Self-study: Information Metrics and Statistical Divergences

Tianpei Xie

Aug. 19th., 2022

Contents

1 Statistical Divergence

1.1 Definitions

- **Definition** Given a *differentiable manifold* $\mathcal{M}$ of dimension n, a ***divergence*** on $\mathcal{M}$ is a C^2-function $\mathbb{D} : \mathcal{M} \times \mathcal{M} \to [0, \infty)$ satisfying:

  1. (***non-negativity***) $\mathbb{D}\left(p \,\|\, q\right) \geq 0$ for all $p, q \in \mathcal{M}$;
  2. (***positivity***) $\mathbb{D}\left(p \,\|\, q\right) = 0$ if and only if $p = q$;
  3. At every point $p \in \mathcal{M}$, $\mathbb{D}\left(p \,\|\, p + dp\right)$ is a ***positive-definite*** quadratic form for infinitesimal displacements dp from p.

  The last property means that divergence defines an *inner product* on the ***tangent space*** $T_p\mathcal{M}$ for every $p \in \mathcal{M}$. Since $\mathbb{D}$ is C^2 on $\mathcal{M}$, this defines a ***Riemannian metric*** g on $\mathcal{M}$.

- **Definition** Let p, q be $\mathbb{R}^d \supset \mathcal{M}_0 :\to \mathbb{R}$ density functions and let $\alpha \in \mathbb{R} \setminus \{1\}$. The ***Rényi divergence*** of order α or **$\alpha-$divergence** of a distribution p from a distribution q is defined to be

$$\mathbb{D}^{\alpha}\left(p \,\|\, q\right) = \frac{1}{\alpha - 1} \log\left(\mathbb{E}_Q\left[\left(\frac{dP}{dQ}\right)^{\alpha}\right]\right) = \frac{1}{\alpha - 1}\log\left(\int_{\mathcal{M}_0} p^{\alpha}(x) q^{1-\alpha}(x)\, \mu(dx)\right) \tag{1}$$

- **Definition** Let P and Q be two probability distributions over a space Ω, such that $P \ll Q$, that is, P is ***absolutely continuous*** with respect to Q. Then, for a ***convex function*** $f : [0, +\infty) \to (-\infty, +\infty]$ such that $f(x)$ is finite for all $x > 0$, $f(1) = 0$, and $f(0) = \lim_{t \to 0^+} f(t)$ (which could be infinite), the ***f-divergence*** of P from Q is defined as

$$\mathbb{D}^{f}\left(P \,\|\, Q\right) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right] = \int_{\Omega} f\left(\frac{dP}{dQ}\right) dQ = \int_{\Omega} q(x) f\left(\frac{p(x)}{q(x)}\right) \mu(dx) \tag{2}$$

  The convex function f is referred as ***generator function***.

- **Definition** Let $F : \mathcal{X} \to \mathbb{R}$ be a *continuously-differentiable*, ***strictly convex*** function defined on a convex set $\mathcal{X}$. The ***Bregman divergence*** associated with F for points $p, q \in \mathcal{X}$ is the difference between the value of F at point p and the value of the *first-order Taylor expansion* of F around point q evaluated at point p:

$$\mathbb{D}^{F}\left(p \,\|\, q\right) = F(p) - F(q) - \left\langle \nabla F(q)\,,\, p - q \right\rangle \tag{3}$$

- **Definition** We suppose $\mathcal{X} = \mathcal{Y}$ and that for some $p \geq 1$, $c(x, y) = d(x, y)^p$, where d is a distance on $\mathcal{X}$, the **p-Wasserstein distance** between measures α, β on $\mathcal{X}$ is $\mathcal{W}_p(\alpha, \beta)$, where

$$(\mathcal{W}_p(\alpha, \beta))^p := \min_{\substack{(X,Y)\sim\pi;\\ X_{\#}\pi = \alpha,\\ Y_{\#}\pi=\beta}} \mathbb{E}_{(X,Y)}\left[d(X, Y)^p\right] \tag{4}$$

1.2 KL Divergence for Exponential Families

- The canonical representation of ***exponential famlity*** of distribution has the following form

$$
\begin{aligned}
p(x_1, \ldots, x_m) = p(\boldsymbol{x}; \boldsymbol{\eta}) &= \exp\left(\langle \boldsymbol{\eta}\,,\, \boldsymbol{\phi}(\boldsymbol{x}) \rangle - A(\boldsymbol{\eta})\right) h(\boldsymbol{x}) \nu(d\boldsymbol{x}) \\
&= \exp\left(\sum_{\alpha} \eta_\alpha \phi_\alpha(\boldsymbol{x}) - A(\boldsymbol{\eta})\right) \qquad (5)
\end{aligned}
$$

  where ϕ is a feature map and $\boldsymbol{\phi}(\boldsymbol{x})$ defines a set of ***sufficient statistics*** (or ***potential functions***). The normalization factor is defined as

$$
A(\boldsymbol{\eta}) := \log \int \exp\left(\langle \boldsymbol{\eta}\,,\, \boldsymbol{\phi}(\boldsymbol{x}) \rangle\right) h(\boldsymbol{x}) \nu(d\boldsymbol{x}) = \log Z(\boldsymbol{\eta})
$$

  $A(\boldsymbol{\eta})$ is also referred as ***log-partition function*** or *cumulant function*. The parameters $\boldsymbol{\eta} = (\eta_\alpha)$ are called ***natural parameters*** or *canonical parameters*. The canonical parameter $\{\eta_\alpha\}$ forms a **natural (canonical) parameter space**

$$
\Omega = \left\{ \boldsymbol{\eta} \in \mathbb{R}^d : A(\boldsymbol{\eta}) < \infty \right\} \qquad (6)
$$

- The exponential family is the unique solution of ***maximum entropy estimation*** problem:

$$
\begin{aligned}
\min_{q \in \Delta} \quad & \mathbb{KL}\left(q \,\|\, p_0\right) & (7)\\
\text{s.t.} \quad & \mathbb{E}_q\left[\phi_\alpha(X)\right] = \mu_\alpha \quad \forall\, \alpha \in \mathcal{I} & (8)
\end{aligned}
$$

  where $\mathbb{KL}\left(q \,\|\, p_0\right) = \int \log(\frac{q}{p_0}) q dx = \mathbb{E}_q\left[\log \frac{q}{p_0}\right]$ is the relative entropy or the Kullback-Leibler divergence of q w.r.t. p_0.

  Here $\boldsymbol{\mu} = (\mu_\alpha)_{\alpha \in \mathcal{I}}$ is a set of ***mean parameters***. The space of mean parameters $\mathcal{M}$ is a *convex polytope* spanned by potential functions $\{\phi_\alpha\}$.

$$
\mathcal{M} := \left\{ \boldsymbol{\mu} \in \mathbb{R}^d : \exists q \text{ s.t. } \mathbb{E}_q\left[\phi_\alpha(X)\right] = \mu_\alpha \quad \forall\, \alpha \in \mathcal{I} \right\} = \operatorname{conv}\left\{\phi_\alpha(x),\ x \in \mathcal{X},\ \alpha \in \mathcal{I}\right\} \qquad (9)
$$

- Moreover $A(\boldsymbol{\eta})$ has a variational form

$$
A(\boldsymbol{\eta}) = \sup_{\boldsymbol{\mu} \in \mathcal{M}} \left\{ \langle \boldsymbol{\eta}\,,\, \boldsymbol{\mu} \rangle - A^*(\boldsymbol{\mu}) \right\} \qquad (10)
$$

  where $A^*(\boldsymbol{\mu})$ is the conjugate dual function of A and it is defined as

$$
A^*(\boldsymbol{\mu}) := \sup_{\boldsymbol{\eta} \in \Omega} \left\{ \langle \boldsymbol{\mu}\,,\, \boldsymbol{\eta} \rangle - A(\boldsymbol{\eta}) \right\} \qquad (11)
$$

  It is shown that $A^*(\boldsymbol{\mu}) = -H(q_{\boldsymbol{\eta}(\boldsymbol{\mu})})$ for $\boldsymbol{\mu} \in \mathcal{M}^\circ$ which is the negative entropy. $A^*(\boldsymbol{\mu})$ is also the optimal value for the **maximum likelihood estimation** problem on p. The exponential family can be reparameterized according to its mean parameters $\boldsymbol{\mu}$ via backward mapping $(\nabla A)^{-1} : \mathcal{M}^\circ \to \Omega$, called **mean parameterization**.

- We can formulate the **KL divergence** between two distributions in exponential family Ω using its primal and dual form

– **Primal-form**: given $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \Omega$

$$\begin{aligned}\mathbb{KL}\left(p_{\boldsymbol{\eta}_1} \,\|\, p_{\boldsymbol{\eta}_2}\right) \equiv \mathbb{KL}\left(\boldsymbol{\eta}_1 \,\|\, \boldsymbol{\eta}_2\right) &= A(\boldsymbol{\eta}_2) - A(\boldsymbol{\eta}_1) - \langle \boldsymbol{\mu}_1 \,, \boldsymbol{\eta}_2 - \boldsymbol{\eta}_1 \rangle \\ &\equiv A(\boldsymbol{\eta}_2) - A(\boldsymbol{\eta}_1) - \langle \nabla A(\boldsymbol{\eta}_1) \,, \boldsymbol{\eta}_2 - \boldsymbol{\eta}_1 \rangle\end{aligned} \tag{12}$$

– **Primal-dual form**: given $\boldsymbol{\mu}_1 \in \mathcal{M}, \boldsymbol{\eta}_2 \in \Omega$

$$\mathbb{KL}\left(\boldsymbol{\mu}_1 \,\|\, \boldsymbol{\eta}_2\right) = A(\boldsymbol{\eta}_2) + A^*(\boldsymbol{\mu}_1) - \langle \boldsymbol{\mu}_1 \,, \boldsymbol{\eta}_2 \rangle \tag{13}$$

– **Dual-form**: given $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathcal{M}$

$$\begin{aligned}\mathbb{KL}\left(\boldsymbol{\mu}_1 \,\|\, \boldsymbol{\mu}_2\right) &= A^*(\boldsymbol{\mu}_1) - A^*(\boldsymbol{\mu}_2) - \langle \boldsymbol{\eta}_2 \,, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \rangle \\ &\equiv A^*(\boldsymbol{\mu}_1) - A^*(\boldsymbol{\mu}_2) - \langle \nabla A^*(\boldsymbol{\mu}_2) \,, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \rangle\end{aligned} \tag{14}$$

- The dual form is related to the *Bregman divergence*, which induce the **projection operation**. We see that dual form $\mathbb{KL}\left(\boldsymbol{\mu}_1 \,\|\, \boldsymbol{\mu}_2\right) = \mathbb{D}^{A^*}\left(\boldsymbol{\mu}_1 \,\|\, \boldsymbol{\mu}_2\right)$, where $F = A^*$ is the negative entropy.

1.3 α-Divergence Properties

See papers in [Hero et al., 2001, Nielsen and Nock, 2011, Póczos and Schneider, 2011].

- $\mathbb{D}^{\alpha}\left(p \,\|\, q\right) = \mathbb{D}^{1-\alpha}\left(q \,\|\, p\right)$
- $\frac{\alpha}{1-\alpha}\mathbb{D}^{1-\alpha}\left(p \,\|\, q\right) = \mathbb{D}^{\alpha}\left(q \,\|\, p\right)$
- If $\alpha = -1$, $\mathbb{D}^{(-1)}\left(p \,\|\, q\right) = \mathbb{D}^{(1)}\left(q \,\|\, p\right) = \mathbb{KL}\left(p \,\|\, q\right) \equiv \int_x p(x) \log \frac{p(x)}{q(x)} dx$ is the **Kullback-Leibler divergence**.
- For $p_{\boldsymbol{\eta}_1}, q_{\boldsymbol{\eta}_2}$ exponential families, α-divergence has closed form expression:

$$\mathbb{D}^{\alpha}\left(p_{\boldsymbol{\eta}_1} \,\|\, q_{\boldsymbol{\eta}_2}\right) = \frac{1}{1-\alpha}\left(\alpha A(\boldsymbol{\eta}_1) + (1-\alpha)A(\boldsymbol{\eta}_2) - A(\alpha\boldsymbol{\eta}_1 + (1-\alpha)\boldsymbol{\eta}_2)\right) \tag{15}$$

  where $A(\boldsymbol{\eta})$ is the ***log-partition function*** or *cumulant function*.

1.4 f-Divergence Properties

For more details see tutorials in [Csiszár et al., 2004, Liese and Vajda, 2006] and see lecture notes in [Polyanskiy and Wu, 2014].

- $\mathbb{D}^{f_1+f_2}\left(p \,\|\, q\right) = \mathbb{D}^{f_1}\left(p \,\|\, q\right) + \mathbb{D}^{f_2}\left(p \,\|\, q\right)$
- $\mathbb{D}^{f}\left(p \,\|\, q\right) = \mathbb{D}^{g}\left(p \,\|\, q\right)$ if $f(x) = g(x) + c(x-1)$ for some $c \in \mathbb{R}$
- ***Reversal by convex inversion***: for any function f, its ***convex inversion*** is defined as $g(t) := tf(1/t)$. If f satisfies condition for f-divergence, then g satisfies the condition as well and $\mathbb{D}^{g}\left(Q \,\|\, P\right) = \mathbb{D}^{f}\left(P \,\|\, Q\right)$.
- ***Data processing inequality***: if κ is an arbitrary transition probability that transforms measures P and Q into P_κ and Q_κ correspondingly, then

$$\mathbb{D}^{f}\left(P \,\|\, Q\right) \geq \mathbb{D}^{f}\left(P_\kappa \,\|\, Q_\kappa\right). \tag{16}$$

The equality here holds if and only if the transition is induced from a ***sufficient statistic*** with respect to $\{P, Q\}$.

- ***Joint Convexity***: for any $0 \le \lambda \le 1$,

$$\mathbb{D}^f\left(\lambda P_1 + (1-\lambda)P_2 \,\|\, \lambda Q_1 + (1-\lambda)Q_2\right) \le \lambda \mathbb{D}^f\left(P_1 \,\|\, Q_1\right) + (1-\lambda)\mathbb{D}^f\left((P_2 \,\|\, Q_2\right). \tag{17}$$

  This follows from the convexity of the mapping $(p,q) \mapsto q\, f(p/q)$ on $\mathbb{R}_+^2$.

- **Theorem 1.1** *(**Variational representations**) [Polyanskiy and Wu, 2014, Wan et al., 2020]*
  Let f^* be the **convex conjugate** *of f. Let* $\mathrm{effdom}(f^*)$ *be the effective domain of f^*, that is,* $\mathrm{effdom}(f^*) = \{y : f^*(y) < \infty\}$. *Then we have* two **variational representations** *of* $\mathbb{D}^f\left(p \,\|\, q\right)$*:*

$$\mathbb{D}^f\left(P \,\|\, Q\right) = \sup_{g:\,\Omega \to \mathrm{effdom}(f^*)} \mathbb{E}_P\left[g\right] - \mathbb{E}_Q\left[f^* \circ g\right] \tag{18}$$

- Special cases:
  1. ***KL divergence*** if $f(x) = x\log(x)$:

$$\mathbb{D}^f\left(P \,\|\, Q\right) = \int_\Omega dQ \frac{dP}{dQ} \log\left(\frac{dP}{dQ}\right) = \int_\Omega dP \log\left(\frac{dP}{dQ}\right) = \mathbb{E}_P\left[\log\left(\frac{dP}{dQ}\right)\right] = \mathbb{KL}\left(P \,\|\, Q\right)$$

  2. <u>***Total Variation divergence***</u> if $f(x) = \frac{1}{2}|x-1|$:

$$\mathbb{D}^f\left(P \,\|\, Q\right) = \frac{1}{2}\mathbb{E}_Q\left[\left|\left(\frac{dP}{dQ}\right) - 1\right|\right] = \frac{1}{2}\int |dP - dQ| := TV(P \,\|\, Q) \tag{19}$$

     It has *variational representation*

$$TV(P \,\|\, Q) = \sup_{f \in \mathrm{Lip}_1} \mathbb{E}_P\left[f(X)\right] - \mathbb{E}_Q\left[f(X)\right] = \mathcal{W}_1(P, Q) \tag{20}$$

     where $\mathrm{Lip}_1 := \{f : \mathcal{X} \to \mathbb{R} : \|f\|_\infty \le 1\}$ is Lipschitz function. It is also equal to the Wasserstein-1 distance.

  3. <u>χ^2***-divergence***</u> if $f(x) = (x-1)^2$:

$$\mathbb{D}^f\left(P \,\|\, Q\right) = \mathbb{E}_Q\left[\left(\frac{dP}{dQ} - 1\right)^2\right] = \int_\Omega \frac{(dP - dQ)^2}{dQ} := \chi^2(P \,\|\, Q) \tag{21}$$

  4. <u>***Squared Hellinger distance***</u>: $f(x) = (1 - \sqrt{x})^2$

$$\begin{aligned}\mathbb{D}^f\left(P \,\|\, Q\right) &= \mathbb{E}_Q\left[\left(1 - \sqrt{\frac{dP}{dQ}}\right)^2\right] \\ &= \int_\Omega \left(\sqrt{dP} - \sqrt{dQ}\right)^2 = 2 - 2\int \sqrt{dP\, dQ} := H^2(P \,\|\, Q)\end{aligned} \tag{22}$$

5. ***Jensen-Shannon divergence***: $f(x) = x\log(\frac{2x}{x+1}) + \log(\frac{2}{x+1})$,

$$\mathbb{D}^{f}\left(P \,\|\, Q\right) = \mathbb{KL}\left(P \,\|\, \frac{P+Q}{2}\right) + \mathbb{KL}\left(Q \,\|\, \frac{P+Q}{2}\right) := \mathbb{D}^{JS}\left(P \,\|\, Q\right) \tag{23}$$

6. ***Hellinger α-divergence*** $\mathbb{D}^{f_\alpha}\left(p \,\|\, q\right)$ is defined by generator

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

For $\alpha = \pm 1$, it is the KL divergence. For $\alpha \neq \pm 1$, the corresponding divergence is

$$\mathbb{D}^{f^{(\alpha)}}\left(p \,\|\, q\right) = \frac{4}{(1-\alpha^2)}\left\{1 - \int_{\mathcal{X}} (p(x))^{\frac{1+\alpha}{2}} (q(x))^{\frac{1-\alpha}{2}} dx\right\} \tag{24}$$

The Rényi divergence and Hellinger α-divergence has one-to-one correspondence

$$\mathbb{D}^{\frac{\alpha+1}{2}}\left(p \,\|\, q\right) = \frac{2}{\alpha - 1}\log\left(1 - \left(\frac{1-\alpha^2}{4}\right)\mathbb{D}^{f^{(\alpha)}}\left(P \,\|\, Q\right)\right).$$

Note that Rényi divergence itself is **not f-divergence**.

We can formulate the ***dual*** of Hellinger α-divergence using ***the conjugate dual*** of $(f^{(\alpha)})^* = f^{(-\alpha)}$. When $\alpha = 1$, it is the KL divergence.

7. ***Bregman divergence***: The only f-divergence that is also a Bregman divergences is the **KL divergence**.

- f-divergence is a **generalization** of KL divergence from ***information theorectial perspective*** [Cover and Thomas, 2006]. Bregman divergence is a generalization of KL divergence from the ***projection perspective*** as well as *Generalized Pythagorean Theorem.*

2 Divergence on Statistical Manifolds

2.1 Dual Connections

- **Definition** Let (S, g) be a Riemannian manifold and ∇ and ∇^* are two connections on TS. If for all vector fields $X, Y, Z \in \mathfrak{X}(S)$,

$$Z\left\langle X\,,\,Y\right\rangle = \left\langle \nabla_Z X\,,\,Y\right\rangle + \left\langle X\,,\,\nabla^*_Z(Y)\right\rangle \tag{25}$$

holds, then we say that ∇ and ∇^* are ***duals*** to each other with respect to the Riemannian metric g. We call one either ***the dual connection*** or ***the conjugate connection***.

We call the triple (g, ∇, ∇^*) ***a dualistic structure*** on S.

- We see that the coefficients $\Gamma_{i,j;k}$ and $\Gamma^*_{i,j;k}$ for ∇ and ∇^* have the relationship:

$$\partial_k\, g_{i,j} = \Gamma_{k,i;j} + \Gamma^*_{k,j;i}$$

- Similarly, define ***the covariant derivative*** of vector field *along curve* with respect to ∇ and its dual connection ∇^* as D_t and D_t^*, then

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

- For ***the parallel transport map*** Π_γ and Π_γ^* along the curve γ (from t_0 to t_1) with respect to ∇ and its dual ∇^*, we have

$$\left\langle \Pi_\gamma(X)\,,\,\Pi_\gamma^*(Y)\right\rangle_q = \langle X\,,\,Y\rangle_p\,.$$

where $p = \gamma(t_0)$ and $q = \gamma(t_1)$. This is a generalization of "***the <u>invariance</u> of the inner product under <u>parallel translation</u> with respect to <u>metric connections</u>***."

- Also ***the Riemannian curvature tensor*** with respect to ∇ and its dual ∇^* has the relationship

$$\langle R(X,Y)Z\,,\,W\rangle = -\langle R^*(X,Y)Z\,,\,W\rangle\,.$$

Thus $Rm = -Rm^*$, so $R = 0 \Leftrightarrow R^* = 0$.

In other word, *a Riemannian manifold S with dualistic structure* (g, ∇, ∇^*) *is <u>**flat** in ∇</u>* ***if and only*** *if it is <u>**flat** in its **dual connection** ∇^*</u>.*

- It is clear that if ∇ is ***a metric connection***, then $\nabla = \nabla^*$. The concept of dual connections (∇, ∇^*) is a generalization of the metric connection. Moreover, $\frac{1}{2}(\nabla + \nabla^*)$ becomes *a metric connection*.

- Within α-connections, $(\nabla^{(-\alpha)}, \nabla^{(\alpha)})$ are ***duals*** to each other with respect to *the Fisher metric*. Specifically, $(\nabla^{(m)}, \nabla^{(e)})$, i.e. ***the mixture connection*** *and* ***the exponential connection*** *are* ***duals*** *to each other*.

From above statement, we see that

$$S \text{ is } (\alpha)\text{-flat} \;\Leftrightarrow S \text{ is } (-\alpha)\text{-flat} \tag{26}$$

That (S, g, ∇, ∇^*) is called <u>***a dually flat space***</u>

- **Remark** *The exponential family* is *a dually flat space* since it is both 1-***flat*** and (-1)-**flat**. The former corresponds to <u>***the natural parameterization***</u> (ξ^i) which is $\nabla^{(e)}$-***affine*** and the latter corresponds to <u>***the mean parameterization***</u> (μ_i) which is $\nabla^{(m)}$-***affine***. It has ***two mutually dual coordinate systems***.

2.2 Divergence as General Contrast Function

- **Definition** Let S be a statistical manifold. D is a smooth function $D = \mathbb{D}\left(\cdot \,\|\, \cdot\right) : S \times S \to \mathbb{R}$ satisfying for any $p, q \in S$

$$\mathbb{D}\left(p \,\|\, q\right) > 0, \text{ and } \mathbb{D}\left(p \,\|\, q\right) = 0, \text{ iff } p = q. \tag{27}$$

- The divergence function usually does not define a *distance function* since it does not satisfy the ***symmetry*** and ***triangle inequality*** condition.

- Given smooth chart $(U, (\xi^i))$ in S, let us represent ***a pair of points*** $(p, \widetilde{p}) \in S \times S$ by a pair of coordinates $((\xi^i), (\widetilde{\xi}^i))$ and denote ***the partial derivatives*** of $\mathbb{D}\,(p \,\|\, \widetilde{p})$ with respect to p and $\widetilde{p}$ by

$$\widehat{\mathbb{D}}\left((\partial_i)_p \,\Big\|\, \widetilde{p}\right) := \widehat{\mathbb{D}}\left(\frac{\partial}{\partial \xi^i}\Big|_p \,\Big\|\, \widetilde{p}\right) := \frac{\partial}{\partial \xi^i}\Big|_p \mathbb{D}\,(p \,\|\, \widetilde{p})$$

$$\widehat{\mathbb{D}}\left((\partial_i)_p \,\Big\|\, (\widetilde{\partial}_j)_{\widetilde{p}}\right) := \widehat{\mathbb{D}}\left(\frac{\partial}{\partial \xi^i}\Big|_p \,\Big\|\, \frac{\partial}{\partial \widetilde{\xi}^j}\Big|_{\widetilde{p}}\right) := \frac{\partial}{\partial \xi^i}\Big|_p \frac{\partial}{\partial \widetilde{\xi}^j}\Big|_{\widetilde{p}} \mathbb{D}\,(p \,\|\, \widetilde{p})$$

$$\widehat{\mathbb{D}}\left((\partial_i \partial_j)_p \,\Big\|\, (\widetilde{\partial}_k)_{\widetilde{p}}\right) := \widehat{\mathbb{D}}\left(\frac{\partial}{\partial \xi^i}\frac{\partial}{\partial \xi^j}\Big|_p \,\Big\|\, \frac{\partial}{\partial \widetilde{\xi}^k}\Big|_{\widetilde{p}}\right) := \left(\frac{\partial}{\partial \xi^i}\frac{\partial}{\partial \xi^j}\right)\Big|_p \frac{\partial}{\partial \widetilde{\xi}^k}\Big|_{\widetilde{p}} \mathbb{D}\,(p \,\|\, \widetilde{p})$$

$$\cdots$$

Here with abuse of notations, we consider the function $\widehat{\mathbb{D}}(\,\cdot \,\|\, \cdot\,)$ as $T_pS \times T_{\widetilde{p}}S \to \mathbb{R}$, where the derivation operation on p and $\widetilde{p}$ are separated into two sides. Similarly, we have

$$\widehat{\mathbb{D}}\left((X_1, \ldots, X_l)_p \,\Big\|\, \widetilde{p}\right) \quad \text{and} \quad \widehat{\mathbb{D}}\left(p \,\Big\|\, (Y_1, \ldots, Y_m)_{\widetilde{p}}\right)$$
$$\text{and } \widehat{\mathbb{D}}\left((X_1, \ldots, X_l)_p \,\Big\|\, (Y_1, \ldots, Y_m)_{\widetilde{p}}\right)$$

Now we consider ***their restrictions onto the diagonal*** $\{(p, p) : p \in S\} \subset S \times S$ and denote the functions induced on S by

$$\widehat{\mathbb{D}}\left[X_1, \ldots, X_l \,\Big\|\, \cdot\right] : p \mapsto \widehat{\mathbb{D}}\left((X_1, \ldots, X_l)_p \,\Big\|\, p\right)$$
$$\widehat{\mathbb{D}}\left[\cdot \,\Big\|\, Y_1, \ldots, Y_m\right] : p \mapsto \widehat{\mathbb{D}}\left(p \,\Big\|\, (Y_1, \ldots, Y_m)_p\right)$$
$$\text{and } \widehat{\mathbb{D}}\left[X_1, \ldots, X_l \,\Big\|\, Y_1, \ldots, Y_m\right] : p \mapsto \widehat{\mathbb{D}}\left((X_1, \ldots, X_l)_p \,\Big\|\, (Y_1, \ldots, Y_m)_p\right)$$

It follows from the definition that at $p = q$ is the ***miniminer*** of $\mathbb{D}\,(p \,\|\, q)$ and $\mathbb{D}\,(q \,\|\, p)$ so

$$\widehat{\mathbb{D}}\,[\partial_i \,\|\, \cdot] = \widehat{\mathbb{D}}\,[\cdot \,\|\, \partial_i] = 0, \quad i = 1, \ldots, n \tag{28}$$

The ***Hessian*** *of function* $\mathbb{D}$ is defined as

$$\widehat{\mathbb{D}}\,[\partial_i \partial_j \,\|\, \cdot] = \frac{\partial}{\partial \xi^i}\frac{\partial}{\partial \xi^j}\Big|_{p=q} \mathbb{D}\,(p \,\|\, q) := g^D_{i,j}(q) \tag{29}$$

We can also show that

$$\widehat{\mathbb{D}}\,[\partial_i \partial_j \,\|\, \cdot] = \widehat{\mathbb{D}}\,[\cdot \,\|\, \partial_i \partial_j] = -\widehat{\mathbb{D}}\,[\partial_i \,\|\, \partial_j]$$

- **Definition** Let S be a statistical manifold. a ***(statistical) divergence*** or ***contrast function*** is a smooth function $\mathbb{D} = \mathbb{D}\,(\cdot \,\|\, \cdot) : S \times S \to \mathbb{R}$ satisfying for any $p, q \in S$

  1. $\mathbb{D}\,(p \,\|\, q) > 0$
  2. $\mathbb{D}\,(p \,\|\, q) = 0$ iff $p = q$
  3. At each $p = q \in S$, ***the Hessian matrix*** of $\mathbb{D}\,(p \,\|\, q)$, $[g^D_{i,j}]_p$ is ***strictly positive definite*** where

$$g^{(D)}_{i,j} := \widehat{\mathbb{D}}\,[\partial_i \partial_j \,\|\, \cdot] = \widehat{\mathbb{D}}\,[\cdot \,\|\, \partial_i \partial_j] = -\widehat{\mathbb{D}}\,[\partial_i \,\|\, \partial_j]$$

- For a divergence $\mathbb{D}$, a ***unique Riemannian metric*** $g^{(D)} = \langle \, , \, \rangle^{(D)}$ on S is defined by $g^{(D)}_{i,j} := \langle \partial_i \, , \, \partial_j \rangle^{(D)}$, or equivalently by, for $X, Y \in \mathfrak{X}(S)$,

$$\langle X \, , \, Y \rangle^{(D)} = -\widehat{\mathbb{D}} \left[X \parallel Y \right] \tag{30}$$

- This metric gives ***the second order approximation*** of $\mathbb{D}$ as

$$\mathbb{D}\left(p \parallel q\right) = \frac{1}{2} g^{(D)}_{i,j}(q) \Delta\xi^i \, \Delta\xi^j + o(\|\Delta\xi\|_2^2) \tag{31}$$

  where $\Delta\xi^i := \xi^i(p) - \xi^i(q)$ and $o(\|\Delta\xi\|_2^2)$ is a term vanishing faster than $\|\Delta\xi\|_2^2$ as p tends to q.

- Given a ***divergence*** $\mathbb{D}$, we can also define ***an affine connection*** $\nabla^{(D)}$ with coefficients $\Gamma^{(D)}_{i,j;k}$ by

$$\Gamma^{(D)}_{i,j;k} := -\widehat{\mathbb{D}} \left[\partial_i \, \partial_j \parallel \partial_k \right], \tag{32}$$

  or equivalently by

$$\left\langle \nabla^{(D)}_X Y \, , \, Z \right\rangle^{(D)} = -\widehat{\mathbb{D}} \left[XY \parallel Z \right]. \tag{33}$$

- Note that $\nabla^{(D)}$ is ***necessarily symmetric***

$$\Gamma^{(D)}_{i,j;k} = \Gamma^{(D)}_{j,i;k}$$

- Combined with the metric $g^{(D)}$, the connection $\nabla^{(D)}$ gives ***the third order approximation*** of the divergence $\mathbb{D}$: where

$$\mathbb{D}\left(p \parallel q\right) = \frac{1}{2} g^{(D)}_{i,j}(q) \Delta\xi^i \, \Delta\xi^j + \frac{1}{6} h^{(D)}_{i,j,k}(q) \Delta\xi^i \, \Delta\xi^j \Delta\xi^k + o(\|\Delta\xi\|_2^3) \tag{34}$$

  where

$$h^{(D)}_{i,j,k} := \widehat{\mathbb{D}} \left[\partial_i \partial_j \partial_k \parallel \cdot \right] \tag{35}$$

  Indeed, the coefficients $h^{(D)}_{i,j,k}$ are determined from $g^{(D)}$ and $\Gamma^{(D)}_{i,j;k}$ by

$$h^{(D)}_{i,j,k} = \partial_i \, g^{(D)}_{j,k} + \Gamma^{(D)}_{j,k;i}$$

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

$$\Gamma^{(D^*)}_{i,j;k} := -\widehat{\mathbb{D}} \left[\partial_k \parallel \partial_i \, \partial_j \right] \tag{36}$$

  Now it is easy to see the following theorem.

  Theorem 2.1 *$\nabla^{(D)}$ and $\nabla^{(D^*)}$ are **dual** with respect to $g^{(D)}$.*

- Moreover, we see that

$$\begin{aligned}\mathbb{D}\left(p \,\|\, q\right) = \mathbb{D}^*\left(q \,\|\, p\right) &= \frac{1}{2} g_{i,j}^{(D^*)}(p)(-\Delta\xi^i)\,(-\Delta\xi^j) + \frac{1}{6} h_{i,j,k}^{(D^*)}(p)(-\Delta\xi^i)\,(-\Delta\xi^j)(-\Delta\xi^k) + o(\|\Delta\xi\|_2^3) \\ &= \frac{1}{2} g_{i,j}^{(D)}(p)\Delta\xi^i\,\Delta\xi^j - \frac{1}{6} h_{i,j,k}^{(D^*)}(p)\Delta\xi^i\Delta\xi^j\Delta\xi^k + o(\|\Delta\xi\|_2^3)\end{aligned}$$

  Thus

$$h_{i,j,k}^{(D^*)} := \widehat{\mathbb{D}}\left[\cdot \,\|\, \partial_i\partial_j\partial_k\right] = \partial_i g_{j,k}^{(D)} + \Gamma_{j,k;i}^{(D^*)}$$

  We thus see that <u>***any divergence induces a torsion-free dualistic structure***</u>.

- ***Conversely***, *any triple* $(g^{(D)}, \nabla, \nabla^*)$ *of a* ***metric*** *and* ***mutually dual symmetric connections*** *are* ***induced from a divergence***. [Amari and Nagaoka, 2007].

- **Remark** For each ***divergence*** $\mathbb{D}$ and its dual $\mathbb{D}^*$, we can construct *a* <u>***dualist structure***</u> $(g^{(D)}, \nabla^{(D)}, \nabla^{(D^*)})$ *on statistical manifold* S, where ***the Riemannian metric*** $g^{(D)}$ is the ***Hessian*** of $\mathbb{D}$ at $p = q$, and the ***coefficients*** of connections $\Gamma_{i,j;k}^{(D)}$ and $\Gamma_{i,j;k}^{(D^*)}$ are computed in (32) and (36), respectively.

2.3 Induced Connections from KL-Divergence and f-Divergence

- **Example** Consider the KL divergence:

$$\begin{aligned}\mathbb{KL}\left(p(x;\xi) \,\|\, q(x;\widetilde{\xi})\right) &= \int_{\mathcal{X}} p(x;\xi)\log p(x;\xi)dx - \int_{\mathcal{X}} p(x;\xi)\log q(x;\widetilde{\xi})dx \\ \widehat{\mathbb{D}}^{KL}\left[\partial_i \,\|\, \cdot\right] &= (\partial_i)_p\left(\int_{\mathcal{X}} p(x;\xi)\log p(x;\xi)dx - \int_{\mathcal{X}} p(x;\xi)\log q(x;\widetilde{\xi})dx\right) \\ &= \int_{\mathcal{X}} ((\partial_i)_p p(x;\xi))\log p(x;\xi)dx + \int_{\mathcal{X}} ((\partial_i)_p \log p(x;\xi))\,p(x;\xi)dx \\ &\quad - \int_{\mathcal{X}} ((\partial_i)_p p(x;\xi))\log q(x;\widetilde{\xi})dx \\ &\text{since } \int_{\mathcal{X}} (\partial_i \log p)p dx = \int_{\mathcal{X}} (\partial_i p)\,p^{-1}p dx = \int_{\mathcal{X}} (\partial_i p)\,dx = 0 \\ &= \int_{\mathcal{X}} ((\partial_i)_p p(x;\xi))\log p(x;\xi)dx - \int_{\mathcal{X}} ((\partial_i)_p p(x;\xi))\log q(x;\widetilde{\xi})dx \\ \widehat{\mathbb{D}}^{KL}\left[\partial_i \,\|\, \widetilde{\partial}_j\right] &= (\widetilde{\partial}_j)_p\left[\int_{\mathcal{X}} (\partial_i p(x;\xi))\log p(x;\xi)dx - \int_{\mathcal{X}} (\partial_i p(x;\xi))\log q(x;\widetilde{\xi})dx\right] \\ &= -\int_{\mathcal{X}} ((\partial_i)_p p(x;\xi))\left((\widetilde{\partial}_j)_p \log q(x;\widetilde{\xi})\right)dx \\ &= -\int_{\mathcal{X}} ((\partial_i)_p \log p(x;\xi))\left((\widetilde{\partial}_j)_p \log q(x;\widetilde{\xi})\right)p(x;\xi)dx \\ \Rightarrow g_{i,j}^{KL} = -\widehat{\mathbb{D}}^{KL}\left[\partial_i \,\|\, \partial_j\right] &= \int_{\mathcal{X}} ((\partial_i)_p \log p(x;\xi))\,((\partial_j)_p \log p(x;\xi))\,p(x;\xi)dx = g_{i,j}.\end{aligned}$$

- **Example** Consider the f-divergence, where f is convex i.e. $f''(x) > 0$ and $f(1) = 0$

$$\mathbb{D}^f\left(p(x;\xi) \parallel q(x;\widetilde{\xi})\right) = \int_{\mathcal{X}} q(x;\widetilde{\xi}) f\left(\frac{p(x;\xi)}{q(x;\widetilde{\xi})}\right) dx$$
$$\widehat{\mathbb{D}}^f\left[\partial_i \parallel \cdot\right] = -\int_{\mathcal{X}} ((\partial_i)_p p(x;\xi)) f'\left(\frac{p(x;\xi)}{q(x;\widetilde{\xi})}\right) dx$$
$$\widehat{\mathbb{D}}^f\left[\partial_i \parallel \widetilde{\partial}_j\right] = -\int_{\mathcal{X}} \{(\partial_i)_p p(x;\xi)\}\{(\widetilde{\partial}_j)_p q(x;\widetilde{\xi})\}\left(\frac{p(x;\xi)}{q^2(x;\widetilde{\xi})}\right) f''\left(\frac{p(x;\xi)}{q(x;\widetilde{\xi})}\right) dx$$
$$= -\int_{\mathcal{X}} \{(\partial_i)_p \log p(x;\xi)\}\{(\widetilde{\partial}_j)_p \log q(x;\widetilde{\xi})\}\left(\frac{p(x;\xi)}{q(x;\widetilde{\xi})}\right)^2 f''\left(\frac{p(x;\xi)}{q(x;\widetilde{\xi})}\right) q(x;\widetilde{\xi}) dx$$
$$= -\mathbb{E}_q\left[\{(\partial_i)_p \log p(x;\xi)\}\{(\widetilde{\partial}_j)_p \log q(x;\widetilde{\xi})\}\left(\frac{p(x;\xi)}{q(x;\widetilde{\xi})}\right)^2 f''\left(\frac{p(x;\xi)}{q(x;\widetilde{\xi})}\right)\right]$$
$$\Rightarrow g_{i,j}^{D^f} = -\widehat{\mathbb{D}}^f\left[\partial_i \parallel \partial_j\right] := f''(1)\int_{\mathcal{X}} \{(\partial_i)_p \log p(x;\xi)\}\{(\partial_j)_p \log p(x;\xi)\} p(x;\xi) dx = f''(1) g_{i,j} \tag{37}$$

- **Example** We can check on the connection induced by the KL divergence and f-divergence:

  1. For **KL-divergence**, *the induced Riemannian metric is the* ***Fisher metric*** $g_{i,j}$ and *the induced affine connection* induced by is

  $$\Gamma_{i,j;k}^{(KL)} := -\widehat{\mathbb{D}}^{KL}\left[\partial_i\,\partial_j \parallel \partial_k\right] = \mathbb{E}_p\left[\{\partial_i\partial_j\ell + (\partial_i\ell)(\partial_j\ell)\}(\partial_k\ell)\right] = \Gamma_{i,j;k}^{(-1)} \tag{38}$$

  It is ***the mixture connection*** $\nabla^{(-1)} = \nabla^{(m)}$ with respect to *the Fisher metric*.

  2. For ***f*-divergence**, *the induced Riemannian metric is the* ***(scaled) Fisher metric*** with scaling factor $f''(1)$.

  $$-\widehat{\mathbb{D}}^f\left[\partial_i\partial_j \parallel \widetilde{\partial}_k\right] = \partial_i \int_{\mathcal{X}} \left(\frac{p_\xi}{q_{\widetilde{\xi}}}\right)^2 f''\left(\frac{p_\xi}{q_{\widetilde{\xi}}}\right)\{(\partial_j)_p \log p_\xi\}\{(\widetilde{\partial}_k)_p \log q_{\widetilde{\xi}}\} q_{\widetilde{\xi}} dx$$
  $$= \int_{\mathcal{X}} \left\{\left[2\left(\frac{p_\xi}{q_{\widetilde{\xi}}}\right) f''\left(\frac{p_\xi}{q_{\widetilde{\xi}}}\right) + \left(\frac{p_\xi}{q_{\widetilde{\xi}}}\right)^2 f^{(3)}\left(\frac{p_\xi}{q_{\widetilde{\xi}}}\right)\right]\left(\frac{p_\xi}{q_{\widetilde{\xi}}}\right)\{(\partial_i)_p \log p_\xi\}\right\} \times$$
  $$\{(\partial_j)_p \log p_\xi\}\{(\widetilde{\partial}_k)_p \log q_{\widetilde{\xi}}\} q_{\widetilde{\xi}} dx$$
  $$+ \int_{\mathcal{X}} \left(\frac{p_\xi}{q_{\widetilde{\xi}}}\right)^2 f''\left(\frac{p_\xi}{q_{\widetilde{\xi}}}\right)\{(\partial_i\partial_j)_p \log p_\xi\}\{(\widetilde{\partial}_k)_p \log q_{\widetilde{\xi}}\} q_{\widetilde{\xi}} dx$$

  The ***induced affine connection*** by f-divergence is

  $$\Gamma_{i,j;k}^{(D_f)} := -\widehat{\mathbb{D}}^f\left[\partial_i\partial_j \parallel \partial_k\right] = \mathbb{E}_p\left[\left\{f''(1)\partial_i\partial_j\ell + \left(2f''(1) + f^{(3)}(1)\right)(\partial_i\ell)(\partial_j\ell)\right\}(\partial_k\ell)\right] \tag{39}$$

- **Example** As for the ***dual divergence*** and ***dual connections***, we have the following statement:

1. For **KL-divergence**, its *dual* $\mathbb{KL}^*(p \,\|\, q) = \mathbb{KL}(q \,\|\, p)$, ***the affine connection*** induced by $\mathbb{KL}^*$ is ***the exponential connection*** $\nabla^{KL^*} = \nabla^{(1)} = \nabla^{(e)}$.

2. For **f-divergence**, its dual $\mathbb{D}^g(p \,\|\, q) = \mathbb{D}^f(q \,\|\, p)$ where $g(t) = tf(1/t)$ is ***the convex inversion*** of f. Thus the induced connection

$$\Gamma^{(D_g)}_{i,j;k} := -\widehat{\mathbb{D}}^g[\partial_i\partial_j \,\|\, \partial_k] = -\widehat{\mathbb{D}}^f[\partial_k \,\|\, \partial_i\partial_j] = \Gamma^{(D_f^*)}_{i,j;k}$$

Note that $g'(t) = f(1/t) - (1/t)\, f'(1/t)$, $g''(t) = (1/t)^3 f''(1/t)$, $g^{(3)}(t) = -3t^{-4}f''(t^{-1}) - t^{-5}f^{(3)}(t^{-1})$ so $g''(1) = f''(1)$ and $g^{(3)}(1) = -3f''(1) - f^{(3)}(1)$. So ***the dual connection*** is

$$\Gamma^{(D_f^*)}_{i,j;k} := \mathbb{E}_p\left[\left\{f''(1)\partial_i\partial_j\ell - \left(f''(1) + f^{(3)}(1)\right)(\partial_i\ell)(\partial_j\ell)\right\}(\partial_k\ell)\right]$$

2.4 Hellinger α-Divergence and α-Connection

- Now consider the f-divergence $\mathbb{D}^{f^{(\beta)}}(p \,\|\, q)$ with the following f function:

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

This is the ***Hellinger α-divergence*** as discussed above. (Note that in [Amari and Nagaoka, 2007] the definition of f-divergence is the dual of the standard f-divergence definition. As a result, the Hellinger α-divergence is the book is also the dual of the standard one. We need to replace $\beta = -\alpha$ to recover the book's definition.) For $\beta = 1$, it is the KL divergence and $\beta = -1$ it is the dual of KL divergence. For $\beta \neq \pm 1$, the corresponding divergence is

$$\mathbb{D}^{f^{(\beta)}}(p \,\|\, q) = \frac{4}{(1-\beta^2)}\left\{1 - \int_{\mathcal{X}} (p(x))^{\frac{1+\beta}{2}}(q(x))^{\frac{1-\beta}{2}}dx\right\}$$

Then for $\beta \neq \pm 1$, $\frac{d}{dt}f^{(\beta)} = -\frac{2}{1-\beta}x^{\frac{\beta-1}{2}}$ and $\frac{d^2}{dt^2}f^{(\beta)} = x^{\frac{\beta-3}{2}}$ so that $f''(1) = 1$. $\frac{d^3}{dt^3}f^{(\beta)} = \frac{\beta-3}{2}x^{\frac{\beta-5}{2}}$ and $f^{(3)}(1) = \frac{\beta-3}{2}$.

Substitute the formula $f^{(\beta)}(x)$ into the (39)

$$\begin{aligned}\Gamma^{(D_f^{(\beta)})}_{i,j;k} &= \mathbb{E}_p\left[\left\{f''(1)\partial_i\partial_j\ell + \left(2f''(1) + f^{(3)}(1)\right)(\partial_i\ell)(\partial_j\ell)\right\}(\partial_k\ell)\right] \\ &= \mathbb{E}_p\left[\left\{\partial_i\partial_j\ell + \frac{1+\beta}{2}(\partial_i\ell)(\partial_j\ell)\right\}(\partial_k\ell)\right] \end{aligned} \tag{40}$$

For $\beta = 1$, we reconstruct the same formula as in (38).

- Recall that ***the α-connections*** [Amari and Nagaoka, 2007] $\nabla^{(\alpha)}$ as ***a family of affine connections*** on the tangent bundle TS. The ***coefficient of the α-connection*** under ***the Fisher metric*** is formulated as

$$\Gamma^{(\alpha)}_{i,j;k} = \mathbb{E}_p\left[\left(\partial_i\partial_j\ell + \frac{1-\alpha}{2}(\partial_i\ell)(\partial_j\ell)\right)(\partial_k\ell)\right] \tag{41}$$

- **Remark** Thus we show that ***the α-connection with respect to the Fisher metric*** g ***is the induced affine connection by the the Hellinger α-divergence*** in (24). And ***the induced dualistic structure*** $(g^{(D_f^{(\alpha)})}, \nabla^{(D_f^{(\alpha)})}, \nabla^{(D_f^{(-\alpha)})})$ *is equal to* $(g, \nabla^{(-\alpha)}, \nabla^{(\alpha)})$.

2.5 Dual Coordinate System

- **Remark** Recall that ***the exponential family*** is *a dually flat space* since it is both 1-***flat*** and (-1)-**flat**. The former corresponds to ***the natural parameterization*** (ξ^i) which is $\nabla^{(e)}$-***affine*** and the latter corresponds to ***the mean parameterization*** (μ_i) which is $\nabla^{(m)}$-***affine***. It has ***two mutually dual coordinate systems***.

  Specifically, we have two coordinate systems (ξ^i) and (η_j):

  1. ***The canonical representation*** of exponential famlity of distribution has the following form
  $$p(x;\xi) = \exp\left(\sum_i \xi^i\, \phi_i(x) - A(\xi)\right) h(x)d\mu(x)$$
     where (ϕ_i) defines a set of *sufficient statistics* (or ***potential functions***). The normalization factor is defined as
  $$A(\xi) := \log \int \exp\left(\sum_i \xi^i\, \phi_i(x) - A(\xi)\right) h(x)d\mu(x) = \log Z(\xi)$$
     $A(\boldsymbol{\eta})$ is the log-partition function. The parameterization (ξ^i) are called ***natural parameters*** or ***canonical parameters***.

     The natural coordinates (ξ^i) ***is a*** 1-***affine coordinate system***. The canonical parameter $\{\xi i\}$ forms a **natural (canonical) parameter space**
  $$\Omega = \{\xi \in \mathbb{R}^n : A(\xi) < \infty\}$$

  2. ***The mean representation*** is related to the unique solution of the ***maximum entropy estimation*** problem:
  $$\begin{aligned} \min_{q \in \Delta} \quad & \mathbb{KL}\left(q \,\|\, p_0\right) \\ \text{s.t.} \quad & \mathbb{E}_q\left[\phi_j(X)\right] = \mu_j \quad \forall\, j \in \mathcal{I}. \end{aligned}$$
     Here (μ_j) is a set of ***mean parameters***, which forms (-1)-***affine coordinate system***. The space of mean parameters $\mathcal{M}$ is a ***convex polytope*** spanned by potential functions $\{\phi_i\}$.
  $$\mathcal{M} := \{\mu \in \mathbb{R}^n : \exists q \text{ s.t. } \mathbb{E}_q\left[\phi_j(X)\right] = \mu_j \quad \forall\, j \in \mathcal{I}\} = \text{conv}\left\{\phi_j(x),\; x \in \mathcal{X},\; j \in \mathcal{I}\right\}$$

- We can see that this is not unique to exponential families. In fact, ***the existance of mutually dual coordinate systems is the characteristics of a dually flat space***.

- **Definition** For any dually flat space with structure (g, ∇, ∇^*), let (ξ^i) be a coordinate system that is ∇-***flat*** (i.e. $\Gamma_{i,j;k} = 0$ under (ξ^i)), and (μ_j) be a coordinate system that is ∇^*-***flat*** (i.e. $\Gamma^*_{i,j;k} = 0$ under (μ_j)).

  Denote $\partial_i \equiv \frac{\partial}{\partial \xi^i}$ and $\partial^j \equiv \frac{\partial}{\partial \mu_j}$. Since ∂_i is a ∇-***flat vector field*** and ∂^j is a ∇^*-***flat vector field***, we see from property of *dual connections* that $\left\langle \partial_i\,,\, \partial^j \right\rangle_g$ is ***constant*** on S. Moreover, for a particular ∇-***affine coordinate system*** (ξ^i), one may ***choose*** a corresponding ∇^*-***affine coordinate system*** (η_j) such that

$$\left\langle \partial_i \, , \, \partial^j \right\rangle_g = \delta_i^j \tag{42}$$

In general, if two coordinate systems (ξ^i) and (μ_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.

- **Remark** We can see similar ***duality*** *structure between* ***vector fields*** *and* ***covector fields***. In Riemannian manifold, $\partial^j = (\epsilon^j)^\sharp$ can be seen as obtained from some covector fields $\epsilon^j \in \mathfrak{X}^*(S)$ by ***raising an index*** [Lee, 2018].

- Note that the Euclidean coordinate system is *self-dual. In general, there do* ***not exist dual coordinate systems for a Riemannian manifold*** (S, g)***. Conversely***, if for a Riemannian manifold (S, g) there exists such coordinate systems (ξ^i) and (μ_j), then the connections ∇ and ∇^* for which they are affine are determined, and (g, ∇, ∇^*) ***is a dually flat space***.

- Moreover, we see that

$$g_{i,j} = \langle \partial_i \, , \, \partial_j \rangle \, , \quad g^{i,j} = \left\langle \partial^i \, , \, \partial^j \right\rangle . \tag{43}$$

By considering the coordinate transformation between (ξ^i) and (μ_j), we have ***the change of coordinate***

$$\partial_i = (\partial_i \, \mu_j) \, \partial^j, \quad \partial^j = (\partial^j \xi^i) \partial_i$$

From this we see that Equation (42) is equivalent to

$$\frac{\partial \mu_j}{\partial \xi^i} = g_{i,j}, \quad \frac{\partial \xi^i}{\partial \mu_j} = g^{i,j} \tag{44}$$

and therefore $g_{i,j} g^{j,k} = \delta_i^k$, which is consistent with Equation (42).

- Now suppose that we are given mutually dual coordinate systems (ξ^i) and (μ_j), and consider the following ***partial differential equation*** for a function $\psi : S \to \mathbb{R}$:

$$\partial_i \psi = \mu_i. \tag{45}$$

Note that $\psi \equiv A$ which is ***the log-partition function*** for exponential family. We may rewrite this as $d\psi = \mu_i d\xi^i$, and a solution exists if and only if $\partial_i \mu_j = \partial_j \mu_i$. Since from Equation (44) we see that $\partial_i \mu_j = g_{i,j} = \partial_j \mu_i$, in the context of our discussion a solution ψ always exists. Thus

$$\partial_i \partial_j \psi = g_{i,j}. \tag{46}$$

Hence the second derivatives of ψ form a *positive definite matrix*, and therefore ψ ***is a strictly convex function*** of $(\xi^1 \, , \ldots , \xi^n)$. Similarly, a solution φ to

$$\partial^i \varphi = \xi^i \tag{47}$$

exists. In particular, using a solution ψ to Equation (45), let

$$\varphi = \xi^i \, \mu_i - \psi \tag{48}$$

Then we have

$$\begin{aligned} d\varphi &= \xi^i d\mu_i + \mu_i d\xi^i - d\psi \\ &= \xi^i d\mu_i \end{aligned}$$

we see that φ satisfies

$$\partial^i \partial^j \varphi = g^{i,j}, \tag{49}$$

and hence it is a ***strictly convex function*** *of* $(\mu_1, \ldots, \mu_n)$. Furthermore, it follows from the convexity of ψ and Equations (46) and (48) that for every $q \in S$

$$\varphi(q) = \sup_{p \in S} \left\{ \xi^i(p)\, \mu_i(q) - \psi(p) \right\} \tag{50}$$

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

$$\psi(p) = \sup_{q \in S} \left\{ \xi^i(p)\, \mu_i(q) - \varphi(q) \right\} \tag{51}$$

- **Definition** In general, those coordinate transformations (ξ^i) and (μ_j) which may be expressed in the form given in Equations (46) through (51) are called ***Legendre transformations***, and ψ and φ are called their ***potentials***.
- Note also that

$$\Gamma^*_{i,j;k} := \left\langle \nabla^*_{\partial_i} \partial_j\, ,\, \partial_k \right\rangle = \partial_i\, \partial_j\, \partial_k\, \psi, \tag{52}$$

$$\Gamma^{i,j;k} := \left\langle \nabla_{\partial^i} \partial^j\, ,\, \partial^k \right\rangle = \partial^i\, \partial^j\, \partial^k\, \varphi, \tag{53}$$

which are derived from Equation

$$\partial_k g_{i,j} = \Gamma_{k,i;j} + \Gamma^*_{k,j;i}$$

combined with the fact that (ξ^i) and (μ_j) are ∇-affine and ∇^*-affine so $\Gamma_{i,j;k} = \Gamma^{*i,j;k} = 0$.

Theorem 2.2 ***(The Existance of Dual Coordinate System in Dually Flat Space)*** *[Amari and Nagaoka, 2007]*
*Let (ξ^i) be a **∇-affine** coordinate system on a **dually flat space** (S, g, ∇, ∇^*). Then with respect to g there exists a **dual coordinate system** (μ_j) of (ξ^i), where (μ_j) 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 (46) through (51). 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 (46) and (49).*

- **Remark A similar analysis** can be found in [Wainwright et al., 2008] (see *probablistic graphical model self-learning note*) based on ***convex analysis***. On the other hand, the analysis in this section is from ***the differential geometry point of view***, and it applies to ***all dually flat spaces*** *with respect to* (g, ∇, ∇^*). It also **generalize** the concept of *canonical representation* and *mean representation* of exponential family to ***the dual coordinate systems*** with respect to Riemannian metric g.

2.6 Canonical Divergence

- **Remark** We have seen that every divergence D induces a torsion-free dualistic structure $(g, \nabla^D, \nabla^{D*})$ on the statistical manifold S. On the other hand, the correpsonding between divergence and dualistic structure is not one-to-one, i.e. ***there exists many divergence to the same dualistic structure***. In this section, we will present one divergence that is ***uniquely*** defined on a dually flat space.

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

$$\mathbb{D}\left(p \,\|\, q\right) := \psi(p) + \varphi(q) - \xi^i(p)\, \mu_i(q). \tag{54}$$

From (50) and (51) we see that $\mathbb{D}\left(p \,\|\, q\right) \geq 0$ with equality holds iff $p = q$. Moreover, we see that

$$\mathbb{D}\left((\partial_i \partial_j)_p \,\|\, p\right) = g_{i,j}(p), \quad \mathbb{D}\left(p \,\|\, (\partial^i \partial^j)_p\right) = g^{i,j}(p). \tag{55}$$

This implies that D is a *divergence* that induces the metric g. This divergence is called ***the canonical divergence*** of (S, g, ∇, ∇^*) or the (g, ∇)***-divergence*** on S.

- **Remark** After change of coordinates, (see [Amari and Nagaoka, 2007],) we see that the canonical divergence D in (54) is ***uniquely defined*** from (S, g, ∇, ∇^*)

- **Remark** D is (g, ∇)-divergence if and only its *dual* D^* is (g, ∇^*)-divergence

- **Example** (***KL-divergence is Canonical Divergence***)
Compare it to the **primal-dual form** (13), we see that the ***KL-divergence*** is ***the canonical divergence of*** $(S, \nabla^{(m)}, \nabla^{(e)})$ (or ***KL-divergence is*** $(g, \nabla^{(-1)})$***-divergence***)

$$\mathbb{KL}\left(\mu(p) \,\|\, \xi(q)\right) = A^*(\mu(p)) + A(\xi(q)) - \mu_i(p) \xi^i(q).$$

Thus, the KL-divergence is ***uniquely*** determined on $(S, \nabla^{(m)}, \nabla^{(e)})$.

- **Remark** For ***Riemannian connection*** $\nabla = \nabla^*$, the *dually flat space* becomes ***flat space*** and there exists a Euclidean coordinate system (ξ^i) such that $\varphi = \psi = \frac{1}{2} \|\xi\|_2^2$

$$\begin{aligned}\mathbb{D}\left(p \,\|\, q\right) &= \psi(p) + \varphi(q) - \xi^i(p)\, \mu_i(q) = \frac{1}{2} \sum_i \left((\xi^i(p))^2 + (\xi^i(q))^2 - 2\xi^i(p)\xi^i(q)\right) \\ &= \frac{1}{2}(d(p,q))^2,\end{aligned}$$

where $d(p, q)$ is *the Euclidean distance* between the coordinates of p and q.

- The following is the important characteristic of the canonical divergence:

Theorem 2.3 ***(Characterization of Canonical Divergence)*** *[Amari and Nagaoka, 2007] Let $\{(\xi^i), (\mu_j)\}$ be mutually dual affine coordinate systems of a dually flat space (S, g, ∇, ∇^*), and let D be a divergence on S. Then a* ***necessary and sufficient condition*** *for D to be the (g, ∇)-divergence is that for all $p, q, r \in S$ the following* ***triangular relation*** *holds:*

$$\mathbb{D}\left(p \,\|\, q\right) + \mathbb{D}\left(q \,\|\, r\right) - \mathbb{D}\left(p \,\|\, r\right) = \{\xi^i(p) - \xi^i(q)\}\{\mu_i(r) - \mu_i(q)\} \tag{56}$$

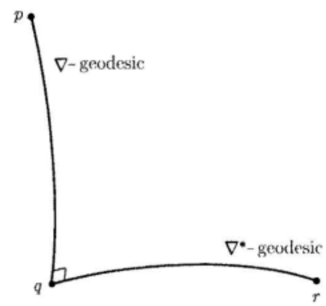

Figure 1: The Pythagorean relation for canonical divergence [Amari and Nagaoka, 2007]

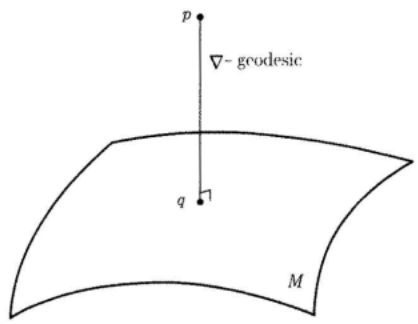

Figure 2: The Projection theorem for canonical divergence [Amari and Nagaoka, 2007]

- **Theorem 2.4** ***(Pythagorean Theorem for Canonical Divergence)*** *[Amari and Nagaoka, 2007]*
  Let p, q, and r be three points in S . 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*** *(Fig 1)*

$$\mathbb{D}\left(p \,\|\, r\right) = \mathbb{D}\left(p \,\|\, q\right) + \mathbb{D}\left(q \,\|\, r\right) \tag{57}$$

- **Corollary 2.5** ***(Projection Theorem)*** *[Amari and Nagaoka, 2007]*
  *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*

$$\mathbb{D}\left(p \,\|\, q\right) = \min_{r \in M} \mathbb{D}\left(p \,\|\, r\right)$$

  is for the ***∇-geodesic*** *connecting p and q to be* ***orthogonal*** *to M at q.*

- **Definition** The point q in the theorem above is called the ***∇-projection of p onto M***.

- **Remark** The ***maximum likelihood estimation***

$$\min_{r \in M} \mathbb{KL}\left(p \,\|\, r\right)$$

  is the ***$\nabla^{(m)}$-projection*** or ***m-projection*** onto M. In other words, the process of maximum likelihood estimation is to ***match the mean*** of features from the model to the mean of the features from the data.

On the other hand, ***the maximum entropy estimation***

$$\min_{r \in M} \mathbb{KL}\left(r \,\|\, p\right)$$

is the $\nabla^{(e)}$***-projection*** or ***e-projection*** onto M. In other word, *the process of maximum entropy estimatoin* is to ***project*** of *the prior distribution* into the ***exponential family***.

- **Theorem 2.6** *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 $\mathbb{D}\left(p \,\|\, \cdot\right) : r \mapsto \mathbb{D}\left(p \,\|\, r\right)$ 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.*

- **Corollary 2.7** *Given a point p in S and a positive number c, suppose that the "D-sphere" $M = \{q \in S : \mathbb{D}\left(p \,\|\, q\right) = c\}$ forms a* ***hypersurface*** *in S. Then every* ∇***-geodesic*** *passing through the center p* ***orthogonally*** *intersects M.*

- **Remark** Similarly, ***the Hellinger*** α***-divergence in*** (24) ***is a*** $(g, \nabla^{(-\alpha)})$***-divergence***. It is ***the canonical divergence*** *with respect to* ***dualistic structure*** $(S, g, \nabla^{(-\alpha)}, \nabla^{(\alpha)})$ where g is the Fisher metric.

- **Remark** The ***KL-divergence*** ($\alpha = \pm 1$) is the ***only*** f***-divergence*** that fits the Pythagorean relation (57). The other canonical divergence w.r.t. $(S, g, \nabla^{(-\alpha)}, \nabla^{(\alpha)})$ has similar formula but has an additional product term $\mathbb{D}^{(\alpha)}\left(p \,\|\, q\right) \mathbb{D}^{(\alpha)}\left(q \,\|\, r\right)$.

References

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006. ISBN 978-0-471-24195-9.

Imre Csiszár, Paul C Shields, et al. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528, 2004.

Alfred O Hero, Bing Ma, Olivier Michel, and John Gorman. Alpha-divergence for classification, indexing and retrieval. In *University of Michigan*. Citeseer, 2001.

John M Lee. *Introduction to Riemannian manifolds*, volume 176. Springer, 2018.

Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

Frank Nielsen and Richard Nock. On Rényi and Tsallis entropies and divergences for exponential families. *arXiv preprint arXiv:1105.3259*, 2011.

Barnabás Póczos and Jeff Schneider. On the estimation of alpha-divergences. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 609–617. JMLR Workshop and Conference Proceedings, 2011.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7, 2014.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Neng Wan, Dapeng Li, and Naira Hovakimyan. f-divergence variational inference. *Advances in neural information processing systems*, 33:17370–17379, 2020.