Chapter 11

Contrast Functions Geometry

Contrast functions, called also divergence functions, are distance-like quantities which measure the asymmetric "proximity" of two probability density functions on a statistical manifold or statistical model $\mathcal{S}$. A contrast function, $D(p||q)$, for density functions $p, q \in \mathcal{S}$, is a smooth, non-negative function that vanishes for $p = q$. Eguchi [38, 39, 41] has shown that a contrast function D induces a Riemannian metric by its second order derivatives, and a pair of dual connections by its third order derivatives.

This chapter introduces contrast functionals on statistical manifolds, which are natural extensions of Kullback–Leibler relative entropy from statistical models, and analyzes their corresponding geometric structures and how these interact with the dualistic structure of a statistical manifold. The chapter also investigates the geometry generated by a contrast functional on the space of probability distributions of a statistical model and provides examples of contrast functions.

It has been shown in Chap. 4 that Kullback–Leibler relative entropy is positive, non-degenerate, its first variation along the diagonal $\xi^0 = \xi$ vanishes, and the Hessian along the diagonal defines the Fisher metric.

O. Calin and C. Udrişte, *Geometric Modeling in Probability and Statistics*,
DOI 10.1007/978-3-319-07779-6_11,

The contrast functions mimic the aforementioned properties of the Kullback–Leibler relative entropy. The only difference in the new context is that there are no density functions and no formula of expectation type can be used here.

We overcome this flaw by defining the contrast functions abstractly in two stages: (i) on an open set of $\mathbb{R}^k$; (ii) on a smooth manifold $\mathcal{S}$.

11.1 Contrast Functions on $\mathbb{R}^k$

Consider an open set $\mathbb{E}$ in $\mathbb{R}^k$, and let $\xi_1, \xi_2 \in \mathbb{E}$. A *contrast function* on $\mathbb{E}$ is a smooth function $D(\,\cdot\,\|\,\cdot\,) : \mathbb{E}\times\mathbb{E} \to \mathbb{R}$ satisfying the following properties:

(i) *positive:* $D(\xi_1||\xi_2) \geq 0$, $\forall \xi_1, \xi_2 \in \mathbb{E}$;

(ii) *non-degenerate:* $D(\xi_1||\xi_2) = 0 \iff \xi_1 = \xi_2$;

(iii) *the first variation along the diagonal* $\{\xi_1 = \xi_2\}$ *vanishes:*

$$\partial_{\xi_1^i} D(\xi_1||\xi_2)_{|\xi_1=\xi_2} = \partial_{\xi_2^i} D(\xi_1||\xi_2)_{|\xi_1=\xi_2} = 0;$$

(iv) *the Hessian along the diagonal* $\xi^0 = \xi$

$$g_{ij}(\xi_1) = \partial_{\xi_2^i}\partial_{\xi_2^j} D(\xi_1||\xi_2)_{|\xi_2=\xi_1}$$

is strictly positive definite and smooth with respect to ξ_1.

Some comments regarding the notation are worthy to make. Even if the function $D(\xi_1||\xi_2)$ is not a distance (the symmetry and the triangle inequality are not satisfied), it is a useful distance-like measure of the separation between two points ξ_1, ξ_2. The separation notation is represented by the symbol $||$.

Another observation worthy to make is the redundancy of part (iii) of the definition; this is a consequence of parts (i) and (ii) as follows:

$$\lim_{\epsilon \searrow 0} \frac{D(\xi_1+\epsilon||\xi_1) - D(\xi_1||\xi_1)}{\epsilon} = \lim_{\epsilon \searrow 0} \frac{D(\xi_1+\epsilon||\xi_1)}{\epsilon} \geq 0$$
$$\lim_{\epsilon \nearrow 0} \frac{D(\xi_1+\epsilon||\xi_1) - D(\xi_1||\xi_1)}{\epsilon} = \lim_{\epsilon \nearrow 0} \frac{D(\xi_1+\epsilon||\xi_1)}{\epsilon} \leq 0,$$

which implies the limit equal to 0. We assumed $\xi_1 \in \mathbb{R}$ for the sake of notation simplicity, but the result holds true in multiple dimensions. We note two facts, which are direct consequences of the definition:

(1) *The point ξ_0 is a global minimum of the map $\xi \to D(\xi_0||\xi)$.*

(2) *The quadratic approximation of a contrast function is given by*

$$D(\xi_1||\xi_2) = \frac{1}{2}\sum_{i,j} g_{ij}(\xi_1)(\xi_1^i - \xi_2^i)(\xi_1^j - \xi_2^j) + o(\|\Delta(\xi_1 - \xi_2)\|^2) \tag{11.1.1}$$

when $\xi_2 - \xi_1 \to 0$.

Hence, for any two close enough neighbor vectors $\xi_1, \xi_2 \in \mathbb{E}$, the contrast function is approximated by half the length of their difference measured in the inner product induced by the matrix g_{ij}

$$D(\xi_1||\xi_2) \approx \frac{1}{2}\langle \xi_1 - \xi_2, \xi_1 - \xi_2\rangle_g = \frac{1}{2}\|\xi_1 - \xi_2\|_g^2.$$

In the following we show how a contrast function can be induced by a strictly convex function.

Proposition 11.1.1 *Let $\varphi : \mathbb{E} \to \mathbb{R}$ be a strictly convex function. Then*

$$\begin{aligned} D(\xi_0||\xi) &= \varphi(\xi) - \varphi(\xi_0) - \sum_j \partial_j\varphi(\xi_0)(\xi^j - \xi_0^j) \quad (11.1.2)\\ &= \varphi(\xi) - \varphi(\xi_0) - \langle \partial\varphi(\xi_0), \xi - \xi_0\rangle \end{aligned}$$

is a contrast function on $\mathbb{E}$.

Proof:

(i) Positivity: since the graph of the strictly convex function φ is above the tangent plane at each point, we have

$$\varphi(\xi) \geq \varphi(\xi_0) + \sum_j \partial_j\varphi(\xi_0)(\xi^j - \xi_0^j). \tag{11.1.3}$$

This implies $D(\xi_0||\xi) \geq 0$.

(ii) Non-degenerate: Since the equality in (11.1.3) occurs only for $\xi = \xi_0$, it follows that $D(\xi_0||\xi) = 0$ implies $\xi = \xi_0$.

(iii) Differentiating with respect to ξ_i yields

$$\partial_{\xi_i} D(\xi_0||\xi) = \partial_{\xi_i}\varphi(\xi) - \partial_{\xi_i}\varphi(\xi_0),$$

and hence $\partial_{\xi_i} D(\xi_0||\xi)_{|\xi=\xi_0} = 0$.

(*iv*) Since the function φ is strictly convex, and

$$\partial_{\xi_i}\partial_{\xi_j} D(\xi_0||\xi) = \partial_{\xi_i}\partial_{\xi_j}\varphi(\xi) \tag{11.1.4}$$

it follows that $\partial_{\xi_i}\partial_{\xi_j} D(\xi_0||\xi)$ is strictly positive definite. Hence $D(\xi_0||\xi)$ satisfies the properties of a contrast function.

■

We shall discuss in the following a few particular cases.

Example 11.1.2 (Exponential Model) Consider the convex function $\varphi(\xi) = -\ln\xi$, with $\xi > 0$. The induced contrast function is given by

$$D(\xi_0||\xi) = \frac{\xi}{\xi_0} - \ln\frac{\xi}{\xi_0} - 1,$$

which is exactly the Kullback–Leibler relative entropy for the exponential distribution. It is worth noting that the convex function $\varphi(\xi) = \xi - \ln\xi$ induces the same contrast function. Hence, there is no one-to-one correspondence between convex functions and contrast functions.

Example 11.1.3 The convex function $\varphi(\xi) = \xi^2 - \ln\xi$, with $\xi > 0$, induces the contrast function

$$D(\xi_0||\xi) = (\xi - \xi_0)^2 + \frac{\xi}{\xi_0} - \ln\frac{\xi}{\xi_0} - 1.$$

Example 11.1.4 If consider $\varphi(\xi) = \xi^2$, with $\xi > 0$, the induced contrast function is

$$D(\xi_0||\xi) = (\xi - \xi_0)^2.$$

Not all contrast functions are induced by strictly convex functions. For instance, one can show that

$$D(\xi_0||\xi) = \frac{(\xi - \xi_0)^2}{\xi_0\xi^2}$$

is a contrast function on $(0,\infty)^2$, which cannot be written in the form of formula (11.1.2). We make the note that this contrast function is related to the problem of minimum chi-squared estimator, as described in Kass and Vos [49], p.244. There are many other contrast functions that are not in the form (11.1.2), for instance most

f-divergences, see Sect. 12.2. It can be shown that a contrast function derived from a strictly convex function by formula (11.1.2) is a dually flat contrast function.

It is worth noting that the definition of the contrast function adopted by Kass and Vos [49], p.240, is slightly modified, replacing condition (iv) by the following condition:

(iv') *the matrix*

$$g_{ij}(\xi_1) = \partial_{\xi_1^i}\partial_{\xi_1^j} D(\xi_1||\xi_2)$$

is positive definite and a smooth function of ξ_1 *alone.*

The contrast function given by formula (11.1.2) is sometimes called *Bregman divergence*, see Bregman [20], and it is widely used in convex optimization, see Bauschke [14], Bauschke and Combettes [16], and Bauschke et al. [15].

The term of "contrast function" has been defined slightly different by other authors, and under different names (divergence, yoke, etc.) see Eguchi [40], Rao [72] and Barndorff-Nielsen [11].

11.2 Contrast Functions on a Manifold

Let $\mathcal{S}$ be a smooth manifold. A *contrast function* on $\mathcal{S}$ is a smooth mapping $D_{\mathcal{S}}(\cdot\,\|\,\cdot) : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$, such that any parametrization $\phi : \mathbb{E} \to \mathcal{S}$ makes

$$D(\xi_1||\xi_2) = D_{\mathcal{S}}\big(\phi(\xi_1)||\phi(\xi_2)\big)$$

a contrast function on $\mathbb{E}$. This definition was given for the first time in Amari [5].

We note the local character of a contrast function on a manifold. If $p_1, p_2 \in \mathcal{S}$ belong to the same coordinate chart, there are $\xi_1, \xi_2 \in \mathbb{E}$ such that $\phi(p_i) = \xi_i$ and then we have $D(\xi_1||\xi_2) = D_{\mathcal{S}}\big(p_1||p_2\big)$. Since there might be no coordinate charts to include both points p_1, p_2, then the contrast function $D_{\mathcal{S}}(\cdot\,\|\,\cdot)$ makes sense only locally. In general, there might be no global defined contrast functions on a manifold $\mathcal{S}$.

The invariance of the contrast function with respect to charts is given in the following result.

Theorem 11.2.1 *Consider two local parametrizations* $\phi : \mathbb{E}_\xi \to U$, $\varphi : \mathbb{E}_\eta \to V$ *on the manifold* $\mathcal{S}$. *If*

$$D(\xi_1||\xi_2) = D_{\mathcal{S}}\big(\phi(\xi_1)||\phi(\xi_2)\big)$$

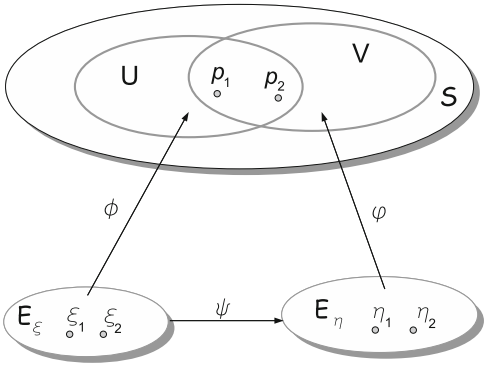

Figure 11.1: The parameterizations ϕ and φ on a manifold $\mathcal{S}$

is a contrast function on the parameter space $\mathbb{E}_\xi$*, then*

$$D(\eta_1||\eta_2) = D_{\mathcal{S}}\big(\varphi(\eta_1)||\varphi(\eta_2)\big)$$

is also a contrast function on the parameter space $\mathbb{E}_\eta$*.*

Proof: For any two points $p_1, p_2 \in U \cap V \subset \mathcal{S}$ denote $p_1 = \phi(\xi_1) = \varphi(\eta_1)$, $p_2 = \phi(\xi_2) = \varphi(\eta_2)$. Let $\psi : \mathbb{E}_\xi \to \mathbb{E}_\eta$, $\psi(\xi) = \eta$ be the change of parametrization map, which is invertible as a composition of invertible maps $\psi = \varphi^{-1} \circ \phi$, see Fig. 11.1.

(*i*) The positivity follows obviously from

$$D(\eta_1||\eta_2) = D_{\mathcal{S}}\big(p_1||p_2\big) = D(\xi_1||\xi_2) \geq 0.$$

(*ii*) To check the non-degeneracy we note that $D(\eta_1||\eta_2) = 0$ implies $D(\xi_1||\xi_2) = 0$, and hence $\xi_1 = \xi_2$, or $\psi^{-1}(\eta_1) = \psi^{-1}(\eta_2)$. Since ψ^{-1} is one-to-one, we obtain $\eta_1 = \eta_2$.

(*iii*) The fact that the first variation along the diagonal $\{\eta_1 = \eta_2\}$ vanishes is a consequence of (*i*) and (*ii*).

(*iv*) We investigate first how does g_{ij} change when changing the parameter ξ into η

$$\begin{aligned} g_{ij}(\xi) &= g(\partial_{\xi^i}, \partial_{\xi^j}) = g\Big(\frac{\partial \eta^r}{\partial \xi^i}\partial_{\eta^r}, \frac{\partial \eta^k}{\partial \xi^j}\partial_{\eta^k}\Big) \\ &= \frac{\partial \eta^r}{\partial \xi^i}\frac{\partial \eta^k}{\partial \xi^j} g(\partial_{\eta^r}, \partial_{\eta^k}) = \frac{\partial \eta^r}{\partial \xi^i}\frac{\partial \eta^k}{\partial \xi^j}\bar{g}_{rk}(\eta), \end{aligned}$$

and hence

$$g_{ij}(\xi) = \frac{\partial \eta^r}{\partial \xi^i} \frac{\partial \eta^k}{\partial \xi^j} \bar{g}_{rk}(\eta). \tag{11.2.5}$$

Consider the points p_1 and p_2 infinitesimally close. Then writing the quadratic approximation formula (11.1.1) in differential form for $D(\xi_1||\xi_2)$ and $D(\eta_1||\eta_2)$ and combining with (11.2.5) and the chain rule yields

$$\begin{aligned} D(\xi_1||\xi_2) &= \frac{1}{2} \sum_{i,j} g_{ij}(\xi_1) d\xi^i d\xi^j \\ &= \frac{1}{2} \sum_{i,j} \sum_{r,k} \bar{g}_{rk}(\eta_1) \frac{\partial \eta^r}{\partial \xi^i} \frac{\partial \eta^k}{\partial \xi^j} d\xi^i d\xi^j \qquad (11.2.6) \\ D(\eta_1||\eta_2) &= \frac{1}{2} \sum_{r,k} h_{rk}(\eta_1) d\eta^r d\eta^k \\ &= \frac{1}{2} \sum_{i,j} \sum_{r,k} h_{rk}(\eta_1) \frac{\partial \eta^r}{\partial \xi^i} \frac{\partial \eta^k}{\partial \xi^j} d\xi^i d\xi^j. \qquad (11.2.7) \end{aligned}$$

Comparing (11.2.6) and (11.2.7) yields $\bar{g}_{rk}(\eta) = h_{rk}(\eta)$. Since $\bar{g}_{rk}(\eta)$ is strictly positive definite, then $h_{rk}(\eta)$ is the same. Hence $D(\eta_1, \eta_2)$ verifies all the conditions of a contrast function.

■

Corollary 11.2.2 *The diagonal part of the Hessians*

$$g_{ij}(\xi_1) = \partial_{\xi_2^i} \partial_{\xi_2^j} D(\xi_1||\xi_2)_{|\xi_2=\xi_1}$$

$$h_{ij}(\eta_1) = \partial_{\eta_2^i} \partial_{\eta_2^j} D(\eta_1||\eta_2)_{|\eta_2=\eta_1}$$

are related by the following relation

$$g_{ij}(\xi_1) = \frac{\partial \eta^r}{\partial \xi^i} \frac{\partial \eta^k}{\partial \xi^j} h_{rk}(\eta_1). \tag{11.2.8}$$

11.3 Induced Riemannian Metric

One of the useful consequences of the invariance property given by Theorem 11.2.1 is that a contrast function provides a unique Riemannian metric on the manifold $\mathcal{S}$. This metric is the inner product $g_p : T_p\mathcal{S} \times T_p\mathcal{S} \to \mathbb{R}$ defined in a particular chart as

$$g_p(\partial_i, \partial_j) = \partial_{\xi_2^i}\partial_{\xi_2^j} D(\xi_1||\xi_2)_{|\xi_2=\xi_1}, \tag{11.3.9}$$

for any coordinate vector fields ∂_i, ∂_j on $\mathcal{S}$ about p.

In the following we shall develop two formulas equivalent with (11.3.9). Consider the notation $\rho(\xi_1, \xi_2) = D(\xi_1||\xi_2)$. By (ii) we have

$$\begin{aligned} \partial_{\xi_1^i}\rho(\xi_1,\xi_2)_{|\xi_1=\xi_2=\xi} &= \partial_{\xi_1^i}\rho(\xi,\xi) = 0 \\ \partial_{\xi_2^i}\rho(\xi_1,\xi_2)_{|\xi_1=\xi_2=\xi} &= \partial_{\xi_2^i}\rho(\xi,\xi) = 0. \end{aligned}$$

Denote $\partial_j = \frac{\partial}{\partial \xi^j}$. Differentiating the function $\varphi(\xi) = \partial_{\xi_1^i}\rho(\xi,\xi)$ with respect to ∂_j we get

$$0 = \partial_j\varphi(\xi) = \partial_{\xi_1^j}\partial_{\xi_1^i}\rho(\xi,\xi) + \partial_{\xi_2^j}\partial_{\xi_1^i}\rho(\xi,\xi),$$

which implies

$$\partial_{\xi_1^j}\partial_{\xi_1^i}\rho(\xi,\xi) = -\partial_{\xi_2^j}\partial_{\xi_1^i}\rho(\xi,\xi). \tag{11.3.10}$$

Differentiating the function $\phi(\xi) = \partial_{\xi_2^i}\rho(\xi,\xi)$ with respect to ∂_j we obtain

$$0 = \partial_j\phi(\xi) = \partial_{\xi_1^j}\partial_{\xi_2^i}\rho(\xi,\xi) + \partial_{\xi_2^j}\partial_{\xi_2^i}\rho(\xi,\xi),$$

which implies

$$\partial_{\xi_2^j}\partial_{\xi_2^i}\rho(\xi,\xi) = -\partial_{\xi_1^j}\partial_{\xi_2^i}\rho(\xi,\xi). \tag{11.3.11}$$

Assuming $\rho(\cdot\,,\cdot)$ smooth enough, the partial derivatives commute and using (11.3.10) and (11.3.11) we arrive at the following equivalent local formulas for the induced Riemannian metric:

$$\begin{aligned} g_{ij}(\xi) &= \partial_{\xi_1^i}\partial_{\xi_1^j} D(\xi_1||\xi_2)_{|\xi_2=\xi_1} && (11.3.12)\\ &= \partial_{\xi_2^i}\partial_{\xi_2^j} D(\xi_1||\xi_2)_{|\xi_2=\xi_1} && (11.3.13)\\ &= -\partial_{\xi_1^i}\partial_{\xi_2^j} D(\xi_1||\xi_2)_{|\xi_2=\xi_1} && (11.3.14)\\ &= -\partial_{\xi_1^j}\partial_{\xi_2^i} D(\xi_1||\xi_2)_{|\xi_2=\xi_1}. && (11.3.15) \end{aligned}$$

Another relation which will be useful in a later section is obtained by differentiating with respect to $\partial_k (= \frac{\partial}{\partial \xi^k})$ in relation (11.3.11) and applying the chain rule

$$\begin{aligned} \partial_k\partial_{\xi_2^j}\partial_{\xi_2^i}\rho(\xi,\xi) &= -\partial_k\partial_{\xi_1^j}\partial_{\xi_2^i}\rho(\xi,\xi) \Longleftrightarrow \\ \partial_{\xi_1^k}\partial_{\xi_2^j}\partial_{\xi_2^i}\rho(\xi,\xi) + \partial_{\xi_2^k}\partial_{\xi_2^j}\partial_{\xi_2^i}\rho(\xi,\xi) &= -\partial_{\xi_1^k}\partial_{\xi_1^j}\partial_{\xi_2^i}\rho(\xi,\xi) \\ &\quad -\partial_{\xi_2^k}\partial_{\xi_1^j}\partial_{\xi_2^i}\rho(\xi,\xi). \end{aligned} \tag{11.3.16}$$

The following notation is adopted for the representation of a vector field X on $\mathcal{S}$ with respect to two local coordinate systems (ξ_1^i) and (ξ_2^i)

$$X_{(\xi_1)} = \sum_i X^i(\xi_1)\partial_{\xi_1^i}, \quad X_{(\xi_2)} = \sum_i X^i(\xi_2)\partial_{\xi_2^i}.$$

We note that for any vector field X we have

$$X_{(\xi_1)}D(\xi_1||\xi_2)_{|\xi_1=\xi_2} = X_{(\xi_2)}D(\xi_1||\xi_2)_{|\xi_1=\xi_2} \quad = \quad 0.$$

Next we provide the global definition of the induced Riemannian metric.

Proposition 11.3.1 *The inner product of two vector fields is given by the following equivalent formulas*

$$\begin{aligned} g(X,Y) &= X_{(\xi_1)}Y_{(\xi_1)}D(\xi_1||\xi_2)_{|\xi_1=\xi_2} \\ &= X_{(\xi_2)}Y_{(\xi_2)}D(\xi_1||\xi_2)_{|\xi_1=\xi_2} \\ &= -X_{(\xi_1)}Y_{(\xi_2)}D(\xi_1||\xi_2)_{|\xi_1=\xi_2} \\ &= -X_{(\xi_2)}Y_{(\xi_1)}D(\xi_1||\xi_2)_{|\xi_1=\xi_2}. \end{aligned}$$

Proof: The proof follows from the bilinearity of g and an application of relations (11.3.12)–(11.3.15). For instance, the first relation can be shown as

$$\begin{aligned} g(X,Y) &= \sum_{i,j} X^iY^j g(\partial_i,\partial_j) \\ &= \sum_{i,j} X^iY^j \partial_{\xi_1^i}\partial_{\xi_1^j}D(\xi_1||\xi_2)_{|\xi_1=\xi_2} \\ &= X_{(\xi_1)}Y_{(\xi_1)}D(\xi_1||\xi_2)_{|\xi_1=\xi_2}. \end{aligned}$$

■

11.4 Dual Contrast Function

If D is a contrast function on $\mathbb{R}^k$, then the associated dual contrast function is defined by

$$D^*(\xi_1||\xi_2) = D(\xi_2||\xi_1).$$

The fact that D^* satisfies properties (i)–(iv) from the definition of a contrast function follows obviously from the fact that D satisfies the

same properties. Similarly, we can define the dual contrast function on a manifold by

$$D^*_{\mathcal{S}}(p||q) = D_{\mathcal{S}}(q||p), \qquad \forall p, q \in \mathcal{S}.$$

It is worthy to note that the contrast functions D and D^* induce the same Riemannian metric on the manifold $\mathcal{S}$. However, the connections induced by D and D^* play a central role in the geometry of contrast functions, as we shall see in the next couple of sections.

11.5 Induced Primal Connection

Let g be the Riemannian metric on $\mathcal{S}$ induced by the contrast function $D_{\mathcal{S}}$. Consider the operator $\nabla^{(D)}$ given by

$$g(\nabla^{(D)}_X Y, Z) \;=\; -X_{(\xi_1)} Y_{(\xi_1)} Z_{(\xi_2)} D(\xi_1||\xi_2)_{|\xi_1=\xi_2}, \quad (11.5.17)$$

for any vector fields X, Y, Z defined on the overlap of the chart neighborhoods associated with the coordinate systems (ξ^i_1) and (ξ^i_2). We shall check that $\nabla^{(D)}$ satisfies the properties of a connection. The $\mathbb{R}$-bilinearity is obvious. Let $f \in \mathcal{F}(\mathcal{S})$ be an arbitrary smooth function. Then

$$g(\nabla^{(D)}_{fX} Y, Z) \;=\; -f X_{(\xi_1)} Y_{(\xi_1)} Z_{(\xi_2)} D(\xi_1||\xi_2)_{|\xi_1=\xi_2} = g(f\nabla^{(D)}_X Y, Z),$$

and dropping the Z-argument implies $\nabla^{(D)}_{fX} Y = f\nabla^{(D)}_X Y$. Next we check Leibniz rule in the second argument

$$\begin{aligned} g(\nabla^{(D)}_X fY, Z) &= -X_{(\xi_1)}(fY_{(\xi_1)}) Z_{(\xi_2)} D(\xi_1||\xi_2)_{|\xi_1=\xi_2} \\ &= -f X_{(\xi_1)} Y_{(\xi_1)} Z_{(\xi_2)} D(\xi_1||\xi_2)_{|\xi_1=\xi_2} \\ &\quad -X_{(\xi_1)}(f)\, Y_{(\xi_1)} Z_{(\xi_2)} D(\xi_1||\xi_2)_{|\xi_1=\xi_2} \\ &= f g(\nabla^{(D)}_X fY, Z) + X_{(\xi_1)}(f) g(Y, Z) \\ &= g(f\nabla^{(D)}_X fY + X(f)Y, Z), \end{aligned}$$

so $\nabla^{(D)}_X fY = f\nabla^{(D)}_X fY + X(f)Y$.

Writing formula (11.5.17) in local coordinates we obtain the components of the linear connection $\nabla^{(D)}$ as in the following

$$\Gamma^{(D)}_{ij,k} = g(\nabla^{(D)}_{\partial_i} \partial_j, \partial_k) = -\partial_{\xi^i_1} \partial_{\xi^j_1} \partial_{\xi^k_2} D(\xi_1||\xi_2)_{|\xi_1=\xi_2}. \quad (11.5.18)$$

The commutativity of the partial derivatives imply $\Gamma^{(D)}_{ij,k} = \Gamma^{(D)}_{ji,k}$, and hence the connection $\nabla^{(D)}$ has zero torsion. We can arrive to the same result in the following equivalent way. Starting from the global definition of the connection and Riemannian metric we write

$$\begin{aligned} g(\nabla^{(D)}_X Y - \nabla^{(D)}_Y X, Z) &= -X_{(\xi_1)} Y_{(\xi_1)} Z_{(\xi_2)} D(\xi_1||\xi_2)_{\xi_1=\xi_2} \\ &\quad + Y_{(\xi_1)} X_{(\xi_1)} Z_{(\xi_2)} D(\xi_1||\xi_2)_{\xi_1=\xi_2} \\ &= -[X,Y]_{(\xi_1)} Z_{(\xi_2)} D(\xi_1||\xi_2)_{\xi_1=\xi_2} \\ &= g([X,Y],Z). \end{aligned}$$

Dropping the Z-argument implies $\nabla^{(D)}_X Y - \nabla^{(D)}_Y X = [X,Y]$, i.e., the torsion of connection $\nabla^{(D)}$ is zero.

11.6 Induced Dual Connection

The dual connection $\nabla^{(D^*)}$ is the connection induced by the dual contrast function D^*, i.e., it is given by

$$\begin{aligned} g(\nabla^{(D^*)}_X Y, Z) &= -X_{(\xi_2)} Y_{(\xi_2)} Z_{(\xi_1)} D^*(\xi_2||\xi_1)_{|\xi_1=\xi_2} \\ &= -X_{(\xi_2)} Y_{(\xi_2)} Z_{(\xi_1)} D(\xi_1||\xi_2)_{|\xi_1=\xi_2}, \end{aligned}$$

for any vector fields X, Y, Z. This can be written locally as

$$\Gamma^{(D^*)}_{ij,k} = g(\nabla^{(D^*)}_{\partial_i} \partial_j, \partial_k) = -\partial_{\xi_2^i} \partial_{\xi_2^j} \partial_{\xi_1^k} D(\xi_1||\xi_2)_{|\xi_1=\xi_2}.$$

Theorem 11.6.1 *The connections $\nabla^{(D)}$ and $\nabla^{(D^*)}$ are torsion-less dual connections.*

Proof: The fact that the torsions vanish follows from the symmetry in the first two indices of the connection components $\Gamma^{(D)}_{ij,k} = \Gamma^{(D)}_{ji,k}$ and $\Gamma^{(D^*)}_{ij,k} = \Gamma^{(D^*)}_{ji,k}$. The duality relation will be shown in local coordinates. Differentiating with respect to $\partial_k = \partial_{\xi^k}$ in relation $g_{ij}(\xi) = -\partial_{\xi_1^i} \partial_{\xi_2^j} D(\xi||\xi)$ we obtain

$$\begin{aligned} \partial_k g_{ij} &= -\partial_{\xi_1^k} \partial_{\xi_1^i} \partial_{\xi_2^j} D(\xi||\xi) \\ &\quad - \partial_{\xi_2^k} \partial_{\xi_1^i} \partial_{\xi_2^j} D(\xi||\xi) \\ &= \Gamma^{(D)}_{ki,j} + \Gamma^{(D^*)}_{kj,i}, \end{aligned}$$

which is equivalent with the duality of D and D^*. ∎

Therefore, a contrast function D on a manifold $\mathcal{S}$ induces a statistical structure $(g, \nabla^{(D)}, \nabla^{(D^*)})$. Hence, $(\mathcal{S}, g, \nabla^{(D)}, \nabla^{(D^*)})$ becomes the statistical manifold induced by the contrast function D.

Proposition 11.6.2 *The Levi–Civita connection of the Riemannian space $(\mathcal{S}, g)$ is given by*

$$\nabla^{(0)} = \frac{1}{2}\big(\nabla^{(D)} + \nabla^{(D^*)}\big).$$

Proof: Since $\nabla^{(D)}$ and $\nabla^{(D^*)}$ have zero torsion, the same applies to $\nabla^{(0)}$. Using the duality relation we show that $\nabla^{(0)}$ is a metrical connection

$$\begin{aligned} Xg(Y,Z) &= \frac{1}{2}Xg(Y,Z) + \frac{1}{2}Xg(Y,Z) \\ &= \frac{1}{2}\Big\{g(\nabla_X^{(D)}Y, Z) + g(Y, \nabla_X^{(D^*)}Z)\Big\} \\ &= \frac{1}{2}\Big\{g(\nabla_X^{(D^*)}Y, Z) + g(Y, \nabla_X^{(D)}Z)\Big\} \\ &= g\Big(\frac{\nabla_X^{(D)}Y + \nabla_X^{(D^*)}Y}{2}, Z\Big) + g\Big(Y, \frac{\nabla_X^{(D)}Z + \nabla_X^{(D^*)}Z}{2}\Big) \\ &= g(\nabla_X^{(0)}Y, Z) + g(Y, \nabla_X^{(0)}Z). \end{aligned}$$

■

11.7 Skewness Tensor

Besides a Riemannian metric g and a pair of dual connections $\nabla^{(D)}$, $\nabla^{(D^*)}$, a contrast function D also induces the skewness tensor by

$$\begin{aligned} C^{(D)}(X,Y,Z) &= g\big(\nabla_X^{(D^*)}Y - \nabla_X^{(D)}Y, Z\big) \\ &= \Big(X_{(\xi_1)}Y_{(\xi_1)}Z_{(\xi_2)} - X_{(\xi_2)}Y_{(\xi_2)}Z_{(\xi_1)}\Big)D(\xi_1||\xi_2)_{|\xi_1=\xi_2}. \end{aligned}$$

In local coordinates this becomes

$$\begin{aligned} C_{ijk}^{(D)} &= \Gamma_{ij,k}^{(D^*)} - \Gamma_{ij,k}^{(D)} \\ &= \partial_{\xi_1^i}\partial_{\xi_1^j}\partial_{\xi_2^k}D(\xi_1||\xi_2)_{|\xi_1=\xi_2} - \partial_{\xi_2^i}\partial_{\xi_2^j}\partial_{\xi_1^k}D(\xi_1||\xi_2)_{|\xi_1=\xi_2}. \end{aligned}$$

In the virtue of identities (11.3.12)–(11.3.15), the tensor $C_{ijk}^{(D)}$ becomes completely symmetric.

11.8 Third Order Approximation of $D(p||\cdot)$

This section will present the third order approximation of a contrast function $D_{\mathcal{S}}$ on a manifold $\mathcal{S}$. Let $p, q \in \mathcal{S}$ be two points in the same chart with coordinates $\xi_1 = \phi^{-1}(p)$ and $\xi_2 = \phi^{-1}(q)$. Denote $\Delta\xi^i = \xi_2^i - \xi_1^i$. The third order approximation of $D_{\mathcal{S}}(p||\cdot)$ about p is given by

$$\begin{aligned} D_{\mathcal{S}}(p||q) &= D_{\mathcal{S}}(p||p) + \partial_{\xi_2^i} D(\xi_1||\xi_2)_{|\xi_1=\xi_2=\xi}\, \Delta\xi^i \\ &\quad + \frac{1}{2}\partial_{\xi_2^i}\partial_{\xi_2^j} D(\xi_1||\xi_2)_{|\xi_1=\xi_2=\xi}\, \Delta\xi^i\Delta\xi^j \\ &\quad + \frac{1}{6}\partial_{\xi_2^i}\partial_{\xi_2^j}\partial_{\xi_2^k} D(\xi_1||\xi_2)_{|\xi_1=\xi_2=\xi}\, \Delta\xi^i\Delta\xi^j\Delta\xi^k + o(\|\Delta\xi\|^2), \end{aligned}$$

where $o(\|\Delta\xi\|^2)$ is a term which converges to 0 faster than $\|\Delta\xi\|^2$ does, as $p \to q$. Since from the definition of a contrast function the first two terms are zero, then

$$D_{\mathcal{S}}(p||q) = \frac{1}{2} g_{ij}(\xi_1)\Delta\xi^i\Delta\xi^j + \frac{1}{6} h_{ijk}(\xi_1)\Delta\xi^i\Delta\xi^j\Delta\xi^k + o(\|\Delta\xi\|^2),$$

where g_{ij} is the induced Riemannian metric. It suffices to compute the coefficients

$$h_{ijk}(\xi_1) = \partial_{\xi_2^i}\partial_{\xi_2^j}\partial_{\xi_2^k} D(\xi_1||\xi_2)_{|\xi_1=\xi_2=\xi}.$$

Writing relation (11.3.16) in terms of the induced connections components, see formula (11.5.18), we have

$$-\Gamma^*_{ij,k} + h_{ijk} = \Gamma_{jk,i} + \Gamma^*_{ik,j}$$

from where

$$\begin{aligned} h_{ijk} &= \Gamma^*_{ij,k} + \Gamma_{jk,i} + \Gamma^*_{ik,j} \\ &= \partial_j g_{ik} + \Gamma^*_{ik,j} \\ &= \partial_k g_{ij} + \Gamma^*_{ij,k}. \end{aligned}$$

The last two identities follow from formula (8.1.2). A similar argument can be used to show also the relation

$$h_{ijk} = \partial_i g_{kj} + \Gamma^*_{jk,i}.$$

This relations imply the total symmetry of h_{ijk}

$$h_{ijk} = h_{ikj} = h_{kji} = h_{jik}.$$

It is worthy to mention that if $D(\cdot\,||\,\cdot)$ induces a dually flat statistical manifold (i.e., $\Gamma = \Gamma^* = 0$), then $h_{ijk} = 0$.

We have seen that any contrast function induces a dualistic structure $(g^{(D)}, \nabla^{(D)}, \nabla^{(D^*)})$ on $\mathcal{S}$. Next we consider the converse implication, which states that any triple (g, ∇, ∇^*), which consists in a metric and two dual torsion-free connections, is induced from a divergence. The divergence can be given locally by

$$D(p||q) = \frac{1}{2} g_{ij}(p)\Delta\xi^i\Delta\xi^j + \frac{1}{6} h_{ijk}(p)\Delta\xi^i\Delta\xi^j\Delta\xi^k, \tag{11.8.19}$$

where $\Delta\xi^i = \xi^i(q) - \xi^i(p)$ and $h_{ijk} = \partial_i g_{kj} + \Gamma^*_{jk,i}$. The existence of a globally defined contrast function is proved in Matumoto [56].

However, the contrast function is not unique. An alternative construction for (11.8.19) is

$$D(p||q) = \frac{1}{2} g_{ij}(p)\Delta\xi^i\Delta\xi^j - \frac{1}{6} h^*_{ijk}(p)\Delta\xi^i\Delta\xi^j\Delta\xi^k,$$

where $h^*_{ijk} = \partial_i g_{jk} + \Gamma^*_{jk,i}$.

11.9 Hessian Geometry

Assume now that there is a local coordinate chart with respect to which the contrast function $D_{\mathcal{S}}$ is induced locally by a convex function φ via formula (11.1.2). We make the remark that it is not necessarily true that there is always a local system of coordinates in which the contrast function is induced by a convex function. However, when this occurs, it defines a dually flat structure of statistical manifold, as we shall see next. This type of contrast function is sometimes called *Bregman divergence*, see Bregman [20], and it is widely used in convex optimization, see Bauschke [14–16]. For a generalization of this contrast function to an α-family, see Zhang [86].

Using (11.1.4) we obtain that the metric is given by the Hessian of the strictly convex potential function φ

$$g_{ij}(\xi) = \partial_{\xi^i}\partial_{\xi^j}\varphi(\xi). \tag{11.9.20}$$

A straightforward computation shows that the components of the induced dual connections $\nabla^{(D)}$ and $\nabla^{(D^*)}$ are given by

$$\Gamma^{(D)}_{ij,k}(\xi) = 0, \qquad \Gamma^{(D^*)}_{ij,k}(\xi) = \partial_{\xi^i}\partial_{\xi^j}\partial_{\xi^k}\varphi(\xi). \tag{11.9.21}$$

A further computation shows that the Riemann curvature tensors are $R = R^* = 0$, i.e., the connections are dually flat.

It is worth noting that there are topological obstructions to the existence of dually flat structures. Ay and Tuschmann [10] proved that if $(\mathcal{S}, g, \nabla, \nabla^*)$ is dually flat and $\mathcal{S}$ is compact, then the first fundamental group $\pi_1(\mathcal{S})$ must be finite.

The skewness tensor is given by the third order derivatives as

$$C^{(D)}_{ijk} = \partial_{\xi^i}\partial_{\xi^j}\partial_{\xi^k}\varphi(\xi).$$

This geometry is commonly referred to in the literature as the *Hessian geometry*. Some authors considered weaker conditions than strictly convexity for the potential function φ, see Shima [74] and Shima and Yagi [75]. For more details on hessian metrics, the reader is referred to Bercu [17] and Corcodel [29].

11.10 Problems

11.1. Let $\gamma : (a, b) \to (M, g)$ be a regular curve, i.e., $\dot\gamma \neq 0$. Define

$$D(s||t) = \int_s^t (t-u)|\dot\gamma(u)|_g^2\, du.$$

Show that $D(\cdot\,||\,\cdot)$ is a contrast function on (a, b).

11.2. Let $\mathcal{S}$ be a statistical model and consider two distributions $p_0, p_1 \in \mathcal{S}$. Define the following curves in $\mathcal{S}$

$$p_t^{(m)} = (1-t)p_0 + tp_1, \quad p_t^{(e)} = C_t p_0^{1-t} p_1^t, \quad 0 \le t \le 1,$$

where C_t is a normalization function. Denote by $g^{(m)}(t)$ and $g^{(e)}(t)$ the Fisher metrics along the aforementioned curves. Let

$$D^{(m)}(p_1||p_0) = \int_0^1 (1-s)g^{(m)}(s)\, ds,$$

$$D^{(e)}(p_1||p_0) = \int_0^1 (1-s)g^{(e)}(s)\, ds.$$

(*a*) Prove that $D^{(m)}(\cdot\,||\,\cdot)$ and $D^{(e)}(\cdot\,||\,\cdot)$ are contrast functions on $\mathcal{S}$.

(*b*) What is the relationship between $D^{(m)}(\cdot\,||\,\cdot)$ and $D^{(e)}(\cdot\,||\,\cdot)$?

11.3. Let (M, g, ∇, ∇^*) be a dually flat statistical manifold and (x^i) and (ζ_α) a pair of dual coordinate systems associated with potentials φ and ψ (i.e., $x^i = \partial_{\zeta_i}\varphi(\zeta)$, $\zeta_j = \partial_{x^j}\psi(x)$). Define $D : M \times M \to \mathbb{R}$ as

$$D(p||q) = \psi\big(x(p)\big) + \varphi\big(\zeta(q)\big) - x^i(p)\zeta_i(q).$$

(*a*) Prove that $D(\,\cdot\,||\,\cdot\,)$ is a contrast function (called the **canonical divergence** of (M, g, ∇, ∇^*)).

(*b*) Find the dual contrast function $D^*(\,\cdot\,||\,\cdot\,)$.

(*c*) Show that for any $p, q, r \in M$ the following relation holds

$$D(p||q) + D(q||r) = D(p||r) - \big(x^i(q) - x^i(p)\big)\big(\zeta_i(q) - \zeta_i(q)\big).$$

(*d*) Let θ be the angle made at q by the ∇-geodesic joining p and q, γ_{pq}, and the ∇^*-geodesic joining q and r, γ^*_{qr}. Show that

$$D(p||q) + D(q||r) = D(p||r) - \|\dot\gamma_{pq}\| \cdot \|\dot\gamma^*_{qr}\| \cos(\pi - \theta).$$

(*e*) If $\theta = \dfrac{\pi}{2}$ show the following Pythagorean relation:

$$D(p||r) = D(p||q) + D(q||r).$$

(*f*) Find the skewness tensor associated with $D(\,\cdot\,||\,\cdot\,)$.

11.4. Consider the Euclidean space $(M, g) = (\mathbb{R}^n, \delta_{ij})$, with $\nabla = \nabla^*$ given by $\nabla_U V = U(V^j)e_j$, for any $U, V \in \mathcal{X}(M)$.

(*a*) Show that the Euclidean coordinates system is self-dual, i.e., $x^i = \zeta_i$.

(*b*) Show that in this case the potential functions are given by

$$\psi(x) = \frac{1}{2}\sum_i (x^i)^2, \qquad \phi(x) = \frac{1}{2}\sum_i (\zeta_i)^2.$$

(*c*) Prove that the canonical divergence is given by $D(p||q) = \frac{1}{2}d^2_E(p, q)$, where $d_E(p, q)$ denotes the Euclidean distance between p and q.

11.5. How many of the previous requirements still hold on a Riemannian manifold (M, g, ∇) with a flat Levi–Civita connection ∇?

11.6. Let (M, g, ∇, ∇^*) be a dually flat statistical manifold, and denote by $D(\cdot\,||\,\cdot)$ the associated canonical divergence. Consider the D-sphere centered at $p \in M$ of radius ρ, defined by

$$S^{(D)} = \{q \in M; D(p||q) = \rho\}.$$

Show that every ∇-geodesic starting at the center p intersects $S^{(D)}$ orthogonally.

11.7. Consider the exponential family $p(x;\xi) = e^{C(x)+\xi^i F_i(x)-\psi(\xi)}$, $x \in \mathcal{X}$, with $\{F_i(x)\}$ linearly independent on $\mathcal{X}$. Define $\eta_j = E_\xi[F_j]$, $1 \le j \le n$.

(*a*) Show that $\eta_j = \partial_j \psi(\xi)$.

(*b*) Prove that (ξ^i) and (η_j) are dual systems of coordinates.

(*c*) Verify that (ξ^i) is a 1-affine coordinate system and (η_j) is a (-1)-affine coordinate system.

(*d*) Let $\varphi(\eta)$ be the potential associated with ξ, i.e., $\xi^j = \partial_{\eta^j}\varphi(\eta)$. Show that $\varphi(\eta) = E_\xi[\ln p_\xi(x) - C(x)]$.

(*e*) Let $H(p)$ be the entropy of distribution p. Validate the relation

$$H(p_\xi) = -\varphi(\xi) - E_\xi[C(x)].$$

(*f*) Let $\widehat{\eta_j} = F_j(x)$. Show that $\hat{\eta}$ is an unbiased estimator for η, and that the covariance matrix provides the Fisher metric, i.e., $V_\eta(\hat{\eta}) = g_{ij}$.

(*g*) Find the contrast function given by the canonical divergence associated with the dual system of coordinates (ξ^i), (η_i). What is its relationship with the Kullback–Leibler relative entropy?

11.8. Consider the statistical model given by the Poisson distribution $p(x;\xi) = e^{-\xi}\frac{\xi^x}{x!}$, $x \in \{0, 1, 2, \dots\}$, $\xi > 0$. Consider $\eta = \xi$ and $\theta = \ln \xi$.

(*a*) Prove that η and θ are dual coordinates.

(*b*) Find the canonical divergence associated with the above dual coordinates.

11.9. Consider the statistical model given by the normal family

$$p(x;\xi) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ \mu \in \mathbb{R}, \sigma > 0.$$

Show that (θ^i) are (η_i) are dual systems of coordinates, where

$$\eta_1 = \mu, \qquad \eta_2 = \mu^2 + \sigma^2$$

$$\frac{\theta^1}{2\theta^2} = -\mu, \qquad \frac{(\theta^1)^2 - 2\theta^2}{4(\theta^2)^2} = \mu^2 + \sigma^2.$$

11.10. Consider the statistical model given by the exponential distribution $p(x;\xi) = \xi e^{-\xi x}$, $x \geq 0$, $\xi > 0$.

(a) Find a pair of dual coordinates on the above statistical model.

(b) Find the potentials ψ and φ associated with the dual coordinates obtained at (a).

(c) Deduct the expression for the Fisher metric.

(d) Find the canonical divergence associated with the dual coordinates obtained at (a).