *facial* if there exists a vector $\vartheta \in \mathbb{R}^{d+1}$ such that

$$\sum_{k=0}^{d} \vartheta_k f_k(i) = 0 \quad \text{for all } i \in F, \qquad \sum_{k=0}^{d} \vartheta_k f_k(i) \geq 1 \quad \text{for all } i \in I \setminus F. \tag{2.141}$$

Lemma 2.7 *Fix a subset $F \subseteq I$. Then we have*:

(1) *F is facial if and only if for any $u \in \mathcal{L}^{\perp}$*:

$$\mathrm{supp}\big(u^+\big) \subseteq F \quad \Leftrightarrow \quad \mathrm{supp}\big(u^-\big) \subseteq F \tag{2.142}$$

(*here, we consider u as an element of $\mathcal{F}(I)$, and for any $f \in \mathcal{F}(I)$, $\mathrm{supp}(f) := \{i \in I : f(i) \neq 0\}$*).

(2) *If μ is a solution to* (2.139), *then* $\mathrm{supp}(\mu)$ *is facial*.

Proof One direction of the first statement is straightforward: Let $u \in \mathcal{L}^{\perp}$ and suppose that $\mathrm{supp}(u^+) \subseteq F$. Then

$$\sum_{i \in F} u(i) f_k(i) = -\sum_{i \notin F} u(i) f_k(i), \quad k = 0, 1, \ldots, d,$$

and therefore

$$0 = \sum_{i \in F} u(i) \sum_{k=0}^{d} \vartheta_k f_k(i) = -\sum_{i \notin F} u(i) \sum_{k=0}^{d} \vartheta_k f_k(i).$$

Since $\sum_{k=0}^{d} \vartheta_k f_k(i) > 1$ and $u(i) \leq 0$ for $i \notin F$ it follows that $u(i) = 0$ for $i \notin F$, proving one direction of the first statement.

The opposite direction is a bit more complicated. Here, we present a proof using elementary arguments from polytope theory (see, e.g., [261]). For an alternative proof using Farkas' Lemma see [103]. Assume that F is not facial. Let

F' be the smallest facial set containing F. Let P_F and $P_{F'}$ be the convex hulls of $\{(f_0(i), \dots, f_d(i)) : i \in F\}$ and $\{(f_0(i), \dots, f_d(i)) : i \in F'\}$. Then P_F contains a point g from the relative interior of $P_{F'}$. Therefore g can be represented as $g_k = \sum_{i\in F} \alpha(i) f_k(i) = \sum_{i\in F'} \beta(i) f_k(i)$, where $\alpha(i) \geq 0$ for all $i \in F$ and $\beta(i) > 0$ for all $i \in F'$. Hence $u(i) := \alpha(i) - \beta(i)$ (where $\alpha(i) := 0$ for $i \notin F$ and $\beta(i) := 0$ for $x \notin F'$) defines a vector $u \in \mathcal{L}^\perp$ such that $\operatorname{supp}(u^+) \subseteq F$ and $\operatorname{supp}(u^-) \cap (I \setminus F) = F' \setminus F \neq \emptyset$.

The second statement now follows immediately: If μ satisfies (2.139) for some $u \in \mathcal{L}^\perp$, then the LHS of (2.139) vanishes if and only if the RHS vanishes, and by the first statement this implies that $\operatorname{supp}(\mu)$ is facial. □

Theorem 2.5 *A distribution μ is an element of the closure of $\mathcal{E}(\mu_0, \mathcal{L})$ if and only if it satisfies* (2.139).

Proof The first thing to note is that it is enough to prove the theorem when $\mu_{0,i} = 1$ for all $i \in I$. To see this observe that $\mu \in \overline{\mathcal{E}}(\mathcal{L})$ if and only if $\lambda \sum_i \mu_{0,i}\mu_i \, \delta^i \in \overline{\mathcal{E}}(\mu_0, \mathcal{L})$, where $\lambda > 0$ is a normalizing constant, which does not appear in (2.139) since they are homogeneous.

Let $Z_\mathcal{L}$ be the set of solutions of (2.139). The derivation of Eqs. (2.139) was based on the requirement that $\mathcal{E}(\mathcal{L}) \subseteq Z_\mathcal{L}$, which also implies $\overline{\mathcal{E}}(\mathcal{L}) \subseteq \overline{Z}_\mathcal{L} = Z_\mathcal{L}$. It remains to prove the reversed inclusion $\overline{\mathcal{E}}(\mathcal{L}) \supseteq Z_\mathcal{L}$. Let $\mu \in Z_\mathcal{L} \setminus \mathcal{E}(\mathcal{L})$ and put $F := \operatorname{supp}(\mu)$. We construct a sequence $\mu^{(n)}$ in $\mathcal{E}(\mathcal{L})$ that converges to μ as $n \to \infty$. We claim that the system of equations

$$\sum_{k=0}^{d} b_k f_k(i) = \log \mu_i \quad \text{for all } i \in F \tag{2.143}$$

in the variables b_k, $k = 0, 1, \dots, d$, has a solution. Otherwise we can find a function $v(i)$, $i \in I$, such that $\sum_{i\in I} v(i) \log \mu_i \neq 0$ and $\sum_{i\in I} v(i) f_k(i) = 0$ for all k. This leads to the contradiction $\mu^{v^+} \neq \mu^{v^-}$. Fix a vector $\vartheta \in \mathbb{R}^{d+1}$ with property (2.141). For any $n \in \mathbb{N}$ define

$$\mu^{(n)} := \frac{1}{Z} \sum_i e^{-n\sum_k \vartheta_k f_k(i)} e^{\sum_k b_k f_k(i)} \, \delta^i \in \mathcal{E}(\mathcal{L}), \tag{2.144}$$

where Z is a normalization factor. By (2.141) and (2.143) it follows that $\lim_{n\to\infty} \mu^{(n)} = \mu$. This proves the theorem. □

The last statement of Lemma 2.7 can be generalized by the following explicit description of the closure of an exponential family.

Theorem 2.6 (Closure of an exponential family) *Let $\mathcal{L}$ be a linear subspace of $\mathcal{F}(I)$, and let $S(\mathcal{L})$ denote the set of non-empty facial subsets of I (see Definition* 2.11, *and Lemma* 2.7). *Define for each set $F \in S(\mathcal{L})$ the* truncated exponential

family *as*

$$\mathcal{E}_F := \mathcal{E}_F(\mu_0, \mathcal{L}) := \left\{ \frac{1}{\sum_{j\in F} \mu_j} \sum_{i\in F} \mu_i\, \delta^i : \mu = \sum_{i\in I} \mu_i\, \delta^i \in \mathcal{E}(\mu_0, \mathcal{L}) \right\}. \quad (2.145)$$

Then the closure of the exponential family $\mathcal{E}$ is given by

$$\overline{\mathcal{E}}(\mu_0, \mathcal{L}) = \bigcup_{F\in S(\mathcal{L})} \mathcal{E}_F. \quad (2.146)$$

Proof "$\subseteq$" Let μ be in the closure of $\mathcal{E}(\mu_0, \mathcal{L})$. Clearly, μ satisfies Eqs. (2.139) and therefore, by Lemma 2.7, its support set F is facial. Furthermore, the same reasoning that underlies Eq. (2.143) yields a solution of the equations

$$\sum_{k=0}^{d} \vartheta_k\, f_k(i) = \log \frac{\mu_i}{\mu_{0,i}}, \quad i \in F.$$

Using these ϑ values for $k = 1, \ldots, d$, we extend μ by

$$\widetilde{\mu}_i := \frac{1}{Z(\vartheta)} \exp\left(\sum_{k=1}^{d} \vartheta_k\, f_k(i) \right), \quad i \in I,$$

to a distribution $\widetilde{\mu}$ with full support. Obviously, $\widetilde{\mu}$ defines μ through truncation.

"$\supseteq$" Let $\mu \in \mathcal{E}_F$ for some non-empty facial set F. Then μ has a representation

$$\mu_i := \begin{cases} \mu_{2,i} \exp(\sum_{k=0}^{d} \vartheta_k\, f_k(i)), & \text{if } i \in F, \\ 0, & \text{otherwise.} \end{cases}$$

With a vector $\vartheta' = (\vartheta'_0, \vartheta'_1, \ldots, \vartheta_d) \in \mathbb{R}^{d+1}$ that satisfies (2.141), the sequence

$$\mu_i^{(n)} := \exp\left(\sum_{k=0}^{d} \left(\vartheta_k - n\,\vartheta'_k\right) f_k(i) \right) \in \mathcal{E}(\mu_0, \mathcal{L})$$

converges to μ, proving $\mu \in \overline{\mathcal{E}}(\mu_0, \mathcal{L})$. □

Example 2.4 (Support sets of an exponential family) In this example, we apply Theorem 2.6 to the exponential families shown in Fig. 2.7. These are families of distributions on $I = \{1, 2, 3, 4\}$, and we write the elements of $\mathcal{F}(I)$ as vectors (x_1, x_2, x_3, x_4). In order to determine the individual facial subsets of I, we use the criterion given in the first part of Lemma 2.7.

(1) Let us start with the exponential family shown in Fig. 2.7(A). The space $\mathcal{L}$ of this exponential family is the linear hull of the orthogonal vectors

$$(1, 1, 1, 1), \quad (1, 1, -2, 0), \quad (1, -1, 0, 0).$$

Its one-dimensional orthogonal complement $\mathcal{L}^\perp$ is spanned by the vector $(1, 1, 1, -3)$. This implies the following pairs $(\mathrm{supp}(u^+), \mathrm{supp}(u^-))$, $u = u^+ - u^- \in \mathcal{L}^\perp$, of disjoint support sets:

$$(\emptyset, \emptyset), \quad \big(\{1, 2, 3\}, \{4\}\big), \quad \big(\{4\}, \{1, 2, 3\}\big). \tag{2.147}$$

The criterion (2.142) for a subset F of I to be a facial set simply means that for any of the support set pairs (M, N) in (2.147), either M and N are both contained in F or neither of them is. This yields the following set of facial sets:

$$\emptyset,\ \{1\},\ \{2\},\ \{3\},\ \{1, 2\},\ \{1, 3\},\ \{2, 3\},\ \{1, 2, 3, 4\}.$$

Obviously, these sets, except $\emptyset$, are exactly the support sets of distributions that are in the closure of the exponential family (see Fig. 2.7(A)).

(2) Now let us come to the exponential family shown in Fig. 2.7(B). Its linear space $\mathcal{L}$ is spanned by

$$(1, 1, 1, 1), \quad (1, 1, -1, -1), \quad (1, -1, -1, 1),$$

with the orthogonal complement $\mathcal{L}^\perp$ spanned by $(1, -1, 1, -1)$. As possible support set pairs, we obtain

$$(\emptyset, \emptyset), \quad \big(\{1, 3\}, \{2, 4\}\big), \quad \big(\{2, 4\}, \{1, 3\}\big).$$

Applying criterion (2.142) finally yields the facial sets

$$\emptyset,\ \{1\},\ \{2\},\ \{3\},\ \{4\},\ \{1, 2\},\ \{1, 4\},\ \{2, 3\},\ \{3, 4\},\ \{1, 2, 3, 4\}.$$

Also in this example, these sets, except $\emptyset$, are the possible support sets of distributions that are in the closure of the exponential family (see Fig. 2.7(B)).

Theorem 2.5 provides an implicit description of the closure of an exponential family $\mathcal{E}(\mu_0, \mathcal{L})$. Here, however, we have to test the equations with infinitely many elements n of the orthogonal complement $\mathcal{L}^\perp$ of $\mathcal{L}$. Now we ask the question whether the test can be reduced to a finite number of vectors $n \in \mathcal{L}^\perp$. For an element $\mu \in \mathcal{E}(\mu_0, \mathcal{L})$, clearly the orthogonality (2.137) has to be tested only for a basis $n_1, \ldots, n_c$, $c = |I| - \dim(\mathcal{L})$, of $\mathcal{L}$, which is equivalent to

$$\mu^{n_k^+} \mu_0^{n_k^-} = \mu^{n_k^-} \mu_0^{n_k^+} \quad \text{for all } k = 1, \ldots, c. \tag{2.148}$$

This criterion is sufficient for the elements of $\mathcal{E}(\mu_0, \mathcal{L})$. It turns out, however, that it is not sufficient for describing elements in the boundary of $\mathcal{E}(\mu_0, \mathcal{L})$. But it is still possible to reduce the number of equations to a finite number. In order to do so, we have to replace the basis $n_1, \ldots, n_c$ by a so-called circuit basis, which is still a generating system but contains in general more than c elements.

Definition 2.12 A *circuit vector* of the space $\mathcal{L}$ is a nonzero vector $n \in \mathcal{L}^\perp$ with inclusion minimal support, i.e., if $n' \in \mathcal{L}^\perp$ satisfies $\operatorname{supp}(n') \subseteq \operatorname{supp}(n)$, then $n' = \lambda n$ for some $\lambda \in \mathbb{R}$. A *circuit* is the support set of a circuit vector. A *circuit basis* is a subset of $\mathcal{L}^\perp$ containing precisely one circuit vector for every circuit.

This definition allows us to prove the following theorem.

Theorem 2.7 *Let $\mathcal{E}(\mu_0, \mathcal{L})$ be an exponential family. Then its closure $\overline{\mathcal{E}}(\mu_0, \mathcal{L})$ equals the set of all probability distributions that satisfy*

$$\mu^{c^+}\mu_0{}^{c^-} = \mu^{c^-}\mu_0{}^{c^+} \quad \text{for all } c \in C, \tag{2.149}$$

where C is a circuit basis of $\mathcal{L}$.

The proof is based on the following two lemmas.

Lemma 2.8 *For every vector $n \in \mathcal{L}^\perp$ there exists a sign-consistent circuit vector $c \in \mathcal{L}^\perp$, i.e., if $c(i) \neq 0 \neq n(i)$ then $\operatorname{sign} c(i) = \operatorname{sign} n(i)$, for all $i \in I$.*

Proof Let c be a vector with inclusion-minimal support that is sign-consistent with n and satisfies $\operatorname{supp}(c) \subseteq \operatorname{supp}(n)$. If c is not a circuit vector, then there exists a circuit vector c' with $\operatorname{supp}(c') \subseteq \operatorname{supp}(c)$. A suitable linear combination $c + \alpha c'$, $\alpha \in \mathbb{R}$, gives a contradiction to the minimality of c. □

Lemma 2.9 *Every vector $n \in \mathcal{L}^\perp$ is a finite sign-consistent sum of circuit vectors $n = \sum_{k=1}^r c_k$, i.e., if $c_k(i) \neq 0$ then $\operatorname{sign} c_k(i) = \operatorname{sign} n(i)$, for all $i \in I$.*

Proof Use induction on the size of $\operatorname{supp}(n)$. In the induction step, use a sign-consistent circuit vector, as in the last lemma, to reduce the support. □

Proof of Theorem 2.7 Again, we can assume $\mu_{0,i} = 1$ for all $i \in I$. By Theorem 2.5 it suffices to show the following: If μ satisfies (2.149), then it also satisfies $\mu^{n^+} = \mu^{n^-}$ for all $n \in \mathcal{L}^\perp$. Write $n = \sum_{k=1}^r c_k$ as a sign-consistent sum of circuit vectors c_k, as in the last lemma. Without loss of generality, we can assume $c_k \in C$ for all k. Then $n^+ = \sum_{k=1}^r c_k^+$ and $n^- = \sum_{k=1}^r c_k^-$. Hence μ satisfies

$$\mu^{n^+} - \mu^{n^-} = \mu^{\sum_{k=2}^r c_k^+}\left(\mu^{c_1^+} - \mu^{c_1^-}\right) + \left(\mu^{\sum_{k=2}^r c_k^+} - \mu^{\sum_{k=2}^r c_k^-}\right)\mu^{c_1^-},$$

so the theorem follows easily by induction. □

Example 2.5 Let $I := \{1, 2, 3, 4\}$, and consider the vector space $\mathcal{L}$ spanned by the following two functions (here, we write functions on I as row vectors of length 4):

$$f_0 = (1, 1, 1, 1) \quad \text{and} \quad f_1 = (-\alpha, 1, 0, 0),$$

where $\alpha \notin \{0, 1\}$ is arbitrary. This generates a one-dimensional exponential family $\mathcal{E}(\mathcal{L})$. The kernel of $\mathcal{L}$ is then spanned by

$$n_1 = (1, \alpha, -1, -\alpha) \quad \text{and} \quad n_2 = (1, \alpha, -\alpha, -1),$$

but these two vectors do not form a circuit basis: They correspond to the two relations

$$\mu_1 {\mu_2}^\alpha = \mu_3 {\mu_4}^\alpha \quad \text{and} \quad \mu_1 {\mu_2}^\alpha = {\mu_3}^\alpha \mu_4. \tag{2.150}$$

It follows immediately that

$$\mu_3 {\mu_4}^\alpha = {\mu_3}^\alpha \mu_4. \tag{2.151}$$

If $\mu_3 \mu_4$ is not zero, then we conclude that $\mu_3 = \mu_4$. However, on the boundary this does not follow from Eqs. (2.150): Possible solutions to these equations are given by

$$\mu^{(a)} = (0, a, 0, 1 - a) \quad \text{for } 0 \leq a < 1. \tag{2.152}$$

However, $\mu^{(a)}$ does not lie in the closure of the exponential family $\mathcal{E}(\mathcal{L})$, since all members of $\mathcal{E}(\mathcal{L})$ satisfy $\mu_3 = \mu_4$.

A circuit basis of A is given by the vectors

$$(0, 0, 1, -1), \quad (1, \alpha, 0, -1 - \alpha), \quad \text{and} \quad (1, \alpha, -1 - \alpha, 0),$$

which have the following corresponding equations:

$$\mu_3 = \mu_4, \quad \mu_1 {\mu_2}^\alpha = {\mu_4}^{1+\alpha}, \quad \text{and} \quad \mu_1 {\mu_2}^\alpha = {\mu_3}^{1+\alpha}. \tag{2.153}$$

By Theorem 2.7, these three equations characterize $\overline{\mathcal{E}}(\mathcal{L})$.

Using arguments from matroid theory, the number of circuits can be shown to be less than or equal to $\binom{m}{d+2}$, where $m = |I|$ is the size of the state space and d is the dimension of $\mathcal{E}(\mu_0, \mathcal{L})$, see [83]. This gives an upper bound on the number of implicit equations describing $\overline{\mathcal{E}}(\mu_0, \mathcal{L})$. Note that $\binom{m}{d+2}$ is usually much larger than the codimension $m - d - 1$ of $\mathcal{E}(\mu_0, \mathcal{L})$ in the probability simplex. In contrast, if we only want to find an implicit description of all probability distributions of $\mathcal{E}(\mu_0, \mathcal{L})$, which have full support, then $m - d - 1$ equations are enough.

It turns out that even in the boundary the number of equations can be further reduced: In general we do not need all circuits for the implicit description of $\overline{\mathcal{E}}(\mu_0, \mathcal{L})$. For instance, in Example 2.5, the second and third equation of (2.153) are equivalent given the first one, i.e., we only need two of the three circuits to describe $\overline{\mathcal{E}}(\mu_0, \mathcal{L})$.

2.8.3 Information Projections

In Sect. 2.7.2 we have introduced the relative entropy. It turns out to be the right divergence function for projecting probability measures onto exponential families.

These projections, referred to as *information projections*, are closely related to large-deviation theory and maximum-likelihood estimation in statistics. The foundational work on information projections is due to Chentsov [65] and Csiszár [75] (see also the tutorial by Csiszár and Shields [78]). Csiszár and Matúš revisited the classical theory of information projections within a much more general setting [76]. The differential-geometric study of these projections and their generalizations based on dually flat structures (see Sect. 4.3) is due to Amari and Nagaoka [8, 16, 194].

In order to treat the most general case, where probability distributions do not have to be strictly positive, we have to extend the relative entropy or KL-divergence (2.101) of Definition 2.8 so that it is defined for general probability distributions $\mu, \nu \in \mathcal{P}(I)$. It turns out that, although D_{KL} is continuous on the product $\mathcal{P}_+(I) \times \mathcal{P}_+(I)$, there is no continuous extension to the Cartesian product $\mathcal{P}(I) \times \mathcal{P}(I)$. As D_{KL} is used for minimization problems, it is reasonable to consider the following lower semi-continuous extension of D_{KL} with values in the extended line $\overline{\mathbb{R}}_+ := \{x \in \mathbb{R} : x \geq 0\} \cup \{\infty\}$ (considered as a topological space where $U \subseteq \overline{\mathbb{R}}_+$ is a neighborhood of ∞ if it contains an interval (x, ∞)):

$$D_{KL}(\mu \parallel \nu) := \begin{cases} \sum_{i \in I} \mu_i \log \frac{\mu_i}{\nu_i}, & \text{if } \operatorname{supp}(\mu) \subseteq \operatorname{supp}(\nu), \\ \infty, & \text{otherwise.} \end{cases} \tag{2.154}$$

Here, we use the convention $\mu_i \log \frac{\mu_i}{\nu_i} = 0$ whenever $\mu_i = 0$. Defining $\mu_i \log \frac{\mu_i}{\nu_i}$ to be ∞ if $\mu_i > \nu_i = 0$ allows us to rewrite (2.154) as $D_{KL}(\mu \parallel \nu) = \sum_{i \in I} \mu_i \log \frac{\mu_i}{\nu_i}$. Whenever appropriate, we use this concise expression.

We summarize the basic properties of this (extended) relative entropy or KL-divergence.

Proposition 2.14 *The function* (2.154) *satisfies the following properties*:

(1) $D_{KL}(\mu \parallel \nu) \geq 0$, *and* $D_{KL}(\mu \parallel \nu) = 0$ *if and only if* $\mu = \nu$.
(2) *The functions* $D_{KL}(\mu \parallel \cdot)$ *and* $D_{KL}(\cdot \parallel \nu)$ *are continuous for all* $\mu, \nu \in \mathcal{P}(I)$.
(3) D_{KL} *is lower semi-continuous, that is, for all* $(\mu^{(k)}, \nu^{(k)}) \to (\mu, \nu)$, *we have*

$$D_{KL}(\mu \parallel \nu) \leq \liminf_{k \to \infty} D_{KL}\big(\mu^{(k)} \parallel \nu^{(k)}\big). \tag{2.155}$$

(4) D_{KL} *is jointly convex, that is, for all* $\mu^{(j)}, \nu^{(j)}, \lambda_j \in [0, 1]$, $j = 1, \ldots, n$, *satisfying* $\sum_{j=1}^n \lambda_j = 1$,

$$D_{KL}\left(\sum_{j=1}^n \lambda_j \, \mu^{(j)} \middle\| \sum_{j=1}^n \lambda_j \, \nu^{(j)}\right) \leq \sum_{j=1}^n \lambda_j \, D_{KL}\big(\mu^{(j)} \parallel \nu^{(j)}\big). \tag{2.156}$$

The proof of Proposition 2.14 involves the following basic inequality.

Lemma 2.10 (log-sum inequality) *For arbitrary non-negative real numbers* $a_1, \ldots, a_m$ *and* $b_1, \ldots, b_m$, *we have*

$$\sum_{k=1}^{m} a_k \log \frac{a_k}{b_k} \geq \left(\sum_{k=1}^{m} a_k \right) \log \frac{\sum_{k=1}^{m} a_k}{\sum_{k=1}^{m} b_k}, \tag{2.157}$$

where equality holds if and only if $\frac{a_k}{b_k}$ *is independent of* k. *Here,* $a \log \frac{a}{b}$ *is defined to be* 0 *if* $a = 0$ *and* ∞ *if* $a > b = 0$.

Proof We set $a := \sum_{k=1}^{m} a_k$ and $b := \sum_{k=1}^{m} b_k$. With the strict convexity of the function $f : [0, \infty) \to \mathbb{R}$, $f(x) := x \log x$ for $x > 0$ and $f(0) = 0$, we obtain

$$\begin{aligned} \sum_{k=1}^{m} a_k \log \frac{a_k}{b_k} &= \sum_{k=1}^{m} b_k \frac{a_k}{b_k} \log \frac{a_k}{b_k} = b \sum_{k=1}^{m} \frac{b_k}{b} f\left(\frac{a_k}{b_k} \right) \\ &\geq b\, f\left(\sum_{k=1}^{m} \frac{b_k}{b} \frac{a_k}{b_k} \right) = b f\left(\frac{a}{b} \right) = a \log \frac{a}{b}. \end{aligned}$$

□

Proof of Proposition 2.14

(1) In the case of probability measures, the log-sum inequality (2.157) implies

$$\sum_{i \in I} \mu_i \log \frac{\mu_i}{\nu_i} \geq \left(\sum_{i \in I} \mu_i \right) \log \frac{\sum_{i \in I} \mu_i}{\sum_{i \in I} \nu_i} = 0,$$

where equality holds if and only if $\mu_i = c\, \nu_i$ for some constant c, which, in this case, has to be equal to one.

(2) This follows directly from the continuity of the functions $\log x$, where $\log 0 := -\infty$, and $x \log x$, where $0 \log 0 := 0$.

(3) If $\nu_i > 0$, we have

$$\lim_{k \to \infty} \mu_i^{(k)} \log \frac{\mu_i^{(k)}}{\nu_i^{(k)}} = \mu_i \log \frac{\mu_i}{\nu_i}. \tag{2.158}$$

If $\nu_i = 0$ and $\mu_i > 0$,

$$\lim_{k \to \infty} \mu_i^{(k)} \log \frac{\mu_i^{(k)}}{\nu_i^{(k)}} = \infty = \mu_i \log \frac{\mu_i}{\nu_i}. \tag{2.159}$$

Finally, if $\nu_i = 0$ and $\mu_i = 0$,

$$\liminf_{k \to \infty} \mu_i^{(k)} \log \frac{\mu_i^{(k)}}{\nu_i^{(k)}} \geq \liminf_{k \to \infty} \left(\mu_i^{(k)} - \nu_i^{(k)} \right) = 0 = \mu_i \log \frac{\mu_i}{\nu_i}, \tag{2.160}$$

since $\log x \geq 1 - \frac{1}{x}$ for $x > 0$. Altogether, (2.158), (2.159), and (2.160) imply

$$\liminf_{k\to\infty} \sum_{i\in I} \mu_i^{(k)} \log \frac{\mu_i^{(k)}}{\nu_i^{(k)}} \geq \sum_{i\in I} \liminf_{k\to\infty} \mu_i^{(k)} \log \frac{\mu_i^{(k)}}{\nu_i^{(k)}} \geq \sum_{i\in I} \mu_i \log \frac{\mu_i}{\nu_i},$$

which equals ∞ whenever there is at least one $i \in I$ satisfying $\mu_i > \nu_i = 0$.

(4) We use again the log-sum inequality (2.157):

$$\begin{aligned} D_{KL}\left(\sum_{j=1}^n \lambda_j \mu^{(j)} \middle\| \sum_{j=1}^n \lambda_j \nu^{(j)}\right) &= \sum_{i\in I}\left(\sum_{j=1}^n \lambda_j \mu_i^{(j)}\right) \log \frac{\sum_{j=1}^n \lambda_j \mu_i^{(j)}}{\sum_{j=1}^n \lambda_j \nu_i^{(j)}} \\ &\leq \sum_{i\in I}\sum_{j=1}^n \lambda_j \mu_i^{(j)} \log \frac{\lambda_j \mu_i^{(j)}}{\lambda_j \nu_i^{(j)}} \\ &= \sum_{j=1}^n \lambda_j \sum_{i\in I} \mu_i^{(j)} \log \frac{\mu_i^{(j)}}{\nu_i^{(j)}} \\ &= \sum_{j=1}^n \lambda_j D_{KL}\big(\mu^{(j)} \,\|\, \nu^{(j)}\big). \end{aligned}$$

□

We consider information projections onto exponential and corresponding mixture families. They are assigned to a linear subspace $\mathcal{L}$ of $\mathcal{F}(I)$ and measures $\mu_1 \in \mathcal{P}(I)$, $\mu_2 \in \mathcal{P}_+(I)$. Without loss of generality, we assume $\mathbb{1} \in \mathcal{L}$ and choose a basis $f_0 := \mathbb{1}, f_1, \ldots, f_d$ of $\mathcal{L}$. The mixture family through μ_1 is simply the set of distributions that have the same expectation values of the f_k as the distribution μ_1, that is,

$$\mathcal{M} := \mathcal{M}(\mu_1, \mathcal{L}) := \big\{\nu \in \mathcal{P}(I) : \nu(f_k) = \mu_1(f_k), k = 1, \ldots, d\big\}. \tag{2.161}$$

The corresponding exponential family $\mathcal{E} := \mathcal{E}(\mu_2, \mathcal{L})$ through μ_2 is given as the image of the parametrization

$$\vartheta = (\vartheta_1, \ldots, \vartheta_d) \mapsto \frac{1}{Z(\vartheta)} \sum_{i\in I} \mu_{2,i} \exp\left(\sum_{k=1}^d \vartheta_k f_k(i)\right)\delta^i, \tag{2.162}$$

where

$$Z(\vartheta) := \sum_j \mu_{2,j} \exp\left(\sum_{k=1}^d \vartheta_k f_k(j)\right).$$

Theorem 2.8 *For any distribution $\hat{\mu} \in \mathcal{P}(I)$, the following statements are equivalent*:

(1) $\hat{\mu} \in \mathcal{M} \cap \overline{\mathcal{E}}$.

(2) *For all* $\nu_1 \in \mathcal{M}$, $\nu_2 \in \overline{\mathcal{E}}$: $D_{KL}(\nu_1 \,\|\, \hat{\mu}) < \infty$, $D_{KL}(\nu_1 \,\|\, \nu_2) < \infty$ *iff* $D_{KL}(\hat{\mu} \,\|\, \nu_2) < \infty$, *and*

$$D_{KL}(\nu_1 \,\|\, \nu_2) = D_{KL}(\nu_1 \,\|\, \hat{\mu}) + D_{KL}(\hat{\mu} \,\|\, \nu_2). \tag{2.163}$$

In particular, the intersection $\mathcal{M} \cap \overline{\mathcal{E}}$ *consists of the single point* $\hat{\mu}$.

(3) $\hat{\mu} \in \mathcal{M}$, *and* $D_{KL}(\hat{\mu} \,\|\, \nu_2) = \inf_{\nu \in \mathcal{M}} D_{KL}(\nu \,\|\, \nu_2)$ *for all* $\nu_2 \in \overline{\mathcal{E}}$.

(4) $\hat{\mu} \in \overline{\mathcal{E}}$, *and* $D_{KL}(\nu_1 \,\|\, \hat{\mu}) = \inf_{\nu \in \overline{\mathcal{E}}} D_{KL}(\nu_1 \,\|\, \nu)$ *for all* $\nu_1 \in \mathcal{M}$.

Furthermore, there exists a unique distribution $\hat{\mu}$ *that satisfies one and therefore all of these conditions.*

Proof (1) $\Rightarrow$ (2) We choose $\nu_1 \in \mathcal{M}$ and $\nu_2 \in \mathcal{E}$ (strict positivity). As $\hat{\mu} \in \overline{\mathcal{E}}$, there is a sequence $\mu^{(n)} \in \mathcal{E}$, $\mu^{(n)} \to \hat{\mu}$. This implies

$$\sum_{i \in I} (\nu_{1,i} - \hat{\mu}_i) \log \frac{\mu_i^{(n)}}{\nu_{2,i}} = 0. \tag{2.164}$$

This is because $\log \frac{d\mu^{(n)}}{d\nu_2} \in \mathcal{L}$, and $\nu_1, \hat{\mu} \in \mathcal{M}$. By continuity,

$$\sum_{i \in I} (\nu_{1,i} - \hat{\mu}_i) \log \frac{\hat{\mu}_i}{\nu_{2,i}} = 0. \tag{2.165}$$

This equality is equivalent to

$$D_{KL}(\nu_1 \,\|\, \nu_2) = D_{KL}(\nu_1 \,\|\, \hat{\mu}) + D_{KL}(\hat{\mu} \,\|\, \nu_2). \tag{2.166}$$

As we assumed ν_2 to be strictly positive, this means that $D_{KL}(\nu_1 \,\|\, \nu_2)$ and $D_{KL}(\hat{\mu} \,\|\, \nu_2)$ are finite, so that $D_{KL}(\nu_1 \,\|\, \hat{\mu})$ has to be finite. This is only the case if $\operatorname{supp}(\hat{\mu}) \supseteq \operatorname{supp}(\nu_1)$ for all $\nu_1 \in \mathcal{M}$. By continuity, (2.166) also holds for $\nu_2 \in \overline{\mathcal{E}}$. Note, however, that we do not exclude the case where $D_{KL}(\nu_1 \,\|\, \nu_2)$ and $D_{KL}(\hat{\mu} \,\|\, \nu_2)$ become infinite when ν_2 does not have full support. We finally prove uniqueness: Assume $\hat{\mu}' \in \mathcal{M} \cap \overline{\mathcal{E}}$. Then the Pythagorean relation (2.163) implies for $\nu_1 = \nu_2 = \hat{\mu}'$

$$0 = D_{KL}(\nu_1 \,\|\, \nu_2) = D_{KL}(\nu_1 \,\|\, \hat{\mu}) + D_{KL}(\hat{\mu} \,\|\, \nu_2) = D_{KL}\big(\hat{\mu}' \,\big\|\, \hat{\mu}\big) + D_{KL}\big(\hat{\mu} \,\big\|\, \hat{\mu}'\big),$$

and therefore $\hat{\mu}' = \hat{\mu}$, that is, $\mathcal{M} \cap \overline{\mathcal{E}} = \{\hat{\mu}\}$.

(2) $\Rightarrow$ (3) For all $\nu_1 \in \mathcal{M}$ and $\nu_2 \in \overline{\mathcal{E}}$, we obtain

$$D_{KL}(\nu_1 \,\|\, \nu_2) = D_{KL}(\nu_1 \,\|\, \hat{\mu}) + D_{KL}(\hat{\mu} \,\|\, \nu_2) \geq D_{KL}(\hat{\mu} \,\|\, \nu_2) \geq \inf_{\nu \in \mathcal{M}} D_{KL}(\nu \,\|\, \nu_2).$$

This implies

$$\inf_{\nu_1 \in \mathcal{M}} D_{KL}(\nu_1 \,\|\, \nu_2) \geq D_{KL}(\hat{\mu} \,\|\, \nu_2) \geq \inf_{\nu \in \mathcal{M}} D_{KL}(\nu \,\|\, \nu_2).$$

(2) $\Rightarrow$ (4) For all $\nu_1 \in \mathcal{M}$ and $\nu_2 \in \overline{\mathcal{E}}$, we obtain

$$D_{KL}(\nu_1 \,\|\, \nu_2) = D_{KL}(\nu_1 \,\|\, \hat{\mu}) + D_{KL}(\hat{\mu} \,\|\, \nu_2) \geq D_{KL}(\nu_1 \,\|\, \hat{\mu}) \geq \inf_{\nu \in \overline{\mathcal{E}}} D_{KL}(\nu_1 \,\|\, \nu).$$

This implies

$$\inf_{\nu_2 \in \overline{\mathcal{E}}} D_{KL}(\nu_1 \,\|\, \nu_2) \geq D_{KL}(\nu_1 \,\|\, \hat{\mu}) \geq \inf_{\nu \in \overline{\mathcal{E}}} D_{KL}(\nu_1 \,\|\, \nu).$$

(3) $\Rightarrow$ (1) We consider the KL-divergence for $\nu_2 = \mu_2$, the base measure of the exponential family $\mathcal{E} = \mathcal{E}(\mu_2, \mathcal{L})$ which we have assumed to be strictly positive:

$$\mathcal{M} \to \mathbb{R}, \qquad \nu \mapsto D_{KL}(\nu \,\|\, \mu_2). \tag{2.167}$$

This function is strictly convex and therefore has $\hat{\mu} \in \mathcal{M}$ as its unique minimizer. In what follows, we prove that $\hat{\mu}$ is contained in the (relative) interior of $\mathcal{M}$ so that we can derive a necessary condition for $\hat{\mu}$ using the method of Lagrange multipliers. This necessary condition then implies $\hat{\mu} \in \overline{\mathcal{E}}$. We structure this chain of arguments in three steps:

Step 1: Define the curve $[0, 1] \to \mathcal{M}$, $t \mapsto \nu(t) := (1 - t)\,\hat{\mu} + t\,\nu$, and consider its derivative

$$\frac{d}{dt} D_{KL}\big(\nu(t) \,\big\|\, \mu_2\big)\Big|_{t=t_0} = \sum_{i \in I} (\nu_i - \hat{\mu}_i) \log \frac{\nu_i(t_0)}{\mu_{2,i}} \tag{2.168}$$

for $t_0 \in (0, 1)$. If $\hat{\mu}_i = 0$ for some i with $\nu_i > 0$ then the derivative (2.168) converges to $-\infty$ when $t_0 \to 0$. As $\hat{\mu}$ is the minimizer of $D_{KL}(\nu(\cdot) \,\|\, \mu_2)$, this is ruled out, proving

$$\operatorname{supp}(\nu) \subseteq \operatorname{supp}(\hat{\mu}), \quad \text{for all } \nu \in \mathcal{M}. \tag{2.169}$$

Step 2: Let us now consider the particular situation where $\mathcal{M}$ has a non-empty intersection with $\mathcal{P}_+(I)$. This is obviously the case, when we choose μ_1 to be strictly positive, as $\mu_1 \in \mathcal{M} = \mathcal{M}(\mu_1, \mathcal{L})$ by definition. In that case $\operatorname{supp}(\hat{\mu}) = I$, by (2.169). We consider the restriction of the function (2.167) to $\mathcal{M} \cap \mathcal{P}_+(I)$ and introduce Lagrange multipliers $\vartheta_0, \vartheta_1, \ldots, \vartheta_d$, in order to obtain a necessary condition for $\hat{\mu}$ to be its minimizer. More precisely, differentiating

$$\sum_{i \in I} \nu_i \log \frac{\nu_i}{\mu_{2,i}} - \vartheta_0 \left(1 - \sum_{i \in I} \nu_i\right) - \sum_{k=1}^{d} \vartheta_k \left(\mu_1(f_k) - \sum_{i \in I} \nu_i f_k(i)\right) \tag{2.170}$$

with respect to ν_i leads to the necessary condition

$$\log \nu_i + 1 - \log \mu_{2,i} - \vartheta_0 - \sum_k \vartheta_k f_k(i) = 0, \quad i \in I,$$

which is equivalent to

$$\nu_i = \mu_{2,i} \exp\left(\vartheta_0 - 1 + \sum_{k=1}^{d} \vartheta_k f_k(i)\right), \quad i \in I. \tag{2.171}$$

As the minimizer, $\hat{\mu}$ has this structure and is therefore contained in $\mathcal{E}$, proving $\hat{\mu} \in \mathcal{M} \cap \mathcal{E}$.

Step 3: In this final step, we drop the assumption that μ_1 is strictly positive and consider the sequence

$$\mathcal{M}^{(n)} := \mathcal{M}\big(\mu_1^{(n)}, \mathcal{L}\big), \tag{2.172}$$

where $\mu_1^{(n)} = (1 - \frac{1}{n})\,\mu_1 + \frac{1}{n}\,\mu_2$, $n \in \mathbb{N}$. Each of these distributions $\mu_1^{(n)}$ is strictly positive so that, according to Step 2, we have a corresponding sequence $\hat{\mu}^{(n)}$ of distributions in $\mathcal{M}^{(n)} \cap \mathcal{E}$. The limit of any convergent subsequent is an element of $\mathcal{M} \cap \overline{\mathcal{E}}$ and, by uniqueness, coincides with $\hat{\mu}$.

(4) $\Rightarrow$ (1) Define $S := \operatorname{supp}(\hat{\mu})$. Then $\hat{\mu}$ is contained in the family $\mathcal{E}_S$ (see Theorem 2.6), defined in terms of the parametrization

$$\nu_i(\vartheta) := \nu_i(\vartheta_1, \ldots, \vartheta_d) := \begin{cases} \frac{1}{Z_S(\vartheta)}\,\mu_{2,i} \exp\left(\sum_{k=1}^{d} \vartheta_k\, f_k(i)\right), & \text{if } i \in S, \\ 0, & \text{otherwise,} \end{cases}$$

with

$$Z_S(\vartheta) := \sum_{j \in S} \mu_{2,j} \exp\left(\sum_{k=1}^{d} \vartheta_k\, f_k(j)\right).$$

Note that $\nu(\hat{\vartheta}) = \hat{\mu}$ for some $\hat{\vartheta}$. With this parametrization, we obtain the function

$$D_{KL}\big(\nu_1 \,\big\|\, \nu(\vartheta)\big) = D_{KL}(\nu_1 \,\|\, \mu_2) - \sum_{k=1}^{d} \vartheta_k\, \nu_1(f_k) + \log\big(Z_S(\vartheta)\big),$$

and its partial derivatives

$$\frac{\partial D_{KL}(\nu_1 \,\|\, \nu(\cdot))}{\partial \vartheta_k} = \nu(\vartheta)(f_k) - \nu_1(f_k), \quad k = 1, \ldots, d. \tag{2.173}$$

As $\hat{\mu} = \nu(\hat{\vartheta})$ is the minimizer, $\hat{\vartheta}$ satisfies Eqs. (2.173), which implies that $\hat{\mu}$ is contained in $\mathcal{M}$.

Existence: We proved the equivalence of the conditions for any distribution $\hat{\mu}$, which, in particular, implies the uniqueness of a distribution that satisfies one and therefore all of these conditions. To see that there exists such a distribution, consider the function (2.167) and observe that it has a unique minimizer $\hat{\mu} \in \mathcal{M} \cap \overline{\mathcal{E}}$ (see the proof of the implication "(3) $\Rightarrow$ (1)"). □

Let us now use Theorem 2.8 in order to define projections onto mixture and exponential families, referred to as the *I-projection* and *rI-projection*, respectively (see [76, 78]).

Let us begin with the I-projection. Consider a mixture family $\mathcal{M} = \mathcal{M}(\mu_1, \mathcal{L})$ as defined by (2.161). Following the criterion (3) of Theorem 2.8, we define the *distance from* $\mathcal{M}$ by

$$D_{KL}(\mathcal{M} \,\|\, \cdot) : \mathcal{P}(I) \to \mathbb{R}, \qquad \mu \mapsto D_{KL}(\mathcal{M} \,\|\, \mu) := \inf_{\nu \in \mathcal{M}} D_{KL}(\nu \,\|\, \mu). \tag{2.174}$$

Theorem 2.8 implies that there is a unique point $\hat{\mu} \in \mathcal{M}$ that satisfies $D_{KL}(\hat{\mu} \,\|\, \mu) = D_{KL}(\mathcal{M} \,\|\, \mu)$. It is obtained as the intersection of $\mathcal{M}(\mu_1, \mathcal{L})$ with the closure of the exponential family $\mathcal{E}(\mu, \mathcal{L})$. This allows us to define the I-projection $\pi_{\mathcal{M}} : \mathcal{P}(I) \to \mathcal{M}$, $\mu \mapsto \hat{\mu}$.

Now let us come to the analogous definition of the rI-projection, which will play an important role in Sect. 6.1. Consider an exponential family $\mathcal{E} = \mathcal{E}(\mu_2, \mathcal{L})$, and, following criterion (4) of Theorem 2.8, define the *distance from* $\mathcal{E}$ by

$$D_{KL}(\cdot \,\|\, \mathcal{E}) : \mathcal{P}(I) \to \mathbb{R}, \qquad \mu \mapsto D_{KL}(\mu \,\|\, \mathcal{E}) := \inf_{\nu \in \mathcal{E}} D_{KL}(\mu \,\|\, \nu) = \inf_{\nu \in \overline{\mathcal{E}}} D_{KL}(\mu \,\|\, \nu), \tag{2.175}$$

where the last equality follows from the continuity of $D_{KL}(\mu \,\|\, \cdot)$. Theorem 2.8 implies that there is a unique point $\hat{\mu} \in \overline{\mathcal{E}}$ that satisfies $D_{KL}(\mu \,\|\, \hat{\mu}) = D_{KL}(\mu \,\|\, \mathcal{E})$. It is obtained as the intersection of the closure of $\mathcal{E}(\mu_2, \mathcal{L})$ with the mixture family $\mathcal{M}(\mu, \mathcal{L})$. This allows us to define the projection $\pi_{\mathcal{E}} : \mathcal{P}(I) \to \overline{\mathcal{E}}$, $\mu \mapsto \hat{\mu}$.

Proposition 2.15 *Both information distances, $D_{KL}(\mathcal{M} \,\|\, \cdot)$ and $D_{KL}(\cdot \,\|\, \mathcal{E})$, are continuous functions on $\mathcal{P}(I)$.*

Proof We prove the continuity of $D_{KL}(\mathcal{M} \,\|\, \cdot)$. One can prove the continuity of $D_{KL}(\cdot \,\|\, \mathcal{E})$ following the same reasoning.

Let μ be a point in $\mathcal{P}(I)$ and $\mu_n \in \mathcal{P}(I)$, $n \in \mathbb{N}$, a sequence that converges to μ. For all $\nu \in \mathcal{M}$, we have $D_{KL}(\mathcal{M} \,\|\, \mu_n) \leq D_{KL}(\nu \,\|\, \mu_n)$, $n \in \mathbb{N}$, and by the continuity of $D_{KL}(\nu \,\|\, \cdot)$ we obtain

$$\limsup_{n \to \infty} D_{KL}(\mathcal{M} \,\|\, \mu_n) \leq \limsup_{n \to \infty} D_{KL}(\nu \,\|\, \mu_n) = \lim_{n \to \infty} D_{KL}(\nu \,\|\, \mu_n) = D_{KL}(\nu \,\|\, \mu). \tag{2.176}$$

From the lower semi-continuity of the KL-divergence D_{KL} (Lemma 2.14), we obtain the lower semi-continuity of the distance $D_{KL}(\mathcal{M} \,\|\, \cdot)$ (see [228]). Taking the infimum of the RHS of (2.176) then leads to

$$\limsup_{n \to \infty} D_{KL}(\mathcal{M} \,\|\, \mu_n) \leq \inf_{\nu \in \mathcal{M}} D_{KL}(\nu \,\|\, \mu) = D_{KL}(\mathcal{M} \,\|\, \mu) \leq \liminf_{n \to \infty} D_{KL}(\mathcal{M} \,\|\, \mu_n), \tag{2.177}$$

proving $\lim_{n \to \infty} D_{KL}(\mathcal{M} \,\|\, \mu_n) = D_{KL}(\mathcal{M} \,\|\, \mu)$. □

In Sects. 2.9 and 6.1, we shall study exponential families that contain the uniform distribution, say $\mu_{0,i} = \frac{1}{|I|}$, $i \in I$. In that case, the projection $\hat{\mu} = \pi_{\mathcal{E}}(\mu)$ coincides

with the so-called *maximum entropy estimate* of μ. To be more precise, let us consider the function

$$\nu \mapsto D_{KL}(\nu \,\|\, \mu_0) = \log |I| - H(\nu), \tag{2.178}$$

where

$$H(\nu) := -\sum_{i \in I} \nu_i \log \nu_i . \tag{2.179}$$

The function $H(\nu)$ is the basic quantity of Shannon's theory of information [235], and it is known as the *Shannon entropy* or simply the *entropy*. It is continuous and strictly concave, because the function $f : [0, \infty) \to \mathbb{R}$, $f(x) := x \log x$ for $x > 0$ and $f(0) = 0$, is continuous and strictly convex. Therefore, H assumes its maximal value in a unique point, subject to any linear constraint. Clearly, ν minimizes (2.178) in a linear family $\mathcal{M}$ if and only if it maximizes the entropy on $\mathcal{M}$. Therefore, assuming that the uniform distribution is an element of $\mathcal{E}$, the distribution $\hat{\mu}$ of Theorem 2.8 is the one that maximizes the entropy, given the linear constraints of the set $\mathcal{M}$. This relates the information projection to the *maximum entropy method*, which has been proposed by Jaynes [128, 129] as a general inference method, based on statistical mechanics and information theory. In order to motivate this method, let us briefly elaborate on the information-theoretic interpretation of the entropy as an information gain, due to Shannon [235]. We assume that the probability measure ν represents our expectation about the outcome of a random experiment. In this interpretation, the larger ν_i is the higher our confidence that the events i will be the outcome of the experiment. With expectations, there is always associated a surprise. If an event i is not expected prior to the experiment, that is, if ν_i is small, then it should be surprising to observe it as the outcome of the experiment. If that event, on the other hand, is expected to occur with high probability, that is, if ν_i is large, then it should not be surprising at all to observe it as the experiment's outcome. It turns out that the right measure of surprise is given by the function $\nu_i \mapsto -\log \nu_i$. This function quantifies the extent to which one is surprised by the outcome of an event $i \in I$, if i was expected to occur with probability ν_i. The entropy of ν is the expected (or mean) surprise and is therefore a measure of the subjective uncertainty about the outcome of the experiment. The higher that uncertainty, the less information is contained in the probability distribution about the outcome of the experiment. As the uncertainty about this outcome is reduced to zero after having observed the outcome, this uncertainty reduction can be interpreted as information gain through the experiment. In his influential work [235], Shannon provided an axiomatic characterization of information based on this intuition.

The interpretation of entropy as uncertainty allows us to interpret the distance between μ and its maximum entropy estimate $\hat{\mu} = \pi_{\mathcal{E}}(\mu)$ as reduction of uncertainty.

Lemma 2.11 *Let $\hat{\mu}$ be the maximum entropy estimate of μ. Then*

$$D_{KL}(\mu \,\|\, \mathcal{E}) = D_{KL}(\mu \,\|\, \hat{\mu}) = H(\hat{\mu}) - H(\mu). \tag{2.180}$$

Proof It follows from (2.163) that

$$D_{KL}(\mu \,\|\, \mu_0) = D_{KL}(\mu \,\|\, \hat{\mu}) + D_{KL}(\hat{\mu} \,\|\, \mu_0).$$

If we choose μ_0 to be the uniform distribution, this amounts to

$$\big(\log |I| - H(\mu)\big) = D_{KL}(\mu \,\|\, \hat{\mu}) + \big(\log |I| - H(\hat{\mu})\big). \qquad \square$$

This will be further explored in Sect. 6.1.2. See also the discussion in Sect. 4.3 of the duality between exponential and mixture families.

2.9 Hierarchical and Graphical Models

We have derived and studied exponential families $\mathcal{E}(\mu_0, \mathcal{L})$ from a purely geometric perspective (see Definition 2.10). On the other hand, they naturally appear in statistical physics, known as families of Boltzmann–Gibbs distributions. In that context, there is an energy function which we consider to be an element of a linear space $\mathcal{L}$. One typically considers a number of particles that interact with each other so that the energy is decomposed into a family of interaction terms, the *interaction potential*. The strength of interaction, for instance, then parametrizes the space $\mathcal{L}$ of energies, and one can study how particular system properties change as result of a parameter change. This mechanistic description of a system consisting of interacting units has inspired corresponding models in many other fields, such as the field of neural networks, genetics, economics, etc.

It is remarkable that purely geometric information about the system can reveal relevant features of the physical system. For instance, the Riemannian curvature with respect to the Fisher metric can help us to detect critical parameter values where a phase transition occurs [54, 218]. Furthermore, the Gibbs–Markov equivalence in statistical physics [191], the equivalence of particular mechanistic and phenomenological properties of a system, can be interpreted as an equivalence of two perspectives of the same geometric object, its explicit parametrization and its implicit description as the solution set of corresponding equations. This close connection between geometry and physical systems allows us to assign to the developed geometry a mechanistic interpretation.

In this section we want to present a more refined view of exponential families that are naturally defined for systems of interacting units, so-called hierarchical models. Of particular interest are graphical models, where the interaction is compatible with a graph. For these models, the Gibbs–Markov equivalence is then stated by the Hammersley–Clifford theorem. The material of this section is mainly based on Lauritzen's monograph [161] on graphical models. However, we shall confine ourselves to discrete models with the counting measure as the base measure. The next section on interaction spaces incorporates work of Darroch and Speed [80].

2.9.1 Interaction Spaces

Consider a finite set V of *units* or *nodes*. To simplify the notation we sometimes choose V to be the set $[N] = \{1, \ldots, N\}$. We assign to each node a corresponding (non-empty and finite) set of *configurations* I_v. For every subset $A \subseteq V$, the configurations on A are given by the Cartesian product

$$I_A := \bigtimes_{v \in A} I_v. \tag{2.181}$$

Note that in the case where A is the empty set, the product space consists of the empty sequence ϵ, that is, $I_\emptyset = \{\epsilon\}$. We have the natural projections

$$X_A : I_V \to I_A, \qquad (i_v)_{v \in V} \mapsto (i_v)_{v \in A}. \tag{2.182}$$

Given a distribution $p \in \mathcal{P}(I)$, the X_A become random variables and we denote the X_A-image of p by p_A and use the shorthand notation

$$p(i_A) := p_A(i_A) = \sum_{i_{V \setminus A}} p(i_A, i_{V \setminus A}), \quad i_A \in I_A. \tag{2.183}$$

Given an i_A with $p(i_A) > 0$, we define

$$p(i_B | i_A) := \frac{p(i_A, i_B)}{p(i_A)}. \tag{2.184}$$

By $\mathcal{F}_A$ we denote the algebra of real-valued functions $f \in \mathcal{F}(I_V)$ that only depend on A, that is, the image of the algebra homomorphism $\mathcal{F}(I_A) \to \mathcal{F}(I_V)$, $g \mapsto g \circ X_A$ (see the first paragraph of Sect. 2.1). This is called the space of *A-interactions*. Clearly this space has dimension $\prod_{v \in A} |I_v|$. Note that we recover the one-dimensional space of constant functions on I_V for $A = \emptyset$.

We consider the canonical scalar product on $\mathcal{F}_V = \mathcal{F}(I_V)$, defined by $\langle f, g \rangle := \sum_i f^i g^i$, which coincides with the scalar product (2.10) for the counting measure $\mu = \sum_i \delta^i$. With the *A-marginal*

$$f(i_A) := \sum_{i'_{V \setminus A}} f\left(i_A, i'_{V \setminus A}\right), \quad i_A \in I_A, \tag{2.185}$$

of a function $f \in \mathcal{F}_V$, the orthogonal projection P_A onto $\mathcal{F}_A$ with respect to $\langle \cdot, \cdot \rangle$ has the following form.

Proposition 2.16

$$P_A(f)(i_A, i_{V \setminus A}) = \frac{f(i_A)}{|I_{V \setminus A}|}, \quad i_A \in I_A,\ i_{V \setminus I} \in I_{V \setminus I}. \tag{2.186}$$

(Note that, according to our convention, $|I_\emptyset| = 1$.)

Proof We have to show

$$\begin{aligned}
&\big\langle f - P_A(f), g\big\rangle = 0 \quad \text{for all } g \in \mathcal{F}_A.\\
&\big\langle f - P_A(f), g\big\rangle = \langle f, g\rangle - \big\langle P_A(f), g\big\rangle\\
&\qquad = \sum_{i_A, i_{V\setminus A}} f(i_A, i_{V\setminus A})\, g\big(i_A, i'_{V\setminus A}\big) - \sum_{i_A, i_{V\setminus A}} \frac{f(i_A)}{|I_{V\setminus A}|}\, g\big(i_A, i'_{V\setminus A}\big)\\
&\qquad = \sum_{i_A} g\big(i_A, i'_{V\setminus A}\big) \underbrace{\sum_{i_{V\setminus A}} f(i_A, i_{V\setminus A})}_{= f(i_A)} - \sum_{i_A} \frac{f(i_A)}{|I_{V\setminus A}|} \sum_{i_{V\setminus A}} g\big(i_A, i'_{V\setminus A}\big)\\
&\qquad = \sum_{i_A} g\big(i_A, i'_{V\setminus A}\big)\, f(i_A) - \sum_{i_A} \frac{f(i_A)}{|I_{V\setminus A}|}\, |I_{V\setminus A}|\, g\big(i_A, i'_{V\setminus A}\big)\\
&\qquad = 0. \qquad \square
\end{aligned}$$

Note that

$$P_A P_B = P_B P_A = P_{A\cap B} \quad \text{for all } A, B \subseteq V. \tag{2.187}$$

We now come to the notion of pure interactions. The vector space of *pure A-interactions* is defined as

$$\widetilde{\mathcal{F}}_A := \mathcal{F}_A \cap \Big(\bigcap_{B \subsetneq A} \mathcal{F}_B{}^{\perp} \Big). \tag{2.188}$$

Here, the orthogonal complements are taken with respect to the scalar product $\langle \cdot, \cdot \rangle$. The space $\widetilde{\mathcal{F}}_A$ consists of functions that depend on arguments in A but not only on arguments of a proper subset B of A. Obviously, the following holds:

$$f \in \widetilde{\mathcal{F}}_A \quad \Leftrightarrow \quad f \in \mathcal{F}_A \text{ and } f(i_B) = 0 \text{ for all } B \subsetneq A \text{ and } i_B \in I_B, \tag{2.189}$$

where $f(i_B)$ is defined by (2.185). We denote the orthogonal projection of f onto $\widetilde{\mathcal{F}}_A$ by $\widetilde{P}_A$. The following holds:

$$P_A \widetilde{P}_B = \begin{cases} \widetilde{P}_B, & \text{if } B \subseteq A,\\ 0, & \text{otherwise.} \end{cases} \tag{2.190}$$

The first case is obvious. The second case follows from (2.189) (and $B \nsubseteq A \Rightarrow A \cap B \subsetneq B$):

$$P_A \widetilde{P}_B = P_A P_B \widetilde{P}_B = P_{A\cap B} \widetilde{P}_B = 0.$$

Proposition 2.17

(1) *The spaces* $\widetilde{\mathcal{F}}_A$, $A \subseteq V$, *of pure interactions are mutually orthogonal.*
(2) *For all* $A \subseteq V$, *we have*

$$\widetilde{P}_A = \sum_{B \subseteq A} (-1)^{|A \setminus B|} P_B, \qquad P_A = \sum_{B \subseteq A} \widetilde{P}_B. \tag{2.191}$$

(3) *The space of A-interactions,* $A \subseteq V$, *has the following orthogonal decomposition into spaces of pure interactions*:

$$\mathcal{F}_A = \bigoplus_{B \subseteq A} \widetilde{\mathcal{F}}_B. \tag{2.192}$$

(4) *For* $A \subseteq V$, *the dimension of the space of pure A-interactions is given by*

$$\dim(\widetilde{\mathcal{F}}_A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} \prod_{v \in B} |I_v| = \prod_{v \in A} \big(|I_v| - 1\big). \tag{2.193}$$

For the proof of Proposition 2.17 we need the Möbius inversion formula, which we state and prove first.

Lemma 2.12 (Möbius inversion) *Let* Ψ *and* Φ *be functions defined on the set of subsets of a finite set* V, *taking values in an Abelian group. Then the following statements are equivalent*:

(1) *For all* $A \subseteq V$, $\Psi(A) = \sum_{B \subseteq A} \Phi(B)$.
(2) *For all* $A \subseteq V$, $\Phi(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} \Psi(B)$.

Proof

$$\begin{aligned}
\sum_{B \subseteq A} \Phi(B) &= \sum_{B \subseteq A} \sum_{D \subseteq B} (-1)^{|B \setminus D|} \Psi(D) \\
&= \sum_{D \subseteq A,\, C \subseteq A \setminus D} (-1)^{|C|} \Psi(D) \\
&= \sum_{D \subseteq A} \Psi(D) \sum_{C \subseteq A \setminus D} (-1)^{|C|} \\
&= \Psi(A).
\end{aligned}$$

The last equality results from the fact that the inner sum equals 1, if $A \setminus D = \emptyset$. In the case $A \setminus D \neq \emptyset$ we set $n := |A \setminus D|$ and get

$$\begin{aligned}
\sum_{C \subseteq A \setminus D} (-1)^{|C|} &= \sum_{k=1}^{n} \big|\{C \subseteq A \setminus D : |C| = k\}\big| (-1)^k \\
&= \sum_{k=0}^{n} \binom{n}{k} (-1)^k
\end{aligned}$$

$$= (1-1)^n$$
$$= 0.$$

For the proof of the second implication, we use the same arguments:

$$\begin{aligned}\sum_{B\subseteq A}(-1)^{|A\setminus B|}\Psi(B) &= \sum_{D\subseteq B\subseteq A}(-1)^{|A\setminus B|}\Phi(D)\\ &= \sum_{D\subseteq A}\Phi(D)\sum_{C\subseteq A\setminus D}(-1)^{|C|}\\ &= \Phi(A).\end{aligned}$$

□

The Möbius inversion formula is of independent interest within discrete mathematics and has various generalizations (see [1]). We now come to the proof of the above proposition.

Proof of Proposition 2.17

(1) First observe that

$$\bigcap_{B\subsetneq A}\mathcal{F}_B{}^{\perp} = \bigcap_{v\in A}\mathcal{F}_{A\setminus\{v\}}{}^{\perp}. \tag{2.194}$$

Here, the inclusion "$\subseteq$" follows from the corresponding inclusion of the index sets: $\{A\setminus\{v\} : v\in A\}\subseteq\{B\subseteq A : B\neq A\}$. The opposite inclusion "$\supseteq$" follows from the fact that any $B\subsetneq A$ is contained in some $A\setminus\{v\}$, which implies $\mathcal{F}_B\subseteq\mathcal{F}_{A\setminus\{v\}}$ and therefore $\mathcal{F}_B{}^{\perp}\supseteq\mathcal{F}_{A\setminus\{v\}}{}^{\perp}$.

From (2.194) we obtain

$$\widetilde{\mathcal{F}}_A = \mathcal{F}_A\cap\bigcap_{B\subsetneq A}\mathcal{F}_B{}^{\perp} = \mathcal{F}_A\cap\bigcap_{v\in A}\mathcal{F}_{A\setminus\{v\}}{}^{\perp},$$

and therefore

$$\widetilde{P}_A = P_A\prod_{v\in A}(\mathrm{id}_{\mathcal{F}_V} - P_{A\setminus\{v\}}). \tag{2.195}$$

This implies that $\widetilde{P}_A$ and P_B commute. As a consequence, we derive

$$\widetilde{P}_A\widetilde{P}_B = P_A\widetilde{P}_A\widetilde{P}_B = \widetilde{P}_AP_A\widetilde{P}_B = 0 \quad \text{if } A\neq B.$$

The last equality follows from $P_A\widetilde{P}_B = 0$ according to (2.190), where $A\neq B$ implies $A\not\subseteq B$ or $B\not\subseteq A$. This yields

$$\widetilde{P}_A\widetilde{P}_B = \widetilde{P}_B\widetilde{P}_A \quad \text{if } A\neq B,$$

and therefore the spaces $\widetilde{\mathcal{F}}_A$, $A\subseteq V$, are mutually orthogonal.

(2) We use (2.195):

$$\begin{aligned}
\widetilde{P}_A &= P_A \prod_{v\in A} (\mathrm{id}_{\mathcal{F}_V} - P_{A\setminus\{v\}}) \\
&= P_A \sum_{B\subseteq A} (-1)^{|B|} \prod_{v\in B} P_{A\setminus\{v\}} \quad \text{(by direct multiplication)} \\
&= P_A \sum_{B\subseteq A} (-1)^{|B|} P_{A\setminus B} \\
&\quad \left(\text{iteration of (2.187), together with } \bigcap_{v\in B} \big(A\setminus\{v\}\big) = A\setminus B \right) \\
&= P_A \sum_{B\subseteq A} (-1)^{|A\setminus B|} P_B \quad \text{(change of summation index)} \\
&= \sum_{B\subseteq A} (-1)^{|A\setminus B|} P_A P_B \\
&= \sum_{B\subseteq A} (-1)^{|A\setminus B|} P_B. \quad (\mathcal{F}_B \text{ subspace of } \mathcal{F}_A)
\end{aligned}$$

This proves the first part of the statement. For the second part, we use the Möbius inversion of Lemma 2.12. It implies

$$P_A = \sum_{B\subseteq A} \widetilde{P}_B,$$

which is the second part of the statement.

(3) The inclusion "$\supseteq$" is clear. We prove the opposite inclusion "$\subseteq$":

$$f \in \mathcal{F}_A \quad \Rightarrow \quad f = P_A(f) = \sum_{B\subseteq A} \underbrace{\widetilde{P}_B(f)}_{\in \widetilde{\mathcal{F}}_B} \in \bigoplus_{B\subseteq A} \widetilde{\mathcal{F}}_B.$$

(4) From (2.192) we know

$$\dim(\mathcal{F}_A) = \sum_{B\subseteq A} \dim(\widetilde{\mathcal{F}}_B), \quad A \subseteq V.$$

The Möbius inversion formula implies

$$\begin{aligned}
\dim(\widetilde{\mathcal{F}}_A) &= \sum_{B\subseteq A} (-1)^{|A\setminus B|} \dim(\mathcal{F}_B) \\
&= \sum_{B\subseteq A} (-1)^{|A\setminus B|} \prod_{v\in B} |I_v| \\
&= \prod_{v\in A} \big(|I_v| - 1\big).
\end{aligned}$$

□

In the remaining part of this section, we concentrate on binary nodes $v \in V$ with state spaces $I_v = \{0, 1\}$ for all $v \in V$. In this case, by (2.193),

$$\dim(\widetilde{\mathcal{F}}_A) = \prod_{v \in A} \big(|I_v| - 1\big) = 1.$$

We are now going to define a family of vectors $e_A : I_V = \{0, 1\}^V \to \mathbb{R}$ that span the individual spaces $\widetilde{\mathcal{F}}_A$ and thereby form an orthogonal basis of $\mathcal{F}_V$. In order to do so, we interpret the symbols 0 and 1 as the elements of the group $\mathbb{Z}_2$, with group law determined by $1 + 1 = 0$. (Below, we shall interpret 0 and 1 as elements of $\mathbb{R}$, which will then lead to a different basis of $\mathcal{F}_V$.) The set of group homomorphisms from the product group $I_V = {\mathbb{Z}_2}^V$ into the unit circle of the complex plane forms a group with respect to pointwise multiplication, called the character group. The elements of that group, the characters of I_V, can be easily specified in our setting. For each $v \in V$, we first define the function

$$\xi_v : I_V = \{0, 1\}^V \to \{-1, 1\}, \qquad \xi_v(i) := (-1)^{X_v(i)} = \begin{cases} 1, & \text{if } X_v(i) = 0, \\ -1, & \text{if } X_v(i) = 1. \end{cases}$$

The characters of I_V are then given by the real-valued functions

$$e_A(i) := \prod_{v \in A} \xi_v(i), \quad A \subseteq V. \tag{2.196}$$

With $E(A, i) := |\{v \in A : X_v(i) = 1\}|$, we can rewrite (2.196) as

$$e_A(i) = (-1)^{E(A,i)}, \quad A \subseteq V. \tag{2.197}$$

These vectors are known as *Walsh vectors* [127, 253] and are used in various applications. In particular, they play an important role within Holland's genetic algorithms [110, 166]. If follows from general character theory (see [120, 158]) that the e_A form an orthogonal basis of the real vector space $\mathcal{F}_V$. We provide a more direct derivation of this result.

Proposition 2.18 (Walsh basis) *The vector e_A spans the one-dimensional vector space $\widetilde{\mathcal{F}}_A$ of pure A-interactions. In particular, the family e_A, $A \subseteq V$, of Walsh vectors forms an orthogonal basis of $\mathcal{F}_V$.*

Proof For $A = \emptyset$, we have $e_\emptyset = \sum_i 1 \cdot e_i = \mathbb{1}$ which spans $\widetilde{\mathcal{F}}_\emptyset$.

Now let $A \neq \emptyset$, and observe

$$f \in \widetilde{\mathcal{F}}_A \quad \Leftrightarrow \quad f \in \mathcal{F}_A \text{ and } P_B(f) = 0 \text{ for all } B \subsetneq A. \tag{2.198}$$

Below, we verify (2.198) by using

$$\sum_{i \in \{0,1\}^V} (-1)^{E(A,i)} = 0. \tag{2.199}$$

To see this, let v be an element of A and define

$$I_- := \{i : X_v(i) = 1\}, \qquad I_+ := \{i : X_v(i) = 0\}.$$

Obviously, $E(A, i) = E(A \setminus \{v\}, i)$ if $i \in I_+$. This implies (2.199):

$$\sum_i (-1)^{E(A,i)} = \sum_{i \in I_+} (-1)^{E(A\setminus\{v\},i)} - \sum_{i \in I_-} (-1)^{E(A\setminus\{v\},i)} = 0.$$

Now we verify (2.198). For $e_A \in \mathcal{F}_A$ we have

$$\begin{aligned} P_A(e_A)(i_A, i_{V\setminus A}) &= \frac{1}{2^{|V\setminus A|}} \sum_{i'_{V\setminus A} \in I_{V\setminus A}} e_A\big(i_A, i'_{V\setminus A}\big) \\ &= e_A(i_A, i_{V\setminus A}), \end{aligned}$$

which follows from the fact that $E(A, i)$ does not depend on $i_{V\setminus A}$. Furthermore, $P_B(e_A) = 0$ for $B \subsetneq A$:

$$\begin{aligned} &P_B(e_A)(i_B, i_{V\setminus B}) \\ &= \frac{1}{2^{|V\setminus B|}} \sum_{i'_{V\setminus B}} e_A(i_B, i'_{V\setminus B}) \\ &= \frac{1}{2^{|V\setminus B|}} \sum_{i'_{V\setminus B}} (-1)^{E(A,(i_B,i'_{V\setminus B}))} \\ &= \frac{1}{2^{|V\setminus B|}} \sum_{i'_{V\setminus B}} (-1)^{\{E(A,(i_B,i_{V\setminus B}))+E(A,(i_B,i'_{V\setminus B}))\}} \\ &= \frac{1}{2^{|V\setminus B|}} (-1)^{E(B,(i_B,i_{V\setminus B}))} \cdot 2^{|V\setminus B|} \cdot \underbrace{\sum_{i'_{V\setminus B}} (-1)^{E(A\setminus B,(i_B,i'_{V\setminus B}))}}_{\substack{=0, \\ \text{according to (2.199), since } A\setminus B \neq \emptyset}} \\ &= 0. \end{aligned}$$

□

Remark 2.3 (Characters of finite groups for the non-binary case) Assuming that each set I_v has the structure of the group $\mathbb{Z}_{n_v} = \mathbb{Z}/n_v\mathbb{Z}$, that is, $I_v = \{0, 1, \ldots, n_v - 1\}$, $n_v = |I_v|$, we denote its n_v characters by

$$\chi_{v,i_v} : I_v \to \mathbb{C}, \qquad j_v \mapsto \chi_{v,i_v}(j_v) := \exp\left(\mathrm{i}\,\frac{2\pi\, i_v\, j_v}{n_v}\right).$$

(Here, i denotes the imaginary number in $\mathbb{C}$.) We can write any function $f : I_V \to \mathbb{R}$ on the Abelian group $I_V = \bigtimes_{v\in V} \mathbb{Z}_{n_v}$ uniquely in the form

$$f = \sum_{i=(i_v)_{v\in V}\in I_V} \vartheta_i \prod_{v\in V} \chi_{v,i_v}$$

with complex coefficients ϑ_i satisfying $\vartheta_{-i} = \bar{\vartheta}_i$ (the bar denotes the complex conjugation). This allows us to decompose f into a unique sum of *l-body interactions*

$$\sum_{A\subseteq V,|A|=l} \sum_{\substack{i=(i_v)_{v\in V}\in I_V \\ i_v\neq 0 \text{ iff } v\in A}} \vartheta_i \prod_{v\in A} \chi_{v,i_v}.$$

In many applications, functions are decomposed as

$$f = \sum_{A\subseteq V} f_A, \quad \text{for suitable} \quad f_A \in \mathcal{F}_A, A \subseteq V, \tag{2.200}$$

where the family f_A, $A \subseteq V$, of functions is referred to as the *interaction potential*. However, it is not always natural to assume that the f_A are elements of the spaces $\widetilde{\mathcal{F}}_A$, $A \subseteq V$. One frequently used way of decomposing f assumes for each node $v \in V$ a distinguished state $o_v \in I_v$, the so-called *vacuum state*. Having this state, one requires that the family $(f_A)_{A\subseteq V}$ is normalized in the following sense:

$$\begin{aligned} &\text{(i)} \quad f_\emptyset = 0, \text{ and} \\ &\text{(ii)} \quad f_A(i) = 0 \text{ if there is a } v \in A \text{ satisfying } X_v(i) = o_v. \end{aligned} \tag{2.201}$$

With these requirements, the decomposition (2.200) is unique and can be obtained in terms of the Möbius inversion (Lemma 2.12, for details see [258], Theorem 3.3.3).

If we work with binary values 0 and 1, interpreted as elements in $\mathbb{R}$, it is natural to set $o_v = 0$ for all v. In that case, any function f can be uniquely decomposed as

$$f = \sum_{A\subseteq V} \vartheta_A \prod_{v\in A} X_v. \tag{2.202}$$

Obviously, $f_A = \vartheta_A \prod_{v\in A} X_v \in \mathcal{F}_A$ and the conditions (2.201) are satisfied. Note that, despite the similarity between the monomials $\prod_{v\in A} X_v$ and those defined by (2.196), the decomposition (2.202) is not an orthogonal one, and, for $A \neq \emptyset$, $\prod_{v\in A} X_v \notin \widetilde{\mathcal{F}}_A$. Clearly, by rewriting $\xi_v = (-1)^{X_v}$ as $1 - 2X_v$ (where the values of X_v are now interpreted as real numbers), we can transform one representation of a function into the other.

2.9.2 Hierarchical Models

We are now going to describe the structure of the interactions in a system. This structure sets constraints on the interaction potentials and the corresponding Gibbs

distributions. Although pairwise interactions are most commonly used in applications, which is associated with the edges of a graph, we have to incorporate, in particular, higher-order interactions by considering generalizations of graphs.

Definition 2.13 (Hypergraph) Let V be a finite set, and let $\mathfrak{S}$ be a set of subsets of V. We call the pair $(V, \mathfrak{S})$ a *hypergraph* and the elements of $\mathfrak{S}$ *hyperedges*. (Note that, at this point, we do not exclude the cases $\mathfrak{S} = \emptyset$ and $\mathfrak{S} = \{\emptyset\}$.) When V is fixed we usually refer to the hypergraph only by $\mathfrak{S}$. A hypergraph $\mathfrak{S}$ is a *simplicial complex* if it satisfies

$$A \in \mathfrak{S}, B \subseteq A \quad \Rightarrow \quad B \in \mathfrak{S}. \tag{2.203}$$

We denote the set of all inclusion maximal elements of a hypergraph $\mathfrak{S}$ by $\mathfrak{S}^{\max}$. A simplicial complex $\mathfrak{S}$ is determined by $\mathfrak{S}^{\max}$. Finally, we can extend any hypergraph $\mathfrak{S}$ to the simplicial complex $\overline{\mathfrak{S}}$ by including any subset A of V that is contained in a set $B \in \mathfrak{S}$.

For a hypergraph $\mathfrak{S}$, we consider the sum

$$\mathcal{F}_{\mathfrak{S}} := \sum_{A \in \mathfrak{S}} \mathcal{F}_A. \tag{2.204}$$

Note that for $\mathfrak{S} = \emptyset$, this space is zero-dimensional and consists of the constant function $f \equiv 0$. (This follows from the usual convention that the empty sum equals zero.) For $\mathfrak{S} = \{\emptyset\}$, we have the one-dimensional space of constant functions on I_V. In order to evaluate the dimension of $\mathcal{F}_{\mathfrak{S}}$ in the general case, we extend $\mathfrak{S}$ to the simplicial complex $\overline{\mathfrak{S}}$ and represent the vector space as the inner sum of the orthogonal sub-spaces $\widetilde{\mathcal{F}}_A$ (see Proposition 2.17),

$$\mathcal{F}_{\mathfrak{S}} = \bigoplus_{A \in \overline{\mathfrak{S}}} \widetilde{\mathcal{F}}_A, \tag{2.205}$$

which directly implies

$$\dim(\mathcal{F}_{\mathfrak{S}}) = \sum_{A \in \overline{\mathfrak{S}}} \prod_{v \in A} \big(|I_v| - 1\big). \tag{2.206}$$

For the particular case of binary nodes we obtain the number $|\overline{\mathfrak{S}}|$ of hyperedges of the simplicial complex $\overline{\mathfrak{S}}$ as the dimension of $\mathcal{F}_{\mathfrak{S}}$.

Definition 2.14 (Hierarchical model) For a hypergraph $\mathfrak{S}$, we define a *hierarchical model* as the exponential family generated by the vector space $\mathcal{F}_{\mathfrak{S}}$ (see Definition 2.10):

$$\mathcal{E}_{\mathfrak{S}} := \mathcal{E}(\mathcal{F}_{\mathfrak{S}}) = \left\{ \sum_{i \in I_V} \frac{e^{f(i)}}{\sum_{i' \in I_V} e^{f(i')}} \delta^i : f \in \mathcal{F}_{\mathfrak{S}} \right\}. \tag{2.207}$$

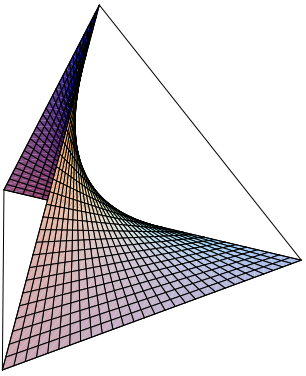

Fig. 2.8 Exponential family of product distributions

Note that for $\mathfrak{S} = \emptyset$ and $\mathfrak{S} = \{\emptyset\}$, the hierarchical model $\mathcal{E}_{\mathfrak{S}}$ consists of only one element, the uniform distribution, and is therefore zero-dimensional. In order to remove this ambiguity, we always assume $\mathfrak{S} \neq \emptyset$ in the context of hierarchical models (but still allow $\mathfrak{S} = \{\emptyset\}$). With this assumption, the dimension of $\mathcal{E}_{\mathfrak{S}}$ is one less than the dimension of $\mathcal{F}_{\mathfrak{S}}$:

$$\dim(\mathcal{E}_{\mathfrak{S}}) = \sum_{\emptyset \neq A \in \overline{\mathfrak{S}}} \prod_{v \in A} \big(|I_v| - 1\big). \tag{2.208}$$

Example 2.6

(1) (Independence model) A hierarchical model is particularly simple if the hypergraph is a partition $\mathfrak{S} = \{A_1, \ldots, A_n\}$ of V. The corresponding simplicial complex is then given by

$$\overline{\mathfrak{S}} = \bigcup_{k=1}^{n} 2^{A_k}. \tag{2.209}$$

The hierarchical model $\mathcal{E}_{\mathfrak{S}}$ consists of all strictly positive distributions that factorize according to the partition $\mathfrak{S}$, that is,

$$p(i) = \prod_{k=1}^{n} p(i_{A_k}). \tag{2.210}$$

We refer to this hierarchical model as an *independence model*. In the special case of binary units, it has the dimension

$$\dim(\mathcal{E}_{\mathfrak{S}}) = \sum_{k=1}^{n} \big(2^{|A_k|} - 1\big). \tag{2.211}$$

(2) (Interaction model of order k) In this example, we want to model interactions up to order k. For that purpose, it is sufficient to consider the hypergraph consisting

of the subsets of cardinality k:

$$\mathfrak{S}_k := \binom{V}{k}. \tag{2.212}$$

The corresponding simplicial complex is given by

$$\overline{\mathfrak{S}}_k := \{A \subseteq V : 0 \leq |A| \leq k\} = \bigcup_{l=0}^{k} \binom{V}{l}. \tag{2.213}$$

This defines the hierarchical model

$$\mathcal{E}^{(k)} := \mathcal{E}_{\mathfrak{S}_k} \tag{2.214}$$

with dimension

$$\dim\left(\mathcal{E}^{(k)}\right) = \sum_{l=1}^{k} \binom{N}{l}, \tag{2.215}$$

and we obviously have

$$\mathcal{E}^{(1)} \subseteq \mathcal{E}^{(2)} \subseteq \cdots \subseteq \mathcal{E}^{(N)}. \tag{2.216}$$

The information geometry of this hierarchy has been developed in more detail by Amari [10]. The exponential family $\mathcal{E}^{(1)}$ consists of the product distributions (see Fig. 2.8). The extension to $\mathcal{E}^{(2)}$ incorporates pairwise interactions, and $\mathcal{E}^{(N)}$ is nothing but the whole set of strictly positive distributions. Within the field of artificial neural networks, $\mathcal{E}^{(2)}$ is known as a *Boltzmann machine* [15].

In Sect. 6.1, we shall study the relative entropy distance of a distribution p from a hierarchical model $\mathcal{E}_{\mathfrak{S}}$, that is, $\inf_{q \in \mathcal{E}_{\mathfrak{S}}} D(p \,\|\, q)$. This distance can be evaluated with the corresponding maximum entropy estimate $\hat{p}$ (see Sect. 2.8.3). More precisely, consider the set of distributions q that have the same A-marginals as p for all $A \in \mathfrak{S}$:

$$\mathcal{M} := \mathcal{M}(p, \mathfrak{S}) := \Bigg\{ q \in \mathcal{P}(I_V) : \sum_{i_{V \setminus A}} q(i_A, i_{V \setminus A}) = \sum_{i_{V \setminus A}} p(i_A, i_{V \setminus A}) \\ \text{for all } A \in \mathfrak{S} \text{ and all } i_A \in I_A \Bigg\}.$$

Obviously, $\mathcal{M}$ is a closed and convex subset of $\mathcal{P}(I_V)$. Therefore, the restriction of the continuous and strictly concave Shannon entropy $H(q) = -\sum_i q(i) \log q(i)$ to $\mathcal{M}$ attains its maximal value in a unique distribution $\hat{p} \in \mathcal{M}$. We refer to this distribution $\hat{p}$ as the *maximum entropy estimate (of p) with respect to* $\mathfrak{S}$. The following lemma provides a sufficient condition for a distribution to be the maximum entropy estimate with respect to $\mathfrak{S}$.

Lemma 2.13 (Maximum entropy estimation for hierarchical models) *Let $\mathfrak{S}$ be a non-empty hypergraph and let $p \in \mathcal{P}(I_V)$. If a distribution $\hat{p}$ satisfies the following two conditions then it is the maximum entropy estimate of p with respect to $\mathfrak{S}$:*

(1) *There exist functions $\phi_A \in \mathcal{F}_A$, $A \in \mathfrak{S}^{\max}$, satisfying*

$$\hat{p}(i_V) = \prod_{A \in \mathfrak{S}^{\max}} \phi_A(i_V). \tag{2.217}$$

(2) *For all $A \in \mathfrak{S}^{\max}$, the A-marginal of $\hat{p}$ coincides with the A-marginal of p, that is,*

$$\sum_{i_{V \setminus A}} \hat{p}(i_A, i_{V \setminus A}) = \sum_{i_{V \setminus A}} p(i_A, i_{V \setminus A}), \quad \text{for all } i_A \in I_A. \tag{2.218}$$

Proof We prove that $\hat{p}$ is in the closure of $\mathcal{E}_{\mathfrak{S}}$. As $\hat{p}$ is also in $\mathcal{M}(p, \mathfrak{S})$, the statement follows from Theorem 2.8.

$$\hat{p}(i_V) = \prod_{A \in \mathfrak{S}^{\max}} \phi_A(i_V) = \lim_{\varepsilon \to 0} \frac{\prod_{A \in \mathfrak{S}^{\max}} (\phi_A(i_V) + \varepsilon)}{\sum_{j_V} \prod_{A \in \mathfrak{S}^{\max}} (\phi_A(j_V) + \varepsilon)} = \lim_{\varepsilon \to 0} p^{(\varepsilon)}(i_V),$$

where obviously $p^{(\varepsilon)} \in \mathcal{E}_{\mathfrak{S}}$. □

For a strictly positive distribution p, the conditions (2.217) and (2.218) of Lemma 2.13 are necessary and sufficient for a distribution $\hat{p}$ to be the maximum entropy estimate of p with respect to $\mathfrak{S}$. This directly follows from Theorem 2.8. In general, however, Lemma 2.13 provides only a sufficient condition, but not a necessary one, as follows from the following observation. If the maximum entropy estimate $\hat{p}$ lies in the boundary of the hierarchical model $\mathcal{E}_{\mathfrak{S}}$, it does not necessarily have to admit the product structure (2.217).

In Sect. 6.1, we shall use Lemma 2.13 for the evaluation of a number of examples.

2.9.3 Graphical Models

Graphs provide a compact way of representing a particular kind of hierarchical models, the so-called graphical models [161]. In this section we present the Hammersley–Clifford theorem which is central within the theory of graphical models.

We consider an undirected graph $G = (V, E)$, with *node set* V and *edge set* $E \subseteq \binom{V}{2}$. If $\{v, w\} \in E$ then we write $v \sim w$ and call v and w *neighbors*. Given a node v the set $\{w \in V : v \sim w\}$ of neighbors is called the *boundary* of v and denoted by $\mathrm{bd}(v)$. For an arbitrary subset A of V, we define the boundary $\mathrm{bd}(A) := \cup_{v \in A} (\mathrm{bd}(v) \setminus A)$ of A and its *closure* $\mathrm{cl}(A) := A \cup \mathrm{bd}(A)$. Note that according to this definition, A and $\mathrm{bd}(A)$ are always disjoint.

A *path* in V is a sequence $\gamma = (v_1, \dots, v_n)$ satisfying $v_i \sim v_{i+1}$ for all $i = 1, \dots, n-1$. Let A, B, and S be three disjoint subsets of V. We say that S *separates* A from B if for every path $\gamma = (v_1, \dots, v_n)$ with $v_1 \in A$ and $v_n \in B$ there is a $v_i \in S$. Note that $\mathrm{bd}(A)$ separates A from $V \setminus \mathrm{cl}(A)$.

A subset C of V is called a *clique* (of G), if for all $v, w \in C$, $v \neq w$, it holds that $v \sim w$. The set of cliques is a simplicial complex in the sense of Definition 2.13, which we denote by $\mathfrak{C}(G)$. A hierarchical model that only includes interactions of nodes within cliques of a graph has very special properties.

Definition 2.15 Let G be a graph, and let $\mathfrak{C}(G)$ be the simplicial complex consisting of the cliques of G. Then the hierarchical model $\mathcal{E}(G) := \mathcal{E}_{\mathfrak{C}(G)}$, as defined by (2.207), is called a *graphical model*.

A graph encodes natural conditional independence properties, so-called *Markov properties*, which are satisfied by all distributions of the corresponding graphical model. In Definition 2.16 below, we shall use the notation $X_A \perp\!\!\!\perp X_B \,|\, X_C$ for the conditional independence statement "X_A and X_B are stochastically independent given X_C." This clearly depends on the underlying distribution which is not mentioned explicitly in this notation. Formally, this conditional independence with respect to a distribution $p \in \mathcal{P}(I_V)$ is expressed by

$$p(i_A, i_B \,|\, i_C) = p(i_A \,|\, i_C)\, p(i_B \,|\, i_C) \quad \text{whenever } p(i_C) > 0, \tag{2.219}$$

where we apply the definition (2.184). If we want to use marginals only, we can rewrite this condition as

$$p(i_A, i_B, i_C) = \frac{p(i_A, i_C)\, p(i_B, i_C)}{p(i_C)} \quad \text{whenever } p(i_C) > 0. \tag{2.220}$$

This is equivalent to the existence of two functions $f \in \mathcal{F}_{A \cup C}$ and $g \in \mathcal{F}_{B \cup C}$ such that

$$p(i_A, i_B, i_C) = f(i_A, i_C)\, g(i_B, i_C), \tag{2.221}$$

where we can ignore "whenever $p(i_C) > 0$" in (2.219) and (2.220).

Definition 2.16 Let G be a graph with node set V, and let p be a distribution on I_V. Then we say that p satisfies the

(G) *global Markov property*, with respect to G, if

$$A, B, S \subseteq V \text{ disjoint, } S \text{ separates } A \text{ from } B \quad \Rightarrow \quad X_A \perp\!\!\!\perp X_B \,|\, X_S;$$

(L) *local Markov property*, with respect to G, if

$$v \in V \quad \Rightarrow \quad X_v \perp\!\!\!\perp X_{V \setminus v} \,|\, X_{\mathrm{bd}(v)};$$

(P) *pairwise Markov property*, with respect to G, if

$$v, w \in V, v \nsim w \quad \Rightarrow \quad X_v \perp\!\!\!\perp X_w \,|\, X_{V \setminus \{v,w\}}.$$

Proposition 2.19 *Let G be a graph with node set V, and let p be a distribution on I_V. Then the following implications hold*:

$$\text{(G)} \quad \Rightarrow \quad \text{(L)} \quad \Rightarrow \quad \text{(P)}.$$

Proof (G) $\Rightarrow$ (L) This is a consequence of the fact that $\mathrm{bd}(v)$ separates v from $V \setminus \mathrm{cl}(v)$, as noted above.

(L) $\Rightarrow$ (P) Assuming that $v, w \in V$ are not neighbors, one has $w \in V \setminus \mathrm{cl}(w)$ and therefore

$$\mathrm{bd}(v) \cup \big(\big(V \setminus \mathrm{cl}(v)\big) \setminus \{w\}\big) = V \setminus \{v, w\}. \tag{2.222}$$

From the local Markov property (L), we know that

$$X_v \perp\!\!\!\perp X_{V \setminus \mathrm{cl}(v)} \,|\, X_{\mathrm{bd}(v)}. \tag{2.223}$$

We now use the following general rule for conditional independence statements:

$$X_A \perp\!\!\!\perp X_B \,|\, X_S, C \subseteq B \quad \Rightarrow \quad X_A \perp\!\!\!\perp X_B \,|\, X_{S \cup C}.$$

With (2.223) and $(V \setminus \mathrm{cl}(v)) \setminus \{w\} \subseteq V \setminus \mathrm{cl}(v)$, this rule implies

$$X_v \perp\!\!\!\perp X_{V \setminus \mathrm{cl}(v)} \,|\, X_{\mathrm{bd}(v) \cup [(V \setminus \mathrm{cl}(v)) \setminus \{w\}]}.$$

Because of (2.222), this is equivalent to

$$X_v \perp\!\!\!\perp X_{V \setminus cl(v)} \,|\, X_{V \setminus \{v,w\}}.$$

With $w \in V \setminus \mathrm{cl}(v)$ we finally obtain

$$X_v \perp\!\!\!\perp X_w \,|\, X_{V \setminus \{v,w\}},$$

which proves the pairwise Markov property. □

Now we provide a criterion for a distribution p that is sufficient for the global Markov property. We say that *p factorizes according to G* or satisfies the

(F) *factorization property*, with respect to G, if there exist functions $f_C \in \mathcal{F}_C$, C a clique, such that

$$p(i) = \prod_{C \text{ clique}} f_C(i). \tag{2.224}$$

Proposition 2.20 *Let G be a graph with node set V, and let p be a distribution on I_V. Then the factorization property implies the global Markov property, that is*,

$$\text{(F)} \quad \Rightarrow \quad \text{(G)}.$$

Proof We assume that S separates A from B and have to show

$$X_A \perp\!\!\!\perp X_B \,|\, X_S.$$

The complement $V \setminus S$ of S in V can be written as the union of its connected components V_i, $i = 1, \ldots, n$:

$$V \setminus S = V_1 \cup \cdots \cup V_n.$$

We define

$$\widetilde{A} := \bigcup_{\substack{i \in \{1,\ldots,n\} \\ V_i \cap A \neq \emptyset}} V_i, \qquad \widetilde{B} := \bigcup_{\substack{i \in \{1,\ldots,n\} \\ V_i \cap A = \emptyset}} V_i.$$

Obviously, $A \subseteq \widetilde{A}$, and $B \subseteq \widetilde{B}$ (S separates A from B). Furthermore, a clique C is contained in $\widetilde{A} \cup S$ or $\widetilde{B} \cup S$. Therefore, we can split the factorization of p as follows:

$$\begin{aligned} p(i) &= \prod_{C \text{ clique}} f_C(i) \\ &= \prod_{\substack{C \text{ clique} \\ C \subseteq \widetilde{A} \cup S}} f_C(i) \cdot \prod_{\substack{C \text{ clique} \\ C \not\subseteq \widetilde{A} \cup S}} f_C(i) \\ &=: g(i_{\widetilde{A}}, i_S) \cdot h(i_{\widetilde{B}}, i_S). \end{aligned}$$

With (2.221), this proves $X_{\widetilde{A}} \perp\!\!\!\perp X_{\widetilde{B}} \,|\, X_S$. Since $A \subseteq \widetilde{A}$ and $B \subseteq \widetilde{B}$, we finally obtain $X_A \perp\!\!\!\perp X_B \,|\, X_S$. □

Generally, there is no equivalence of the above Markov conditions. On the other hand, if we assume strict positivity of a distribution p, that is, $p \in \mathcal{P}_+(I_V)$, then we have equivalence. This is the content of the Hammersley–Clifford theorem of graphical model theory.

Theorem 2.9 (Hammersley–Clifford theorem) *Let G be a graph with node set V, and let p be a strictly positive distribution on I_V. Then p satisfies the pairwise Markov property if and only if it factorizes according to G.*

Proof We only have to prove that (P) implies (F), since the opposite implication directly follows from Propositions 2.19 and 2.20.

We assume that p satisfies the pairwise Markov property (P). As p is strictly positive, we can consider $\log p(i)$. We choose one configuration $i^* \in I_V$ and define

$$H_A(i) := \log p\big(i_A, i^*_{V \setminus A}\big), \quad A \subseteq V.$$

Here $(i_A, i^*_{V\setminus A})$ coincides with i on A and with i^* on $V \setminus A$. Clearly, H_A does not depend on $i_{V\setminus A}$, and therefore $H_A \in \mathcal{F}_A$. We define

$$\phi_A(i) := \sum_{B\subseteq A} (-1)^{|A\setminus B|} H_B(i).$$

Also, the ϕ_A depend only on A. With the Möbius inversion formula (Lemma 2.12), we obtain

$$\log p(i) = H_V(i) = \sum_{A\subseteq V} \phi_A(i). \tag{2.225}$$

In what follows we use the pairwise Markov property of p in order to prove that in the representation (2.225), $\phi_A = 0$ whenever A is not a clique. This obviously implies that p factorizes according to G and completes the proof.

Assume A is not a clique. Then there exist $v, w \in A$, $v \neq w$, $v \nsim w$. Consider $C := A \setminus \{v, w\}$. Then

$$\begin{aligned}
\phi_A(i) &= \sum_{B\subseteq A} (-1)^{|A\setminus B|} H_B(i) \\
&= \sum_{\substack{B\subseteq A \\ v,w\notin B}} (-1)^{|A\setminus B|} H_B(i) + \sum_{\substack{B\subseteq A \\ v\in B,\, w\notin B}} (-1)^{|A\setminus B|} H_B(i) \\
&\quad + \sum_{\substack{B\subseteq A \\ v\notin B,\, w\in B}} (-1)^{|A\setminus B|} H_B(i) + \sum_{\substack{B\subseteq A \\ v,w\in B}} (-1)^{|A\setminus B|} H_B(i) \\
&= \sum_{B\subseteq C} (-1)^{|C\setminus B|+2} H_{B\cup\{v,w\}}(i) + \sum_{B\subseteq C} (-1)^{|C\setminus B|+1} H_{B\cup\{v\}}(i) \\
&\quad + \sum_{B\subseteq C} (-1)^{|C\setminus B|+1} H_{B\cup\{w\}}(i) + \sum_{B\subseteq C} (-1)^{|C\setminus B|} H_B(i) \\
&= \sum_{B\subseteq C} (-1)^{|C\setminus B|} \big(H_B(i) - H_{B\cup\{v\}}(i) - H_{B\cup\{w\}}(i) + H_{B\cup\{v,w\}}(i)\big).
\end{aligned} \tag{2.226}$$

We now set $D := V \setminus \{v, w\}$ and use the pairwise Markov property (P) in order to show that (2.226) vanishes:

$$\begin{aligned}
&H_{B\cup\{v,w\}}(i) - H_{B\cup\{v\}}(i) \\
&\quad = \log \frac{p(i_B, i_v, i_w, i^*_{D\setminus B})}{p(i_B, i_v, i^*_w, i^*_{D\setminus B})} \\
&\quad = \log \frac{p(i_B, i_w, i^*_{D\setminus B}) \cdot p(i_v \,|\, i_B, i_w, i^*_{D\setminus B})}{p(i_B, i^*_w, i^*_{D\setminus B}) \cdot p(i_v \,|\, i_B, i^*_w, i^*_{D\setminus B})}
\end{aligned}$$

$$
\begin{aligned}
&= \log \frac{p(i_B, i_w, i^*_{D\setminus B}) \cdot p(i_v \,|\, i_B, i^*_{D\setminus B})}{p(i_B, i^*_w, i^*_{D\setminus B}) \cdot p(i_v \,|\, i_B, i^*_{D\setminus B})} \\
&= \log \frac{p(i_B, i_w, i^*_{D\setminus B}) \cdot p(i^*_v \,|\, i_B, i^*_{D\setminus B})}{p(i_B, i^*_w, i^*_{D\setminus B}) \cdot p(i^*_v \,|\, i_B, i^*_{D\setminus B})} \\
&= \log \frac{p(i_B, i_w, i^*_{D\setminus B}) \cdot p(i^*_v \,|\, i_B, i_w, i^*_{D\setminus B})}{p(i_B, i^*_w, i^*_{D\setminus B}) \cdot p(i^*_v \,|\, i_B, i^*_w i^*_{D\setminus B})} \\
&= \log \frac{p(i_B, i^*_v, i_w, i^*_{D\setminus B})}{p(i_B, i^*_v, i^*_w, i^*_{D\setminus B})} \\
&= H_{B\cup\{w\}}(i) - H_B(i).
\end{aligned}
$$

This implies $\phi_A(i) = 0$ and, with the representation (2.225), we conclude that p factorizes according to G. □

The Hammersley–Clifford theorem implies that for strictly positive distributions, all Markov properties of Definition 2.16 are equivalent. The set of strictly positive distributions that satisfy one of these properties, and therefore all of them, is given by the graphical model $\mathcal{E}(G)$. Its closure $\overline{\mathcal{E}}(G)$ is sometimes referred to as the *extended graphical model*. We want to summarize the results of this section by an inclusion diagram. In order to do so, for each property (prop) $\in \{(\mathrm{F}), (\mathrm{G}), (\mathrm{L}), (\mathrm{P})\}$, we define

$$
\mathcal{M}^{(\mathrm{prop})} := \big\{ p \in \mathcal{P}(I_V) : p \text{ satisfies (prop)} \big\},
$$

and

$$
\mathcal{M}^{(\mathrm{prop})}_+ := \mathcal{M}^{(\mathrm{prop})} \cap \mathcal{P}_+(I_V).
$$

Clearly, the set of strictly positive distributions that factorize according to G, that is, $\mathcal{M}^{(\mathrm{F})}_+$, coincides with the graphical model $\mathcal{E}(G)$. One can easily verify that $\mathcal{M}^{(\mathrm{G})}$, $\mathcal{M}^{(\mathrm{L})}$, and $\mathcal{M}^{(\mathrm{P})}$ are closed subsets of $\mathcal{P}(I_V)$ that contain the extended graphical model $\overline{\mathcal{E}}(G)$ as a subset. However, the set $\mathcal{M}^{(\mathrm{F})}$ is not necessarily closed: limits of factorized distributions do not have to be factorized. Furthermore, it is contained in $\overline{\mathcal{E}}(G)$ (see the proof of Lemma 2.13).

These considerations, Propositions 2.19 and 2.20, and the Hammersley–Clifford theorem (Theorem 2.9) can be summarized in terms of the following diagram.

$$
\begin{array}{ccccccccc}
\mathcal{M}^{(\mathrm{F})}_+ & = & \mathcal{E}(G) & = & \mathcal{M}^{(\mathrm{G})}_+ & = & \mathcal{M}^{(\mathrm{L})}_+ & = & \mathcal{M}^{(\mathrm{P})}_+ \\
\cap\mathrel{|} & & \cap\mathrel{|} & & \cap\mathrel{|} & & \cap\mathrel{|} & & \cap\mathrel{|} \\
\mathcal{M}^{(\mathrm{F})} & \subseteq & \overline{\mathcal{E}}(G) & \subseteq & \mathcal{M}^{(\mathrm{G})} & \subseteq & \mathcal{M}^{(\mathrm{L})} & \subseteq & \mathcal{M}^{(\mathrm{P})}
\end{array}
\tag{2.227}
$$

The upper row of equalities in this diagram summarizes the content of the Hammersley–Clifford theorem (Theorem 2.9). The lower row of inclusions in this diagram follows from Propositions 2.19 and 2.20. Each of the horizontal inclusions

can be strict in the sense that there exists a graph G for which the inclusion is strict. Corresponding examples are given in Lauritzen's monograph [161], referring to work by Moussouris [191], Matúš [174], and Matúš and Studený [181].

Remark 2.4

(1) Conditional independence statements give rise to a natural class of models, referred to as *conditional independence models* (see the monograph of Studený [241]). This class includes graphical models as prime examples. By the Hammersley–Clifford theorem, on the other hand, graphical models are also special within the class of hierarchical models. A surprising and important result of Matúš highlights the uniqueness of graphical models within both classes ([178], Theorem 1). If a hierarchical model is specified in terms of conditional independence statements, then it is already graphical. This means that one cannot specify any other hierarchical model in terms of conditional independence statements. Furthermore, the result of Matúš also provides a new proof of the Hammersley–Clifford theorem ([178], Corollary 1). It is not based on the Möbius inversion of the classical proof which we have used in our presentation.

(2) Graphical model theory has been further developed by Geiger, Meek, and Sturmfels using tools from algebraic statistics [103]. In their work, they develop a refined geometric understanding of the results presented in this section. This understanding is not restricted to graphical models but also applies to general hierarchical models. Let us briefly sketch their perspective. All conditional independence statements that appear in Definition 2.16 can be reformulated in terms of polynomial equations. Each Markov property can then be associated with a corresponding set of polynomial equations, which generates an ideal $\mathscr{I}$ in the polynomial ring $\mathbb{R}[x_1, \ldots, x_{|I_V|}]$. (The indeterminates are the coordinates $p(i)$, $i \in I_V$, of the distributions in $\mathcal{P}(I_V)$.) This leads to the ideals $\mathscr{I}^{(\mathrm{G})}$, $\mathscr{I}^{(\mathrm{L})}$, and $\mathscr{I}^{(\mathrm{P})}$ that correspond to the individual Markov properties, and we obviously have

$$\mathscr{I}^{(\mathrm{P})} \subseteq \mathscr{I}^{(\mathrm{L})} \subseteq \mathscr{I}^{(\mathrm{G})}. \tag{2.228}$$

Each of these ideals fully characterizes the graphical model as its associated variety in $\mathcal{P}_+(I_V)$, which follows from the Hammersley–Clifford theorem. The respective varieties $\mathcal{M}^{(\mathrm{G})}$, $\mathcal{M}^{(\mathrm{L})}$, and $\mathcal{M}^{(\mathrm{P})}$ in the *full* simplex $\mathcal{P}(I_V)$ differ in general and contain the extended graphical model $\overline{\mathcal{E}}(G)$ as a proper subset. On the other hand, one can fully specify $\overline{\mathcal{E}}(G)$ in terms of polynomial equations using Theorems 2.5 and 2.7. Denoting the corresponding ideal by $\mathscr{I}(G)$, we can extend the above inclusion chain (2.228) by

$$\mathscr{I}^{(\mathrm{G})} \subseteq \mathscr{I}(G). \tag{2.229}$$

Stated differently, the ideal $\mathscr{I}(G)$ encodes all Markov properties in terms of elements of an appropriate ideal basis. Let us now assume that we are given an arbitrary hierarchical model $\mathcal{E}_{\mathfrak{S}}$ with respect to a hypergraph $\mathfrak{S}$, and let us denote by $\mathscr{I}_{\mathfrak{S}}$ the ideal that is generated by the corresponding Eqs. (2.139) or,

equivalently, (2.149). Geiger, Meek, and Sturmfels interpret the elements of a finite ideal basis of $\mathscr{I}_{\mathfrak{S}}$ as generalized conditional independence statements and prove a version of the Hammersley–Clifford theorem for hierarchical models. Note, however, that the above mentioned result of Matúš implies that there is a correspondence between ideal basis elements and actual conditional independence statements only for graphical models.