UNIVERSITAT DE BARCELONA

A CHARACTERIZATION OF MONOTONE AND REGULAR DIVERGENCES

by

J.M. Corcuera and *F. Giummolè*

AMS 1980 Subjects Classifications: 62F10, 62B10, 62A99

Mathematics Preprint Series No. 213
July 1996

A characterization of monotone and regular divergences

J.M. Corcuera
Departament d'Estadística
Universitat de Barcelona
Gran Vía de les Corts Catalanes, 585
08007 Barcelona, Spain
corcuera@cerber.mat.ub.es

F. Giummolè
Dipartimento di Scienze Statistiche
Università degli Studi di Padova
Via S. Francesco, 33
35121 Padova, Italy
giummole@hal.stat.unipd.it

Abstract

In this paper we characterize monotone and regular divergences, which include f-divergences as a particular case, by giving their Taylor expansion up to fourth order. We extend a previous result obtained by Čencov, using the invariant properties of Amari's α-connections.

AMS 1980 subjects classifications: 62F10, 62B10, 62A99.

Keywords and phrases: differential geometry, divergence, embedding invariance, Markov embedding, string, α-connection.

1 Introduction

The necessity of measuring how different two populations are appears in many statistical problems. A wide class of indices or divergences has been used with such a finality (for a comprehensive exposition see Burbea (1983)). We are not able to give an universal rule for the choice of a divergence in each practical case. Anyway, we can investigate the general properties that an index of discrepancy should possess in order to describe a meaningful dissimilarity between populations. For instance, suppose to assemble the individuals of two finite populations in classes $A_1, \ldots, A_m$. Let $D(P_1, P_2)$ be a convenient function of the proportions $P_i = (P_i(A_1), \ldots, P_i(A_m))$, $i = 1, 2$, of individuals belonging to the different groups in the two populations. We can now decide to join several classes, obtaining $B_1, \ldots, B_l$, $l < m$. If $\tilde{P}_i = (P_i(B_1), \ldots, P_i(B_l))$, it is natural to demand that $D(\tilde{P}_1, \tilde{P}_2) \leq D(P_1, P_2)$, since the new classification brings less information than the previous one. It is also necessary to ask that the introduction of new artificial classes does not cause any change in the divergence. Formally, if $Q_i = (q_{11}P_i(A_1), \ldots, q_{1r_1}P_i(A_1), \ldots, q_{m1}P_i(A_m), \ldots, q_{mr_m}P_i(A_m))$, $i = 1, 2$, where $\sum_{k=1}^{r_j} q_{jk} = 1$, $j = 1, \ldots, m$, then $D(P_1, P_2) = D(Q_1, Q_2)$.

Divergences satisfying these two properties have already been studied by Čencov (1972). He gives their Taylor expansion up to second order, by means of the invariance of the Fisher metric. In this paper we extend Čencov's result to fourth order, using the invariance properties of the α-connections. Following Campbell (1986), we do not use the language of the categories.

An additional property allows us to extend a divergence to the case of when the individuals are classified in an infinite number of groups. This property expresses a sort of continuity of the divergence, when we let the number of classes tend to infinity.

2 Some basic definitions

In this section, we introduce operators representing an index of discrepancy between probability measures defined on the same measurable space.

In the sequel, we indicate with $(\mathcal{X}, a)$ a measurable space. $a_\alpha \subseteq a$ is a finite subσ-field of a and P_α is the restriction of P, defined on $(\mathcal{X}, a)$, to a_α.

Definition 2.1 *A* divergence $D(P,Q)$ *is a real-valued function whose arguments are two probability measures defined on the same measurable space.*

Definition 2.2 *Let* $(\mathcal{X}_1, a_1)$ *and* $(\mathcal{X}_2, a_2)$ *be two measurable spaces. We say that* $K : \mathcal{X}_1 \times a_2 \to [0,1]$ *is a* Markov kernel, $K \in Stoch\{(\mathcal{X}_1, a_1), (\mathcal{X}_2, a_2)\}$, *if it satisfies the following properties:*

1. $\forall A_2 \in a_2,\ K(\cdot, A_2)$ *is a measurable map;*
2. $\forall x_1 \in \mathcal{X}_1,\ K(x_1, \cdot)$ *is a probability on* $(\mathcal{X}_2, a_2)$*;*

If P is a probability measure on $(\mathcal{X}_1, a_1)$, then K induces a probability measure on $(\mathcal{X}_2, a_2)$, KP, defined by

$$KP(A_2) = \int_{\mathcal{X}_1} K(x, A_2) P(dx), \quad \forall A_2 \in a_2.$$

Let $D(\cdot,\cdot)$ be a divergence and $(\mathcal{X}_1, a_1)$ and $(\mathcal{X}_2, a_2)$ be two measurable spaces.

Definition 2.3 $D(P,Q)$ *is said to be* monotone *with respect to Markov kernels if*

$$-\infty < D(KP, KQ) \leq D(P,Q) \leq +\infty, \tag{2.1}$$

for every P, Q *probability measures on* $(\mathcal{X}_1, a_1)$, *and for every Markov kernel* $K \in Stoch\{(\mathcal{X}_1, a_1), (\mathcal{X}_2, a_2)\}$.

As observed in the introduction, (2.1) is a natural property to require, since a transformation through a Markov kernel will, in general, cause a loss of information that is well explained by a decreasing of the divergence.

Monotonicity of a divergence function implies its invariance under a particular class of Markov kernels. Let $\mathcal{P}$ be a family of probabilities on $(\mathcal{X}_1, a_1)$.

Definition 2.4 $K \in Stoch\{(\chi_1, a_1), (\chi_2, a_2)\}$ *is said to be* Blackwell sufficient *(B-sufficient) with respect to* $\mathcal{P}$ *if there exists* $N \in Stoch\{(\chi_2, a_2), (\chi_1, a_1)\}$ *such that* $N(KP) = P$, $\forall P \in \mathcal{P}$. *We say that* K *is B-sufficient if* $\mathcal{P}$ *is the family of all probability measures on* (χ_1, a_1).

Proposition 2.1 *If* D *is a monotone function with respect to Markov kernels, then, for every B-sufficient* K,

$$D(P,Q) = D(KP, KQ), \quad \forall P, Q. \tag{2.2}$$

Proof:

$$D(P,Q) = D(N(KP), N(KQ)) \leq D(KP, KQ),$$

that, together with the monotonicity, gives (2.2). ■

This is also natural since a B-sufficient Markov kernel does not cause any loss of information.

Corollary 2.1 *The value of* $D(P,P)$ *is independent of* P *and it is a minimum value of the function* D :

$$D(P,P) = D_0 \leq D(Q,R), \quad \forall P, Q, R.$$

Proof: Given a probability measure P, it always exists a Markov kernel K, taking every probability measure on P: $K(x, \cdot) = P(\cdot)$, $\forall x \in \chi$. Then,

$$D(Q,R) \geq D(KQ, KR) = D(P,P), \quad \forall Q, R,$$

proving that $D(P,P)$ is a minimum value for D. Now, for every probability measure P', K is B-sufficient with respect to the family $\mathcal{P} = \{P, P'\}$, since $N(x, \cdot) = P'(\cdot)$, $\forall x \in \chi$, transforms P in P'. By (2.2),

$$D(P,P) = D(P', P') = D_0, \quad \forall P, P'.$$

■

Definition 2.5 *$D(P, Q)$ is said to be* regular *if*

$$D(P,Q) = \lim_{\alpha} D(P_\alpha, Q_\alpha), \tag{2.3}$$

for every P and Q probability measures on $(\mathcal{X}, a)$, where the limit is taken over the filter of all finite subσ-fields a_α of a, that is, over any increasing sequence $\{a_n\}$ such that $\sigma(\bigcup_n a_n) = a$.

Remark. Since the restriction of a probability measure to a subσ-field is a particular case of Markov kernel, for monotone divergences the limit in (2.3) is a supremum.

The regularity condition enables us to extend to the general case a divergence originally defined on probability measures over finite σ-fields.

3 The multinomial case

In the present section we consider equivalent probability measures defined on the measurable space $(\mathcal{X}, a_m)$, where a_m is a finite subσ-field of a generated by the m atoms $A_1, \ldots, A_m$. Every probability measure P on $(\mathcal{X}, a)$ induces a probability measure on $(\mathcal{X}, a_m)$, defined by m values, $x_1 \ldots, x_m$, with $x_i = P(A_i) > 0$ and $\sum_{i=1}^m x_i = 1$. Thus, every P corresponds to a point of the simplex

$$S_{m-1} = \left\{ x \in \mathbb{R}^m : \sum_{i=1}^m x_i = 1,\ x_i > 0,\ i = 1, \ldots, m \right\}.$$

S_{m-1} can be regarded as a surface in the differentiable manifold $\mathbb{R}^m$.

There is a tangent space M_x, with base $\left\{ X_i = \dfrac{\partial}{\partial x_i},\ i = 1, \ldots, m \right\}$, associated to every point $x \in \mathbb{R}^m$. If $x \in S_{m-1}$, the derivative of a function $h(x_1, \ldots, x_m)$ along a curve $x_i = \psi_i(t)$, $i = 1, \ldots, m$, tangent to S_{m-1}, takes the form $\sum_{i=1}^m \psi_i'(t) \dfrac{\partial h}{\partial x_i}$. Since $\sum_{i=1}^m \psi_i(t) = 1$, then $\sum_{i=1}^m \psi_i'(t) = 0$ and every vector X tangent to the simplex S_{m-1} can be represented as $X = \sum_{i=1}^m a_i X_i$, with $\sum_{i=1}^m a_i = 0$.

For $n \geq m$, let $B_1, \ldots, B_n$ be a partition of $\mathcal{X}$ such that $A_i = \bigcup_{j \in I_i} B_j$, $i = 1, \ldots, m$, where $I_1, \ldots, I_m$ is a partition of $\{1, \ldots, n\}$. For any probability measure P on $(\mathcal{X}, a)$, let $x_i = P(A_i)$, $i = 1, \ldots, m$, and $y_j = P(B_j)$, $j = 1, \ldots, n$. Thus

$$x_i = \sum_{j \in I_i} y_j, \quad i = 1, \ldots, m. \tag{3.1}$$

Conversely, define

$$q_{ij} = P(B_j | A_i) = \begin{cases} \dfrac{y_j}{x_i} & \text{if } j \in I_i, \\ 0 & \text{if } j \notin I_i. \end{cases}$$

Then

$$y_j = \sum_{i=1}^{m} q_{ij} x_i, \quad j = 1, \ldots, n. \tag{3.2}$$

Note that $(q_{ij})_{ij}$ is a stochastic matrix, that is: $q_{ij} \geq 0$, $\forall i, j$; $\sum_{j=1}^{n} q_{ij} = 1$, $\forall i$; and $q_{rj} q_{sj} = 0$ if $r \neq s$, $\forall j$. (3.2) defines a Markov kernel $f : S_{m-1} \to S_{n-1}$ and (3.1) an inverse of f, so that f is B-sufficient with respect to the family of all probability measures on $(\mathcal{X}, a_m)$. In fact, it is easy to prove that any B-sufficient Markov kernel with respect to the family of all probability measures on $(\mathcal{X}, a_m)$, can be written in the form (3.2). We call f a *Markov embedding.*

The jacobian map associated to f, $f^* : M_x \to M_y$, is defined by

$$f^* X_i = \sum_{j=1}^{n} q_{ij} Y_j, \quad i = 1, \ldots, m. \tag{3.3}$$

3.1 Embedding invariant structures

In the present section we consider geometrical structures defined on the simplex, that are invariant with respect to Markov embeddings. We characterize invariant Riemannian metrics and affine connections, showing that, up to constant factors, they coincides respectively with the Fisher metric and the α-connections.

3.1.1 Invariant Riemannian metrics

Definition 3.1 *If*

$$\langle U, V \rangle_m(x) = \langle f^* U, f^* V \rangle_n(y), \quad \forall U, V \in M_x, \tag{3.4}$$

where $x \in S_{m-1}$ and $y = f(x) \in S_{n-1}$, we say that f is an isometry and that $\langle \cdot, \cdot \rangle$ is invariant under f. $\langle \cdot, \cdot \rangle$ is said to be embedding invariant *if it is invariant under every Markov embedding.*

The following result was first given by Čencov (1972). Anyway, we refer the reader to Campbell (1986) for an easier proof.

Theorem 3.1 *The only embedding invariant Riemannian metrics are of the form*

$$\langle X_i, X_j \rangle_m(x) = A \frac{\delta_{ij}}{x_i}, \tag{3.5}$$

where $A > 0$ and δ_{ij} is the Kronecker delta.

Remark. (The Fisher metric) Theorem 3.1 states the unicity, up to a multiplicative constant, of embedding invariant metrics. Let $u_i = X_i - X_m$, $i = 1, \ldots, m-1$. Then

$$\begin{aligned} \langle u_i, u_j \rangle_m(x) &= \langle X_i - X_m, X_j - X_m \rangle_m(x) \\ &= A \left(\frac{\delta_{ij}}{x_i} - \frac{\delta_{im}}{x_i} - \frac{\delta_{mj}}{x_j} + \frac{1}{x_m} \right) = A \left(\frac{\delta_{ij}}{x_i} + \frac{1}{x_m} \right), \end{aligned} \tag{3.6}$$

that is the same, up to a constant factor, that the Fisher metric in the multinomial case, see Amari (1985, p. 31).

3.1.2 Invariant affine connections

A similar characterization can be given for embedding invariant affine connections. For this purpose, even though the tangent space M_x has dimension $m-1$, it will be better for us to work with an overdefined system of m vectors, $v_1, \ldots, v_m$, such that:

1. $\{v_{i_1}, \ldots, v_{i_{m-1}}\}$ is a base of M_x, $\forall \{i_1, \ldots, i_{m-1}\} \subset \{1, \ldots, m\}$;
2. $\sum_{i=1}^{m} v_i = 0$.

We choose $v_i = X_i - \frac{1}{m}\sum_{j=1}^{m} X_j$, $i = 1, \ldots, m$, where $X_i = \frac{\partial}{\partial x_i}$, $i = 1, \ldots, m$ is the usual base of M_x in $\mathbb{R}^m$. Notice that $v_i \in M_x$, $x \in S_{m-1}$, $\forall i = 1, \ldots, m$, since $v_i = \sum_{j=1}^{m} a_j X_j$, with $a_j = \delta_{ij} - \frac{1}{m}$, and $\sum_{j=1}^{m} a_j = 0$. We suppose the tangent space is equipped with an embedding invariant Riemannian metric.

We will describe an affine connection on S_{m-1} through the coefficients ${\gamma_{ij}}^k$ obtained by evaluating it on the vectors $v_1, \ldots, v_m$:

$$\nabla_{v_i} v_j = \sum_{k=1}^{m} {\gamma_{ij}}^k v_k, \quad \forall i, j = 1, \ldots, m. \tag{3.7}$$

Since $\{v_1, \ldots, v_m\}$ is not a base of M_x, ${\gamma_{ij}}^k$'s are not Christoffel symbols for the connection ∇ and, moreover, they are not uniquely determined. We choose the ${\gamma_{ij}}^k$'s satisfying the following condition:

$$\sum_{k=1}^{m} {\gamma_{ij}}^k = 0, \quad \forall i, j = 1, \ldots, m-1. \tag{3.8}$$

Condition (3.8) guarantees that expression (3.7) is a good definition for ∇. To see this, let

$$\sum_{k=1}^{m} {\gamma'_{ij}}^k v_k = \sum_{k=1}^{m} {\gamma_{ij}}^k v_k,$$

that is

$$\sum_{k=1}^{m} ({\gamma'_{ij}}^k - {\gamma_{ij}}^k) v_k = 0.$$

Since

$$0 = \sum_{k=1}^{m} a_k v_k = \sum_{k=1}^{m-1} (a_k - a_m) v_k \quad \Rightarrow \quad a_k = a_m, \quad \forall k = 1, \ldots, m-1,$$

then

$${\gamma'_{ij}}^k - {\gamma_{ij}}^k = {\gamma'_{ij}}^m - {\gamma_{ij}}^m, \quad \forall k = 1, \ldots, m-1.$$

By (3.8),

$$\sum_{k=1}^{m} ({\gamma'_{ij}}^k - {\gamma_{ij}}^k) = m({\gamma'_{ij}}^m - {\gamma_{ij}}^m) = 0,$$

that is ${\gamma'_{ij}}^m = {\gamma_{ij}}^m$ and

$${\gamma'_{ij}}^k = {\gamma_{ij}}^k, \quad \forall k = 1, \ldots, m-1.$$

¿From the ${\gamma_{ij}}^k$'s, $i, j, k = 1, \ldots, m-1$, we can easily obtain Christoffel symbols ${\Gamma_{ij}}^k$ for ∇, with respect to the base $v_1, \ldots, v_{m-1}$:

$$\begin{aligned} \nabla_{v_i} v_j &= \sum_{k=1}^{m} {\gamma_{ij}}^k v_k = \sum_{k=1}^{m-1} ({\gamma_{ij}}^k - {\gamma_{ij}}^m) v_k = \\ &= \sum_{k=1}^{m-1} ({\gamma_{ij}}^k + \sum_{l=1}^{m-1} {\gamma_{ij}}^l) v_k = \sum_{k=1}^{m-1} {\Gamma_{ij}}^k v_k. \end{aligned}$$

Thus

$$(3.9) \qquad {\Gamma_{ij}}^k = {\gamma_{ij}}^k + \sum_{l=1}^{m-1} {\gamma_{ij}}^l, \quad i, j, k = 1, \ldots, m-1.$$

Let $\overset{m}{\nabla}$, $\overset{m}{\gamma}_{ij}{}^k$ and $\overset{m}{\Gamma}_{ij}{}^k$ denote respectively the affine connection ∇ on S_{m-1}, its coefficients in (3.7) and the corresponding Christoffel symbols in (3.9). We will omit the superscript m when not necessary.

If f is a Markov embedding between S_{m-1} and S_{n-1}, $m < n$, then $\overset{m}{\nabla}$ induces through f an affine connection on $f(S_{m-1})$, an $(m-1)$-dimensional submanifold of S_{n-1}, defined by

$$f^* \left(\overset{m}{\nabla}_U \, V(x) \right), \quad \forall U, V \in M_x,$$

where $f^* : M_x \to M_y$ is the jacobian function corresponding to f. Moreover, as a submanifold of the Riemannian manifold S_{n-1}, $f(S_{m-1})$ naturally inherits the affine connection of S_{n-1}:

$$\overset{n}{\bar{\nabla}}_{f^* U} \, f^* V(y), \quad \forall U, V \in M_x,$$

where $\overset{n}{\bar{\nabla}}$ is the orthogonal projection of $\overset{n}{\nabla}$ on the tangent space to $f(S_{m-1})$. We can thus give the following definition of invariance:

Definition 3.2 *An affine connection ∇ is said to be* embedding invariant *if the affine connection induced by $\overset{m}{\nabla}$ on $f(S_{m-1})$ through f coincides with that induced* *on* $f(S_{m-1})$ *by* $\overset{n}{\nabla}$*, that is*

(3.10) $$f^*\left(\overset{m}{\nabla}_U V(x)\right) = \overset{\overset{n}{=}}{\nabla}_{f^*U} f^*V(y),$$

for every Markov embedding f*.*

Next theorem gives a characterization of the affine connections defined on the probability simplex, that are invariant under Markov embeddings. Following Campbell's characterization of invariant metrics, we give a proof that does not use Čencov's language of categories.

Theorem 3.2 *The only affine connections that are embedding invariant are of the form*

$$\gamma_{ij}{}^k(x) = I\left(x_k - \frac{\delta_{ik}+\delta_{jk}}{2}\right) + G\frac{\delta_{ij}}{x_i}\left(x_k - \delta_{ik}\right),$$

where I *and* G *are constants and* δ_{ij} *is the Kronecker delta.*

Proof: Condition (3.10) of invariance for ∇ can be written in terms of the vectors $v_1, \ldots, v_m$ of M_x as

(3.11) $$f^*\left(\overset{m}{\nabla}_{v_i} v_j(x)\right) = \overset{\overset{n}{=}}{\nabla}_{f^*v_i} f^*v_j(y), \quad i,j = 1, \ldots, m.$$

Consider first the case of $m = n$. In this situation, $f(S_{m-1}) = S_{n-1}$ and $\overset{\overset{n}{=}}{\nabla}$ coincides with $\overset{n}{\nabla}$. If f interchanges x_r and x_s, then f^* interchanges X_r and X_s, and thus v_r and v_s. Let $x_1 = \ldots = x_m$, that is $x = \left(\frac{1}{m}, \ldots, \frac{1}{m}\right)$; then $f(x) = x$.

If $i = j$, by (3.11) we obtain:

$$f^*\left(\nabla_{v_i} v_i(x)\right) = f^*\left(\gamma_{ii}{}^k(x) v_k\right) = \gamma_{ii}{}^k(x) f^* v_k = \nabla_{f^*v_i} f^* v_i(x).$$

It follows that:

$$\gamma_{rr}{}^r(x) = \gamma_{ss}{}^s(x),$$
$$\gamma_{ii}{}^r(x) = \gamma_{ii}{}^s(x), \quad i \notin \{r, s\},$$

and

$$\gamma_{rr}{}^{i}(x) = \gamma_{ss}{}^{i}(x), \quad i \notin \{r, s\}.$$

Since r and s may be chosen arbitrarly, we may write

$$\gamma_{ii}{}^{i}(x) = F_m$$

and

$$\gamma_{ii}{}^{k}(x) = G_m, \quad i \neq k.$$

If $i \neq j$,

$$f^* \left(\nabla_{v_i} v_j(x)\right) = f^* \left(\gamma_{ij}{}^{k}(x) v_k\right) = \gamma_{ij}{}^{k}(x) f^* v_k = \nabla_{f^* v_i} f^* v_j(x).$$

It follows:

$$\begin{aligned} \gamma_{rs}{}^{k}(x) &= \gamma_{sr}{}^{k}(x), \\ \gamma_{kr}{}^{s}(x) &= \gamma_{ks}{}^{r}(x), \\ \gamma_{sk}{}^{r}(x) &= \gamma_{rk}{}^{s}(x), \\ \gamma_{ij}{}^{r}(x) &= \gamma_{ij}{}^{s}(x), \end{aligned}$$

where $k, i, j \notin \{r, s\}$, which imply

$$\gamma_{ij}{}^{k}(x) = I_m, \quad i \neq j \neq k.$$

Similarly,

$$\begin{aligned} \gamma_{ir}{}^{i}(x) &= \gamma_{is}{}^{i}(x), \\ \gamma_{ri}{}^{r}(x) &= \gamma_{si}{}^{s}(x), \\ \gamma_{ir}{}^{r}(x) &= \gamma_{is}{}^{s}(x), \\ \gamma_{ri}{}^{i}(x) &= \gamma_{si}{}^{i}(x), \end{aligned}$$

where $i \notin \{r, s\}$, together with the simmetry condition

$$\gamma_{ij}{}^{k} = \gamma_{ji}{}^{k},$$

imply

$$\gamma_{ij}{}^{k}(x) = H_m, \quad k \in \{i, j\}.$$

Thus, for $x = \left(\frac{1}{m}, \ldots, \frac{1}{m}\right)$,

$$\overset{m}{\gamma}_{ij}{}^{k}(x) = \begin{cases} F_m & i = j = k \\ G_m & i = j \neq k \\ H_m & i \neq j, \quad k \in \{i, j\} \\ I_m & i \neq j, \quad k \notin \{i, j\}. \end{cases}$$

Next, let $n = hm$, where h is an integer bigger than one. If f_h is the Markov embedding defined by $y = f_h(x) = \left(\frac{x_1}{h}, \ldots, \frac{x_1}{h}, \ldots, \frac{x_m}{h}, \ldots, \frac{x_m}{h}\right)$, each component being repeated h times, then $f^*X_i = \frac{1}{h}\left(Y_{(i-1)h+1} + \ldots + Y_{ih}\right)$ and $f^*v_i = \frac{1}{h}\left(u_{(i-1)h+1} + \ldots + u_{ih}\right) = \frac{1}{h}\sum_{r \in R_i} u_r$, where $u_i = Y_i - \frac{1}{n}\sum_{j=1}^{n} Y_j$ and $R_i = \{(i-1)h+1, \ldots, ih\}$, $\quad i = 1, \ldots, m$.

For $i = j$, we have

$$\begin{aligned} f^*\left(\overset{m}{\nabla}_{v_i} v_i(x)\right) &= f^*\left(\overset{m}{\gamma}_{ii}{}^{k}(x) v_k\right) = \overset{m}{\gamma}_{ii}{}^{k}(x) f^* v_k = \frac{1}{h}\, \overset{m}{\gamma}_{ii}{}^{k}(x) \sum_{r \in R_k} u_r \\ &= \frac{1}{h}\left(\overset{m}{\gamma}_{ii}{}^{i}(x) \sum_{r \in R_i} u_r + \sum_{k \neq i} \overset{m}{\gamma}_{ii}{}^{k}(x) \sum_{r \in R_k} u_r\right) \\ &= \frac{1}{h}\left(F_m \sum_{r \in R_i} u_r + G_m \sum_{r \notin R_i} u_r\right), \end{aligned}$$

and

$$\begin{aligned} &\overset{n}{\nabla}_{f^*v_i} f^* v_i(y) = \\ &= \frac{1}{h^2} \sum_{r,s \in R_i} \overset{n}{\nabla}_{u_r} u_s(y) = \frac{1}{h^2} \sum_{r,s \in R_i} \overset{n}{\gamma}_{rs}{}^{k}(y) u_k \\ &= \frac{1}{h^2}\left[\sum_{r \in R_i} \overset{n}{\gamma}_{rr}{}^{k}(y) u_k + \sum_{r \in R_i} \sum_{\substack{s \in R_i \\ s \neq r}} \overset{n}{\gamma}_{rs}{}^{k}(y) u_k\right] \\ &= \frac{1}{h^2}\left[\sum_{r \in R_i} \overset{n}{\gamma}_{rr}{}^{r}(y) u_r + \sum_{r \in R_i} \sum_{k \neq r} \overset{n}{\gamma}_{rr}{}^{k}(y) u_k + \sum_{r \in R_i} \sum_{\substack{s \in R_i \\ s \neq r}} \sum_{k \neq r,s} \overset{n}{\gamma}_{rs}{}^{k}(y) u_k\right. \end{aligned}$$

$$+\sum_{r\in R_i}\sum_{\substack{s\in R_i\\ s\neq r}}\left(\overset{n}{\gamma}_{rs}{}^{r}(y)u_r+\overset{n}{\gamma}_{rs}{}^{s}(y)u_s\right)\Bigg]$$

$$= \frac{1}{h^2}\Bigg[F_n\sum_{r\in R_i}u_r + hG_n\sum_{r\notin R_i}u_r + (h-1)G_n\sum_{r\in R_i}u_r$$

$$+h(h-1)I_n\sum_{r\notin R_i}u_r + (h-1)(h-2)I_n\sum_{r\in R_i}u_r + 2(h-1)H_n\sum_{r\in R_i}u_r\Bigg].$$

Since $\overset{n}{\nabla}_{f^*v_i} f^*v_i(y)$ belongs to the tangent space to $f(S_{m-1})$ in y, we do not need to project it to find $\overset{\underline{n}}{\nabla}_{f^*v_i} f^*v_i(y)$, that is, $\overset{\underline{n}}{\nabla}=\overset{n}{\nabla}$.

Condition (3.11) implies:

$$G_m = G_n + (h-1)I_n \tag{3.12}$$

and

$$F_m = \frac{F_n}{h} + \frac{h-1}{h}G_n + \frac{(h-1)(h-2)}{h}I_n + 2\frac{h-1}{h}H_n. \tag{3.13}$$

For $i \neq j$, we have:

$$\begin{aligned} f^*\left(\overset{m}{\nabla}_{v_i} v_j(x)\right) &= f^*\left(\overset{m}{\gamma}_{ij}{}^{k}(x)v_k\right) = \overset{m}{\gamma}_{ij}{}^{k}(x)f^*v_k = \frac{1}{h}\,\overset{m}{\gamma}_{ij}{}^{k}(x)\sum_{r\in R_k}u_r \\ &= \frac{1}{h}\left(\sum_{k\notin\{i,j\}}\overset{m}{\gamma}_{ij}{}^{k}(x)\sum_{r\in R_k}u_r + \overset{m}{\gamma}_{ij}{}^{i}(x)\sum_{r\in R_i}u_r + \overset{m}{\gamma}_{ij}{}^{j}(x)\sum_{r\in R_j}u_r\right) \\ &= \frac{1}{h}\left(I_m\sum_{r\notin R_i\cup R_j}u_r + H_m\sum_{r\in R_i\cup R_j}u_r\right), \end{aligned}$$

and

$$\begin{aligned} &\overset{n}{\nabla}_{f^*v_i} f^*v_j(y) = \\ &= \frac{1}{h^2}\sum_{r\in R_i}\sum_{s\in R_j}\overset{n}{\nabla}_{u_r} u_s(y) = \frac{1}{h^2}\sum_{r\in R_i}\sum_{s\in R_j}\overset{n}{\gamma}_{rs}{}^{k}(y)u_k \\ &= \frac{1}{h^2}\left[\sum_{r\in R_i}\sum_{s\in R_j}\sum_{k\neq r,s}\overset{n}{\gamma}_{rs}{}^{k}(y)u_k + \sum_{r\in R_i}\sum_{s\in R_j}\left(\overset{n}{\gamma}_{rs}{}^{r}(y)u_r+\overset{n}{\gamma}_{rs}{}^{s}(y)u_s\right)\right] \end{aligned}$$

$$= \frac{1}{h^2}\left[h^2 I_n \sum_{r\notin R_i\cup R_j} u_r + h(h-1)I_n \sum_{r\in R_i\cup R_j} u_r + hH_n \sum_{r\in R_i\cup R_j} u_r\right].$$

$\overset{n}{\nabla}_{f^*v_i} f^*v_j(y)$ still belongs to the tangent space to $f(S_{m-1})$ in y and $\overset{\underline{n}}{\nabla}$ coincides with $\overset{n}{\nabla}$.

By (3.11),

(3.14) $$I_m = hI_n$$

and

(3.15) $$H_m = H_n + (h-1)I_n.$$

Now, by (3.14), $mI_m = nI_n = I$, that is

$$I_m = \frac{I}{m}.$$

By (3.15), $H_m - I_m = H_n - I_n = H$, so

$$H_m = H + \frac{I}{m}.$$

A similar expression holds, by (3.12), for G:

$$G_m = G + \frac{I}{m}.$$

Finally, (3.13) implies that $\frac{1}{m}(F_m - G - 2H - I_m) = \frac{1}{n}(F_n - G - 2H - I_n) = F$, that is

$$F_m = mF + \frac{I}{m} + G + 2H.$$

We can then write, for $x = \left(\frac{1}{m},\ldots,\frac{1}{m}\right)$,

$$\overset{m}{\gamma}_{ij}{}^k(x) = \begin{cases} mF + I/m + G + 2H & i = j = k \\ G + I/m & i = j \neq k \\ H + I/m & i \neq j, \quad k \in \{i,j\} \\ I/m & i \neq j, \quad k \notin \{i,j\}. \end{cases}$$

Let now $x = \left(\frac{r_1}{n}, \ldots, \frac{r_m}{n}\right)$, where $\sum_{i=1}^m r_i = n$, and all r_i are positive integers. Define

$$q_{ij} = \begin{cases} \frac{1}{r_i} & \text{if } j \in R_i \\ 0 & \text{otherwise,} \end{cases}$$

where $R_i = \{r_1 + \ldots + r_{i-1} + 1, \ldots, r_1 + \ldots + r_i\}$, $i = 1, \ldots, m$. The corresponding Markov embedding f maps x to $y = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$. Moreover

$$f^* X_i = \frac{1}{r_i}\left(Y_{r_1+\ldots+r_{i-1}+1} + \ldots + Y_{r_1+\ldots+r_i}\right)$$

and thus

$$f^* v_i = \frac{1}{r_i}\left(u_{r_1+\ldots+r_{i-1}+1} + \ldots + u_{r_1+\ldots+r_i}\right) = \frac{1}{r_i}\sum_{r\in R_i} u_r.$$

For $i = j$, we have

$$f^*\left(\overset{m}{\nabla}_{v_i} v_i(x)\right) = \overset{m}{\gamma}_{ii}{}^k(x) f^* v_k = \frac{1}{r_k}\overset{m}{\gamma}_{ii}{}^k(x)\sum_{r\in R_k} u_r,$$

and

$$\begin{aligned}
\overset{n}{\nabla}_{f^*v_i} f^* v_i(y) &= \frac{1}{r_i^2}\sum_{r,s\in R_i}\overset{n}{\nabla}_{u_r} u_s(y) = \frac{1}{r_i^2}\sum_{r,s\in R_i}\overset{n}{\gamma}_{rs}{}^k(y)u_k \\
&= \frac{1}{r_i^2}\left[\sum_{r\in R_i}\overset{n}{\gamma}_{rr}{}^k(y)u_k + \sum_{\substack{r,s\in R_i\\ s\neq r}}\overset{n}{\gamma}_{rs}{}^k(y)u_k\right] \\
&= \frac{1}{r_i^2}\left[\left(nF + \frac{I}{n} + G + 2H\right)\sum_{r\in R_i} u_r + r_i\left(G + \frac{I}{n}\right)\sum_{r\notin R_i} u_r\right. \\
&\quad + (r_i - 1)\left(G + \frac{I}{n}\right)\sum_{r\in R_i} u_r + r_i(r_i - 1)\frac{I}{n}\sum_{r\notin R_i} u_r \\
&\quad \left. + (r_i - 1)(r_i - 2)\frac{I}{n}\sum_{r\in R_i} u_r + 2(r_i - 1)\left(H + \frac{I}{n}\right)\sum_{r\in R_i} u_r\right].
\end{aligned}$$

Since $\overset{n}{\nabla}_{f^*v_i} f^* v_j(y)$ belongs to the tangent space to $f(S_{m-1})$ in y, $\overset{\underline{n}}{\nabla} = \overset{n}{\nabla}$.

Invariance condition (3.11) gives:

$$\overset{m}{\gamma}_{ii}{}^{i}(x) = \frac{1}{r_i}\left[nF + r_iG + 2r_iH + \frac{I}{n}\right] = \frac{F}{x_i} + G + 2H + Ix_i$$

and

$$\overset{m}{\gamma}_{ii}{}^{k}(x) = \frac{r_k}{r_i}G + \frac{r_k}{n}I = G\frac{x_k}{x_i} + Ix_k, \quad k \neq i.$$

For $x = \left(\frac{r_1}{n}, \ldots, \frac{r_m}{n}\right)$, we have:

$$\overset{m}{\gamma}_{ij}{}^{k}(x) = \begin{cases} \frac{F}{x_i} + Ix_i + G + 2H & i = j = k \\ G\frac{x_k}{x_i} + Ix_k & i = j \neq k \\ H + Ix_k & i \neq j, \quad k \in \{i, j\} \\ Ix_k & i \neq j, \quad k \notin \{i, j\}. \end{cases}$$

Finally, any $x \in S_{m-1}$ can be approximated arbitrarily well by an x of the form $\left(\frac{r_1}{n}, \ldots, \frac{r_m}{n}\right)$. Since the $\overset{m}{\gamma}_{ij}{}^{k}$'s are C^∞ functions, then

$$\overset{m}{\gamma}_{ij}{}^{k}(x) = Ix_k + H(\delta_{ik} + \delta_{jk}) + \delta_{ij}\left[G\frac{x_k}{x_i} + F\frac{\delta_{ik}}{x_i}\right], \quad \forall x \in S_{m-1}. \tag{3.16}$$

We can now impose condition (3.8) on the coefficients $\overset{m}{\gamma}_{ij}{}^{k}$'s. If $i \neq j$, we obtain $H = -\frac{I}{2}$; for $i = j$, we have $F = -G$. Substituting these conditions in (3.16) gives the result. ∎

Remark. (α-connections) By (3.9), we can obtain the Christoffel symbols of any embedding invariant connection on the simplex S_{m-1}, with respect to the base $\{v_1, \ldots, v_{m-1}\}$:

$$\overset{m}{\Gamma}_{ij}{}^{k}(x) = I\left(x_k - x_m - \frac{\delta_{ik} + \delta_{jk}}{2}\right) + G\frac{\delta_{ij}}{x_i}(x_k - x_m - \delta_{ik}). \tag{3.17}$$

Let us now show that the affine connections characterized by (3.17) are in fact, up to a constant factor, Amari's α-connections.

Let $u_i = v_i - v_m = X_i - X_m$, $i = 1, \dots, m-1$. Then $v_m = -\frac{1}{m}\sum_{i=1}^{m-1} u_i$. Using the repeated index convention and avoiding the superindex m,

$$\begin{aligned}
\nabla_{u_i} u_j &= \nabla_{v_i - v_m}(v_j - v_m) \\
&= \nabla_{v_i} v_j - \nabla_{v_i} v_m - \nabla_{v_m} v_j + \nabla_{v_m} v_m \\
&= \gamma_{ij}{}^k v_k - \gamma_{im}{}^k v_k - \gamma_{jm}{}^k v_k + \gamma_{mm}{}^k v_k \\
&= \left(\gamma_{ij}{}^k - \gamma_{im}{}^k - \gamma_{jm}{}^k + \gamma_{mm}{}^k\right)\left(u_k - \frac{1}{m}\sum_{i=1}^{m-1} u_i\right) \\
&= \left(\gamma_{ij}{}^k - \gamma_{im}{}^k - \gamma_{jm}{}^k + \gamma_{mm}{}^k - \frac{1}{m}\sum_{k=1}^{m}\left(\gamma_{ij}{}^k - \gamma_{im}{}^k - \gamma_{jm}{}^k + \gamma_{mm}{}^k\right)\right) u_k,
\end{aligned}$$

for $i, j = 1, 2, \dots, m-1$. Now, by (3.8) and since

$$\sum_{k=1}^{m} \gamma_{im}{}^k = I \sum_{k=1}^{m}\left(x_k - \frac{\delta_{ik} + \delta_{mk}}{2}\right) + G\frac{\delta_{im}}{x_i}\sum_{k=1}^{m}(x_k - \delta_{ik}) = 0$$

for $i = 1, \dots, m$, we have

$$\Gamma_{ij}{}^k = \gamma_{ij}{}^k - \gamma_{im}{}^k - \gamma_{jm}{}^k + \gamma_{mm}{}^k,$$

for $i, j, k = 1, 2, \dots, m-1$. That is,

$$\begin{aligned}
\Gamma_{ij}{}^k &= I\left(x_k - \frac{\delta_{ik} + \delta_{jk}}{2} - x_k + \frac{\delta_{ik} + \delta_{mk}}{2} - x_k + \frac{\delta_{jk} + \delta_{mk}}{2} + x_k - \frac{\delta_{mk} + \delta_{mk}}{2}\right) \\
&\quad + G\left(\frac{\delta_{ij}}{x_i}(x_k - \delta_{ik}) - \frac{\delta_{im}}{x_i}(x_k - \delta_{ik}) - \frac{\delta_{jm}}{x_j}(x_k - \delta_{jk}) + \frac{1}{x_m}(x_k - \delta_{mk})\right) \\
&= G\left(x_k\left(\frac{\delta_{ij}}{x_i} + \frac{1}{x_m}\right) - \frac{\delta_{ij}\delta_{ik}}{x_i}\right).
\end{aligned}$$

Finally,

$$\begin{aligned}
\Gamma_{ijk} &= \Gamma_{ij}{}^r g_{rk} = G\sum_{r=1}^{m-1}\left(x_r\left(\frac{\delta_{ij}}{x_i} + \frac{1}{x_m}\right) - \frac{\delta_{ij}\delta_{ir}}{x_i}\right)\left(\frac{\delta_{rk}}{x_r} + \frac{1}{x_m}\right) \\
&= G\left(\frac{1}{x_m^2} - \frac{\delta_{ijk}}{x_i^2}\right), \qquad (3.18)
\end{aligned}$$

that coincides, up to a constant factor, with the coefficient of the α-connections in the multinomial case, see Amari (1985, p. 43).

3.2 An expansion for D

To study the local behaviour of a divergence D, suppose it is weakly smooth, that is, at any point of S_{m-1} it admits an expression in local coordinates that is differentiable up to necessary order.

Theorem 3.3 *At each point P of S_{m-1}, any monotone divergence $D(P,Q)$, admits the expansion*

$$\begin{aligned} D(P,Q) &= D_0 + D_1 \sum_{i=1}^{m} \frac{[Q(A_i)-P(A_i)]^2}{P(A_i)} + D_2 \sum_{i=1}^{m} \frac{[Q(A_i)-P(A_i)]^3}{P(A_i)^2} \\ &+ D_3 \sum_{i=1}^{m} \frac{[Q(A_i)-P(A_i)]^4}{P(A_i)^3} + D_4 \left(\sum_{i=1}^{m} \frac{[Q(A_i)-P(A_i)]^2}{P(A_i)} \right)^2 \\ &+ o(\|Q-P\|^4), \end{aligned}$$

where D_0, D_1, D_2, D_3, D_4 are constants, $D_1 > 0$.

Proof: Consider in S_{m-1} the system of local coordinates that to each $P \in S_{m-1}$ associates a vector $p = (p_1, \ldots, p_{m-1})$, where $p_i = P(A_i)$, $i = 1, \ldots, m$. Then $D(P,Q) = d(p,q)$ is the expression of D in local coordinates. In a neighborhood of p, we have:

$$\begin{aligned} d(p,q) &= d(p,p) + d_{;i}(p,p)(q_i - p_i) + \frac{1}{2} d_{;ij}(p,p)(q_i-p_i)(q_j-p_j) \\ &+ \frac{1}{6} d_{;ijk}(p,p)(q_i-p_i)(q_j-p_j)(q_k-p_k) \\ &+ \frac{1}{24} d_{;ijkh}(p,p)(q_i-p_i)(q_j-p_j)(q_k-p_k)(q_h-p_h) + o(\|q-p\|^4), \end{aligned}$$

where $d_{.;.}$ are the partial derivatives of d with respect to the two arguments. For any divergence we have that:

$$d(p,p) = d_0,$$

$$d_{;i}(p,p) = 0$$

and

$$d_{;ij}(p,p) \geq 0,$$

since d takes minimum value in the diagonal. Moreover, by Proposition 2.1, monotonicity of D implies its invariance; then, if it does not vanish, that is the only interesting case, $d_{;ij}(p,p)$ defines an inner product on $\mathbb{R}^{m-1}$ which is embedding invariant. We take $d_{;ij}(p,p)$ as the metric tensor and denote by $d^{;ij}$ the inverse matrix of $d_{;ij}$. By (3.6), we obtain

$$d_{;ij}(p,p) = A\left(\frac{\delta_{ij}}{p_i} + \frac{1}{p_m}\right), \quad i,j = 1,\ldots,m-1, \quad A > 0. \tag{3.19}$$

Thus,

$$\begin{aligned}
\sum_{i,j=1}^{m-1} d_{;ij}(p,p)(q_i - p_i)(q_j - p_j) &= \\
&= A\sum_{i,j=1}^{m-1}\left(\frac{\delta_{ij}}{p_i} + \frac{1}{p_m}\right)(q_i - p_i)(q_j - p_j) \\
&= A\left(\sum_{i=1}^{m-1}\frac{(q_i - p_i)^2}{p_i} + \sum_{i,j=1}^{m-1}\frac{(q_i - p_i)(q_j - p_j)}{p_m}\right) \\
&= A\left(\sum_{i=1}^{m-1}\frac{(q_i - p_i)^2}{p_i} + \frac{(q_m - p_m)^2}{p_m}\right) \\
&= A\sum_{i=1}^{m}\frac{(q_i - p_i)^2}{p_i},
\end{aligned} \tag{3.20}$$

since $q_m - p_m = -\sum_{i=1}^{m-1}(q_i - p_i)$.

As regards the third order term, by differentiating (3.19), we obtain that

$$d_{;ijk}(p,p) + d_{k;ij}(p,p) = A\left(\frac{1}{p_m^2} - \frac{\delta_{ijk}}{p_i^2}\right). \tag{3.21}$$

On the other hand, $d_{k;ij}(p,p)$ behaves as the Christoffel symbols of an affine connection, see Amari (1985, p. 98). Since D is an invariant divergence, the affine connection is embedding invariant and by (3.18) we can write

$$d_{k;ij}(p,p) = H\left(\frac{1}{p_m^2} - \frac{\delta_{ijk}}{p_i^2}\right), \quad i,j,k = 1,\ldots,m-1.$$

By (3.21),

$$d_{;ijk}(p,p) = (A-H)\left(\frac{1}{p_m^2} - \frac{\delta_{ijk}}{p_i^2}\right) = B\left(\frac{1}{p_m^2} - \frac{\delta_{ijk}}{p_i^2}\right). \tag{3.22}$$

Thus,

$$\begin{aligned}\sum_{i,j,k=1}^{m-1} d_{;ijk}(p,p)(q_i-p_i)(q_j-p_j)(q_k-p_k) = \\ = \; B\sum_{i,j,k=1}^{m-1}\left(\frac{1}{p_m^2} - \frac{\delta_{ijk}}{p_i^2}\right)(q_i-p_i)(q_j-p_j)(q_k-p_k) \\ = \; -B\sum_{i=1}^{m}\frac{(q_i-p_i)^3}{p_i^2}.\end{aligned} \tag{3.23}$$

Let us now study the fourth order term. By differentiating (3.22), we obtain

$$d_{;ijkh}(p,p) + d_{h;ijk}(p,p) = B\left(\frac{\delta_{ijkh}}{p_i^3} + \frac{1}{p_m^3}\right). \tag{3.24}$$

Since $d^t_{;ijk}(p,p) = d^{;ht} d_{h;ijk}(p,p)$, behaves as the components of a connection string (see Blaesild (1988)), we can write

$$d^t_{;ijk} e_t = \nabla_i \nabla_j e_k,$$

for some covariant derivative ∇ and any vector e. Moreover, by the invariance of D, ∇ must be embedding invariant, that is, the corresponding Christoffel symbols are of the form

$$\Gamma_{ijk} = G\left(\frac{1}{p_m^2} - \frac{\delta_{ijk}}{p_i^2}\right).$$

In order to calculate $d^t_{;ijk}$, we will need the following expressions:

$$\partial_i g_{st} = A\left(\frac{1}{p_m^2} - \frac{\delta_{ist}}{p_i^2}\right),$$

$$\partial_i \Gamma_{jkr} = 2G\left(\frac{1}{p_m^3} + \frac{\delta_{ijkr}}{p_i^3}\right),$$

$$d^{;rs} = A^{-1}(\delta_{rs}p_r - p_r p_s),$$

$$\sum_{r,s=1}^{m-1} d^{;rs} = A^{-1}p_m(1-p_m),$$

$$\sum_{r,s=1}^{m-1} d^{;rs}\delta_{jkr} = A^{-1}\delta_{jk}p_j p_m,$$

$$\sum_{r,s=1}^{m-1} d^{;rs}\delta_{jkr}\delta_{ist} = A^{-1}(\delta_{ijkt}p_i - \delta_{jk}\delta_{it}p_i p_j).$$

We thus have

$$\begin{aligned} d^t_{;ijk}e_t &= \nabla_i\nabla_j e_k = \nabla_i({\Gamma_{jk}}^s e_s) = \partial_i({\Gamma_{jk}}^s)e_s + {\Gamma_{jk}}^s\nabla_i e_s \\ &= (\partial_i{\Gamma_{jk}}^t + {\Gamma_{jk}}^s{\Gamma_{is}}^t)e_t. \end{aligned}$$

Now,

$$\begin{aligned} &\partial_i{\Gamma_{jk}}^t + {\Gamma_{jk}}^s{\Gamma_{is}}^t = \\ &= -d^{;rs}d^{;ht}\partial_i d_{;sh}\Gamma_{jkr} + d^{;rs}d^{;ht}\Gamma_{jkr}\Gamma_{ish} + d^{;ht}\partial_i\Gamma_{jkh} \\ &= d^{;ht}\left[d^{;rs}\Gamma_{jkr}\left(-\partial_i d_{;sh} + \Gamma_{ish}\right) + \partial_i\Gamma_{jkh}\right] \\ &= d^{;ht}\left[d^{;rs}G(G-A)\left(\frac{1}{p_m^2} - \frac{\delta_{ish}}{p_i^2}\right)\left(\frac{1}{p_m^2} - \frac{\delta_{jkr}}{p_j^2}\right) + 2G\left(\frac{1}{p_m^3} + \frac{\delta_{ijkh}}{p_i^3}\right)\right] \\ &= d^{;ht}\left[C\left(\frac{1}{p_m} + \frac{\delta_{ih}}{p_i}\right)\left(\frac{1}{p_m} + \frac{\delta_{jk}}{p_j}\right) + F\left(\frac{1}{p_m^3} + \frac{\delta_{ijkh}}{p_i^3}\right)\right]. \end{aligned}$$

Then

$$d_{h;ijk} = C\left(\frac{1}{p_m} + \frac{\delta_{jk}}{p_j}\right)\left(\frac{1}{p_m} + \frac{\delta_{ih}}{p_i}\right) + F\left(\frac{1}{p_m^3} + \frac{\delta_{ijkh}}{p_i^3}\right).$$

Finally, we substitute the preceding expression in (3.24), obtaining

$$(3.25)\qquad d_{;ijkh}(p,p) = C\left(\frac{1}{p_m} + \frac{\delta_{jk}}{p_j}\right)\left(\frac{1}{p_m} + \frac{\delta_{ih}}{p_i}\right) + D\left(\frac{1}{p_m^3} + \frac{\delta_{ijkh}}{p_i^3}\right).$$

Thus,

$$(3.26)\qquad \sum_{i,j,k,h=1}^{m-1} d_{;ijkh}(p,p)(q_i-p_i)(q_j-p_j)(q_k-p_k)(q_h-p_h) =$$

$$
\begin{aligned}
&= C\left[\sum_{j,k=1}^{m-1}\left(\frac{1}{p_m}+\frac{\delta_{jk}}{p_j}\right)(q_j-p_j)(q_k-p_k)\sum_{i,h=1}^{m-1}\left(\frac{1}{p_m}+\frac{\delta_{ih}}{p_i}\right)(q_i-p_i)(q_h-p_h)\right] \\
&\quad +D\sum_{i=1}^{m}\frac{(q_i-p_i)^4}{p_i^3} \\
&= C\left(\sum_{i=1}^{m}\frac{(q_i-p_i)^2}{p_i}\right)^2+D\sum_{i=1}^{m}\frac{(q_i-p_i)^4}{p_i^3}
\end{aligned}
$$

By (3.20), (3.23) and (3.26), we can finally write

$$
\begin{aligned}
d(p;q) &= d_0+d_1\sum_{i=1}^{m}\frac{(q_i-p_i)^2}{p_i}+d_2\sum_{i=1}^{m}\frac{(q_i-p_i)^3}{p_i^2}+d_3\sum_{i=1}^{m}\frac{(q_i-p_i)^4}{p_i^3} \\
&\quad +d_4\left(\sum_{i=1}^{m}\frac{(q_i-p_i)^2}{p_i}\right)^2+o(\|q-p\|^4),
\end{aligned}
$$

$d_1 > 0$, that, written in terms of D, gives the result. ■

4 The general case

In the preceding section, we obtained a local expression for monotone divergences defined on multinomial distributions. Using the regularity condition, we are able to extend this expansion to the general case. We need the following result:

Theorem 4.1 *Let $f : \mathbb{R} \to \mathbb{R}_+$ be a convex function. For any P and Q, equivalent probability measures on $(\mathcal{X}, a)$, there exists a non negative integral*

$$
\int_{\mathcal{X}} f\left(\frac{dQ}{dP}(x)\right)P(dx)=\lim_{\alpha}\sum_i f\left(\frac{Q(A_i)}{P(A_i)}\right)P(A_i),
$$

where the limit is taken over the filter of all finite subσ-fields a_α of a.

Proof: Since

$$
\sum_i f\left(\frac{Q(A_i)}{P(A_i)}\right)P(A_i)=\int_{\mathcal{X}} f\left(\frac{dQ_\alpha}{dP_\alpha}(x)\right)P_\alpha(dx),
$$

the thesis can be written in the form

$$\lim_{\alpha} \int_{\mathcal{X}} f\left(\frac{dQ_\alpha}{dP_\alpha}(x)\right) P_\alpha(dx) = \int_{\mathcal{X}} f\left(\frac{dQ}{dP}(x)\right) P(dx).$$

It is sufficient to prove it for any increasing sequence $\{a_n\}$ of finite subσ-field of a, such that $\sigma(\bigcup_n a_n) = a$. Since f-divergences are monotone, see Heyer (1982, Theorem 22.9, p.169), and by the remark following Definition 2.5, we obtain:

$$\lim_{n} \int_{\mathcal{X}} f\left(\frac{dQ_n}{dP_n}(x)\right) P_n(dx) \leq \int_{\mathcal{X}} f\left(\frac{dQ}{dP}(x)\right) P(dx).$$

We show now that the reverse inequality also holds. Since

$$\int_{\mathcal{X}} \frac{dQ}{dP}(x) P(dx) = 1 < \infty,$$

we can apply a well known theorem of convergence of martingales, thus obtaining

$$\frac{dQ_n}{dP_n} = E\left(\frac{dQ}{dP}\bigg| a_n\right) \overset{a.s.}{\to} E\left(\frac{dQ}{dP}\bigg| a\right) = \frac{dQ}{dP}.$$

Since f is a continuous function, the convergence still holds:

$$f\left(\frac{dQ_\alpha}{dP_\alpha}\right) \overset{a.s.}{\to} f\left(\frac{dQ}{dP}\right).$$

By the Fatou lemma,

$$\int_{\mathcal{X}} f\left(\frac{dQ}{dP}(x)\right) P(dx) \leq \lim_{n} \int_{\mathcal{X}} f\left(\frac{dQ_n}{dP_n}(x)\right) P(dx) = \lim_{n} \int_{\mathcal{X}} f\left(\frac{dQ_n}{dP_n}(x)\right) P_n(dx),$$

and the thesis is proved. ∎

Remark. The preceding theorem can be easily extended to the case of f being any linear combination of non negative convex functions.

We can now prove the main result of the present section:

Theorem 4.2 *If*

$$\int_{\mathcal{X}} \left| \frac{Q(dx) - P(dx)}{P(dx)} \right|^4 P(dx) < \infty,$$

then, at each point P, any monotone and regular divergence $D(P,Q)$, admits the expansion

$$\begin{aligned}(4.1)\ D(P,Q) &= D_0 + D_1 \int_{\mathcal{X}} \frac{[Q(dx) - P(dx)]^2}{P(dx)} + D_2 \int_{\mathcal{X}} \frac{[Q(dx) - P(dx)]^3}{P(dx)^2} \\ &+ D_3 \int_{\mathcal{X}} \frac{[Q(dx) - P(dx)]^4}{P(dx)^3} + D_4 \left(\int_{\mathcal{X}} \frac{[Q(dx) - P(dx)]^2}{P(dx)} \right)^2 \\ &+ o(\|Q - P\|^4),\end{aligned}$$

where D_0, D_1, D_2, D_3, D_4 are constants, $D_1 > 0$.

Proof: By Theorem 3.3, it holds:

$$\begin{aligned} D(P_\alpha, Q_\alpha) &= D_0 + D_1 \sum_i \frac{[Q(A_i) - P(A_i)]^2}{P(A_i)} + D_2 \sum_i \frac{[Q(A_i) - P(A_i)]^3}{P(A_i)^2} \\ &+ D_3 \sum_i \frac{[Q(A_i) - P(A_i)]^4}{P(A_i)^3} + D_4 \left(\sum_i \frac{[Q(A_i) - P(A_i)]^2}{P(A_i)} \right)^2 \\ &+ o(\|Q_\alpha - P_\alpha\|^4),\end{aligned}$$

for every P_α and Q_α, restrictions of P and Q to the finite dimensional subσ-field a_α of a. We can now pass to the limit. The terms with coefficients D_1 and D_4 can be obtained by applying Theorem 4.1 with $f(x) = (x-1)^2$. The same holds for the term with coefficient D_3, with $f(x) = (x-1)^4$. For the third order term we can use the remark following Theorem 4.1, since $f(x) = (x-1)^3$ can be written as the difference of two non negative convex functions:

$$f(x) = f_1(x) - f_2(x),$$

where

$$f_1(x) = \begin{cases} 0 & x \le 1 \\ (x-1)^3 & x > 1 \end{cases}$$

and

$$f_2(x) = \begin{cases} -(x-1)^3 & x \le 1 \\ 0 & x > 1. \end{cases}$$

The regularity of D guarantees the result. ■

4.1 The parametric case

Suppose now that P and Q belong to some regular parametric model, that is, P and Q are equivalent probability measures with densities $p(x;\theta)$ and $p(x;\theta')$, θ, $\theta' \in \Theta \subset \mathbb{R}^k$, with respect to some common dominating measure μ. By Theorem 4.2, we have that any monotone and regular divergence between P and Q can be expanded as:

$$\begin{aligned}(4.2)\quad D(P,Q) &= D(\theta,\theta') \\ &= D_0 + D_1 \int \left(\frac{p(x;\theta') - p(x;\theta)}{p(x;\theta)}\right)^2 p(x;\theta)\mu(dx) \\ &\quad + D_2 \int \left(\frac{p(x;\theta') - p(x;\theta)}{p(x;\theta)}\right)^3 p(x;\theta)\mu(dx) + o(|\theta' - \theta|^3).\end{aligned}$$

Moreover,

$$\begin{aligned}p(x;\theta') &= p(x;\theta) + \partial_i p(x;\theta)(\theta'_i - \theta_i) \\ &\quad + \frac{1}{2}\partial_{ij} p(x;\theta)(\theta'_i - \theta_i)(\theta'_j - \theta_j) + o(|\theta' - \theta|^2),\end{aligned}$$

so that

$$\begin{aligned}&\left(\frac{p(x;\theta') - p(x;\theta)}{p(x;\theta)}\right)^2 = \\ &= \left(\frac{\partial_i p(x;\theta)}{p(x;\theta)}(\theta'_i - \theta_i) + \frac{1}{2}\frac{\partial_{ij} p(x;\theta)}{p(x;\theta)}(\theta'_i - \theta_i)(\theta'_j - \theta_j)\right)^2 + o(|\theta' - \theta|^3) \\ &= \partial_i l(x;\theta)\partial_j l(x;\theta)(\theta'_i - \theta_i)(\theta'_j - \theta_j) \\ &\quad + \partial_k l(x;\theta)\left[\partial_{ij} l(x;\theta) + \partial_i l(x;\theta)\partial_j l(x;\theta)\right](\theta'_i - \theta_i)(\theta'_j - \theta_j)(\theta'_k - \theta_k) \\ &\quad + o(|\theta' - \theta|^3),\end{aligned}$$

and

$$\left(\frac{p(x;\theta') - p(x;\theta)}{p(x;\theta)}\right)^3 = \left(\frac{\partial_i p(x;\theta)}{p(x;\theta)}(\theta'_i - \theta_i)\right)^3 + o(|\theta' - \theta|^3)$$

$$= \partial_i l(x;\theta)\partial_j l(x;\theta)\partial_k l(x;\theta)(\theta'_i - \theta_i)(\theta'_j - \theta_j)(\theta'_k - \theta_k) + o(|\theta' - \theta|^3).$$

By substituting in (4.2), we obtain

$$\begin{aligned} D(\theta,\theta') &= D_0 + D_1 g_{ij}(\theta)(\theta'_i - \theta_i)(\theta'_j - \theta_j) \\ &\quad + D_1 \overset{m}{\Gamma}_{ijk}(\theta)(\theta'_i - \theta_i)(\theta'_j - \theta_j)(\theta'_k - \theta_k) \\ &\quad + D_2 T_{ijk}(\theta)(\theta'_i - \theta_i)(\theta'_j - \theta_j)(\theta'_k - \theta_k) + o(|\theta' - \theta|^3) \\ &= D_0 + D_1 \left[g_{ij}(\theta)(\theta'_i - \theta_i)(\theta'_j - \theta_j) + \overset{\alpha}{\Gamma}_{ijk}(\theta)(\theta'_i - \theta_i)(\theta'_j - \theta_j)(\theta'_k - \theta_k) \right] \\ &\quad + o(|\theta' - \theta|^3), \end{aligned}$$

where $\alpha = -\dfrac{D_1 + 2D_2}{D_1}$ and $g_{ij}(\theta)$ and $\overset{\alpha}{\Gamma}_{ijk}(\theta)$ are respectively the Fisher metric and Amari's α-connections of the parametric model, see Amari (1985, pp. 26 and 39).

References

Amari, S. (1985), *Differential-Geometrical Methods in Statistics*, Vol. 28 of *Lecture Notes in Statistics*, Springer Verlag, New York.

Blaesild, P. (1988), Yokes and tensors derived from yokes, Research reports 173, University of Aarhus.

Burbea, J. (1983), *J-divergences and related concepts*, Vol. 4 of *Encyclopedia of Statistical Sciences*, J. Wiley, pp. 290–296.

Campbell, L. (1986), 'An extended Čencov characterization of the information metric', *Proc. Amer. Math. Soc.* **98**, 135–141.

Heyer, H. (1982), *Theory of Statistical Experiments*, Springer-Verlag, New York.

Čencov, N. (1972), *Statistical Decision Rules and Optimal Infererence*, Nauka, Moscow (in Russian, English version, 1982, Math. Monographs, 53, Amer. Math. Soc., Providence).

Relació dels últims Preprints publicats:

- 195 *Separatrix splitting for a high-frequency perturbation of the pendulum.* V. G. Gelfreich. 1991 Mathematics Subject Classification: 34C37, 34C29, 58F30, 70H05. January 1996.
- 196 *A second order Stratonovich differential equation with boundary conditions.* Aureli Alabert and David Nualart. AMS Subject Classification: 60H10, 60H07. January 1996.
- 197 *A Gentzen system equivalent to the BCK-logic.* Romà J. Adillon and Ventura Verdú. Mathematics Subject Classification: 03B20. January 1996.
- 198 *On Gentzen systems associated with the finite linear MV-algebras.* Àngel J. Gil, Antoni Torrens Torrell and Ventura Verdú. Mathematics Subject Classification: 03B22, 03F03, 03G20, 03G99. February 1996.
- 199 *Exponentially small splitting of separatrices under fast quasiperiodic forcing.* Amadeu Delshams, Vassili Gelfreich, Àngel Jorba and Tere M. Seara. AMS classification scheme numbers: 34C37, 58F27, 58F36, 11J25. February 1996.
- 200 *Existence and regularity of the density for solutions to stochastic differential equations with boundary conditions.* Arturo Kohatsu-Higa and Marta Sanz-Solé. AMS 1990 SCI: 60H07, 60H10, 60H99. March 1996.
- 201 *A forcing construction of thin–tall Boolean algebras.* Juan Carlos Martínez. 1991 Mathematics Subject Classification: 03E35, 06E99, 54612. March 1996.
- 202 *Weighted continuous metric scaling.* C. M. Cuadras and J. Fortiana. AMS Subject Classification: 62H25, 62G99. April 1996.
- 203 *Homoclinic orbits in the complex domain.* V.F. Lazutkin and C. Simó. AMS Subject Classification: 58F05. May 1996.
- 204 *Quasivarieties generated by simple MV-algebras.* Joan Gispert and Antoni Torrens. Mathematics Subject Classification: 03B50, 03G99, 06F99, 08C15. May 1996.
- 205 *Regularity of the law for a class of anticipating stochastic differential equations.* Carles Rovira and Marta Sanz-Solé. AMS Subject Classification: 60H07, 60H10. May 1996.
- 206 *Effective computations in Hamiltonian dynamics.* Carles Simó. AMS Subject Classification: 58F05, 70F07. May 1996.
- 207 *Small perturbations in a hyperbolic stochastic partial differential equation.* David Márquez–Carreras and Marta Sanz–Solé. AMS Subject Classification: 60H15, 60H07. May 1996.
- 208 *Transformed empirical processes and modified Kolmogorov-Smirnov tests for multivariate distributions.* A. Cabaña and E. M. Cabaña. AMS Subject Classification: Primary: 62G30, 62G20, 62G10. Secondary: 60G15. June 1996.
- 209 *Anticipating stochastic Volterra equations.* Elisa Alòs and David Nualart. AMS Subject Classification: 60H07, 60H20. June 1996.
- 210 *On the relationship between α-connections and the asymptotic properties of predictive distributions.* J.M. Corcuera and F. Giummolè. AMS 1980 Subjects Classifications: 62F10, 62B10, 62A99. July 1996.
- 211 *Global efficiency.* J.M. Corcuera and J.M. Oller. AMS 1980 Subjects Classifications: 62F10, 62B10, 62A99. July 1996.
- 212 *Intrinsic analysis of the statistical estimation.* J.M. Oller and J.M. Corcuera. AMS 1980 Subjects Classifications: 62F10, 62B10, 62A99. July 1996.