Dual Connections

Zhengchao Wan

Peking University

June 21, 2016

Overview

- Duality of connections
- Divergences: general contrast functions
- Dually flat spaces
- Canonical divergence
- f-divergence
- Dualistic structure of exponential families

Riemannian connection

Let M be a manifold on which there is given a Riemannian metric $g = \langle , \rangle$. A connection satisfying

$$Z\langle X, Y\rangle = \langle \nabla_Z X, Y\rangle + \langle X, \nabla_Z Y\rangle \tag{1}$$

for all vector fields $X, Y, Z \in \mathcal{T}(M)$ is called the Riemannian connection.

Dual connection

Giving two connections ∇ and ∇^* on M, if for all X, Y, Z

$$Z\langle X, Y\rangle = \langle \nabla_Z X, Y\rangle + \langle X, \nabla^*_Z Y\rangle \tag{2}$$

holds, then we say that ∇ and ∇^* are **duals** of each other with respect to g, and call one either the **dual connection** or the **conjugate connection** of the other. In addition, we call such a triple (g, ∇, ∇^*) a **dualistic structure** on M.

If ∇ is metric, then $\nabla = \nabla^*$. Hence the duality of connections may be considered as a generalization of metric connection.
In a statistical model S, $(g, \nabla^{(\alpha)}, \nabla^{(-\alpha)})$ is a dualistic structure.

Dual connection

Given a local frame $[x^i]$, from Equation (2) we have

$$\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma^*_{kj,i}. \tag{3}$$

Thus, given g and ∇, there exists a unique dual connection ∇^*.

In addition $(\nabla^*)^* = \nabla$ holds. We also see that $(\nabla + \nabla^*)/2$ becomes a metric connection. And conversely, if a connection ∇' has the same torsion as ∇^* and if $(\nabla + \nabla')/2$ is metric, then $\nabla' = \nabla^*$.

Submanifold

Letting N be a submanifold of M, consider ∇_N and ∇_N^*, which are respectively the projections of ∇ and ∇^* onto N with respect to g. These are dual with respect to g_N (the metric on N determined by g). We call $(g_N, \nabla_N, \nabla_N^*)$ the dualistic structure on N **induced** by (g, ∇, ∇^*), or the **induced dualistic structure** on M.

Covariant derivative

Let $\gamma : t \mapsto \gamma(t)$ be a curve in M and let X and Y be vector fields along γ. In addition, let $D_t X$ and $D_t^* Y$ respectively denote the covariant derivatives of X with respect to ∇ and Y with respect to ∇^*. Then from Equation (2), we see that

$$\frac{d}{dt}\langle X(t), Y(t)\rangle = \langle D_t X(t), Y(t)\rangle + \langle X(t), D_t^* Y(t)\rangle \qquad (4)$$

Now suppose that X is parallel with respect to ∇, and that Y is parallel with respect to ∇^*, i.e., $D_t X = D_t^* Y = 0$. Then $\langle X(t), Y(t)\rangle$ is constant on γ.

Parallel transform

Theorem

Letting P_γ and P_γ^* ($:T_p(M) \to T_q(M)$, where p and q are boundary points of γ) respectively denote the parallel translation along γ with respect ro ∇ and ∇^*, then for all $X, Y \in T_p(M)$ we have

$$\langle P_\gamma(X), P_\gamma^*(Y) \rangle_q = \langle X, Y \rangle_p. \tag{5}$$

This is a generalization of "the invariance of the inner product under parallel translation through metric connections" discussed in Chapter 1.

Curvature

The relationship between P_γ and P_γ^* is completely determined by Equation (5). Hence if P_γ is independent of the actual curve joining p and q, and hence may be written as $P_\gamma = P_{p,q}$, then this is true of P_γ^* also.

Theorem

Letting the curvature tensors of ∇ and ∇^* be denoted by R and R^*, respectively, we have

$$R = 0 \iff R^* = 0. \tag{6}$$

This is immediate from

$$\langle R(X,Y)Z, W\rangle = -\langle R^*(X,Y)W, Z\rangle, \quad \forall X, Y, Z, X \in \mathcal{T}(M). \tag{7}$$

Measure

Consider a smooth function $D = D(\cdot\|\cdot) : M \times M \to R$ satisfying for any $p, q \in M$

$$D(p\|q) \geq 0, \quad \text{and} \quad D(p\|q) = 0 \quad \text{iff} \quad p = q. \tag{8}$$

D is a distance-like measure of the separation between two points. However, it does not in general satisfy the axioms of distance (symmetry and the triangle inequality).

Derivatives

Given an arbitrary coordinate system $[x^i]$ of M, let us represent a pair of points $(p, p^{'}) \in M \times M$ by a pair of coordinates $([x^i], [x^{'i}])$ and denote the partial derivatives of $D(p\|p^{'})$ with respect to p and $p^{'}$ by

$$
\begin{aligned}
D((\partial_i)_p\|p^{'}) &\triangleq (\partial_i)_x D(x\|p^{'})|_x = p \\
D((\partial_i)_p\|(\partial_j)_{p'}) &\triangleq (\partial_i)_x(\partial_j^{'})_y D(x\|y)|_{x=p,y=p'} \\
D((\partial_i\partial_j)_p\|(\partial_k)_{p'}) &\triangleq (\partial_i)_x(\partial_j)_x(\partial_k^{'})_y D(x\|y)|_{x=p,y=p'}, etc.,
\end{aligned}
\tag{9}
$$

These definitions are naturally extended to those of $D((X_1 \cdots X_l)_p\|p^{'}), D(p\|(Y_1 \cdots Y_m)_{p'})$ and $D((X_1 \cdots X_l)_p\|(Y_1 \cdots Y_m)_{p'})$ for any vector fields $X_1, \cdots X_l, Y_1, \cdots, Y_m \in \mathcal{T}(M)$.

Divergence

Now consider the restrictions onto the diagonal $\{(p,p)|p \in M\} \subset M \times M$ and denote the functions induced on M by

$$\begin{aligned} D[X_1 \cdots X_l \| \cdot] : p &\mapsto D((X_1 \cdots X_l)_p \| p), \\ D[\cdot \| Y_1 \cdots Y_m] : p &\mapsto D(p \| (Y_1 \cdots Y_m)_p), \\ D[X_1 \cdots X_l \| Y_1 \cdots Y_m] : p &\mapsto D((X_1 \cdots X_l)_p \| (Y_1 \cdots Y_m)_p). \end{aligned} \tag{10}$$

Easily, we have

$$D[\partial_i \| \cdot] = D[\cdot \| \partial_i] = 0, \tag{11}$$

$$D[\partial_i \partial_j \| \cdot] = D[\cdot \| \partial_i \partial_j] = -D[\partial_i \| \partial_j] (\triangleq g_{ij}^{(D)}). \tag{12}$$

Divergence and Riemannian metric

The matrix $[g_{ij}^{(D)}]$ is positive semidefinite (it's the Hessian matrix of the minimum point). When $[g_{ij}^{(D)}]$ is strictly positive definite everywhere on M, we say that D is a **divergence** or a **contrast function** on M.

For a divergence D, a unique Riemannian metric $g^{(D)} = \langle , \rangle^{(D)}$ on M is defined by $\langle \partial_i, \partial_j \rangle^{(D)} = g_{ij}^{(D)}$, or equivalently by

$$\langle X, Y \rangle^{(D)} = -D[X \| Y]. \tag{13}$$

Using Taylor expansion, we have

$$D(p\|q) = \frac{1}{2} g_{ij}^{(D)}(q) \Delta x^i \Delta x^j + o(\|\Delta x\|^2), \tag{14}$$

Divergence and connection

We define a connection $\nabla^{(D)}$ with the coefficients $\Gamma^{(D)}_{ij,k}$ by

$$\Gamma^{(D)}_{ij,k} = -D[\partial_i \partial_j \| \partial_k], \tag{15}$$

or equivalently by

$$\langle \nabla^{(D)}_X Y, Z \rangle^{(D)} = -D[XY \| Z] \tag{16}$$

It's easy to see that $\Gamma^{(D)}_{ij,k} = \Gamma^{(D)}_{ji,k} \Longleftrightarrow \Gamma^{h(D)}_{ij} = \Gamma^{h(D)}_{ji}$.

Divergence and connection

$$D(p\|q) = \frac{1}{2} g_{ij}^{(D)}(q)\Delta x^i \Delta x^j + \frac{1}{6} h_{ijk}^{(D)}(q)\Delta x^i \Delta x^j \Delta x^k + o(\|\Delta x\|^3), \tag{17}$$

where

$$h_{ijk}^{(D)} \triangleq D[\partial_i \partial_j \partial_k \| \cdot] = \partial_i g_{jk}^{(D)} + \Gamma_{jk,i}^{(D)}. \tag{18}$$

Conversely, we see that $g^{(D)}$ and $\nabla^{(D)}$ are determined by the expansion (17) through Equation (18).

Divergence and dual connection

Replace the divergence $D(p\|q)$ with its **dual** $D^*(p\|q) = D(q\|p)$. Then we obtain $g^{(D^*)} = g^{(D)}$ and

$$\Gamma_{ij,k}^{(D^*)} = -D[\partial_k \| \partial_i \partial_j]. \tag{19}$$

Theorem

$\nabla^{(D)}$ and $\nabla^{(D^*)}$ are dual with respect to $g^{(D)}$.

Divergence and dual connection

$$D(p\|q) = D^*(q\|p)$$
$$= \frac{1}{2} g_{ij}^{(D)}(p)\Delta x^i \Delta x^j - \frac{1}{6} h_{ijk}^{(D^*)}(p)\Delta x^i \Delta x^j \Delta x^k + o(\|\Delta x\|^3) \tag{20}$$

where

$$h_{ijk}^{(D^*)} \triangleq D[\cdot\|\partial_i\partial_j\partial_k] = \partial_i g_{jk}^{(D)} + \Gamma_{jk,i}^{(D^*)}. \tag{21}$$

Dual connection and divergence

We see that any divergence induces a torsion-free dualistic structure. Conversely, any triple (g, ∇, ∇^*) are induced from a divergence.
In fact, if we let

$$D(p\|q) \triangleq \frac{1}{2} g_{ij}(q)\Delta x^i \Delta x^j + \frac{1}{6} h_{ijk}(q)\Delta x^i \Delta x^j \Delta x^k, \qquad (22)$$

where

$$h_{ijk} \triangleq \partial_i g_{jk} + \Gamma_{jk,i} = \Gamma_{ij,k} + \Gamma^*_{ik,j} + \Gamma_{jk,i}, \qquad (23)$$

then $(g, \nabla, \nabla^*) = (g^{(D)}, \nabla^{(D)}, \nabla^{(D^*)})$.

Dually flat spaces

Let (g, ∇, ∇^*) be a dualistic structure on a manifold M. If ∇ and ∇^* are both symmetric ($T = T^* = 0$), then from Theorem before we see that ∇-flatness and ∇^*-flatness are equivalent.

We call (M, g, ∇, ∇^*) a **dually flat space** if both dual connections are flat.

Autoparallel

Theorem

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

Dual coordinate

Suppose $(U; \theta^i, \eta_j)$ is a coordinate neighborhood of dually flat space (M, g, ∇, ∇^*), where $[\theta^i]$ and $[\eta_j]$ denote the affine coordinate system for ∇ and ∇^* respectively. We let $\partial_i \triangleq \frac{\partial}{\partial \theta^i}$ and $\partial^j \triangleq \frac{\partial}{\partial \eta_j}$. $\langle \partial_i, \partial^j \rangle$ is constant on U since they are respectively parallel on flat manifold. Thus we can choose particular coordinate systems such that

$$\langle \partial_i, \partial^j \rangle = \delta_i^j. \tag{24}$$

Such two systems are called **mutually dual**. We see then that the Euclidean coordinate system defined as $\langle \partial_i, \partial_j \rangle = \delta_{ij}$ (affine coordinate) is self-dual.

Dual coordinate

Dual coordinate systems do not generally exist for a Riemannian manifold.

If (M, g, ∇, ∇^*) is a dually flat space, then dual coordinate systems exist. Conversely, if for a Riemannian manifold (M, g) there exists such coordinate systems, then ∇ and ∇^* for which they are affine are determined, and (M, g, ∇, ∇^*) is a dually flat space.

Dual coordinate

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

$$g_{ij} \triangleq \langle \partial_i, \partial_j \rangle \quad \text{and} \quad g^{ij} \triangleq \langle \partial^i, \partial^j \rangle. \tag{25}$$

Considering $\partial^j = (\partial^j \theta^i)\partial_i$ and $\partial_i = (\partial_i \eta_j)\partial^j$, the Equation (24) is equivalent to

$$\frac{\partial \eta_j}{\partial \theta^i} = g_{ij} \quad \text{and} \quad \frac{\partial \theta^i}{\partial \eta_j} = g^{ij}, \tag{26}$$

and therefore $g_{ij} g^{jk} = \delta_i^k$.

Legendre transformations

Suppose we are given mutually dual coordinate systems $[\theta^i]$ and $[\eta_j]$, and consider the following partial differential equation for a function $\psi : M \to R$:

$$\partial_i \psi = \eta_i. \tag{27}$$

Rewrite this as $d\psi = \eta_i d\theta^i$, and a solution exists iff $\partial_i \eta_j = \partial_j \eta_i$. Since $\partial_i \eta_j = g_{ij} = \partial_j \eta_i$, a solution ψ always exists.

$$\partial_i \partial_j \psi = g_{ij}, \tag{28}$$

Hessian matrix of ψ is positve definte, thus it's strictly convex of $[\theta^1, \cdots, \theta^m]$.

Legendre transformations

Similarly, a solution φ to

$$\partial^i \varphi = \theta^i \tag{29}$$

exists. In fact, $\varphi = \theta^i \eta_i - \psi$ is a solution.

$$\partial^i \partial^j \varphi = g^{ij}, \tag{30}$$

and hence it's a strictly convex function of $[\eta_1, \cdots, \eta_m]$.

Legendre transformations

From convexity we have

$$\varphi(q) = \max_{p \in M} \{\theta^i(p)\eta_i(q) - \psi(p)\}, \tag{31}$$

$$\psi(p) = \max_{q \in M} \{\theta^i(p)\eta_i(q) - \varphi(q)\}. \tag{32}$$

Sometimes it is more natural to view these relations as

$$\varphi(\eta) = \max_{\theta \in \Theta} \{\theta^i \eta_i - \psi(\theta)\}, \tag{33}$$

$$\psi(\theta) = \max_{\eta \in H} \{\theta^i \eta_i - \varphi(\eta)\}, \tag{34}$$

where ψ and φ are simply convex functions defined on convex regions Θ and H in R^m.

Legendre transformations

Those coordinate transformations expressed in Equations (27) through (32) are called **Legendre transformations**, and ψ and φ are called their **potentials**.

Note also that

$$\Gamma^*_{ij,k} \triangleq \langle \nabla^*_{\partial_i} \partial_j, \partial_k \rangle = \partial_i \partial_j \partial_k \psi, \tag{35}$$

$$\Gamma^{ij,k} \triangleq \langle \nabla_{\partial^i} \partial^j, \partial^k \rangle = \partial^i \partial^j \partial^k \psi, \tag{36}$$

which are derived from Equation (3) combined with $\Gamma_{ij,k} = \Gamma^{*ij,k} = 0$.

Canonical divergence

Let (M, g, ∇, ∇^*) be a dually flat space, on which we are given mutually dual affine coordinate systems $\{[\theta^i], [\eta_i]\}$. The **canonical divergence** or **(g, ∇)-divergence** is defined as

$$D(p\|q) \triangleq \psi(p) + \varphi(q) - \theta^i(p)\eta_i(q). \tag{37}$$

Then from Equation (31) and (32) we see that $D(p\|q) \leq 0$ and $D(p\|q) = 0 \iff p = q$. It is easy to verify the equations

$$D((\partial_i \partial_j)_p \| q) = g_{ij}(p) \quad \text{and} \quad D(p\|(\partial^i \partial^j)_q) = g^{ij}(q) \tag{38}$$

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

Note 1

The canonical divergence is defined globally, though it uses locally defined charts, which is guaranteed by the following lemma.

Lemma

Suppose M is connected and is flat with respect to ∇, then every two or finite points on M can be contained in a single affine chart.

Note 2

If given another set of dual affine coordinate systems expressed by

$$\tilde{\theta}^j = A^j_i \theta^i + B^j, \quad \tilde{\eta}_j = C^i_j \eta_i + D_j, \tag{39}$$

$$\tilde{\psi} = \psi + D_j \tilde{\theta}^j + c, \quad \tilde{\varphi} = \varphi + B^j \tilde{\eta}_j - B^j D_j - c, \tag{40}$$

where $[A^j_i]$ is a regular matrix and $[C^i_j]$ is its inverse, $[B^j]$ and $[D_j]$ are real-valued vectors, and c is a real number, then we have

$$\psi(p) + \varphi(q) - \theta^i(p)\eta_i(q) = \tilde{\psi}(p) + \tilde{\varphi}(q) - \tilde{\theta}^i(p)\tilde{\eta}_i(q), \tag{41}$$

which indicates that the canonical divergence is well defined. On (M, g, ∇^*, ∇), we define the (g, ∇^*)-divergence $D^*(p\|q) = D(q\|p)$.

Example

If ∇ is a Riemannian connection, the condition for "dually flat" reduces to ∇ being flat, and hence there exists a Euclidean coordinate system $[\theta^i]$, which is self dual ($\theta^i = \eta_i$), and its potential is given by $\psi = \varphi = \frac{1}{2}\sum_i (\theta^i)^2$. Hence we obtain

$$\begin{aligned} D(p\|q) &= \frac{1}{2}\sum_i \{(\theta^i(p))^2 + (\theta^i(q))^2 - 2\theta^i(p)\theta^i(q)\} \\ &= \frac{1}{2}\{d(p,q)\}^2, \end{aligned} \tag{42}$$

where d is the Euclidean distance $d(p,q) \triangleq \sqrt{\sum_i \{\theta^i(p) - \theta^i(q)\}^2}$. In general, $D(p\|q)$ on a dually flat space is only approximately equal to $\frac{1}{2}\{d(p,q)\}^2$ in the sense of Equation (14).

Triangular relation

Theorem

Let $\{[\theta^i], [\eta_i]\}$ *be mutually dual affine coordinate systems of a dually flat space* (M, g, ∇, ∇^*)*, and let* D *be a divergence on* M*. Then a necessary and sufficient condition for* D *to be the* (g, ∇)*-divergence is that for all* $p,q,r \in M$ *the following* **triangular relation** *holds:*

$$D(p\|q) + D(q\|r) - D(p\|r) = \{\theta^i(p) - \theta^i(q)\}\{\eta_i(r) - \eta_i(q)\}. \tag{43}$$

Pythagorean relation

Theorem

Let p,q, and r be three points in M. Let γ_1 be the ∇-geodesic connecting p and q, and let γ_2 be the ∇^*-geodesic connecting q and r. If at the intersection q the curves γ_1 and γ_2 are orthogonal (with respect to the inner product g), then we have the **Pythagorean relation**

$$D(p\|r) = D(p\|q) + D(q\|r) \tag{44}$$

Pythagorean relation

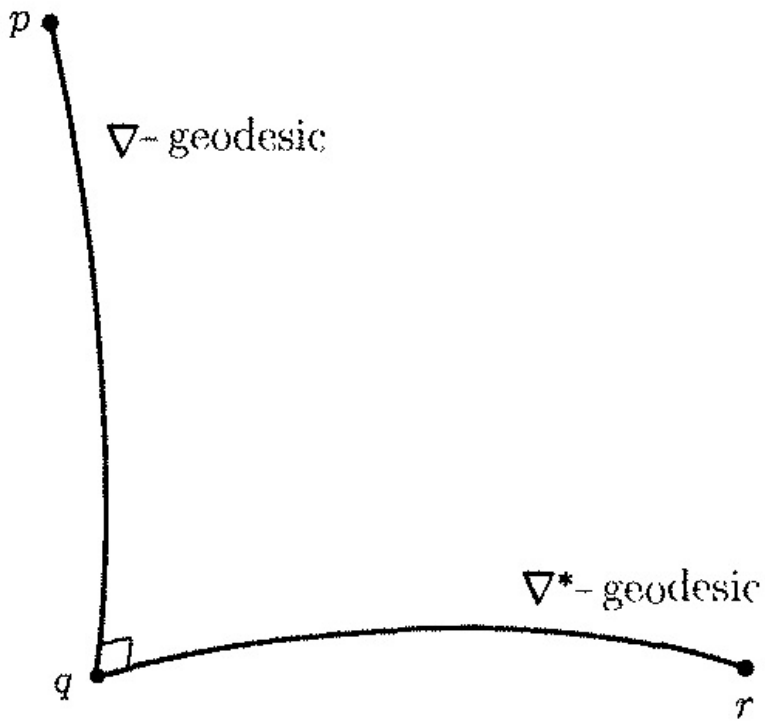

Figure: The Pythagorean relation for (g, ∇)-divergences.

Projection

Corollary

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

The point q is called the **∇-projection of p onto N** when the geodesic connecting p and $q \in N$ is orthogonal to N.

Projection

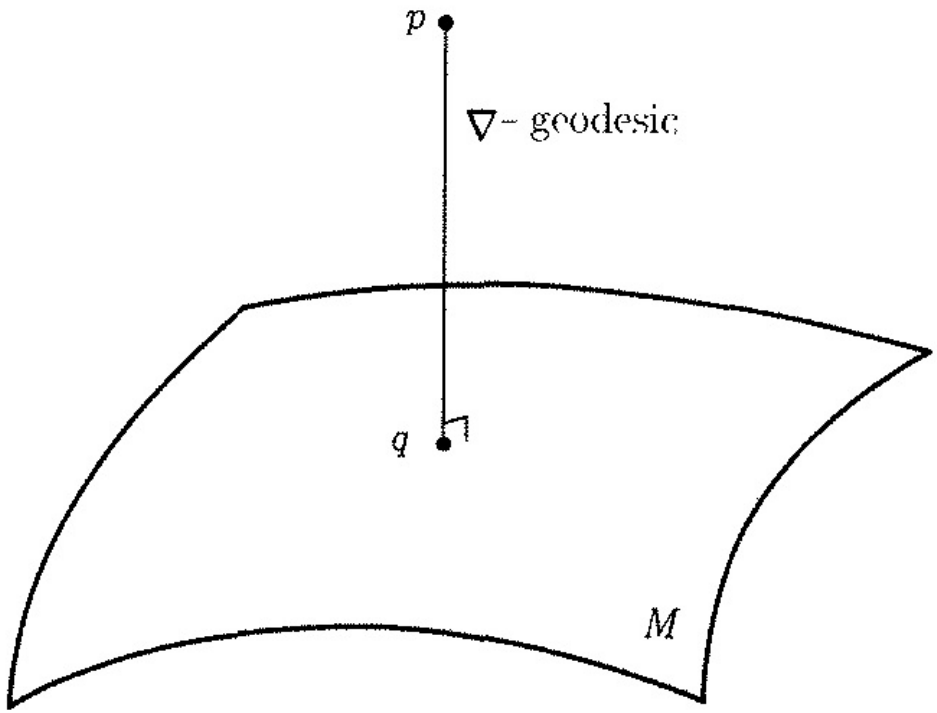

Figure: The projection theorem of (g, ∇)-divergence.

Projection

Theorem

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

Corollary

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

em algorithm

Given two submanifolds K and S in a dually flat M, we define a divergence between K and S by

$$D[K\|S] \triangleq \min_{p\in K, q\in S} D(p\|q) = D(\bar{p}\|\bar{q}), \tag{45}$$

where D is the (g, ∇)-divergence of M and $\bar{p} \in K$ and $\bar{q} \in S$ are the closest pair between K and S.
In order to obtain the closest pair, the following iterative algorithm is proposed.

em algorithm

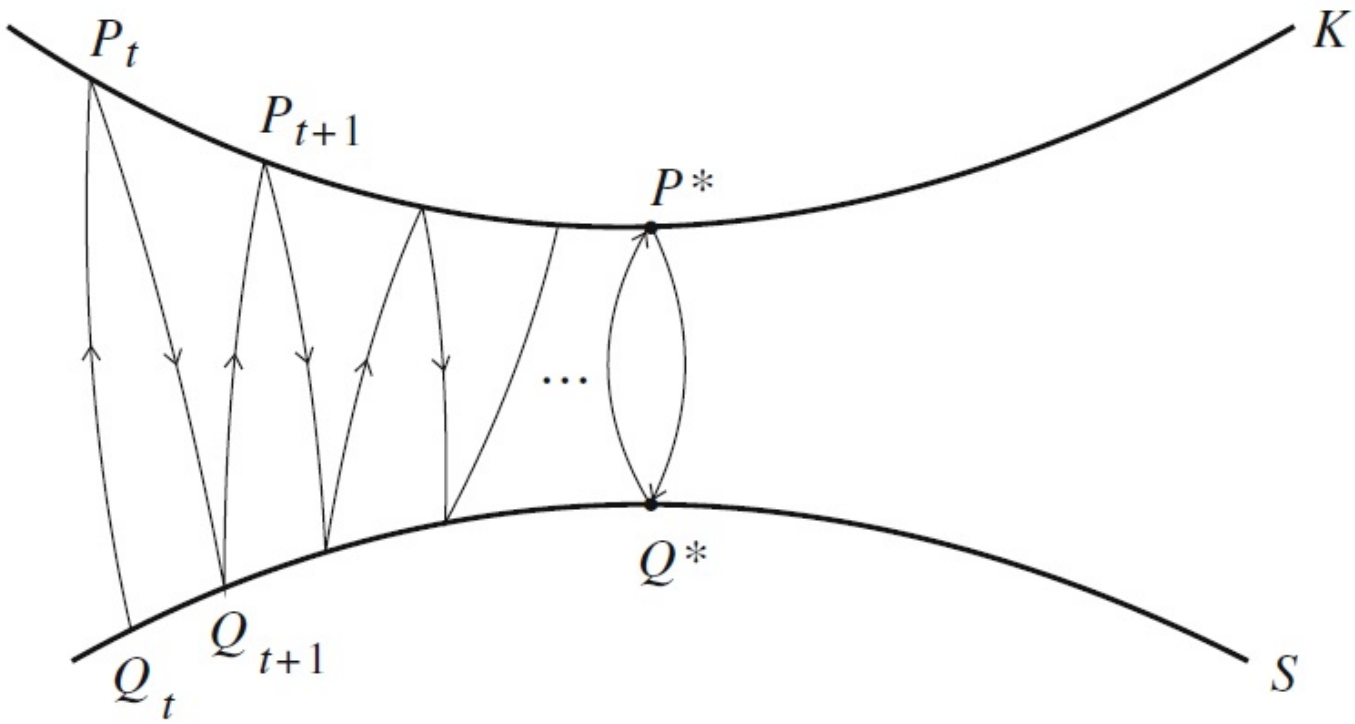

Figure: Iterated dual geodesic projections (em algorithm)

em algorithm

Begin with an arbitrary $Q_t \in S, t = 0, 1, \cdots$ and search for $P \in K$ that minimizes $D(P\|Q_t)$ which is given by the geodesic projection of Q_t to K. Let it be $P_t \in K$. Then search for the point in S that minimizes $D(P_t\|Q)$ which is given by the dual geodesic projection of P_t to S, denoted as Q_{t+1}. Since we have

$$D(P_{t-1}\|Q_t) \geq D(P_t\|Q_t) \geq D(P_t\|Q_{t+1}), \tag{46}$$

the procedure converges. It is unique when S is flat and K is dual flat. Otherwise, the converging point is not necessarily unique.

f-divergence

Let $f(u)$ be a convex function on $u > 0$. For each probability distributions p, q, we define

$$D_f(p\|q) \triangleq \int p(x) f\left(\frac{q(x)}{p(x)}\right) dx \tag{47}$$

and call it the ***f*-divergence**.

Properties of f-divergence

- Using Jensen's inequality we have

$$D_f(p\|q) \geq f\left(\int p(x)\frac{q(x)}{p(x)}dx\right) = f(1), \tag{48}$$

  where the equality holds if $p = q$ and, conversely, the equality implies $p = q$ when $f(u)$ is strictly convex at $u = 1$.
- D_f is invariant when $f(u)$ is replaced with $f(u) + c(u-1)$ for any $c \in R$.

Properties of f-divergence

- $D_f^* = D_{f^*}$, where $f^* = uf(1/u)$.
- **Monotonicity** Let $\kappa = \{\kappa(y|x) \geq 0; x \in \mathcal{X}, y \in \mathcal{Y}\}$ be an arbitrary transition probability distribution such that $\int \kappa(y|x)dy = 1, \forall x$, whereby the value of x is randomly transformed ro y according to the probability $\kappa(y|x)$. Denoting the distributions of y derived from $p(x)$ and $q(x)$ by $p_\kappa(y)$ and $q_\kappa(y)$ respectively, we have

$$D_f(p\|q) \geq D_f(p_\kappa\|q_\kappa) \tag{49}$$

Properties of f-divergence

Proof of monotonicity.

$$\begin{aligned} D_f(p\|q) &= \int\int p(x)\kappa(y|x) f\left(\frac{q(x)}{p(x)}\right) dxdy \\ &= \int\int p_\kappa(y) p_\kappa(x|y) f\left(\frac{q(x)}{p(x)}\right) dxdy \\ &\geq \int p_\kappa(y) f\left(\int p_\kappa(x|y)\frac{q(x)}{p(x)}dx\right) dy \\ &= D_f(p_\kappa\|q_\kappa) \end{aligned} \tag{50}$$

□

The equality holds if $p_\kappa(x|y) = q_\kappa(x|y)$ for all x and y.

Joint convexity

The **joint convexity**

$$\begin{aligned} & D_f(\lambda p_1 + (1-\lambda)p_2 \| \lambda q_1 + (1-\lambda)q_2) \\ & \quad \leq \lambda D_f(p_1 \| q_1) + (1-\lambda) D_f(p_2 \| q_2), \quad 0 \leq \lambda \leq 1 \end{aligned} \tag{51}$$

follows from the convexity of $pf(\frac{q}{p})$

$$((\lambda_1 p_1 + \lambda_2 p_2) f(\frac{\lambda_1 q_1 + \lambda_2 q_2}{\lambda_1 p_1 + \lambda_2 p_2}) =$$

$$(\lambda_1 p_1 + \lambda_2 p_2) f(\frac{\lambda_1 p_1 \frac{q_1}{p_1} + \lambda_2 p_2 \frac{q_2}{p_2}}{\lambda_1 p_1 + \lambda_2 p_2}) \leq \lambda_1 p_1 f(\frac{q_1}{p_1}) + \lambda_2 p_2 f(\frac{q_2}{p_2})).$$

f-divergence

Assume f is strictly convex and smooth and $f(1) = 0$, then D_f becomes a divergence and induces the metric $g^{(D_f)} = g^{(f)}$ and the connection $\nabla^{(D_f)} = \nabla^{(f)}$.

α-divergence

Important examples of smooth f-divergences are given by the **α-divergence** $D^{(\alpha)} = D_{f^{(\alpha)}}$ for a real number α, which is defined by

$$f^{(\alpha)}(u) = \begin{cases} \frac{4}{1-\alpha^2}\{1 - u^{(1+\alpha)/2}\} & (\alpha \neq \pm 1) \\ ulogu & (\alpha = 1) \\ -logu & (\alpha = -1). \end{cases} \tag{52}$$

We have for $\alpha \neq \pm 1$

$$D^{(\alpha)}(p\|q) = \frac{4}{1-\alpha^2}\{1 - \int p(x)^{\frac{1-\alpha}{2}} q(x)^{\frac{1+\alpha}{2}} dx\} \tag{53}$$

α-divergence

and for $\alpha = \pm 1$

$$D^{(-1)}(p\|q) = D^{(1)}(q\|p) = \int p(x) log \frac{p(x)}{q(x)} dx. \tag{54}$$

We can immediately see that the α-divergence $D^{(\alpha)}$ induces $(g^{(f^{(\alpha)})}, \nabla^{(f^{(\alpha)})}) = (g, \nabla^{(\alpha)})$. Note that $D^{(\alpha)}(p\|q) = D^{(-\alpha)}(q\|p)$ generally holds. In particular, $D^{(0)}(p\|q)$ is symmetric, and moreover $\sqrt{D^{(0)}(p\|q)}$ satisfies the axioms of distance, which follows since

$$D^{(0)}(p\|q) = 2 \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx. \tag{55}$$

$\sqrt{D^{(0)}(p\|q)}$ is called the **Hellinger distance**.

Kullback divergence

The ± 1-divergence is called the **Kullback divergence** or **Kullback-Leibler(KL) divergence**. Here we refer to $D^{(-1)}$ as the KL divergence and $D^{(1)}$ its dual. The KL divergence satisfies the chain rule:

$$\begin{aligned} D^{(-1)}(p\|q) =& D^{(-1)}(p_\kappa\|q_\kappa) \\ &+ \int D^{(-1)}(p_\kappa(\cdot|y)\|q_\kappa(\cdot|y))p_\kappa(y)dy. \end{aligned} \tag{56}$$

Expectation parameters

In an exponential family

$$p(x;\theta) = exp[C(x) + \theta^i F_i(x) - \psi(\theta)], \tag{57}$$

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

$$\eta_i = \eta_i(\theta) \triangleq E_\theta[F_i] = \int F_i(x)p(x;\theta)dx, \tag{58}$$

then $\eta_i = \partial_i\psi$ and $\partial_i\partial_j\psi = g_{ij}$. Hence $[\eta_i]$ is a (-1)-affine chart 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**.

Examples

- **Normal Distribution**

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

- **Poisson Distribution**

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

- $\mathcal{P}(\mathcal{X})$ **for finite** $\mathcal{X}$

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

Entropy

The dual potential φ is given by

$$\begin{aligned}\varphi(\eta) &= \theta^i \eta_i - \psi(\theta) \\ &= E_\theta[logp_\theta - C] \\ &= -H(p_\theta) - E_\theta[C],\end{aligned} \tag{59}$$

where H is the **entropy**: $H(p) \triangleq -\int p(x) logp(x) dx$. In addition, we have

$$\varphi(\theta) = \max_{\theta'} \{\theta'^i \eta_i(\theta) - \psi(\theta')\}, \tag{60}$$

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

Canonical divergence

The ± 1-divergence is exactly the canonical $(g, \nabla^{(\pm 1)})$-divergence.

The triangular relation can be rewritten as

$$\begin{aligned} & D(p\|q) + D(q\|r) - D(p\|r) \\ &= \int \{p(x) - q(x)\}\{logr(x) - logq(x)\}dx, \end{aligned} \tag{61}$$

where $D = D^{(-1)}$ is the KL divergence.

Projection

From theorems in canonical divergence, 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 repectively given by the $\nabla^{(m)}$-projection and $\nabla^{(e)}$-projection.

Principle of maximum entropy

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

$$\begin{aligned} &H(p_\theta) + E_{p_\theta}[C] + \theta^i E_{p_\theta}[F_i] - H(q) - E_q[C] - \theta^i E_q[F_i] \\ &\quad = D(q\|p_\theta) \geq 0, \end{aligned} \tag{62}$$

which leads to

$$\begin{aligned} &\max_{q \in \mathcal{P}(\mathcal{X})} \{H(q) + E_q[C] + \theta^i E_q[F_i]\} \\ &\qquad = H(p_\theta) + E_{p_\theta}[C] + \theta^i E_{p_\theta}[F_i] = \psi(\theta). \end{aligned} \tag{63}$$

Principle of maximum entropy

Given a vector $\lambda = (\lambda_1, \cdots, \lambda_n) \in R^n$,let

$$M_\lambda \triangleq \{q \in \mathcal{P} | E_q[F_i] = \lambda_i, i = 1, \cdots, n\}. \tag{64}$$

Now assume $S \cap M_\lambda \neq \emptyset$ and suppose $\theta_\lambda \in \Theta$ s.t. $\eta_i(\theta_\lambda) = E_{p_{\theta_\lambda}}[F_i] = \lambda_i$ for $i = 1, \cdots, n$. Then we have

$$\begin{aligned} \max_{q \in M_\lambda} \{H(q) + E_q[C]\} &= H(p_{\theta_\lambda}) + E_{p_{\theta_\lambda}}[C] \\ &= \psi(\theta_\lambda) - \theta_\lambda^i \lambda_i \\ &= \min_{\theta \in \Theta} \{\psi(\theta) - \theta^i \lambda_i\}, \end{aligned} \tag{65}$$

When $C = 0$ it follows that $max_{q \in M_\lambda} H(q) = H(p_{\theta_\lambda})$, which is called the **principle of maximum entropy**.

Boltzmann-Gibbs distribution

The thermal equilibrium state which maximizes the thermodynamical entropy $S(p) \triangleq kH(p)$, where $k(>0)$ is Boltzmann's constant, under the constraint $E_q[\epsilon] = \bar{\epsilon}$ on the average of the energy function ϵ, is given by the Boltzmann-Gibbs distribution

$$p^*(x) = \frac{1}{Z} e^{-\epsilon(x)/kT}, \tag{66}$$

where T is the temperature and Z is the partition function. This corresponds to the previous situation by letting $C = 0, n = 1, F_i = \epsilon, \lambda = \bar{\epsilon}, \theta_\lambda = -1/kT$ and $\psi(\theta_\lambda) = \log Z$.

Statistical model with hidden variables

Consider a statistical model $M = \{p(\boldsymbol{x}, \boldsymbol{\xi})\}$, where $\boldsymbol{x}$ is divided into two parts $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{h})$ so that $p(\boldsymbol{x}, \boldsymbol{\xi}) = p(\boldsymbol{y}, \boldsymbol{h}; \boldsymbol{\xi})$. When $\boldsymbol{x}$ is not fully observed but $\boldsymbol{y}$ is observed, $\boldsymbol{h}$ is called a hidden variable. In such a case, we estimate $\boldsymbol{\xi}$ from observed $\boldsymbol{y}$. Actually, we want to compute the MLE of $p_Y(\boldsymbol{y}, \boldsymbol{\xi}) = \int p(\boldsymbol{y}, \boldsymbol{h}; \boldsymbol{\xi}) d\boldsymbol{h}$. However, in many cases, the form of $p(\boldsymbol{x}, \boldsymbol{\xi})$ is simple and estimation is tractable in M, but $p_Y(\boldsymbol{y}, \boldsymbol{\xi})$ is complicated and the estimation is computationally intractable.

Empirical distribution

Consider a larger model $S = \{q(\boldsymbol{y}, \boldsymbol{h})\}$ consisting of all probability density functions of $(\boldsymbol{y}, \boldsymbol{h})$. We don't have the empirical distribution $\bar{q}(\boldsymbol{x}) = \frac{1}{N}\sum \delta(\boldsymbol{x} - \boldsymbol{x}_i)$ but only an empirical distribution $\bar{q}_Y(\boldsymbol{y})$ for $\boldsymbol{y}$ only. We use an arbitrary conditional distribution $q(\boldsymbol{h}|\boldsymbol{y})$ and put

$$\bar{q}(\boldsymbol{y}, \boldsymbol{h}) = \bar{q}_Y(\boldsymbol{y})q(\boldsymbol{h}|\boldsymbol{y}). \tag{67}$$

And we take all the candidates of observed points and consider a submanifold

$$D = \{\bar{q}(\boldsymbol{y}, \boldsymbol{h}) | \bar{q}(\boldsymbol{y}, \boldsymbol{h}) = \bar{q}_Y(\boldsymbol{y})q(\boldsymbol{h}|\boldsymbol{y}), q(\boldsymbol{h}|\boldsymbol{y}) \text{ is arbitrary}\}. \tag{68}$$

Empirical distribution

D is the observed submanifold in S specified by the partially observed data $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N$. By using the empirical distribution, it is written as

$$q(\boldsymbol{y}, \boldsymbol{h}) = \frac{1}{N} \sum \delta(\boldsymbol{y} - \boldsymbol{y}_i) q(\boldsymbol{h}|\boldsymbol{y}_i) \tag{69}$$

The data submanifold D is m-flat, because it is linear with respect to $q(\boldsymbol{h}|\boldsymbol{y}_i)$.

MLE and KL-divergence

Consider the minimizer of KL-divergence from data manifold M to the model manifold D,

$$D[D:M] = \min \int \bar{q}_Y(\boldsymbol{y}) q(\boldsymbol{h}|\boldsymbol{y}) \log \frac{\bar{q}_Y(\boldsymbol{y}) q(\boldsymbol{h}|\boldsymbol{y})}{p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{\xi})} d\boldsymbol{y} d\boldsymbol{h} \quad (70)$$

Theorem

The MLE of $p_Y(\boldsymbol{y}, \boldsymbol{\xi})$ is the minimizer of the KL-divergence from D to M.

In fact, we minimize the equation above with respect to both $\boldsymbol{\xi}$ and $q(\boldsymbol{h}|\boldsymbol{y})$ alternately by the *em* algorithm, that is, the alternating use of the e-projection and m-projection.

EM algorithm

Algorithm (EM algorithm)

1. *Choose an initial parameter* ξ_0.
2. **E-step** *e-project* ξ_0 *to* D. *It can be verified that the e-projection is* $q(\boldsymbol{h}|\boldsymbol{y}) = p(\boldsymbol{h}|\boldsymbol{y};\xi_0)$.
3. **M-step** *Maximize a log likelihood*

$$L(\xi, \xi_0) = \frac{1}{N}\sum_i \int p(\boldsymbol{h}|\boldsymbol{y};\xi_0) log p(\boldsymbol{y}_i, \boldsymbol{h}, \xi) d\boldsymbol{h} \qquad (71)$$

   to obtain a new candidate ξ_1 *in* M. *It can be verified that this is the m-projection.*
4. *Repeat step 2 and 3.*

EM algorithm

Theorem

The KL-divergence decreases monotonically by repeating the E-step and the M-step. Hence, the algorithm converges to an equilibrium.

It should be noted that the m-projection is not necessarily unique unless M is e-flat. Hence, there might exist local minima. However, we often come across the exponential family and thus there exists unique solution.