which gives another proof of the monotonicity. In particular, it satisfies the **additivity**:

$$D^{(\pm 1)}\left(p_{12} \parallel q_{12}\right) = D^{(\pm 1)}\left(p_1 \parallel q_1\right) + D^{(\pm 1)}\left(p_2 \parallel q_2\right) \tag{3.28}$$

for product distributions $p_{12}(x_1, x_2) = p_1(x_1)p_2(x_2)$ and $q_{12}(x_1, x_2) = q_1(x_1)q_2(x_2)$. See § 2.2 for the chain rule and the additivity of the Fisher metric.

3.3 Dually flat spaces

Let (g, ∇, ∇^*) be a dualistic structure on a manifold S. If the connections ∇ and ∇^* are both symmetric ($T = T^* = 0$), then from Theorem 3.3 we see that ∇-flatness and ∇^*-flatness are equivalent. For example, since the α-connections are always symmetric, we have for any statistical model (or more generally for any manifold consisting of finite measures) S and for any real number α that

$$S \text{ is } \alpha\text{-flat} \iff S \text{ is } (-\alpha)\text{-flat}. \tag{3.29}$$

In particular, recalling that an exponential family is 1-flat and that a mixture family is (-1)-flat (§2.3), we now see in addition that they are both (± 1)-flat.

In general, we call (S, g, ∇, ∇^*) a **dually flat space** if both duals ∇ and ∇^* are flat.

Theorem 3.5 *Let (S, g, ∇, ∇^*) be a dually flat space. If a submanifold M of S is autoparallel with respect to either ∇ or ∇^*, then M is a dually flat space with respect to the dualistic structure $(g_M, \nabla_M, \nabla_M^*)$ induced on M by (g, ∇, ∇^*).*

Proof: Suppose M is ∇-autoparallel. Then from Theorem 1.1 (§1.8) we know that ∇_M is flat. Hence by Equation (3.5) the curvature tensor of ∇_M^* is 0. On the other hand, since ∇^* is flat, it is a symmetric connection, and hence its projection ∇_M^* is symmetric also. From the above we see that ∇_M^* is flat, and that M is a dually flat space. The argument for the case when M is ∇^*-autoparallel is similar. ∎

For instance, an m-autoparallel submanifold of an exponential family and an e-autoparallel submanifold of a mixture family are both (± 1)-flat, even though they are no longer exponential nor mixture families in general.

Now let us investigate the general structure of a dually flat space (S, g, ∇, ∇^*). First, from the definition it follows that there exist for S a ∇-affine coordinate system $[\theta^i]$ and a ∇^*-affine coordinate system $[\eta_j]$,[2] for which we let $\partial_i \stackrel{\text{def}}{=} \frac{\partial}{\partial \theta^i}$ and $\partial^j \stackrel{\text{def}}{=} \frac{\partial}{\partial \eta_j}$. Since ∂_i is a ∇-flat vector field and ∂^j is a ∇^*-flat vector field, we see from Theorem 3.2 that $\langle \partial_i, \partial^j \rangle$ is constant on S. From Equation (1.41) which describes the degree of freedom in an affine coordinate system under

[2] By superscripting one of the indices and subscripting the other, we obtain forms which are naturally suited to the use of Einstein's convention

regular affine transformations, we see that for a particular ∇-affine coordinate system $[\theta^i]$, one may choose a corresponding ∇^*-affine coordinate system $[\eta_j]$ such that

$$\langle \partial_i, \partial^j \rangle = \delta_i^j. \tag{3.30}$$

In general, if two coordinate systems $[\theta^i]$ and $[\eta_j]$ for a Riemannian manifold (S, g) satisfy the condition above, we call the coordinate systems **mutually dual** (with respect to g), and call one the **dual coordinate system** of the other. We see then that the Euclidean coordinate system defined in Equation (1.70) is self-dual. In general, there do not exist dual coordinate systems for a Riemannian manifold (S, g). However, if (S, g, ∇, ∇^*) is a dually flat space, then such a pair of coordinate systems exist. Conversely, if for a Riemannian manifold (S, g) there exists such coordinate systems $[\theta^i]$ and $[\eta_j]$, then the connections ∇ and ∇^* for which they are affine are determined, and (S, g, ∇, ∇^*) is a dually flat space.

Let the components of g with respect to $[\theta^i]$ and $[\eta_j]$ be defined by

$$g_{ij} \stackrel{\text{def}}{=} \langle \partial_i, \partial_j \rangle \quad \text{and} \quad g^{ij} \stackrel{\text{def}}{=} \langle \partial^i, \partial^j \rangle. \tag{3.31}$$

By considering the coordinate transformation between $[\theta^i]$ and $[\eta_j]$, we have

$$\partial^j = (\partial^j \theta^i)\partial_i \quad \text{and} \quad \partial_i = (\partial_i \eta_j)\partial^j.$$

From this we see that Equation (3.30) is equivalent to

$$\frac{\partial \eta_j}{\partial \theta^i} = g_{ij} \quad \text{and} \quad \frac{\partial \theta^i}{\partial \eta_j} = g^{ij}, \tag{3.32}$$

and therefore $g_{ij} g^{jk} = \delta_i^k$, which is consistent with Equation (1.23).

Now suppose that we are given mutually dual coordinate systems $[\theta^i]$ and $[\eta_j]$, and consider the following partial differential equation for a function $\psi : S \to \mathbb{R}$:

$$\partial_i \psi = \eta_i. \tag{3.33}$$

We may rewrite this as $\mathrm{d}\psi = \eta_i \mathrm{d}\theta^i$, and a solution exists if and only if $\partial_i \eta_j = \partial_j \eta_i$. Since from Equation (3.32) we see that $\partial_i \eta_j = g_{ij} = \partial_j \eta_i$, in the context of our discussion a solution ψ always exists. From Equations (3.33) and (3.32) we see that

$$\partial_i \partial_j \psi = g_{ij}. \tag{3.34}$$

Hence the second derivatives of ψ form a positive definite matrix, and therefore ψ is a strictly convex function of $[\theta^1, \cdots, \theta^n]$. Similarly, a solution φ to

$$\partial^i \varphi = \theta^i \tag{3.35}$$

exists. In particular, using a solution ψ to Equation (3.33), let

$$\varphi = \theta^i \eta_i - \psi. \tag{3.36}$$

Then we have

$$\mathrm{d}\varphi = \theta^i \mathrm{d}\eta_i + \eta_i \mathrm{d}\theta^i - \mathrm{d}\psi.$$

Substituting $\mathrm{d}\psi = \eta_i \mathrm{d}\theta^i$ into this equation, we obtain $\mathrm{d}\varphi = \theta^i \mathrm{d}\eta_i$, which is equivalent to Equation (3.35). From Equations (3.35) and (3.32) we see that φ satisfies

$$\partial^i \partial^j \varphi = g^{ij}, \tag{3.37}$$

and hence it is a strictly convex function of $[\eta_1, \cdots, \eta_n]$. Furthermore, it follows from the convexity of ψ and Equations (3.33) and (3.36) that for every $q \in S$

$$\varphi(q) = \max_{p \in S} \left\{ \theta^i(p)\eta_i(q) - \psi(p) \right\}. \tag{3.38}$$

Similarly, for every $p \in S$ we have

$$\psi(p) = \max_{q \in S} \left\{ \theta^i(p)\eta_i(q) - \varphi(q) \right\}. \tag{3.39}$$

Sometimes it is more natural to view these relations as

$$\varphi(\eta) = \max_{\theta \in \Theta} \left\{ \theta^i \eta_i - \psi(\theta) \right\} \tag{3.40}$$

$$\psi(\theta) = \max_{\eta \in \mathrm{H}} \left\{ \theta^i \eta_i - \varphi(\eta) \right\}, \tag{3.41}$$

where ψ and φ are simply convex functions defined on convex regions Θ and H in $\mathbb{R}^n$.

In general, those coordinate transformations $[\theta^i] \leftrightarrow [\eta_i]$ which may be expressed in the form given in Equations (3.33) through (3.39) are called **Legendre transformations**, and ψ and φ are called their **potentials**. Note also that

$$\Gamma^*_{ij,k} \overset{\text{def}}{=} \left\langle \nabla^*_{\partial_i} \partial_j, \partial_k \right\rangle = \partial_i \partial_j \partial_k \psi \quad \text{and} \tag{3.42}$$

$$\Gamma^{ij,k} \overset{\text{def}}{=} \left\langle \nabla_{\partial^i} \partial^j, \partial^k \right\rangle = \partial^i \partial^j \partial^k \varphi, \tag{3.43}$$

which are derived from Equation (3.2) combined with $\Gamma_{ij,k} = \Gamma^{*ij,k} = 0$.

We summarize the discussion above in the following theorem.

Theorem 3.6 *Let $[\theta^i]$ be a ∇-affine coordinate system on a dually flat space (S, g, ∇, ∇^*). Then with respect to g there exists a dual coordinate system $[\eta_i]$ of $[\theta^i]$, where $[\eta_i]$ turns out to be a ∇^*-affine coordinate system. These two coordinate systems are related by the Legendre transformation given using potentials ψ and φ in Equations (3.33) through (3.39). In addition, the components of the metric g with respect to these coordinate systems are given by the second derivatives of the potentials as given in Equations (3.34) and (3.37).*

3.4 Canonical divergence

In §3.2, we observed that an arbitrary divergence induces a torsion-free dualistic structure and that the converse statement is also true. However, it should be noted that the correspondence between divergences and dualistic structures is not one-to-one in that infinitely many divergences correspond to one dualistic structure. In this section, we show that a kind of canonical divergence is uniquely defined on a dually flat space.

Let (S, g, ∇, ∇^*) be a dually flat space, on which we are given mutually dual affine coordinate systems $\{[\theta^i], [\eta_i]\}$ and their potentials $\{\psi, \varphi\}$. Given two points $p, q \in S$, let

$$D(p \,\|\, q) \stackrel{\text{def}}{=} \psi(p) + \varphi(q) - \theta^i(p)\eta_i(q). \tag{3.44}$$

Then from Equations (3.38) and (3.39) we see that $D(p \,\|\, q) \geq 0$ and $D(p \,\|\, q) = 0 \Leftrightarrow p = q$. Moreover, it is easy to verify the equations

$$D((\partial_i \partial_j)_p \,\|\, q) = g_{ij}(p) \quad \text{and} \quad D\left(p \,\|\, (\partial^i \partial^j)_q\right) = g^{ij}(q) \tag{3.45}$$

which immediately implies that D is a divergence and induces g. We can also conclude from these equations that $\nabla = \nabla^{(D)}$ and $\nabla^* = \nabla^{(D^*)}$ since we have $\Gamma_{ij,k} = \Gamma^{*ij,k} = 0$ due to the ∇-affinity of $[\theta^i]$ and the ∇^*-affinity of $[\eta_i]$.

On a dually flat space (S, g, ∇, ∇^*), the degrees of freedom of the dual affine coordinate systems $\{[\theta^i], [\eta_i]\}$ and the potentials $\{\psi, \varphi\}$ are expressed by

$$\begin{aligned} \tilde{\theta}^j &= A^j_i \theta^i + B^j, & \tilde{\eta}_j &= C^i_j \eta_i + D_j, \\ \tilde{\psi} &= \psi + D_j \tilde{\theta}^j + c, \text{ and} & \tilde{\varphi} &= \varphi + B^j \tilde{\eta}_j - B^j D_j - c. \end{aligned}$$

Here $[A^j_i]$ is a regular matrix, $[C^i_j]$ is its inverse, $[B^j]$ and $[D_j]$ are real-valued vectors, and c is a real number. These degrees of freedom completely cancel each other in Equation (3.44) so that D is uniquely determined from (S, g, ∇, ∇^*). We call this D the **canonical divergence of (S, g, ∇, ∇^*)** or the **(g, ∇)-divergence** on S for short.[3]

By interchanging the roles of ∇ and ∇^* on S, the roles of $[\theta^i]$ and $[\eta_i]$, and also those of ψ and φ are exchanged, and we find the (g, ∇^*)-divergence D^* to be

$$D^*(p \,\|\, q) = D(q \,\|\, p). \tag{3.46}$$

In addition, if we let M be a submanifold which is autoparallel with respect to either ∇ or ∇^*, and consider the induced dually flat structure $(g_M, \nabla_M, \nabla^*_M)$ (see Theorem 3.5), then it is possible to prove that the (g_M, ∇_M)-divergence D_M is given by the restriction $D_M = D|_{M \times M}$, or in other words

$$D_M(p \,\|\, q) = D(p \,\|\, q) \qquad (\forall p, q \in M). \tag{3.47}$$

[3] In some literature (including the original Japanese edition of this book), the canonical divergence is simply called the divergence.

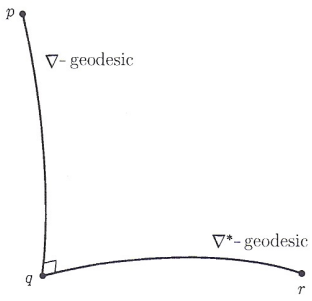

Figure 3.1: The Pythagorean relation for (g, ∇)-divergences.

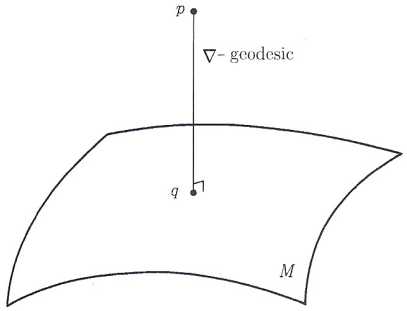

Figure 3.2: The projection theorem of (g, ∇)-divergence

Proof: Since a ∇-geodesic is a straight line with respect to $[\theta^i]$, we may parameterize γ_1 using t as $\theta_t^i = t\theta^i(p) + (1-t)\theta^i(q)$, and obtain $\frac{\mathrm{d}}{\mathrm{d}t}\theta_t^i \partial_i = \left\{\theta^i(p) - \theta^i(q)\right\}\partial_i$ as an expression of the tangent vectors of this curve. Similarly, we may parameterize γ_2 as $\eta_{ti} = t\eta_i(q) + (1-t)\eta_i(r)$, and obtain as its tangent vectors $\frac{\mathrm{d}}{\mathrm{d}t}\eta_{ti}\partial^i = \{\eta_i(q) - \eta_i(r)\}\partial^i$. From Equation (3.30) we see that the inner product of these tangent vectors at the intersection point q may be written as $\left\{\theta^i(p) - \theta^i(q)\right\}\{\eta_i(q) - \eta_i(r)\}$. Therefore when the curves are orthogonal at q, Equation (3.51) follows from Equation (3.49). ∎

The projection theorem given below follows immediately from the theorem above (see Figure 3.2).

Corollary 3.9 *Let p be a point in S and let M be a submanifold of S which is ∇^*-autoparallel. Then a necessary and sufficient condition for a point q in M to satisfy $D(p \parallel q) = \min_{r \in M} D(p \parallel r)$ is for the ∇-geodesic connecting p and q to be orthogonal to M at q.*

The point q in the theorem above is called the **∇-projection of p onto M.** More generally, the following holds for any submanifold M.

Theorem 3.10 *Let p be a point in S and let M be a submanifold of S. A necessary and sufficient condition for a point $q \in M$ to be a stationary point of the function $D(p \,\|\, \cdot) : r \mapsto D(p \,\|\, r)$ restricted on M (in other words, the partial derivatives with respect to a coordinate system of M are all 0) is for the ∇-geodesic connecting p and q to be orthogonal to M at q.*

Proof: Let $\partial_a = \frac{\partial}{\partial u^a}$ be the natural basis of a coordinate system $[u^a]$ of M. Then from Equations (3.44) and (3.35) we have

$$\begin{aligned} D(p \,\|\, (\partial_a)_q) &= (\partial_a \eta_i)_q \, D\left(p \,\middle\|\, (\partial^i)_q\right) \\ &= (\partial_a \eta_i)_q \{\theta^i(q) - \theta^i(p)\} \\ &= \left\langle (\partial_a)_q, \{\theta^i(q) - \theta^i(p)\}(\partial_i)_q \right\rangle, \end{aligned} \tag{3.52}$$

from which the theorem follows. ∎

Corollary 3.11 *Given a point p in S and a positive number c, suppose that the "D-sphere" $M = \{q \in S \mid D(p \,\|\, q) = c\}$ forms a hypersurface in S. Then every ∇-geodesic passing through the center p orthogonally intersects M.*

Before concluding this section, let us take another look at the notion of canonical divergence to illustrate how our geometry modifies the usual Riemannian geometry. Let a Riemannian metric g and an affine connection ∇ be given on a manifold S and let ∇^* be the dual of ∇ with respect to g. We do not assume that (S, g, ∇, ∇^*) is dually flat, and hence the canonical divergence is not generally defined on S. Now, let $\gamma : [a, b] \to S$ $(a < b)$ be a smooth curve in S connecting the points $\gamma(a)$ and $\gamma(b)$, on which the dualistic structure $(g_\gamma, \nabla_\gamma, \nabla^*_\gamma)$ is induced from (g, ∇, ∇^*) by projection. The coefficients of g_γ and ∇_γ corresponding to g_{ab} and $\Gamma^{(\pi)d}_{ab} = \Gamma^{(\pi)}_{ab,c} g^{cd}$ in Equations (1.26) and (1.62) are given by

$$\begin{aligned} g_\gamma(t) &= g_{ij}(\gamma(t))\dot\gamma^i(t)\dot\gamma^j(t), \\ \Gamma_\gamma(t) &= \left\{\dot\gamma^i(t)\dot\gamma^j(t)\Gamma_{ij,k}(\gamma(t)) + \ddot\gamma^j(t) g_{jk}(\gamma(t))\right\} \dot\gamma^k(t)/g_\gamma(t). \end{aligned}$$

Since γ is 1-dimensional, $(\gamma, g_\gamma, \nabla_\gamma, \nabla^*_\gamma)$ is always dually flat and the canonical divergence D_γ is defined on γ. We define $D(\gamma) \stackrel{\text{def}}{=} D_\gamma(\gamma(b) \,\|\, \gamma(a))$ and call it the **(g, ∇)-divergence of the curve γ**. Note that $D(\gamma)$ does not depend on the parametrization $t \mapsto \gamma(t)$ of γ but its orientation, and that the (g, ∇^*)-divergence of γ coincides with the (g, ∇)-divergence of the reversely oriented curve. Some calculation shows that

$$D(\gamma) = \iint_{a \le s \le t \le b} g_\gamma(s) \, \frac{\mu(t)}{\mu(s)} \, \mathrm{d}s \, \mathrm{d}t, \tag{3.53}$$

where

$$\mu(t) \stackrel{\text{def}}{=} \exp\Big[\int_a^t \Gamma_\gamma(s) \, \mathrm{d}s\Big].$$

In particular, if the parameter t is chosen to be ∇_γ-affine, or in other words if $\Gamma_\gamma(t) = 0$ for all t, then we have (cf. Equation (3.45))

$$D(\gamma) = \iint_{a \le s \le t \le b} g_\gamma(s)\, \mathrm{d}s\, \mathrm{d}t = \int_a^b (b-s)\, g_\gamma(s)\, \mathrm{d}s. \tag{3.54}$$

When ∇ is the Riemannian connection, by applying Equation (3.48) to D_γ we see that $D(\gamma) = \frac{1}{2}\|\gamma\|^2$, where $\|\gamma\|$ is the length of γ defined by Equation (1.25). In this respect, the (g, ∇)-divergence of a curve gives a modification of the definition of curve length. Now, suppose that (S, g, ∇, ∇^*) is a dually flat space, on which the canonical divergence D is defined. In the Riemannian case ($\nabla = \nabla^*$), for any geodesic γ we have $\|\gamma\| = d(\gamma(a), \gamma(b))$ and hence $D(\gamma) = D(\gamma(a) \,\|\, \gamma(b)) = D(\gamma(b) \,\|\, \gamma(a))$ by Equation (3.48). (Note that we are only treating local properties and do not consider cases such as cylinders.) On the other hand, in the general dually flat case we have $D(\gamma) = D(\gamma(b) \,\|\, \gamma(a))$ if γ is either a ∇-geodesic or a ∇^*-geodesic (see Equation (3.47)). See Henmi and Kobayashi [108] for an interesting physical interpretation of the canonical divergence.

3.5 The dualistic structure of exponential families

In this section, we investigate the dually flat structure of an exponential family with respect to the (± 1)-connections and the Fisher metric, which are shown to be closely linked to some fundamental aspects of statistics. Let us begin with revisiting the question put at the end of §2.6 about the origin of the 1-flatness of an exponential family, which is now easy to answer. For an exponential family S, its extension $\tilde{S}$ is an 1-affine manifold and hence is 1-flat, and turns out also to be (-1)-flat by the duality. According to Theorem 2.9, S inherits the (-1)-flatness from $\tilde{S}$, and turns out to be 1-flat by the duality again.

We showed in §2.3 that with respect to an exponential family

$$p(x;\theta) = \exp\left[C(x) + \theta^i F_i(x) - \psi(\theta)\right], \tag{3.55}$$

the natural parameters $[\theta^i]$ form a 1-affine coordinate system. Now if we define

$$\eta_i = \eta_i(\theta) \stackrel{\text{def}}{=} E_\theta[F_i] = \int F_i(x) p(x;\theta)\, \mathrm{d}x, \tag{3.56}$$

then from Equations (2.33) and (2.9) we obtain $\eta_i = \partial_i \psi$. Furthermore, from Equations (2.34) and (2.8) we obtain $\partial_i \partial_j \psi = g_{ij}$. Hence $[\eta_i]$ is a (-1)-affine coordinate system dual to $[\theta^i]$, and ψ is the potential of a Legendre transformation. We call this $[\eta_i]$ the **expectation parameters** or the **dual parameters**. For the examples of exponential families given in §2.3, we have the following.

Example 3.1 *(Example 2.5: Normal Distribution)*

$$\eta_1 = \mu = -\frac{\theta^1}{2\theta^2}, \quad \eta_2 = \mu^2 + \sigma^2 = \frac{(\theta^1)^2 - 2\theta^2}{4(\theta^2)^2}$$

Example 3.2 *(Example 2.6: Multivariate Normal Distribution)*

$$\eta_i = \mu_i, \quad \eta_{ij} = (\Sigma)_{ij} + \mu_i \mu_j \quad (i \leq j)$$

$$\eta_A = \mu = -\frac{1}{2}(\theta^B)^{-1}\theta^A,$$

$$\eta_B = \Sigma + \mu\mu^t = -\frac{1}{2}(\theta^B)^{-1} + \frac{1}{4}(\theta^B)^{-1}\theta^A(\theta^A)^t\,(\theta^B)^{-1}$$

Example 3.3 *(Example 2.7: Poisson Distribution)*

$$\eta = \xi = \exp\theta$$

Example 3.4 *(Example 2.8: $\mathcal{P}(\mathcal{X})$ for finite $\mathcal{X}$)*

$$\eta_i = p(x_i) = \xi^i = \frac{\exp\theta^i}{1 + \sum_{j=1}^{n} \exp\theta^j}$$

The dual potential φ in Equation (3.36) is then given by

$$\begin{aligned} \varphi(\theta) &= \theta^i \eta_i(\theta) - \psi(\theta) \\ &= E_\theta[\log p_\theta - C] \\ &= -H(p_\theta) - E_\theta[C], \end{aligned} \tag{3.57}$$

where H is the **entropy**: $H(p) \overset{\text{def}}{=} -\int p(x)\log p(x)\mathrm{d}x$. In addition, from Equation (3.38) we have

$$\varphi(\theta) = \max_{\theta'} \left\{ \theta'^i \eta_i(\theta) - \psi(\theta') \right\}, \tag{3.58}$$

where the maximum is attained by $\theta' = \theta$.

From the definition of the Fisher information matrix (i.e., Equation (2.6)) we have

$$g_{ij}(\theta) = E_\theta[(F_i - \eta_i)(F_j - \eta_j)]. \tag{3.59}$$

Now let us regard the function F_i as an estimator for the parameter η_i and denote it by $\hat{\eta}_i(x) = F_i(x)$. Then Equation (3.56) means that $\hat{\eta} = [\hat{\eta}_1, \cdots, \hat{\eta}_n]$ is an unbiased estimator for the coordinate system $\eta = [\eta_1, \cdots, \eta_n]$, while Equation (3.59) means that the covariance matrix $V_\eta[\hat{\eta}]$ is equal to $G = [g_{ij}]$. It should be noted that G is the Fisher information matrix for the coordinate system $\theta = [\theta^i]$ and, at the same time, is the inverse of the Fisher information matrix for $\eta = [\eta_i]$

by Equations (3.31) and (3.32). Hence, $\hat{\eta}$ attains the equality in the Cramér-Rao inequality (Theorem 2.2), or in other words, $\hat{\eta}$ is an efficient estimator. We have thus seen that an m-affine coordinate system of an exponential family always has an efficient estimator.

Conversely, if a coordinate system $\xi = [\xi^i]$ of a model $S = \{p_\xi\}$ has an efficient estimator, then S is an exponential family and ξ is an m-affine coordinate system composed of expectation parameters. We give a geometrical proof to this statement on the assumption that $\mathcal{X}$ is a finite set. Recall that we have already proved in §2.5 that if there is an efficient estimator $\hat{\xi} = [\hat{\xi}^i]$, then S is an exponential family and there are $n(= \dim S)$ linearly independent e-parallel vector fields on S, say $\{X^1, \cdots, X^n\}$, such that their e-representations are $(X^i_\xi)^{(\mathrm{e})} = \hat{\xi}^i - \xi^i$. Letting $\partial_i = \frac{\partial}{\partial \xi^i}$, we have (cf. Equation (2.49))

$$\langle \partial_i, X^j \rangle = \partial_i E_\xi \left[\hat{\xi}^j \right] = \partial_i \xi^j = \delta^j_i. \tag{3.60}$$

In other words, the inner product between ∂_i and an arbitrary e-parallel vector field is always constant on S. By the duality of e- and m-connections, this means that ∂_i is m-parallel and consequently $[\xi^i]$ is m-affine. An essentially same argument applies to the case when $\mathcal{X}$ is infinite, and we obtain the following theorem.

Theorem 3.12 *A necessary and sufficient condition for a coordinate system ξ of a model $S = \{p_\xi\}$ to have an efficient estimator is that S is an exponential family and ξ is* m*-affine.*

Let us proceed to investigate the canonical divergence. Substituting Equations (3.56) and (3.57) into Equation (3.44) we see that the $(g, \nabla^{(1)})$-divergence on the exponential family $S = \{p_\theta\}$ is given by

$$D^{(1)}(p_\theta \,\|\, p_{\theta'}) = E_{\theta'}[\log p_{\theta'} - \log p_\theta],$$

which is the 1-divergence defined by Equation (3.26), or in other words, the dual of Kullback divergence, and consequently the $(g, \nabla^{(-1)})$-divergence is the Kullback divergence $D^{(-1)}$. The triangular relation (3.49) in this case is essentially equivalent to the following relation for the Kullback divergence $D = D^{(-1)}$:

$$\begin{aligned} &D(p \,\|\, q) + D(q \,\|\, r) - D(p \,\|\, r) \\ &\quad = \int \{p(x) - q(x)\}\{\log r(x) - \log q(x)\} \mathrm{d}x, \end{aligned} \tag{3.61}$$

which is elementary but often useful in applications.

From Corollary 3.9 and Theorem 3.10, the solutions to the minimization problems

$$\min_{q \in M} D(p \,\|\, q) \quad \text{and} \quad \min_{q \in M} D(q \,\|\, p)$$

are respectively given by the $\nabla^{(\mathrm{m})}$-projection and $\nabla^{(\mathrm{e})}$-projection, both of which are important in many applications. The former problem frequently appears in

statistics in connection with the maximum likelihood estimation (see Equation (4.38)), while the latter plays a crucial role in the large deviation theory via Sanov's theorem (see §6.2). An example of the latter problem will be given below, for which we begin with a slightly wider setting as follows.

Given $(n+1)$ functions $C, F_1, \ldots, F_n : \mathcal{X} \to \mathbb{R}$, let $S = \{p_\theta \mid \theta \in \Theta\}$ be the n-dimensional exponential family represented by Equation (3.55). Then for any $\theta \in \Theta$ and any $q \in \mathcal{P}(\mathcal{X})$ we have

$$\begin{aligned} &H(p_\theta) + E_{p_\theta}[C] + \theta^i E_{p_\theta}[F_i] - H(q) - E_q[C] - \theta^i E_q[F_i] \\ &= D(q \,\|\, p_\theta) \geq 0, \end{aligned}$$

which leads to

$$\begin{aligned} &\max_{q \in \mathcal{P}(\mathcal{X})} \{H(q) + E_q[C] + \theta^i E_q[F_i]\} \\ &\quad = H(p_\theta) + E_{p_\theta}[C] + \theta^i E_{p_\theta}[F_i] = \psi(\theta). \end{aligned} \tag{3.62}$$

Given a vector $\lambda = (\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n$, let

$$M_\lambda \stackrel{\text{def}}{=} \{q \in \mathcal{P} \mid E_q[F_i] = \lambda_i,\ i = 1, \ldots, n\}. \tag{3.63}$$

Since M_λ is defined by a linear constraint on the elements, it is a mixture family. Now let us assume that $S \cap M_\lambda \neq \phi$, or equivalently that there exists an element of Θ, say θ_λ, such that $\eta_i(\theta_\lambda) = E_{p_{\theta_\lambda}}[F_i] = \lambda_i$ for $i = 1, \ldots, n$. Then we have:

$$\begin{aligned} \max_{q \in M_\lambda} \{H(q) + E_q[C]\} &= H(p_{\theta_\lambda}) + E_{p_{\theta_\lambda}}[C] \\ &= \psi(\theta_\lambda) - \theta^i_\lambda \lambda_i \\ &= \min_{\theta \in \Theta} \{\psi(\theta) - \theta^i \lambda_i\}, \end{aligned} \tag{3.64}$$

where the first and second equalities follow from Equation (3.62), while the last follows from Equations (3.57) and (3.58).

When $C = 0$ it follows that $\max_{q \in M_\lambda} H(q) = H(p_{\theta_\lambda})$, which is often referred to as the **principle of maximum entropy** . This has the origin in statistical physics; the thermal equilibrium state which maximizes the thermodynamical entropy $S(p) \stackrel{\text{def}}{=} kH(p)$, where $k(> 0)$ is Boltzmann's constant, under the constraint $E_q[\varepsilon] = \bar{\varepsilon}$ on the average of the energy function ε, is given by the Boltzmann-Gibbs distribution

$$p^*(x) = \frac{1}{Z} e^{-\varepsilon(x)/kT},$$

where T is the temperature and Z is the partition function. This corresponds to the previous situation by letting $C = 0$, $n = 1$, $F_i = \varepsilon$, $\lambda = \bar{\varepsilon}$, $\theta_\lambda = -1/kT$ and $\psi(\theta_\lambda) = \log Z$. Assuming $T > 0$, p^* is also characterized as the distribution minimizing Helmholtz's free energy $E_q[\varepsilon] - TS(q)$, which may be regarded as a special case of Equation (3.62).

When $C(x) = \log p(x)$ for a given distribution $p \in \mathcal{P}$, on the other hand, Equation (3.64) may be rewritten as

$$\begin{aligned} \min_{q \in M_\lambda} D(q \,\|\, p) &= D(p_{\theta_\lambda} \,\|\, p) \\ &= \theta^i_\lambda \lambda_i - \psi(\theta_\lambda) = \max_{\theta \in \Theta} \{\theta^i \lambda_i - \psi(\theta)\}, \end{aligned} \tag{3.65}$$

where p_θ and ψ are now represented as

$$p(x;\theta) = p(x) \exp[\theta^i F_i(x) - \psi(\theta)] \quad \text{and} \tag{3.66}$$

$$\psi(\theta) = \log E_p\left[e^{\theta^i F_i} \right]. \tag{3.67}$$

This $\psi(\theta)$ is commonly called the **logarithmic moment generating function** or the **cumulant generating function** of p with respect to the random variables $F_1, \ldots, F_n$. Note that the mixture family M_λ and the exponential family $S = \{p_\theta\}$ intersects orthogonally at p_{θ_λ}, and therefore for all $q \in M_\lambda$ and all θ the following Pythagorean relation holds:

$$D(q \,\|\, p_\theta) = D(q \,\|\, p_{\theta_\lambda}) + D(p_{\theta_\lambda} \,\|\, p_\theta). \tag{3.68}$$

The first equality of Equation (3.65) may be viewed as a consequence of this relation for $\theta = 0$.

Now consider the case when $n = 1$. Given a probability distribution $p \in \mathcal{P}(\mathcal{X})$, a random variable $F : \mathcal{X} \to \mathbb{R}$ and a closed interval $I \subset \mathbb{R}$, Equation (3.65) leads to

$$R(I) \stackrel{\text{def}}{=} \min_{q : E_q[F] \in I} D(q \,\|\, p) = \min_{\lambda \in I} \max_{\theta} \{\theta\lambda - \psi(\theta)\}, \tag{3.69}$$

where $\psi(\theta) = E_p\left[e^{\theta F} \right]$. The probabilistic meaning of this quantity is given by the large deviation theory, which tells us that

$$\lim_{N \to \infty} \frac{1}{N} \log \Pr\left\{ \frac{1}{N} \sum_{t=1}^{N} F(X_t) \in I \right\} = -R(I),$$

where $X_1, X_2, \cdots$ are $\mathcal{X}$-valued random variables that are independent and identically distributed according to $p(x)$. Equation (3.69) may now be considered to be a bridge between two famous large deviation theorems — Sanov's theorem and Cramér's theorem (see e.g. [78]).

We conclude this section by noting that, as examples of Equation (3.54), the Kullback divergence has two mutually dual integral representations. For arbitrary distributions p_0 and p_1, let us define $p_t^{(\mathrm{m})} \stackrel{\text{def}}{=} (1-t)p_0 + tp_1$ and $p_t^{(\mathrm{e})} \stackrel{\text{def}}{=} p_0^{1-t} p_1^t / Z_t$, where Z_t is the normalizing constant. Then $\{p_t^{(\mathrm{m})}\}$ and $\{p_t^{(\mathrm{e})}\}$ form a mixture family and an exponential family, respectively, which

are two different curves connecting p_0 and p_1. Now, letting $g^{(\mathrm{m})}(t)$ and $g^{(\mathrm{e})}(t)$ respectively denote the Fisher informations of $\{p_t^{(\mathrm{m})}\}$ and $\{p_t^{(\mathrm{e})}\}$, we have

$$D\left(p_1 \,\|\, p_0\right) = \iint_{0\le s\le t\le 1} g^{(\mathrm{m})}(s)\,\mathrm{d}s\,\mathrm{d}t = \int_0^1 (1-s)\,g^{(\mathrm{m})}(s)\,\mathrm{d}s \tag{3.70}$$

$$D\left(p_0 \,\|\, p_1\right) = \iint_{0\le s\le t\le 1} g^{(\mathrm{e})}(s)\,\mathrm{d}s\,\mathrm{d}t = \int_0^1 (1-s)\,g^{(\mathrm{e})}(s)\,\mathrm{d}s. \tag{3.71}$$

3.6 The dualistic structure of α-affine manifolds and α-families

Let us turn our attention to the general α-connections and try to extend some of the results in the previous section to their α-versions, using the framework of §2.6. Let us fix α to a particular value and let $S = \{p_\theta\}$ $(\subset \tilde{\mathcal{P}}(\mathcal{X}))$ be an α-affine manifold represented as Equation (2.65). Then as mentioned in §2.6, S is α-flat and $[\theta^i]$ forms an α-affine coordinate system, while from Equation (3.29) we now see that S is also $(-\alpha)$-flat. In other words, $(S, g, \nabla^{(\alpha)}, \nabla^{(-\alpha)})$ is a dually flat space. Now if we define

$$\eta_i \stackrel{\mathrm{def}}{=} \int F_i(x)\ell^{(-\alpha)}(x;\theta)\mathrm{d}x, \tag{3.72}$$

Equations (2.60) and (2.65) lead to

$$\partial_j\eta_i = \int F_i\partial_j\ell^{(-\alpha)}\mathrm{d}x = \int \partial_i\ell^{(\alpha)}\partial_j\ell^{(-\alpha)}\mathrm{d}x = g_{ij}. \tag{3.73}$$

Since this satisfies Equation (3.32), $[\theta^i]$ and $[\eta_i]$ are mutually dual, and hence $[\eta_i]$ is a $(-\alpha)$-affine coordinate system. In addition, letting for an arbitrary $p \in \tilde{\mathcal{P}}(\mathcal{X})$

$$\Psi^{(\alpha)}(p) \stackrel{\mathrm{def}}{=} \begin{cases} \dfrac{2}{1+\alpha}\displaystyle\int p\,(x)\,\mathrm{d}x & (\alpha \neq -1) \\ \displaystyle\int p\,(x)\,\{\log p\,(x) - 1\}\,\mathrm{d}x & (\alpha = -1) \end{cases}, \tag{3.74}$$

we may easily confirm that the potential functions of the Legendre transformation satisfying Equations (3.33) through (3.36) are given by

$$\psi(\theta) = \Psi^{(\alpha)}(p_\theta), \tag{3.75}$$

$$\varphi(\theta) = \Psi^{(-\alpha)}(p_\theta) - \int C(x)\ell^{(-\alpha)}(x;\theta)\mathrm{d}x. \tag{3.76}$$

For arbitrary $p, q \in \tilde{\mathcal{P}}(\mathcal{X})$ and $\alpha \in \mathbb{R}$, let

$$D^{(\alpha)}\left(p \,\|\, q\right) \stackrel{\mathrm{def}}{=} \Psi^{(\alpha)}(p) + \Psi^{(-\alpha)}(q) - \int L^{(\alpha)}(p(x))L^{(-\alpha)}(q(x))\mathrm{d}x. \tag{3.77}$$