Information Geometry: A Review

Research Report for STA497H1

Zhixu Tao
Department of Mathematics
University of Toronto

Supervisor: Prof. Ting-Kam Leonard Wong
Department of Statistical Sciences
University of Toronto

Abstract

This report summarizes (i) the relationship among dually flat manifolds, Bregman divergence and exponential family, and (ii) the relationship among dually projective flat manifolds, $L^{(\pm\alpha)}$ divergence and $F^{(\pm\alpha)}$ family. Also, this report summarizes how (i) and (ii) are related with each other.

Contents

1 Introduction **3**

2 Dually Flat Manifold and Bregman Divergence **3**
2.1 Dualistic structure of Bregman Divergence 3
2.2 Geometry of Dually Flat Manifolds 6
2.3 Dual flatness from Bregman Divergence 9
2.4 Orthogonal Foliation 12

3 Exponential Families of Probability Distributions **14**
3.1 Flat Structure 14
3.2 Principal Component Analysis 17

4 Dually Projective Flat Manifold and $L^{(\pm\alpha)}$ Divergence **19**
4.1 $L^{(\pm\alpha)}$ Divergence 19
4.2 Dualistic Structure of $L^{(\pm\alpha)}$ Divergence 22
4.3 Dually Projective Flat Manifold 24
4.4 Orthogonal Foliation 28

5 Generalizations of Exponential Family **29**
5.1 $F^{(\alpha)}$ Family of Probability Distributions 29
5.2 Principal Component Analysis 32

6 Acknowledgment **32**

1 Introduction

Information Geometry is an interdisciplinary research field that uses Riemannian Geometry to study statistical manifold. The modern theory and applications of Information Geometry is mainly due to a series of work [5], [4], [6] etc. In this report, we will discuss about several important parts from Information Geometry, namely, manifolds with special geometric structure induced by some divergences and special family of probability distributions. The first special manifold we will explore has a dually flat structure, which is induced by Bregman divergence. Bregman divergence is a useful class of loss functions [9]. The geometric structure induced by Bregman divergence was first studied by Nagaoka and Amari in [14]. It is natural to use Bregman divergence when we analyze exponential family of probability distributions, so how the exponential family and Bregman divergence are related is of our interests. In fact, the exponential family and Bregman divergence has a one-to-one correspondence which was studied in [7]. The second special manifold in this report has dually projective flat structure, which corresponds to a larger family of distributions, namely, $F^{(\pm\alpha)}$ family, and $F^{(\pm\alpha)}$ family is induced by another type of divergence, $L^{(\pm\alpha)}$ divergence. We regard $L^{(\pm\alpha)}$ divergence as a new extension of Bregman divergence. These have already been studied in a series of papers [27], [23], [25], [24], [28], [29], [30]. There also exist some other divergences related with Bregman divergence. Readers may refer to a series of work [13], [32], [33], [3], [34], [21], [22] for more details. We will introduce some fundamental theorems in both of two parts such as Generalized Pythagorean Theorem. Also, we will discuss some applications of Bregman divergence and $L^{(\pm\alpha)}$ divergence, one of which is Principal Component Analysis [11]. There are many other applications of Bregman divergence discussed in [12], [26], [31], [8] and [7]. Moreover, we will discuss how those two parts are related. Specifically, we will see how Bregman divergence and exponential family can be generalized to $L^{(\pm\alpha)}$ divergence and $F^{(\pm\alpha)}$ family.

2 Dually Flat Manifold and Bregman Divergence

In this section, we are going to study manifolds with a dually flat structure. First, we will introduce basic notations of differentiable manifolds. A manifold M of dimension k is a set of points such that $\forall a \in M$, there is an open neighbourhood U of a, open set $W \subset \mathbb{R}^k$ and C^r mapping $\varphi : W \to \mathbb{R}^n$ such that φ is one-to-one and $\varphi(W) = M \cap U$. This means a manifold M is locally isomorphic to a k-dimensional Euclidean space. Therefore, we can introduce a local coordinate system $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k)$. Since coordinate system is not unique, we can introduce another coordinate system $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)$. There is a C^r mapping f, which is called transition map, between two coordinate systems such that $\boldsymbol{\xi} = f(\eta_1, \ldots, \eta_k)$ and $\boldsymbol{\eta} = f^{-1}(\xi_1, \ldots, \xi_k)$.

2.1 Dualistic structure of Bregman Divergence

Bregman divergence is related with convex function. Recall that a real-valued function f defined on a convex set U is convex if $\forall x_1, x_2 \in U$ and $t \in [0, 1]$, we have $f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$. A Bregman divergence measures distance between the value of f at point x and the hyperplane defined around another point x_0.

Definition 2.1. *(Bregman Divergence) Given a convex function $\varphi : U \to \mathbb{R}$, the Bregman*

divergence associated with two points $x, y \in U$ is written as

$$D_\varphi[x : y] = \varphi(x) - \varphi(y) - \nabla\varphi(y) \cdot (x - y).$$

To introduce dual divergence of Bregman divergence, we will first define Legendre Transformation. Legendre transformation is defined for a real-valued convex function.

Definition 2.2. *(Legendre Transformation) Given a convex function $\varphi(\xi)$, the Legendre transformation of $\varphi(\xi)$ is defined as*

$$\varphi^*(\xi^*) = \max_{\xi'}\{\xi' \cdot \xi^* - \varphi(\xi')\}.$$

For a differentiable convex function φ with an invertible first derivative, we can specify its Legendre transformation. Denote the gradient of $\varphi(\xi)$ by $\xi^* = \nabla\varphi(\xi)$ and define a new function of ξ^* by

$$\varphi^*(\xi^*) = \xi \cdot \xi^* - \varphi(\xi) \tag{1}$$

Note that ξ^* is defined as the gradient of φ at ξ, so ξ is not a free variable in the function we just defined. Actually, ξ is the inverse function of $\xi^* = \nabla\varphi(\xi^*)$. If we differentiate equation (1), by chain rule, we have

$$\begin{aligned}
\nabla\varphi^*(\xi^*) &= \frac{\partial}{\partial\xi^*}\{\xi \cdot \xi^* - \varphi(\xi)\} \\
&= \xi + \frac{\partial\xi}{\partial\xi^*}\xi^* - \nabla\varphi(\xi)\frac{\partial\xi}{\partial\xi^*} \\
&= \xi + \frac{\partial\xi}{\partial\xi^*}\xi^* - \xi^*\frac{\partial\xi}{\partial\xi^*} \\
&= \xi
\end{aligned}$$

As we see, $\xi = \nabla\varphi^*(\xi^*)$ implies that ξ is the critical point of the function $\xi' \cdot \xi^* - \varphi(\xi')$. Therefore, the maximum of $\xi' \cdot \xi^* - \varphi(\xi')$ is achieved by ξ, and $\varphi^*(\xi^*) = \xi \cdot \xi^* - \varphi(\xi)$ is the Legendre transformation of convex function φ. Moreover, this gives us a dualistic structure.

Proposition 2.3. *(Dualistic structure of Legendre transformation) Given a convex function $\varphi(\xi)$ and its Legendre transformation $\varphi^*(\xi^*)$, we have a dualistic structure*

$$\xi^* = \nabla\varphi(\xi),\ \xi = \nabla\varphi^*(\xi^*).$$

The Legendre transformation $\varphi^*(\xi^*)$ also has convexity.

Proposition 2.4. *(Convexity of Legendre transformation) If φ is a convex function, its Legendre transformation $\varphi^*(\xi^*) = \xi \cdot \xi^* - \varphi(\xi)$ is also convex.*

Proof. The Hessian of $\varphi^*(\xi^*)$ is given by

$$\nabla(\nabla\varphi^*(\xi^*)) = \nabla(\xi) = \frac{\partial\xi}{\partial\xi^*} = (\frac{\partial\xi^*}{\partial\xi})^{-1} = (\nabla(\nabla\varphi(\xi)))^{-1}$$

Since φ is convex, then its Hessian $\nabla(\nabla\varphi(\xi))$ is positive definite, so is its inverse. Therefore, the Hessian of $\varphi^*(\xi^*)$ is positive definite, which means $\varphi^*(\xi^*)$ is convex. □

Because of the convexity of φ^*, we can define Bregman divergence D_{φ^*} in terms of φ^*, which is called dual divergence [4] of D_φ.

Definition 2.5. *(Dual Divergence) For a convex function φ, its dual divergence is given by*

$$D_{\varphi^*}[x^* : y^*] = \varphi^*(x^*) - \varphi^*(y^*) - \nabla\varphi^*(y^*) \cdot (x^* - y^*)$$

where φ^* is Legendre transformation of φ.

The dual divergence and Bregman divergence has the following relation.

Proposition 2.6. *The Bregman divergence induced by φ and its dual divergence satisfy*

$$D_{\varphi^*}[x^* : y^*] = D_\varphi[y : x]$$

where $x^* = \nabla\varphi(x)$ and $y^* = \nabla\varphi(y)$.

Proof.

$$\begin{aligned} D_{\varphi^*}[x^* : y^*] &= \varphi^*(x^*) - \varphi^*(y^*) - \nabla\varphi^*(y^*) \cdot (x^* - y^*) \\ &= x \cdot x^* - \varphi(x) - y \cdot y^* + \varphi(y) - y \cdot (x^* - y^*) \\ &= \varphi(y) - \varphi(x) - y \cdot x^* + x \cdot x^* \\ &= \varphi(y) - \varphi(x) - x^* \cdot (y - x) \\ &= \varphi(y) - \varphi(x) - \nabla\varphi(x) \cdot (y - x) \\ &= D_\varphi[y : x] \end{aligned}$$

□

Because of the dualistic structure of Legendre transformation, we can derive a self-dual expression of Bregman divergence by using two coordinate systems ξ and ξ^* [4].

Theorem 2.7. *(Self-dual expression of Bregman divergence) The divergence from P to Q derived from a convex function $\varphi(\xi)$ is written as*

$$D_\varphi[P : Q] = \varphi(\xi_P) + \varphi^*(\xi_Q^*) - \xi_P \cdot \xi_Q^*$$

where ξ_P is the coordinates of P in ξ coordinate system and ξ_Q^* is the coordinates of Q in ξ^* coordinate system.

Proof.

$$\begin{aligned} D_\varphi[P : Q] &= \varphi(\xi_P) - \varphi(\xi_Q) - \nabla\varphi(\xi_Q) \cdot (\xi_P - \xi_Q) \\ &= \varphi(\xi_P) + \varphi^*(\xi_Q^*) - \xi_Q \cdot \xi_Q^* - \xi_Q^* \cdot (\xi_P - \xi_Q) \\ &= \varphi(\xi_P) + \varphi^*(\xi_Q^*) - \xi_Q^* \cdot \xi_P \end{aligned}$$

□

This self-dual expression will play an important role later in the proof of Generalized Pythagorean Theorem.

2.2 Geometry of Dually Flat Manifolds

Before we discuss the geometry of dually flat manifolds, we first review some definitions and theorems from Riemannian Geometry [10].

Definition 2.8. *(Riemannian metric) A Riemannian metric on a differentiable manifold M is a correspondence which associates to each point p of M an inner product $\langle\ ,\ \rangle_p$ on the tangent space T_pM, which varies differentiably.*

Definition 2.9. *(Riemannian manifold) A differentiable manifold equipped with a Riemannian metric is called Riemannian manifold.*

From now on, denote the set of C^∞ vector fields on a Riemannian manifold M by $\mathfrak{X}(M)$.

Definition 2.10. *An affine connection ∇ on a differentiable manifold M is a mapping*

$$\nabla : \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$$

which is denoted by $(X, Y) \xrightarrow{\nabla} \nabla_X Y$ and which satisfies the following properties:
i. $\nabla_{fX+gY} Z = f\nabla_X Z + g\nabla_Y Z$
ii. $\nabla_X(Y + Z) = \nabla_X Y + \nabla_X Z$
iii. $\nabla_X(fZ) = f\nabla_X Y + X(f)Y$.

Definition 2.11. *(Compatible metric) A connection ∇ on a Riemannian manifold M is compatible with the metric if and only if*

$$X\langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle$$

for all vector fields X, Y, Z on M.

Definition 2.12. *(Symmetric connection).An affine connection ∇ on a smooth manifold M is symmetric when*

$$\nabla_X Y - \nabla_Y X = [X, Y]\ \forall X, Y \in \mathfrak{X}(M)$$

where $[\ ,\ ]$ is the Lie bracket of vector fields.

Theorem 2.13. *(Levi-Civita) Given a Riemannian manifold M, there exists a unique affine connection ∇ on M such that ∇ is symmetric and ∇ is compatible with the Riemannian metric.*

Definition 2.14. *(Curvature tensor) The curvature R of a Riemannian manifold M is a correspondence that associates to every pair $X, Y \in \mathfrak{X}(M)$ a mapping $R(X, Y) : \mathfrak{X}(M) \to \mathfrak{X}$ given by*

$$R(X, Y)Z = \nabla_Y \nabla_X Z - \nabla_X \nabla_Y Z + \nabla_{[X,Y]} Z$$

where ∇ is the Riemannian connection of M.

Remark 2.15. We can view the above definition of curvature tensor in another way. Suppose we have a coordinate system $\{x_i\}$ near a point $p \in M$. Then the basis vector field is given by $\{\frac{\partial}{\partial x_i}\}$. Since the Riemannian connection is symmetric, $[\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j}] = 0$. Therefore, we obtain

$$R(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j})\frac{\partial}{\partial x_k} = (\nabla_{\frac{\partial}{\partial x_j}} \nabla_{\frac{\partial}{\partial x_i}} - \nabla_{\frac{\partial}{\partial x_i}} \nabla_{\frac{\partial}{\partial x_j}})\frac{\partial}{\partial x_k} \tag{2}$$

With those basic notations from Riemannian geometry for general Riemannian manifolds, we can define our notations for dually flat manifolds [5]. The first important thing is to define flatness. Intuitively, flatness means no curvature, so the flat manifold will be similar as Euclidean space. In the space with no curvature such as Euclidean space, if we parallel transport a vector along any loop, we will always get the original vector. This implies the parallel transport of a vector does not depend on the path [4]. Therefore, we have the following definitions [5].

Definition 2.16. *(Affine coordinate system) Let ξ^i be a coordinate system of a manifold M and suppose that with respect to this coordinate system, the n basis vector fields $\frac{\partial}{\partial \xi^i}, i = 1, \ldots, n$ are parallel on M. Then we call ξ^i an affine coordinate system for the connection ∇.*

Remark 2.17. Notice that when we require $\frac{\partial}{\partial \xi^i}$ to be parallel on M, it's equivalent to say$\nabla_{\frac{\partial}{\partial \xi^i}} \frac{\partial}{\partial \xi^j} = 0$. It's also equivalent to say all the Christoffel symbol are identically zero.

Definition 2.18. *(Flatness) Given a Riemannian connection ∇ on M, if there exists an affine coordinate system for the connection ∇, we say that ∇ is flat and M is flat with respect to ∇.*

Remark 2.19. If a manifold is flat, then by equation (2) from remark 2.15, we can obtain $R(\frac{\partial}{\partial \xi^i}, \frac{\partial}{\partial \xi^j})\frac{\partial}{\partial \xi^k} = 0$. Therefore, flatness implies the curvature R is identically zero.

Definition 2.20. *(Dual connection) Let M be a manifold with Riemannian metric $g = \langle\ ,\ \rangle$ and two affine connections ∇ and ∇^*. If for all vector fields $X, Y, Z \in \mathfrak{X}(M)$,*

$$Z\langle X, Y\rangle = \langle \nabla_Z X, Y\rangle + \langle X, \nabla^*_Z Y\rangle$$

holds true, then we say ∇ and ∇^* are duals of each other with respect to g. In addition, we call such a triple (g, ∇, ∇^*) a dualistic structure on M.

The invariance of inner product under parallel transport also holds true for dual connections, so we have the following theorem.

Theorem 2.21. *(Invariance of inner product) Let Π_γ and Π^*_γ denote the parallel transport along a curve γ which connects two points p and q with respect to ∇ and ∇^* respectively. Then for all $X, Y \in T_pM$, we have*

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

Proof. Let γ be a curve in M and we want to parallel transport X, Y along the curve γ.
Let $\frac{DX}{dt}$ and $\frac{D^*Y}{dt}$ denote the covariant derivatives of X, Y with respect to ∇, ∇^* respectively.
Since ∇ and ∇^* are dual connections, we have

$$\frac{d}{dt}\langle \Pi_\gamma(X), \Pi^*_\gamma(Y)\rangle_q = \langle \frac{D\Pi_\gamma(X)}{dt}, \Pi^*_\gamma(Y)\rangle + \langle \Pi_\gamma(X), \frac{D\Pi^*_\gamma(Y)}{dt}\rangle. \tag{3}$$

Since $\Pi_\gamma(X)$ and $\Pi^*_\gamma(Y)$ are parallel along γ, we have $\frac{D\Pi_\gamma(X)}{dt} = 0$ and $\frac{D\Pi^*_\gamma(Y)}{dt} = 0$.
Then,

$$\frac{d}{dt}\langle \Pi_\gamma(X), \Pi^*_\gamma(Y)\rangle_q = 0$$

This implies $\langle \Pi_\gamma(X), \Pi^*_\gamma(Y)\rangle_q = \textit{constant}$ along the curve γ.
Therefore, the theorem holds true. □

In fact, the theorem above completely determines the relationship between Π_γ and Π^*_γ. Once we know Π_γ, we immediately know Π^*_γ and vice versa. Therefore, we have the following theorem, which is the highlight of this subsection.

Theorem 2.22. *(Dually flat manifold) Let the curvature tensors of ∇ and ∇^* be denoted by R and R^* respectively. Then $R = 0$ if and only if $R^* = 0$. In other words, the manifold is flat with respect to ∇ if and only if it's flat with respect to ∇^*.*

Proof. We only prove for one direction since the proof for another direction is similar.
Let $p, q \in M$ and γ be a curve connecting them. Let $X, Y \in T_pM$.
Suppose $R = 0$. Then the parallel transport won't change any vector field, so we have

$$X = \Pi_\gamma(X).$$

By theorem 2.21, we have

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

Since this holds true for all Y, we have $Y = \Pi^*_\gamma(Y)$.
Since this holds true for any curve, we have $Y = \Pi^*(Y)$.
Hence, $R^* = 0$. □

From this theorem, we see that a manifold is dually flat when it's flat with respect to one connection. Now suppose we have a dually flat manifold M equipped with (g, ∇, ∇^*), then from the definition, there exists an affine coordinate system $\{\theta_i\}$ with respect to connection ∇ and an affine coordinate system $\{\eta^j\}$ with respect to ∇^*. For convenience, we denote basis vector fields by $e_i = \frac{\partial}{\partial\theta_i}$ and $e^j = \frac{\partial}{\partial\eta^j}$. Since $\{\theta_i\}$ and $\{\eta^j\}$ are affine coordinate systems, by definition, both of them are parallel on M. Then, by theorem 2.21, $\langle e_i, e^j\rangle$ is constant on M. Therefore, we have the following definition.

Definition 2.23. *(Dual coordinate system) If two coordinate systems $\{\theta_i\}$ and $\{\eta^j\}$ for a Riemannian manifold satisfy the condition $\langle e_i, e^j\rangle = \delta^j_i$ where δ is the Kronecker delta, then we say these two coordinate systems are mutually dual and one is the dual coordinate system of the other.*

Remark 2.24. If for a Riemannian manifold (M, g), there exists such mutually dual coordinate systems, then the connections ∇ and ∇^* for which they are affine are determined and (M, g, ∇, ∇^*) is a dually flat manifold.

We define components of the Riemannian metric in the mutually dual coordinates as the following:

$$g_{ij} = \langle e_i, e_j\rangle \quad \text{and} \quad g^{*ij} = \langle e^i, e^j\rangle \tag{4}$$

In fact, g_{ij} and g^{*ij} have nice relationships.

Theorem 2.25. *(Relationships between g and g^*) $g_{ij}g^{*jk} = \delta^k_i$ where δ is the Kronecker delta.*

Proof. Notice that

$$\frac{\partial}{\partial\eta^j} = \frac{\partial\theta_i}{\partial\eta^j}\frac{\partial}{\partial\theta_i} \quad \text{and} \quad \frac{\partial}{\partial\theta_i} = \frac{\partial\eta^j}{\partial\theta_i}\frac{\partial}{\partial\eta^j}. \tag{5}$$

Therefore,

$$g_{ij} = \langle e_i, e_j \rangle = \langle \frac{\partial}{\partial \theta_i}, \frac{\partial}{\partial \theta_j} \rangle = \langle \frac{\partial \eta^j}{\partial \theta_i} \frac{\partial}{\partial \eta^j}, \frac{\partial}{\partial \theta_j} \rangle = \frac{\partial \eta^j}{\partial \theta_i} \langle e^j, e_j \rangle$$

Since $\{\theta_i\}$ and $\{\eta^j\}$ are mutually flat, then $\langle e^j, e_j \rangle = 1$. Therefore, we obtain

$$g_{ij} = \frac{\partial \eta^j}{\partial \theta_i}. \tag{6}$$

Similarly,

$$g^{*ij} = \frac{\partial \theta_i}{\partial \eta^j}. \tag{7}$$

Then,

$$g_{ij} g^{*jk} = \frac{\partial \eta^j}{\partial \theta_i} \frac{\partial \theta_j}{\partial \eta^k} = \delta_i^k$$

□

Based on this relationship, we can consider the following partial differential equation

$$\frac{\partial}{\partial \theta_i} \varphi = \eta^i.$$

If we rewrite this PDE as $d\varphi = \eta^j d\theta_i$, we will see the solution to this PDE exists if and only if $\frac{\partial \eta^j}{\partial \theta_i} = \frac{\partial \theta_i}{\partial \eta^j}$. Because of equation (6) and (7), we see both of them are g_{ij}, so solution to the PDE exists. Moreover,

$$\frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_j} \varphi = \frac{\partial}{\partial \theta_i} \eta^j = g_{ij}.$$

Since the metric is positive definite, this implies the Hessian of φ is positive definite. Therefore, φ is a convex function. Based on this observation, we will discuss the dual flatness from Bregman divergence.

2.3 Dual flatness from Bregman Divergence

Recall the Bregman divergence D_φ and its dual divergence D_φ^* we defined before. Since D and D^* are defined in terms of convex functions $\varphi(\xi)$ and $\varphi^*(\xi^*)$, based on the discussion from section 2.2, it's nature to define the Riemannian metric by

$$g_{ij} = \frac{\partial}{\partial \xi_i} \frac{\partial}{\partial \xi_j} \varphi(\xi) \quad \text{and} \quad g^{*ij} = \frac{\partial}{\partial \xi_i^*} \frac{\partial}{\partial \xi_j^*} \varphi^*(\xi^*).$$

Recall that in theorem 2.7, we proved the self-dual expression of Bregman divergence. If we define following notations

$$D_{i;} = \frac{\partial}{\partial \xi_{P_i}} D[P:Q]_{P=Q} \quad \text{and} \quad D_{;i} = \frac{\partial}{\partial \xi_{Q_i}^*} D[P:Q]_{P=Q}$$

by using self-dual expression, it's easy to verify that $g_{ij} = -D_{i;j}$. Moreover, if we denote $e_i = \frac{\partial}{\partial \xi_i}$ and $e^{*j} = \frac{\partial}{\partial \xi_j^*}$, then $\langle e_i, e^{*j} \rangle = \delta_i^j$. Therefore, $\{\xi_i\}$ and $\{\xi_i^*\}$ are two affine coordinates and they are mutually dual coordinates. They give the Riemannian manifold M a dually flat structure.

Now since $\{\xi_i\}$ and $\{\xi_i^*\}$ are affine coordinate systems, we can consider each coordinate axis of ξ_i, ξ_i^* as a straight line. Moreover, any curve in the form of $\boldsymbol{\xi}(t) = at + b$ where a, b are constants and t is a parameter is a straight line in ξ coordinates. Similarly, $\boldsymbol{\xi^*}(t) = at + b$ is a straight line in ξ^* coordinates. Here, we call $\boldsymbol{\xi}(t)$ a geodesic and $\boldsymbol{\xi^*}(t)$ a dual geodesic.

Remark 2.26. Here, geodesic doesn't mean the shortest curve that connects two points.

Dually flat manifold is still different from Euclidean space. It's a generalization of Euclidean space [4]. The orthogonality is different from orthogonality in the Euclidean space.

Definition 2.27. *(Orthogonality) Two curves $\theta_1(t)$ and $\theta_2(t)$ intersect orthogonally when their tangent vectors*

$$\dot{\theta}_1(t) = \frac{d}{dt}\theta_1(t)$$
$$\dot{\theta}_2(t) = \frac{d}{dt}\theta_2(t)$$

are orthogonal, that is,

$$\langle \dot{\theta}_1(t), \dot{\theta}_2(t) \rangle = \sum_{ij} g_{ij} \dot{\theta}_{1i}(t) \dot{\theta}_{2j}(t) = 0$$

at the intersection point t when $\theta_1(t) = \theta_2(t)$.

Since dually flat manifold is a generalization of Euclidean space, the generalized Pythagorean theorem [4] holds true.

Theorem 2.28. *(Generalized Pythagorean Theorem) Let P, Q, R be in a dually flat manifold M. When triangle PQR is orthogonal such that the geodesic connecting P and Q is orthogonal to the dual geodesic connecting Q and R, the following generalized Pythagorean relation holds:*

$$D_\psi[P : R] = D_\psi[P : Q] + D_\psi[Q : R].$$

Proof. By using the self-dual expression of D_ψ, we have

$$\begin{aligned}
&D_\psi[P : Q] + D_\psi[Q : R] - D_\psi[P : R] \\
&= \psi(\theta_P) + \psi^*(\theta_Q^*) - \theta_P \cdot \theta_Q^* + \psi(\theta_Q) + \psi^*(\theta_R^*) - \theta_Q \cdot \theta_R^* - \psi(\theta_P) - \psi^*(\theta_R^*) + \theta_P \cdot \theta_R^* \\
&= \psi^*(\theta_Q^*) - \theta_P \cdot \theta_Q^* + \psi(\theta_Q) - \theta_Q \cdot \theta_R^* + \theta_P \cdot \theta_R^*
\end{aligned}$$

Since $\psi^*(\theta^*) = \theta \cdot \theta^* - \psi(\theta)$, we have

$$\begin{aligned}
&D_\psi[P : Q] + D_\psi[Q : R] - D_\psi[P : R] \\
&= \theta_Q \cdot \theta_Q^* - \psi(\theta_Q) - \theta_P \cdot \theta_Q^* + \psi(\theta_Q) - \theta_Q \cdot \theta_R^* + \theta_P \cdot \theta_R^* \\
&= (\theta_Q^* - \theta_R^*) \cdot (\theta_Q - \theta_P)
\end{aligned}$$

The dual geodesic connecting Q and R can be represented as

$$\theta_{QR}^*(t) = t(\theta_Q^* - \theta_R^*) + \theta_R^*$$

and the geodesic connecting Q and P can be represented as

$$\theta_{QP} = t(\theta_Q - \theta_P) + \theta_P.$$

Therefore, we have $\dot{\theta}^*_{QR}(t) = \theta^*_Q - \theta^*_R$ and $\dot{\theta}_{QP} = \theta_Q - \theta_P$.
Since the triangle PQR is orthogonal, the tangent vector of dual geodesic is orthogonal to the tangent vector of geodesic.
Therefore, $(\theta^*_Q - \theta^*_R) \cdot (\theta_Q - \theta_P) = 0$.
Hence, the Pythagorean relation holds true. □

As we proved before, dual divergence of Bregman divergence satisfies proposition 2.6. Therefore, we have the following Dual Pythagorean Theorem.

Theorem 2.29. *(Dual Pythagorean Theorem) When triangle PQR is orthogonal such that the dual geodesic connecting P and Q is orthogonal to the geodesic connecting Q and R, the dual of the generalized Pythagorean relation holds,*

$$D_{\psi^*}[P : R] = D_{\psi^*}[P : Q] + D_{\psi^*}[Q : R].$$

Proof. By using the self-dual expression $D_{\psi^*}[P : Q] = \psi^*(\theta^*_P) + \psi(\theta_Q) - \theta^*_P \cdot \theta_Q$ and orthogonality of the triangle, the proof of Dual Pythagorean Theorem is same as the proof of Generalized Pythagorean Theorem. □

Now consider a point p in a dually flat manifold M and a submanifold S of M. We can define the divergence from this point p to submanifold S [4].

Definition 2.30. *(Divergence betwee a point and submanifold) The divergence from a point P to a submanifold S is defined by*

$$D_\psi[P : S] = \min_{R \in S} D_\psi[P : R].$$

The Pythagorean theorem we just proved will be useful in finding minimum [4].

Definition 2.31. *$\hat{P}_S$ is the geodesic projection of P to S when the geodesic connecting P and $\hat{P}_S \in S$ is orthogonal to S. Dually, $\hat{P}^*_S$ is the dual geodesic projection of P to S, when the dual geodesic connecting P and $\hat{P}^*_S \in S$ is orthogonal to S.*

Then we have the following projection theorem.

Theorem 2.32. *(Projection Theorem) Given $P \in M$ and a smooth submanifold $S \subset M$, the point $\hat{P}^*_S$ that minimizes the divergence $D_{\psi^*}[P : R], R \in S$ is the dual geodesic projection of P to S. The point $\hat{P}_S$ that minimizes the dual divergence $D_\psi[P : R], R \in S$ is the geodesic projection of P to S.*

Proof. Let $\hat{P}^*_S$ be the geodesic projection of P to S and let Q be a point infinitely close to $\hat{P}^*_S$. Then these three points $P, \hat{P}^*_S, Q$ form an orthogonal triangle. By the Dual Pythagorean Theorem, we know that

$$D_{\psi^*}[P : \hat{P}^*_S] + D_{\psi^*}[\hat{P}^*_S : Q] = D_{\psi^*[P:Q]}$$

which means $\hat{P}^*_S$ is the critical point.
The same idea can be applied to $\hat{P}_S$. □

Generalized Pythagorean Theorem and Projection Theorem are highlights in this subsection, which will be pretty useful when we discuss about orthogonal foliation.

2.4 Orthogonal Foliation

In this subsection, we are going to discuss about one important application of properties of dually flat manifolds. Because of the dual flatness, we can consider a series of flat hierarchical structures [2]. Let $S \subset M$ be a submanifold of a dually flat manifold M. Suppose we have an e-affine coordinates $\{\theta_i\}$ and an m-affine coordinates $\{\eta_i\}$.

Definition 2.33. *(e-flatness and m-flatness) The submanifold S is called an e-flat submanifold if it is a linear subspace in $\{\theta_i\}$ coordinates and it is called an m-flat submanifold if it is a linear subspace in $\{\eta_i\}$ coordinates.*

Remark 2.34. Note that an e-flat submanifold S is an e-flat manifold itself. Because of theorem 2.22, we see that a manifold is dually flat if it is flat with respect to one connection. Therefore, S is also an m-flat manifold. However, S may not be linear in η coordinates, so being m-flat manifold doesn't mean S is an m-flat submanifold of M. This definition will show up again in a more rigorous discussion in section 3.1.

It's more convenient to introduce a new coordinate system, k-cut coordinate system, in order to define orthogonal foliations [2].

Definition 2.35. *(k-cut) A new coordinate system called the k-cut mixed ones*

$$\boldsymbol{\xi_k} = (\boldsymbol{\eta}_{k^-}; \boldsymbol{\theta}_{k^+}) = (\eta_1, \ldots, \eta_k; \theta_{k+1}, \ldots, \theta_n)$$

consists of a pair of complementary parts of η and θ, namely,

$$\boldsymbol{\eta}_{k^-} = (\eta_1, \eta_2, \ldots, \eta_k) \tag{8}$$
$$\boldsymbol{\theta}_{\boldsymbol{k}^+} = (\theta_{k+1}, \theta_{k+2}, \ldots, \theta_n). \tag{9}$$

This is defined for any non-negative integer k.

With this notation, we define a subset $\boldsymbol{E_k}(\boldsymbol{c_{k^+}})$ where $\boldsymbol{c_{k^+}} = (c_{k+1}, \ldots, c_n)$ is a constant vector. The subset $\boldsymbol{E_k}(\boldsymbol{c_{k^+}})$ consists of all probability distributions with the same $\boldsymbol{\theta_{k^+}}$ coordinates. Their $\boldsymbol{\theta_{k^+}}$ coordinates are specified by $\boldsymbol{c_{k^+}} = (c_{k+1}, \ldots, c_n)$ while the other θ coordinates are still free. Here, the integer k actually represents the number of θ coordinates that are still free. In other words,

$$\boldsymbol{E_k}(\boldsymbol{c_{k^+}}) = \{p(\boldsymbol{x}, \boldsymbol{\theta}) | \boldsymbol{\theta_{k^+}} = \boldsymbol{c_{k^+}}\}. \tag{10}$$

This is an e-flat submanifold. If we union all different $\boldsymbol{c_{k^+}}$, then we will get the whole manifold M. This means

$$\bigcup_{\boldsymbol{c_{k^+}}} \boldsymbol{E_k}(\boldsymbol{c_{k^+}}) = M. \tag{11}$$

Then, we derive the definition for hierarchical e-structure [2].

Definition 2.36. *(Hierarchical e-structure) The hierarchical e-structure is introduced in M by putting $\boldsymbol{c_{k^+}} = 0$*

$$\boldsymbol{E}_1(0) \subset \boldsymbol{E}_2(0) \subset \cdots \subset \boldsymbol{E}_n(0) = S. \tag{12}$$

The above definitions can also be defined for η coordinates. The subset $\boldsymbol{M_k}(\boldsymbol{d_{k^-}}) = \{p(\boldsymbol{x}, \boldsymbol{\theta}) | \boldsymbol{\eta_{k^-}} = \boldsymbol{d_{k^-}}\}$ where $\boldsymbol{d_{k^-}} = (d_1, \ldots, d_k)$ is a constant vector consists of all probability distributions with the same $\boldsymbol{\eta_{k^-}}$ coordinates. Their $\boldsymbol{\eta_{k^-}}$ coordinates are specified by $d_{k^-} = (d_1, \ldots, d_k)$ while the other η coordinates are free. here, the integer k represents the number of η coordinates that are fixed. This is an m-flat submanifold. The two foliations $\boldsymbol{M_k}$ and $\boldsymbol{E_k}$ are actually orthogonal [2].

Theorem 2.37. *(Orthogonal Foliation) Any two foliations $\boldsymbol{M_k}$ and $\boldsymbol{E_k}$ are orthogonal in the sense that $\boldsymbol{M_k}$ and $\boldsymbol{E_k}$ are complementary and orthogonal at any point.*

Proof. Note that

$$\boldsymbol{E_k} = \{p(\boldsymbol{x}, \boldsymbol{\theta}) | (\theta_1, \ldots, \theta_k, \theta_{k+1}, \ldots, \theta_m) = (\theta_1, \ldots, \theta_k, c_{k+1}, \ldots, c_n)\}$$

and

$$\boldsymbol{M_k} = \{p(\boldsymbol{x}, \boldsymbol{\theta}) | (\eta_1, \ldots, \eta_k, \eta_{k+1}, \ldots, \eta_n) = (d_1, \ldots, d_k, \eta_{k+1}, \ldots, \eta_n)\}.$$

The only intersection point of $\boldsymbol{M_k}$ and $\boldsymbol{E_k}$ is, in k-cut coordinates, $\boldsymbol{\xi_k} = (d_1, \ldots, d_k; c_{k+1}, \ldots, c_n)$. The basis of tangent space of the intersection point in θ coordinates is given by $\{e_1, \ldots, e_k\}$ and the basis in η coordinates is given by $\{e^{k+1}, \ldots, e^n\}$.
Since M is dually flat, $\langle e_i, e^j \rangle = 0$ for different i, j. Therefore, the inner product between any vector in $\text{span}\{e_1, \ldots, e_k\}$ and any vector in $\text{span}\{e^{k+1}, \ldots, e^n\}$ are orthogonal. Therefore, $\boldsymbol{M_k}$ and $\boldsymbol{E_k}$ are orthogonal. □

Since $\boldsymbol{M_k}$ and $\boldsymbol{E_k}$ are orthogonal, we can apply Generalized Pythagorean Theorem here to decompose a probability distribution into different foliations. Suppose we have a probability distribution $p(\boldsymbol{x}, \boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. We want to decompose it into different $\boldsymbol{E_k}(0)$ where $\boldsymbol{E_k}(0)$ represents the submanifold that includes distributions that do not have effect higher than k because $(\theta_{k+1}, \ldots, \theta_k) = 0$. Recall that in definition 2.30, we defined the divergence between a point and a submanifold. Now we define information projection in a similar manner.

Definition 2.38. *Given a probability distribution p, the information projection of p to $\boldsymbol{E_k}(0)$ is defined by*

$$p^{(k)}(x) = \arg \min_{q \in \boldsymbol{E_k}(0)} D[p(x), q(x)]$$

This projection point $p^{(k)}$ is the closest point in $\boldsymbol{E_k}(0)$ to p in sense of divergence. Since there is always an $\boldsymbol{M_k}$ that contains p and this $\boldsymbol{M_k}$ is orthogonal to $\boldsymbol{E_k}(0)$, there is always an m-geodesic that connects p and $p^{(k)}$. Moreover, for any point $q \in \boldsymbol{E_k}(0)$, there is always an e-geodesic that connects $p^{(k)}$ and q since $\boldsymbol{E_k}(0)$ is a linear space in θ coordinates. Therefore, by Generalized Pythagorean Theorem, we obtain

$$D[p : q] = D[p : p^{(k)}] + D[p^{(k)} : q]. \tag{13}$$

Based on this idea, we have the following orthogonal decomposition.

Theorem 2.39. *(Orthogonal Decomposition) Suppose θ coordinates for p is $(\theta_1, \ldots, \theta_n)$ and the θ coordinates for $p^(k) \in \boldsymbol{E_k}(0)$ is $(\theta_1, \ldots, \theta_k, 0, \ldots, 0)$. Let $(\theta_1, \ldots, \theta_j, 0, \ldots, 0)$ denote the coordinate of $p^{(j)} \in \boldsymbol{E_j}(0)$ where $k \leq j \leq n-1$. Then we have*

$$D[p : p^{(k)}] = \sum_{j=k}^{n-1} D[p^{(j+1)} : p^{(j)}]$$

where $p^{(n)} = p$.

Proof. Note that $p^{(j)} \in \boldsymbol{E_{j+1}}(0)$ for all j. Then by what we discussed before,

$$D[p : p^{(j)}] = D[p : p^{(j+1)}] + D[p^{(j+1)} : p^{(j)}].$$

Then, by induction,

$$D[p : p^{(k)}] = \sum_{j=k}^{n-1} D[p^{(j+1)} : p^{(j)}]$$

holds true. □

Orthogonal foliation is useful in sense that it gives us a way to single out the amount of k-th order effects in $p(\boldsymbol{x}, \boldsymbol{\theta})$.

3 Exponential Families of Probability Distributions

In this section, we are going to discuss about exponential family of probability distributions since it is closely related with Bregman divergence. The standard form of an exponential family is defined as the following [4].

Definition 3.1. *(Exponential family) The standard form of an exponential family is given by*

$$p(x, \boldsymbol{\theta}) = \exp\{\theta^i h_i(x) + k(x) - \psi(\boldsymbol{\theta})\} \tag{14}$$

where x *is a random variable.* $\boldsymbol{\theta} = (\theta^1, \ldots, \theta^n)$ *is an* n*-dimensional vector parameter to specify a distribution,* $h_i(x)$ *are* n *functions of* x *which are linearly independent,* $k(x)$ *is a function of* x*,* ψ *corresponds to the normalization factor and the Einstein summation convention is working.*

Exponential family of probability distributions include many well-known family of probability distributions such as Gaussian distributions and discrete probability distributions. Therefore, the exponential family of probability distributions is of particular importance.

3.1 Flat Structure

Exponential family of probability distributions is associated with a convex function which is called cumulant generating function in statistics. If we introduce a new variable $\boldsymbol{x} = (x_1, \ldots, x_n)$ by letting $x_i = h_i(x)$ and introduce a measure in the sample space by $d\mu(\boldsymbol{x}) = \exp\{k(x)\}dx$, then we can rewrite the probability distribution as

$$p(\boldsymbol{x}, \boldsymbol{\theta})dx = \exp\{\boldsymbol{\theta} \cdot \boldsymbol{x} - \psi(\boldsymbol{\theta})\}d\mu(x). \tag{15}$$

Then, with respect to measure $d\mu(x)$, the exponential family of probability distributions is in form of

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\{\boldsymbol{\theta} \cdot \boldsymbol{x} - \psi(\boldsymbol{\theta})\}. \tag{16}$$

Notice that different $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ specifies different probability distributions. Therefore, the family of distributions $M = \{p(\boldsymbol{x}, \boldsymbol{\theta})\}$ forms a manifold of dimension n and $\boldsymbol{\theta}$ is a global coordinate system.

Next, we want to explore the geometry of this manifold M. Notice that a probability density satisfies the normalization condition

$$\int p(\boldsymbol{x}, \boldsymbol{\theta}) d\mu(x) = 1.$$

From this condition, we see

$$\int \exp\{\boldsymbol{\theta} \cdot \boldsymbol{x} - \psi(\boldsymbol{\theta})\} d\mu(x) = 1,$$

so we derive the following convex function

$$\psi(\boldsymbol{\theta}) = \log \int \exp(\boldsymbol{\theta} \cdot \boldsymbol{x}) d\mu(x) = 1. \tag{17}$$

This convex function $\psi(\boldsymbol{\theta})$ is the cumulant generating function [7]. By using $\psi(\boldsymbol{\theta})$, we can introduce a dually flat structure to the manifold M. The affine coordinate system induced by $\psi(\theta)$ is given by $\boldsymbol{\theta}$. The Legendre transformation will give another affine coordinate system

$$\boldsymbol{\eta} = \nabla \psi(\boldsymbol{\theta}).$$

Therefore, $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are two affine coordinate systems connected by the Legendre transformation for the manifold M [4]. These two affine coordinate systems, by definition, give a dually flat structure to M.

Now we want to find the Riemannian metric of the manifold M. Let the Legendre transformation of $\psi(\boldsymbol{\theta})$ be denoted by $\phi(\boldsymbol{\eta})$.

Proposition 3.2. *The Legendre dual $\phi(\boldsymbol{\eta})$ is given by the negative entropy,*

$$\phi(\boldsymbol{\eta}) = \int p(\boldsymbol{x}, \boldsymbol{\theta}) \log p(\boldsymbol{x}, \boldsymbol{\theta}) dx \tag{18}$$

where $\boldsymbol{\theta}$ is regarded as function of $\boldsymbol{\eta}$ through $\boldsymbol{\eta} = \nabla \psi(\boldsymbol{\theta})$ [4].

Proof. In order to obtain $\phi(\boldsymbol{\eta})$, we calculate the negative entropy of p, which is

$$E[\log p(\boldsymbol{x}, \boldsymbol{\theta})] = \int p(\boldsymbol{x}, \boldsymbol{\theta}) \log p(\boldsymbol{x}, \boldsymbol{\theta}) d\mu(x) = \boldsymbol{\theta} \cdot \boldsymbol{\eta} - \psi(\boldsymbol{\theta}).$$

Notice that the right-hand side of the equation is maximized when $\boldsymbol{\eta} = \nabla \psi(\boldsymbol{\theta})$ as we proved before. Therefore, when $\boldsymbol{\eta} = \nabla \psi(\boldsymbol{\theta})$, we see that $\phi(\boldsymbol{\eta})$ is given by the negative entropy. □

One more interesting result is about the relationship between KL-divergence and Bregman divergence, which can be easily derived from computations.

Proposition 3.3. *(KL-divergence and Bregman divergence) The Bregman divergence from $p(\boldsymbol{x}, \boldsymbol{\theta}')$ to $p(\boldsymbol{x}, \boldsymbol{\theta})$ satisfies*

$$D_\psi[\boldsymbol{\theta}' : \boldsymbol{\theta}] = D_{KL}[\boldsymbol{\theta} : \boldsymbol{\theta}'].$$

As we discussed in section 2.3, if we have a pair of convex functions $\psi(\boldsymbol{\theta})$ and $\phi(\boldsymbol{\eta})$, it's nature to define the Riemannian metric by

$$g_{ij} = \frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\psi(\boldsymbol{\theta}) \quad \text{and} \quad g^{ij} = \frac{\partial}{\partial\eta_i}\frac{\partial}{\partial\eta_j}\phi(\boldsymbol{\eta}).$$

In fact, the Riemannian metric here is the Fisher information matrix [4].

Theorem 3.4. *The Riemannian metric in an exponential family is the Fisher information matrix defined by*

$$g_{ij} = E[\partial_i \log p(\boldsymbol{x}, \boldsymbol{\theta})\partial_j p(\boldsymbol{x}, \boldsymbol{\theta})]$$

where $\partial_i = \frac{\partial}{\partial\theta_i}$.

Proof.

$$\partial_i \log p(\boldsymbol{x}, \boldsymbol{\theta}) = \partial_i\{\boldsymbol{x} \cdot \boldsymbol{\theta} - \psi(\boldsymbol{\theta})\} = x_i - \partial_i\psi(\boldsymbol{\theta}) = x_i - \eta_i$$

Therefore,

$$E[\partial_i \log p(\boldsymbol{x}, \boldsymbol{\theta})\partial_j \log p(\boldsymbol{x}, \boldsymbol{\theta})] = E[(x_i - \eta_i)(x_j - \eta_j)].$$

Since $\nabla\psi(\boldsymbol{\theta}) = \boldsymbol{\eta} = E[\boldsymbol{x}]$, then the ij-th entry of $\nabla\nabla\psi(\boldsymbol{\theta})$ is given by $E[(x_i - \eta_i)(x_j - \eta_j)]$. Hence, the Riemannian metric is the Fisher information matrix. □

Now we will consider the geodesic, which is the straight line as defined in section 2.3, in this manifold M. Since $\boldsymbol{\theta}$ is an affine coordinate system, then the geodesic connection two distributions $p(\boldsymbol{x}, \boldsymbol{\theta_1})$ and $p(\boldsymbol{x}, \boldsymbol{\theta_2})$ is given by

$$\boldsymbol{\theta}(t) = (1 - t)\boldsymbol{\theta_1} + t\boldsymbol{\theta_2} \tag{19}$$

in $\boldsymbol{\theta}$ coordinates. Now the probability distribution that is on this geodesic can be represented as

$$p(\boldsymbol{x}, \boldsymbol{\theta}(t)) = \exp\{\boldsymbol{x} \cdot ((1 - t)\boldsymbol{\theta}_1 + t\boldsymbol{\theta}_2) - \psi(\boldsymbol{\theta}(t))\}. \tag{20}$$

Since $\boldsymbol{\theta(t)}$ is totally dependent on the parameter t, we can rewrite equation (20) as

$$p(\boldsymbol{x}, t) = \exp\{t(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) \cdot \boldsymbol{x} + \boldsymbol{\theta}_1 \cdot \boldsymbol{x} - \psi(t)\}. \tag{21}$$

Therefore, a geodesic itself is a one-dimensional exponential family where t is the parameter [4]. By taking the logarithm, we will see that

$$\log p(\boldsymbol{x}, t) = (1 - t)\log p(\boldsymbol{x}, \boldsymbol{\theta}_1) + t \log p(\boldsymbol{x}, \boldsymbol{\theta}_2) - \psi(t). \tag{22}$$

Since equation (22) is in a linear form, we motivate the following definition which has already shown up in section 2.4.

Definition 3.5. *(e-flatness) An e-geodesic in an exponential family is in the form of*

$$\log p(\boldsymbol{x}, t) = (1 - t)\log p(\boldsymbol{x}, \boldsymbol{\theta}_1) + t \log p(\boldsymbol{x}, \boldsymbol{\theta}_2) - \psi(t).$$

More generally, a submanifold which is defined by constraints in $\boldsymbol{\theta}$ is said to be e-flat. The affine parameter $\boldsymbol{\theta}$ is called the e-affine parameter.

This finishes the discussion about flatness of manifolds of exponential family.

3.2 Principal Component Analysis

One important application of Bregman divergence is the Principal Component Analysis (PCA) of the exponential family of probability distributions. Readers can refer to [15] for a most recent survey. There is a series of papers [1], [20], [18], [17], [16] discussing about PCA of the exponential family and its extensions. Let $\{P_{\boldsymbol{\theta}}\}$ be any set of probability distributions from the exponential family. Given data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, the goal is to find parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ which lie in a low-dimensional subspace and for which the log-likelihood $\sum_i \log P_{\boldsymbol{\theta}_i}(\boldsymbol{x}_i)$ is maximized [11]. Therefore, we will first take a look at the conditional probability of a value x given parameter value θ in the exponential family of distributions where x and θ are both in $\mathbb{R}$. Recall expression of the exponential family (14).

$$P(x|\theta) = \exp\{x\theta + k(x) - \varphi(\theta)\}$$

$$\log P(x|\theta) = x\theta + k(x) - \varphi(\theta)$$

Notice that when $\theta = 0$, $P(x|0) = \exp\{k(x)\}$. Therefore, we denote $\log P_0(x) = k(x)$. Then we rewrite the log-likelihood as the following:

$$\log P(x|\theta) = \log P_0(x) + x\theta - \varphi(\theta). \tag{23}$$

Recall that if $\psi(\eta)$ is the Legendre transformation of convex function $\varphi(\theta)$, then $\psi(\eta) = \theta\eta - \varphi(\theta)$. Here $\eta = \nabla\varphi(\theta) = f(\theta)$, so $\theta = f^{-1}(\eta) = g(\eta)$. Therefore,

$$\psi(f(\theta)) + \varphi(\theta) = \theta f(\theta) \tag{24}$$

which is an equation only dependent on θ.

Theorem 3.6. *The negative log-likelihood of $P(x, \theta)$ from the exponential family can be expressed as*

$$-\log P(x|\theta) = -\log P_0(x) - \psi(x) + D_\psi[x : f(\theta)].$$

Proof. From equation (23),

$$-\log P(x|\theta) = -\log P_0(x) - x\theta + \varphi(\theta).$$

Because of equation (24),

$$-\log P(x|\theta) = -\log P_0(x) - x\theta + \theta f(\theta) - \psi(f(\theta)).$$

Notice that $\theta = g(\eta)$ and $\eta = f(\theta)$, so $\theta = g(f(\theta))$.
Therefore,

$$\begin{aligned} -\log P(x|\theta) &= -\log P_0(x) - xg(f(\theta)) + g(f(\theta))f(\theta) - \psi(f(\theta)) \\ &= -\log P_0(x) - \psi(x) + \psi(x) - xg(f(\theta)) + g(f(\theta))f(\theta) - \psi(f(\theta)) \\ &= -\log P_0(x) - \psi(x) + \psi(x) - \psi(f(\theta)) - g(f(\theta))(x - f(\theta)) \end{aligned}$$

Since $g = f^{-1}$, then by dualistic structure of Legendre transformation, we have $g(\eta) = \nabla\psi(\eta)$. Therefore, $g(f(\theta)) = \nabla\psi(f(\theta))$.
Hence,

$$\begin{aligned} -\log P(x|\theta) &= -\log P_0(x) - \psi(x) + \psi(x) - \psi(f(\theta)) - \nabla\psi(f(\theta))(x - f(\theta)) \\ &= -\log P_0(x) - \psi(x) + D_\psi[x : f(\theta)]. \end{aligned}$$

□

Remark 3.7. Note that here all variables x, θ, η are in $\mathbb{R}$. By the above theorem, negative log-likelihood can always be written as Bregman divergence plus a term that is constant with respect to θ and which therefore can be ignored [11].

Now we want to extend this idea to divergences between vectors and matrices. Therefore, we introduce the following notations [11].

Notation 3.8. *If $\boldsymbol{x}, \boldsymbol{y}$ are vectors and $\boldsymbol{A}, \boldsymbol{B}$ are matrices, then we use the notation*

$$D_\psi[\boldsymbol{x} : \boldsymbol{y}] = \sum_i D_\psi[x_i : y_i]$$

and

$$D_\psi[\boldsymbol{A} : \boldsymbol{B}] = \sum_i \sum_j D_\psi[a_{ij} : b_{ij}]$$

where x_i, y_i, a_{ij}, b_{ij} are entries of $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{A}, \boldsymbol{B}$.

Now we extend our idea to general cases. As we discussed before, we want to find some $\boldsymbol{\theta}_i$'s which are close to $\boldsymbol{x}_i$'s and those $\boldsymbol{\theta}_i$'s belong to a lower dimensional subspace of parameter space. Therefore, we need to find a basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_l\}$ which spans a linear subspace in $\mathbb{R}^d$ such that $\boldsymbol{\theta}_i = \sum_k a_{ik}\boldsymbol{v}_k$. Let $\boldsymbol{X}$ denote the $n \times d$ matrix whose i'th row is $\boldsymbol{x}_i$. Let $\boldsymbol{V}$ be the $l \times d$ matrix whose k'th row is $\boldsymbol{v}_k$ and let $\boldsymbol{A}$ be the $n \times l$ matrix with elements a_{ik}.Then $\boldsymbol{\Theta} = \boldsymbol{AV}$ is an $n \times d$ matrix whose i'th row is $\boldsymbol{\theta}_i$. Since we want to find the set of parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ that maximizes the likelihood of the data, we define the following loss function

$$L(\boldsymbol{V}, \boldsymbol{A}) = -\log P(\boldsymbol{X}|\boldsymbol{V}, \boldsymbol{A}) = -\log P(\boldsymbol{X}|\boldsymbol{\Theta}) = -\sum_i \sum_j \log P(x_{ij}|\theta_{ij}) \tag{25}$$

Recall equation (23) which is the log-likelihood for x_{ij}, θ_{ij}. Then equation (25) becomes

$$L(\boldsymbol{V}, \boldsymbol{A}) = -\sum_i \sum_j \log P_0(x_{ij}) + x_{ij}\theta_{ij} - \varphi(\theta_{ij}) = C + \sum_i \sum_j (-x_{ij}\theta_{ij} + \varphi(\theta_{ij})) \tag{26}$$

where C is a constant term. From theorem 3.5, we can further rewrite the loss function as

$$L(\boldsymbol{V}, \boldsymbol{A}) = \sum_i \sum_j D_\psi[x_{ij} : f(\theta_{ij})]. \tag{27}$$

Once we find $\boldsymbol{V}$ and $\boldsymbol{A}$ for $\boldsymbol{x}_i$'s, we can represent $\boldsymbol{x}_i$ as a vector $\boldsymbol{a}_i \in \mathbb{R}^l$. Since $\boldsymbol{A}$ should minimize the loss function L and $\boldsymbol{\Theta} = \boldsymbol{AV}$, then $\boldsymbol{\theta}_i = \boldsymbol{a}_i\boldsymbol{V}$ and

$$\boldsymbol{a}_i = \arg\min_{\boldsymbol{a} \in \mathbb{R}^l} D_\psi[\boldsymbol{x}_i : f(\boldsymbol{aV})] \tag{28}$$

where $\boldsymbol{a}_i, \boldsymbol{a}, \boldsymbol{x}_i$ are all vectors. Once we find all $\boldsymbol{a}_i$'s, we get the matrix $\boldsymbol{A}$. Knowing $\boldsymbol{\Theta}$ and $\boldsymbol{A}$, we will search a low dimensional basis (matrix $\boldsymbol{V}$) which defines a surface that is close to all the data points $\boldsymbol{x}_i$ [11].

In summary, when we have a probability distribution P from the exponential family, we will automatically get a convex function $\varphi(\theta)$. Given this $\varphi(\theta)$ and $\boldsymbol{\Theta} = \boldsymbol{AV}$, we can define the loss function and minimize the loss function to find $\boldsymbol{A}$. Once we know $\boldsymbol{A}$, we will search for a matrix $\boldsymbol{V}$ which defines a surface that is close to all data points. This completes PCA for exponential family of probability distributions.

4 Dually Projective Flat Manifold and $L^{(\pm\alpha)}$ Divergence

In this section, we are going to discuss about dually projective flat manifold, which is an analogy to dually flat manifold. Moreover, we will also explore $L^{(\pm\alpha)}$ divergence, which is a generalization of Bregman divergence. The relationship between dually projective flat manifolds and $L^{(\pm\alpha)}$ divergence is same as the one between dually flat manifolds and Bregman divergence. In other words, $L^{(\pm\alpha)}$ has dualistic structure and will induce a dually projective flat structure on a manifold M. Now we assume all conditions are nice enough so that definitions and theorems that we will state later make sense.

4.1 $L^{(\pm\alpha)}$ Divergence

We will introduce $L^{(\pm\alpha)}$ divergence in a more general and abstract way, which is defined by c-divergence which is induced by a continuous cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Now, we will define the following definitions [28].

Definition 4.1. *Let $f : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ and $g : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$.*
(i) The c-transforms of f and g are defined respectively by

$$f^c(y) = \inf_{x \in \mathcal{X}} (c(x, y) - f(x)), \quad y \in \mathcal{Y}$$

and

$$g^c(x) = \inf_{y \in \mathcal{Y}} (c(x, y) - g(x)), \quad x \in \mathcal{X}.$$

(ii) We say that f (respectively g) is c-concave if $f^{cc} = f$ (respectively $g^{cc} = g$).
(iii) If f and g are c-concave, their c-superdifferentials are defined by

$$\partial^c f = \{(x, y) \in \mathcal{X} \times \mathcal{Y} : f(x) + f^c(y) = c(x, y)\}$$

and

$$\partial^c g = \{(x, y) \in \mathcal{X} \times \mathcal{Y} : g^c(x) + g(y) = c(x, y)\}.$$

Notice that if f is c-concave, then $f^c \leq c(x, y) - f(x)$. Therefore, we have the following inequality, which is called generalized Fenchel inequality

$$f(x) + f^c(y) \leq c(x, y). \tag{29}$$

The equality in equation (29) holds if and only if $(x, y) \in \partial^c f$. Therefore, we motivate the following definition [28].

Definition 4.2. *(c-gradient) Assume that f is c-differentiable in the following sense: for each $x \in \mathcal{X}$ there is a unique element $y \in \mathcal{Y}$ such that $(x, y) \in \partial^c f$. We call $y = D^c f(x)$ the c-gradient of f.*

Remark 4.3. We will see that the definition of c-transforms is analogous to the definition of Legendre transform. In addition, recall the Legendre transform of a convex function $\varphi(\xi)$ is $\varphi^*(\xi^*)$ where ξ^* is the gradient of $\varphi(\xi)$ and $\varphi^*(\xi^*) = \xi \cdot \xi^* - \varphi(\xi)$. Here, the c-transform of a c-concave function $f(x)$ is $f^c(y)$ where y is the c-gradient of $f(x)$ and $f^c(y) = c(x, y) - f(x)$. Because of such analogy, we will see later that Bregman divergence is a special case of $L^{(\pm\alpha)}$ divergence.

Now we can define c-divergence which is induced by a c-concave function [28].

Definition 4.4. *Let f be a c-concave and c-differentiable function on $\mathcal{X}$. The c-divergence of f is the functional $\boldsymbol{D}_f : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ defined by*

$$\boldsymbol{D}_f[x : x'] = c(x, y') - c(x', y') - (f(x) - f(x')), \quad x, x' \in \mathcal{X}, \quad y = D^c f(x').$$

If g is a c-differentiable and c-concave function on $\mathcal{Y}$, then we define the dual c-divergence $\boldsymbol{D}_g^* : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ by

$$\boldsymbol{D}_g^*[y : y'] = c(x', y) - c(x', y') - (g(y) - g(y')), \quad y, y' \in \mathcal{Y}, \quad x' = D^c g(y').$$

As Bregman divergence has a self-dual representation, c-divergence also has a self-dual representation.

Theorem 4.5. *(Self-dual representation) Let $y' = D^c f(x')$. Then*

$$\boldsymbol{D}_f[x : x'] = c(x, y') - f(x) - f^c(y').$$

Similarly, if $x = D^c g(y)$, then

$$\boldsymbol{D}_g^*[y : y'] = c(x', y) - g(y) - g^c(y').$$

Proof. Since $y = D^c f(x')$, then by Fenchel identity, $f(x') + f^c(y') = c(x', y')$.
Therefore,

$$\begin{aligned} D_f[x : x'] &= c(x, y') - c(x', y') - (f(x) - f(x')) \\ &= c(x, y') - c(x', y') - f(x) + c(x', y') - f^c(y') \\ &= c(x, y') - f(x) - f^c(y'). \end{aligned}$$

By the similar computation, we will obtain the self-dual expression of $\boldsymbol{D}_g^*$. □

Recall that for Bregman divergence and its dual divergence, we have the equality in proposition 2.6. Here, the asymmetry also holds true.

Corollary 4.6. *(Asymmetry of c-divergence) Suppose f^c is c-differentiable on the range of $D^c f$. Let $y = D^c f(x)$ and $y' = D^c f(x')$. Then*

$$\boldsymbol{D}_f[x' : x] = \boldsymbol{D}_{f^c}^*[y : y'].$$

With all these notations and theorems from c divergence, we can analysis $L^{(\pm\alpha)}$ divergence. In fact, $L^{(\pm\alpha)}$ divergence is a special case of c-divergence. Now we only consider $L^{(\alpha)}$ divergence since the analysis of $L^{(-\alpha)}$ is similar.

For $\alpha > 0$, we consider a specific continuous cost function on $\mathbb{R}^d$ given by

$$c^{(\alpha)}(x, y) = \frac{1}{\alpha} \log(1 + \alpha x \cdot y). \tag{30}$$

Then we can talk about c-concavity and c-gradient in terms of the cost function above [28].

Theorem 4.7. *Consider the cost function $c = c^{(\alpha)}(x, y)$ given by (30).*
(i) f is c-concave if and only if $e^{\alpha f}$ is concave on $\mathbb{R}^d$.
(ii) Suppose f is c-concave. The c-gradient of f at x, if exists, is given by

$$D^c f(x) = \frac{Df(x)}{1 - \alpha Df(x) \cdot x}.$$

As we have seen before, given a c-concave function f and a cost function c, we can define c-divergence. Now since we specify our cost function c, we can derive an explicit formula for the induced divergence. By using definition 4.4, we obtain

$$\begin{aligned} D_f[x : x'] &= \frac{1}{\alpha} \log(x + \alpha x \cdot y') - \frac{1}{\alpha}(1 + \alpha x' \cdot y') - (f(x) - f(x')) \\ &= \frac{1}{\alpha} \log\left(\frac{1 + \alpha x \cdot y'}{1 + \alpha x' \dot{y}'}\right) - (f(x) - f(x')). \end{aligned}$$

Since $y' = \frac{Df(x')}{1 - \alpha Df(x') \cdot x'}$, then

$$\begin{aligned} D_f[x : x'] &= \frac{1}{\alpha} \log\left(\frac{1 + \alpha x \cdot \frac{Df(x')}{1 - \alpha Df(x') \cdot x'}}{1 + \alpha x' \cdot \frac{Df(x')}{1 - \alpha Df(x') \cdot x'}}\right) - (f(x) - f(x')) \\ &= \frac{1}{\alpha} \log(1 - \alpha Df(x') \cdot x' + \alpha x \cdot Df(x')) - (f(x) - f(x')) \\ &= \frac{1}{\alpha} \log(1 + \alpha Df(x') \cdot (x - x')) - (f(x) - f(x')). \end{aligned}$$

Then, we can motivate the formal definition of exponential concavity, $L^{(\alpha)}$ divergence and α-gradient (which is same as the c-gradient in theorem 4.7 (ii)) [28].

Definition 4.8. *(Exponential concavity) Let $\Omega \subset \mathbb{R}^d$ be an open convex set, $\varphi : \Omega \to \mathbb{R}$, and fix $\alpha > 0$. Let $\Phi = e^{\alpha\varphi}$. We say that φ is α-exponentially concave if Φ is concave on Ω.*

Definition 4.9. *($L^{(\alpha)}$ divergence) If φ is a differentiable and α-exponentially concave function, we define the $L^{(\alpha)}$ divergence of φ by*

$$\boldsymbol{D}^{(\alpha)}[\xi : \xi'] = \frac{1}{\alpha} \log(1 + \alpha D\varphi(\xi') \cdot (\xi - \xi')) - (\varphi(\xi) - \varphi(\xi')), \xi, \xi' \in \Omega.$$

Remark 4.10. When $\alpha \to 0^+$, the $L^{(\alpha)}$ divergence is reduced to Bregman divergence defined for concave function. Notice that the Bregman divergence defined for concave function $\varphi(\xi)$ is given by

$$\boldsymbol{D}_\varphi[\xi : \xi'] = D\varphi(\xi') \cdot (\xi - \xi') - (\varphi(\xi) - \varphi(\xi')).$$

Definition 4.11. *(α-gradient) We define the α-gradient of φ at $\xi \in \Omega$ bt*

$$D^{(\alpha)}\varphi(\xi) = \frac{D\varphi(\xi)}{1 - \alpha D\varphi(\xi) \cdot \xi}$$

where $D\varphi$ is the Euclidean gradient.

Remark 4.12. Notice that when $\alpha = 0$, $D^{(0)}\varphi(\xi) = D\varphi(\xi)$ which is reduced to normal Euclidean gradient.

4.2 Dualistic Structure of $L^{(\pm\alpha)}$ Divergence

As Bregman divergence has dualistic structure, $L^{(\pm\alpha)}$ divergence also has dualistic structure. We will focus our discussion on $L^{(\alpha)}$ when $\alpha > 0$ since $L^{(-\alpha)}$ is similar. When we discussed about Bregman divergence and Legendre transformations, we defined dual coordinates by gradient. For $L^{(\alpha)}$ divergence, we can define dual coordinates by α-gradient.

Definition 4.13. *(Dual coordinate) We define the dual coordinate η by α-gradient*

$$\eta = D^{(\alpha)}\varphi(\xi).$$

Similarly, we can define α-conjugate (which first appears in definition 4.1 (i)) analogous to Legendre transformations by using η.

Definition 4.14. *(α-conjugate) Given the function $c^{(\alpha)} = \frac{1}{\alpha}\log(1+\alpha x\cdot y)$ defined for $x, y \in \mathbb{R}^d$ such that $1+\alpha x\cdot y > 0$, we define the α-conjugate ψ of φ on the open set $\Omega' = D^{(\alpha)}\varphi(\Omega)$ by*

$$\psi(\eta) = c^{(\alpha)}(\xi,\eta) - \varphi(\xi), \xi = (D^{(\alpha)}\varphi)^{-1}(\eta).$$

Next, we are going to show duality of α-conjugate.

Theorem 4.15. *(α-duality) For $\xi \in \Omega, \eta \in \Omega'$, we have*

$$\varphi(\xi) = \min_{\eta'\in\Omega'}(c^{(\alpha)}(\xi,\eta') - \psi(\eta')) \tag{31}$$

$$\psi(\eta) = \min_{\xi'\in\Omega}(c^{(\alpha)}(\xi',\eta) - \varphi(\xi')) \tag{32}$$

where $\eta = D^{(\alpha)}\varphi(\xi)$ and $\xi = (D^{(\alpha)}\varphi)^{-1}(\eta)$.

Proof. Recall that φ is exponentially concave if and only if $e^{\alpha\varphi}$ is concave. By concavity of $e^{\alpha\varphi}$, we have

$$e^{\alpha\varphi(\xi)} = \min_{\xi'\in\Omega} e^{\alpha\varphi(\xi')}(1-\alpha D\varphi(\xi')\cdot\xi' + \alpha D\varphi(\xi')\cdot\xi).$$

Taking logarithm in above equation, we have

$$\alpha\varphi(\xi) = \min_{\xi'\in\Omega}\{\alpha\varphi(\xi') + \log(1-\alpha D\varphi(\xi')\cdot\xi' + \alpha D\varphi(\xi')\cdot\xi)\}$$

$$\begin{aligned}
\varphi(\xi) &= \min_{\xi'\in\Omega}\{\varphi(\xi') + \frac{1}{\alpha}\log(1-\alpha D\varphi(\xi')\cdot\xi' + \alpha D\varphi(\xi')\cdot\xi)\} \\
&= \min_{\xi'\in\Omega}\{\varphi(\xi') + \frac{1}{\alpha}\log\left(\frac{1-\alpha D\varphi(\xi')\cdot\xi' + \alpha D\varphi(\xi')\cdot\xi}{1-\alpha D\varphi(\xi')\cdot\xi'}(1-\alpha D\varphi(\xi')\cdot\xi')\right)\} \\
&= \min_{\xi'\in\Omega}\{\varphi(\xi') + \frac{1}{\alpha}\log\left(1+\alpha\frac{D\varphi(\xi')}{1-\alpha D\varphi(\xi')\cdot\xi'}\cdot\xi\right) + \frac{1}{\alpha}\log(1-\alpha D\varphi(\xi')\cdot\xi')\}
\end{aligned}$$

Since $\eta' = D^{(\alpha)}\varphi(\xi') = \frac{D\varphi(\xi')}{1-\alpha D\varphi(\xi')\cdot\xi'}$, then we have

$$\frac{1}{\alpha}\log\left(1+\alpha\frac{D\varphi(\xi')}{1-\alpha D\varphi(\xi')\cdot\xi'}\cdot\xi\right) = c^{(\alpha)}(\xi,\eta')$$

and

$$1 + \alpha\xi' \cdot \eta' = \frac{1}{1 - \alpha D\varphi(\xi') \cdot \xi'}. \tag{33}$$

Therefore,

$$\begin{aligned}
\varphi(\xi) &= \min_{\xi' \in \Omega}\{\varphi(\xi') + c^{(\alpha)}(\xi, \eta') - \frac{1}{\alpha}\log(1 - \alpha D\varphi(\xi') \cdot \xi')^{-1}\} \\
&= \min_{\xi' \in \Omega}\{\varphi(\xi') + c^{(\alpha)}(\xi, \eta') - \frac{1}{\alpha}\log(1 + \alpha\eta' \cdot \xi')\} \\
&= \min_{\xi' \in \Omega}\{\varphi(\xi') + c^{(\alpha)}(\xi, \eta') - c^{(\alpha)}(\xi', \eta')\} \\
&= \min_{\xi' \in \Omega}\{c^{(\alpha)}(\xi, \eta') - (c^{(\alpha)}(\xi', \eta') - \varphi(\xi'))\} \\
&= \min_{\eta' \in \Omega'}\{c^{(\alpha)}(\xi, \eta') - \psi(\eta')\}
\end{aligned}$$

which gives us equation (31).
Now since $\varphi(\xi) = \min_{\eta' \in \Omega'}(c^{(\alpha)}(\xi, \eta') - \psi(\eta'))$, we have

$$\varphi(\xi) \leq c^{(\alpha)}(\xi, \eta') - \psi(\eta'),$$

so

$$\varphi(\xi') \leq c^{(\alpha)}(\xi', \eta) - \psi(\eta).$$

Therefore,

$$\psi(\eta) \leq c^{(\alpha)}(\xi', \eta) - \varphi(\xi').$$

By definition, we know that $\psi(\eta) = c^{(\alpha)}(\xi, \eta) - \varphi(\xi)$, then

$$\psi(\eta) = \min_{\xi' \in \Omega}(c^{(\alpha)}(\xi', \eta) - \varphi(\xi')).$$

□

As Bregman divergence has a self-dual representation, $L^{(\alpha)}$ divergence also has a self-dual representation. Because of α-conjugate gives us equation (33), we have the following proposition, which is useful in the derivation of the self-dual representation.

Proposition 4.16. *$L^{(\alpha)}$ divergence has the following representation*

$$\boldsymbol{D}^{(\alpha)}[\xi : \xi'] = c^{(\alpha)}(\xi, \eta') - c^{(\alpha)}(\xi', \eta') - (\varphi(\xi) - \varphi(\xi')).$$

Proof.

$$\begin{aligned}
c^{(\alpha)}(\xi, \eta') - c^{(\alpha)}(\xi', \eta') &= \frac{1}{\alpha}\log\left(\frac{1 + \alpha\xi \cdot \eta'}{1 + \alpha\xi' \cdot \eta'}\right) \\
&= \frac{1}{\alpha}\log\left(\frac{1}{1 + \alpha\xi' \cdot \eta'} + \alpha\xi \cdot \frac{\eta'}{1 + \alpha\xi' \cdot \eta'}\right)
\end{aligned}$$

Equation (33) gives us $\frac{1}{1+\alpha\xi'\cdot\eta'} = 1 - \alpha D\varphi(\xi')\cdot\xi'$. Therefore,

$$\begin{aligned} c^{(\alpha)}(\xi,\eta') - c^{(\alpha)}(\xi',\eta') &= \frac{1}{\alpha}\log(1-\alpha D\varphi(\xi')\cdot\xi' + \alpha\xi\cdot\eta'(1-\alpha D\varphi(\xi')\cdot\xi')) \\ &= \frac{1}{\alpha}\log\left(1-\alpha D\varphi(\xi')\cdot\xi' + \alpha\xi\cdot\frac{D\varphi(\xi')}{1-\alpha D\varphi(\xi')\cdot\xi'}(1-\alpha D\varphi(\xi')\cdot\xi')\right) \\ &= \frac{1}{\alpha}\log(1-\alpha D\varphi(\xi')\cdot\xi' + \alpha\xi\cdot D\varphi(\xi')) \\ &= \frac{1}{\alpha}\log(1+\alpha D\varphi(\xi')\cdot(\xi-\xi')). \end{aligned}$$

Hence, the equation holds true. □

By proposition above, we will get the self-dual representations of $L^{(\alpha)}$ divergence.

Theorem 4.17. *(Self-dual representation) The $L^{(\alpha)}$ divergence of φ admits the self-dual representation*

$$\boldsymbol{D}^{(\alpha)}[\xi:\xi'] = c^{(\alpha)}(\xi,\eta') - \varphi(\xi) - \psi(\eta'), \quad \xi,\xi'\in\Omega.$$

Proof. By proposition 4.16 and definition 4.14, we obtain

$$\begin{aligned} \boldsymbol{D}^{(\alpha)}[\xi:\xi'] &= c^{(\alpha)}(\xi,\eta') - c^{(\alpha)}(\xi',\eta') - (\varphi(\xi)-\varphi(\xi')) \\ &= c^{(\alpha)}(\xi,\eta') - \psi(\eta') - \varphi(\eta') - \varphi(\eta) + \varphi(\xi') \\ &= c^{(\alpha)}(\xi,\eta') - \psi(\eta') - \varphi(\xi). \end{aligned}$$

□

4.3 Dually Projective Flat Manifold

Because of the duality of $L^{(\alpha)}$ divergence, the geometric structure induced by it is of particular interest. Let M be a manifold and let $\boldsymbol{\xi} = (\xi_1,\dots,\xi_d)$ be a coordinate system on M. Suppose φ is an α-exponentially concave function on M. Then $\boldsymbol{\eta} = D^{(\alpha)}\varphi(\boldsymbol{\xi}) = (\eta_1,\dots,\eta_d)$ is the dual coordinate system given by the α-gradient. To simply notations, we let

$$\Pi(\boldsymbol{\xi},\boldsymbol{\eta}') = \Pi^{(\alpha)}(\boldsymbol{\xi},\boldsymbol{\eta}') = 1 + \alpha\boldsymbol{\xi}\cdot\boldsymbol{\eta}' \tag{34}$$

where $\boldsymbol{\xi},\boldsymbol{\eta}'$ are column vectors [28]. Because of duality, it suffices to consider the primal connection ∇ in the coordinate system $\{\xi\}$. Let $\{\frac{\partial}{\partial\xi_1},\dots,\frac{\partial}{\partial\xi_d}\}$ be the coordinate vector fields. Then by definition, the coordinate representations of (g,∇,∇^*) under ξ are given by [28]

$$g_(ij)(\xi) = \langle\frac{\partial}{\partial\xi_i},\frac{\partial}{\partial\xi_j}\rangle = -\frac{\partial}{\partial\xi_i}\frac{\partial}{\partial\xi_j}\boldsymbol{D}[\xi:\xi']|_{\xi=\xi'}, \tag{35}$$

$$\Gamma_{ijk}(\xi) = \langle\nabla_{\frac{\partial}{\partial\xi_i}}\frac{\partial}{\partial\xi_j},\frac{\partial}{\partial\xi_k}\rangle = -\frac{\partial^2}{\partial\xi_i\partial\xi_j}\frac{\partial}{\partial\xi'_k}\boldsymbol{D}[\xi:\xi']|_{\xi=\xi'}, \tag{36}$$

$$\Gamma^*_{ijk}(\xi) = \langle\nabla^*_{\frac{\partial}{\partial\xi_i}}\frac{\partial}{\partial\xi_j},\frac{\partial}{\partial\xi_k}\rangle = -\frac{\partial^2}{\partial\xi'_i\partial\xi'_j}\frac{\partial}{\partial\xi'_k}\boldsymbol{D}[\xi:\xi']|_{\xi=\xi'}. \tag{37}$$

If we use self-dual representation of $L^{(\alpha)}$ divergence, by computation which we will omit here, we will get the coordinate representation of the Riemannian metric in ξ.

Proposition 4.18. *The coordinate representation $G(\xi) = (g_{ij}(\xi))$ of the Riemannian metric g is given by*

$$G(\xi) = \frac{-1}{\Pi^{(\alpha)}(\xi,\eta)}(I - \frac{\alpha}{\Pi^{(\alpha)}(\xi,\eta)}\eta\xi^T)\frac{\partial\eta}{\partial\xi}.$$

Its inverse is given by

$$G^{-1}(\xi) = (g^{*ij}(\xi)) = -\Pi^{(\alpha)}(\xi,\eta)\frac{\partial\xi}{\partial\eta}(I + \alpha\eta\xi^T).$$

Remark 4.19. If we directly differentiate $\boldsymbol{D}^{(\alpha)}$ without using self-dual representations, we will get

$$g_{ij}(\xi) = -\frac{\partial^2\varphi}{\partial\xi_i\partial\xi_j}(\xi) - \alpha\frac{\partial\varphi}{\partial\xi_i}(\xi)\frac{\partial\varphi}{\partial\xi_j}(\xi). \tag{38}$$

Proposition 4.20. *The Christoffel symbols of the primal connection are given by*

$$\Gamma^k_{ij}(\xi) = -\alpha(\frac{\partial\varphi}{\partial\xi_i}\delta^k_j + \frac{\partial\varphi}{\partial\xi_j}\delta^k_i).$$

With all these notations available, we are going to derive some geometric properties of dually projective flat manifolds. The following definition is taken from [19].

Definition 4.21. *(Projectively flatness) Let ∇ and $\tilde{\nabla}$ be torsion-free affine connections on a smooth manifold. They are projectively equivalent if there exists a differential 1-form τ such that*

$$\nabla_X Y = \widetilde{\nabla}_X Y + \tau(X)Y + \tau(Y)X$$

for any smooth vector fields X and Y. We say that ∇ is projectively flat if ∇ is projectively equivalent to a flat connection (a connection whose Riemann-Christoffel curvature tensor vanishes).

Remark 4.22. Notice that if we let $\tau = a_i(\xi)d\xi_i$ which is a differential one form in ξ coordinate, then the equation in definition 4.20 will be equivalent to

$$\Gamma^k_{ij}(\xi) = \widetilde{\Gamma}^k_{ij}(\xi) + a_i(\xi)\delta^k_j + a_j(\xi)\delta^k_i. \tag{39}$$

Theorem 4.23. *For any $\boldsymbol{D}^{(\pm\alpha)}$, the induced primal connection ∇ and the dual connection ∇^* are both projectively flat.*

Proof. We only prove for primal connection. Let $\widetilde{\nabla}$ be the flat Euclidean connection with respect to the coordinate system ξ such that all Christoffel symbols $\widetilde{\Gamma}^k_{ij}(\xi)$ are zero. If we can find a differential one form τ such that Γ^k_{ij} satisfies equation (39), then the primal connection is projectively flat.
Consider the differential one form $\tau(X) = -\alpha X\varphi$ where $X\varphi$ means the directional derivative of φ in direction of X. Then in ξ coordinates,

$$\tau = -\sum_{j=1}^{d}\alpha\frac{\partial\varphi}{\partial\xi_i}d\xi_i$$

and

$$\tau(X)Y + \tau(Y)X = -\alpha\frac{\partial}{\partial\xi_i}\delta^k_j - \alpha\frac{\partial\varphi}{\partial\xi_j}\delta^k_i = \Gamma^k_{ij}(\xi).$$

Therefore, the primal connection ∇ is projectively flat. □

Since the primal and dual connections are both projectively flat, we say that the induced dualistic structure are dually projective flat. Now we want to know the geodesic in dually projective flat manifolds. We discussed that in dually flat manifold, a geodesic is just a straight line. In fact, for a dually projective flat manifold, a geodesic is a still straight line under suitable parameterization. Recall that a smooth curve γ is a primal geodesic if $\nabla_{\dot{\gamma}(t)}\dot{\gamma}(t) = 0$. Equivalently, we have a set of geodesic equations

$$\ddot{\xi}_k + \dot{\xi}_i(t)\dot{\xi}_j(t)\Gamma^k_{ij}(\xi(t)) = 0, \quad k = 1, \ldots, d. \tag{40}$$

Based on this set of ordinary differential equations, we have the following corollary.

Corollary 4.24. *For the dualistic structure induced by $\boldsymbol{D}^{(\pm\alpha)}$, if γ is a primal (dual) geodesic, then its trace in the primal (dual) coordinate system is a straight line.*

Proof. By the equation in proposition 4.20, we rewrite equation (40) as

$$\ddot{\xi}_k(t) + \dot{\xi}_i\dot{\xi}_j(-\alpha\frac{\partial\varphi}{\partial\xi_i}\delta^k_j - \alpha\frac{\partial\varphi}{\partial\xi_j}\delta^k_i) = 0$$

Note that δ^k_j, δ^k_i are 1 if and only if $i = j = k$. Therefore,

$$\ddot{\xi}_k(t) - 2\alpha\dot{\xi}_k\frac{d}{dt}\varphi(\xi_k(t)) = 0.$$

Hence, we get a single vector equation

$$\ddot{\xi}(t) = 2\alpha\dot{\xi}(t)\frac{d}{dt}\varphi(\xi(t)). \tag{41}$$

From this ODE, we see that the trace of the primal geodesic is a straight line under the coordinate ξ. □

An interesting question is the relationship between the primal and dual coordinate vector fields. Recall that in dually flat manifolds, the primal coordinate vector field e_i and the dual coordinate vector field e^*_j satisfy

$$\langle e_i, e^*_j\rangle = \delta^j_i.$$

Proposition 4.25. *For $\boldsymbol{D}^{(\alpha)}$, the inner product between $\frac{\partial}{\partial\xi_i}$ and $\frac{\partial}{\partial\eta_j}$ is*

$$\langle\frac{\partial}{\partial\xi_i}, \frac{\partial}{\partial\eta_j}\rangle = \frac{-1}{\Pi(\xi,\eta)}\delta^j_i + \frac{\alpha}{\Pi(\xi,\eta)^2}\xi_j\eta_i.$$

Remark 4.26. Since the proof of this proposition is pure computation, we will omit the proof here.

When we discussed about dually flat manifolds, the Generalized Pythagorean Theorem holds true and it is a fundamental result of information geometry. Now in dually projective flat manifolds, we also have generalized Pythagorean theorem [28].

Theorem 4.27. *(Generalized Pythagorean Theorem) Consider any $L^{(\pm\alpha)}$ divergence $\boldsymbol{D} = \boldsymbol{D}^{(\alpha)}$ and the induced dualistic structure. Let $p, q, r \in \Omega$ and suppose that the dual geodesic from q to p exists. Then the generalized Pythagorean relation*

$$\boldsymbol{D}[q : p] + \boldsymbol{D}[r : q] = \boldsymbol{D}[r : p]$$

holds if and if only the primal geodesic from q to r and the dual geodesic from q to p meet g-orthogonally at q.

Proof. Again, we only consider the case of $\boldsymbol{D}^{(\alpha)}$ since the proof of $\boldsymbol{D}^{(-\alpha)}$ is similar.
By the self-dual representation of $L^{(\alpha)}$ divergence, we have

$$\boldsymbol{D}[q : p] = \frac{1}{\alpha}\log(1 + \alpha\xi_p \cdot \eta_p) - \varphi(\xi_q) - \psi(\eta_p)$$

and similarly for $\boldsymbol{D}[r : q]$ and $\boldsymbol{D}[r : p]$.
With these expressions, the generalized Pythagorean relation holds if and only if

$$\begin{aligned}&\frac{1}{\alpha}(1 + \alpha\xi_q \cdot \eta_p) - \varphi(\xi_p) - \psi(\eta_p) + \frac{1}{\alpha}\log(1 + \alpha\xi_r \cdot \eta_q) - \varphi(\xi_r) - \psi(\eta_q)\\ &= \frac{1}{\alpha}\log(1 + \alpha\xi_r \cdot \eta_p) - \varphi(\xi_r) - \psi(\eta_p). \qquad (42)\end{aligned}$$

Simply equation (42) to get the following equation

$$\frac{1}{\alpha}\log[(1 + \alpha\xi_q \cdot \eta_p)(1 + \alpha\xi_r \cdot \eta_q)] = \frac{1}{\alpha}\log[(1 + \alpha\xi_r \cdot \eta_p)(1 + \alpha\xi_q \cdot \eta_q)]. \qquad (43)$$

Therefore,

$$(1 + \alpha\xi_q \cdot \eta_p)(1 + \alpha\xi_r \cdot \eta_q) = (1 + \alpha\xi_r \cdot \eta_p)(1 + \alpha\xi_q \cdot \eta_q).$$

Expanding and simplifying, we get

$$(\xi_r - \xi_q) \cdot (\eta_p - \eta_q) = \alpha(\xi_q \cdot \eta_p)(\xi_r \cdot \eta_q) - \alpha(\xi_r \cdot \eta_p)(\xi_q \cdot \eta_q). \qquad (44)$$

Consider the primal geodesic connecting q and r and the dual geodesic connecting p and q. By projective flatness, the initial velocity v_1 of the primal geodesic is proportional to $\xi_r - \xi_q$ and the initial velocity v_2 of the dual geodesic is proportional to $\eta_p - \eta_q$. In other words, $v_1 = a(\xi_r - \xi_q)$ and $v_2 = b(\eta_p - \eta_q)$. Since v_1, v_2 are tangent vectors, then we can write them as

$$v_1 = a\sum_i (\xi_r^i - \xi_q^i)\frac{\partial}{\partial \xi^i}$$

and

$$v_2 = b\sum_j (\eta_p^j - \eta_q^j)\frac{\partial}{\partial \eta^j}.$$

By orthogonality, we know $\langle v_1, v_2\rangle = 0$. It's equivalent to say

$$\begin{aligned}0 &= \langle\sum_i (\xi_r^i - \xi_q^i)\frac{\partial}{\partial \xi^i}, \sum_j (\eta_p^j - \eta_q^j)\frac{\partial}{\partial \eta^j}\rangle\\ &= \sum_{i,j}(\xi_r^i - \xi_q^i)(\eta_p^j - \eta_q^j)\langle\frac{\partial}{\partial \xi^i}, \frac{\partial}{\partial \eta^j}\rangle\end{aligned}$$

By proposition 4.25, we have

$$\begin{aligned}0 &= \sum_{i,j}(\xi_r^i - \xi_q^i)(\eta_p^j - \eta_q^j)(\frac{-1}{\Pi(\xi,\eta)}\delta_i^j + \frac{\alpha}{\Pi(\xi,\eta)^2}\eta_q^i\xi_q^j) \\ &= \frac{-(\xi_r - \xi_q)\cdot(\eta_p - \eta_q)}{1+\alpha\xi_q\cdot\eta_q} + \sum_{i,j}\frac{\alpha}{(1+\alpha\xi_q\cdot\eta_q)^2}((\xi_r^i - \xi_q^i)\eta_q^i)((\eta_p^j - \eta_q^j)\xi_q^j)\end{aligned}$$

Therefore,

$$\begin{aligned}(1+\alpha\xi_q\cdot\eta_q)(\xi_r - \xi_q)\cdot(\eta_p - \eta_q) &= \sum_{i,j}\alpha((\xi_r^i - \xi_q^i)\eta_q^i)((\eta_p^j - \eta_q^j)\xi_q^j) \\ &= \alpha((\xi_r - \xi_q)\cdot\eta_q)((\eta_p - \eta_q)\cdot\xi_q).\end{aligned}$$

Note that this equation is equivalent to equation (44). In other words, two geodesics are orthogonal if and only if equation (44) holds true.
Hence, the generalized Pythagorean relation holds true. □

This finishes the discussion about geometry of dually projective flat manifolds.

4.4 Orthogonal Foliation

Recall that we derived orthogonal foliations for a dually flat manifold. Analogously, a dually projective flat manifold also has orthogonal foliations. However, the conditions and conclusion are more complicated and still under development. Here, we only introduce the basic idea of orthogonal foliations of dually flat manifolds without any proofs given. We will use same notations as in section 2.4.

Claim 4.28. *For every submanifold $E_k(c_{k+1},\dots,c_n)$, there exists an submanifold M_k such that M_k and E_k are orthogonal.*

Remark 4.29. In fact, M_k and E_k only have one intersection point, which is the same as in orthogonal foliations of dually flat manifolds. What is different is that such M_k is unique. In other words, given a submanifold E_k and an intersection point p, there exists a unique M_k that is orthogonal with E_k and intersects with E_k only at point p. However, in dually flat manifolds, for a given E_k and a point p, there may have infinitely many M_k's that satisfy the condition.

The complicated part about orthogonal foliations in a dually projective flat manifold is that M_k is not in the form of $\{p(\boldsymbol{x},\boldsymbol{\theta})|(\eta_1,\dots,\eta_k) = (d_1,\dots,d_k)\}$ where $d_1,\dots,d_k$ are constants. We have the following claim.

Claim 4.30. *Let $\hat{p} \in E_k(c_{k+1},\dots,c_n)$ and let M_k be the unique orthogonal complement of E_k at point $\hat{p}$. Suppose the θ-coordinates for $\hat{p}$ is $\theta_{\hat{p}} = (\hat{\theta}_1,\dots,\hat{\theta}_k,c_{k+1},\dots,c_n)$ and the η-coordinates for $\hat{p}$ is $\eta_{\hat{p}} = (\hat{\eta}_1,\dots,\hat{\eta}_k,d_{k+1},\dots,d_n)$. Then $M_k = \{p : \eta_p - \eta_{\hat{p}} \in S\}$ where*

$$S = span\left\{\begin{pmatrix} A\hat{\eta}_1 c_{k+1} \\ \vdots \\ A\hat{\eta}_k c_{k+1} \\ 1 \\ \vdots \\ 0 \end{pmatrix},\dots,\begin{pmatrix} A\hat{\eta}_1 c_n \\ \vdots \\ A\hat{\eta}_k c_n \\ 0 \\ \vdots \\ 1 \end{pmatrix}\right\} \text{ and } A = \tfrac{-\alpha}{\Pi}(1 - \alpha\hat{\eta}_1\hat{\theta}_1 - \cdots - \alpha\hat{\eta}_k\hat{\theta}_k).$$

Since orthogonal foliations exist and the Generalized Pythagorean Theorem still holds true in dually projective flat manifolds, we have the following claim, which is analogous to theorem 2.39.

Claim 4.31. *Suppose the θ-coordinate for p is $(\theta_1, \ldots, \theta_n)$ and the θ-coordinate for $p_k \in E_k(c_{k+1}, \ldots, c_n)$ is $(\theta_1^{(k)}, \ldots, \theta_k^{(k)}, c_{k+1}, \ldots, c_n)$. Let $(\theta_1^{(j)}, \ldots, \theta_j^{(j)}, c_{j+1}, \ldots, c_n)$ denote the coordinate of $p_j \in E_j(c_{j+1}, \ldots, c_n)$ where $k \leq j \leq n-1$. Then*

$$D[p : p_k] = \sum_{j=k}^{n-1} D[p_{j+1} : p_j]$$

where $p_n = p$.

We see that the basic idea of orthogonal foliations is similar as in the case of dually flat manifolds. The full set of theories is still under development.

5 Generalizations of Exponential Family

We know that Bregman divergence are used naturally in the exponential family of probability distributions. Analogous to exponential family, we have a more generalized family of probability distributions, where $L^{(\pm\alpha)}$ divergence plays an important role. We call this family of probability distributions $F^{(\pm\alpha)}$ family. As before, we focus our discussion on $F^{(\alpha)}$ family.

5.1 $F^{(\alpha)}$ Family of Probability Distributions

Our motivation to define $F^{(\alpha)}$ family comes from the relationship between Bregman divergence and the exponential family. Recall that a density $p(x, \xi)$ from the exponential family is of the form

$$\log p(x, \xi) = \xi \cdot h(x) - \varphi(\xi)$$

where φ is a convex function. Moreover, we have

$$\log p(x, \xi) = -\boldsymbol{D}[\xi : \xi'] + \psi(h(x))$$

where $\boldsymbol{D}$ is Bregman divergence and ψ is the convex conjugate of φ. Based on this idea, recall the self-dual representation of $L^{(\alpha)}$ divergence is given by

$$\boldsymbol{D}^{(\alpha)}[\xi : \xi'] = \frac{1}{\alpha} \log(1 + \alpha\xi \cdot \eta') - \varphi(\xi) - \psi(\eta').$$

We consider the following parameterized density of the form

$$\log p(x, \xi) = -\boldsymbol{D}^{(\alpha)}[\xi : \xi'] - \psi(h(x)) = -\frac{1}{\alpha} \log(1 + \alpha\xi \cdot h(x)) + \varphi(\xi)$$

where $h(x) = \eta'$ [28]. Rearranging the above equation, we have

$$p(x, \xi) = (1 + \alpha\xi \cdot h(x))^{-\frac{1}{\alpha}} e^{\varphi(\xi)}. \tag{45}$$

By integrating $p(x, \xi)$, we see the normalization function $\varphi(\xi)$ is given by

$$\varphi(\xi) = \log \int (1 + \alpha\xi \cdot h(x))^{\frac{1}{\alpha}} d\mu(x) \tag{46}$$

where μ is the suitable measure.

Definition 5.1. *($F^{(\alpha)}$ Family) The $F^{(\alpha)}$ family of probability distribution is in the form of*

$$p(x,\xi) = (1+\alpha\xi\cdot h(x))^{-\frac{1}{\alpha}} e^{\varphi(\xi)}.$$

Now we will see an important relationship between Rényi entropy, Rényi divergence [28] and $L^{(\alpha)}$ divergence for $F^{(\alpha)}$ family.

Definition 5.2. *(Rényi entropy and Rényi divergence) Let $\tilde{\alpha} \in (0,1)\cup(1,\infty)$. Let P, Q be probability measures on $\mathcal{X}$ that are absolutely continuous with respect μ. We let $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$ be their densities. (i) The Rényi entropy of P of order $\tilde{\alpha}$ is defined by*

$$\boldsymbol{H}_{\tilde{\alpha}} = \frac{1}{1-\tilde{\alpha}} \log \int p^{\tilde{\alpha}} d\mu.$$

(ii) The Rényi divergence of order $\tilde{\alpha}$ between P and Q is defined by

$$\boldsymbol{D}_{\tilde{\alpha}}(P||Q) = \frac{1}{\tilde{\alpha}-1} \log \int p^{\tilde{\alpha}} q^{1-\tilde{\alpha}} d\mu.$$

The following is the main result of this subsection [28].

Theorem 5.3. *Consider an $F^{(\alpha)}$ family with $\alpha > 0$ and potential function φ. Then the $L^{(\alpha)}$ divergence of φ is the Rényi divergence of order $\tilde{\alpha} := 1+\alpha$:*

$$\boldsymbol{D}^{(\alpha)}[\xi : \xi'] = \boldsymbol{D}_{\tilde{\alpha}}(p(\cdot,\xi')||p(\cdot,\xi)).$$

Moreover, the α-conjugate ψ of φ is the Rényi entropy of order $\tilde{\alpha}$:

$$\psi(\eta) = \boldsymbol{H}_{\tilde{\alpha}}(p(\cdot,\xi)), \quad \eta = D^{(\alpha)}\varphi(\xi).$$

Proof. Define a random variable

$$Z_\xi = \frac{h(X)}{1+\alpha\xi\cdot h(x)}.$$

By some computation, which we omit here, we will obtain

$$D\varphi(\xi) = \mathbb{E}_\xi[Z_\xi].$$

Then,

$$\begin{aligned} 1+\alpha D\varphi(\xi')\cdot(\xi-\xi') &= 1+\alpha\mathbb{E}_{\xi'}[Z_{\xi'}]\cdot(\xi-\xi') \\ &= 1+\alpha\mathbb{E}_{\xi'}\left[\frac{h(X)}{1+\alpha\xi'\cdot h(x)}\cdot(\xi-\xi')\right] \\ &= \mathbb{E}_{\xi'}\left[\frac{1+\alpha\xi\cdot h(X)}{1+\alpha\xi'\cdot h(X)}\right]. \end{aligned}$$

Since $p(x,\xi) = (1+\alpha\xi\cdot h(x))^{\frac{-1}{\alpha}} e^{\varphi(\xi)}$, we have

$$1+\alpha\xi'\cdot h(X) = \frac{e^{\alpha\varphi(\xi')}}{P(X,\xi')^\alpha}.$$

Therefore,

$$\begin{aligned}\boldsymbol{D}^{(\alpha)}[\xi:\xi'] &= \frac{1}{\alpha}\log(1+\alpha D\varphi(\xi')\cdot(\xi-\xi')) - (\varphi(\xi)-\varphi(\xi')) \\ &= \frac{1}{\alpha}\mathbb{E}_{\xi'}\left[\frac{e^{\alpha\varphi(\xi)}/p(X,\xi)^\alpha}{e^{\alpha\varphi(\xi')}/p(X,\xi;)^\alpha}\right] - (\varphi(\xi)-\varphi(\xi')) \\ &= \frac{1}{\alpha}\log\mathbb{E}_{\xi'}\left[\frac{p(X,\xi')^\alpha}{p(X,\xi)^\alpha}\right].\end{aligned}$$

Now we will express the divergence as an integral.

$$\begin{aligned}\boldsymbol{D}^{(\alpha)}[\xi:\xi'] &= \frac{1}{\alpha}\log\int \xi' p(x,\xi')^\alpha p(x,\xi)^{-\alpha}d\mu \\ &= \frac{1}{\alpha}\log\int p(x,\xi')^{1+\alpha}p(x,\xi)^{-\alpha}d\mu \\ &= \frac{1}{\tilde{\alpha}-1}\log\int p(x,\xi')^{\tilde{\alpha}}p(x,\xi)^{1-\tilde{\alpha}}d\mu \\ &= \boldsymbol{D}_{\tilde{\alpha}}(p(\cdot,\xi')||p(\cdot,\xi))\end{aligned}$$

where $\tilde{\alpha} = 1+\alpha$. This proves the first part of the theorem.
Next, we will prove for the α-conjugate.
Recall

$$\psi(\eta) = \frac{1}{\alpha}\log(1+\alpha\xi\cdot\eta) - \varphi(\xi).$$

Recall the expression (33). Then

$$1+\alpha\xi\cdot\eta = \frac{1}{1-\alpha D\varphi(\xi)\cdot\xi}$$

Moreover, we derived that

$$1+\alpha D\varphi(\xi')\cdot(\xi-\xi') = 1+\alpha\mathbb{E}_{\xi'}\left[\frac{h(X)}{1+\alpha\xi'\cdot h(X)}\cdot(\xi-\xi')\right].$$

Therefore,

$$\alpha D\varphi(\xi)\cdot\xi = \alpha\mathbb{E}_\xi\left[\frac{h(x)}{1+\alpha\xi\cdot h(x)}\cdot\xi\right].$$

Then,

$$\begin{aligned}1+\alpha\xi\cdot\eta &= (\mathbb{E}_\xi\left[\frac{1}{1+\alpha\xi\cdot h(X)}\right])^{-1} \\ &= (\mathbb{E}_\xi\left[p(X,\xi)^\alpha e^{-\alpha\varphi(\xi)}\right])^{-1}.\end{aligned}$$

From this, we obtain

$$\begin{aligned}\psi(\eta) &= \frac{-1}{\alpha}\log\int p(x,\xi)^{1+\alpha}e^{-\alpha\varphi(\xi)}d\mu - \varphi(\xi) \\ &= \frac{1}{1-\tilde{\alpha}}\log\int p(x,\xi)^{\tilde{\alpha}}d\mu \\ &= \boldsymbol{H}_{\tilde{\alpha}}(p(\cdot,\xi)).\end{aligned}$$

This proved the second part of the theorem. □

5.2 Principal Component Analysis

As Bregman divergence can be used to generalize PCA to the exponential family, $L^{(\alpha)}$ divergence can be used to generalize PCA to the $F^{(\alpha)}$ family. Since theories about PCA for $F^{(\alpha)}$ family are still under development, we will only discuss about the basic idea without giving any rigorous proof. The basic idea is to use the self-dual representations of $L^{(\alpha)}$ divergence to rewrite the negative log-likelihood function.

Let $\psi(\eta)$ be the α-conjugate of $\varphi(\theta)$ where $\eta = D^{(\alpha)}\varphi(\theta)$ and $\psi(\eta) = \frac{1}{\alpha}\log(1+\alpha\eta\cdot\theta) - \varphi(\theta)$. Let $p(x,\theta) = (1+\alpha\theta\cdot h(x))^{\frac{-1}{\alpha}} e^{\varphi(\theta)}$ which is from the $F^{(\alpha)}$ family such that the negative log-likelihood is given by $-\log p(x,\theta) = \frac{1}{\alpha}\log(1+\alpha\theta\cdot h(x)) - \varphi(\theta)$. Then because of the self-dual expression of $L^{(\alpha)}$ divergence, we can claim

$$-\log p(x,\theta) = D_{\psi}^{(\alpha)}[h(x):\eta] + \psi(h(x)).$$

Now we can extend the idea in PCA for the exponential family to the $F^{(\alpha)}$ family by minimizing negative log-likelihood function.

6 Acknowledgment

I would like to thank Professor Ting-Kam Leonard Wong from Department of Statistical Sciences, Univeristy of Toronto for supervising me over the Winter semester of 2020, and also thank the Department of Statistical Sciences in University of Toronto for providing me with this opportunity to research on this topic.

References

[1] Shotaro Akaho. The e-pca and m-pca: Dimension reduction of parameters by information geometry. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*, volume 1, pages 129–134. IEEE, 2004.

[2] S-I Amari. Information geometry on hierarchy of probability distributions. *IEEE transactions on information theory*, 47(5):1701–1711, 2001.

[3] Shun-Ichi Amari. Alpha-divergence is unique, belonging to both f-divergence and bregman divergence classes. *IEEE Transactions on Information Theory*, 55(11):4925–4931, 2009.

[4] Shun-Ichi Amari. *Information Geometry and Its Applications*. Springer, 2016.

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

[6] Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. *Information geometry*, volume 64. Springer, 2017.

[7] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

[8] Jean-Daniel Boissonnat, Frank Nielsen, and Richard Nock. Bregman voronoi diagrams. *Discrete & Computational Geometry*, 44(2):281–307, 2010.

[9] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[10] Manfredo Perdigao do Carmo. *Riemannian geometry*. Birkhäuser, 1992.

[11] Michael Collins, Sanjoy Dasgupta, and Robert E Schapire. A generalization of principal components analysis to the exponential family. In *Advances in neural information processing systems*, pages 617–624, 2002.

[12] Michael Collins, Robert E Schapire, and Yoram Singer. Logistic regression, adaboost and bregman distances. *Machine Learning*, 48(1-3):253–285, 2002.

[13] Imre Csiszar et al. Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. *The annals of statistics*, 19(4):2032–2066, 1991.

[14] Shun-Ichi Amari Hiroshi Nagaoka. Differential geometry of smooth families of probability distributions. 1982.

[15] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[16] Atsuya Kumagai. Spectral dimensionality reduction for bregman information. *Information Geometry*, 2(2):273–282, 2019.

[17] Meng Lu, Kai He, Jianhua Z Huang, and Xiaoning Qian. Principal component analysis for exponential family data. In *Advances in Principal Component Analysis*, pages 193–223. Springer, 2018.

[18] Meng Lu, Jianhua Z Huang, and Xiaoning Qian. Sparse exponential family principal component analysis. *Pattern recognition*, 60:681–691, 2016.

[19] Hiroshi Matsuzoe et al. Geometry of contrast functions and conformal geometry. *Hiroshima Mathematical Journal*, 29(1):175–191, 1999.

[20] Shakir Mohamed, Zoubin Ghahramani, and Katherine A Heller. Bayesian exponential family pca. *Advances in neural information processing systems*, 21:1089–1096, 2008.

[21] Jan Naudts and Jun Zhang. Rho–tau embedding and gauge freedom in information geometry. *Information geometry*, 1(1):79–115, 2018.

[22] Tomohiro Nishiyama. Generalized bregman and jensen divergences which include some f-divergences. *arXiv preprint arXiv:1808.06148*, 2018.

[23] Soumik Pal and Ting-Kam Leonard Wong. The geometry of relative arbitrage. *Mathematics and Financial Economics*, 10(3):263–293, 2016.

[24] Soumik Pal and Ting-Kam Leonard Wong. Multiplicative schrödinger problem and the dirichlet transport. *Probability Theory and Related Fields*, 178(1):613–654, 2020.

[25] Soumik Pal, Ting-Kam Leonard Wong, et al. Exponentially concave functions and a new information geometry. *The Annals of Probability*, 46(2):1070–1113, 2018.

[26] Suvrit Sra and Inderjit Dhillon. Generalized nonnegative matrix approximations with bregman divergences. *Advances in neural information processing systems*, 18:283–290, 2005.

[27] Ting-Kam Leonard Wong. Optimization of relative arbitrage. *Annals of Finance*, 11(3-4):345–382, 2015.

[28] Ting-Kam Leonard Wong. Logarithmic divergences from optimal transport and rényi geometry. *Information Geometry*, 1(1):39–78, 2018.

[29] Ting-Kam Leonard Wong. Information geometry in portfolio theory. In *Geometric Structures of Information*, pages 105–136. Springer, 2019.

[30] Ting-Kam Leonard Wong and Jiaowen Yang. Optimal transport and information geometry. *arXiv preprint arXiv:1906.00030*, 2019.

[31] Chunming Zhang, Yuan Jiang, and Zuofeng Shang. New aspects of bregman divergence in regression and classification with parametric and nonparametric estimation. *Canadian Journal of Statistics*, 37(1):119–139, 2009.

[32] Jun Zhang. Divergence function, duality, and convex analysis. *Neural computation*, 16(1):159–195, 2004.

[33] Jun Zhang. Referential duality and representational duality on statistical manifolds. In *Proceedings of the Second International Symposium on Information Geometry and Its Applications, Tokyo, Japan*, volume 1216, page 5867, 2005.

[34] Jun Zhang. Reference duality and representation duality in information geometry. In *AIP Conference Proceedings*, volume 1641, pages 130–146. American Institute of Physics, 2015.