GEOMETRIC STRUCTURES ON A STATISTICAL MANIFOLD AND GEOMETRY OF ESTIMATION

A Thesis submitted

in partial fulfillment for the Degree of

Doctor of Philosophy

by

HARSHA K. V.

Department of Mathematics

INDIAN INSTITUTE OF SPACE SCIENCE AND TECHNOLOGY

THIRUVANANTHAPURAM

AUGUST 2015

CERTIFICATE

This is to certify that the thesis entitled **Geometric Structures on a Statistical Manifold and Geometry of Estimation** submitted by **Harsha K. V.** to the Indian Institute of Space Science and Technology, Thiruvananthapuram, in partial fulfillment for the award of the degree of **Doctor of Philosophy** is a *bona fide* record of research work carried out by her under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institution or University for the award of any degree or diploma.

Dr. Subrahamanian Moosath K. S.
Supervisor
Department of Mathematics

Thiruvananthapuram
August 2015

Counter signature of HOD with seal

DECLARATION

I declare that this thesis entitled **Geometric Structures on a Statistical Manifold and Geometry of Estimation** submitted in partial fulfillment of the Degree of Doctor of Philosophy is a record of original work carried out by me under the supervision of **Dr. Subrahamanian Moosath K. S.**, and has not formed the basis for the award of any other degree or diploma, in this or any other Institution or University. In keeping with the ethical practice in reporting scientific information, due acknowledgements have been made wherever the findings of others have been cited.

Thiruvananthapuram-695547	Harsha K. V.
August 2015	(SC11D017)

ACKNOWLEDGEMENTS

First of all, let me take this opportunity to express my sincere gratitude to my supervisor Dr. Subrahamanian Moosath K. S. for his guidance, exceptional teaching and immense patience. He has been a constant source of inspiration and support for me personally and professionally throughout my journey in this research work. Apart from the academic excellence he is a wonderful human being who has been there for me in my difficult times. Also I acknowledge the tireless discussions and presentations that I had with him without which this work would not have been possible.

I would like to acknowledge my deep debt of gratitude to Prof. Shun-ichi Amari, the legend of Information Geometry, for his valuable suggestions and encouragement. I was always wondered by his tremendous knowledge and intuitions which have been a great motivation for a beginner like me to carry on my research in the area of information geometry. Thank you very much sir for your guidance and mentoring without which this work would not have seen the light of the day.

I sincerely thank my doctoral committee member Prof. Sitaramayya M. for his continuous support, enthusiasm and fruitful discussions. I am indebted to him for reading the manuscripts and results and encouraging me to move forward. I wish to thank my department HOD, my doctoral committee member Prof. Raju K. George for his immediate help and support that I received whenever I approached him. I also thank my other doctoral committee members, Dr. Rajan M. P., Dr. Deepak Mishra and Dr. Deepak T. G. for their helpful suggestions and feedback on my research work.

I would like to thank Director, IIST, Dr. K. S. Dasgupta for providing me an opportunity to be a part of this institute. I also thank all faculty members of IIST, especially Dr. Kuruvilla Josheph and Dr. Kurien Isaac for their support and goodwill. I would like to thank IIST for the financial support during my research work. I acknowledge the financial support by the Department of Science & Technology for participating in an international conference held in Amboise, France as part of my research work. I also thank each and every members of library and administrative staffs for their support and immediate help. I wholeheartedly thank my department staff members Anish, Nisha, Karim and Pournami for their timely help, love and support.

I also take this opportunity to thank all my dear teachers for their advise and guidance. Especially I would like to thank Dr. Rajan Sundaravaradan, Amrita University who motivated me to have a career in mathematics which I thought would be impossible. Thank you very much sir for being a great source of inspiration for me by your amazing teaching ability and immense knowledge. I also thank Dr. Krishna Kumar V., NISER for his guidance, love and support.

I thank all my friends for their love, encouragement and criticism. Especially, I thank Rashmi for being there for me throughout my Ph.D days as a best companion. It is my pleasure to acknowledge the never ending discussions I had with Ms. Dhanya, my teacher and my friend. I also thank all my fellow research scholars, in particular, Raja, Bhaskar, Haripadmam, Sarika, Sarah, Dhanya, Aneesha, Jalaja, Ruchi, Sreenivas, Arun and Gopakumar for their support and care.

Last but not the least, I would like to thank my parents for making me who I am today, encouraging me to believe in myself without whom none of my success would be possible. I am running out of words to thank my wonderful husband Jithin for his immense love, criticism and encouragement throughout this work. He has been my backbone, my motivation without whose support and care, I would not have been able to make it this far. I also thank my in-laws for understanding me so well and supporting me all the way during this work. I wish to thank my brother for his endless love and encouragement.

Harsha K. V.

ABSTRACT

The main objective of this thesis is to study the various geometric structures on a statistical manifold and the geometry of parameter estimation. This study comes under the area of Information Geometry which is the geometric study of a statistical model of probability distributions. A statistical model equipped with a Riemannian metric and a pair of dual affine connections is called a statistical manifold. Amari's α-geometry is an important geometric structure on a statistical manifold which plays a major role in the asymptotic theory of estimation.

In Chapter 2 we introduce a generalized class of geometric structures on a statistical manifold called the (F, G)-geometry using a general embedding function F and a positive smooth function G. In Section 2.2 the Fisher information metric and the α-connections are computed for a statistical manifold defined on finite sets. In Theorem 2.3.5 we prove a necessary and sufficient condition for two (F, G)-connections to be dual with respect to the G-metric. In Theorem 2.3.6 we show that the α-geometry is a special case of the (F, G)-geometry. Thus we obtain a generalized dualistic structure on a statistical manifold which includes the α-geometry as a special case. Further the G-metric and the (F, G)-connections are computed for statistical manifold defined on finite sets in Section 2.3.

In Chapter 3 we study the invariance properties of various geometric structures on a statistical manifold and classify them into invariant and non-invariant classes. The covariance under reparametrization of the (F, G)-geometric structures are shown in Theorems 3.2.3 and 3.2.4. Then in Theorem 3.2.5 we prove that the (F, G)-geometry is not invariant under smooth one to one transformations of the random variable in general. In Corollary 3.2.6 we prove that the α-geometry is the only (F, G)-geometry which is invariant under smooth one to one transformations of the random variable. In Theorems 3.2.7 and 3.2.8 we show that the (α, ρ, τ)-geometry is covariant under reparametrization and is not invariant under smooth one to one transformations of the random variable in general. Also the α-geometry is the only (α, ρ, τ)-geometry which is invariant under smooth one to one transformations of the random variable. Further the relation between

the (F, G) and (α, ρ, τ)-geometries are given in Theorem 3.2.11.

In Chapter 4 first we give the (± 1)-conformal equivalence of the α-geometry and the geometry induced from the conformal transformation of the α-divergence in Propositions 4.2.3 and 4.2.4. In Corollary 4.2.6 we prove that the q-structure is the conformal flattening of the α-geometry. Then we discuss the importance of non-invariant (F, G)-geometry in the study of the dually flat geometries of the deformed exponential family. There are two dually flat geometries on a deformed exponential family, the U-geometry and the χ-geometry. In Theorem 4.3.4 we show that the U-geometry is the (F, G)-geometry for suitable choices of F and G. Further we prove that the χ-geometry is the conformal flattening of the (F, G)-geometry for suitable choices of F and G in Theorems 4.3.16, 4.3.17 and 4.3.18.

In Chapter 5 we consider the parameter estimation problem based on a mismatched model. In Theorems 5.3.1 and 5.3.2 we prove a necessary and sufficient condition for the estimator based on a mismatched model to be consistent and first order efficient. Further a theoretical formulation of the maximum likelihood estimation problem based on a mismatched model in an exponential family is given. We prove a necessary and sufficient condition for an MLE based on a mismatched model to be consistent and efficient in Theorems 5.3.8 and 5.3.9.

In Chapter 6 we define certain generalized notions like F-product, F- independence of random variables and maximum F-likelihood estimator (F-MLE) in Section 6.1. In Theorem 6.1.6 we show that the F-MLE is a MAP estimator with a prior. Then using the F-escort probability distribution we define two generalized notions of MLE, the $\mathbf{x}_N$-based F-escort MLE and the F-escort MLE based on the product of F-escort distribution of the marginal probability density of single observations in Section 6.2. In Theorem 6.2.3 we give a characterization of the q-escort MLE among the $\mathbf{x}_N$ based F-escort MLE as a Bayesian MAP estimator with a prior. Further an analytic proof of the F-version of the maximum entropy theorem is given in Theorem 6.2.5. In Theorem 6.3.2 a proof of the generalized Cramer-Rao bound defined by Naudts is given. Further we show that the U-estimator for the dual coordinate in the U-geometry of the deformed exponential family is optimal with respect to this bound in Theorem 6.3.3. This chapter ends with an open problem regarding the properties of the F-MLE in a deformed exponential family.

TABLE OF CONTENTS

CERTIFICATE **v**

DECLARATION **vii**

ACKNOWLEDGEMENTS **ix**

ABSTRACT **xi**

1 Introduction **1**

2 Geometric Structures on a Statistical Manifold **11**

2.1 Statistical Manifold 12

2.1.1 Affine structure of the family of measures 12

2.1.2 Statistical manifold 13

2.2 α-Geometry 15

2.2.1 α-affine manifold and α-family 20

2.3 (F, G)-Geometry 28

2.3.1 F-affine manifold and F-family 35

2.4 Summary 39

3 Invariant and Non-invariant Geometric Structures **41**

3.1 Divergence and the Induced Geometry 42

3.1.1 f-divergence 43

3.1.2 Bregman divergence 43

3.1.3 (α, ρ, τ)-divergence 44

3.1.4 U-divergence 45

3.2 Invariant and Non-invariant Geometries on a Statistical Manifold . 47

3.2.1 Invariance of the (F, G)-geometry 49

3.2.2 Invariance of the (α, ρ, τ)-geometry 53

3.2.3 (F, G) and (α, ρ, τ)-geometries 59

3.3 Summary 61

4 Deformed Exponential Family 63

4.1 Dually Flat Spaces 64

4.2 Exponential Family and q-Exponential Family 66

4.2.1 Dually flat structure of the exponential family 66

4.2.2 q-Exponential family and the q-structure 68

4.3 Dually Flat Geometries on a Deformed Exponential Family 74

4.3.1 Dually flat U-geometry of the F-exponential family 76

4.3.2 Dually flat χ-geometry of the deformed exponential family . 79

4.4 Summary 88

5 Geometry of Estimation 89

5.1 Parameter Estimation in a Statistical Manifold 89

5.2 Estimation in Exponential Family 92

5.2.1 Estimation in a curved exponential family 94

5.3 Mismatched Estimation in a Curved Exponential Family 100

5.3.1 MLE based on a mismatched model 105

5.4 Summary 112

6 Generalized Estimators **113**

6.1 Maximum F-Likelihood Estimator 114

6.2 F-Escort Maximum Likelihood Estimator 118

6.2.1 F-Maximum entropy theorem 121

6.3 Estimation in F-Exponential Family 124

6.3.1 U-estimator in F-exponential family 124

6.3.2 F-MLE in a F-exponential family 130

6.4 Summary 132

Concluding Remarks **133**

REFERENCES **135**

LIST OF PUBLICATIONS **141**

CHAPTER 1

Introduction

Information geometry emerged from the geometric study of a statistical model of probability distributions. The information geometric tools are widely applied to various fields such as statistics, information theory, stochastic processes, neural networks, statistical physics, neuroscience etc. The importance of the differential geometric approach to the field of statistics was first noticed by Rao [1]. On a statistical model of probability distributions he introduced a Riemannian metric defined by the Fisher information known as the Fisher information metric, see also [2], [3].

One of the major developments in the history of information geometry was the seminal work by Chentsov [4] in which he introduced a family of affine connections on a statistical model defined on finite sets. Efron [5], [6] introduced the notion of statistical curvature of a statistical manifold and mentioned the role of statistical curvature in the asymptotic theory of statistical estimation. His theory used a new affine connection (exponential connection) implicitly. Dawid [7] as a continuation of Efron's work defined another affine connection (mixture connection), see also [8], [9].

Motivated by the works of Efron and Dawid, Amari [10], [11] introduced a one parameter family of affine connections called the α-connections indexed by a real parameter α. These connections are equivalent to the connections introduced by Chentsov [4] on finite sets. This family has a property that the α-connection and the $(-\alpha)$-connection are dual connections with respect to the Fisher information metric. A statistical model of probability distributions endowed with a Riemannian metric and a pair of dual affine connections is called a statistical manifold. A theoretical formulation of information geometry was initially given by Amari [12] and further enriched by Murray and Rice [13], Amari and Nagaoka [14].

The α-geometry consisting of the Fisher information metric and the $(\pm\alpha)$- connections is a significant tool in the higher order asymptotic theory of inference [12]. Amari [12] defined the α-geometry using a particular family of functions called the

α-embedding. Burbea [15] introduced the concept of weighted Fisher information metric using a positive continuous function. Motivated by these works, we considered a general embedding function F and a positive smooth function G to define a more generalized geometric structure on a statistical manifold called the (F, G)-geometry which is an extension of the α-geometry [16].

In Chapter 2 first we describe the manifold structure of a statistical model and the α-geometry. Then the Fisher information metric and the α-connections are computed for statistical manifold defined on finite sets. We define the (F, G)-geometric structures, the G-metric and the dual (F, G)-connections, on a statistical manifold. Then we prove a necessary and sufficient condition for two (F, G)-connections to be dual with respect to the G-metric. We show that the α-geometry is a special case of (F, G)-geometry, thus obtained a generalized class of geometric structures on a statistical manifold which extends the α-geometry. Further the G-metric and the (F, G)-connections are computed for statistical manifold defined on finite sets.

Eguchi [17] introduced a method to define geometric structures on a statistical manifold using a divergence function. The f-divergence and the Bregman divergence are two important classes of divergence functions [18–20]. A more general family of divergences called the (α, ρ, τ)-divergence was introduced by Zhang [21] using a real parameter α and two representations ρ and τ of densities which are conjugate with respect to a strictly convex function. Another class of divergence called the U-divergence was introduced by Murata et al. [22] using a generator function U.

On a statistical manifold one can consider two kinds of invariance of the geometric structures, covariance under reparametrization of the parameter of the manifold and the invariance under the smooth one to one transformations of the random variable [12], [14]. Chentsov [4] proved that the Fisher information metric and the α-connections are unique in the family of probability distributions defined on finite sets with respect to the categorical invariance, see also [23], [24]. Amari [12] conjectured that the Fisher information metric and the α-connections are the only metric and affine connections which are invariant under any coordinate transformations of the sample space and of the parameter. These works motivated us to study the invariance properties of the (F, G)-geometry in which the α-geometry is a special case. We gave a partial answer to Amari's conjecture by proving that the α-geometry is the only geometry among (F, G)-geometries which is both covariant under reparametrization of the parameter and invariant under the smooth one to one transformations of the random variable [16]. Ay et al. [25] studied this problem in the infinite dimensional case also.

Chapter 3 provides an overview of various divergence functions on a statistical manifold and the geometric structures induced by them. We study the invariance properties of the geometric structures on a statistical manifold and classify them into two categories, invariant and non-invariant. We prove that the (F, G)-geometry and the (α, ρ, τ)-geometry are non-invariant geometries on a statistical manifold. First we show that these geometries are covariant under reparametrization of the parameter of the manifold. Then prove that both the (F, G)-geometry and the (α, ρ, τ)-geometry are not invariant under smooth one to one transformations of the random variable in general. Also we show that the α-geometry is the only invariant geometry in the category of both (F, G) and (α, ρ, τ)-geometries. Further the relation between these two geometries are discussed in detail. We show that the (α, ρ, τ)-geometry can always be expressed as (F, G)-geometry and the converse is true only under certain conditions. Some examples are given to illustrate this point.

An exponential family is an important statistical model which is attracted by many of the researchers from Physics, Mathematics and Statistics. Many of the phenomena in the statistical mechanics are modeled by an exponential class of distributions. It is relevant in the context of the Boltzmann-Gibbs entropy maximization problem. It is also equally important from the information geometric point of view. A finite dimensional exponential family has a dually flat structure with respect to (± 1)-connection defined by Amari [12]. Tsallis [26] introduced the notion of non-extensive entropy called the q-entropy or Tsallis entropy which is a generalization of the Botlzmann-Gibbs entropy. This led to the non-extensive statistical mechanics which uses power functions instead of the exponential functions. This motivated many researchers to consider a generalized exponential family called the q-exponential family which is relevant in the q-entropy maximization problem. An information geometric foundation is given to the q-exponential family by Amari and Ohara [27]. A q-exponential family has a dually flat structure called the q-structure which is the conformal flattening of the α-geometry [27].

Naudts [28] introduced a more generalized notion of exponential family called the deformed exponential family and defined a dually flat structure on it, the U-geometry.

Many authors studied the geometry of the deformed exponential family, [29–36]. Amari et al. [37] also considered this family and defined a dually flat structure called the χ-geometry, which is different from the U-geometry. In the case of exponential family the invariant α-geometry gives a dually flat structure. For the deformed exponential family one has to look at the geometric structures other than the invariant α-geometry. In our work we present the role of the non-invariant (F, G)-geometry in the study of dually flat structures of the deformed exponential family [38].

In Chapter 4 first we describe the general structure of a dually flat space. Then the dually flat geometries of the exponential family and the q-exponential family are described in detail. The geometry induced by the conformal transformation of the α-divergence is considered and prove that it is (± 1)-conformally equivalent to the α-geometry. As a corollary to this we obtain that the q-geometry on a q-exponential family is the conformal flattening of the α-geometry. Then we investigate the dually flat structures of a deformed exponential family in detail and provide a clear picture of the state of the art. A description of the two dually flat structures, the U-geometry by Naudts [28] and the χ-geometry by Amari et al. [37], are given. We show that the U-geometry is the (F, G)-geometry and the χ-geometry is the conformal flattening of the (F, G)-geometry for suitable choices of F and G [38]. Thus our study validates the role of non-invariant (F, G)-geometry in the dually flat structures of the deformed exponential family.

Amari [12] demonstrated the significance of α-connection, α-curvature and the duality of connections in the higher order asymptotic theory of inference. He gave a differential geometric framework for the estimation theory. Many researchers have studied the importance of geometric approach in the theory of inference [14], [39–46]. In mismatched neural decoding problem one uses a mismatched model or an unfaithful model instead of the original model [47, 48]. This may be the case when the true model is not observable or may be a simpler model is preferred over the original model for the computational convenience. In Ozumi et al. [48] an information geometric approach to the maximum likelihood estimation based on a mismatched model is described. Motivated by this we construct an information geometric framework for a general estimation problem based on a mismatched model in an exponential family.

In Chapter 5 we discuss the geometric theory of parameter estimation problem in an exponential family and in a curved exponential family. We detail Amari's differential geometric formulation of the asymptotic properties of an estimator in a curved exponential family [12]. Then we describe parameter estimation problem based on a mismatched model in an exponential family. We prove a necessary and sufficient condition for an estimator based on a mismatched model to be consistent and first order efficient. Ozumi et al. [48] stated certain conditions for the maximum likelihood estimator (MLE) based on a mismatched model to be consistent and efficient. We give a theoretical formulation of these results in a curved exponential family and a detailed proof of the same.

The statistical estimation in an exponential family is well studied and the role of the invariant α-geometry in this context is also well established. Naturally, one may think of the estimation problem in a deformed exponential family. What is the role of the non-invariant (F, G)-geometry in the theory of estimation in a deformed exponential family? Recall the theorem by Amari and Nagaoka [14] which states that an estimator on a statistical model $\mathcal{S} = \{p(x; \theta)\}$ is finite sample efficient iff $\mathcal{S}$ is an exponential family and θ is a m-affine coordinate system. Therefore it is natural to expect that the deformed exponential family may not have a finite sample efficient estimator in general. So for the estimation in a deformed exponential family one has to consider certain generalized notions of independence of random variables, MLE, Cramer-Rao lower bound etc.

In the context of the nonextensive thermostatistics Umarov et al. [49] defined the q-independence and the q-central limit theorem using a generalized product called the q-product, see also [50], [51]. Ferrari and Yang [52] defined a maximum L_q-estimator (ML$_q$E) based on the q-entropy and studied its asymptotic behavior in the case of an exponential family. Matsuzoe and Ohara [53] also considered a generalized q-likelihood estimator and studied its geometry in a q-exponential family. Fujimoto and Murata [54] defined a more generalized notion of independence called the U-independence using a smooth strictly convex function U.

Eguchi et al. [36] defined the U-estimator which is a generalization of the MLE and showed that it is consistent and asymptotically normal. They studied the U-estimator in a deformed exponential family also. In general, the U-estimator is not asymptotically efficient. Naudts [28] defined a generalized Cramer-Rao bound using an escort probability distribution and gave sufficient condition for the optimality. He showed that this bound is optimal in a deformed exponential family. That is, a deformed exponential family naturally has an estimator which attains equality in the generalized Cramer-Rao lower bound. It is well known that the MLE for an exponential family is closely related to the dually flat structure. Motivated by this, we explore the relation between an estimator and the two dually flat structures, the U-geometry and the χ-geometry, of a deformed exponential family.

In Chapter 6 first we define F-product, F-independence using a function F and its inverse function Z. Then a generalized MLE called the maximum F-likelihood estimator (F-MLE) is defined and discussed its property as a MAP estimator with a prior. Further using the F-escort probability distribution we define two generalized notions of MLE, the $\mathbf{x}_N$ based F-escort MLE and the F-escort MLE based on the product of F-escort distribution of the marginal probability density of single observations. Also we give a characterization of the q-escort MLE among the $\mathbf{x}_N$ based F-escort MLE as a Bayesian MAP estimator with a prior. Then an analytic proof of the F-version of the maximum entropy theorem is given. Next we give a proof of the generalized Cramer-Rao bound defined by Naudts. Then we show that the U-estimator for the dual coordinate in the U-geometry of a deformed exponential family is optimal with respect to this bound. Further we consider the F-MLE in a deformed exponential family which is given in terms of the dual coordinate in the χ-geometry. To analyze the properties of the F-MLE one need to have certain generalized notions of consistency and efficiency, which is an open problem.

Preliminaries

Here we give the necessary differential geometric tools for the geometric study of statistics [13], [14].

Definition 1.0.1. *An n-dimensional **topological manifold** M is a second countable Hausdorff topological space which is locally Euclidean. So for every point $p \in M$, there exist an open set $U \subset M$ containing p and a homeomorphism $\phi : U \longrightarrow U'$, where U' is an open subset of $\mathbb{R}^n$.*
*(U, ϕ) is called a **coordinate chart** on M around p and $\phi = (x^i),\ i = 1, \cdots, n$ are*

called **local coordinates** *on U. When $U = M$, (U, ϕ) is called a* **global chart** *and we obtain a* **global coordinate system** *on M.*

If we have two charts (U, φ) and (V, ψ) on M such that $U \cap V \neq \varnothing$, the composite map $\psi \circ \varphi^{-1} : \varphi(U \cap V) \longrightarrow \psi(U \cap V)$ is called the **transition map**. The two charts (U, φ) and (V, ψ) are said to be **smoothly compatible** if either $U \cap V = \varnothing$ or the transition map $\psi \circ \varphi^{-1}$ is a diffeomorphism.

An **atlas** $\mathcal{A}$ for M is the collection of charts whose domain cover M and $\mathcal{A}$ is said to be a **smooth atlas** if any two charts in $\mathcal{A}$ are smoothly compatible with each other. $\mathcal{A}$ is a **maximal atlas** if any chart that is smoothly compatible with every charts in $\mathcal{A}$ is in $\mathcal{A}$. A **smooth structure** on any topological manifold is a maximal smooth atlas on M. A **smooth manifold** is a pair $(M, \mathcal{A})$, where M is a topological manifold and $\mathcal{A}$ is a smooth structure on M.

Definition 1.0.2. *Let M be a smooth manifold. A function $f : M \longrightarrow \mathbb{R}$ is said to be* ***smooth*** *if $f \circ \varphi^{-1}$ is smooth for some smooth chart (U, φ) around each point. The set of all smooth functions from M to $\mathbb{R}$ is denoted by $C^\infty(M)$ which is a vector space over $\mathbb{R}$.*

Definition 1.0.3. *A linear map $X : C^\infty(M) \longrightarrow \mathbb{R}$ is called a* ***derivation*** *of $C^\infty(M)$ at p if it satisfies the following*

$$X(fg) = f(p)Xg + g(p)Xf, \quad \forall \quad f, g \in C^\infty(M) \tag{1.1}$$

Let M be a smooth manifold and let $p \in M$. The **tangent space** to M at p, denoted by T_pM, is defined as the set of all derivations of $C^\infty(M)$ at p.

Let $(U, \phi = (x^i))$ be a smooth chart on M around p. Then T_pM is a vector space of dimension n with basis $\{\frac{\partial}{\partial x^i}|_p, \ i = 1, \cdots, n\}$. Each element in T_pM is called a **tangent vector** at p. Let T_p^*M denote the dual space of T_pM which is also an n-dimensional vector space and $\{dx^i|_p, \ i = 1, \cdots, n\}$ forms a basis. Elements of T_p^*M are called **cotangent vectors** at p.

A **tangent bundle** TM on M is the disjoint union of tangent spaces at all points of M.

$$TM = \bigcup_{p \in M} T_pM \tag{1.2}$$

A **cotangent bundle** T^*M on M is the disjoint union of cotangent spaces at all points of M

$$T^*M = \bigcup_{p \in M} T_p^*M \tag{1.3}$$

A **vector field** X on a smooth manifold M is a map $X : M \longrightarrow TM$ which associates to each point $p \in M$ a tangent vector $X_p \in T_pM$. X is said to be a **smooth vector field** if it is smooth as a map from M to TM. Let $\Gamma(TM)$ denote the set of all smooth vector fields on M.

Definition 1.0.4. *Let M be an n-dimensional smooth manifold. A **Riemannian metric** $g =<, >$ on M is a smooth symmetric 2-tensor field which is positive definite at each point. So for every $p \in M$, $g_p =<, >_p : T_pM \times T_pM \longrightarrow \mathbb{R}$ is bilinear, symmetric and positive definite.*

A **Riemannian manifold** is a manifold equipped with a Riemannian metric.

Definition 1.0.5. *Let M be an n-dimensional smooth manifold. A **linear** or an **affine connection** on M is defined as a map $\nabla : \Gamma(TM) \times \Gamma(TM) \longrightarrow \Gamma(TM)$ which satisfies the following*

1. $\nabla_X(Y + Z) = \nabla_X Y + \nabla_X Z$
2. $\nabla_{(X+Y)}Z = \nabla_X Z + \nabla_Y Z$
3. $\nabla_X(fY) = f\nabla_X Y + (Xf)Y$
4. $\nabla_{fX}Y = f\nabla_X Y$

for all $f \in C^\infty(M)$ and $X, Y, Z \in \Gamma(TM)$.

Let $(U, \phi = (x^i))$ be a smooth chart in M. Then $\{\partial_i = \frac{\partial}{\partial x^i},\ i = 1, \cdots, n\}$ are smooth vector fields on U called **coordinate vector fields** on U. The affine connection ∇ can be locally determined by n^3 functions Γ_{ij}^k given by

$$\nabla_{\partial_i}\partial_j = \sum_k \Gamma_{ij}^k \partial_k \tag{1.4}$$

where Γ_{ij}^k are called the **Christoffel symbols** of the affine connection ∇ with respect to the coordinates $(x^i),\ i = 1, \cdots, n$.

If ∇ is an affine connection on a Riemannian manifold M with a Riemannian metric $g =<, >$

$$< \nabla_{\partial_i}\partial_j, \partial_m >= \sum_k \Gamma_{ij}^k < \partial_k, \partial_m >= \sum_k \Gamma_{ij}^k g_{km} \tag{1.5}$$

where $g_{km} =< \partial_k, \partial_m >$.

It is often convenient to express the Christoffel symbols of the affine connection ∇ by

$$\Gamma_{ijm} = \sum_k \Gamma_{ij}^k g_{km} =< \nabla_{\partial_i}\partial_j, \partial_m > \tag{1.6}$$

The n^3 functions Γ_{ijm} are called the **components** of the affine connection with respect to the coordinate (x^i).

Definition 1.0.6. *Let M be a Riemannian manifold with a Riemannian metric g. A connection ∇ is said to be a **metric connection** if it satisfies*

$$d(g(X, Y)) = g(\nabla X, Y) + g(X, \nabla Y) \tag{1.7}$$

where d is the differential operator.

Definition 1.0.7. *A connection is said to be **symmetric** or **torsion free** if the torsion tensor $T(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y]$ vanishes. That is, $\Gamma_{ij}^k = \Gamma_{ji}^k$.*

An affine connection which is both symmetric and metric is called the **Riemannian connection** or **Levi-Civita connection** with respect to g. Given a metric g, there exist a unique Levi-Civita connection ∇ with respect to g given by

$$\Gamma_{ijk} = \frac{1}{2}\left(\partial_i g_{jk} + \partial_j g_{ki} + \partial_k g_{ij}\right) \tag{1.8}$$

Definition 1.0.8. *Let M be a smooth manifold and let ∇ be an affine connection on M. If there exists a coordinate system $\theta = (\theta^i)$ such that $\nabla_{\partial_i}\partial_j = 0$ then we say that the connection ∇ is a **flat connection** or M is flat with respect to ∇. Then the coordinate system θ is called an **affine coordinate system** for M or we say that θ is ∇-affine.*

Definition 1.0.9. *Let (M, g) be a Riemannian manifold with a Riemannian metric g and a coordinate system $(x^1, \cdots, x^n)$. Let $\gamma : [a, b] \longrightarrow \mathcal{S}$ be a curve in M. Define*

$\gamma^i(t) = x^i(\gamma(t))$. *Then* ***the tangent vector (velocity vector)*** *to* γ *is*

$$\dot{\gamma}(t) = \sum_{i=1}^{n} \dot{\gamma}^i(t)\, \partial_i, \quad \text{where } \dot{\gamma}^i(t) = \frac{d}{dt}\gamma^i(t) \tag{1.9}$$

A curve γ in M is said to be a **geodesic** for an affine connection ∇ if its velocity is constant according to ∇. That is,

$$\nabla_{\dot{\gamma}}\dot{\gamma} = 0. \tag{1.10}$$

CHAPTER 2

Geometric Structures on a Statistical Manifold

In this chapter a generalized geometric structure called the (F, G)-geometry is introduced on a statistical manifold which includes Amari's α-geometry as a special case [16]. A statistical manifold of probability distributions is equipped with a Riemannian metric and a pair of dual affine connections [1], [4], [12–14]. It was Rao [1] who first explicitly introduced a Riemannian metric on a statistical manifold called the Fisher information metric. Chentsov [4] introduced a family of affine connections in a statistical manifold defined on finite sets. Amari [12] introduced a family of affine connections called α-connections using a one parameter family of functions, the α-embeddings, see also [5–7], [14]. These α-connections are equivalent to those defined by Chentsov [4]. Burbea [15] introduced the concept of weighted Fisher information metric using a positive continuous function. Motivated by these works, to define more general geometric structures on a statistical manifold, we considered a general embedding function F and a positive smooth function G and defined a geometry called the (F, G)-geometry [16]. The α-geometry turned out to be a special case of (F, G)-geometry.

In Section 2.1 we describe the affine structure of the family of measures and the manifold structure of a statistical model of probability distributions. In Section 2.2 a short account of Amari's α-geometric structure is presented and the Fisher information metric and the α-connections are computed for the statistical manifold defined on finite sets. Then in Section 2.3 we give a detailed description of the dualistic (F, G)-geometry on a statistical manifold and prove the necessary and sufficient conditions for two (F, G)-connections to be dual with respect to the G-metric. Also prove that the α-geometry is a special case of the (F, G)-geometry. Further the G-metric and the (F, G)-connections are computed for statistical manifold defined on finite sets.

2.1 Statistical Manifold

In this section we first discuss about the affine structure of the family of measures defined on a measurable space under certain regularity conditions (refer [13] for more details). Then we consider the family of probability measures and describe the manifold structure of a statistical model of probability distributions.

2.1.1 Affine structure of the family of measures

Definition 2.1.1. *Let V be an n-dimensional real vector space. An n-**dimensional affine space** over the vector space V is a non-empty set $\mathcal{E}$ together with a translation map $+ : V \times \mathcal{E} \longrightarrow \mathcal{E}$, $(v, p) \longmapsto v + p$ which satisfies*

1. $v + (w + p) = (v + w) + p, \quad \forall\, v, w \in V,\ \forall\, p \in \mathcal{E}.$
2. *For any two points $p, q \in \mathcal{E}$, $\ \exists$ a unique vector $v \in V$ such that $q = p + v$.*

An affine space can be thought of as a set which becomes a vector space by selecting a point to be the origin.

Let $(\mathcal{X}, \mathcal{B})$ be a measurable space, where $\mathcal{X}$ is a non-empty set and $\mathcal{B}$ is the σ-field of subsets of $\mathcal{X}$. Consider the family $\mathcal{A}$ of non-negative σ-finite measures on $(\mathcal{X}, \mathcal{B})$. Define an equivalence relation $\sim$ on $\mathcal{A}$ by two measures in $\mathcal{A}$ are equivalent if they are absolutely continuous with respect to each other. Let $\mathcal{M}$ denote one of the equivalence classes of $\mathcal{A}$.

Let $R_{\mathcal{X}}$ be the set of all real valued measurable functions defined on $(\mathcal{X}, \mathcal{B})$. In general, $R_{\mathcal{X}}$ is an infinite dimensional vector space. Then $\mathcal{M}$ is an affine space over the vector space $R_{\mathcal{X}}$ under the translation map defined by

$$\nu + f = e^f \nu, \quad \forall\, f \in R_{\mathcal{X}},\ \nu \in \mathcal{M}. \tag{2.1}$$

1. For any $\mu \in \mathcal{M}$ and $f \in R_{\mathcal{X}}$, $\ \nu = e^f \mu$ is a non-negative σ-finite measure. Whenever $\mu(E) = 0$ for $E \in \mathcal{B}$, $\nu(E) = \int\limits_E e^f d\mu = 0$ and hence $e^f \mu \in \mathcal{M}$. Hence the translation map is well defined.

2. For any two measures $\nu, \mu \in \mathcal{M}$, $\exists$ a unique function $f = \frac{d\nu}{d\mu} \in R_{\mathcal{X}}$ (f is the Radon-Nikodym derivative) which translates μ to ν. We often call e^f as the density function with respect to the measure μ.

3. $\forall\ f, g\ \in R_{\mathcal{X}},\ \forall \mu \in \mathcal{M}$, $(\mu + f) + g = e^f \mu + g = e^g e^f \mu = e^{f+g} \mu = \mu + (f + g)$ (Note that the same symbol $+$ is used for vector addition and affine space translation map).

Remark 2.1.2. *Since $\mathcal{M}$ is an affine space over $R_{\mathcal{X}}$, by choosing an origin μ, $\mathcal{M}$ can be identified with the vector space $R_{\mathcal{X}}$. It is equivalent to saying that any measure in $\mathcal{M}$ can be expressed as densities with respect to a base measure.*

2.1.2 Statistical manifold

Let Ω be the sample space associated with some random experiment and $\mathcal{F}$ be the σ-field of subsets of Ω. Then a probability measure P on $(\Omega, \mathcal{F})$ is a measure satisfying $P(\Omega) = 1$ and $(\Omega, \mathcal{F}, P)$ is called a **probability space**.

Now consider a measurable space $(\mathcal{X}, \mathcal{B})$, where $\mathcal{B}$ is the σ-field of subsets of $\mathcal{X}$. The $(\mathcal{X}, \mathcal{B})$-valued random variable X is defined as a $(\mathcal{F}, \mathcal{B})$-measurable function from $\Omega \longrightarrow \mathcal{X}$. The probability measure P on $(\Omega, \mathcal{F})$ induces a probability measure X_*P on $(\mathcal{X}, \mathcal{B})$ defined by

$$X_*P(B) = P(X^{-1}(B)), \quad \forall\ B \in \mathcal{B}. \tag{2.2}$$

Assume that X_*P is absolutely continuous with respect to a σ-finite measure μ on $(\mathcal{X}, \mathcal{B})$. Then the density of X with respect to μ is the Radon-Nikodym derivative p given by

$$p = \frac{dX_*P}{d\mu}. \tag{2.3}$$

That is $p : \mathcal{X} \longrightarrow [0, \infty)$ is a measurable function such that

$$X_*P(B) = \int_{X^{-1}(B)} dP = \int_B p\, d\mu. \tag{2.4}$$

In most of the applications $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{B} = \mathcal{B}(\mathbb{R}^n)$ is the σ-algebra of Borel subsets of $\mathbb{R}^n$. We know that measures can be expressed as densities with respect to some base measure. In this case the base measure can be taken as the Lebesgue measure on $\mathbb{R}^n$. So any probability measure on $\mathcal{X}$ can be represented in terms of density function with respect to Lebesgue measure.

A probability distribution on $\mathcal{X}$ is a function $p : \mathcal{X} \longrightarrow \mathbb{R}$ satisfying

$$p(x) \geq 0, \quad \forall\, x \in \mathcal{X} \quad \text{and} \tag{2.5}$$

$\sum_{x\in\mathcal{X}} p(x) = 1$ if $\mathcal{X}$ is a discrete set (finite or countably infinite) and $\int\limits_{\mathcal{X}} p(x)dx = 1$ (Note that if $n \geq 2$, then $\int$ denotes a multiple integral) if $\mathcal{X} = \mathbb{R}^n$.

Definition 2.1.3. *Consider a family $\mathcal{S}$ of probability distributions on $\mathcal{X}$. Suppose each element of $\mathcal{S}$ can be parametrized using n real-valued variables $(\theta^1, \cdots, \theta^n)$ so that*

$$\mathcal{S} = \{p_\theta = p(x;\theta) \;/\; \theta = (\theta^1, \cdots, \theta^n) \in \mathbb{E}\} \tag{2.6}$$

where $\mathbb{E}$ is a subset of $\mathbb{R}^n$ and the mapping $\theta \mapsto p_\theta$ is injective. Such a family $\mathcal{S}$ is called an n-dimensional ***statistical model*** *or a* ***parametric model*** *or simply a* ***model*** *on θ. We often write it as $\mathcal{S} = \{p_\theta\}$.*

Now we state certain regularity conditions regarding the statistical model $\mathcal{S} = \{p_\theta\}$ which are required for our geometric theory [12], [14].

Regularity conditions

1. $\mathbb{E}$ is an open subset of $\mathbb{R}^n$ and for each $x \in \mathcal{X}$, the function $\theta \mapsto p(x;\theta)$ is of class c^∞.

2. Let $\ell(x;\theta) = \log p(x;\theta)$ and $\partial_i = \frac{\partial}{\partial\theta^i}$. For every fixed θ, n functions in x $\{\partial_i \ell(x;\theta),\ i = 1, \cdots, n\}$ are linearly independent and are known as **scores**.

3. The order of integration and differentiation may be freely rearranged.

4. The moments of scores exists upto necessary orders.

5. The supp(p_θ) does not vary with respect to θ, where supp$(p_\theta) := \{x \;/\; p(x;\theta) > 0\}$. Then we can redefine $\mathcal{X}$ to be supp(p_θ). This is equivalent to $p(x;\theta) > 0$ holds for all $\theta \in \mathbb{E}$ and all $x \in \mathcal{X}$. So the model $\mathcal{S}$ is a subset of

$$\mathcal{P}(\mathcal{X}) := \{p : \mathcal{X} \longrightarrow \mathbb{R} \;/\; p(x) > 0 \;(\forall\, x \in \mathcal{X}),\; \int\limits_{\mathcal{X}} p(x)dx = 1\} \tag{2.7}$$

Definition 2.1.4. *For a model $\mathcal{S} = \{p_\theta \ / \ \theta \in \mathbb{E}\}$, the mapping $\varphi : \mathcal{S} \longrightarrow \mathbb{R}^n$ defined by $\varphi(p_\theta) = \theta$ allows us to consider $\varphi = (\theta^i)$ as a coordinate system for $\mathcal{S}$. Suppose there is a c^∞ diffeomorphism $\psi : \mathbb{E} \longrightarrow \psi(\mathbb{E})$, where $\psi(\mathbb{E})$ is an open subset of $\mathbb{R}^n$. Then if we use $\rho = \psi(\theta)$ instead of θ as the parameter we obtain $\mathcal{S} = \{p_{\psi^{-1}(\rho)} \mid \rho \in \psi(\mathbb{E})\}$. This expresses the same family of probability distributions $\mathcal{S} = \{p_\theta\}$. Then $\mathcal{S}$ is a c^∞ differentiable manifold by considering parametrizations which are c^∞ diffeomorphic to each other to be equivalent and is called a* ***statistical manifold****. Note that (θ^i) is a global coordinate system on $\mathcal{S}$.*

Example 2.1.5. (Normal Distribution)
$\mathcal{X} = \mathbb{R}$, $n = 2$, $\theta = (\mu, \sigma)$, $E = \{(\mu, \sigma) \ / \ -\infty < \mu < \infty, 0 < \sigma < \infty\}$

$$N(\mu, \sigma) = \{p(x; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \ / \ \theta = (\mu, \sigma) \in E\}. \tag{2.8}$$

This is a 2-dimensional manifold which can be identified with the upper half plane.

Note 2.1.6. *In this thesis we will be considering only finite dimensional statistical manifolds.*

2.2 α-Geometry

A statistical manifold naturally has a Riemannian metric called the Fisher information metric introduced by Rao [1]. Amari [12] defined a one parameter family of affine connections on a statistical manifold called the α-connection using a family of functions called the α-embedding. This family of connections has a property that the α-connection and the $(-\alpha)$-connection are dual connections with respect to the Fisher information metric. The α-geometry consisting of the $(\pm\alpha)$-connections together with the Fisher information metric is an important tool in the geometric theory of statistical estimation [12]. Here we present a short description of the α-geometry on a statistical manifold.

Let $\mathcal{S} = \{p_\theta \ / \ \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$ be a statistical manifold. The tangent space $T_\theta(\mathcal{S})$ to $\mathcal{S}$ at a point p_θ is given by

$$T_\theta(\mathcal{S}) = \{\sum_{i=1}^{n} \alpha^i \partial_i \ / \ \alpha^i \in \mathbb{R}\}. \tag{2.9}$$

There is a more convenient way of representing the tangent space to a statistical manifold. The set of scores $\{\partial_i \ell(x;\theta),\ i = 1, \cdots, n\}$ is linearly independent by the assumption, so define an n-dimensional vector space spanned by the scores as

$$T_\theta^1(\mathcal{S}) = \{A(x) \ / \ A(x) = \sum_{i=1}^{n} A^i \partial_i \ell(x;\theta),\ A^i \in \mathbb{R}\}. \tag{2.10}$$

Then there is a natural isomorphism between the two vector spaces $T_\theta(\mathcal{S})$ and $T_\theta^1(\mathcal{S})$ given by

$$\partial_i \in T_\theta(\mathcal{S}) \longleftrightarrow \partial_i \ell(x;\theta) \in T_\theta^1(\mathcal{S}). \tag{2.11}$$

Any tangent vector $A = \sum_{i=1}^{n} A^i \partial_i \in T_\theta(\mathcal{S})$ corresponds to a random variable $A(x) = \sum_{i=1}^{n} A^i \partial_i \ell(x;\theta) \in T_\theta^1(\mathcal{S})$ having the same coefficients A^i. Note that $T_\theta(\mathcal{S})$ is the differentiation operator representation of the tangent space, while $T_\theta^1(\mathcal{S})$ is the random variable representation of the same tangent space. The space $T_\theta^1(\mathcal{S})$ is called the **1-representation of the tangent space**.

Define expectation with respect to the distribution $p(x;\theta)$ as

$$E_\theta(f) = \int f(x) p(x;\theta) dx. \tag{2.12}$$

Note that $E_\theta[\partial_i \ell(x;\theta)] = 0$ since $\int p(x;\theta)dx = 1$. Hence for any random variable $A(x) \in T_\theta^1(\mathcal{S})$, $E_\theta[A(x)] = 0$.

This expectation induces an inner product on $T_\theta \mathcal{S}$ in a natural way. Let A and B be two tangent vectors in $T_\theta(\mathcal{S})$ and $A(x)$ and $B(x)$ be the corresponding 1-representations. Then the inner product $g =<,>$ is defined as

$$g(A,B)(\theta) =< A, B >_\theta = \ E_\theta[A(x)B(x)]. \tag{2.13}$$

Denote the inner product of the basis vectors ∂_i and ∂_j by $g_{ij}(\theta)$ which is given by

$$g_{ij}(\theta) =< \partial_i, \partial_j >_\theta = E_\theta[\partial_i \ell(x;\theta) \partial_j \ell(x;\theta)] = \int \partial_i \ell(x;\theta) \partial_j \ell(x;\theta) p(x;\theta) dx \tag{2.14}$$

Here we assume that the integral in Equation (2.14) exists for all $\theta \in \mathbb{E}$.

It is clear that the matrix $G(\theta) = (g_{ij}(\theta))$ is symmetric.

For any n-dimensional vector $c = [c^1, \cdots, c^n]^t$

$$c^t G(\theta) c = \int \{\sum_{i=1}^{n} c^i \partial_i \ell(x;\theta)\}^2 p(x;\theta) dx \geqslant 0. \tag{2.15}$$

Since $\{\partial_i \ell, \ i = 1, \cdots, n\}$ is linearly independent, from Equation (2.15) it follows that $G(\theta)$ is positive definite for all θ. Hence $g =<, >$ defined in Equation (2.14) is a Riemannian metric on the statistical manifold $\mathcal{S}$, called the **Fisher information metric**. The matrix $G(\theta)$ is called the **Fisher information matrix** of $\mathcal{S}$ at the point p_θ [12].

Example 2.2.1. Normal distribution

For the normal family

$$\mathcal{S} = N(\mu, \sigma) = \{p(x;\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \ / \ \theta = (\mu, \sigma) \in E\} \tag{2.16}$$

with parameters $\theta = (\mu, \sigma)$, the log-likelihood function is given by

$$\ell(x, \theta) = -\frac{(x-\mu)^2}{2\sigma^2} - \log \sqrt{2\pi}\sigma. \tag{2.17}$$

Let $\partial_1 = \frac{\partial}{\partial \mu}$ and $\partial_2 = \frac{\partial}{\partial \sigma}$. The tangent space $T^1_\theta \mathcal{S}$ is spanned by

$$\partial_1 \ell = \frac{(x-\mu)}{\sigma^2}, \quad \partial_2 \ell = -\frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma}. \tag{2.18}$$

Then the Fisher information matrix $G(\theta) = (g_{ij}(\theta))$ is

$$G(\theta) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix} \tag{2.19}$$

Definition 2.2.2. *Let $\mathcal{S} = \{p(x;\theta) \ / \ \theta \in \mathbb{E}\}$ be an n-dimensional statistical manifold with the Fisher information metric $g =<, >$. Define n^3 functions Γ^1_{ijk} by*

$$\Gamma^1_{ijk}(\theta) = E_\theta[(\partial_i \partial_j \ell(x;\theta)) \partial_k \ell(x;\theta)] \tag{2.20}$$

which uniquely determine an affine connection ∇^1 on the statistical manifold $\mathcal{S}$ called the ***1-connection*** *or the* ***exponential connection*** *given by*

$$\Gamma^1_{ijk}(\theta) =< \nabla^1_{\partial_i} \partial_j, \partial_k >_\theta . \tag{2.21}$$

For defining the 1-connection Amari [12] used $\ell(x;\theta)$, the logarithm of the density function $p(x;\theta)$. To obtain more general geometric structures on $\mathcal{S}$ Amari [12] used a one parameter family of functions called the α-embedding instead of $\ell(x;\theta)$.

Definition 2.2.3. *The α-**embedding** $L_\alpha(p)$ is a one parameter family of functions defined by*

$$L_\alpha(p) = \begin{cases} \frac{2}{1-\alpha} p^{\frac{1-\alpha}{2}}, & \alpha \neq 1 \\ \log p, & \alpha = 1 \end{cases} \tag{2.22}$$

*called the α-**representation** of the density function $p(x;\theta)$.*

Let $\ell_\alpha(x;\theta) = L_\alpha(p(x;\theta))$. Note that the 1-representation $\ell_1(x;\theta)$ is the log-likelihood function $\ell(x;\theta)$ and the (-1)-representation $\ell_{-1}(x;\theta)$ is the density function $p(x;\theta)$ itself.

Let $T_\theta^\alpha(\mathcal{S})$ be an n-dimensional vector space spanned by n linearly independent functions $\{\partial_i \ell_\alpha(x;\theta),\ i = 1, \cdots, n\}$ in x,

$$T_\theta^\alpha(\mathcal{S}) = \{A_\alpha(x) \ / \ A_\alpha(x) = \sum_{i=1}^{n} A^i \partial_i \ell_\alpha(x;\theta), A^i \in \mathbb{R}\}. \tag{2.23}$$

There is a natural isomorphism between the two vector spaces $T_\theta(\mathcal{S})$ and $T_\theta^\alpha(\mathcal{S})$ given by

$$\partial_i \in T_\theta(\mathcal{S}) \longleftrightarrow \partial_i \ell_\alpha(x;\theta) \in T_\theta^\alpha(\mathcal{S}). \tag{2.24}$$

The vector space $T_\theta^\alpha(\mathcal{S})$ is called the **α-representation of the tangent space** $T_\theta(\mathcal{S})$. The α-representation of a vector $A = \sum_{i=1}^n A^i \partial_i \in T_\theta(\mathcal{S})$ is the random variable

$$A_\alpha(x) = \sum_{i=1}^{n} A^i \partial_i \ell_\alpha(x;\theta). \tag{2.25}$$

We have the relations

$$\partial_i \ell_\alpha = p^{\frac{(1-\alpha)}{2}} \partial_i \ell \tag{2.26}$$

$$\partial_i \partial_j \ell_\alpha = p^{\frac{(1-\alpha)}{2}} (\partial_i \partial_j \ell + \frac{1-\alpha}{2} \partial_i \ell \partial_j \ell) \tag{2.27}$$

Define the α-**expectation** of a random variable f with respect to the density $p(x;\theta)$ as

$$E_\theta^\alpha(f) = \int f(x)(p(x;\theta))^\alpha dx. \tag{2.28}$$

This induces an inner product on $\mathcal{S}$ given by

$$< A, B >_\theta^\alpha = E_\theta^\alpha[A_\alpha(x) B_\alpha(x)] \tag{2.29}$$

where $A_\alpha(x), B_\alpha(x)$ are the α-representations of $A, B \in T_\theta(\mathcal{S})$.

Using Equation (2.26), the inner product of the basis vectors is given by

$$\begin{aligned} < \partial_i, \partial_j >_\theta^\alpha &= \int \partial_i \ell_\alpha \, \partial_j \ell_\alpha \, p^\alpha \, dx = \int \partial_i \ell_\alpha \, \partial_j \ell_{-\alpha} \, dx & (2.30) \\ &= \int \partial_i \ell \, \partial_j \ell \, p \, dx = g_{ij}(\theta) & (2.31) \end{aligned}$$

which is the Fisher information metric g.

That is, the α-expectation induces the Fisher information metric on $\mathcal{S}$.

Definition 2.2.4. *Let $\mathcal{S} = \{p(x;\theta) \ / \ \theta \in \mathbb{E}\}$ be a statistical manifold with the Fisher information metric $g =<,>$. Using the α-representation of the density function define n^3 functions Γ_{ijk}^α for each $\alpha \in \mathbb{R}$ as*

$$\begin{aligned} \Gamma_{ijk}^\alpha &= \int \partial_i \partial_j \ell_\alpha(x;\theta) \partial_k \ell_{-\alpha}(x;\theta) dx & (2.32) \\ &= \int (\partial_i \partial_j \ell + \frac{1-\alpha}{2} \partial_i \ell \, \partial_j \ell) \, \partial_k \ell \, p \, dx & (2.33) \end{aligned}$$

where the last equation follows from Equations (2.26) *and* (2.27).

*These Γ_{ijk}^α uniquely determine an affine connection ∇^α on the statistical manifold $\mathcal{S}$ called the **α-connection** given by*

$$\Gamma_{ijk}^\alpha = < \nabla_{\partial_i}^\alpha \partial_j, \partial_k > . \tag{2.34}$$

Thus the one parameter family of functions $L_\alpha(p)$ defines a family of connections ∇^α, $\alpha \in \mathbb{R}$ on the statistical manifold $\mathcal{S}$.

Definition 2.2.5. *Let M be a Riemannian manifold with a Riemannian metric $g =<,>$. Two affine connections ∇ and ∇^* on M are said to be **dual connections with respect to the metric** g if*

$$d(g(X,Y))(Z) = g(\nabla_Z X, Y) + g(X, \nabla_Z^* Y) \tag{2.35}$$

for all $X, Y, Z \in \Gamma(TM)$, where d is the differential operator.
Then the triple (g, ∇, ∇^*) is called a ***dualistic structure*** *on M.*

Letting $\Gamma_{ijk} = < \nabla_{\partial_i} \partial_j, \partial_k >, \ \Gamma^*_{ijk} = < \nabla^*_{\partial_i} \partial_j, \partial_k >$, Equation (2.35) can be written in terms of the basis vectors as

$$\partial_i g_{jk} = \Gamma_{ijk} + \Gamma^*_{ikj}. \tag{2.36}$$

Note that every affine connection has a unique dual with respect to a Riemannian metric and if the affine connection is metric, then it is self dual.
Amari [12] proved the following theorem,

Theorem 2.2.6. *The α-connection ∇^α and the $(-\alpha)$-connection $\nabla^{-\alpha}$ are dual with respect to the Fisher information metric g. In particular, the 0-connection is the Levi-Civita connection with respect to g.*

Remark 2.2.7. *On a statistical manifold $\mathcal{S}$, the triple $(g, \nabla^\alpha, \nabla^{-\alpha})$ consisting of $(\pm\alpha)$-connections $\nabla^{\pm\alpha}$ with the Fisher information metric g defines a dualistic structure.*

2.2.1 α-affine manifold and α-family

Amari [12] defined the notions of α-affine manifold and α-family and described them on a statistical manifold defined on finite sets. Here first we give an overview of his work and then compute the Fisher information metric and α-connections on a statistical manifold defined on finite sets.

Let $\mathbb{R}_\mathcal{X}$ be the set of all real valued measurable functions on $\mathcal{X}$. Consider the set of all finite positive measures $\tilde{\mathcal{P}}(\mathcal{X})$ on $\mathcal{X}$ given by

$$\tilde{\mathcal{P}}(\mathcal{X}) := \{p : \mathcal{X} \longrightarrow \mathbb{R} \ / \ p(x) > 0 \ (\forall \ x \in \mathcal{X}); \int_\mathcal{X} p(x)dx < \infty\} \subset \mathbb{R}_\mathcal{X}. \tag{2.37}$$

The set of all probability distributions $\mathcal{P}(\mathcal{X})$ on $\mathcal{X}$ is a subset of $\tilde{\mathcal{P}}(\mathcal{X})$ determined as

$$\mathcal{P}(\mathcal{X}) := \{p(x) \in \tilde{\mathcal{P}}(\mathcal{X}); \ \int_\mathcal{X} p(x)dx = 1\}. \tag{2.38}$$

For an n-dimensional statistical manifold $\mathcal{S} = \{p(x;\xi) \ / \ \xi = (\xi^1, \cdots, \xi^n) \in \mathbb{E} \subseteq$

$\mathbb{R}^n\} \subseteq \mathcal{P}(\mathcal{X})$ define the **denormalization** $\tilde{\mathcal{S}}$ of $\mathcal{S}$ by

$$\tilde{\mathcal{S}} = \{c\, p(x;\xi) \;/\; c > 0,\; p(x;\xi) \in \mathcal{S}\} \;\subseteq \tilde{\mathcal{P}}(\mathcal{X}). \tag{2.39}$$

Note that $\tilde{\mathcal{S}}$ is an $(n+1)$-dimensional manifold and $\mathcal{S}$ is a submanifold of $\tilde{\mathcal{S}}$.

Definition 2.2.8. *Let $\mathcal{S} = \{p(x;\xi) \;/\; \xi = (\xi^1, \cdots, \xi^n) \in \mathbb{E} \subseteq \mathbb{R}^n\}$ be an n-dimensional statistical manifold. If for some coordinate system $\theta = (\theta^i),\; i = 1, \cdots, n$*

$$\partial_i \partial_j \ell_\alpha(x;\theta) = 0 \tag{2.40}$$

then from Equation (2.32) *θ is a ∇^α-affine coordinate system (referred as α-affine coordinate system) and that $\mathcal{S} = \{p_\theta\}$ is ∇^α-flat (referred as α-flat). Then $\mathcal{S}$ is said to be an **α-affine manifold**.*

The above condition is equivalent to the existence of the functions $C, F_1, \cdots, F_n$ on $\mathcal{X}$ such that

$$\ell_\alpha(x;\theta) = C(x) + \sum_{i=1}^{n} \theta^i F_i(x). \tag{2.41}$$

Definition 2.2.9. *A statistical manifold $\mathcal{S} = \{p(x;\xi) \;/\; \xi = (\xi^1, \cdots, \xi^n) \in \mathbb{E} \subseteq \mathbb{R}^n\}$ is said to be an **α-family** if its denormalization $\tilde{\mathcal{S}}$ is an α-affine manifold.*

For an n-dimensional α-family $\mathcal{S}$, there exists a coordinate system $\theta = (\theta^i)$ and functions $C_1(x), \cdots, C_n(x), \psi(\theta)$ such that

$$\ell_\alpha(x;\theta) = \sum_{i=1}^{n} \theta^i C_i(x) - \psi(\theta) \tag{2.42}$$

where $\psi(\theta)$ is obtained from the normalization condition $\int_{\mathcal{X}} p(x;\theta)dx = 1$.

Remark 2.2.10. *When $\mathcal{X}$ is infinite, $\tilde{\mathcal{P}}(\mathcal{X})$ and $\mathcal{P}(\mathcal{X})$ are infinite dimensional spaces. Hence the manifold structure of these spaces cannot be described in the usual way. Here we consider $\tilde{\mathcal{P}}(\mathcal{X})$ and $\mathcal{P}(\mathcal{X})$ for finite $\mathcal{X}$.*

Let $\mathcal{X} = \{x_1, \cdots, x_n\}$ be a finite set with cardinality n. Now consider the measurable space $(\mathcal{X}, \wp(\mathcal{X}))$, where $\wp(\mathcal{X})$ be the power set of $\mathcal{X}$. Let $\mathbb{R}_{\mathcal{X}}$ be the space of all real valued measurable functions defined on $(\mathcal{X}, \wp(\mathcal{X}))$. Any real valued measurable function m on $\mathcal{X}$ can be specified by n-real numbers $m_1 = m(x_1), \cdots, m_n = m(x_n)$.

Hence $\mathbb{R}_{\mathcal{X}}$ can be identified with $\mathbb{R}^n$ with coordinates $(m_1, \cdots, m_n)$. Then the set $\tilde{\mathcal{P}}(\mathcal{X})$ can be identified with the first orthant in $\mathbb{R}^n$. That is. $\tilde{\mathcal{P}}(\mathcal{X})$ can be identified with the subset $\{(m_1, \cdots, m_n) \ / \ m_i > 0, \ \forall \ i = 1, \cdots, n \ \}$ of $\mathbb{R}^n$.

Theorem 2.2.11. *For a finite set $\mathcal{X}$ of cardinality n, $\tilde{\mathcal{P}}(\mathcal{X})$ is an α-affine manifold for any $\alpha \in \mathbb{R}$.*

Proof. Let $\mathcal{X} = \{x_1, \cdots, x_n\}$ be a finite set constituting n elements. Let $F_i : \mathcal{X} \longrightarrow \mathbb{R}$ be the functions defined by $F_i(x_j) = \delta_{ij}$ for $i, j = 1, \cdots, n$. Then any $p(x) \in \tilde{\mathcal{P}}(\mathcal{X})$ can be written as

$$p(x) = \sum_{i=1}^{n} p(x_i) F_i(x) \tag{2.43}$$

Define n coordinates $\theta^i = L_\alpha(p(x_i))$. Then

$$L_\alpha(p(x)) = \sum_{i=1}^{n} \theta^i F_i(x) \tag{2.44}$$

Therefore $\tilde{\mathcal{P}}(\mathcal{X})$ is an α-affine manifold for any α. □

Remark 2.2.12. *Note that for any $\alpha \in \mathbb{R}$, $\mathcal{P}(\mathcal{X})$ is an α-family since its denormalization $\tilde{\mathcal{P}}(\mathcal{X})$ is an α-affine manifold for any α.*

For computational purpose let us restrict ourselves to a finite set $\mathcal{X}$ with three elements. Let $\mathcal{X} = \{x_1, x_2, x_3\}$. Then $\tilde{\mathcal{P}}(\mathcal{X})$ can be identified with the first octant $\{(m_1, m_2, m_3) \ / \ m_i > 0, \ \forall \ i = 1, 2, 3 \ \}$ of $\mathbb{R}^3$. Hence $\tilde{\mathcal{P}}(\mathcal{X})$ is a 3-dimensional manifold with global coordinates (m_1, m_2, m_3).

α-Geometry on $\tilde{\mathcal{P}}(\mathcal{X})$ and $\mathcal{P}(\mathcal{X})$

Since $\tilde{\mathcal{P}}(\mathcal{X})$ is an α-affine manifold for any α, any measure $m(x) \in \tilde{\mathcal{P}}(\mathcal{X})$ can be expressed as

$$L_\alpha(m(x)) = \sum_{i=1}^{3} u_i F_i(x) \tag{2.45}$$

where $F_i(x_j) = \delta_{ij}$ for $i, j = 1, 2, 3$ and $u_i = L_\alpha(m(x_i))$ are α-affine coordinates.

Now we calculate the Fisher information metric and the α-connection on $\tilde{\mathcal{P}}(\mathcal{X})$ and $\mathcal{P}(\mathcal{X})$.

Case 1: $\alpha = 1$

Consider $\alpha = 1$ embedding which takes $m \longmapsto \log m, \ \forall \ m \in \tilde{\mathcal{P}}(\mathcal{X})$. Since $\tilde{\mathcal{P}}(\mathcal{X})$ is a 1-affine manifold,

$$L_1(m(x)) = \log m(x) = \sum_{i=1}^{3} u_i F_i(x) \tag{2.46}$$

where $u_i = \log m(x_i)$.
Let $\partial_i = \frac{\partial}{\partial u_i}$ and $\partial_{ij} = \frac{\partial^2}{\partial u_i \partial u_j}$. The 1-connection on $\tilde{\mathcal{P}}(\mathcal{X})$

$$\Gamma^1_{ijk} = \sum_{\mathcal{X}} \partial_i \partial_j L_1(m(x)) \partial_k L_{-1}(m(x)), \quad i, j, k = 1, 2, 3 \tag{2.47}$$

From Equation (3.70) it follows that $\tilde{\mathcal{P}}(\mathcal{X})$ is flat with respect to ∇^1-connection, that is $\Gamma^1_{ijk} = 0$ and the coordinate (u_i) is 1-affine.
Denote the components of the Fisher information metric $\tilde{g}$ on $\tilde{\mathcal{P}}(\mathcal{X})$ with respect to the coordinates u_i by $\tilde{g_{ij}}$.

$$\tilde{g_{ij}} = \sum_{\mathcal{X}} \partial_i L_1(m(x)) \partial_j L_1(m(x)) m(x) \tag{2.48}$$

$$= \sum_{\mathcal{X}} F_i(x) F_j(x) m(x) \tag{2.49}$$

Thus

$$\tilde{g_{ij}} = \begin{cases} \exp(u_i), & i = j \\ 0, & i \neq j \end{cases} \tag{2.50}$$

The Fisher information matrix for $\tilde{\mathcal{P}}(\mathcal{X})$ is

$$\tilde{G} = \begin{bmatrix} \exp(u_1) & 0 & 0 \\ 0 & \exp(u_2) & 0 \\ 0 & 0 & \exp(u_3) \end{bmatrix}$$

Remark 2.2.13. *It is easy to see that this metric can be suitably transformed to the Euclidean metric via a coordinate transformation $u_i \longmapsto 2\exp(u_i/2)$. Then we get $\tilde{g_{ij}} = \delta_{ij}$.*

Now consider $\mathcal{P}(\mathcal{X})$. We can identify
$\mathcal{P}(\mathcal{X}) \sim \{(u_1, u_2, u_3) \ / \ \exp(u_1) + \exp(u_2) + \exp(u_3) = 1\}$.

Then for $p(x) \in \mathcal{P}(\mathcal{X})$

$$\log p(x) = u_1 F_1(x) + u_2 F_2(x) + u_3 F_3(x) \tag{2.51}$$

where $u_3 = \log(1 - \exp(u_1) - \exp(u_2))$.

Take $v_1 = u_1, v_2 = u_2$. Note that (v_1, v_2) is a coordinate system for $\mathcal{P}(\mathcal{X})$ and hence $\mathcal{P}(\mathcal{X})$ is a two dimensional submanifold of $\tilde{\mathcal{P}}(\mathcal{X})$. Hence

$$\log p(x) = v_1 F_1(x) + v_2 F_2(x) + \log(1 - \exp(v_1) - \exp(v_2)) F_3(x). \tag{2.52}$$

Denote the components of the Fisher information metric on $\mathcal{P}(\mathcal{X})$ by g_{ij}. Let $\partial_i = \frac{\partial}{\partial v_i}$.

$$g_{ij} = \sum_{\mathcal{X}} \partial_i \log p(x) \partial_j \log p(x) p(x). \tag{2.53}$$

Let $w = 1 - \exp(v_1) - \exp(v_2)$, then

$$\partial_1 \log p(x) = F_1(x) - \frac{\exp(v_1) F_3(x)}{w} \tag{2.54}$$

$$\partial_2 \log p(x) = F_2(x) - \frac{\exp(v_2) F_3(x)}{w}. \tag{2.55}$$

From Equation (2.53)

$$g_{11} = \frac{\exp(v_1)(1 - \exp(v_2))}{w} \quad ; \quad g_{22} = \frac{\exp(v_2)(1 - \exp(v_1))}{w}. \tag{2.56}$$

$$g_{12} = g_{21} = \frac{\exp(v_1 + v_2)}{w}. \tag{2.57}$$

The Fisher information matrix G for $\mathcal{P}(\mathcal{X})$ is

$$G = \begin{bmatrix} \frac{\exp(v_1)(1-\exp(v_2))}{w} & \frac{\exp(v_1+v_2)}{w} \\ \frac{\exp(v_1+v_2)}{w} & \frac{\exp(v_2)(1-\exp(v_1))}{w} \end{bmatrix} \tag{2.58}$$

Let $\partial_{ij} = \frac{\partial^2}{\partial v_i \partial v_j}$. Then the 1-connection on $\mathcal{P}(\mathcal{X})$ is

$$\begin{aligned} \Gamma^1_{ijk} &= \sum_{\mathcal{X}} \partial_{ij} L_1(p(x)) \partial_k L_{-1}(p(x)) && (2.59) \\ &= \sum_{\mathcal{X}} \partial_{ij} \log(p(x)) \partial_k (p(x)) && (2.60) \end{aligned}$$

where $i, j, k = 1, 2$.

$$\begin{aligned} \partial_{11} \log p(x) &= -F_3(x) \left[\frac{\exp(2v_1)}{w^2} + \frac{\exp(v_1)}{w} \right] && (2.61) \\ \partial_{22} \log p(x) &= -F_3(x) \left[\frac{\exp(2v_2)}{w^2} + \frac{\exp(v_2)}{w} \right] && (2.62) \\ \partial_{12} \log p(x) &= \partial_{21} \log p(x) = -F_3(x) \frac{\exp(v_1 + v_2)}{w^2}. && (2.63) \end{aligned}$$

Also

$$\begin{aligned} \partial_1 p(x) &= p(x) \left[F_1(x) - \frac{\exp(v_1) F_3(x)}{w} \right] && (2.64) \\ \partial_2 p(x) &= p(x) \left[F_2(x) - \frac{\exp(v_2) F_3(x)}{w} \right]. && (2.65) \end{aligned}$$

Thus we get the components of 1-connection as

$$\Gamma^1_{111} = \left[\frac{\exp(3v_1)}{w^2} + \frac{\exp(2v_1)}{w} \right] \quad ; \quad \Gamma^1_{222} = \left[\frac{\exp(3v_2)}{w^2} + \frac{\exp(2v_2)}{w} \right] \qquad (2.66)$$

$$\begin{aligned} \Gamma^1_{112} &= \left[\frac{\exp(2v_1 + v_2)}{w^2} + \frac{\exp(2v_1 + v_2)}{w} \right] && (2.67) \\ \Gamma^1_{221} &= \left[\frac{\exp(2v_2 + v_1)}{w^2} + \frac{\exp(2v_1 + v_2)}{w} \right] && (2.68) \end{aligned}$$

$$\Gamma^1_{121} = \Gamma^1_{211} = \frac{\exp(2v_1 + v_2)}{w^2} \quad ; \quad \Gamma^1_{122} = \Gamma^1_{212} = \frac{\exp(2v_2 + v_1)}{w^2}. \qquad (2.69)$$

For any $\alpha \neq 1$

Consider the α-embedding which takes $m \longmapsto \frac{2}{1-\alpha} m^{\frac{1-\alpha}{2}},\ \forall\ m \in \tilde{\mathcal{P}}(X)$. Since $\tilde{\mathcal{P}}(\mathcal{X})$ is an α-affine manifold, for any measure $m \in \tilde{\mathcal{P}}(\mathcal{X})$

$$L_\alpha(m(x)) = \frac{2}{1-\alpha} m^{\frac{1-\alpha}{2}} = \sum_{i=1}^{3} u_i F_i(x). \tag{2.70}$$

where $u_i = \frac{2}{1-\alpha} m_i^{\frac{1-\alpha}{2}}$.
Let $\partial_i = \frac{\partial}{\partial u_i}$ and $\partial_{ij} = \frac{\partial^2}{\partial u_i \partial u_j}$. Then the α-connection on $\tilde{\mathcal{P}}(\mathcal{X})$ is

$$\Gamma^\alpha_{ijk} = \sum_{\mathcal{X}} \partial_i \partial_j L_\alpha(m(x)) \partial_k L_{-\alpha}(m(x)), \quad i, j, k = 1, 2, 3. \tag{2.71}$$

From Equation (2.70) it follows that $\tilde{\mathcal{P}}(\mathcal{X})$ is flat with respect to ∇^α-connection, that is $\Gamma^\alpha_{ijk} = 0$ and the coordinate (u_i) is α-affine.
Denote the components of the Fisher information metric $\tilde{g}$ on $\tilde{\mathcal{P}}(\mathcal{X})$ with respect to the coordinates u_i by $\tilde{g_{ij}}$.

$$\begin{aligned} \tilde{g_{ij}} &= \sum_{\mathcal{X}} \partial_i L_\alpha(m(x)) \partial_j L_\alpha(m(x)) m(x)^\alpha & (2.72) \\ &= \sum_{\mathcal{X}} F_i(x) F_j(x) m(x)^\alpha. & (2.73) \end{aligned}$$

Thus

$$\tilde{g_{ij}} = \begin{cases} (\frac{1-\alpha}{2} u_i)^{\frac{2\alpha}{1-\alpha}}, & i = j \\ 0, & i \neq j \end{cases} \tag{2.74}$$

The Fisher information matrix for $\tilde{\mathcal{P}}(\mathcal{X})$ is

$$\tilde{G} = \begin{bmatrix} (\frac{1-\alpha}{2} u_1)^{\frac{2\alpha}{1-\alpha}} & 0 & 0 \\ 0 & (\frac{1-\alpha}{2} u_2)^{\frac{2\alpha}{1-\alpha}} & 0 \\ 0 & 0 & (\frac{1-\alpha}{2} u_3)^{\frac{2\alpha}{1-\alpha}} \end{bmatrix} \tag{2.75}$$

Remark 2.2.14. *This metric can be transformed to Euclidean metric via the coordinate transformation* $u_i \longmapsto 2(\frac{1-\alpha}{2} u_i)^{\frac{1}{1-\alpha}}$. *Then we get* $\tilde{g_{ij}} = \delta_{ij}$.

Now consider $\mathcal{P}(\mathcal{X})$. Under the α-embedding, we can identify
$\mathcal{P}(\mathcal{X}) \sim \{(u_1, u_2, u_3)\ /\ u_1^{\frac{2}{1-\alpha}} + u_2^{\frac{2}{1-\alpha}} + u_3^{\frac{2}{1-\alpha}} = (\frac{1-\alpha}{2})^{\frac{2}{\alpha-1}}\}$.

For any $p(x) \in \mathcal{P}(\mathcal{X})$,

$$L_\alpha(p(x)) = \frac{2}{1-\alpha} m(x)^{\frac{1-\alpha}{2}} = u_1 F_1(x) + u_2 F_2(x) + u_3 F_3(x) \tag{2.76}$$

where $u_3 = \left[(\frac{1-\alpha}{2})^{\frac{2}{\alpha-1}} - u_1^{\frac{2}{1-\alpha}} - u_2^{\frac{2}{1-\alpha}} \right]^{\frac{1-\alpha}{2}}$.

Take $v_1 = u_1, v_2 = u_2$. Then $\mathcal{P}(\mathcal{X})$ is a two dimensional submanifold of $\tilde{\mathcal{P}}(\mathcal{X})$ with a coordinate system (v_1, v_2). We have

$$L_\alpha(p(x)) = v_1 F_1(x) + v_2 F_2(x) + \left[(\frac{1-\alpha}{2})^{\frac{2}{\alpha-1}} - v_1^{\frac{2}{1-\alpha}} - v_2^{\frac{2}{1-\alpha}} \right]^{\frac{1-\alpha}{2}} F_3(x). \tag{2.77}$$

Denote the components of the Fisher information metric on $\mathcal{P}(\mathcal{X})$ with respect to (v_1, v_2) by g_{ij}. Denote $\partial_i = \frac{\partial}{\partial v_i}$.

$$g_{ij} = \sum_{\mathcal{X}} \partial_i L_\alpha(p(x)) \partial_j L_\alpha(p(x)) p(x)^\alpha. \tag{2.78}$$

Let $w = (\frac{1-\alpha}{2})^{\frac{2}{\alpha-1}} - u_1^{\frac{2}{1-\alpha}} - u_2^{\frac{2}{1-\alpha}}$. Then

$$\begin{aligned} \partial_1 L_\alpha(p(x)) &= F_1(x) - F_3(x)\, v_1^{\frac{1+\alpha}{1-\alpha}} w^{-(\frac{1+\alpha}{2})} && (2.79) \\ \partial_2 L_\alpha(p(x)) &= F_2(x) - F_3(x)\, v_2^{\frac{1+\alpha}{1-\alpha}} w^{-(\frac{1+\alpha}{2})} && (2.80) \end{aligned}$$

$$\begin{aligned} g_{11} &= w^{-1}\, v_1^{\frac{2\alpha}{1-\alpha}} \left((\frac{1-\alpha}{2})^{-2} - (\frac{1-\alpha}{2})^{\frac{2\alpha}{1-\alpha}} v_2^{\frac{2}{1-\alpha}} \right) && (2.81) \\ g_{22} &= w^{-1}\, v_2^{\frac{2\alpha}{1-\alpha}} \left((\frac{1-\alpha}{2})^{-2} - (\frac{1-\alpha}{2})^{\frac{2\alpha}{1-\alpha}} v_1^{\frac{2}{1-\alpha}} \right) && (2.82) \\ g_{12} &= g_{21} = w^{-1} (\frac{1-\alpha}{2})^{\frac{2\alpha}{1-\alpha}}\, v_1^{\frac{1+\alpha}{1-\alpha}}\, v_2^{\frac{1+\alpha}{1-\alpha}} && (2.83) \end{aligned}$$

Let $\partial_{ij} = \frac{\partial^2}{\partial v_i \partial v_j}$. Then the α-connection on $\mathcal{P}(\mathcal{X})$ is

$$\Gamma^\alpha_{ijk} = \sum_{\mathcal{X}} \partial_{ij} L_\alpha(p(x)) \partial_k L_{-\alpha}(p(x)) \tag{2.84}$$

where $i, j, k = 1, 2$.

$$\partial_{11} L_\alpha(p(x)) = \frac{1+\alpha}{\alpha-1} \, w^{-\frac{(\alpha+3)}{2}} \, v_1^{\frac{2(1+\alpha)}{1-\alpha}} \, F_3(x) \tag{2.85}$$

$$\partial_{22} L_\alpha(p(x)) = \frac{1+\alpha}{\alpha-1} \, w^{-\frac{(\alpha+3)}{2}} \, v_2^{\frac{2(1+\alpha)}{1-\alpha}} \, F_3(x) \tag{2.86}$$

$$\partial_{12} L_\alpha(p(x)) = \partial_{21} L_\alpha(p(x)) \tag{2.87}$$

$$= \frac{1+\alpha}{\alpha-1} \, w^{-\frac{(\alpha+3)}{2}} \, v_2^{\frac{2(1+\alpha)}{1-\alpha}} \, F_3(x) \tag{2.88}$$

Also

$$\partial_1 L_{-\alpha}(p(x)) = p(x)^\alpha \left[F_1(x) - F_3(x) \, w^{-\frac{(\alpha+1)}{2}} \, v_1^{\frac{1+\alpha}{1-\alpha}} \right] \tag{2.89}$$

$$\partial_2 L_{-\alpha}(p(x)) = p(x)^\alpha \left[F_2(x) - F_3(x) \, w^{-\frac{(\alpha+1)}{2}} \, v_2^{\frac{1+\alpha}{1-\alpha}} \right] \tag{2.90}$$

Thus we get the components of α-connection as

$$\Gamma^\alpha_{111} = \frac{1+\alpha}{2} \, (\frac{1-\alpha}{2})^{\frac{3\alpha-1}{1-\alpha}} \, w^{-2} \, v_1^{\frac{3(1+\alpha)}{1-\alpha}} \tag{2.91}$$

$$\Gamma^\alpha_{222} = \frac{1+\alpha}{2} \, (\frac{1-\alpha}{2})^{\frac{3\alpha-1}{1-\alpha}} \, w^{-2} \, v_2^{\frac{3(1+\alpha)}{1-\alpha}} \tag{2.92}$$

$$\Gamma^\alpha_{112} = \frac{1+\alpha}{2} \, (\frac{1-\alpha}{2})^{\frac{3\alpha-1}{1-\alpha}} \, w^{-2} \, v_1^{\frac{2(1+\alpha)}{1-\alpha}} \, v_2^{\frac{1+\alpha}{1-\alpha}} \tag{2.93}$$

$$\Gamma^\alpha_{221} = \frac{1+\alpha}{2} \, (\frac{1-\alpha}{2})^{\frac{3\alpha-1}{1-\alpha}} \, w^{-2} \, v_2^{\frac{2(1+\alpha)}{1-\alpha}} \, v_1^{\frac{1+\alpha}{1-\alpha}} \tag{2.94}$$

and

$$\Gamma^\alpha_{211} = \Gamma^\alpha_{121} = \Gamma^\alpha_{112} \tag{2.95}$$

$$\Gamma^\alpha_{212} = \Gamma^\alpha_{122} = \Gamma^\alpha_{221} \tag{2.96}$$

2.3 (F, G)-Geometry

Amari [12] defined the α-geometry using a particular family of functions called the α-embedding. We considered a general embedding function F instead of the α-embedding and also a positive smooth function G to obtain more general geometric structures on a statistical manifold called the (F, G)-geometry [16]. The α-geometry is a special case of the (F, G)-geometry. Now we describe the (F, G)-geometric structure in detail.

Let $F : (0, \infty) \longrightarrow \mathbb{R}$ be a function which is atleast twice differentiable. Assume that $F'(u) \neq 0 \ \forall \ u \in (0, \infty)$. Then F is an embedding of $\mathcal{S}$ into $\mathbb{R}_{\mathcal{X}}$ which takes each $p(x; \theta) \longmapsto F(p(x; \theta))$. Denote $F(p(x; \theta))$ by $F(x; \theta)$. $\partial_i F(x; \theta)$ can be written as

$$\partial_i F(x; \theta) = p(x; \theta) F'(p(x; \theta)) \partial_i \ell(x; \theta) \tag{2.97}$$

It is clear that for every θ, the set of n functions $\{\partial_i F(x; \theta), \quad i = 1, \cdots, n\}$ in x is linearly independent since $\{\partial_i \ell(x; \theta), \ i = 1, \cdots, n\}$ is linearly independent.
Let $T_{F(p_\theta)} F(\mathcal{S})$ be the n-dimensional vector space spanned by $\{\partial_i F(x; \theta), \ i = 1, \cdots, n\}$.

$$T_{F(p_\theta)} F(\mathcal{S}) = \{A^F(x) \ / \ A^F(x) = \sum_{i=1}^{n} A^i \partial_i F(x; \theta), \ A^i \in \mathbb{R}\}. \tag{2.98}$$

Let the tangent space $T_{F(p_\theta)}(F(\mathcal{S}))$ to $F(\mathcal{S})$ at the point $F(p_\theta)$ be denoted by $T_\theta^F(\mathcal{S})$. There is a natural isomorphism between the two vector spaces $T_\theta(\mathcal{S})$ and $T_\theta^F(\mathcal{S})$ given by

$$\partial_i \in T_\theta(\mathcal{S}) \longleftrightarrow \partial_i F(x; \theta) \in T_\theta^F(\mathcal{S}). \tag{2.99}$$

The vector space $T_\theta^F(\mathcal{S})$ is called the **F-representation** of the tangent space $T_\theta(\mathcal{S})$. The F-representation of the tangent vector $A = \sum_{i=1}^{n} A^i \partial_i \in T_\theta(\mathcal{S})$ is the random variable

$$A^F(x) = \sum_{i=1}^{n} A^i \partial_i F \in T_\theta^F(\mathcal{S}). \tag{2.100}$$

Remark 2.3.1. *Burbea [15] introduced the concept of weighted Fisher information metric which is a generalized notion of Fisher information metric. He used a positive continuous function to define the weighted metric. We consider a positive smooth function G together with an embedding function F to define more general geometric structures on a statistical manifold called the (F, G)-geometry.*

Definition 2.3.2. *Let $G : (0, \infty) \longrightarrow \mathbb{R}$ be a positive smooth function and F be the embedding function. Then the (F, G)-expectation of a random variable f with respect to the distribution $p(x; \theta)$ is defined as*

$$E_\theta^{F,G}(f) = \int f(x) \frac{1}{p(F'(p))^2} G(p) \, dx. \tag{2.101}$$

(here we assume that the above integral exists.)

We can use this (F, G)-expectation to define an inner product in $\mathbb{R}_{\mathcal{X}}$ by

$$< f, g >_{\theta}^{F,G} = E_{\theta}^{F,G}[f(x)g(x)]. \tag{2.102}$$

which induces a Riemannian metric on $\mathcal{S}$ given by

$$< A, B >_{\theta}^{F,G} = E_{\theta}^{F,G}[A^F(x)B^F(x)], \; quad A, B \in T_{\theta}(\mathcal{S}). \tag{2.103}$$

In terms of the basis vectors

$$\begin{aligned} < \partial_i, \partial_j >_{\theta}^{F,G} &= \int \partial_i F \; \partial_j F \frac{G(p)}{p(F'(p))^2} dx && (2.104) \\ &= \int \partial_i \ell \; \partial_j \ell \; G(p) \; p \; dx && (2.105) \end{aligned}$$

Since this metric do not depend on F, let us call this metric as G**-metric**. Denote it by $g^G =<, >^G$ and its components by g_{ij}^G.

$$g_{ij}^G(\theta) =< \partial_i, \partial_j >_{\theta}^G = \int \partial_i \ell \; \partial_j \ell \; G(p) \; p \; dx. \tag{2.106}$$

The matrix $[g_{ij}^G(\theta)]$ is called the G**-matrix**.

Definition 2.3.3. *Let $\pi_{|p_\theta}^{F,G} : \mathbb{R}_{\mathcal{X}} \longrightarrow T_{\theta}^F(\mathcal{S})$ be the projection map. The affine connection induced by this map on $\mathcal{S}$, the (F, G)**-connection** $\nabla^{F,G}$, is defined as*

$$\begin{aligned} \nabla_{\partial_i}^{F,G} \partial_j &= \pi_{|p_\theta}^{F,G}(\frac{\partial^2 F}{\partial \theta^i \partial \theta^j}) && (2.107) \\ &= \sum_n \sum_m g^{G(mn)} < \frac{\partial^2 F}{\partial \theta^i \partial \theta^j}, \frac{\partial F}{\partial \theta^m} >_{\theta}^{F,G} \partial_n && (2.108) \end{aligned}$$

where $[g^{G(mn)}(\theta)]$ is the inverse of the G-matrix $[g_{mn}^G(\theta)]$.
Note that the (F, G)-connections are symmetric.

Lemma 2.3.4. *The (F, G)-connection and its components can be written in terms of scores as*

$$\nabla_{\partial_i}^{F,G} \partial_j = \sum_n \sum_m g^{G(mn)} E_{\theta} \left[\left(\partial_i \partial_j \ell + (1 + \frac{pF''(p)}{F'(p)}) \partial_i \ell \; \partial_j \ell \right) \partial_m \ell \; G(p) \right] \partial_n \tag{2.109}$$

and

$$\Gamma^{F,G}_{ijk}(\theta) = \int \left(\partial_i\partial_j\ell + (1+\frac{pF''(p)}{F'(p)})\partial_i\ell\,\partial_j\ell\right)\partial_k\ell\,G(p)\,p\,dx. \tag{2.110}$$

Proof. From Equation (2.97),

$$\partial_i\partial_j F = pF'(p)\partial_i\partial_j\ell + [pF'(p) + p^2F''(p)]\,\partial_i\ell\,\partial_j\ell. \tag{2.111}$$

Therefore

$$\begin{aligned} <\partial_i\partial_j F, \partial_m F>^{F,G}_\theta &= \int \partial_i\partial_j F\,\partial_m F \frac{G(p)}{p(F'(p))^2}dx & (2.112)\\ &= \int [pF'(p) + p^2F''(p)]\,\partial_i\ell\,\partial_j\ell\,\partial_m\ell\,\frac{G(p)}{F'(p)}dx \\ &\quad + \int \partial_i\partial_j\ell\,G(p)\,p\,dx & (2.113)\\ &= E_\theta\left[(\partial_i\partial_j\ell + (1+\frac{pF''(p)}{F'(p)})\partial_i\ell\,\partial_j\ell)\partial_m\ell\,G(p)\right] & (2.114)\end{aligned}$$

Hence

$$\begin{aligned} \nabla^{F,G}_{\partial_i}\partial_j &= \pi^{F,G}_{|p_\theta}(\partial_i\partial_j F) & (2.115)\\ &= \sum_n\sum_m g^{G(mn)} E_\theta\left[\left(\partial_i\partial_j\ell + (1+\frac{pF''(p)}{F'(p)})\partial_i\ell\,\partial_j\ell\right)\partial_m\ell\,G(p)\right]\partial_n & (2.116)\end{aligned}$$

Then the Christoffel symbols of the (F,G)-connection are

$$\Gamma^n_{ij} = \sum_m g^{G(mn)} E_\theta\left[(\partial_i\partial_j\ell + (1+\frac{pF''(p)}{F'(p)})\partial_i\ell\,\partial_j\ell)\partial_m\ell\,G(p)\right] \tag{2.117}$$

and components of the (F,G)-connection are

$$\begin{aligned} \Gamma^{F,G}_{ijk}(\theta) &= <\nabla^{F,G}_{\partial_i}\partial_j, \partial_k>^G_\theta & (2.118)\\ &= \int\left(\partial_i\partial_j\ell + (1+\frac{pF''(p)}{F'(p)})\partial_i\ell\,\partial_j\ell\right)\partial_k\ell\,G(p)\,p\,dx & (2.119)\end{aligned}$$

□

Theorem 2.3.5. *Let F and H be two embeddings of $\mathcal{S}$ into $\mathbb{R}_\mathcal{X}$ and G be a positive smooth function on $(0,\infty)$. Then the (F,G)-connection $\nabla^{F,G}$ and the (H,G)-*

connection $\nabla^{H,G}$ are dual connections with respect to the G-metric iff the functions F and H satisfy

$$H'(p) = \frac{G(p)}{pF'(p)}. \tag{2.120}$$

*We call such an embedding H as a G-**dual embedding** of F.*

The components of the dual connection $\nabla^{H,G}$ can be written as

$$\begin{aligned}\Gamma^{H,G}_{ijk}(\theta) &= \int \left(\partial_i\partial_j\ell + (1+\frac{pH''(p)}{H'(p)})\partial_i\ell\;\partial_j\ell\right)\partial_k\ell\;G(p)\;p\;dx & (2.121)\\ &= \int \left(\partial_i\partial_j\ell + (\frac{pG'(p)}{G(p)} - \frac{pF''(p)}{F'(p)})\partial_i\ell\;\partial_j\ell\right)\partial_k\ell\;G(p)\;p\;dx. & (2.122)\end{aligned}$$

Proof. $\nabla^{F,G}$ and $\nabla^{H,G}$ are dual connections with respect to the G-metric means

$$\partial_k < \partial_i, \partial_j >^G = < \nabla^{F,G}_{\partial_k}\partial_i, \partial_j >^G + < \partial_i, \nabla^{H,G}_{\partial_k}\partial_j >^G. \tag{2.123}$$

for any basis vectors $\partial_i, \partial_j, \partial_k \in T_\theta(\mathcal{S})$.

$$\begin{aligned}\partial_k < \partial_i, \partial_j >^G &= \int \partial_k\partial_j\ell\;\partial_i\ell\;pG(p)dx + \int \partial_k\partial_i\ell\;\partial_j\ell\;pG(p)dx\\ &+ \int (1+\frac{pG'(p)}{G(p)})\partial_i\ell\;\partial_j\ell\;\partial_k\ell\;pG(p)dx. & (2.124)\end{aligned}$$

$$\begin{aligned}< \nabla^{F,G}_{\partial_k}\partial_i, \partial_j >^G + < \partial_i, \nabla^{H,G}_{\partial_k}\partial_j >^G &= \int \partial_k\partial_i\ell\;\partial_j\ell\;pG(p)dx\\ &+ \int 1+\frac{pF''(p)}{F'(p)}\partial_i\ell\;\partial_j\ell\;\partial_k\ell\;pG(p)dx\\ &+ \int 1+\frac{pH''(p)}{H'(p)}\partial_i\ell\;\partial_j\ell\;\partial_k\ell\;pG(p)dx\\ &+ \int \partial_k\partial_j\ell\;\partial_i\ell\;pG(p)dx & (2.125)\end{aligned}$$

Then Equation (2.123) holds iff

$$\begin{aligned}&\int [2+\frac{pF''(p)}{F'(p)} + \frac{pH''(p)}{H'(p)}]\partial_i\ell\;\partial_j\ell\;\partial_k\ell\;pG(p)dx =\\ &\qquad \int [1+\frac{pG'(p)}{G(p)}]\partial_i\ell\;\partial_j\ell\;\partial_k\ell\;pG(p)dx & (2.126)\end{aligned}$$

$$\iff [2+\frac{pF''(p)}{F'(p)} + \frac{pH''(p)}{H'(p)}] = 1+\frac{pG'(p)}{G(p)}. \tag{2.127}$$

$$\iff 1+\frac{pH''(p)}{H'(p)} = \frac{pG'(p)}{G(p)} - \frac{pF''(p)}{F'(p)} \tag{2.128}$$

$$\iff \frac{H''(p)}{H'(p)} = \frac{G'(p)}{G(p)} - \frac{F''(p)}{F'(p)} - \frac{1}{p} \iff H'(p) = \frac{G(p)}{pF'(p)}. \tag{2.129}$$

Hence $\nabla^{F,G}$ and $\nabla^{H,G}$ are dual connections with respect to the G-metric iff Equation (2.120) holds.

From Equation (2.128) we can rewrite the components of dual connection $\nabla^{H,G}$ as

$$\begin{aligned}
\Gamma_{ijk}^{H,G}(\theta) &= \int \left(\partial_i\partial_j\ell + (1+\frac{pH''(p)}{H'(p)})\partial_i\ell\;\partial_j\ell\right)\partial_k\ell\; G(p)\;p\;dx \qquad (2.130)\\
&= \int \left(\partial_i\partial_j\ell + (\frac{pG'(p)}{G(p)} - \frac{pF''(p)}{F'(p)})\partial_i\ell\;\partial_j\ell\right)\partial_k\ell\; G(p)p\;dx. \qquad (2.131)
\end{aligned}$$

□

Theorem 2.3.6. *Amari's α-geometry is a special case of the (F,G)-geometry.*

Proof. Let $F(p) = L_\alpha(p)$, the α-embedding of Amari and $G(p) = 1$. Then

$$\begin{aligned}
F'(p) &= L'_\alpha(p) = p^{-\left(\frac{1+\alpha}{2}\right)} \qquad (2.132)\\
F''(p) &= L''_\alpha(p) = -\frac{1+\alpha}{2}p^{-\left(\frac{3+\alpha}{2}\right)} \qquad (2.133)\\
1+\frac{pF''(p)}{F'(p)} &= 1+\frac{pL''_\alpha(p)}{L'_\alpha(p)} = \frac{1-\alpha}{2} \qquad (2.134)
\end{aligned}$$

Then from Equation (2.120), the G-dual embedding of F is obtained as $H(p) = L_{-\alpha}(p)$. Also

$$1+\frac{pH''(p)}{H'(p)} = \frac{1+\alpha}{2} \tag{2.135}$$

Thus

$$\begin{aligned}
\Gamma_{ijk}^{F,G}(\theta) &= E_\theta\left[(\partial_i\partial_j\ell + (1+\frac{pF''(p)}{F'(p)})\partial_i\ell\;\partial_j\ell)\partial_k\ell\;G(p)\right] \qquad (2.136)\\
&= E_\theta\left[(\partial_i\partial_j\ell + \frac{1-\alpha}{2}\partial_i\ell\;\partial_j\ell)(\partial_k\ell)\right] \qquad (2.137)\\
&= \Gamma^\alpha_{ijk}(\theta) \qquad (2.138)
\end{aligned}$$

and

$$\begin{aligned}
\Gamma^{H,G}_{ijk}(\theta) &= E_\theta\left[(\partial_i\partial_j\ell + (1+\frac{pH''(p)}{H'(p)})\partial_i\ell\,\partial_j\ell)\partial_k\ell\,G(p)\right] && (2.139)\\
&= E_\theta\left[(\partial_i\partial_j\ell + \frac{1+\alpha}{2}\partial_i\ell\,\partial_j\ell)(\partial_k\ell)\right] && (2.140)\\
&= \Gamma^{-\alpha}_{ijk}(\theta) && (2.141)
\end{aligned}$$

Hence (F,G)-connection reduces to the α-connection and the (H,G)-connection reduces to the $(-\alpha)$-connection.

Also the G-metric is

$$\begin{aligned}
g^G_{ij}(\theta) &= \int \partial_i\ell\,\partial_j\ell\,G(p)\,p\,dx && (2.142)\\
&= \int \partial_i\ell\,\partial_j\ell\,p\,dx && (2.143)
\end{aligned}$$

which is the Fisher information metric g.

Thus the α-geometry is a special case of the (F,G)-geometry. □

Remark 2.3.7. *The Levi-Civita connection ∇^G with respect to the G-metric is given by*

$$\begin{aligned}
\Gamma^G_{ijk}(\theta) &= \frac{1}{2}\left(\partial_i g^G_{jk} + \partial_j g^G_{ki} + \partial_k g^G_{ij}\right) && (2.144)\\
&= \int \left(\partial_i\partial_j\ell + \frac{1}{2}(1+\frac{pG'(p)}{G(p)})\,\partial_i\ell\,\partial_j\ell\right)\partial_k\ell\,G(p)\,p\,dx. && (2.145)
\end{aligned}$$

∇^G is a (F,G)-connection $\nabla^{F,G}$ with the embedding function F given by

$$F'(p) = \frac{\sqrt{G(p)}}{\sqrt{p}}. \qquad (2.146)$$

When $G(p) = 1$ the connection ∇^G reduces to Amar's 0-connection $\nabla^{(0)}$, which is the Levi-Civita connection with respect to the Fisher information metric.

Example 2.3.8. Let $F(x) = x\ln x - x$ and let $G(x) = \ln x$. Then from equation (2.120) the G-dual embedding H of F is defined by

$$H'(x) = \frac{G(x)}{xF'(x)} = \frac{\ln x}{x\ln x} = \frac{1}{x}. \qquad (2.147)$$

Thus $H(x) = \ln x$. Then from Equations (2.106), (2.110), (2.122) the G-metric and dual (F,G) and (H,G)-connections are given by

$$g^G_{ij}(\theta) = \int \partial_i\ell\,\partial_j\ell\,\ln p\,p\,dx. \tag{2.148}$$

$$\Gamma^{F,G}_{ijk}(\theta) = \int \left(\partial_i\partial_j\ell + (1+\frac{1}{\ln p})\partial_i\ell\,\partial_j\ell\right)\partial_k\ell\,\ln p\,p\,dx. \tag{2.149}$$

$$\Gamma^{H,G}_{ijk}(\theta) = \int \partial_i\partial_j\ell\,\partial_k\ell\,\ln p\,p\,dx. \tag{2.150}$$

Remark 2.3.9. *Zhang [21] considered a generalized α-representation of density function called ρ-representation and defined a divergence function called the (α,ρ,τ)- divergence. Using this divergence he obtained a geometry on a statistical manifold called the (α,ρ,τ)-geometry which is a generalization of α-geometry. In Chapter 3 we will detail his work and discuss the relation between the (F,G)-geometry and the (α,ρ,τ)-geometry.*

2.3.1 F-affine manifold and F-family

Using the α-representation of density function, Amari [12] defined the notion of α-affine manifold and α-family. Using a generalized ρ-representation of density function Zhang [21] considered a ρ-affine family which is a generalization of α-affine manifold. We consider the same family using the embedding function F and compute the G-metric and the (F,G)-connections in the finite case.

Definition 2.3.10. *Let $\mathcal{S} = \{p(x;\xi)\ /\ \xi = (\xi^1,\cdots,\xi^n) \in \mathbb{E} \subseteq \mathbb{R}^n\}$ be an n-dimensional statistical manifold. If for some coordinate system $\theta = (\theta^i),\ i = 1,\cdots,n$*

$$\partial_i\partial_j F(x;\theta) = 0 \tag{2.151}$$

then from Equation (2.107) *θ is an $\nabla^{F,G}$-affine coordinate system and that $\mathcal{S} = \{p_\theta\}$ is $\nabla^{F,G}$-flat. We call such an $\mathcal{S}$ as an F**-affine manifold**.*

The above condition is equivalent to the existence of the functions $C, F_1,\cdots,F_n$ on $\mathcal{X}$ such that

$$F(x;\theta) = C(x) + \sum_{i=1}^{n}\theta^i F_i(x). \tag{2.152}$$

Definition 2.3.11. *A statistical manifold $\mathcal{S} = \{p(x;\xi)\ /\ \xi = (\xi^1,\cdots,\xi^n) \in \mathbb{E} \subseteq \mathbb{R}^n\}$ is said to an F**-family** if its denormalization $\tilde{\mathcal{S}}$ is an F-affine manifold.*

For an n-dimensional F-family $\mathcal{S}$ there exists a coordinate system $\theta = (\theta^i)$ and functions $C_1(x), \cdots, C_n(x), \psi(\theta)$ such that

$$F(x;\theta) = \sum_{i=1}^{n} \theta^i C_i(x) - \psi(\theta) \tag{2.153}$$

where $\psi(\theta)$ is obtained from the normalization condition $\int_{\mathcal{X}} p(x;\theta)dx = 1$.

Theorem 2.3.12. *For any embedding F, $\tilde{\mathcal{P}}(\mathcal{X})$ is an F-affine manifold for finite $\mathcal{X}$.*

Proof. Let $\mathcal{X} = \{x_1, \cdots, x_n\}$. Let $F_i : \mathcal{X} \longrightarrow \mathbb{R}$ be functions defined by $F_i(x_j) = \delta_{ij}$ for $i, j = 1, \cdots, n$. Then any $p(x) \in \tilde{\mathcal{P}}(\mathcal{X})$ can be written as

$$p(x) = \sum_{i=1}^{n} p(x_i) F_i(x) \tag{2.154}$$

Define n coordinates $\theta^i = F(p(x_i))$. Then

$$F(p(x)) = \sum_{i=1}^{n} \theta^i F_i(x) \tag{2.155}$$

Therefore $\tilde{\mathcal{P}}(\mathcal{X})$ is an F-affine manifold for any F. □

Remark 2.3.13. *$\mathcal{P}(\mathcal{X})$ is a F-family for any F since $\tilde{\mathcal{P}}(\mathcal{X})$ is an F-affine manifold for any F. The F-family is a generalization of the exponential family. Hence it would be appropriate to call it as F-exponential family instead of F-family. The geometry of the F-exponential family will be discussed in Chapter 4.*

(F, G)-Geometry on $\tilde{\mathcal{P}}(\mathcal{X})$ and $\mathcal{P}(\mathcal{X})$

Let $\mathcal{X} = \{x_1, x_2, x_3\}$. Since $\tilde{\mathcal{P}}(\mathcal{X})$ is an F-affine manifold for any F, any measure $m(x) \in \tilde{\mathcal{P}}(\mathcal{X})$ can be expressed as

$$F(m(x)) = \sum_{i=1}^{3} u_i F_i(x) \tag{2.156}$$

where $F_i(x_j) = \delta_{ij}$ for $i, j = 1, 2, 3$ and $u_i = F(m(x_i)) = F(m_i)$.
Let Z be the inverse function of F and then $m_i = Z(u_i)$.

Let $\partial_i = \frac{\partial}{\partial u_i}$ and $\partial_{ij} = \frac{\partial^2}{\partial u_i \partial u_j}$. The (F, G)-connection on $\tilde{\mathcal{P}}(\mathcal{X})$ is

$$\Gamma_{ijk}^{F,G} = \sum_{\mathcal{X}} \partial_i \partial_j F(m(x))\, \partial_k H(m(x)), \quad i, j, k = 1, 2, 3 \tag{2.157}$$

where H is the G-dual embedding of F.

From Equation (2.156) it follows that $\tilde{\mathcal{P}}(\mathcal{X})$ is flat with respect to $\nabla^{F,G}$-connection. Thai is, $\Gamma_{ijk}^{F,G} = 0$ and the coordinate (u_i) is $\nabla^{F,G}$-affine.

Denote the components of the G-metric $\tilde{g}^G$ on $\tilde{\mathcal{P}}(\mathcal{X})$ with respect to the coordinates u_i by $\tilde{g}_{ij}^G$.

$$\begin{aligned} \tilde{g}_{ij}^G &= \sum_{\mathcal{X}} \partial_i F(m(x))\, \partial_j H(m(x)) && (2.158) \\ &= \sum_{\mathcal{X}} \frac{G(m(x))}{m(x)(F'(m(x)))^2}\, F_i(x)\, F_j(x) && (2.159) \end{aligned}$$

$$\tilde{g}_{ij}^G = \begin{cases} \frac{G(m_i)}{m_i(F'(m_i))^2} & i = j \quad ; \quad \text{where } m_i = Z(u_i) \\ 0 & i \neq j \end{cases} \tag{2.160}$$

Now consider $\mathcal{P}(\mathcal{X})$. We can identify $\mathcal{P}(\mathcal{X}) \sim \{(u_1, u_2, u_3) \;/\; Z(u_1) + Z(u_2) + Z(u_3) = 1\}$. Then for $p(x) \in \mathcal{P}(\mathcal{X})$,

$$F(p(x)) = u_1 F_1(x) + u_2 F_2(x) + u_3 F_3(x) \tag{2.161}$$

where $u_3 = F(1 - Z(u_1) - Z(u_2))$.

Take $v_1 = u_1, v_2 = u_2$. Note that (v_1, v_2) is a coordinate system for $\mathcal{P}(\mathcal{X})$ and hence $\mathcal{P}(\mathcal{X})$ is a two dimensional submanifold of $\tilde{\mathcal{P}}(\mathcal{X})$. Hence

$$F(p(x)) = v_1 F_1(x) + v_2 F_2(x) + F(1 - Z(v_1) - Z(v_2)) F_3(x) \tag{2.162}$$

Denote the components of the G-metric on $\mathcal{P}(\mathcal{X})$ by g_{ij}^G. Let $\partial_i = \frac{\partial}{\partial v_i}$.

$$g_{ij}^G = \sum_{\mathcal{X}} \partial_i F(p(x))\, \partial_j H(p(x)) \tag{2.163}$$

Let $w = 1 - Z(v_1) - Z(v_2)$. Then

$$\begin{aligned} \partial_1 F(p(x)) &= F_1(x) - F_3(x)\ F'(w)\ Z'(v_1) && (2.164) \\ \partial_2 F(p(x)) &= F_2(x) - F_3(x)\ F'(w)\ Z'(v_2) && (2.165) \end{aligned}$$

and

$$\begin{aligned} \partial_1 H(p(x)) &= H'(p(x))\ Z'(F(p(x)))\,[F_1(x) - F_3(x)\ F'(w)\ Z'(v_1)] && (2.166) \\ \partial_2 H(p(x)) &= H'(p(x))\ Z'(F(p(x)))\,[F_2(x) - F_3(x)\ F'(w)\ Z'(v_2)] && (2.167) \end{aligned}$$

Then

$$\begin{aligned} g_{11}^G &= H'(Z(v_1))\ Z'(v_1) + \frac{G(w)}{w}(Z'(v_1))^2 && (2.168) \\ g_{22}^G &= H'(Z(v_2))\ Z'(v_2) + \frac{G(w)}{w}(Z'(v_2))^2 && (2.169) \\ g_{12}^G &= g_{21}^G = \frac{G(w)}{w} Z'(v_1)\ Z'(v_2) && (2.170) \end{aligned}$$

Let $\partial_{ij} = \frac{\partial^2}{\partial v_i \partial v_j}$. Then the (F,G)-connection on $\mathcal{P}(\mathcal{X})$ is

$$\Gamma_{ijk}^{F,G} = \sum_{\mathcal{X}} \partial_{ij} F(p(x))\ \partial_k H(p(x)) \qquad (2.171)$$

where $i, j, k = 1, 2$.

$$\begin{aligned} \partial_{11} F(p(x)) &= -F_3(x)\left[F'(w)\ Z''(v_1) - F''(w)\ (Z'(v_1))^2\right] && (2.172) \\ \partial_{22} F(p(x)) &= -F_3(x)\left[F'(w)\ Z''(v_2) - F''(w)\ (Z'(v_2))^2\right] && (2.173) \\ \partial_{12} F(p(x)) &= \partial_{21} F(p(x)) = F_3(x)\ F''(w)\ Z'(v_1)\ Z'(v_2) && (2.174) \end{aligned}$$

Thus we get the components of (F,G)-connection as

$$\begin{aligned} \Gamma_{111}^{F,G} &= \frac{G(w)}{w} Z''(v_1)\ Z'(v_1) - \frac{G(w)F''(w)}{wF'(w)}\ (Z'(v_1))^3 && (2.175) \\ \Gamma_{222}^{F,G} &= \frac{G(w)}{w} Z''(v_2)\ Z'(v_2) - \frac{G(w)F''(w)}{wF'(w)}\ (Z'(v_2))^3 && (2.176) \\ \Gamma_{112}^{F,G} &= \frac{G(w)Z'(v_2)}{wF'(w)}\left[F'(w)\ Z''(v_1) - F''(w)\ (Z'(v_1))^2\right] && (2.177) \\ \Gamma_{221}^{F,G} &= \frac{G(w)Z'(v_1)}{wF'(w)}\left[F'(w)\ Z''(v_2) - F''(w)\ (Z'(v_2))^2\right] && (2.178) \end{aligned}$$

$$\Gamma_{121}^{F,G} = \frac{-G(w)F''(w)}{wF'(w)} \left(Z'(v_1)\right)^2 Z'(v_2) \tag{2.179}$$

$$\Gamma_{122}^{F,G} = \frac{-G(w)F''(w)}{wF'(w)} \left(Z'(v_2)\right)^2 Z'(v_1) \tag{2.180}$$

and $\quad \Gamma_{211}^{F,G} = \Gamma_{121}^{F,G};\ \Gamma_{212}^{F,G} = \Gamma_{122}^{F,G}.$

2.4 Summary

In this chapter we described the affine structure of family of measures, the manifold structure of a statistical model and the α-geometry on a statistical manifold. Then the Fisher information metric and the α-connections are computed for statistical manifold defined on finite sets. Further a detailed description of the (F, G)-geometry on a statistical manifold is presented. We proved a necessary and sufficient condition for two (F, G)-connections to be dual with respect to the G-metric. Also we showed that the α-geometry is a special case of the (F, G)-geometry. Further the G-metric and the (F, G)-connections are computed for statistical manifold defined on finite sets.

CHAPTER 3

Invariant and Non-invariant Geometric Structures

In this chapter we study the invariance properties of a statistical manifold. First we discuss about various geometric structures on a statistical manifold induced from a two point function called divergence function. A divergence function measures the amount of discrepancy or asymmetric distance between two probability distributions. Eguchi [17] introduced a method of obtaining geometric structures on a statistical manifold using the divergence function, see also [12], [14]. There are various classes of divergence functions; f-divergence, Bregman divergence, (α, ρ, τ)-divergence, U-divergence etc. These divergence functions give rise to various geometries on a statistical manifold [18–22], [55]

Chentsov [4] proved the uniqueness of the Fisher information metric and the α-connections on a statistical manifold defined on finite sets with respect to the categorical invariance, see also [23], [24]. Amari [12] conjectured that the Fisher information metric and the α-connections are the only metric and affine connections which are invariant under any coordinate transformations of the sample space and of the parameter. Recently, Ay et al. [25] addressed the invariance problem in the infinite dimensional case also. The (F, G)-geometry is a generalized geometric structure on a statistical manifold which includes the α-geometry as a special case. In this chapter we study the invariance properties of the geometric structures and show that the α-geometry is the only invariant geometry among the (F, G)-geometries [16].

Section 3.1 gives an overview of various divergence functions on a statistical manifold and their induced geometric structures. In section 3.2 we describe the invariance properties of the α-geometry, the (F, G)-geometry and the (α, ρ, τ)-geometry on a statistical manifold. First we show that all these geometries are co-variant under reparametrization of the parameter of the manifold. Then prove that both the (F, G)-geometry and the (α, ρ, τ)-geometry and are not invariant under smooth one to one transformations of the random variable in general. Also prove that the α-geometry is the only invariant geometry in the category of both (F, G) and (α, ρ, τ)-geometries. Further the relation between these two geometries is discussed.

3.1 Divergence and the Induced Geometry

In this section Eguchi's method of defining geometric structures using a divergence function is given. Then various classes of divergence functions and the geometric structures induced by them are discussed.

Definition 3.1.1. *Let $\mathcal{S}$ be an n-dimensional manifold with coordinate system $\theta = (\theta^1, \cdots, \theta^n) = (\theta^i)$. Let the coordinates of the points p, q be $(\theta^i), (\theta'^i)$ respectively. A divergence function $D : \mathcal{S} \times \mathcal{S} \longrightarrow \mathbb{R}$ is a smooth function satisfying the following conditions*

1. *$D(p, q) \geq 0$ for any $p, q \in \mathcal{S}$ with equality holding iff $p = q$*
2. *$\partial_i \partial_{j'} D(p, q)|_{p=q}$ is negative definite.*

where $\partial_i = \frac{\partial}{\partial \theta^i}$ and $\partial_{j'} = \frac{\partial}{\partial \theta'^j}$.

Eguchi [17] defined a unique Riemannian metric g^D and an affine connection ∇^D from a divergence D as

$$g^D_{ij}(\theta) = < \partial_i, \partial_j >^D_\theta = -\partial_i \partial_{j'} D(p, q)|_{p=q} \tag{3.1}$$

$$\Gamma^D_{ijk}(\theta) = < \nabla^D_{\partial_i} \partial_j, \partial_k >^D_\theta = -\partial_i \partial_j \partial_{k'} D(p, q)|_{p=q} \tag{3.2}$$

Dual of the divergence D^* of D is defined as $D^*(p, q) = D(q, p)$. The metric and the affine connection induced from D^* are given by

$$g^{D^*} = g^D \tag{3.3}$$

$$\Gamma^{D^*}_{ijk} = -\partial_{i'} \partial_{j'} \partial_k D(p, q)|_{p=q} \tag{3.4}$$

Note that the connections ∇^D and ∇^{D^*} are dual with respect to the metric g^D [17]. Hence a divergence function D induces a dualistic structure $(g^D, \nabla^D, \nabla^{D^*})$ on a statistical manifold.

Now we describe various classes of divergence functions and the geometric structures induced by them.

3.1.1 f-divergence

The most commonly used class of divergence is the f-divergence introduced by Csiszar [18]. Ali and Silvey [19] independently studied the f-divergence class. Let $f : (0, \infty) \rightarrow \mathbb{R}$ be any convex function satisfying $f(1) = 0$ and $f'(1) = 0$.

$$D_f(p, q) = \int f(\frac{q}{p})\; p\; dx. \tag{3.5}$$

An important example of the f-divergence is the α-divergence introduced by Amari [12] which is generated by the function f^α given by

$$f^\alpha(u) = \begin{cases} \frac{4}{1-\alpha^2}\{1 - u^{\frac{1+\alpha}{2}}\}, & \alpha \neq \pm 1 \\ \quad u \log u, & \alpha = 1 \\ \quad -\log u, & \alpha = -1. \end{cases} \tag{3.6}$$

For $\alpha \neq \pm 1$,

$$D^\alpha(p, q) = \frac{4}{1-\alpha^2}\left\{1 - \int p^{\frac{1-\alpha}{2}} q^{\frac{1+\alpha}{2}} dx\right\}. \tag{3.7}$$

and for $\alpha = \pm 1$,

$$D^{-1}(p, q) = D^1(q, p) = \int p\; \log \frac{p}{q}\; dx. \tag{3.8}$$

D^{-1} is the **Kullback-Leibler divergence** or **relative entropy**.

On a statistical manifold the f-divergence induces a Riemannian metric proportional to the Fisher information metric with constant of proportionality $f''(1)$ and affine connection equal to the α-connection with

$$\alpha = 3 + 2\frac{f'''(1)}{f''(1)}. \tag{3.9}$$

3.1.2 Bregman divergence

Bregman [20] introduced another class of divergences called the Bregman divergence. Let $\phi : \Omega \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ be a smooth real valued and strictly convex function defined on a closed convex set Ω. The Bregman divergence associated with the function ϕ is defined as

$$D_\phi(x,y) = \phi(y) - \phi(x) - \nabla\phi(x).(y-x), \quad \forall \quad x,y \in \Omega. \tag{3.10}$$

Let $\mathcal{S} = \{p(x;\theta) \ / \ \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$ be an n-dimensional statistical manifold. Let $\phi : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ be a smooth real valued and strictly convex function defined on a closed convex set $\Omega \subseteq E \subseteq \mathbb{R}^n$. The Bregman divergence associated with ϕ is given by

$$D_\phi(p_\theta, p_{\theta'}) = D_\phi(\theta,\theta') = \phi(\theta') - \phi(\theta) - \nabla\phi(\theta).(\theta'-\theta), \quad \forall \quad \theta,\theta' \in \Omega. \tag{3.11}$$

The metric and dual affine connections induced from the Bregman divergence are

$$\begin{aligned} g_{ij}(\theta) &= \partial_i\partial_j\phi(\theta) & (3.12)\\ \Gamma_{ijk}(\theta) &= \partial_i\partial_j\partial_k\phi(\theta) & (3.13)\\ \Gamma^*_{ijk}(\theta) &= 0. & (3.14) \end{aligned}$$

The Bregman divergence is important in the study of dually flat spaces and in turn in the asymptotic theory of statistical inference which will be discussed in the subsequent chapters.

3.1.3 (α, ρ, τ)-divergence

Zhang [21] introduced a divergence function (or a functional) called the (α,ρ,τ)- divergence using a real parameter α and a conjugate ρ,τ-representations of the density function with respect to a convex function f. This ρ-representation is a generalized notion of the α-representation.

Let $\rho : (0,\infty) \to \mathbb{R}$ be a strictly monotone increasing function and let $f : \mathbb{R} \to \mathbb{R}$ be a smooth strictly convex function. A τ-representation of the density function is said to be conjugate to the ρ-representation with respect to f if

$$\begin{aligned} \tau(p) &= f'(\rho(p)) = ((f')^*)^{-1}(\rho(p)) & (3.15)\\ \rho(p) &= (f')^{-1}(\tau(p)) = (f^*)'(\tau(p)). & (3.16) \end{aligned}$$

Using the ρ-representation of densities and using a real parameter α, a divergence functional $D^{(\alpha)}_{f,\rho}$ is defined as

$$\begin{aligned} D^{(\alpha)}_{f,\rho}(p,q) &= \frac{4}{1-\alpha^2}\int \left[\frac{1-\alpha}{2}f(\rho(p)) + \frac{1+\alpha}{2}f(\rho(q))\right. && (3.17)\\ &\quad \left. -f\left(\frac{1-\alpha}{2}\rho(p)+\frac{1+\alpha}{2}\rho(q)\right)\right] dx && (3.18)\end{aligned}$$

with

$$\begin{aligned} D^{(1)}_{f,\rho}(p,q) &= D^{(-1)}_{f,\rho}(q,p) = D^{(1)}_{f^*,\tau}(q,p) = D^{(-1)}_{f^*,\tau}(p,q) && (3.19)\\ &= \int \left[f(\rho(p)) + f^*(\tau(q)) - \rho(p)\tau(q)\right] dx. && (3.20)\end{aligned}$$

For a parametric model $\mathcal{S}$, this can be written as

$$\begin{aligned} D^{\alpha}_{f,\rho}(\theta_p,\theta_q) &= \frac{4}{1-\alpha^2}\int \left[\frac{1-\alpha}{2}f(\rho(\theta_p)) + \frac{1+\alpha}{2}f(\rho(\theta_q))\right. && (3.21)\\ &\quad \left. -f\left(\frac{1-\alpha}{2}\rho(\theta_p)+\frac{1+\alpha}{2}\rho(\theta_q)\right)\right] dx. && (3.22)\end{aligned}$$

This induces a metric g' and dual connections $\nabla'^{(\alpha)}$, $\nabla'^{*(\alpha)}$ on a statistical model $\mathcal{S}$ given by

$$\begin{aligned} g'_{ij}(\theta) &= \int \frac{\partial\tau}{\partial\theta^i}\frac{\partial\rho}{\partial\theta^j}\, dx. && (3.23)\\ \Gamma'^{(\alpha)}_{ijk}(\theta) &= \int \left[\frac{1-\alpha}{2}\frac{\partial^2\tau}{\partial\theta^i\partial\theta^j}\frac{\partial\rho}{\partial\theta^k} + \frac{1+\alpha}{2}\frac{\partial^2\rho}{\partial\theta^i\partial\theta^j}\frac{\partial\tau}{\partial\theta^k}\right] dx. && (3.24)\\ \Gamma'^{*(\alpha)}_{ijk}(\theta) &= \int \left[\frac{1+\alpha}{2}\frac{\partial^2\tau}{\partial\theta^i\partial\theta^j}\frac{\partial\rho}{\partial\theta^k} + \frac{1-\alpha}{2}\frac{\partial^2\rho}{\partial\theta^i\partial\theta^j}\frac{\partial\tau}{\partial\theta^k}\right] dx. && (3.25)\end{aligned}$$

3.1.4 U-divergence

Murata et al. [22] introduced a generalized class of divergence function called the U-divergence. Eguchi et al. [36] discussed the geometry induced from the U-divergence and its applications. U-divergence is a generalization of Kullback-Leibler divergence and is defined using a generator function U.

Let $U : \mathbb{R} \to \mathbb{R}_+$ be an increasing convex function and let U^* be the convex conjugate of U given by $U^*(t) = t\xi(t) - U(\xi(t))$, where $\xi(t)$ is the inverse function of the derivative of $U(t)$, i.e. $\frac{d}{dt}U^*(t) = \xi(t)$.

The U-divergence is defined as

$$\begin{aligned} D_U(p,q) &= \int [U^*(p) - p\xi(q) + U(\xi(q))]\,dx && (3.26)\\ &= \int [U^*(p) - U^*(q) - \xi(q)(p-q)]\,dx. && (3.27)\end{aligned}$$

The geometry induced from the U-divergence, the U-geometry, is given by [36]

$$\begin{aligned} g^U_{ij}(\theta) &= \int \partial_i p(x;\theta)\ \partial_j \xi(p(x;\theta))\ dx && (3.28)\\ \Gamma^U_{ijk}(\theta) &= \int \partial_i\partial_j p(x;\theta)\ \partial_k \xi(p(x;\theta))\ dx && (3.29)\\ {}^*\Gamma^U_{ijk}(\theta) &= \int\ \partial_k p(x;\theta)\ \partial_i\partial_j \xi(p(x;\theta))\ dx. && (3.30)\end{aligned}$$

Proposition 3.1.2. *The U-geometry is a special case of both the (F,G) and (α,ρ,τ)-geometries.*

Proof. Let us take $F(p) = \xi(p) = (U^*)'(p)$, $H(p) = p$ and $G(p) = p\ \xi'(p)$. Then from Equations (3.28), (3.29) and (3.30), the U-geometric structures can be written as

$$\begin{aligned} g^U_{ij}(\theta) &= \int \partial_i\ell\ \partial_j\ell\ p\ \xi'(p)\ p\ dx = g^G(\theta) && (3.31)\\ \Gamma^U_{ijk}(\theta) &= \int (\partial_i\partial_j\ell + \partial_i\ell\ \partial_j\ell)\,\partial_k\ell\ p\ \xi'(p)\ p\ dx = \Gamma^{H,G}_{ijk}(\theta) && (3.32)\\ {}^*\Gamma^U_{ijk}(\theta) &= \int \left(\partial_i\partial_j\ell + (1 + \frac{p\xi''(p)}{\xi'(p)})\partial_i\ell\ \partial_j\ell\right)\partial_k\ell\ p\ \xi'(p)\ p\ dx && (3.33)\\ &= \Gamma^{F,G}_{ijk}(\theta) && (3.34)\end{aligned}$$

Thus the U-geometry is a special case of the (F,G)-geometry.

The U-geometry is a (α,ρ,τ)-geometry with $f = U$, $\rho(p) = \xi(p)$, $\tau(p) = p$ and the parameter $\alpha = \pm 1$.

$$\begin{aligned} g^U_{ij}(\theta) &= \int \partial_i p(x;\theta)\ \partial_j\xi(p(x;\theta))\ dx = g'_{ij}(\theta) && (3.35)\\ \Gamma^U_{ijk}(\theta) &= \int \partial_i\partial_j p(x;\theta)\ \partial_k\xi(p(x;\theta))\ dx = \Gamma^{'(-1)}_{ijk}(\theta) = \Gamma^{'*(1)}_{ijk}(\theta) && (3.36)\\ {}^*\Gamma^U_{ijk}(\theta) &= \int\ \partial_k p(x;\theta)\ \partial_i\partial_j\xi(p(x;\theta))\ dx = \Gamma^{'(1)}_{ijk}(\theta) = \Gamma^{'*(-1)}_{ijk}(\theta). && (3.37)\end{aligned}$$

□

3.2 Invariant and Non-invariant Geometries on a Statistical Manifold

For a statistical manifold $\mathcal{S} = \{p(x;\theta) \ / \ \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$ the parameters are merely labels attached to each point $p \in \mathcal{S}$. So the intrinsic geometric properties should be independent of these labels. Hence it is natural to consider the invariance properties of the geometric structures under suitable transformations of the variables in a statistical manifold. There are two kinds of invariance of the geometric structures, covariance under reparametrization of the parameter of the manifold and invariance under the smooth one to one transformations of the random variable [12], [14].

On a statistical manifold defined on finite sets, Chentsov [4] proved that the α-geometry can be characterized by the invariance with respect to the sufficient statistic, see also [23], [24]. Amari [12] conjectured that the Fisher information metric and the α-connections are the only metric and affine connections which are invariant under any coordinate transformations of the sample space and of the parameter. In this section we show that the α-geometry is the only invariant geometry among the generalized (F,G)-geometry class. Picard [56] also studied statistical morphisms and related invariance properties. Ay et al. [25] studied this problem in the infinite dimensional case also.

In the previous section we described various classes of divergence functions and their induced geometries. The f-divergence induces the α-geometry and the (α,ρ,τ)-divergence induces the (α,ρ,τ)-geometry. The U-geometry comes under both the (F,G)-geometry and the (α,ρ,τ)-geometry. In this section we discuss the invariance properties of the α-geometry, the (F,G)-geometry and the (α,ρ,τ)-geometry. All these geometries are covariant under reparametrization. But the (F,G) and the (α,ρ,τ)-geometries are in general not invariant under smooth one to one transformations of the random variable. The α-geometry is the only geometry among these geometries which is both covariant under reparametrization and invariant under smooth one to one transformations of the random variable.

Definition 3.2.1. *Let (θ^i) and (η_j) be two coordinate systems on S which are related by an invertible transformation $\eta = \eta(\theta)$. Let the coordinate expressions of the metric g with respect to θ^i and η_i be given by $g_{ij} =< \partial_i, \partial_j >$ and $\tilde{g}_{ij} =< \partial^i, \partial^j >$ respectively, where $\partial_i = \frac{\partial}{\partial \theta^i}$ and $\partial^j = \frac{\partial}{\partial \eta_j}$. Let the components of the connection ∇ with respect to* *the coordinates (θ^i) and (η_j) be given by Γ_{ijk}, $\tilde{\Gamma}_{ijk}$ respectively.*

*Then the **covariance under the reparametrization of the metric and the connection** is defined as [14]*

$$\begin{aligned} \tilde{g}_{ij} &= \sum_m \sum_n \frac{\partial\theta^m}{\partial\eta_i}\frac{\partial\theta^n}{\partial\eta_j} g_{mn} && (3.38) \\ \tilde{\Gamma}_{ijk} &= \sum_{m,n,h} \frac{\partial\theta^m}{\partial\eta_i}\frac{\partial\theta^n}{\partial\eta_j}\frac{\partial\theta^h}{\partial\eta_k}\Gamma_{mnh} + \sum_{m,h}\frac{\partial\theta^h}{\partial\eta_k}\frac{\partial^2\theta^m}{\partial\eta_i\partial\eta_j} g_{mh}. && (3.39) \end{aligned}$$

The covariance under reparametrization actually means that the metric and connections are coordinate independent.

Definition 3.2.2. *Let $\mathcal{S} = \{p(x;\theta) \ / \ \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$ be a statistical manifold defined on a sample space $\mathcal{X}$. Let x, y be random variables defined on sample spaces $\mathcal{X}, \mathcal{Y}$ respectively and ϕ be a smooth one to one transformation of x to y. Assume that this transformation induces a model $\bar{\mathcal{S}} = \{q(y;\theta) \ / \ \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$ on $\mathcal{Y}$. Let $\lambda : \mathcal{S} \longrightarrow \bar{\mathcal{S}}$ be a diffeomorphism defined as*

$$\lambda(p_\theta) = q_\theta. \tag{3.40}$$

*Let $g, \bar{g}$ be Riemannian metrics and $\nabla, \bar{\nabla}$ be affine connections on $\mathcal{S}$ and $\bar{\mathcal{S}}$ respectively. The **invariance under smooth one to one transformation of the random variable** is defined as [14]*

$$\begin{aligned} g(X,Y)_p &= \bar{g}(\lambda_*(X), \lambda_*(Y))_{\lambda(p)} && (3.41) \\ \lambda_*(\nabla_X Y) &= \bar{\nabla}_{\lambda_*(X)}\lambda_*(Y), \quad \forall\, X, Y \in T_\theta(\mathcal{S}) && (3.42) \end{aligned}$$

where λ_* is the push forward map associated with the map λ defined by

$$\lambda_*(X)_{\lambda(p)} = (d\lambda)_p(X). \tag{3.43}$$

Invariant α-geometry

The Fisher information metric and the α-connections are invariant under smooth one to one transformations of random variable and covariant under reparametrization [4], [12], [23–25].

3.2.1 Invariance of the (F, G)-geometry

Here the invariance property of the (F, G)-geometric structure is discussed.

Theorem 3.2.3. *The G-metric g^G is covariant under reparametrization.*

Proof. The components of the G-metric g^G with respect to the coordinate system (θ^i) are

$$g_{ij}^G(\theta) = < \partial_i, \partial_j >_\theta = \int \partial_i p(x;\theta) \partial_j p(x;\theta) \frac{G(p)}{p(x;\theta)} dx. \tag{3.44}$$

Let $\tilde{p}(x;\eta) = p(x;\theta(\eta))$. Then the components of the Fisher information metric with respect to the coordinate system (η_j) are given by

$$\tilde{g}_{ij}(\eta) = < \partial^i, \partial^j >_\eta = \int \partial^i \tilde{p}(x;\eta) \partial^j \tilde{p}(x;\eta) \frac{G(\tilde{p})}{\tilde{p}(x;\eta)} dx. \tag{3.45}$$

Since

$$\partial^i \tilde{p}(x;\eta) = \sum_m \frac{\partial\theta^m}{\partial\eta_i} \frac{\partial p(x;\theta(\eta))}{\partial\theta^m} \tag{3.46}$$

the components of the metric are

$$\begin{aligned}
\tilde{g}_{ij}(\eta) &= \int \partial^i \tilde{p}(x;\eta) \partial^j \tilde{p}(x;\eta) \frac{G(\tilde{p})}{\tilde{p}(x;\eta)} dx && (3.47)\\
&= \int \sum_m \frac{\partial\theta^m}{\partial\eta_i} \frac{\partial p(x;\theta)}{\partial\theta^m} \sum_n \frac{\partial\theta^n}{\partial\eta_j} \frac{\partial p(x;\theta)}{\partial\theta^n} \frac{G(p)}{p(x;\theta)} dx && (3.48)\\
&= \sum_m \sum_n \frac{\partial\theta^m}{\partial\eta_i} \frac{\partial\theta^n}{\partial\eta_j} \int \partial_m p(x;\theta) \partial_n p(x;\theta) \frac{G(p)}{p(x;\theta)} dx. && (3.49)\\
&= \left[\sum_m \sum_n \frac{\partial\theta^m}{\partial\eta_i} \frac{\partial\theta^n}{\partial\eta_j} g_{mn}^G(\theta) \right]_{\theta=\theta(\eta)} && (3.50)
\end{aligned}$$

Hence the G-metric g^G is covariant under reparametrization. □

Theorem 3.2.4. *The (F, G)-connection $\nabla^{F,G}$ is covariant under reparametrization.*

Proof. Let the components of $\nabla^{F,G}$ with respect to the coordinates (θ^i) and (η_j) be given by $\Gamma_{ijk}^{F,G}$, $\tilde{\Gamma}_{ijk}^{F,G}$ respectively.

Let $\tilde{p}(x;\eta) = p(x;\theta(\eta))$. Denote $\log p(x;\theta)$ by ℓ_θ, $\log \tilde{p}(x;\eta)$ by $\tilde{\ell}_\eta$, $p(x;\theta)$ by p_θ and $\tilde{p}(x;\eta)$ by $\tilde{p}_\eta$.

The components of the (F, G)-connection $\nabla^{F,G}$ with respect to the coordinate system

(θ^i) are

$$\Gamma^{F,G}_{ijk} = \int \left(\partial_i \partial_j \ell_\theta + (1 + \frac{pF''(p)}{F'(p)}) \partial_i \ell_\theta \; \partial_j \ell_\theta \right) \partial_k \ell_\theta \; G(p_\theta) \; p_\theta \; dx \tag{3.51}$$

The components of $\nabla^{F,G}$ with respect to the coordinate system (η_j) are

$$\tilde{\Gamma}^{F,G}_{ijk} = \int \left(\partial^i \partial^j \tilde{\ell}_\eta + (1 + \frac{\tilde{p}F''(\tilde{p})}{F'(\tilde{p})}) \partial^i \tilde{\ell}_\eta \; \partial^j \tilde{\ell}_\eta \right) \partial^k \tilde{\ell}_\eta \; G(\tilde{p}_\eta) \; \tilde{p}_\eta \; dx \tag{3.52}$$

Since,

$$\partial^i \tilde{\ell}_\eta = \sum_m \frac{\partial \theta^m}{\partial \eta_i} \frac{\partial \ell_{\theta(\eta)}}{\partial \theta^m} \tag{3.53}$$

then

$$\partial^i \partial^j \tilde{\ell}_\eta = \sum_{m,n} \frac{\partial \theta^m}{\partial \eta_i} \frac{\partial \theta^n}{\partial \eta_j} \frac{\partial^2 \ell_{\theta(\eta)}}{\partial \theta^m \partial \theta^n} + \sum_m \frac{\partial^2 \theta^m}{\partial \eta_i \partial \eta_j} \frac{\partial \ell_{\theta(\eta)}}{\partial \theta^m} \tag{3.54}$$

$$\partial^i \tilde{\ell}_\eta \; \partial^j \tilde{\ell}_\eta = \sum_{m,n} \frac{\partial \theta^m}{\partial \eta_i} \frac{\partial \theta^n}{\partial \eta_j} \frac{\partial \ell_{\theta(\eta)}}{\partial \theta^m} \frac{\partial \ell_{\theta(\eta)}}{\partial \theta^n} \tag{3.55}$$

$$\partial^k \tilde{\ell}_\eta = \sum_h \frac{\partial \theta^h}{\partial \eta_k} \frac{\partial \ell_{\theta(\eta)}}{\partial \theta^h} \tag{3.56}$$

Hence

$$\begin{aligned} \tilde{\Gamma}^{F,G}_{ijk} &= \int (1 + \frac{pF''(p)}{F'(p)}) \sum_{m,n,h} \frac{\partial \theta^m}{\partial \eta_i} \frac{\partial \theta^n}{\partial \eta_j} \frac{\partial \theta^h}{\partial \eta_k} \frac{\partial \ell_{\theta(\eta)}}{\partial \theta^m} \frac{\partial \ell_{\theta(\eta)}}{\partial \theta^n} \frac{\partial \ell_{\theta(\eta)}}{\partial \theta^h} G(p_{\theta(\eta)}) \; p_{\theta(\eta)} \; dx \\ &\quad + \int \sum_{m,h} \frac{\partial^2 \theta^m}{\partial \eta_i \partial \eta_j} \frac{\partial \theta^h}{\partial \eta_k} \frac{\partial \ell_{\theta(\eta)}}{\partial \theta^m} \frac{\partial \ell_{\theta(\eta)}}{\partial \theta^h} G(p_{\theta(\eta)}) \; p_{\theta(\eta)} \; dx \\ &\quad + \int \sum_{m,n,h} \frac{\partial \theta^m}{\partial \eta_i} \frac{\partial \theta^n}{\partial \eta_j} \frac{\partial \theta^h}{\partial \eta_k} \frac{\partial^2 \ell_{\theta(\eta)}}{\partial \theta^m \partial \theta^n} \frac{\partial \ell_{\theta(\eta)}}{\partial \theta^h} G(p_{\theta(\eta)}) \; p_{\theta(\eta)} \; dx \end{aligned} \tag{3.57}$$

$$\begin{aligned} &= \sum_{m,n,h} \frac{\partial \theta^m}{\partial \eta_i} \frac{\partial \theta^n}{\partial \eta_j} \frac{\partial \theta^h}{\partial \eta_k} \int (1 + \frac{pF''(p)}{F'(p)}) \frac{\partial \ell_{\theta(\eta)}}{\partial \theta^m} \frac{\partial \ell_{\theta(\eta)}}{\partial \theta^n} \frac{\partial \ell_{\theta(\eta)}}{\partial \theta^h} G(p_{\theta(\eta)}) \; p_{\theta(\eta)} \; dx \\ &\quad + \sum_{m,h} \frac{\partial^2 \theta^m}{\partial \eta_i \partial \eta_j} \frac{\partial \theta^h}{\partial \eta_k} \int \frac{\partial \ell_{\theta(\eta)}}{\partial \theta^m} \frac{\partial \ell_{\theta(\eta)}}{\partial \theta^h} G(p_{\theta(\eta)}) \; p_{\theta(\eta)} \; dx \\ &\quad + \sum_{m,n,h} \frac{\partial \theta^m}{\partial \eta_i} \frac{\partial \theta^n}{\partial \eta_j} \frac{\partial \theta^h}{\partial \eta_k} \int \frac{\partial^2 \ell_{\theta(\eta)}}{\partial \theta^m \partial \theta^n} \frac{\partial \ell_{\theta(\eta)}}{\partial \theta^h} G(p_{\theta(\eta)}) \; p_{\theta(\eta)} \; dx \end{aligned} \tag{3.58}$$

$$= \sum_{m,n,h} \frac{\partial \theta^m}{\partial \eta_i} \frac{\partial \theta^n}{\partial \eta_j} \frac{\partial \theta^h}{\partial \eta_k} \Gamma^{F,G}_{mnh} + \sum_{m,h} \frac{\partial \theta^h}{\partial \eta_k} \frac{\partial^2 \theta^m}{\partial \eta_i \partial \eta_j} g^G_{mh} \tag{3.59}$$

Thus the (F, G)-connection is covariant under reparametrization of the parameter. □

Now we prove that the (F, G)-geometry is not invariant under smooth one to one transformations of the random variable in general and the α-geometry is the only invariant geometry among the (F, G)-geometries.

Theorem 3.2.5. *The (F, G)-geometric structures, the G-metric and the (F, G)- connection, are not invariant under smooth one to one transformations of the random variable in general.*

Proof. Consider a statistical manifold $\mathcal{S} = \{p(x; \theta) \ / \ \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$ defined on a sample space $\mathcal{X}$. Let ϕ be a smooth one to one transformation of the random variable x to y. This induces a model $\bar{\mathcal{S}} = \{q(y; \theta) \ / \ \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$ on the sample space $\mathcal{Y}$. Then

$$\begin{aligned} q(y : \theta) &= p(w(y); \theta) w'(y) && (3.60) \\ p(x; \theta) &= q(\phi(x); \theta) \phi'(x) && (3.61) \\ \partial_i \ell(x; \theta) &= \partial_i \ell(\phi(x); \theta) && (3.62) \end{aligned}$$

where w is a function such that $x = w(y)$ and $\phi'(x) = \frac{1}{w'(\phi(x))}$.
For convenience, denote $p(x; \theta)$ by p_x, $q(y; \theta)$ by q_y, $\log(p(x; \theta))$ by $\ell(p_x)$ and $\log(q(y; \theta))$ by $\ell(q_y)$. For any function h, $h(p(x; \theta))$ be denoted by $h(p_x)$ and $h(q(y; \theta))$ be denoted by $h(q_y)$.
Let $g^G, \bar{g}^G$ be the G-metrics defined on $\mathcal{S}$ and $\bar{\mathcal{S}}$ respectively. Then

$$\begin{aligned} g_{ij}^G(\theta) &= \int \partial_i \ell(p_x) \, \partial_j \ell(p_x) \, G(p_x) \, p_x \, dx && (3.63) \\ \bar{g}_{ij}^G(\theta) &= \int \partial_i \ell(q_y) \, \partial_j \ell(q_y) \, G(q_y) \, q_y \, dy && (3.64) \\ &= \int \partial_i \ell(p_x) \, \partial_j \ell(p_x) \, G(q_{\phi(x)}) \, p_x \, dx. \end{aligned}$$

The condition for invariance of the G-metric is

$$\int \partial_i \ell(p_x) \, \partial_j \ell(p_x) \, G(q_{\phi(x)}) \, p_x \, dx = \int \partial_i \ell(p_x) \, \partial_j \ell(p_x) \, G(p_x) \, p_x \, dx \qquad (3.65)$$

This implies that $G(p) = c$, where c is a constant. Thus the G-metric is not invariant in general. It is invariant only if $G(p)$ is a constant.
Now let us look at the invariance of the (F, G)-connection.
Let $\Gamma_{ijk}^{F,G}$ and $\bar{\Gamma}_{ijk}^{F,G}$ be the components of the of (F, G)-connection in $\mathcal{S}$ and $\bar{\mathcal{S}}$ respec-

tively. Then

$$\Gamma^{F,G}_{ijk}(\theta) = \int (1+\frac{p_x F''(p_x)}{F'(p_x)})\partial_i\ell(p_x)\,\partial_j\ell(p_x)\,\partial_k\ell(p_x)\,G(p_x)\,p_x\,dx + \int \partial_i\partial_j\ell(p_x)\,\partial_k\ell(p_x)\,G(p_x)\,p_x\,dx \tag{3.66}$$

$$\bar{\Gamma}^{F,G}_{ijk}(\theta) = \int (1+\frac{q_y F''(q_y)}{F'(q_y)})\partial_i\ell(q_y)\,\partial_j\ell(q_y)\,\partial_k\ell(q_y)\,G(q_y)\,q_y\,dy + \int \partial_i\partial_j\ell(q_y)\,\partial_k\ell(q_y)\,G(q_y)\,q_y\,dy \tag{3.67}$$

$$= \int (1+\frac{q_{\phi(x)} F''(q_{\phi(x)})}{F'(q_{\phi(x)})})\partial_i\ell(p_x)\,\partial_j\ell(p_x)\,\partial_k\ell(p_x)\,G(q_{\phi(x)})\,p_x\,dx + \int \partial_i\partial_j\ell(p_x)\,\partial_k\ell(p_x)\,G(p_x)\,p_x\,dx \tag{3.68}$$

The condition for invariance of the (F,G)-connection is

$$\int (1+\frac{q_{\phi(x)} F''(q_{\phi(x)})}{F'(q_{\phi(x)})})\partial_i\ell(p_x)\,\partial_j\ell(p_x)\,\partial_k\ell(p_x)\,G(q_{\phi(x)})\,p_x\,dx + \int \partial_i\partial_j\ell(p_x)\,\partial_k\ell(p_x)\,G(p_x)\,p_x\,dx = \int (1+\frac{p_x F''(p_x)}{F'(p_x)})\partial_i\ell(p_x)\,\partial_j\ell(p_x)\,\partial_k\ell(p_x)\,G(p_x)\,p_x\,dx + \int \partial_i\partial_j\ell(p_x)\,\partial_k\ell(p_x)\,G(p_x)\,p_x\,dx \tag{3.69}$$

Then

$$\frac{pF''(p)}{F'(p)} = k; \quad G(p) = k_1 \tag{3.70}$$

where k, k_1 are real constants.

Thus in general, the (F,G)-connection $\nabla^{F,G}$ is not invariant. It is invariant if only if Equation (3.70) holds.

Hence the (F,G)-geometry is not invariant under smooth one to one transformations of the random variable in general. □

Corollary 3.2.6. *The only (F,G)-geometry which is invariant under smooth one to one transformations of the random variable is the α-geometry.*

Proof. Using Euler's homogeneous function theorem, it follows from equation (3.70) that the function F' is a positive homogeneous function in p of degree k. Hence

$$F'(\lambda p) = \lambda^k F'(p), \quad \text{for } \lambda > 0. \tag{3.71}$$

Since F' is a positive homogeneous function in the single variable p, without loss of generality take

$$F'(p) = p^k. \tag{3.72}$$

Therefore

$$F(p) = \begin{cases} \frac{p^{k+1}}{k+1}, & k \neq -1 \\ \log p, & k = -1 \end{cases} \tag{3.73}$$

Let

$$k = \frac{-(1+\alpha)}{2}, \ \alpha \in \mathbb{R}. \tag{3.74}$$

Then

$$F(p) = \begin{cases} \frac{2}{1-\alpha} p^{\frac{1-\alpha}{2}}, & \alpha \neq 1 \\ \log p, & \alpha = 1 \end{cases} \tag{3.75}$$

which is Amari's α-embeddings $L_\alpha(p)$.

Also without loss of generality take $k_1 = 1$. Then $G(p) = 1$. Thus the (F, G)-connection reduces to the α-connection and the G-metric reduces to the Fisher information metric.

Hence we obtain that the α-geometry is the only (F, G)-geometry which is invariant under smooth one to one transformations of the random variable. □

3.2.2 Invariance of the (α, ρ, τ)-geometry

Zhang [21] introduced the (α, ρ, τ)-divergence which induces a dualistic structure called the (α, ρ, τ)-geometry.

$$g'_{ij}(\theta) = \int \partial_i \tau \, \partial_j \rho \, dx. \tag{3.76}$$

$$\Gamma'^{(\alpha)}_{ijk}(\theta) = \int \left[\frac{1-\alpha}{2} \partial_i \partial_j \tau \, \partial_k \rho + \frac{1+\alpha}{2} \partial_i \partial_j \rho \, \partial_k \tau \right] dx. \tag{3.77}$$

$$\Gamma'^{*(\alpha)}_{ijk}(\theta) = \int \left[\frac{1+\alpha}{2} \partial_i \partial_j \tau \, \partial_k \rho + \frac{1-\alpha}{2} \partial_i \partial_j \rho \, \partial_k \tau \right] dx. \tag{3.78}$$

where $\partial_i = \frac{\partial}{\partial \theta^i}$.

Now we show that the (α, ρ, τ)-geometry is covariant under reparametrization and not invariant under smooth one to one transformations of the random variable in general.

Theorem 3.2.7. *The (α, ρ, τ)-geometric structures, the metric g' and the affine connection $\nabla'^{(\alpha)}$, are covariant under reparametrization.*

Proof. Let the components of g' with respect to the coordinates (θ^i) and (η_j) be given by g'_{ij}, $\tilde{g}'_{ij}$ respectively.

Let $\tilde{p}(x;\eta) = p(x;\theta(\eta))$. Denote $\tau(p(x;\theta))$ by $\tau(x;\theta)$, $\tau(\tilde{p}(x;\eta))$ by $\tilde{\tau}(x;\eta)$, $\rho(p(x;\theta))$ by $\rho(x;\theta)$ and $\rho(\tilde{p}(x;\eta))$ by $\tilde{\rho}(x;\eta)$.

We have

$$\begin{aligned}
\partial^i \tilde{\tau}(x;\eta) &= \sum_m \frac{\partial\theta^m}{\partial\eta_i}\frac{\partial\tau(x;\theta(\eta))}{\partial\theta^m} && (3.79)\\
\partial^i \tilde{\rho}(x;\eta) &= \sum_m \frac{\partial\theta^m}{\partial\eta_i}\frac{\partial\rho(x;\theta(\eta))}{\partial\theta^m} && (3.80)\\
\partial^i\partial^j \tilde{\tau}(x;\eta) &= \sum_{m,n} \frac{\partial\theta^m}{\partial\eta_i}\frac{\partial\theta^n}{\partial\eta_j}\frac{\partial^2\tau(x;\theta(\eta))}{\partial\theta^m\partial\theta^n} + \sum_m \frac{\partial^2\theta^m}{\partial\eta_i\partial\eta_j}\frac{\partial\tau(x;\theta(\eta))}{\partial\theta^m} && (3.81)\\
\partial^i\partial^j \tilde{\rho}(x;\eta) &= \sum_{m,n} \frac{\partial\theta^m}{\partial\eta_i}\frac{\partial\theta^n}{\partial\eta_j}\frac{\partial^2\rho(x;\theta(\eta))}{\partial\theta^m\partial\theta^n} + \sum_m \frac{\partial^2\theta^m}{\partial\eta_i\partial\eta_j}\frac{\partial\rho(x;\theta(\eta))}{\partial\theta^m}. && (3.82)
\end{aligned}$$

The components of g' with respect to θ are

$$g'_{ij}(\theta) = \int \partial_i\tau(x;\theta)\ \partial_j\rho(x;\theta)\ dx. \qquad (3.83)$$

The components of g' with respect to η can be written as

$$\begin{aligned}
\tilde{g}'_{ij}(\eta) &= \int \partial^i\tilde{\tau}(x;\eta)\ \partial^j\tilde{\rho}(x;\eta)\ dx. && (3.84)\\
&= \int \sum_m \frac{\partial\theta^m}{\partial\eta_i}\frac{\partial\tau(x;\theta)}{\partial\theta^m}\sum_n \frac{\partial\theta^n}{\partial\eta_j}\frac{\partial\rho(x;\theta)}{\partial\theta^n}dx && (3.85)\\
&= \sum_m\sum_n \frac{\partial\theta^m}{\partial\eta_i}\frac{\partial\theta^n}{\partial\eta_j}\int \partial_m\tau(x;\theta)\partial_n\rho(x;\theta)dx. && (3.86)\\
&= \left[\sum_m\sum_n \frac{\partial\theta^m}{\partial\eta_i}\frac{\partial\theta^n}{\partial\eta_j}g'_{mn}(\theta)\right]_{\theta=\theta(\eta)} && (3.87)
\end{aligned}$$

Thus the metric g' is covariant under reparametrization.

Let the components of $\nabla'^{(\alpha)}$ with respect to the coordinates (θ^i) and (η_j) be $\Gamma'^{(\alpha)}_{ijk}$, $\tilde{\Gamma}'^{(\alpha)}_{ijk}$

respectively.

$$
\begin{align}
\Gamma'^{(\alpha)}_{ijk}(\theta) &= \frac{1-\alpha}{2}\int \partial^i\partial^j\tilde{\tau}(x;\theta)\;\partial^k\tilde{\rho}(x;\theta)\;dx \nonumber\\
&\quad + \frac{1+\alpha}{2}\int \partial^i\partial^j\tilde{\rho}(x;\theta)\;\partial^k\tilde{\tau}(x;\theta)\;dx. \tag{3.88}\\
&= \frac{1-\alpha}{2}\left[\int\sum_{m,n,h}\frac{\partial\theta^m}{\partial\eta_i}\frac{\partial\theta^n}{\partial\eta_j}\frac{\partial\theta^h}{\partial\eta_k}\frac{\partial^2\tau(x;\theta(\eta))}{\partial\theta^m\partial\theta^n}\frac{\partial\rho(x;\theta(\eta))}{\partial\theta^h}dx\right. \nonumber\\
&\quad \left.+\int\sum_{m,h}\frac{\partial^2\theta^m}{\partial\eta_i\partial\eta_j}\frac{\partial\theta^h}{\partial\eta_k}\frac{\partial\tau(x;\theta(\eta))}{\partial\theta^m}\frac{\partial\rho(x;\theta(\eta))}{\partial\theta^h}\;dx\right] \nonumber\\
&\quad \frac{1+\alpha}{2}\left[\int\sum_{m,n,h}\frac{\partial\theta^m}{\partial\eta_i}\frac{\partial\theta^n}{\partial\eta_j}\frac{\partial\theta^h}{\partial\eta_k}\frac{\partial^2\rho(x;\theta(\eta))}{\partial\theta^m\partial\theta^n}\frac{\partial\tau(x;\theta(\eta))}{\partial\theta^h}dx\right. \nonumber\\
&\quad \left.+\int\sum_{m,h}\frac{\partial^2\theta^m}{\partial\eta_i\partial\eta_j}\frac{\partial\theta^h}{\partial\eta_k}\frac{\partial\rho(x;\theta(\eta))}{\partial\theta^m}\frac{\partial\tau(x;\theta(\eta))}{\partial\theta^h}\;dx\right] \tag{3.89}\\
&= \sum_{m,n,h}\frac{\partial\theta^m}{\partial\eta_i}\frac{\partial\theta^n}{\partial\eta_j}\frac{\partial\theta^h}{\partial\eta_k}\Gamma'^{\alpha}_{mnh} + \sum_{m,h}\frac{\partial\theta^h}{\partial\eta_k}\frac{\partial^2\theta^m}{\partial\eta_i\partial\eta_j}g'_{mh}. \tag{3.90}
\end{align}
$$

Thus the connection $\nabla'^{(\alpha)}$ is covariant under reparametrization. Hence the (α,ρ,τ)-geometry is covariant under reparametrization. □

Theorem 3.2.8. *The (α,ρ,τ)-geometric structures g' and $\nabla'^{(\alpha)}$ are not invariant under smooth one to one transformations of the random variable in general.*

Proof. Consider a statistical manifold $\mathcal{S} = \{p(x;\theta)\ /\ \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$ defined on a sample space $\mathcal{X}$. Let ϕ be a smooth one to one transformation of the random variable x to y. This induces a model $\bar{\mathcal{S}} = \{q(y;\theta)\ /\ \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$ on $\mathcal{Y}$. Then

$$
\begin{align}
q(y:\theta) &= p(w(y);\theta)w'(y) \tag{3.91}\\
p(x;\theta) &= q(\phi(x);\theta)\phi'(x) \tag{3.92}\\
\partial_i\ell(x;\theta) &= \partial_i\ell(\phi(x);\theta) \tag{3.93}
\end{align}
$$

where w is a function such that $x = w(y)$ and $\phi'(x) = \frac{1}{w'(\phi(x))}$.
For convenience, denote $p(x;\theta)$ by p_x, $q(y;\theta)$ by q_y, $\log(q(y;\theta))$ by $\ell(q_y)$ and $\log(q(y;\theta))$ by $\ell(q_y)$. Also for any function h, $h(p(x;\theta))$ be denoted by $h(p_x)$ and $h(q(y;\theta))$ be denoted by $h(q_y)$.

Let $g', \bar{g}'$ be the metrics defined on $\mathcal{S}$ and $\bar{\mathcal{S}}$ respectively. Then

$$\begin{aligned} g'_{ij}(\theta) &= \int \partial_i \tau(p_x)\, \partial_j \rho(p_x)\, dx && (3.94)\\ &= \int p_x\, \tau'(p_x)\, \rho'(p_x)\, \partial_i \ell(p_x)\, \partial_j \ell(p_x)\, p_x\, dx && (3.95)\\ \bar{g}'_{ij}(\theta) &= \int \partial_i \tau(q_y)\, \partial_j \rho(q_y)\, dy && (3.96)\\ &= \int q_y\, \tau'(q_y)\, \rho'(q_y)\, \partial_i \ell(q_y)\, \partial_j \ell(q_y)\, q_y\, dy && (3.97)\\ &= \int q_{\phi(x)}\, \tau'(q_{\phi(x)})\, \rho'(q_{\phi(x)})\, \partial_i \ell(p_x)\, \partial_j \ell(p_x)\, p_x\, dx. && (3.98) \end{aligned}$$

The condition for invariance of the metric is

$$\begin{aligned} &\int q_{\phi(x)}\, \tau'(q_{\phi(x)})\, \rho'(q_{\phi(x)})\, \partial_i \ell(p_x)\, \partial_j \ell(p_x)\, p_x\, dx = && (3.99)\\ &\qquad \int p_x\, \tau'(p_x)\, \rho'(p_x)\, \partial_i \ell(p_x)\, \partial_j \ell(p_x)\, p_x\, dx. && (3.100) \end{aligned}$$

This implies that

$$p\, \tau'(p)\, \rho'(p) = c \tag{3.101}$$

where c is a constant.

That is the metric g' is not invariant in general. It is invariant if only if $p\ \tau'(p)\ \rho'(p)$ is a constant.

Now let us look at the invarinace of the connection.

Let $\Gamma'^{(\alpha)}_{ijk}$ and $\bar{\Gamma}'^{(\alpha)}_{ijk}$ be the components of the connection in $\mathcal{S}$ and $\bar{\mathcal{S}}$ respectively.

$$\begin{aligned} \Gamma'^{(\alpha)}_{ijk}(\theta) &= \frac{1-\alpha}{2}\Big[\int (1 + \frac{p_x \tau''(p_x)}{\tau'(p_x)}) \partial_i \ell(p_x) \partial_j \ell(p_x) \partial_k \ell(p_x) \tau'(p_x) \rho'(p_x) p_x^2\, dx \\ &\quad + \int \partial_i \partial_j \ell(p_x) \partial_k \ell(p_x) \tau'(p_x) \rho'(p_x) p_x^2 dx\Big] \\ &\quad + \frac{1+\alpha}{2}\Big[\int (1 + \frac{p_x \rho''(p_x)}{\rho'(p_x)}) \partial_i \ell(p_x) \partial_j \ell(p_x) \partial_k \ell(p_x) \tau'(p_x) \rho'(p_x) p_x^2\, dx \\ &\quad + \int \partial_i \partial_j \ell(p_x) \partial_k \ell(p_x) \tau'(p_x) \rho'(p_x) p_x^2 dx\Big]. \end{aligned} \tag{3.102}$$

$$\begin{aligned}\bar{\Gamma}'^{(\alpha)}_{ijk}(\theta) &= \frac{1-\alpha}{2}\bigg[\int (1+\frac{q_y\tau''(q_y)}{\tau'(q_y)})\partial_i\ell(q_y)\partial_j\ell(q_y)\partial_k\ell(q_y)\tau'(q_y)\rho'(q_y)q_y^2\,dy \\ &\quad + \int \partial_i\partial_j\ell(q_y)\partial_k\ell(q_y)\tau'(q_y)\rho'(q_y)q_y^2 dy\bigg] \\ &\quad + \frac{1+\alpha}{2}\bigg[\int (1+\frac{q_y\rho''(q_y)}{\rho'(q_y)})\partial_i\ell(q_y)\partial_j\ell(q_y)\partial_k\ell(q_y)\tau'(q_y)\rho'(q_y)q_y^2\,dy \\ &\quad + \int \partial_i\partial_j\ell(q_y)\partial_k\ell(q_y)\tau'(q_y)\rho'(q_y)q_y^2 dy\bigg] \qquad (3.103) \\ &= \frac{1-\alpha}{2}\bigg[\int \partial_i\ell(p_x)\partial_j\ell(p_x)\partial_k\ell(p_x)\tau'(q_{\phi(x)})\rho'(q_{\phi(x)})q_{\phi(x)}p_x\,dx \\ &\quad + \frac{q_{\phi(x)}\tau''(q_{\phi(x)})}{\tau'(q_{\phi(x)})}\partial_i\ell(p_x)\partial_j\ell(p_x)\partial_k\ell(p_x)\tau'(q_{\phi(x)})\rho'(q_{\phi(x)})q_{\phi(x)}p_x\,dx \\ &\quad \frac{1+\alpha}{2}\bigg[\int \partial_i\ell(p_x)\partial_j\ell(p_x)\partial_k\ell(p_x)\tau'(q_{\phi(x)})\rho'(q_{\phi(x)})q_{\phi(x)}p_x\,dx \\ &\quad + \frac{q_{\phi(x)}\rho''(q_{\phi(x)})}{\rho'(q_{\phi(x)})}\partial_i\ell(p_x)\partial_j\ell(p_x)\partial_k\ell(p_x)\tau'(q_{\phi(x)})\rho'(q_{\phi(x)})q_{\phi(x)}p_x\,dx \\ &\quad + \int \partial_i\partial_j\ell(p_x)\partial_k\ell(p_x)\tau'(q_{\phi(x)})\rho'(q_{\phi(x)})q_{\phi(x)}p_x dx\bigg]. \qquad (3.104)\end{aligned}$$

From the condition for invariance of the connection we get

$$\frac{p\tau''(p)}{\tau'(p)} = c_1; \quad \frac{p\rho''(p)}{\rho'(p)} = c_2, \qquad (3.105)$$

$$p\,\tau'(p)\,\rho'(p) = c_3 \qquad (3.106)$$

where c_1, c_2, c_3 are real constants.

Thus in general, the connection $\nabla'^{(\alpha)}$ is not invariant. It is invariant if and only if the Equations (3.105) and (3.106) hold.

Hence the (α, ρ, τ)-geometry is not invariant under smooth one to one transformations of the random variable in general. □

Corollary 3.2.9. *The only (α, ρ, τ)-geometry which is invariant under smooth one to one transformations of the random variable is the α-geometry.*

Proof. Using the homogeneous function theorem, it follows from the Equation (3.105) that

$$\tau'(p) = p^{c_1}, \quad \rho'(p) = p^{c_2} \qquad (3.107)$$

From the Equation (3.106) it follows that

$$c_1 + c_2 + 1 = 0 \quad \text{or} \quad c_2 = 1 - c_1. \qquad (3.108)$$

Hence

$$\tau(p) = \frac{p^{c_1+1}}{c_1+1}, \quad \rho(p) = \frac{p^{-c_1}}{-c_1} \tag{3.109}$$

Now let

$$c_1 = -\frac{(1-\beta)}{2}, \quad \beta \in \mathbb{R} \tag{3.110}$$

$$\tau(p) = \frac{2}{1+\beta} p^{\frac{1+\beta}{2}}, \quad \rho(p) = \frac{2}{1-\beta} p^{\frac{1-\beta}{2}} \tag{3.111}$$

Thus the connection $\nabla'^{(\alpha)}$ reduces to

$$\begin{aligned} \Gamma'^{(\alpha)}(\theta) &= \int \left(\partial_i \partial_j \ell + \frac{1-\alpha\beta}{2} \partial_i \ell \, \partial_j \ell \right) \partial_k \ell \, p \, dx & (3.112) \\ &= \Gamma^{(\alpha\beta)}(\theta). & (3.113) \end{aligned}$$

This is Amari's α-connection with parameter $\alpha\beta$.

From Equations (3.95) and (3.106)

$$\begin{aligned} g'_{ij}(\theta) &= c_3 \int \partial_i \ell(x;\theta) \, \partial_j \ell(x;\theta) \, p(x;\theta) \, dx \\ &= c_3 \, g_{ij}(\theta). & (3.114) \end{aligned}$$

That is the metric g' reduces to a constant times the Fisher information metric g.

Thus the only (α, ρ, τ)-geometry which is invariant under smooth one to one transformations of the random variable is the α-geometry. □

Remark 3.2.10. *Zhang [21] showed that the only measure invariant divergence function associated with quasi-linear mean operator which is scale invariant is a two parameter family of divergence $D^{\alpha,\beta}$ given by*

$$\begin{aligned} D^{\alpha,\beta}(p,q) &= \frac{4}{1-\alpha^2} \frac{2}{1+\beta} \int \left[\frac{1-\alpha}{2} p + \frac{1+\alpha}{2} q \right. \\ &\quad \left. - \left(\frac{1-\alpha}{2} p^{\frac{1-\beta}{2}} + \frac{1+\alpha}{2} q^{\frac{1-\beta}{2}} \right)^{\frac{2}{1-\beta}} \right] dx & (3.115) \end{aligned}$$

where $\alpha, \beta \in [-1, 1]$.

This divergence function induces Fisher information metric and Amari's α-connection with parameter $\alpha\beta$.

3.2.3 (F, G) and (α, ρ, τ)-geometries

The (F, G)-geometry and the (α, ρ, τ)-geometry come under the category of generalized geometries on a statistical manifold which are non-invariant. The (F, G)- geometry is derived naturally by embedding the manifold into the space of random variables $\mathbb{R}_{\mathcal{X}}$ and suitably defining the inner product on $\mathbb{R}_{\mathcal{X}}$. This is done using an embedding function F and a positive smooth function G. The α-geometry is a special case of this (F, G)-geometry and is the only invariant geometry in that category.

The (α, ρ, τ)-geometry is induced from a divergence function ((α, ρ, τ)-divergence). This divergence function is defined using the conjugate representations ρ and τ of densities with respect to a convex function f. The α-geometry is the only invariant geometry among the (α, ρ, τ)-geometries. Zhang [57] claimed that the (F, G)-geometry and the (α, ρ, τ)-geometry are the same. But in the definition of (α, ρ, τ)-geometry the two representations used are conjugate with respect to a strictly convex function f, which is indeed a strong condition. In this context we have the following theorem and examples.

Theorem 3.2.11. *The (α, ρ, τ)-geometry can always be expressed as (F, G)-geometry. Conversely, if the function*

$$f(x) = \int_a^x \left(\int_b^{F^{-1}(t)} \frac{G(u)}{uF'(u)} \, du \right) dt \tag{3.116}$$

exists then the (F, G)-geometry can be expressed as (α, ρ, τ)-geometry.

Proof. For the (α, ρ, τ)-geometry the conjugate representations ρ and τ with respect to a convex function f are given by

$$\tau(p) = f'(\rho(p)) = ((f')^*)^{-1}(\rho(p)) \tag{3.117}$$

$$\rho(p) = (f')^{-1}(\tau(p)) = (f^*)'(\tau(p)). \tag{3.118}$$

Now take $F(p) = \rho(p)$ and $H(p) = \tau(p)$. Then G is determined as follows

$$G(p) = pF'(p)H'(p) \tag{3.119}$$

$$= pf''(F(p))(F'(p))^2. \tag{3.120}$$

Since ρ and τ are smooth strictly increasing functions the function G is well defined and positive. Thus the (α, ρ, τ)-geometry can always be expressed as the (F, G)-geometry.

Conversely, for the dualistic (F, G)-geometry $(g^G, \nabla^{F,G}, \nabla^{H,G})$ the dual embedding relation is given by

$$H'(p) = \frac{G(p)}{pF'(p)}. \tag{3.121}$$

Take $\rho(p) = F(p)$ and $\tau(p) = H(p)$ and define a convex function f as

$$f(x) = \int_a^x \left(\int_b^{F^{-1}(t)} \frac{G(u)}{uF'(u)}\, du \right) dt = \int_a^x H(F^{-1}(t))\, dt \tag{3.122}$$

This integral need not exist in general. So the function f need not exist in general even if the functions F, H, G exist. Thus the (F, G)-geometry can not be expressed as (α, ρ, τ)-geometry in general.

If the function f exists then the (F, G)-geometry can be expressed as (α, ρ, τ)-geometry.

□

Remark 3.2.12. *Thus the (α, ρ, τ)-geometry can always be expressed as (F, G)-geometry. Further, if we assume that the (F, G)-geometry can be expressed as (α, ρ, τ)-geometry (that is, the convex function f exist) then we have from Equation* (3.24)

$$\nabla'^{(\alpha)} = \frac{1-\alpha}{2}\nabla^{H,G} + \frac{1+\alpha}{2}\nabla^{F,H}. \tag{3.123}$$

Example 3.2.13. Let $F(x) = \ln x$ and $G(x) = \ln x$, Then the G-dual embedding of F is $H(x) = x \ln x - x$.

Now let $\rho(x) = F(x) = \ln x$, $\tau(x) = H(x) = x \ln x - x$. Then from Equation (3.117), f is defined by

$$f'(\ln x) = x \ln x - x \tag{3.124}$$

Let $y = \ln x$. Then $f'(y) = e^y(y - 1)$. Thus

$$\begin{aligned} f(y) &= \int_0^y e^t(t-1)\, dt & (3.125)\\ &= e^y(y-2) + 2. & (3.126) \end{aligned}$$

In this example, since the function f is well defined, the duality of F and H with respect to G can be interpreted in terms of the conjugacy of ρ, τ with respect to f. Thus the (F, G)-geometry can be expressed as (α, ρ, τ)-geometry.

Example 3.2.14. Let $F(x) = x\ln x - x$ and

$$G(x) = \frac{x(\ln x - 1)}{\ln x} \tag{3.127}$$

The G-dual embedding of F is

$$H(x) = \frac{x}{\ln x} \tag{3.128}$$

Let us try to find a convex function f to obtain a (α, ρ, τ)-representation. Take $\rho(x) = F(x)$ and $\tau(x) = H(x)$. The function f is defined by

$$f'(\rho(x)) = \tau(x) \tag{3.129}$$

That is,

$$f'(x\ln x - x) = \frac{x}{\ln x} \tag{3.130}$$

Note that in this case we cannot find an explicit expression of f with respect to which the ρ and τ are conjugate and thus elucidating the Theorem 3.2.11.

3.3 Summary

In this chapter first we described various classes of divergence functions and the geometry induced by them. Then we obtained the U-geometry is a special case of both the (F, G) and (α, ρ, τ)-geometries. We studied the invariance properties of the α-geometry, (F, G)-geometry and (α, ρ, τ)-geometry on a statistical manifold and classified them into two categories; invariant and non-invariant. We showed that all these geometries are covariant under reparametrization. Further we showed that both the (F, G)-geometry and the (ρ, τ)-geometry are not invariant in general. As a partial answer to Amari's conjecture we showed that the α-geometry is the only invariant geometry in the category of the generalized (F, G)-geometry. The (α, ρ, τ)-geometry can be expressed as the (F, G)-geometry and the (F, G)-geometry can be expressed as (α, ρ, τ)-geometry provided the convex function f with respect to which ρ and τ are conjugate exists. Also examples are given to make this point clear.

CHAPTER 4

Deformed Exponential Family

In this chapter we present a clear picture of the state of the art in the study of the dually flat geometries of the deformed exponential family. A dually flat space is an important tool in the geometric study of statistical estimation [12], [14]. An exponential family is an important statistical model which has a dually flat structure with respect to (± 1)-connections [12], [14]. A q-exponential family is generalization of an exponential family which is used in non-extensive statistical mechanics [26], [28]. A q-exponential family has a dually flat structure called the q-structure which is the conformal flattening of the α-geometry [27], [53].

Naudts [28] introduced a more generalized notion of exponential family called the deformed exponential family and defined a dually flat structure on it, the U-geometry. The geometry of the deformed exponential family was extensively studied by many authors [31–36] . Amari et al. [37] also studied this deformed exponential family and obtained a dually flat structure on it called the χ-geometry, which is different from the U-geometry.

In this chapter we discuss the importance of the (F, G)-geometry in the study of the dually flat geometries of the deformed exponential family. In Section 4.1 the general structure of a dually flat space is described. In Section 4.2 a short description of the dually flat geometry of the exponential family and q-exponential family are given. In Section 4.3 the two dually flat geometries, the U-geometry and the χ-geometry, on the deformed exponential family are described. Then the role of non-invariant (F, G)-geometry in the study of a deformed exponential family is presented. We show that the U-geometry is the (F, G)-geometry and χ-geometry is the conformal flattening of (F, G)-geometry for suitable choices of F and G.

4.1 Dually Flat Spaces

On a statistical manifold a divergence function always induces a unique torsion free dualistic structure. Matumoto [58] proved that every torsion-free dualistic structure is induced from a globally defined divergence. But there may be many divergence functions which induces the same dualistic structure. In the case of a dually flat space there exist a unique divergence which generates its geometric structure. Now we give a brief description of a dually flat space.

Let $\mathcal{S}$ be a statistical manifold and (g, ∇, ∇^*) be a dualistic structure on $\mathcal{S}$. Assume that the affine connection ∇ is flat. By duality ∇^* is also flat. Then (g, ∇, ∇^*) is a dually flat structure on $\mathcal{S}$ and $(\mathcal{S}, g, \nabla, \nabla^*)$ is called a **dually flat space**.

Consider a dually flat space $(\mathcal{S}, g, \nabla, \nabla^*)$. By the definition of flat connection, there exists a ∇-affine coordinate system θ for $\mathcal{S}$. Then by the duality of ∇ and ∇^*, one can choose a ∇^*-affine coordinate system η such that

$$< \partial_i, \partial^j >= \delta_{ij}, \quad \text{where} \quad \partial_i = \frac{\partial}{\partial \theta^i}, \; \partial^j = \frac{\partial}{\partial \eta_j}. \tag{4.1}$$

Let the components of the Riemannian metric g with respect to θ and η be

$$g_{ij} =< \partial_i, \partial_j > \quad \text{and} \quad g^{ij} =< \partial^i, \partial^j > \tag{4.2}$$

From Equations (4.1) and (4.2)

$$\frac{\partial \eta_j}{\partial \theta^i} = g_{ij} \quad \text{and} \quad \frac{\partial \theta^i}{\partial \eta_j} = g^{ij} \tag{4.3}$$

Since

$$\partial_i \eta_j = g_{ij} = \partial_j \eta_i \quad \text{and} \quad \partial^j \theta^i = g^{ij} = \partial^i \theta^j \tag{4.4}$$

there exist functions $\psi(\theta)$ and $\phi(\eta)$ corresponding to θ and η such that

$$\eta_i = \partial_i \psi(\theta) \quad \text{and} \quad \theta^i = \partial^i \phi(\eta) \tag{4.5}$$

Since $(\partial_i \partial_j \psi(\theta)) = (g_{ij})$ and $(\partial^i \partial^j \phi(\eta)) = (g^{ij})$ are positive definite matrices, ψ is a strictly convex function of θ and ϕ is a strictly convex function of η. Also it follows that

$$\phi(q) = \max_{p \in \mathcal{S}} \{\theta(p).\eta(q) - \psi(p)\}, \quad \forall\, q \in \mathcal{S} \tag{4.6}$$

and

$$\psi(p) = \max_{q \in \mathcal{S}} \{\theta(p).\eta(q) - \phi(q)\}, \quad \forall\, p \in \mathcal{S} \tag{4.7}$$

This is the Legendre transformation. The convex functions ψ and ϕ are called the **potential functions** corresponding to θ and η respectively.
Then the components of ∇ and ∇^* with respect to θ and η are

$$\Gamma_{ijk} =< \nabla_{\partial_i}\partial_j, \partial_k >= 0 \quad \text{and} \quad \Gamma^*_{ijk} =< \nabla^*_{\partial_i}\partial_j, \partial_k >= \partial_i\partial_j\partial_k\psi \tag{4.8}$$

$$\Gamma^{*ijk} =< \nabla^*_{\partial^i}\partial^j, \partial^k >= 0 \quad \text{and} \quad \Gamma^{ijk} =< \nabla_{\partial^i}\partial^j, \partial^k >= \partial^i\partial^j\partial^k\phi \tag{4.9}$$

Definition 4.1.1. *Let (M, g) be a Riemannian manifold and let ∇ be a flat connection on M. The pair (M, g) is a* ***Hessian structure*** *on M or (M, g, ∇) is a* ***Hessian manifold*** *if there exist a function ψ such that $g = \nabla d\psi$. Let ∇^* be the dual connection of ∇ with respect to the metric g. Then (M, g, ∇) is a Hessian manifold is equivalent to (M, g, ∇, ∇^*) is a dually flat space.*

For a dually flat space $(\mathcal{S}, g, \nabla, \nabla^*)$ there exists a unique divergence called the **canonical divergence** given by

$$D(p, q) = \psi(p) + \phi(q) - \sum \theta^i(p)\eta_i(q). \tag{4.10}$$

The canonical divergence D is also called (g, ∇)**-divergence** on $\mathcal{S}$. The dual divergence or the (g, ∇^*)-divergence is given by

$$D^*(p, q) = D(q, p) \tag{4.11}$$

See Amari and Nagaoka [14] for more details. Note that the Bregman divergence always induces a dually flat structure. For a dually flat space the canonical divergence is the Bregman divergence.
Also for a dually flat space one can have the generalized Pythagorean theorem and the

projection theorem [12], [14].

Definition 4.1.2. *Let $\mathcal{S}$ be an n-dimensional manifold and let M be an m-dimensional submanifold of $\mathcal{S}$. Let ∇ be an affine connection on $\mathcal{S}$. Then M is said to be* ***∇-autoparallel*** *if*

$$\nabla_X Y \in \Gamma(TM), \quad \forall\, X, Y \in \Gamma(TM) \tag{4.12}$$

where $\Gamma(TM)$ is the family of smooth vector fields on M.

Theorem 4.1.3. ***(Pythagorean theorem)*** *Let $(\mathcal{S}, g, \nabla, \nabla^*)$ be a dually flat space and D be the canonical divergence. Given three points $p, q, r \in \mathcal{S}$. Let γ_1 be the ∇-geodesic connecting p and q and let γ_2 be the ∇^*-geodesic connecting q and r. If the curves γ_1 and γ_2 are orthogonal with respect to g at the intersecting point q then*

$$D(p, r) = D(p, q) + D(q, r). \tag{4.13}$$

Theorem 4.1.4. ***(Projection theorem)*** *Let $(\mathcal{S}, g, \nabla, \nabla^*)$ be a dually flat space and let M be a ∇^*-autoparallel submanifold of $\mathcal{S}$. Let D be the canonical divergence of $\mathcal{S}$. Given $p \in \mathcal{S}$, a necessary and sufficient condition for a point $q \in M$ to satisfy $D(p, q) = \min_{r \in M} D(p, r)$ is that the ∇-geodesic connecting p and q is orthogonal to M at q.*

4.2 Exponential Family and q-Exponential Family

Exponential family and q-exponential family are examples of dually flat spaces. In this section we describe the dually flat structure of exponential family and q-exponential family.

4.2.1 Dually flat structure of the exponential family

Exponential family is an important class of probability distributions and most of the common distributions like normal, gamma, exponential, beta, Poisson etc. belong to the exponential class. It is known that a finite dimensional exponential family has a flat structure with respect to 1-connection defined by Amari [12].

An n-dimensional statistical model $\mathcal{S} = \{p(x;\theta) \ / \ \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$ is called an **exponential family** if

$$p(x;\theta) = \exp\{C(x) + \sum_{i=1}^{n} \theta^i C_i(x) - \psi(\theta)\} \tag{4.14}$$

where $C_1, \cdots, C_n, C$ are functions on $\mathcal{X}$ and ψ is a function on $\mathbb{E}$. By renaming random variables $C_i(x)$ as x_i, without loss of generality we can rewrite the above equation in a convenient form (usually called the standard form) with respect to a suitable dominating measure as

$$p(x;\theta) = \exp\{\sum_{i=1}^{n} \theta^i x_i - \psi(\theta)\} \quad \text{or} \quad \log(p(x;\theta)) = \sum_{i=1}^{n} \theta^i x_i - \psi(\theta) \tag{4.15}$$

where $x = (x_1, \cdots, x_n)$ is a set of random variables, $\theta = (\theta^1, \cdots, \theta^n)$ are the canonical parameters and $\psi(\theta)$ is determined from the normalization condition.

The exponential family $(\mathcal{S}, g, \nabla^1, \nabla^{(-1)})$ is a dually flat space, where g is the Fisher information metric, ∇^1 is the 1-connection (exponential connection) and ∇^{-1} is the (-1)-connection (mixture connection).

$$\begin{aligned} g_{ij}(\theta) &= \int \partial_i \ell \ \partial_j \ell \ p \ dx = \partial_i \partial_j \psi(\theta) && (4.16) \\ \Gamma^{-1}_{ijk}(\theta) &= \int (\partial_i \partial_j \ell + \partial_i \ell \ \partial_j \ell) \partial_k \ell \ p \ dx = \partial_i \partial_j \partial_k \psi(\theta) && (4.17) \\ \Gamma^{1}_{ijk}(\theta) &= 0 && (4.18) \end{aligned}$$

where $\ell(x;\theta) = \log p(x;\theta)$ and $\partial_i = \frac{\partial}{\partial \theta^i}$.
The dual coordinate η and the dual potential function $\phi(\eta)$ are

$$\eta_i = \partial_i \psi(\theta) = E(x_i) \tag{4.19}$$

$$\phi(\eta) = E_p[\log p] = -H(p) \tag{4.20}$$

where $H(p) = -\int p \log p \ dx$ is the Shannon entropy.

The (-1)-divergence D_{-1} on $\mathcal{S}$ is the Kullback-Leibler divergence given by

$$\begin{aligned} D_{-1}(p,q) &= \psi(\theta(p)) + \phi(\eta(q)) - \sum_{i=1}^{n} \theta^i(p)\eta_i(q) \qquad (4.21) \\ &= \int (\log p - \log q)\; p\; dx \qquad (4.22) \end{aligned}$$

4.2.2 q-Exponential family and the q-structure

For any $\alpha \in \mathbb{R}$, Amari [12] defined an α-family of probability density functions. $\mathcal{S} = \{p(x;\theta) \;/\; \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$ is said to be an α**-family** if

$$L_\alpha(p(x;\theta)) = \sum_{i=1}^{n} \theta^i x_i - \psi(\theta) \qquad (4.23)$$

where $L_\alpha(p)$ is the α-embedding.

When $\alpha = 1$, the α-family is the exponential family and exponential family is ∇^1-flat. But for $\alpha \neq 1$, α-family is not flat with respect to the α-connection. So how to get dually flat connections on a α-family? q-exponential family originated from the statistical physics gave an answer to this. Amari and Ohara [27] showed that a q-exponential family, which is an α-family with $\alpha = 1 - 2q$, has a dually flat structure called the q-structure. Moreover the q-geometry is the conformal flattening of α-geometry [27].

Definition 4.2.1. *[59], [60] Two statistical manifolds (M, ∇, g) and $(M, \tilde{\nabla}, \tilde{g})$ are said to be β-conformally equivalent if there exist a positive function ϕ on M such that*

$$\tilde{g}(X,Y) = \phi\; g(X,Y) \qquad (4.24)$$

$$\begin{aligned} \tilde{g}(\tilde{\nabla}_X Y, Z) = \phi\; g(\nabla_X Y, Z) + \frac{1-\beta}{2}\{g(Y,Z)d\phi(X) + g(X,Z)d\phi(Y)\} \\ -\frac{1+\beta}{2} g(X,Y)d\phi(Z) \qquad (4.25) \end{aligned}$$

In terms of the basis vectors the above expressions can be written as

$$\tilde{g}(\partial_i, \partial_j) = \tilde{g}_{ij} = \phi\; g(\partial_i, \partial_j) = \phi\; g_{ij} \qquad (4.26)$$

$$\tilde{\Gamma}^{\beta}_{ijk} = \phi\, \Gamma_{ijk} + \frac{1-\beta}{2}\{g_{jk}\partial_i\phi + g_{ik}\partial_j\phi\} - \frac{1+\beta}{2} g_{ij}\partial_k\phi \qquad (4.27)$$

Now let us describe the q-geometry of the q-exponential family.

Define the q-logarithm and its inverse the q-exponential by

$$\begin{aligned} \log_q(u) &= \frac{1}{1-q}(u^{1-q} - 1), \quad q > 0 && (4.28) \\ \exp_q(u) &= \{1 + (1-q)u\}^{\frac{1}{1-q}}, \quad u > \frac{-1}{1-q} && (4.29) \end{aligned}$$

in the limiting case $q \to 1$,

$$\log_q(u) = \log u; \quad \exp_q(u) = \exp u \qquad (4.30)$$

Definition 4.2.2. *A statistical manifold $\mathcal{S} = \{p(x;\theta) \;/\; \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$ is said to be a **q-exponential family** if*

$$\log_q p(x;\theta) = \sum_{i=1}^{n} \theta^i x_i - \psi_q(\theta) \qquad (4.31)$$

where $\psi_q(\theta)$ is obtained from the normalization $\int p(x;\theta)dx = 1$.

Define a functional

$$h_q(\theta) = \int (p(x;\theta))^q dx \qquad (4.32)$$

From the definition of the q-exponential family $\mathcal{S}$

$$\partial_i\partial_j\psi_q(\theta) = \frac{q}{h_q(\theta)} \int (x_i - \partial_i\psi_q(\theta))\, (x_j - \partial_i\psi_q(\theta))p(x;\theta)^{2q-1}\, dx \qquad (4.33)$$

Amari et al. [27] proved that ψ_q is a convex function and further assumed that it is strictly convex to define a divergence of Bregman type called the **q-divergence**,

$$\begin{aligned} D_q(p(x;\theta_1), p(x;\theta_2)) &= \psi_q(\theta_2) - \psi_q(\theta_1) - \nabla\psi_q(\theta_1).(\theta_2 - \theta_1) && (4.34) \\ &= \frac{1}{h_q(\theta)} \int (\log_q(p) - \log_q(r))\, p^q\, dx && (4.35) \end{aligned}$$

On the q-exponential family $\mathcal{S}$ the q-divergence D_q induces a dually flat structure called the q-**structure** $(g^{D_q}, \nabla^{D_q}, \nabla^{D_q^*})$ given by

$$g_{ij}^{D_q} = \partial_i \partial_j \psi_q(\theta) \tag{4.36}$$

$$\Gamma_{ijk}^{D_q} = \partial_i \partial_j \partial_k \psi_q(\theta) \tag{4.37}$$

$$\Gamma_{ijk}^{D_q^*} = 0 \tag{4.38}$$

Let

$$\tilde{D}_q(p, r) = \int (\log_q(p) - \log_q(r))\, p^q \, dx \tag{4.39}$$

Then $\tilde{D}_q$ is a constant multiple of the well known α-divergence [14] with $\alpha = 1 - 2q$. Note that

$$D_q(p, r) = \frac{1}{h_q(\theta)} \tilde{D}_q(p, r) \tag{4.40}$$

which is the conformal transformation of α-divergence, $\alpha = 1-2q$, by a gauge function $\frac{1}{h_q(\theta)}$. Hence the q-structure is the conformal flattening of the α-geometry ($\alpha = 1 - 2q$) by a gauge function $\frac{1}{h_q(\theta)}$.

Geometry induced from the conformal transformation of α-divergence

Let us now look at the geometry obtained by the conformal transformation of the α-divergence by a gauge function $K(\theta)$ [59], [60]. This geometry is (± 1)-conformally equivalent to the α-geometry. Then as a corollary we show that the q-geometry on the q-exponential family is the conformal flattening of the α-geometry.
$\mathcal{S} = \{p(x; \theta) \;/\; \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$ be a statistical manifold. The α-divergence D_α is

$$D_\alpha(p, r) = \frac{4}{1 - \alpha^2} \left[1 - \int p^{\frac{1-\alpha}{2}} r^{\frac{1+\alpha}{2}} dx\right] \tag{4.41}$$

Let $K(\theta)$ be a positive smooth function of θ. Define a divergence function D_K on $\mathcal{S}$ as

$$D_K(p(x; \theta_1), p(x; \theta_2)) = K(\theta_1) D_\alpha(p(x; \theta_1), p(x; \theta_2)) \tag{4.42}$$

Proposition 4.2.3. *The metric and the affine connection ∇^{D_K} induced by the divergence*

D_K are given by

$$g_{ij}^{D_K}(\theta) = K(\theta) g_{ij}(\theta) \tag{4.43}$$

$$\Gamma_{ijk}^{D_K}(\theta) = K(\theta)\Gamma_{ijk}^{\alpha} + \partial_i K(\theta) g_{jk} + \partial_j K(\theta) g_{ik} \tag{4.44}$$

where g is the Fisher information metric and Γ_{ijk}^{α} are the components of α-connection.

Proof. We have

$$\begin{aligned} \partial_i D_K(p,r) &= \partial_i \left[K(\theta_1) D_\alpha(p,r) \right] && (4.45) \\ &= K(\theta_1) \frac{4}{1-\alpha^2} \left[-\frac{1-\alpha}{2} \int p^{\frac{-(1+\alpha)}{2}} r^{\frac{1+\alpha}{2}} \partial_i p \, dx \right] \\ &+ \partial_i K(\theta_1) \frac{4}{1-\alpha^2} \left[1 - \int p^{\frac{1-\alpha}{2}} r^{\frac{1+\alpha}{2}} dx \right] && (4.46) \\ \partial_{j'} \partial_i D_K(p,r) &= -K(\theta_1) \left[\int p^{\frac{-(1+\alpha)}{2}} r^{\frac{\alpha-1}{2}} \partial_i p \, \partial_{j'} r \, dx \right] \\ &+ -\partial_i K(\theta_1) \frac{2}{1-\alpha} \left[\int p^{\frac{1-\alpha}{2}} r^{\frac{\alpha-1}{2}} \partial_{j'} r \, dx \right] && (4.47) \end{aligned}$$

Hence

$$\begin{aligned} g_{ij}^{D_K}(\theta) &= -\partial_i \partial_{j'} D_K(p,r) \mid_{p=r} && (4.48) \\ &= K(\theta) \int \partial_j \ell \, \partial_i \ell \, p \, dx && (4.49) \\ &= K(\theta) g_{ij}(\theta) && (4.50) \end{aligned}$$

Also

$$\begin{aligned} \partial_j \partial_i D_K(p,r) &= -K(\theta_1) \frac{4}{1-\alpha^2} \left[\frac{1-\alpha}{2} \int p^{\frac{-(1+\alpha)}{2}} r^{\frac{1+\alpha}{2}} \partial_i \partial_j p \;\; dx \right] \\ &+ -K(\theta_1) \frac{4}{1-\alpha^2} \left[\frac{1-\alpha}{2} \int p^{\frac{-(3+\alpha)}{2}} r^{\frac{1+\alpha}{2}} \partial_i p \, \partial_j p \;\; dx \right] \\ &+ -\partial_j K(\theta_1) \frac{4}{1-\alpha^2} \left[\int \frac{1-\alpha}{2} p^{\frac{-(1+\alpha)}{2}} r^{\frac{1+\alpha}{2}} \partial_i p \, dx \right] \\ &+ \frac{4}{1-\alpha^2} \left[1 - \int p^{\frac{1-\alpha}{2}} r^{\frac{1+\alpha}{2}} dx \right] \partial_i \partial_j K(\theta_1) \\ &+ -\partial_i K(\theta_1) \frac{4}{1-\alpha^2} \left[\int \frac{1-\alpha}{2} p^{\frac{-(1+\alpha)}{2}} r^{\frac{1+\alpha}{2}} \partial_j p \, dx \right] && (4.51) \end{aligned}$$

$$
\begin{aligned}
\partial_{k'}\partial_j\partial_i D_K(p,r) &= -K(\theta_1)\left[\int p^{\frac{-(1+\alpha)}{2}} r^{\frac{\alpha-1}{2}} \partial_i\partial_j p \, \partial_{k'} r \, dx\right] \\
&+ K(\theta_1)\left[\frac{1+\alpha}{2}\int p^{\frac{-(3+\alpha)}{2}} r^{\frac{\alpha-1}{2}} \partial_i p \, \partial_j p \, \partial_{k'} r \, dx\right] \\
&+ -\partial_j K(\theta_1)\left[\int p^{\frac{-(1+\alpha)}{2}} r^{\frac{\alpha-1}{2}} \partial_i p \, \partial_{k'} r \, dx\right] \\
&+ -\frac{2}{1-\alpha}\left[\int p^{\frac{1-\alpha}{2}} r^{\frac{\alpha-1}{2}} \partial_{k'} r \, dx\right] \partial_i\partial_j K(\theta_1) \\
&+ -\partial_i K(\theta_1)\left[\int p^{\frac{-(1+\alpha)}{2}} r^{\frac{\alpha-1}{2}} \partial_j p \, \partial_{k'} r \, dx\right] \qquad (4.52)
\end{aligned}
$$

Hence

$$
\begin{aligned}
\Gamma_{ijk}^{D_K}(\theta) &= -\partial_i\partial_j\partial_{k'} D_K(p,r) \mid_{p=r} \qquad (4.53) \\
&= K(\theta)\left[\int \partial_i\partial_j \ell \; + \frac{1-\alpha}{2}\partial_i \ell \, \partial_j \ell \right] \partial_k \ell \, p \, dx \\
&+ \partial_i K(\theta) g_{jk} + \partial_j K(\theta) g_{ik} \qquad (4.54) \\
&= K(\theta)\Gamma_{ijk}^{\alpha} + \partial_i K(\theta) g_{jk} + \partial_j K(\theta) g_{ik} \qquad (4.55)
\end{aligned}
$$

where Γ_{ijk}^{α} are the components of the α-connection. □

Proposition 4.2.4. *The affine connection $\nabla^{D_K^*}$ induced by the dual D_K^* of the divergence D_K is given by*

$$\Gamma_{ijk}^{D_K^*}(\theta) = K(\theta)\Gamma_{ijk}^{-\alpha} - \partial_k K(\theta) g_{ij} \tag{4.56}$$

where $\Gamma_{ijk}^{-\alpha}$ are the components of the $(-\alpha)$-connection.

Proof.

$$D_K^*(p,r) = D_K(r,p) = K(\theta_2)\frac{4}{1-\alpha^2}\left[1 - \int r^{\frac{1-\alpha}{2}} p^{\frac{1+\alpha}{2}} dx\right] \tag{4.57}$$

$$
\begin{aligned}
\partial_i D_K^*(p,r) &= \partial_i\left(K(\theta_2)\frac{4}{1-\alpha^2}\left[1 - \int r^{\frac{1-\alpha}{2}} p^{\frac{1+\alpha}{2}}\right] dx\right) \qquad (4.58) \\
&= -K(\theta_2)\frac{4}{1-\alpha^2}\left[\frac{1+\alpha}{2}\int p^{\frac{\alpha-1}{2}} r^{\frac{1-\alpha}{2}} \partial_i p \, dx\right] \qquad (4.59)
\end{aligned}
$$

$$
\begin{aligned}
\partial_j\partial_i D_K^*(p,r) &= -K(\theta_2)\frac{4}{1-\alpha^2}\left[\frac{1+\alpha}{2}\int p^{\frac{\alpha-1}{2}} r^{\frac{1-\alpha}{2}} \partial_i\partial_j p \;\; dx\right] \\
&+ K(\theta_2)\left[\int p^{\frac{\alpha-3}{2}} r^{\frac{1-\alpha}{2}} \partial_i p \, \partial_j p \;\; dx\right] \qquad (4.60)
\end{aligned}
$$

$$\begin{aligned}\partial_{k'}\partial_j\partial_i D_K^*(p,r) &= -K(\theta_2)\left[\int p^{\frac{\alpha-1}{2}} r^{\frac{-(1+\alpha)}{2}} \partial_i\partial_j p \, \partial_{k'} r \, dx\right] \\ &+ -\partial_k K(\theta_2)\left[\frac{2}{1-\alpha}\int p^{\frac{\alpha-1}{2}} r^{\frac{1-\alpha}{2}} \partial_i\partial_j p \, dx\right] \\ &+ K(\theta_2)\left[\frac{1-\alpha}{2}\int p^{\frac{\alpha-3}{2}} r^{\frac{-(1+\alpha)}{2}} \partial_i p \, \partial_j p \, \partial_{k'} r \, dx\right] \\ &+ \partial_k K(\theta_2)\left[\int p^{\frac{\alpha-3}{2}} r^{\frac{1-\alpha}{2}} \partial_i p \, \partial_j p \, dx\right]\end{aligned} \tag{4.61}$$

Hence

$$\begin{align} \Gamma_{ijk}^{D_K^*}(\theta) &= -\partial_i\partial_j\partial_{k'} D_K^*(p,r)\mid_{p=r} \tag{4.62} \\ &= K(\theta)\left[\int \partial_i\partial_j\ell \; + \frac{1+\alpha}{2}\partial_i\ell \, \partial_j\ell\right]\partial_k\ell \, p \, dx - \partial_k K(\theta) g_{ij} \tag{4.63} \\ &= K(\theta)\Gamma_{ijk}^{-\alpha} + \partial_k K(\theta) g_{ij} \tag{4.64} \end{align}$$

where $\Gamma_{ijk}^{-\alpha}$ are the components of the $(-\alpha)$-connection. □

In summary, we proved the following theorem.

Theorem 4.2.5. *$(\mathcal{S}, g, \nabla^{\alpha})$ and $(\mathcal{S}, g^{D_K}, \nabla^{D_K})$ are (-1)-conformally equivalent, where g is the Fisher information metric and ∇^{α} is the α-connection. Also $(\mathcal{S}, g, \nabla^{-\alpha})$ and $(\mathcal{S}, g^{D_K}, \nabla^{D_K^*})$ are 1-conformally equivalent.*

Corollary 4.2.6. *The q-geometry on the q-exponential family is the conformal flattening of the α-geometry by a gauge function $K(\theta) = \frac{q}{h_q(\theta)}$.*

Proof. Take $K(\theta) = \frac{q}{h_q(\theta)}$, then the divergence D_K reduces to q-divergence. Then for the q-exponential family from Equations (4.36) and (4.43)

$$g_{ij}^{D_K}(\theta) = \partial_i\partial_j\psi_q(\theta) = \frac{q}{h_q(\theta)} g_{ij}(\theta) \tag{4.65}$$

Let $\alpha = 1 - 2q$. Then

$$K(\theta)\Gamma_{ijk}^{1-2q} = \frac{q}{h_q(\theta)}\left(\int \partial_i\partial_j\ell \, \partial_k\ell \, p \, dx + q\int \partial_i\ell \, \partial_j\ell \, \partial_k\ell \, p \, dx\right) \tag{4.66}$$

$$\begin{aligned}\partial_i K(\theta) g_{jk} + \partial_j K(\theta) g_{ik} &= \frac{q}{h_q(\theta)}\int \partial_k\partial_j\ell \, \partial_i\ell \, p \, dx + \int \partial_i\partial_k\ell \, \partial_j\ell \, p \, dx \\ &+ (2-2q)\int \partial_i\ell \, \partial_j\ell \, \partial_k\ell \, p \, dx\end{aligned} \tag{4.67}$$

Now from Equation (4.33)

$$\begin{aligned}\partial_i\partial_j\partial_k\psi_q(\theta) &= \frac{q}{h_q(\theta)}\left(\int \partial_i\partial_j\ell\, \partial_k\ell\, p\, dx + (2-q)\int \partial_i\ell\, \partial_j\ell\, \partial_k\ell\, p\, dx\right) \\ &+ \frac{q}{h_q(\theta)}\left(\int \partial_k\partial_j\ell\, \partial_i\ell\, p\, dx + \int \partial_i\partial_k\ell\, \partial_j\ell\, p\, dx\right) \qquad (4.68)\end{aligned}$$

Hence from Equations (4.44) and (4.68)

$$\Gamma^{D_K}_{ijk}(\theta) = \partial_i\partial_j\partial_k\psi_q(\theta) \qquad (4.69)$$

$$(4.70)$$

Hence the result.

A similar proof holds for $\nabla^{D^*_K}$. □

4.3 Dually Flat Geometries on a Deformed Exponential Family

Naudts [28] introduced a generalized notion of exponential family called the deformed exponential family. This is done by replacing the exponential function in the standard exponential family by a deformed exponential function $\exp_\phi$, where ϕ is an increasing positive function on $[0, \infty)$. This deformed exponential family is called the ϕ-exponential family. He defined a ϕ-logarithm by

$$\ln_\phi(u) = \int_1^u \frac{1}{\phi(v)} dv, \quad u > 0 \qquad (4.71)$$

$\exp_\phi$ is the inverse of the ϕ-logarithm. For $\phi(u) = u$, $\ln_\phi(u) = \log(u)$ and $\exp_\phi(u) = \exp u$. For $\phi(u) = u^q$ with $q > 0$

$$\ln_\phi(u) = \begin{cases} \frac{u^{1-q}-1}{1-q}, & q \neq 1 \\ \log u, & q = 1 \end{cases} \qquad (4.72)$$

which is the q-logarithm. A family of probability distributions $\mathcal{S} = \{p(x;\theta)\}$ is said to be **a ϕ-exponential family** if

$$p(x;\theta) = \exp_{\phi}(\psi(\theta) - \sum_{i=1}^{n} \theta^i x_i) \tag{4.73}$$

On a ϕ-exponential family Naudts [28] defined a dually flat structure, the **U-geometry**, using a divergence function. Amari et al. [37] also considered this deformed exponential family formulated as χ-family and defined a dually flat structure called the **χ-geometry** using an escort probability distribution. The κ-exponential family by Kaniadakis et al. [37], U-model by Eguchi et al. [36] are similar formulations of this deformed exponential family.

In this section we describe the two dually flat structures on the deformed exponential family, the U-geometry and the χ-geometry. Then we show that how these two dually flat structures are related to the (F, G)-geometry.

For the sake of notational convenience, deformed exponential family is formulated using the function F and we call it as F-exponential family.

Remark 4.3.1. *Note that ϕ-exponential family, χ-family and F-exponential family are essentially the same family of probability distributions with a generic name deformed exponential family. We may interchangeably use the term deformed exponential family or F-exponential family.*

Definition 4.3.2. *Let $F : (0, \infty) \longrightarrow \mathbb{R}$ be any smooth function satisfying $F'(x) > 0$ and $F''(x) < 0$. Let Z be the inverse function of F. Define the standard form of an n-dimensional F**-exponential family** $\mathcal{S} = \{p(x;\theta)\}$ of probability distributions as*

$$p(x;\theta) = Z(\sum_{i=1}^{n} \theta^i x_i - \psi_F(\theta)) \quad or \quad F(p(x;\theta)) = \sum_{i=1}^{n} \theta^i x_i - \psi_F(\theta) \tag{4.74}$$

where $x = (x_1, \cdots, x_n)$ is a set of random variables, $\theta = (\theta^1, \cdots, \theta^n)$ are the parameters and $\psi_F(\theta)$ is determined from the normalization condition.

Remark 4.3.3. *When $F(p) = \log p$ the F-exponential family is the exponential family and when $F(p) = \log_q p$ the F-exponential family is the q-exponential family.*

4.3.1 Dually flat U-geometry of the F-exponential family

In this section we describe the dually flat U-geometry on the F exponential family [28, 36]. Further we show that the U-geometry on the F-exponential family is the (F, G)-geometry for suitable choices of F and G [38].

Let $\mathcal{S} = \{p(x;\theta) \ / \ \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$ be an n-dimensional F-exponential family. The divergence of Bregman type given by Naudts [28] is

$$D^F(p,q) = \int \left(\int_q^p (F(u) - F(q))du \right) dx \tag{4.75}$$

This divergence is a U-divergence defined by Murata et al. [22], where U is an increasing convex function and

$$D^F(p,q) = D_U(p,q) = \int \left[U^*(p) - p\xi(q) + U(\xi(q)) \right] dx \tag{4.76}$$

with $U^*(t) = \int_1^t F(u)\, du$ and $\xi(t) = \frac{dU^*}{dt}(t)$.

Now consider the dualistic structure $(g^{D^F}, \nabla^{D^F}, \nabla^{D^{*F}})$ induced from the divergence D^F called the U**-geometry** (Eguchi et al. [36]).

$$\begin{aligned} g_{ij}^{D^F}(\theta) &= \int \partial_i p \ \partial_j F(p)\, dx & (4.77) \\ \Gamma_{ijk}^{D^F}(\theta) &= \int \partial_k p \ \partial_i \partial_j F(p)\, dx & (4.78) \\ \Gamma_{ijk}^{D^{*F}}(\theta) &= \int \partial_i \partial_j p \ \partial_k F(p)\, dx & (4.79) \end{aligned}$$

From the definition of $\mathcal{S}$, $\Gamma_{ijk}^{D^F}(\theta) = 0$. Hence the connection ∇^{D^F} is flat. Moreover $(g^{D^F}, \nabla^{D^F}, \nabla^{D^{*F}})$ is a dually flat structure on $\mathcal{S}$.

The dual coordinate (η_i) of the canonical coordinate θ^i are

$$\eta_i = E_p[x_i] = \int x_i \ p(x;\theta)\, dx. \tag{4.80}$$

Define a function

$$v(t) = \int_1^t F(s)\, ds \quad t > 0. \tag{4.81}$$

Assume that $v(o) := \lim_{t \to +0} v(t)$ is finite.

The generalized entropy functional I and generalized Massieu potential Ψ are defined as

$$\begin{aligned} I(p_\theta) &:= -\int [v(p(x;\theta)) + (p(x;\theta)-1)v(o)]\,dx. && (4.82)\\ \Psi_F(\theta) &:= \int p(x;\theta)F(p(x;\theta))\,dx + I(p_\theta) + \psi_F(\theta). && (4.83)\end{aligned}$$

Note that Ψ_F is the potential function corresponding to the canonical coordinate θ and $\eta_i = E_p[x_i] = \partial_i \Psi_F(\theta)$. The dual potential function Φ of the dual coordinate η is given by

$$\Phi(\eta) = -I(p_\theta). \tag{4.84}$$

See [28], [34], for more details.

Next to show that the U-geometry is the (F, G)-geometry for suitable choices of F and G.

Theorem 4.3.4. *For the F-exponential family $\mathcal{S}$ the dually flat U-geometry obtained from the U-divergence is the (F, G)-geometry $(g^G, \nabla^{F,G}, \nabla^{H,G})$ with $G(p) = pF'(p)$ and H is the G-dual embedding of F given by $H(p) = p$.*

Proof. For the F-exponential family $\mathcal{S}$

$$\begin{aligned} \partial_i F &= p\,F'(p)\,\partial_i \ell && (4.85)\\ \partial_i\partial_j F &= p\,F'(p)\,\partial_i\partial_j\ell + [pF'(p) + p^2F''(p)]\,\partial_i\ell\,\partial_j\ell. && (4.86)\end{aligned}$$

Then the Equations (4.77), (4.78) and (4.79) can be written as

$$\begin{aligned} g_{ij}^{D^F}(\theta) &= \int \partial_i p\,\partial_j F(p)\,dx && (4.87)\\ &= \int pF'(p)\ \ \partial_i\ell\,\partial_j\ell\,p\,dx && (4.88)\\ &= g^G(\theta) && (4.89)\end{aligned}$$

which is the G-metric with $G(p) = pF'(p)$ from Equation (2.106).

$$\begin{aligned}
\Gamma_{ijk}^{D^F}(\theta) &= \int \partial_k p \, \partial_i \partial_j F(p) \, dx && (4.90)\\
&= \int \left(\partial_i \partial_j \ell + (1 + \frac{pF''(p)}{F'(p)}) \partial_i \ell \, \partial_j \ell \right) \partial_k \ell \, pF'(p) \, p \, dx && (4.91)\\
&= \Gamma_{ijk}^{F,G}(\theta) && (4.92)
\end{aligned}$$

which is the (F, G)-connection with $G(p) = pF'(p)$ from Equation (2.110)

$$\begin{aligned}
\Gamma_{ijk}^{D^{*F}}(\theta) &= \int \partial_i \partial_j p \, \partial_k F(p) \, dx && (4.93)\\
&= \int (\partial_i \partial_j \ell + \partial_i \ell \, \partial_j \ell) \, \partial_k \ell \, pF'(p) \, p \, dx && (4.94)\\
&= \Gamma_{ijk}^{H,G}(\theta) && (4.95)
\end{aligned}$$

From Equation (2.122), this is the (H, G)-connection, where $G(p) = pF'(p)$, H is the G-dual embedding of F

$$1 + \frac{pH''(p)}{H'(p)} = 1 \Rightarrow \quad H(p) = p. \qquad (4.96)$$

Hence the U-geometry induced from the divergence D^F is the (F, G)-geometry for suitable choices of F and G. □

Remark 4.3.5. *Hence on F-exponential family the dually flat U-geometry obtained from the U-Bregman divergence D^F is the (F, G)-geometry for suitable choices of F and G. The U-geometry on a q-exponential family is the (F, G)-geometry, where $F(p) = \log_q p$ and $G(p) = p^{1-q}$ and $H(p) = p$. Thus*

$$\begin{aligned}
g_{ij}^{D^F}(\theta) &= \int \partial_i \ell \, \partial_j \ell \, p^{2-q} \, dx && (4.97)\\
\Gamma_{ijk}^{D^F}(\theta) &= \int (\partial_i \partial_j \ell + (1-q) \partial_i \ell \, \partial_j \ell) \, \partial_k \ell \, p^{2-q} \, dx && (4.98)\\
\Gamma_{ijk}^{D^{*F}}(\theta) &= \int (\partial_i \partial_j \ell + \partial_i \ell \, \partial_j \ell) \, \partial_k \ell \, p^{2-q} \, dx && (4.99)
\end{aligned}$$

In summary

Theorem 4.3.6. *For a F-exponential family $\mathcal{S}$*

1. *(g^{D_F}, ∇^{D_F}) and $(g^{D_F}, \nabla^{D_F^*})$ are mutually dual Hessian structures on $\mathcal{S}$ equiva-*

lently, $(\mathcal{S}, g^{D_F}, \nabla^{D_F}, \nabla^{D_F^*})$ is a dually flat space.

2. *The canonical coordinate θ is $\nabla^{D_F^*}$-affine and ψ_F is the potential function corresponding to θ.*

3. *The metric $g_{ij}^{D_F}(\theta) = \partial_i \partial_j \psi_F(\theta)$.*

4. *The dual coordinate η is $\eta_i = \partial_i \psi_F(\theta) = E_{\hat{p}_F}[x_i]$ and it is ∇^{D_F}-affine.*

5. *The dual potential function ϕ_F corresponding to the dual coordinate η is $\phi_F(\eta) = E_{\hat{p}_F}(F(p))$.*

6. *$(g^{D_F}, \nabla^{D_F}, \nabla^{D_F^*})$ is the (F, G)-geometry $(g^G, \nabla^{F,G}, \nabla^{H,G})$ with $G(p) = pF'(p)$ and $H(p) = p$.*

4.3.2 Dually flat χ-geometry of the deformed exponential family

Amari et al. [37] also considered deformed exponential family called the χ-exponential family and defined a dually flat geometry called the χ-geometry. This χ-geometry is different from the U-geometry given by Naudts. In this section we show that this dually flat χ-geometry is the conformal flattening of the (F, G)-geometry for suitable F and $G(p) = \frac{-pF''(p)}{F'(p)}$. Here also we follow our previous formulation of the deformed exponential family called the F-exponential family.

χ-Geometry of the F-exponential family

Let $\mathcal{S} = \{p(x;\theta) \ / \ \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$ be a F-exponential family. Then

$$\partial_i F(p(x;\theta)) = x_i - \partial_i \psi_F(\theta) \tag{4.100}$$

$$\partial_i \partial_j F(p(x;\theta)) = -\partial_i \partial_j \psi_F(\theta) \tag{4.101}$$

Define a functional $h_F(\theta)$ as

$$h_F(\theta) = \int \frac{1}{F'(p(x;\theta))} dx \tag{4.102}$$

Theorem 4.3.7. *F-potential function $\psi(\theta)$ is a convex function of θ.*

Proof. We have

$$\partial_i F(p(x;\theta)) = F'(p)\partial_i p(x;\theta) \tag{4.103}$$

From Equation (4.100)

$$\partial_i p = \frac{1}{F'(p)}\partial_i F = \frac{1}{F'(p)}(x_i - \partial_i \psi_F(\theta)) \tag{4.104}$$

By differentiating Equation (4.103) with respect to θ^j

$$\partial_i\partial_j F = F'(p)\partial_i\partial_j p + \frac{F''(p)}{(F'(p))^2}\partial_i F \partial_j F \tag{4.105}$$

Hence

$$\partial_i\partial_j p = \frac{1}{F'(p)}\partial_i\partial_j F - \frac{F''(p)}{(F'(p))^3}\partial_i F \partial_j F \tag{4.106}$$

$$\partial_i\partial_j p = \frac{-\partial_i\partial_j\psi_F(\theta)}{F'(p)} - \frac{F''(p)}{(F'(p))^3}\left(x_i - \partial_i\psi_F(\theta)\right)\left(x_j - \partial_j\psi_F(\theta)\right) \tag{4.107}$$

Since $\int \partial_i p dx = 0$, we have $\int \partial_i\partial_j p dx = 0$. Hence from Equations (4.104) and (4.107)

$$\partial_i\psi_F(\theta) = \frac{1}{h_F(\theta)}\int x_i \frac{1}{F'(p)} dx \tag{4.108}$$

$$\partial_i\partial_j\psi_F(\theta) = \frac{1}{h_F(\theta)}\int \frac{-F''(p)}{(F'(p))^3}\left(x_i - \partial_i\psi_F(\theta)\right)\left(x_j - \partial_j\psi_F(\theta)\right) \tag{4.109}$$

Since F is a concave function $F''(p) < 0$. Thus from Equation (4.109) it follows that $\partial_i\partial_j\psi_F(\theta)$ is positive semidefinite. Hence ψ_F is a convex function of θ. □

Note 4.3.8. *Note that $\partial_i\partial_j\psi_F(\theta)$ in Equation* (4.109) *is positive semidefinite. Further we assume that it is positive definite. Then $\psi_F(\theta)$ is a strictly convex function of θ.*

Definition 4.3.9. *For a probability distribution p parametrized by θ define a probability distribution*

$$\hat{p}_F(x) = \frac{1}{h_F(\theta)F'(p)}, \quad \text{where} \quad h_F(\theta) = \int \frac{1}{F'(p)} dx \tag{4.110}$$

*called the F**-escort probability distribution** related to* p*, see [28], [62] for more details.*

Definition 4.3.10. *Using the escort probability distribution $\hat{p}_F$, the $\hat{F}$-expectation of a random variable is defined as*

$$E_{\hat{p}_F}(f(x)) = \frac{1}{h_F(\theta)} \int \frac{1}{F'(p)} f(x) dx \qquad (4.111)$$

Now using the strictly convex function $\psi_F(\theta)$ define a divergence function which induces a dually flat structure on $\mathcal{S}$.

Definition 4.3.11. *A divergence of Bregman type (χ-divergence in [37]) is defined using $\psi_F(\theta)$ as*

$$D_F(p(x;\theta_1), p(x;\theta_2)) = \psi_F(\theta_2) - \psi_F(\theta_1) - \nabla\psi_F(\theta_1).(\theta_2 - \theta_1) \qquad (4.112)$$

Take two distributions p and r which are parametrized by θ_1 and θ_2 respectively. Then the divergence D_F can be rewritten as

$$\begin{aligned} D_F(p, r) &= \frac{1}{h_F(\theta_1)} \int (F(p) - F(r)) \frac{1}{F'(p)} dx && (4.113) \\ &= E_{\hat{p}_F}(F(p) - F(r)) && (4.114) \end{aligned}$$

Amari et al. [37] showed that the divergence D_F induces a dually flat structure on the F-exponential family.

Theorem 4.3.12. *The metric $g_{ij}^{D_F}$ and the affine connection ∇^{D_F} induced by the divergence D_F are given by*

$$g_{ij}^{D_F}(\theta) = \partial_i \partial_j \psi_F(\theta); \quad \Gamma_{ijk}^{D_F} = \partial_i \partial_j \partial_k \psi_F(\theta). \qquad (4.115)$$

The dual D_F^* of D_F induces an affine connection $\nabla^{D_F^*}$ defined by $\Gamma_{ijk}^{D_F^*} = 0$.

The dual coordinate η is given by

$$\eta_i = \partial_i \psi_F(\theta) = E_{\hat{p}_F}(x_i) \qquad (4.116)$$

Lemma 4.3.13. *The dual potential function $\phi_F(\eta)$ is given by*

$$\phi_F(\eta) = E_{\hat{p}_F}(F(p)) = \frac{1}{h_F(\theta)} \int \frac{F(p)}{F'(p)} dx \tag{4.117}$$

Proof. The dual potential function satisfies

$$\phi_F(\eta) + \psi_F(\theta) - \theta.\eta = 0 \tag{4.118}$$

Hence

$$\begin{aligned}
\phi_F(\eta) &= \theta.\eta - \psi_F(\theta) && (4.119)\\
&= \sum_{i=1}^{n} \theta^i \partial_i \psi(\theta) - \psi_F(\theta) && (4.120)\\
&= \sum_{i=1}^{n} \theta^i \frac{1}{h_F(\theta)} \int x_i \frac{1}{F'(p)} dx - \frac{\psi_F(\theta)}{h_F(\theta)} \int \frac{1}{F'(p)} dx && (4.121)\\
&= \frac{1}{h_F(\theta)} \int (\sum_{i=1}^{n} \theta^i x_i - \psi_F(\theta)) \frac{1}{F'(p)} dx && (4.122)\\
&= \frac{1}{h_F(\theta)} \int \frac{F(p)}{F'(p)} dx && (4.123)\\
&= E_{\hat{p}_F}(F(p)) && (4.124)
\end{aligned}$$

The potential function ψ_F of the canonical parameter (θ^i) is a generalized free energy called the F**-free energy**. The negative of the Legendre dual of the F-free energy

$$H_F(p) = -E_{\hat{p}_F}(F(p)) = \frac{1}{h_F(\theta)} \int \frac{-F(p)}{F'(p)} dx \tag{4.125}$$

is a generalized notion of entropy called the F**-entropy** (χ-entropy in [37]). □

Remark 4.3.14. *On the F-exponential family $\mathcal{S}$ the divergence D_F induces a dually flat structure $(g^{D_F}, \nabla^{D_F}, \nabla^{D_F^*})$ which is the χ-geometry defined by Amari et al. [37].*

Example 4.3.15. Consider a finite set $\mathcal{X} = \{x_0, \cdots x_n\}$. In Chapter 2 we proved that the set $\mathcal{P}(\mathcal{X})$ of all probability distributions defined on $\mathcal{X}$ is an n-dimensional F-exponential family (F-family) for any F. Letting $p_i = p(x = x_i)$, any $p(x) \in \mathcal{P}(\mathcal{X})$ can be written as

$$p(x) = \sum_{i=0}^{n} p_i \delta_i(x), \quad \text{where} \quad \delta_0(x) = 1 - \sum_{i=1}^{n} \delta_i(x),\ p_0 = 1 - \sum_{i=1}^{n} p_i \tag{4.126}$$

Then

$$\begin{aligned} F(p(x)) &= \sum_{i=0}^{n} F(p_i)\delta_i(x) && (4.127)\\ &= \sum_{i=1}^{n}(F(p_i) - F(p_0))\delta_i(x) + F(p_0) && (4.128)\\ &= \sum_{i=1}^{n} \theta^i x_i - \psi_F(\theta) && (4.129)\end{aligned}$$

where the canonical coordinate $\theta^i = F(p_i) - F(p_0)$, $x_i = \delta_i(x)$ and the F-free energy $\psi_F(\theta) = -F(p_0)$.

$$\frac{1}{h_F(\theta)} = \sum_{x\in\mathcal{X}} \frac{1}{F'(p(x))} = \sum_{i=0}^{n} \frac{1}{F'(p_i)} \qquad (4.130)$$

The dual coordinate η and the dual potential function $\phi(\eta)$ are

$$\begin{aligned} \eta_i &= \frac{1}{h_F(\theta)}\frac{1}{F'(p_i)} && (4.131)\\ \phi(\eta) &= \frac{1}{h_F(\theta)}\sum_{i=0}^{n}\frac{F(p_i)}{F'(p_i)} = -H_F(p) && (4.132)\end{aligned}$$

χ-Geometry as a conformal flattening of the (F, G)-geometry

Here we show that the dually flat χ-geometry on a deformed exponential family is the conformal flattening of the (F, G)-geometry for suitable choices of F and G [38], [61], [63]. Matsuzoe and Henmi [34] described the conformal equivalence of the generalized Fisher information metrics on a deformed exponential family.

Next to show that on the F-exponential family $\mathcal{S} = \{p(x;\theta)\ /\ \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$ the metric g^{D_F} induced from the divergence D_F is a conformal transformation of the G-metric g^G.

Theorem 4.3.16. *The metric g^{D_F} induced by the divergence D_F is the conformal transformation of the G-metric g^G, with $G(p) = \frac{-pF''(p)}{F'(p)}$, by a gauge function $K(\theta) = \frac{1}{h_F(\theta)}$. That is,*

$$g_{ij}^{D_F}(\theta) = K(\theta) g_{ij}^G \qquad (4.133)$$

where $G(p) = \frac{-pF''(p)}{F'(p)}$ and $K(\theta) = \frac{1}{h_F(\theta)}$.

Proof. The metric is

$$\begin{aligned} g_{ij}^{D_F}(\theta) &= \partial_i\partial_j\psi_F(\theta) && (4.134)\\ &= \frac{1}{h_F(\theta)}\int \frac{-F''(p)}{(F'(p))^3}\,\partial_i F\,\partial_j F dx && (4.135)\\ &= \frac{1}{h_F(\theta)}\int \frac{-pF''(p)}{F'(p)}\,\partial_i p\,\partial_j p\,\frac{1}{p}dx && (4.136)\end{aligned}$$

The term $\int \frac{-pF''(p)}{F'(p)}\ \partial_i p\ \partial_j p\ \frac{1}{p}dx$ is actually G-metric with $G(p) = \frac{-pF''(p)}{F'(p)}$. Thus $g_{ij}^{D_F}(\theta)$ can be written as

$$g_{ij}^{D_F}(\theta) = K(\theta)g_{ij}^G \qquad (4.137)$$

with $K(\theta) = \frac{1}{h_F(\theta)}$ and $G(p) = \frac{-pF''(p)}{F'(p)}$. Thus the new metric is obtained as a conformal transformation of the G-metric by a gauge function $K(\theta)$. □

Next to show that the connection ∇^{D_F} induced by the divergence D_F is the (-1)-conformal transformation of the (H,G)-connection $\nabla^{H,G}$.

Theorem 4.3.17. *The affine connection ∇^{D_F} induced by the divergence D_F is the (-1)-conformal transformation of the (H,G)-connection $\nabla^{H,G}$ by the gauge function $K(\theta) = \frac{1}{h_F(\theta)}$, where $G(p) = \frac{-pF''(p)}{F'(p)}$ and H is the G-dual embedding of F. That is,*

$$\Gamma_{ijk}^{D_F} = K(\theta)\Gamma_{ijk}^{H,G} + \partial_j K(\theta) g_{ik}^G(\theta) + \partial_i K(\theta) g_{jk}^G(\theta) \qquad (4.138)$$

with $G(p) = \frac{-pF''(p)}{F'(p)}$ and $K(\theta) = \frac{1}{h_F(\theta)}$.

Proof. The components $\Gamma_{ijk}^{D_F}$ of the connection are

$$\begin{aligned} \Gamma_{ijk}^{D_F} &= \partial_i\partial_j\partial_k\psi_F(\theta) && (4.139)\\ &= \frac{1}{h_F(\theta)}\int\left(\frac{-pF''(p)}{F'(p)} - \frac{p^2F'''(p)}{F'(p)} + \frac{2p^2(F''(p))^2}{(F'(p))^2}\right)\partial_i\ell\,\partial_j\ell\,\partial_k\ell\, pdx \\ &+ \frac{1}{h_F(\theta)}\int(\frac{-pF''(p)}{F'(p)})\partial_i\partial_j\ell\,\partial_k\ell\, pdx \\ &+ \frac{1}{h_F(\theta)}\int \partial_j\partial_k\psi_F(\theta)\frac{pF''(p)}{(F'(p))^2}\,\partial_i\ell\, dx \\ &+ \frac{1}{h_F(\theta)}\int \partial_i\partial_k\psi_F(\theta)\frac{pF''(p)}{(F'(p))^2}\,\partial_j\ell\, dx && (4.140)\end{aligned}$$

For any F-embedding the G-dual embedding H of F is

$$H'(p) = \frac{G(p)}{pF'(p)} \tag{4.141}$$

Then the term

$$1 + \frac{pH''(p)}{H'(p)} = \frac{pG'(p)}{G(p)} - \frac{pF''(p)}{F'(p)} \tag{4.142}$$

When $G(p) = \frac{-pF''(p)}{F'(p)}$, the above term reduces

$$1 + \frac{pH''(p)}{H'(p)} = 1 - \frac{2pF''(p)}{F'(p)} + \frac{pF'''(p)}{F'(p)} \tag{4.143}$$

Then the components of the connection $\nabla^{(H,G)}$ are

$$\begin{aligned}
\Gamma^{(H,G)}_{ijk}(\theta) &= \int \left[\partial_i\partial_j\ell \, \partial_k\ell \, + (1 + \frac{pH''(p)}{H'(p)})\partial_i\ell \, \partial_j\ell \, \partial_k\ell \right] G(p) \, p \, dx \qquad (4.144)\\
&= \frac{1}{h_F(\theta)} \int \left(\frac{-pF''(p)}{F'(p)} - \frac{p^2F'''(p)}{F'(p)} + \frac{2p^2(F''(p))^2}{(F'(p))^2} \right) \partial_i\ell \, \partial_j\ell \, \partial_k\ell \, pdx \\
&+ \frac{1}{h_F(\theta)} \int (\frac{-pF''(p)}{F'(p)}) \partial_i\partial_j\ell \, \partial_k\ell \, pdx \qquad (4.145)
\end{aligned}$$

Now for $K(\theta) = \frac{1}{h_F(\theta)}$ and $G(p) = \frac{-pF''(p)}{F'(p)}$,

$$\partial_i K(\theta) g^G_{jk}(\theta) = \frac{-1}{(h_F(\theta))^2} \left(\int \frac{pF''(p)}{(F'(p))^2} \, \partial_i\ell \, dx \right) \int \frac{pF''(p)}{F'(p)} \partial_j\ell \, \partial_k\ell \, p \, dx \tag{4.146}$$

Then the components of the connection ∇^{D_F} can be rewritten as

$$\begin{aligned}
\Gamma^{D_F}_{ijk}(\theta) &= K(\theta) \int \left[\partial_i\partial_j\ell \, \partial_k\ell \, + (1 + \frac{pH''(p)}{H'(p)})\partial_i\ell \, \partial_j\ell \, \partial_k\ell \right] G(p) \, p \, dx \\
&+ \partial_j K(\theta) g^G_{ik}(\theta) + \partial_i K(\theta) g^G_{jk}(\theta) \qquad (4.147)\\
&= K(\theta)\Gamma^{H,G}_{ijk} + \partial_j K(\theta) g^G_{ik}(\theta) + \partial_i K(\theta) g^G_{jk}(\theta) \qquad (4.148)
\end{aligned}$$

with $G(p) = \frac{-pF''(p)}{F'(p)}$ and $K(\theta) = \frac{1}{h_F(\theta)}$.

Hence the connection induced by the divergence function D_F is the (-1)-conformal transformation of the (H, G)-connection $\nabla^{H,G}$ by a gauge function $K(\theta)$. □

Also the connection $\nabla^{D^*_F}$ induced by the dual D^*_F is the 1-conformal transformation of the (F, G)-connection $\nabla^{F,G}$.

Theorem 4.3.18. *The affine connection $\nabla^{D_F^*}$ induced by the dual D_F^* is the 1-conformal transformation of the (F, G)-connection $\nabla^{F,G}$ by a gauge function $K(\theta) = \frac{1}{h_F(\theta)}$, where $G(p) = \frac{-pF''(p)}{F'(p)}$. That is,*

$$\Gamma_{ijk}^{D_F^*}(\theta) \quad = \quad K(\theta)\Gamma_{ijk}^{F,G} - \partial_k K(\theta) g_{ij}^{G}(\theta) \tag{4.149}$$

with $G(p) = \frac{-pF''(p)}{F'(p)}$ and $K(\theta) = \frac{1}{h_F(\theta)}$.

Proof. The components of $\Gamma_{ijk}^{D_F^*}$ of the connection are

$$\begin{aligned}
\Gamma_{ijk}^{D_F^*}(\theta) &= \frac{-1}{(h_F(\theta))^2}\left(\int \frac{\partial_i\partial_j F(p)}{F'(p)} \; dx\right)\left(\int \frac{-F''(p)}{(F'(p))^2}\,\partial_k p \; dx\right) \\
&+ \frac{-1}{h_F(\theta)}\int \partial_i\partial_j F(p)\,\frac{F''(p)}{(F'(p))^2}\,\partial_k p \; dx \qquad (4.150) \\
&= \frac{-1}{h_F(\theta)}\partial_i\partial_j\psi_F(\theta)\int \frac{F''(p)}{(F'(p))^2}\,\partial_k p \; dx \\
&+ \frac{1}{h_F(\theta)}\partial_i\partial_j\psi_F(\theta)\int \frac{F''(p)}{(F'(p))^2}\,\partial_k p \; dx \qquad (4.151) \\
&= 0 \qquad (4.152)
\end{aligned}$$

We have

$$\partial_i\partial_j F = pF'(p)\partial_i\partial_j\ell + [pF'(p) + p^2F''(p)]\;\partial_i\ell\;\partial_j\ell. \tag{4.153}$$

From Equation (4.150) the term

$$\frac{-1}{h_F(\theta)}\int \partial_i\partial_j F(p)\,\frac{F''(p)}{(F'(p))^2}\,\partial_k p \; dx = \tag{4.154}$$

$$\begin{aligned}
&= \frac{1}{h_F(\theta)}\int\left[\partial_i\partial_j\ell + [1 + \frac{pF''(p)}{F'(p)}]\;\partial_i\ell\;\partial_j\ell\right]\frac{-pF''(p)}{(F'(p))^2}\,p\;\partial_k\ell \; dx \qquad (4.155) \\
&= K(\theta)\Gamma_{ijk}^{F,G} \qquad (4.156)
\end{aligned}$$

where $G(p) = \frac{-pF''(p)}{F'(p)}$ and $K(\theta) = \frac{1}{h_F(\theta)}$.
Since $\partial_i\partial_j F(p) = -\partial_i\partial_j\psi_F(\theta)$

$$\frac{-1}{(h_F(\theta))^2}\left(\int \frac{\partial_i\partial_j F(p)}{F'(p)} \; dx\right)\left(\int \frac{-F''(p)}{(F'(p))^2}\,\partial_k p \; dx\right) =$$

$$\begin{aligned} &= \frac{-1}{h_F(\theta)}\partial_i\partial_j\psi_F(\theta)\int \frac{F''(p)}{(F'(p))^2}\,\partial_k p\,dx && (4.157)\\ &= \frac{-1}{(h_F(\theta))^2}\left(\int \frac{pF''(p)}{(F'(p))^2}\,\partial_k\ell\,dx\right)\int \frac{pF''(p)}{F'(p)}\partial_i\ell\,\partial_j\ell\,p\,dx && (4.158)\\ &= -\partial_k K(\theta) g^G_{ik}(\theta) && (4.159)\end{aligned}$$

Then

$$\Gamma^{D_F^*}_{ijk}(\theta) \;=\; K(\theta)\Gamma^{F,G}_{ijk} - \partial_k K(\theta) g^G_{ij}(\theta) \qquad (4.160)$$

with $G(p) = \frac{-pF''(p)}{F'(p)}$ and $K(\theta) = \frac{1}{h_F(\theta)}$.

Hence the connection induced by the divergence function D_F^* is the 1-conformal transformation of the (F,G)-connection $\nabla^{F,G}$ by a gauge function $K(\theta)$. □

In summary we proved the

Theorem 4.3.19. *$(\mathcal{S}, g^G, \nabla^{H,G})$ and $(\mathcal{S}, g^{D_F}, \nabla^{D_F})$ are (-1)-conformally equivalent. Also $(\mathcal{S}, g^G, \nabla^{F,H})$ and $(\mathcal{S}, g^{D_F}, \nabla^{D_F^*})$ are 1-conformally equivalent, with $G(p) = \frac{-pF''(p)}{F'(p)}$ and H is the G-dual embedding of F.*

Remark 4.3.20. *The dually flat χ-geometry on the F-exponential family induced by the divergence D_F is the conformal flattening of the (F,G)-geometry. When $F(p) = \ln_q(p)$ and $G(p) = constant$, the F-exponential family is the q-exponential family and the q-geometry is the conformal flattening of the α-geometry.*

Thus, we have

Theorem 4.3.21. *For a F-exponential family $\mathcal{S}$, let $G(p) = \frac{-pF''(p)}{F'(p)}$ and H is the G-dual embedding of F. Then*

1. *(g^{D_F}, ∇^{D_F}) and $(g^{D_F}, \nabla^{D_F^*})$ are mutually dual Hessian structures on $\mathcal{S}$ equivalently, $(\mathcal{S}, g^{D_F}, \nabla^{D_F}, \nabla^{D_F^*})$ is a dually flat space.*

2. *The canonical coordinate θ is $\nabla^{D_F^*}$-affine and ψ_F is the potential function corresponding to θ.*

3. *The metric $g^{D_F}_{ij}(\theta) = \partial_i\partial_j\psi_F(\theta)$.*

4. *The dual coordinate η is $\eta_i = \partial_i\psi_F(\theta) = E_{\hat{p}_F}[x_i]$ and it is ∇^{D_F}-affine.*

5. *The dual potential function* ϕ_F *corresponding to* η *is* $\phi_F(\eta) = E_{\hat{p}_F}(F(p))$.

6. $(\mathcal{S}, g^{D_F}, \nabla^{D_F})$ *and* $(\mathcal{S}, g^{G}, \nabla^{H,G})$ *are* (-1)*-conformally equivalent.*

7. $(\mathcal{S}, g^{D_F}, \nabla^{D^*_F})$ *and* $(\mathcal{S}, g^{G}, \nabla^{F,H})$ *are* 1*-conformally equivalent.*

4.4 Summary

In this chapter starting with the description of a dually flat space an overview of the dually flat geometry of the exponential family and the q-exponential family is given. Then we described the two dually flat structures on a deformed exponential family, the U-geometry and the χ-geometry. Further the relation between these two dually flat structures and the non-invariant (F, G)-geometry is explored. We showed that U-geometry is the (F, G)-geometry and χ-geometry is the conformal flattening of the (F, G)-geometry for suitable choices of F and G.

CHAPTER 5

Geometry of Estimation

In this chapter we focus on the geometric theory of parameter estimation problem in a statistical model, especially in an exponential family and in a curved exponential family. Amari [12], [11] elucidated the significance of the geometric tools such as the Fisher information metric and the α-connections in the asymptotic theory of estimation. He interpreted the asymptotic properties of an estimator in a curved exponential family in terms of the ancillary manifold, see also [5], [14], [42, 43], [46].

In certain areas of neuroscience a mismatched model or an unfaithful model is often used for statistical inference instead of the original model [47], [48]. Ozumi et al. [48] described the maximum likelihood estimation based on a mismatched model from the information geometric point of view. Here we describe an information geometric approach to a general estimation problem based on a mismatched model in an exponential family.

In Section 5.1 a short account of the statistical properties of an estimator is given. Amari [11], [12] interpreted the consistency and efficiency of an estimator in a curved exponential family in terms of the ancillary manifold. In Section 5.2 we describe his work in detail. In Section 5.3 the parameter estimation problem based on a mismatched model is discussed. We give a necessary and sufficient condition for the estimator based on a mismatched model to be consistent and first order efficient. Further a theoretical formulation of the maximum likelihood estimation problem based on a mismatched model in an exponential family is given with a detailed proof of the same.

5.1 Parameter Estimation in a Statistical Manifold

Consider an n-dimensional statistical manifold $\mathcal{S} = \{p(x;\theta) \ / \ \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$. Let $\mathbf{x}_N = (x^1, \cdots, x^N)$ be N independent observations from the random variable X distributed according to $p(x;\theta) \in \mathcal{S}$. The joint probability density function $p_N(\mathbf{x}_N;\theta)$

is

$$p_N(\mathbf{x}_N;\theta) = \prod_{i=1}^{N} p(x^i;\theta) \tag{5.1}$$

Then the log-likelihood of the density is

$$\ell^N(\mathbf{x}_N;\theta) = \log p_N(\mathbf{x}_N;\theta) = \sum_{i=1}^{N} \log p(x^i;\theta) \tag{5.2}$$

Let $\mathcal{S}_N = \{p_N(\mathbf{x}_N;\theta) \ / \ \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$. Then $\mathcal{S}_N$ is an n-dimensional manifold with a coordinate system θ. The Fisher information metric on $\mathcal{S}_N$ is

$$\begin{aligned} g_{ij}^N(\theta) &= \int \partial_i \ell^N(\mathbf{x}_N;\theta)\, \partial_j \ell^N(\mathbf{x}_N;\theta)\, p_N(\mathbf{x}_N;\theta)\, d\mathbf{x}_N & (5.3)\\ &= N g_{ij}(\theta) & (5.4) \end{aligned}$$

where $\partial_i = \frac{\partial}{\partial\theta^i}$ and $d\mathbf{x}_N = dx^1 \cdots dx^N$ and $g_{ij}(\theta)$ is the components of the Fisher information metric on $\mathcal{S}$.

In parameter estimation one need to estimate the value of an unknown parameter θ based on the observations taken from a random variable x distributed according to $p(x;\theta)$. An estimator $\hat{\theta}_N$ is defined as a function of the N observations of x given by

$$\hat{\theta}_N = \hat{\theta}_N(x^1, \cdots, x^N) = \hat{\theta}_N(\mathbf{x}_N) \tag{5.5}$$

Note 5.1.1. *Note that the estimator $\hat{\theta}_N$ depends on the number of observations N. But for the notational convenience we denote $\hat{\theta}_N$ by $\hat{\theta}$.*

There are certain desired properties that an estimator should possess which reflects the closeness of the estimator to the actual parameter of the distribution in some sense. **Unbiasedness** is one of such conditions which is stated as

$$E_\theta[\hat{\theta}] = \theta, \quad \forall\, \theta \in \mathbb{E} \tag{5.6}$$

where E_θ is the expectation with respect to the distribution $p_N(\mathbf{x}_N;\theta)$.

The **mean square error** of an estimator is expressed as a matrix

$$\mathrm{MSE}(\hat{\theta}) = \left[E_\theta[(\hat{\theta}^i - \theta^i)(\hat{\theta}^j - \theta^j)]\right] \tag{5.7}$$

The variance-covariance matrix $V_\theta(\hat{\theta}) = [v_\theta^{ij}(\hat{\theta})]$ for $\hat{\theta}$ is

$$v_\theta^{ij}(\hat{\theta}) = E_\theta\left[(\hat{\theta}^i - E[\hat{\theta}])(\hat{\theta}^j - E[\hat{\theta}])\right] \tag{5.8}$$

When the estimator $\hat{\theta}$ is unbiased then the mean square error is the variance of the estimator. That is, $MSE(\hat{\theta}) = V_\theta(\hat{\theta})$.

The **Cramer-Rao inequality** gives a lower bound on the variance of an unbiased estimator and is given by

$$V_\theta(\hat{\theta}) \geq G_N^{-1}(\theta) \quad \text{or} \quad [v_\theta^{ij}(\hat{\theta})] \geq \frac{1}{N}[g^{ij}(\theta)] \tag{5.9}$$

where $G_N^{-1}(\theta) = \frac{1}{N}[g^{ij}(\theta)]$ is the inverse of the Fisher information metric on $\mathcal{S}_N$ and $[g^{ij}(\theta)]$ is the inverse of the Fisher information metric on $\mathcal{S}$.

An unbiased estimator $\hat{\theta}$ which achieves Cramer-Rao equality ($[v_\theta^{ij}(\hat{\theta})] = \frac{1}{N}[g^{ij}(\theta)]$)is called the **finite sample efficient estimator**.

The properties of an estimator in the case of fixed number of observations N are described above. In the asymptotic theory of estimation main focus is given to the behavior of an estimator in the limiting case $N \to \infty$. In this case instead of the unbiasedness one has the consistency.

Note 5.1.2. *Note that when describing the finite sample theory, $\hat{\theta}$ is used for $\hat{\theta}_N$. In the case of asymptotic analysis $\{\hat{\theta}_N,\ N = 1, 2, \cdots\}$ is used for the estimator.*

An estimator $\{\hat{\theta}_N,\ N = 1, 2, \cdots\}$ is said to be **consistent** if for all θ the estimator $\hat{\theta}_N(\mathbf{x}_N)$ converges in probability to θ as $N \to \infty$. That is, for all θ and for every $\epsilon > 0$,

$$\lim_{N\to\infty} \Pr_\theta\{|\hat{\theta}_N - \theta| > \epsilon\} = 0 \tag{5.10}$$

The notion of mean consistency is a much more stronger condition than the usual notion of consistency. Under certain regularity conditions, the expectation of $\hat{\theta}_N(\mathbf{x}_N)$ converges to θ uniformly which is the **mean consistency**. That is,

$$\lim_{N\to\infty} E_\theta[\hat{\theta}_N] = \theta, \quad \lim_{N\to\infty} \partial_j E_\theta[\hat{\theta}_N^i] = \partial_j \theta^i = \delta_{ij}. \tag{5.11}$$

Such an estimator is often called an **asymptotically unbiased estimator**.

The mean square error of an asymptotically unbiased estimator satisfies the **asymptotic**

Cramer-Rao inequality

$$\lim_{N\to\infty} N[v_\theta^{ij}(\hat{\theta}_N)] \geq [g^{ij}(\theta)] \tag{5.12}$$

A consistent estimator which attains equality in the above equation is called an **asymptotically efficient estimator** or a **first order efficient estimator** [14].

Definition 5.1.3. *Let $\mathcal{S} = \{p(x;\theta)\ /\ \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$ be an n-dimensional statistical manifold. For N independent observations $\mathbf{x}_N = (x^1, \cdots, x^N)$ from $p(x;\theta) \in \mathcal{S}$ the* ***likelihood function*** *$L^N(\theta)$ is given by*

$$L^N(\theta) = p_N(\mathbf{x}_N;\theta) = \prod_{i=1}^{N} p(x^i;\theta) \tag{5.13}$$

Since $\log$ function is a strictly increasing function, maximizing the likelihood function $L^N(\theta)$ is equivalent to maximizing the log-likelihood function $\log(L^N(\theta))$.
We say that $\hat{\theta}$ is the ***Maximum Likelihood Estimator (MLE)*** *if*

$$\hat{\theta} = \arg\max_{\theta\in E} L^N(\theta) = \arg\max_{\theta\in E} \log(L^N(\theta)) = \arg\max_{\theta\in E} \sum_{i=1}^{N} \log(p(x^i;\theta)) \tag{5.14}$$

Remark 5.1.4. *For an arbitrary model $\mathcal{S} = \{p_\theta\}$ there need not exist a finite sample efficient estimator. Amari and Nagaoka [14] showed that a necessary and sufficient condition for a coordinate system θ of a model $\mathcal{S} = \{p_\theta\}$ to have an efficient estimator is that $\mathcal{S}$ is an exponential family and θ is m-affine. But there always exists an asymptotically efficient estimator for an arbitrary statistical model unlike in the finite case. In fact MLE is an asymptotically efficient estimator [14].*

5.2 Estimation in Exponential Family

Amari [12] constructed a differential geometric framework for the statistical estimation problem in an exponential family and in a curved exponential family, see also [14]. In this section we discuss his work in detail.

Consider an n-dimensional exponential family $\mathcal{S} = \{p(x;\theta) \ / \ \theta \in E \subseteq \mathbb{R}^n\}$

$$p(x;\theta) = \exp\{\sum_{i=1}^{n} \theta^i x_i - \psi(\theta)\}. \tag{5.15}$$

where $x = (x_1, \cdots, x_n)$ is a set of random variables and θ is the canonical coordinate. We have seen that $\mathcal{S}$ is a dually flat space and the dual coordinate $\eta = (\eta_i)$ is $\eta_i = E_\theta[x_i]$.

Now consider an estimator $\hat{\eta} = x$ for η. Then

$$\begin{aligned} E_\theta[x] &= \eta & (5.16) \\ E_\theta[(x_i - \eta_i)(x_j - \eta_j)] &= E_\theta[\partial_i \ell \partial_j \ell] = g_{ij}(\theta) & (5.17) \end{aligned}$$

where $g_{ij}(\theta)$ are the components of the Fisher information metric with respect to the θ coordinate. This implies that variance $V_\eta(\hat{\eta})$ of $\hat{\eta}$ is the Fisher information matrix $G(\theta)$ and by duality, $G(\theta) = G^{-1}(\eta)$. So $V_\eta(\hat{\eta}) = G^{-1}(\eta)$.

Let $\mathbf{x}_N = (x^1, \cdots, x^N)$ be N independent observations from $p(x;\theta) \in \mathcal{S}$. Then the joint probability density function is

$$p_N(\mathbf{x}_N;\theta) = \prod_{j=1}^{N} \exp\{\sum_{i=1}^{n} \theta^i x_i^j - \psi(\theta)\} \tag{5.18}$$

or the log-likelihood function is

$$\ell^N(\mathbf{x}_N;\theta) = N\left[\sum_{i=1}^{n} \theta^i \bar{x}_i - \psi(\theta)\right] \tag{5.19}$$

where $\bar{x} = (\bar{x_1}, \cdots, \bar{x_n})$ is the arithmetic mean given by

$$\bar{x}_i = \frac{x_i^1 + \cdots + x_i^N}{N}, \quad i = 1, \cdots, n \tag{5.20}$$

That is the joint probability density $p_N(\mathbf{x}_N;\theta)$ depends on the N observations $x^1, \cdots, x^N$ through $\bar{x}$. Thus the statistic $\bar{x}$ is a sufficient statistic for the parameter θ and is called the observed point.

Now consider the estimator $\hat{\eta}_N = \bar{x}$ for η. Then

$$\begin{aligned} E_\theta[\bar{x}] &= \eta & (5.21) \\ E_\theta[(\bar{x}_i - \eta_i)(\bar{x}_j - \eta_j)] &= \frac{1}{N} g_{ij}(\theta) & (5.22) \end{aligned}$$

That is, $\hat{\eta}_N = \bar{x}$ is an unbiased estimator and a finite sample efficient estimator for η. Thus a finite dimensional standard exponential family naturally has a sufficient statistic and a finite sample efficient estimate [12], [14].

5.2.1 Estimation in a curved exponential family

Consider an m-dimensional smooth submanifold $M = \{q(x; u) \ / \ u = (u^a) \in \mathbb{R}^m\}$ in an n-dimensional exponential family $\mathcal{S}$. Then M is called an (n, m)**-curved exponential family** and

$$q(x, u) = p(x; \theta(u)). \tag{5.23}$$

Let $\mathbf{x}_N = (x^1, \cdots, x^N)$ be N independent observations from $q(x; u) \in M$. Then the observed point $\bar{x} = (\bar{x}_1, \cdots, \bar{x}_n)$ defines a distribution in $\mathcal{S}$ whose η coordinate is given by $\hat{\eta}_N = \bar{x}$. But this point need not be in the submanifold M. Since $\bar{x}$ is a sufficient statistic for M an estimator $\hat{u}_N$ for $u \in M$ can be regarded as a function of the observed point $\hat{\eta}_N$. That is, the estimator $\hat{u}_N$ is represented as a mapping f_N from $\mathcal{S}$ to M

$$f_N : \mathcal{S} \longrightarrow M \quad \text{where} \quad \hat{\eta} \mapsto \hat{u}_N = f_N(\hat{\eta}_N) \tag{5.24}$$

An **ancillary manifold** or an **estimating submanifold** $A_N(u)$ corresponding to the point $u \in M$ associated with an estimator f_N is defined as

$$A_N(u) = f_N^{-1}(u) = \{\eta = (\eta_i) \in \mathcal{S} \ / \ f_N(\eta) = u\} \tag{5.25}$$

That is, $A_N(u)$ is the set of all points η in $\mathcal{S}$ which are mapped to $u \in M$ by the estimator f_N [11], [12], [14].

Remark 5.2.1. *Amari [11], [12] studied the statistical properties of an estimator and interpreted them geometrically in terms of the estimating submanifold. He first considered an estimator function $f : \mathcal{S} \longrightarrow M$ which does not depend upon the number of* *observations N explicitly and thus the ancillary manifold $A(u)$ at each point $u \in M$ is also independent of N. Then gave geometric interpretations for the consistency and efficiency of an estimator in terms of the estimating submanifold $A(u)$ [11], [12], [14]. In general, the estimator function depends upon the number of observations N explicitly. In that case we have $A_N(u)$ instead of $A(u)$ and we take $A(u)$ to be the limit of $A_N(u)$ as $N \to \infty$. Amari and Nagaoka [14] considered this case also. We detail their work for a better understanding of the mismatched estimation problem discussed in the subsequent sections.*

Let

$$A(u) = \lim_{N \to \infty} A_N(u) \tag{5.26}$$

Note that the estimator f_N is assumed to be a continuous function from $\mathcal{S}$ to M for each N. Also let f be the limiting estimator function which determines the limiting estimating submanifold $A(u)$.

Then the consistency and efficiency of $\{\hat{u}_N,\ N = 1, 2, \cdots\}$ can be interpreted as [11], [12], [14]

Theorem 5.2.2. *Let $M = \{q(x; u)\ /\ u = (u^a) \in \mathbb{R}^m\} \subset \mathcal{S}$ be a curved exponential family. An estimator $\{\hat{u}_N,\ N = 1, 2, \cdots\}$ for $u \in M$ is consistent if and only if $\eta(u) \in M \subset \mathcal{S}$ is in the estimating submanifold $A(u)$.*

Proof. Let $\mathbf{x}_N = (x^1, \cdots, x^N)$ be N independent observations from $q(x; u) \in M$. Then

$$E[x] = \eta(u). \tag{5.27}$$

By the law of large numbers the observed point $\hat{\eta}_N = \bar{x} = (\bar{x}_1, \cdots, \bar{x}_n)$ defined in Equation (5.20), converges in probability (we denote it by $\xrightarrow{p}$) to $\eta(u)$ as $N \to \infty$. That is,

$$\hat{\eta}_N = \bar{x} \xrightarrow{p} \eta(u) \quad \text{as} \quad N \to \infty \tag{5.28}$$

Then

$$\hat{u}_N = f_N(\hat{\eta}_N) \overset{p}{\to} f(\eta(u)) \quad \text{as} \quad N \to \infty \tag{5.29}$$

For the estimator $\{\hat{u}_N,\ N = 1, 2, \cdots\}$ to be consistent

$$\hat{u}_N \xrightarrow{p} u \quad \text{as} \quad N \to \infty \tag{5.30}$$

Thus the estimator $\{\hat{u}_N,\ N = 1, 2, \cdots\}$ is consistent iff

$$f(\eta(u)) = u \tag{5.31}$$

iff

$$\eta(u) \in A(u). \tag{5.32}$$

□

In the proof of the following theorem we use the Einstein summation convention for the sake of convenience (that is, $\sum_i x_i y^i$ is denoted by $x_i y^i$).

Theorem 5.2.3. *Let $M = \{q(x; u)\ /\ u = (u^a) \in \mathbb{R}^m\} \subset \mathcal{S}$ be a curved exponential family. A consistent estimator $\{\hat{u}_N,\ N = 1, 2, \cdots\}$ for $u \in M$ is first order efficient if and only if $A(u)$ is orthogonal to M at the intersecting point $\eta(u) \in M$.*

Proof. Let $\mathbf{x}_N = (x^1, \cdots, x^N)$ be N independent observations from $q(x; u) \in M$. Note that $q(x; u) = p(x; \eta(u))$. By the law of large numbers the observed point $\bar{x}$ converges to $\eta(u)$ as $N \to \infty$. From Equations (5.16) and (5.17) $E[x] = \eta(u)$ and covariance matrix $V(x) = [g_{ij}(\eta(u))] = [g_{ij}(u)]$, where g_{ij} are the components of the Fisher information metric with respect to the θ-coordinate.
Now consider the random variable

$$\tilde{x} = \sqrt{N}(\bar{x} - \eta(u)) \tag{5.33}$$

Then by the central limit theorem $\tilde{x}$ asymptotically follows the normal distribution with mean 0 and covariance $[g_{ij}(u)]$.

Given the estimator $\{\hat{u}_N,\ N = 1, 2, \cdots\}$ is consistent. Then from Theorem 5.2.2 the $(n-m)$-dimensional estimating manifold $A(u)$ passes through the point $\eta(u) \in M$. Now introduce a coordinate system $v = (v^\kappa),\ \kappa = m + 1, \cdots, n$ on $A(u)$ with u as origin. Then $w = (u, v)$ forms another coordinate system for $\mathcal{S}$. Here we use indices such as α, β for the coordinate system w, indices such as a, b for the coordinate system u and indices such as κ, λ for the coordinate system v.

$$w = (w^\alpha) = (u, v) = (u^a, v^\kappa) \tag{5.34}$$

where $\alpha = 1, \cdots, n$, $a = 1, \cdots, m$ and $\kappa = m+1, \cdots, n$.

The η-coordinate for $\mathcal{S}$ can be written in terms of w as

$$\eta(w) = \eta(u, v) \tag{5.35}$$

The w-coordinate of points in M are given by $w = (u, 0)$ or $\eta(u) = \eta(u, 0)$.

The bases of the tangent space $T_{\eta(u)}\mathcal{S}$ of $\mathcal{S}$ at $\eta(u)$ with respect to η-coordinate and w-coordinate are given by

$$\partial^i = \frac{\partial}{\partial \eta_i}, \quad \partial_\alpha = \frac{\partial}{\partial w^\alpha}. \tag{5.36}$$

Note that we can decompose $\{\partial_\alpha\}$ into $\{\partial_a\} \cup \{\partial_\kappa\}$, where

$$\partial_a = \frac{\partial}{\partial u^a}, \quad \partial_\kappa = \frac{\partial}{\partial v^\kappa}. \tag{5.37}$$

The tangent space $T_{\eta(u)}M$ of M is spanned by $\{\partial_a\}$ and the tangent space $T_{\eta(u)}A(u)$ of $A(u)$ is spanned by $\{\partial_\kappa\}$. Also

$$\partial_\alpha = B^i_\alpha \partial^i, \quad \partial_a = B^i_a \partial^i \quad \text{and} \quad \partial_\kappa = B^i_\kappa \partial^i \tag{5.38}$$

where

$$B^i_\alpha = \frac{\partial \eta_i}{\partial w^\alpha}, \quad B^i_a = \frac{\partial \eta_i}{\partial u^a}, \quad B^i_k = \frac{\partial \eta_i}{\partial v^k} \tag{5.39}$$

Let the components of the Fisher information metric with respect to the two bases $\{\partial^i\}$ and $\{\partial_\alpha\}$ be

$$g^{ij} =< \partial^i, \partial^j >; \quad g_{\alpha\beta} =< \partial_\alpha, \partial_\beta >= B^i_\alpha B^j_\beta g^{ij} \tag{5.40}$$

The matrix $[g_{\alpha\beta}]$ can be decomposed as

$$[g_{\alpha\beta}] = \begin{bmatrix} g_{ab} & g_{a\lambda} \\ g_{\kappa b} & g_{\kappa\lambda} \end{bmatrix} \tag{5.41}$$

where

$$g_{ab} =< \partial_a, \partial_b >= B_a^i B_b^j g^{ij} \tag{5.42}$$

$$g_{a\kappa} =< \partial_a, \partial_\kappa >= B_a^i B_\kappa^j g^{ij} \tag{5.43}$$

$$g_{\kappa\lambda} =< \partial_\kappa, \partial_\lambda >= B_\kappa^i B_\lambda^j g^{ij} \tag{5.44}$$

Now let $\hat{w}_N = (\hat{u}_N, \hat{v}_N)$ be the (u, v)-coordinate of the observed point $\hat{\eta}_N = \bar{x}$.

$$\hat{\eta}_N = \eta(\hat{w}_N) = \eta(\hat{u}_N, \hat{v}_N) \tag{5.45}$$

Since $\hat{u}_N$ and $\hat{v}_N$ are close to u and 0 respectively, consider

$$\tilde{u}_N = \sqrt{N}(\hat{u}_N - u), \quad \tilde{v}_N = \sqrt{N}\hat{v}_N, \quad \tilde{w}_N = (\tilde{u}_N, \tilde{v}_N) \tag{5.46}$$

Then

$$\hat{w}_N = w + \frac{1}{\sqrt{N}}\tilde{w}_N \quad \text{with} \quad w = (u, 0). \tag{5.47}$$

By taking the Taylor series expansion of Equation (5.45) around the point $w = (u, 0)$ the i^{th} coordinate of $\hat{\eta}_N$ is

$$\bar{x}_i = \eta_i(u, 0) + \frac{1}{\sqrt{N}}\partial_\alpha \eta_i(u, 0)\tilde{w}_N^\alpha + \frac{1}{2N}\partial_\alpha \partial_\beta \eta_i(u, 0)\tilde{w}_N^\alpha \tilde{w}_N^\beta + O(\frac{1}{N\sqrt{N}}) \tag{5.48}$$

This can be rewritten using Equation (5.33) as

$$\tilde{x}_i = B_\alpha^i(u, 0)\tilde{w}_N^\alpha + \frac{1}{2\sqrt{N}}\partial_\alpha \partial_\beta \eta_i(u, 0)\tilde{w}_N^\alpha \tilde{w}_N^\beta + O(\frac{1}{N}) \tag{5.49}$$

By neglecting the terms smaller than or equal to the order of $\frac{1}{\sqrt{N}}$ in the above equation we obtain the linear equation

$$\tilde{x}_i = B_\alpha^i \tilde{w}_N^\alpha \tag{5.50}$$

Let $[D_i^\alpha]$ be the inverse matrix of $[B_\alpha^i]$. Then from Equation (5.40)

$$D_i^\alpha = \frac{\partial w^\alpha}{\partial \eta_i} = g^{\alpha\beta} g^{ij} B_\beta^j \tag{5.51}$$

where $[g^{\alpha\beta}]$ is the inverse of the matrix $[g_{\alpha\beta}]$.

Then $\tilde{w}_N^\alpha = D_i^\alpha \tilde{x}_i$. Since $\tilde{x}$ asymptotically follows normal distribution with 0 mean and covariance $[g_{ij}]$ then $\tilde{w}_N$ is also normally distributed with 0 mean and covariance $[g^{\alpha\beta}]$. That is,

$$\lim_{N\to\infty} V(\tilde{w}_N) = [g^{\alpha\beta}] \tag{5.52}$$

where $V(\tilde{w}_N)$ is the covariance matrix of $\tilde{w}_N$.

That is,

$$\lim_{N\to\infty} V(\tilde{w}_N) = \lim_{N\to\infty} NV(\hat{w}_N) = [g^{\alpha\beta}] \tag{5.53}$$

Let the asymptotic mean square error of $\{\hat{u}_N,\ N = 1, 2, \cdots\}$ be

$$\lim_{N\to\infty} NE[(\hat{u}_N^a - u^a)(\hat{u}_N^b - u^b)] \tag{5.54}$$

is the $(a, b)^{\text{th}}$ component of $[g^{\alpha\beta}]$ from Equation (5.53) and is denoted by $\bar{g}^{ab}$.

Taking the inverse of the matrix $[g_{\alpha\beta}]$ in Equation (5.41) and also from Equation (5.53)

$$[\bar{g}^{ab}] = [g_{ab} - g_{a\kappa} g^{\kappa\lambda} g_{b\lambda}]^{-1} \quad \Rightarrow \quad [\bar{g}^{ab}] \geq [g^{ab}] \tag{5.55}$$

which is the asymptotic Cramer-Rao inequality, where $[g^{ab}]$ is the inverse of $[g_{ab}]$.

Then the consistent estimator $\{\hat{u}_N,\ N = 1, 2, \cdots\}$ is first order efficient iff

$$[\bar{g}^{ab}] = [g^{ab}] \tag{5.56}$$

which is iff $g_{a\kappa} = 0,\ \forall\ a, \kappa$.

Note that $g_{ak} =< \partial_a, \partial_k >$ is the inner product of the tangent vector $\partial_a \in T_{\eta(u)}M$ and the tangent vector $\partial_k \in T_{\eta(u)}A(u)$. Hence the consistent estimator $\{\hat{u}_N,\ N = 1, 2, \cdots\}$ is first order efficient iff $A(u)$ is orthogonal to M at $\eta(u)$. □

Note 5.2.4. *For a statistical manifold $\mathcal{S} = \{p(x; \theta)\}$ the maximum likelihood estimator $\hat{\theta}$ is an asymptotically efficient estimator. More precisely, the MLE $\hat{\theta}$ asymptotically follows a normal distribution with mean θ and covarinace $[\frac{1}{N} g^{ij}(\theta)]$. Amari and Nagaoka [14] gave a geometric proof of the same in an exponential family using the ancillary manifold and the canonical divergence, see also [11], [12].*

5.3 Mismatched Estimation in a Curved Exponential Family

In the area of population coding in neuroscience a mismatched model or an unfaithful decoding model is often used in place of the original model to quantify the significance of the correlated activities of neurons or for saving the computational cost, see [47], [48] for more details. Oizumi et al. [48] described maximum likelihood inference in a curved exponential family based on a mismatched model and interpreted it from the information geometric point of view. Motivated by their work we consider the problem of mismatched estimation in a curved exponential family for any estimator, not only for the MLE. First we interpret the consistency and efficiency of an estimator based on a mismatched model in terms of the associated ancillary family.

Let $\mathcal{S}$ be an exponential family and $M = \{q(x;u) \ / \ u = (u^a) \in \mathbb{R}^m\}$ be a curved exponential family. Suppose that we have a mismatched model $M^* = \{q'(x;u) \ / \ u = (u^a) \in \mathbb{R}^m\}$ corresponding to the original model M. M^* is a submanifold of $\mathcal{S}$. Let the embedding functions of M and M^* in $\mathcal{S}$ be $\theta(u)$ and $\theta'(u)$ respectively. Let $\eta(u)$ and $\eta'(u)$ be the corresponding dual representations.

Let $\mathbf{x}_N = (x^1, \cdots, x^N)$ be N independent observations from $q(x;u) \in M$. Then the observed point $\bar{x} = (\bar{x}_1, \cdots, \bar{x}_n)$ defines a distribution in $\mathcal{S}$ whose η-coordinate is given by $\hat{\eta}_N = \bar{x}$. For the inference we are using the mismatched model M^* instead of the original model M. Since $\bar{x}$ is a sufficient statistic for $\mathcal{S}$, it is a sufficient statistic for the submanifold M^* also. Then an estimator $\hat{u}'_N$ for M^* can be regarded as a function of the observed point $\hat{\eta}_N$. Thus the estimator $\hat{u}'_N$ is represented as a mapping f'_N from $\mathcal{S}$ to M^*

$$f'_N : \mathcal{S} \longrightarrow M^* \quad \text{where} \quad \hat{\eta}_N \mapsto \hat{u}'_N = f'_N(\hat{\eta}_N) \tag{5.57}$$

The **ancillary manifold** or the **estimating submanifold** $A'_N(u)$ corresponding to the point $u \in M^*$ associated with f'_N is defined as

$$A'_N(u) = f'^{-1}_N(u) = \{\eta = (\eta_i) \in \mathcal{S} \ / \ f'_N(\eta) = u\} \tag{5.58}$$

That is, $A'_N(u)$ is the set of all points η in $\mathcal{S}$ which are mapped to $u \in M^*$ by the estimator f'_N.

Now we analyze the characteristics of an estimator $\{\hat{u}'_N,\ N = 1, 2, \cdots\}$ in M^* using the geometric properties of the ancillary submanifold $A'_N(u)$. Let

$$A'(u) = \lim_{N\to\infty} A'_N(u) \tag{5.59}$$

Note that the estimator f'_N is assumed to be a continuous function from $\mathcal{S}$ to M^* for each N. Also let f' be the limiting estimator function which determines the limiting estimating submanifold $A'(u)$.

Theorem 5.3.1. *An estimator* $\{\hat{u}'_N,\ N = 1, 2, \cdots\}$ *for* $u \in M^*$ *is consistent if and only if* $\eta(u) \in M \subset \mathcal{S}$ *is in the estimating submanifold* $A'(u)$ *attached to the point* $u \in M^*$*.*

Proof. Let $\mathbf{x}_N = (x^1, \cdots, x^N)$ be N independent observations from $q(x; u) \in M$. We have

$$E[x] = \eta(u). \tag{5.60}$$

By the law of large numbers the observed point $\hat{\eta}_N = \bar{x} = (\bar{x}_1, \cdots, \bar{x}_n)$ converges in probability (we denote it by $\xrightarrow{p}$) to $\eta(u)$ as $N \to \infty$. That is,

$$\hat{\eta}_N = \bar{x} \xrightarrow{p} \eta(u) \quad \text{as} \quad N \to \infty \tag{5.61}$$

Then

$$\hat{u}'_N = f'_N(\hat{\eta}_N) \overset{p}{\to} f'(\eta(u)) \quad \text{as} \quad N \to \infty \tag{5.62}$$

For the estimator $\{\hat{u}'_N, N = 1, 2, \cdots\}$ to be consistent

$$\hat{u}'_N \xrightarrow{p} u \quad \text{as} \quad N \to \infty \tag{5.63}$$

Thus the estimator $\{\hat{u}'_N, N = 1, 2, \cdots\}$ is consistent iff

$$f'(\eta(u)) = u \tag{5.64}$$

iff

$$\eta(u) \in A'(u). \tag{5.65}$$

□

In the proof of the following theorem we use the Einstein summation convention for the sake of convenience (that is, $\sum_i x_i y^i$ is denoted by $x_i y^i$).

Theorem 5.3.2. *A consistent estimator* $\{\hat{u}'_N,\ N = 1, 2, \cdots\}$ *for* $u \in M^*$ *is first order efficient if and only if* $A'(u)$ *is orthogonal to* M *at the intersecting point* $\eta(u) \in M$.

Proof. Let $\mathbf{x}_N = (x^1, \cdots, x^N)$ be N independent observations from $q(x; u) \in M$. Note that $q(x; u) = p(x; \eta(u))$. By the law of large numbers the observed point $\bar{x}$ converges to $\eta(u)$ as $N \to \infty$. From Equations (5.16) and (5.17) $E[x] = \eta(u)$ and covariance matrix $V(x) = [g_{ij}(\eta(u))] = [g_{ij}(u)]$, where g_{ij} is the components of the Fisher information metric with respect to the θ-coordinate.
Now consider the random variable

$$\tilde{x} = \sqrt{N}(\bar{x} - \eta(u)) \tag{5.66}$$

By the central limit theorem $\tilde{x}$ asymptotically follows the normal distribution with mean 0 and covariance $[g_{ij}(u)]$.

Given the estimator $\{\hat{u}'_N,\ N = 1, 2, \cdots\}$ is consistent. Thus from Theorem 5.3.1 the $(n-m)$-dimensional estimating manifold $A'(u)$ passes through the point $\eta(u) \in M$. Now introduce a coordinate system $v = (v^\kappa),\ \kappa = m+1, \cdots, n$ on $A'(u)$ with u as the origin. Then $w = (u, v)$ forms another coordinate system for $\mathcal{S}$. Here we use indices such as α, β for the coordinate system w, indices such as a, b for the coordinate system u and indices such as κ, λ for the coordinate system v.

$$w = (w^\alpha) = (u, v) = (u^a, v^\kappa) \tag{5.67}$$

where $\alpha = 1, \cdots, n$, $a = 1, \cdots, m$ and $\kappa = m + 1, \cdots, n$.
The η-coordinate for $\mathcal{S}$ can be written in terms of w as

$$\eta(w) = \eta(u, v) \tag{5.68}$$

The w-coordinate of points in M^* are given by $w = (u, 0)$ or $\eta'(u) = \eta'(u, 0)$. The w-coordinate of points in M are given by $w = (u, v')$ or $\eta(u) = \eta(u, v')$
The bases of the tangent space $T_{\eta(u)}\mathcal{S}$ of $\mathcal{S}$ at $\eta(u) \in M$ with respect to η-coordinate

and w-coordinate are given by

$$\partial^i = \frac{\partial}{\partial \eta_i}, \quad \partial_\alpha = \frac{\partial}{\partial w^\alpha}. \tag{5.69}$$

Note that we can decompose $\{\partial_\alpha\}$ into $\{\partial_a\} \cup \{\partial_\kappa\}$, where

$$\partial_a = \frac{\partial}{\partial u^a}, \quad \partial_\kappa = \frac{\partial}{\partial v^\kappa}. \tag{5.70}$$

The tangent space $T_{\eta(u)}M$ of M is spanned by $\{\partial_a\}$ and the tangent space $T_{\eta(u)}A'(u)$ of $A'(u)$ is spanned by $\{\partial_\kappa\}$. Also

$$\partial_\alpha = B^i_\alpha \partial^i, \quad \partial_a = B^i_a \partial^i \quad \text{and} \quad \partial_\kappa = B^i_\kappa \partial^i \tag{5.71}$$

where

$$B^i_\alpha = \frac{\partial \eta_i}{\partial w^\alpha}, \quad B^i_a = \frac{\partial \eta_i}{\partial u^a}, \quad B^i_k = \frac{\partial \eta_i}{\partial v^k} \tag{5.72}$$

Let the components of the Fisher information metric with respect to the two bases $\{\partial^i\}$ and $\{\partial_\alpha\}$ be

$$g^{ij} =< \partial^i, \partial^j >, \quad g_{\alpha\beta} =< \partial_\alpha, \partial_\beta >= B^i_\alpha B^j_\beta g^{ij} \tag{5.73}$$

The matrix $[g_{\alpha\beta}]$ can be decomposed as

$$[g_{\alpha\beta}] = \begin{bmatrix} g_{ab} & g_{a\lambda} \\ g_{\kappa b} & g_{\kappa\lambda} \end{bmatrix} \tag{5.74}$$

where

$$g_{ab} =< \partial_a, \partial_b >= B^i_a B^j_b g^{ij} \tag{5.75}$$

$$g_{a\kappa} =< \partial_a, \partial_\kappa >= B^i_a B^j_\kappa g^{ij} \tag{5.76}$$

$$g_{\kappa\lambda} =< \partial_\kappa, \partial_\lambda >= B^i_\kappa B^j_\lambda g^{ij} \tag{5.77}$$

Let $\hat{w}'_N = (\hat{u}'_N, \hat{v}'_N)$ be the (u, v)-coordinate of the observed point $\hat{\eta}_N = \bar{x}$.

$$\hat{\eta}_N = \eta(\hat{w}'_N) = \eta(\hat{u}'_N, \hat{v}'_N) \tag{5.78}$$

Since $\hat{u}'_N$ and $\hat{v}'_N$ are close to u and v' respectively, consider

$$\tilde{u}'_N = \sqrt{N}(\hat{u}'_N - u), \quad \tilde{v}'_N = \sqrt{N}(\hat{v}_N - v'), \quad \tilde{w}'_N = (\tilde{u}'_N, \tilde{v}'_N) \tag{5.79}$$

Then

$$\hat{w}'_N = w + \frac{1}{\sqrt{N}}\tilde{w}'_N \quad \text{with} \quad w = (u, v'). \tag{5.80}$$

By taking the Taylor series expansion of Equation (5.78) around the point $w = (u, v')$ the i^{th}-coordinate of $\hat{\eta}_N$ is

$$\bar{x}_i = \eta_i(u, v') + \frac{1}{\sqrt{N}}\partial_\alpha \eta_i(u, v')\tilde{w}'^\alpha_N + \frac{1}{2N}\partial_\alpha\partial_\beta\eta_i(u, v')\tilde{w}'^\alpha_N\tilde{w}'^\beta_N + O(\frac{1}{N\sqrt{N}}) \tag{5.81}$$

This can be rewritten using Equation (5.66) as

$$\tilde{x}_i = B^i_\alpha(u, v')\tilde{w}'^\alpha_N + \frac{1}{2\sqrt{N}}\partial_\alpha\partial_\beta\eta_i(u, v')\tilde{w}'^\alpha_N\tilde{w}'^\beta_N + O(\frac{1}{N}) \tag{5.82}$$

By neglecting the terms smaller than or equal to the order of $\frac{1}{\sqrt{N}}$ in the above equation we obtain the linear equation

$$\tilde{x}_i = B^i_\alpha \tilde{w}'^\alpha_N \tag{5.83}$$

Let $[D^\alpha_i]$ be the inverse matrix of $[B^i_\alpha]$. Then

$$D^\alpha_i = \frac{\partial w^\alpha}{\partial \eta_i} = g^{\alpha\beta}g^{ij}B^j_\beta \tag{5.84}$$

where $[g^{\alpha\beta}]$ is the inverse of the matrix $[g_{\alpha\beta}]$.
Then $\tilde{w}'^\alpha_N = D^\alpha_i \tilde{x}_i$. Since $\tilde{x}$ asymptotically follows normal distribution with 0 mean and covariance $[g_{ij}]$ then $\tilde{w}'_N$ is also normally distributed with 0 mean and covariance $[g^{\alpha\beta}]$. That is,

$$\lim_{N\to\infty} V(\tilde{w}'_N) = [g^{\alpha\beta}] \tag{5.85}$$

where $V(\tilde{w}_N)$ is the covariance matrix of $\tilde{w}_N$. That is,

$$\lim_{N\to\infty} V(\tilde{w}'_N) = \lim_{N\to\infty} NV(\hat{w}'_N) = [g^{\alpha\beta}] \tag{5.86}$$

Let the asymptotic mean square error of $\{\hat{u}'_N,\ N = 1, 2, \cdots\}$ be

$$\lim_{N\to\infty} NE[(\hat{u}'^a_N - u^a)(\hat{u}'^b_N - u^b)] \tag{5.87}$$

is the $(a, b)^{\text{th}}$ component of $[g^{\alpha\beta}]$ from Equation (5.86) and is denoted by $\bar{g}'^{ab}$.
Taking the inverse of the matrix $[g_{\alpha\beta}]$ in Equation (5.74) and also from Equation(5.86) we get

$$[\bar{g}'^{ab}] = [g_{ab} - g_{a\kappa} g^{\kappa\lambda} g_{b\lambda}]^{-1} \quad \Rightarrow \quad [\bar{g}'^{ab}] \geq [g^{ab}] \tag{5.88}$$

which is the asymptotic Cramer-Rao inequality, where $[g^{ab}]$ is the inverse of $[g_{ab}]$.
Then the consistent estimator $\{\hat{u}'_N,\ N = 1, 2, \cdots\}$ is first order efficient iff

$$[\bar{g}'^{ab}] = [g^{ab}] \tag{5.89}$$

which is iff $g_{a\kappa} = 0,\ \forall\ a, \kappa.$
Note that $g_{ak} =< \partial_a, \partial_k >$ is the inner product of the tangent vector $\partial_a \in T_{\eta(u)}M$ and the tangent vector $\partial_k \in T_{\eta(u)}A'(u)$. Hence consistent estimator $\{\hat{u}'_N,\ N = 1, 2, \cdots\}$ is first order efficient iff $A'(u)$ is orthogonal to M at $\eta(u)$. □

Remark 5.3.3. *Theorems 5.3.1 and 5.3.2 give the necessary and sufficient condition for an estimator based on a mismatched model to be consistent and efficient. To achieve this one has to choose the mismatched model suitably for each estimator.*

5.3.1 MLE based on a mismatched model

Here we consider the maximum likelihood estimation based on a mismatched model. Ozumi et al. [48] stated certain conditions for the MLE based on a mismatched model to be consistent and efficient. We give a theoretical formulation of this mismatched maximum likelihood estimation problem along with a detailed proof of the same.

Let $\mathcal{S} = \{p(x; \theta)\ /\ \theta \in \mathbb{R}^n\}$ be an n-dimensional exponential family. Consider a curved exponential family $M = \{q(x; u)\ /\ u = (u^a) \in \mathbb{R}^m\}$ of $\mathcal{S}$ and let $M^* = \{q'(x; u)\ /\ u = (u^a) \in \mathbb{R}^m\}$ be a mismatched model corresponding to the original model M. Let the embedding functions of M and M^* in $\mathcal{S}$ be $\theta(u)$ and $\theta'(u)$ respectively. Let $\eta(u)$ and $\eta'(u)$ be the corresponding dual representations.

Consider N independent observations $\mathbf{x}_N = (x^1, \cdots, x^N)$ from $q(x; u) \in M$. Then the MLE $\hat{u}'_N$ for M^* is determined from the log-likelihood function

$$\ell'(\bar{x}; u) = \log p(\mathbf{x}_N; \theta'(u)) = N \left[\sum_{i=1}^{n} \theta'^i(u)\bar{x}_i - \psi(\theta'(u)) \right] \tag{5.90}$$

as

$$\frac{\partial}{\partial u^a} \log p(\mathbf{x}_N; \theta'(u)) \mid_{\hat{u}'_N} = 0, \ a = 1, \cdots, m \tag{5.91}$$

$$\sum_{i=1}^{n} \frac{\partial \theta'^i}{\partial u^a}(\hat{u}'_N) \left(\bar{x}_i - \eta'_i(\hat{u}'_N)\right) = 0, \ a = 1, \cdots, m. \tag{5.92}$$

Remark 5.3.4. *In general, the estimating function or the ancillary manifold of the MLE depends upon the number of observations N explicitly. But here we consider the case where the estimating function or ancillary manifold of the MLE does not depend upon N explicitly. Thus at each $u \in M^*$, we have $A'(u)$ instead of $A'_N(u)$. Also we denote the estimator $\hat{u}'_N$ by $\hat{u}'$ and $\hat{\eta}_N$ by $\hat{\eta}$.*

From Equation (5.92) the ancillary submanifold $A'(u)$ associated with MLE can be written as

$$A'(u) = \{\eta = (\eta_i) \in \mathcal{S} \ / \ \sum_{i=1}^{n} \frac{\partial \theta'^i}{\partial u^a}(u) \left(\eta_i - \eta'_i(u)\right) = 0, \ \ a = 1, \cdots, m\} \tag{5.93}$$

Note that $\eta'(u) \in A'(u)$ for all $u \in M^*$.

Lemma 5.3.5. *The MLE $\hat{u}'$ in M^* is the point $u \in M^*$ which minimizes (-1)-divergence from the observed point $\hat{\eta}$ to M^*.*

Proof. The (-1)-divergence $D_{-1}(\hat{\theta}, \theta'(u))$ from the observed point $\hat{\theta}$ to the submanifold M^* is given by

$$\begin{aligned} D_{-1}(\hat{\theta}, \theta'(u)) &= D_1(\eta'(u), \hat{\eta}) && (5.94) \\ &= \psi(\theta'(u)) + \phi(\hat{\eta}) - \sum_{i=1}^{n} \theta'^i(u)\hat{\eta}_i && (5.95) \\ &= \phi(\hat{\eta}) - \frac{1}{N}\ell'(\bar{x}; u) && (5.96) \end{aligned}$$

From equation (5.96) it follows that maximizing the log-likelihood ℓ' is same as minimizing the (-1)-divergence from the observed point to M^*. Hence the MLE $\hat{u}'$ in M^*

is the point $u \in M^*$ which minimizes (-1)-divergence from the observed point $\hat{\eta}$ to M^*. □

Corollary 5.3.6. *By the projection theorem the MLE $\hat{u}'$ is the (-1)-projection from the observed point $\hat{\eta}$ to M^*. Thus the ancillary manifold $A'(u)$ contains all (-1)-geodesics which orthogonally intersects M^* at $\eta'(u)$ and $A'(u)$ is orthogonal to M^* at $\eta'(u)$. Then*

$$A'(u) = \{\eta = (\eta_i) \in \mathcal{S} \ / \ \min_{v \in M^*} D_{-1}(\eta, \eta'(v)) = D_{-1}(\eta, \eta'(u))\} \tag{5.97}$$

Now we give an example to show that arbitrary choice of mismatched model M^* may not make the MLE consistent.

Example 5.3.7. Let $\mathcal{S} = \{p(x;\theta)\}$ be a set of normal distributions with mean μ and variance σ^2.

$$\mathcal{S} = N(\mu,\sigma) = \left\{ p(x;\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{(x-\mu)^2}{2\sigma^2} \right) / \mu \in \mathbb{R}, \sigma > 0 \right\} \tag{5.98}$$

For this 2-dimensional exponential family the canonical coordinate $\theta = (\theta^1, \theta^2)$, potential function $\psi(\theta)$ and the dual coordinates $\eta = (\eta_1, \eta_2)$ are given by

$$\theta = (\theta^1, \theta^2) = (\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}) \tag{5.99}$$

$$\eta = (\eta_1, \eta_2) = (\mu, \mu^2 + \sigma^2) \tag{5.100}$$

$$\psi(\theta) = \frac{-(\theta^1)^2}{4\theta^2} - \frac{1}{2}\log(-\theta^2) + \frac{1}{2}\log\pi. \tag{5.101}$$

Let $M = \{q(x;u)\}$ be the set of normal distributions with $\mu = u$ and $\sigma = u$.

$$M = N(u,u) = \left\{ q(x;u) = \frac{1}{\sqrt{2\pi}u} \exp\left(\frac{(x-u)^2}{2u^2} \right) / u > 0 \right\} \tag{5.102}$$

Then M is a smooth submanifold of $\mathcal{S}$ parametrized by $u > 0$.
The embedding functions $\theta(u)$ and $\eta(u)$ of M are given by

$$\theta(u) = (\theta^1(u), \theta^2(u)) = (\frac{1}{u}, \frac{-1}{2u^2}) \tag{5.103}$$

$$\eta(u) = (\eta_1(u), \eta_2(u)) = (u, 2u^2) \tag{5.104}$$

Let $M^* = \{q'(x;u)\}$ be the set of normal distributions with $\mu = u$ and $\sigma = \sqrt{2}u$.

$$M^* = N(u, \sqrt{2}u) = \left\{ q'(x;u) = \frac{1}{2\pi u} \exp\left(\frac{(x-u)^2}{4u^2} \right) / u > 0 \right\} \tag{5.105}$$

Then M^* is a smooth submanifold of $\mathcal{S}$ parametrized by $u > 0$.
The embedding functions $\theta'(u)$ and $\eta'(u)$ of M^* are given by

$$\theta'(u) = (\theta'^1(u), \theta'^2(u)) = (\frac{1}{2u}, \frac{-1}{4u^2}) \tag{5.106}$$

$$\eta'(u) = (\eta'_1(u), \eta'_2(u)) = (u, 3u^2) \tag{5.107}$$

The ancillary submanifolds $A(u)$ and $A'(u)$ associated with MLEs $\hat{u}, \hat{u}'$ in M and M^* respectively are given by

$$A(u) = \{\eta = (\eta_1, \eta_2) \in \mathcal{S} \ / \ u^2 + u\eta_1 - \eta_2 = 0\} \tag{5.108}$$

$$A'(u) = \{\eta = (\eta_1, \eta_2) \in \mathcal{S} \ / \ 2u^2 + u\eta_1 - \eta_2 = 0\} \tag{5.109}$$

We can see that $\eta(u)$ does not belong to $A'(u)$ for all $u > 0$. Thus according to Theorem 5.3.1 the estimator $\hat{u}'$ is not consistent. Hence the mismatched model that we selected is not a good choice for the original model.

Now we describe the conditions for the MLE based on a mismatched model M^* to be consistent and first order efficient.

Theorem 5.3.8. *Let $\hat{u}'$ be the MLE in M^*. Then $\hat{u}'$ is a consistent estimator of u iff*

$$q'(x;u) = \arg \min_{v \in M^*} D_{-1}(q(x;u), q'(x;v)) \tag{5.110}$$

Proof. If

$$q'(x;u) = \min_{v \in M^*} D_{-1}(q(x;u), q'(x;v)) \tag{5.111}$$

then the (-1)-geodesic (∇^m geodesic) connecting $\eta(u)$ and $\eta'(u)$ are orthogonal to M^* at the point $\eta'(u) \in M^*$. Since $A'(u)$ contains all (-1)-geodesics which orthogonally intersect M^* at $\eta'(u)$,

$$\eta(u) \in A'(u) \tag{5.112}$$

Then by Theorem 5.3.1 the MLE $\hat{u}'$ is a consistent estimator.

Conversely, if the MLE $\hat{u}'$ is consistent, $\eta(u) \in A'(u)$. Then from Corollary 5.3.6 we obtain Equation (5.111). □

Theorem 5.3.9. *Let $\hat{u}'$ be the consistent MLE in M^*. Then $\hat{u}'$ is first order efficient iff*

$$q(x;u) = \arg\min_{v \in M} D_{-1}(q'(x;u), q(x;v)) \tag{5.113}$$

Proof. Let $\hat{u}'$ be consistent and Equation (5.113) holds. Then $\eta(u) \in A'(u)$ and (-1)-geodesic connecting $\eta(u)$ and $\eta'(u)$ is orthogonal to M at the point $\eta(u) \in M$. From Corollary 5.3.6 we have (-1)-geodesic connecting $\eta(u)$ and $\eta'(u)$ is orthogonal to M^* at the point $\eta'(u) \in M^*$.

To show that $\hat{u}'$ is first order efficient we show that $A'(u)$ is orthogonal to M at the point $\eta(u) \in M$.

The ancillary submanifold $A'(u)$ associated with $\hat{u}'$ is

$$A'(u) = \{\eta = (\eta_i) \in \mathcal{S} \;/\; \sum_{i=1}^{n} \frac{\partial \theta'^i}{\partial u^a}(u)\,(\eta_i - \eta'_i(u)) = 0, \;\; a = 1, \cdots, m\} \tag{5.114}$$

Thus $A'(u)$ is a linear submanifold in η and hence it is (-1)-flat (∇^m flat) passing through $\eta'(u)$.

Let $Z = \eta - \eta(u)$ and $B'^i_a = \frac{\partial \theta'^i}{\partial u^a}(u)$. Then

$$\sum_{i=1}^{n} B'^i_a Z_i = 0, \; a = 1, \cdots, m \tag{5.115}$$

Let $Z^k, \; k = m+1, \cdots, n$ be the $n-m$ independent solutions of Equation (5.115) and the i^{th} component of the k^{th} independent vector Z^k be denoted by Z_{ki}. Then

$$Z = \eta - \eta(u) = \sum_{k=m+1}^{n} v^k Z^k, \quad \text{for} \quad v^k \in \mathbb{R} \tag{5.116}$$

Thus for any $\eta = (\eta_i) \in A'(u)$,

$$\eta_i = \sum_{k=m+1}^{n} v^k Z_{ki} + \eta_i(u) \tag{5.117}$$

We need to show that

$$< \frac{\partial}{\partial v^k} \mid_{\eta(u)}, \frac{\partial}{\partial u^a} \mid_{\eta(u)} >= 0, \quad \forall \quad a = 1, \cdots, m \text{ and } k = m+1, \cdots, n. \tag{5.118}$$

Let

$$\partial_k = \frac{\partial}{\partial v^k} \;=\; \sum_{i=1}^{n} \frac{\partial \eta_i}{\partial v^k} \partial^i \tag{5.119}$$

$$\partial_a = \frac{\partial}{\partial u^a} \;=\; \sum_{j=1}^{n} \frac{\partial \theta^j}{\partial u^a} \partial_j \tag{5.120}$$

where $\partial^i = \frac{\partial}{\partial \eta_i}$ and $\partial_j = \frac{\partial}{\partial \theta^j}$. From equation (5.117),

$$\frac{\partial \eta_i}{\partial v^k} = Z_{ki} \tag{5.121}$$

Then

$$< \partial_k \mid_{\eta(u)}, \partial_a \mid_{\eta(u)} > \;=\; \sum_{i=1}^{n} \sum_{j=1}^{n} Z_{ki} B_a^j < \partial^i, \partial_j > \tag{5.122}$$

$$=\; \sum_{i=1}^{n} Z_{ki} B_a^i \tag{5.123}$$

where $B_a^j = \frac{\partial \theta^j}{\partial u^a}(u)$.

Now consider the (-1)-geodesic γ connecting $\eta'(u) \in M^*$ and $\eta(u) \in M$

$$\gamma(t) = t\eta'(u) + (1-t)\eta(u), \quad \text{where} \quad t \in [0,1]. \tag{5.124}$$

Then the tangent vector to γ is given by

$$\dot{\gamma} = \sum_{j=1}^{n} (\eta'_j(u) - \eta_j(u)) \partial^j \tag{5.125}$$

Since γ is orthogonal to M at $\eta(u)$

$$L_a =< \sum_{j=1}^{n} (\eta'_j(u) - \eta_j(u)) \partial^j, \partial_a >= 0, \quad \forall \quad a = 1, \cdots, m \tag{5.126}$$

Since $\eta(u) \in A'(u)$, from Equation (5.117)

$$\eta'_i(u) - \eta_i(u) = \sum_{k=m+1}^{n} v_0^k(u) Z_{ki}, \quad \text{for some} \quad v_0^{k.} \tag{5.127}$$

Thus Equation (5.126) can be written as

$$\begin{aligned}
L_a &= < \sum_{j=1}^{n} (\eta'_j(u) - \eta_j(u)) \partial^j, \sum_{i=1}^{n} B_a^i \partial_i > && (5.128)\\
&= < \sum_{j=1}^{n} (\sum_{k=m+1}^{n} v_0^k(u) Z_{kj}) \partial^j, \sum_{i=1}^{n} B_a^i \partial_i > && (5.129)\\
&= \sum_{j=1}^{n} (\sum_{k=m+1}^{n} v_0^k(u) Z_{kj}) \sum_{i=1}^{n} B_a^i < \partial^j, \partial_i > && (5.130)\\
&= \sum_{i=1}^{n} \sum_{k=m+1}^{n} v_0^k(u) Z_{ki} B_a^i && (5.131)\\
&= \sum_{k=m+1}^{n} v_0^k(u) (\sum_{i=1}^{n} Z_{ki} B_a^i) = 0 && (5.132)
\end{aligned}$$

Since $v_o^k(u) \neq 0$

$$\sum_{i=1}^{n} Z_{ki} B_a^i = 0, \quad \forall \quad a = 1, \cdots, m \text{ and } k = m+1, \cdots, n. \tag{5.133}$$

Thus from Equation (5.123)

$$< \partial_k \mid_{\eta(u)}, \partial_a \mid_{\eta(u)} > \; = \; \sum_{i=1}^{n} Z_{ki} B_a^i = 0 \tag{5.134}$$

That is, $A'(u)$ is orthogonal to M at the point $\eta(u) \in M$. Then by Theorem 5.3.2 the consistent MLE $\hat{u}'$ is an efficient estimator.

The converse trivially holds. □

Corollary 5.3.10. *Let $\hat{u}'$ be the MLE in M^*. Let γ be the (-1)-geodesic connecting $q(x;u) \in M$ and $q'(x;u) \in M^*$. Then*

1. *The MLE $\hat{u}'$ is consistent iff γ is orthogonal to M^*.*

2. *The consistent MLE $\hat{u}'$ is first order efficient iff γ is orthogonal to both M and M^*.*

5.4 Summary

In this chapter we first discussed the geometric theory of parameter estimation problem in an exponential family and in a curved exponential family given by Amari [11], [12]. Further the estimation problem based on a mismatched model in an exponential family is considered. We proved a necessary and sufficient condition for an estimator based on a mismatched model to be consistent and efficient. Ozumi et al. [48] stated certain conditions for MLE based on a mismatched model to be consistent and efficient. We gave a theoretical formulation of these results and a detailed proof of the same.

CHAPTER 6

Generalized Estimators

In this chapter first we discuss about certain generalized notions of maximum likelihood estimator. Further we look at the estimation problem in a deformed exponential family. In the context of nonextensive thermostatistics Umarov et al. [49] defined the notion of q-independence and q-central limit theorem using a generalized product called q-product, see also [50], [51]. Ferrari and Yang [52] defined a maximum L_q-estimator (ML$_q$E) based on nonextensive entropy (q-entropy) and studied its asymptotic behavior in the case of an exponential family. Using the q-product Matsuzoe and Ohara [53] also considered the q-independence and the q-likelihood estimator. Fujimoto and Murata [54] defined a more generalized notion of independence called the U-independence using a smooth strictly convex function U. Eguchi et al. [36] defined the U-estimator and discussed its consistency and asymptotic normality. Naudts [28] defined a generalized Cramer-Rao bound and showed that this bound is optimal in a deformed exponential family.

In Section 6.1 we define notions like F-product, F-independence using a function F and its inverse function Z. Then a generalized MLE called the maximum F-likelihood estimator (F-MLE) is defined and discussed its property as a MAP estimator. In Section 6.2 using the F-escort probability distribution we define two generalized notions of MLE, the $\mathbf{x}_N$-based F-escort MLE and the F-escort MLE based on the product of F-escort distribution of the marginal probability density of single observations. Then a characterization of the q-escort MLE among the $\mathbf{x}_N$ based F-escort MLE as a Bayesian MAP estimator with a prior is given. Further an analytic proof of the F-version of the maximum entropy theorem is given. In Section 6.3 first we describe the U-estimator in a deformed exponential family. Then a proof of the generalized Cramer-Rao bound defined by Naudts is given. Further we show that the U-estimator attains equality in this bound. This chapter ends with an open problem regarding the properties of the F-MLE in a deformed exponential family.

6.1 Maximum F-Likelihood Estimator

In this section using a function F and its inverse Z the F-product and the F- independence of random variables are defined. Then the F-MLE is defined on a statistical manifold and show that the F-MLE is a MAP estimator with a prior distribution.

Two random variables X and Y are said to be **independent** if the joint probability density function $p(x, y)$ is given by the product of the marginal probability density functions $p_1(x)$ and $p_2(y)$.

$$p(x, y) = p_1(x)p_2(y) \tag{6.1}$$

Using the properties of the logarithm and exponential functions the above equation can be written as

$$p(x, y) = \exp[\log p_1(x) + \log p_2(y)] \tag{6.2}$$

for positive $p_1(x)$ and $p_2(y)$.

The q**-product** [49] of two positive numbers x, y using the q-logarithm $\log_q$ and the q-exponential $\exp_q$ is defined as

$$x \otimes_q y = \exp_q[\log_q x + \log_q y] = [x^{1-q} + y^{1-q} - 1]^{\frac{1}{1-q}} \tag{6.3}$$

The q-product satisfies the following properties

$$\begin{aligned} \exp_q x \otimes_q \exp_q y &= \exp_q(x + y) && (6.4) \\ \log_q(x \otimes_q y) &= \log_q x + \log_q y && (6.5) \end{aligned}$$

Two random variables X and Y are said to be q**-independent** with normalization if the joint probability density function $p(x, y)$ is given by [53]

$$p(x, y) = \frac{p_1(x) \otimes_q p_2(y)}{K_{p_1,p_2}} \tag{6.6}$$

where K_{p_1,p_2} is the normalization defined by

$$K_{p_1,p_2} = \int_{\mathcal{X}} \int_{\mathcal{Y}} p_1(x) \otimes_q p_2(y) dx\, dy \tag{6.7}$$

Now we define a notion of independence called the F-independence which extends the

q-independence.

Definition 6.1.1. *Let $F : (0, \infty) \longrightarrow \mathbb{R}$ be a smooth function satisfying $F'(x) > 0$ and $F''(x) < 0$ and let Z be its inverse function. Define the F**-product** of two numbers x, y as*

$$x \otimes_F y = Z[F(x) + F(y)] \tag{6.8}$$

(assume that $F(x) + F(y) \in \text{Domain}(Z)$).
The F-product satisfies the following properties

$$\begin{aligned} Z(x) \otimes_F Z(y) &= Z(x + y) && (6.9) \\ F(x \otimes_F y) &= F(x) + F(y) && (6.10) \end{aligned}$$

Define $p_F(x, y)$ as

$$p_F(x, y) = \frac{p_1(x) \otimes_F p_2(y)}{K_{p_1,p_2}} = \frac{Z[F(p_1(x)) + F(p_2(y))]}{K_{p_1,p_2}} \tag{6.11}$$

where K_{p_1,p_2} is the normalization defined by

$$K_{p_1,p_2} = \int_{\mathcal{X}} \int_{\mathcal{Y}} p_1(x) \otimes_F p_2(y) dx dy \tag{6.12}$$

Definition 6.1.2. *Two random variables X and Y are said to be F**-independent** if the joint probability density function $p(x, y)$ is equal to $p_F(x, y)$.*

Note 6.1.3. *Fujimoto and Murata [54] defined a generalized notion of independence called the U-independence using a smooth strictly convex function U. They first defined generalized arithmetic operations called U**-multiplication** and U**-division** as*

$$x \otimes y = u[\xi(x) + \xi(y)], \quad x \oslash y = u[\xi(x) - \xi(y)] \tag{6.13}$$

where $u(.) = U'(.)$ and ξ is the inverse of u.
*Let $p_u(x, y) = u(\xi(p_1(x)) + \xi(p_2(y)) - c_u)$, where c_u is the normalization constant determined from $\sum_{\mathcal{X},\mathcal{Y}} p_u(x, y) = 1$. Then two random variables X, Y are said to be U**-independent** if their joint probability density function $p(x, y)$ is equal to $p_u(x, y)$. That is,*

$$p(x, y) = p_u(x, y) = u\left(\xi(p_1(x)) + \xi(p_2(y)) - c_u\right) \tag{6.14}$$

Note that for the F-independence we divide the F-product of the densities by a normalizing constant, whereas in U-independence the normalizing factor is subtracted.

Definition 6.1.4. *Let $\mathcal{S} = \{p(x;\theta) \ / \ \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$ be an n-dimensional statistical manifold. Let $\mathbf{x}_N = (x^1, \cdots, x^N)$ be N independent observations from a probability density function $p(x;\theta) \in \mathcal{S}$. Define a generalized likelihood function called the **F-likelihood function** $L_F(\theta)$*

$$L_F(\theta) = p(x^1;\theta) \otimes_F \cdots \otimes_F p(x^N;\theta) = Z(\sum_{i=1}^{N} F(p(x^i;\theta))) \tag{6.15}$$

Since F is an increasing function it is equivalent to consider $F(L_F(\theta))$.

$$F(L_F(\theta)) = F(p(x^1;\theta) \otimes_F \cdots \otimes_F p(x^N;\theta)) = \sum_{i=1}^{N} F(p(x^i;\theta)) \tag{6.16}$$

*Estimator $\hat{\theta}_F$ is the **maximum F-likelihood estimator (F-MLE)** if*

$$\hat{\theta}_F = \arg\max_{\theta\in\mathbb{E}} L_F(\theta) = \arg\max_{\theta\in\mathbb{E}} F(L_F(\theta)) \tag{6.17}$$

Definition 6.1.5. *Let $p(x \mid \theta)$ be a distribution of the random variable x which depends on an unobserved population parameter θ and $p(\theta)$ be a prior distribution of θ. Then the posterior distribution $p(\theta \mid x)$ of θ is given by*

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{p(x)} \tag{6.18}$$

where $p(x)$ is the marginal density function of x given by

$$p(x) = \int_{\mathbb{E}} p(x \mid \theta) \ p(\theta) \ d\theta \tag{6.19}$$

*Let $\mathbf{x}_N = (x^1, \cdots, x^N)$ be N independent observations. Then the **maximum a posteriori probability (MAP)** estimator $\hat{\theta}_{\text{MAP}}$ for θ is given by*

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta\in\mathbb{E}} p(\theta \mid \mathbf{x}_N) = \arg\max_{\theta\in\mathbb{E}} p(\mathbf{x}_N \mid \theta)p(\theta) \tag{6.20}$$

Theorem 6.1.6. *Let $\mathbf{x}_N = (x^1, \cdots, x^N)$ be F-independent observations from $p(x \mid \theta)$, where F is a smooth function other than logarithmic function with $F' > 0, F'' < 0$.*

Then the F-MLE is a MAP estimator with the prior distribution $p(\theta)$ given by

$$p(\theta) = \frac{K(\theta)}{K_1}, \quad K_1 = \int K(\theta)\, d\theta < \infty \tag{6.21}$$

where

$$\begin{aligned} K(\theta) &= \int \cdots \int p(x^1;\theta) \otimes_F \cdots \otimes_F p(x^N;\theta)\, d\mathbf{x}_N & (6.22) \\ &= \int \cdots \int Z\left(\sum_{i=1}^{N} F(p(x^i;\theta))\right)\, d\mathbf{x}_N & (6.23) \end{aligned}$$

and $d\mathbf{x}_N = dx_1 \cdots dx_N$.

Proof. Since $\mathbf{x}_N = (x^1, \cdots, x^N)$ are F-independent, from Equations (6.11), (6.12) and (6.15) the joint probability density function $p(d\mathbf{x}_N \mid \theta)$ is

$$p(\mathbf{x}_N \mid \theta) = \frac{p(x^1 \mid \theta) \otimes_F \cdots \otimes_F p(x^N \mid \theta)}{K(\theta)} \tag{6.24}$$

where

$$\begin{aligned} K(\theta) &= \int \cdots \int p(x^1 \mid \theta) \otimes_F \cdots \otimes_F p(x^N \mid \theta)\, d\mathbf{x}_N & (6.25) \\ &= \int \cdots \int Z\left(\sum_{i=1}^{N} F(p(x^i \mid \theta))\right) d\mathbf{x}_N & (6.26) \end{aligned}$$

Let $K_1 = \int K(\theta) d\theta$. The F-MLE $\hat{\theta}_F$ is given by

$$\hat{\theta}_F = \arg\max_{\theta \in \mathbb{E}} L_F(\theta) = \arg\max_{\theta \in \mathbb{E}} p(x^1 \mid \theta) \otimes_F \cdots \otimes_F p(x^N \mid \theta) \tag{6.27}$$

We have

$$\begin{aligned} \frac{L_F(\theta)}{K_1} &= \frac{p(x^1 \mid \theta) \otimes_F \cdots \otimes_F p(x^N \mid \theta)}{K_1} & (6.28) \\ &= \frac{K(\theta)}{K_1} p(\mathbf{x}_N \mid \theta) & (6.29) \end{aligned}$$

Hence

$$\begin{aligned}
\hat{\theta}_F &= \arg\max_{\theta\in\mathbb{E}} L_F(\theta) = \arg\max_{\theta\in\mathbb{E}} \frac{L_F(\theta)}{K_1} && (6.30)\\
&= \arg\max_{\theta\in\mathbb{E}} \frac{K(\theta)}{K_1} p(\mathbf{x}_N \mid \theta) && (6.31)\\
&= \arg\max_{\theta\in\mathbb{E}} p(\theta)\; p(\mathbf{x}_N \mid \theta) && (6.32)\\
&= \hat{\theta}_{\mathrm{MAP}} && (6.33)
\end{aligned}$$

with the prior distribution $p(\theta)$ of θ given by

$$p(\theta) = \frac{K(\theta)}{K_1}, \quad K_1 = \int K(\theta)\; d\theta. \tag{6.34}$$

That is, the F-MLE is a MAP estimator with $p(\theta)$ as prior distribution of θ. □

6.2 F-Escort Maximum Likelihood Estimator

The escort probability distributions are studied in the context of nonextensive statistics and related areas [28–30], [62]. In the study of the geometry of the q-exponential family Amari and Ohara [27] considered an escort distribution called the q-escort distribution and defined an estimator called the q-escort maximum likelihood estimator (q-escort MLE). In Chapter 4 the F-escort probability distribution is defined and using this now we define a generalized maximum likelihood estimator.

Consider a statistical manifold $\mathcal{S} = \{p(x;\theta)\}$. For a distribution $p(x;\theta) \in \mathcal{S}$ the **F-escort probability distribution** $\hat{p}_F$ is defined by

$$\hat{p}_F(x;\theta) = \frac{1}{h_F(\theta)F'(p)}, \quad \text{where} \quad h_F(\theta) = \int \frac{1}{F'(p)} dx \tag{6.35}$$

Let $\mathcal{S}'$ be the manifold consisting of the F-escort probability distributions

$$\mathcal{S}' = \{\hat{p}_F(x;\theta) \;/\; \theta \in \mathbb{E} \subseteq \mathbb{R}^n\} \tag{6.36}$$

One can define an estimator using the F-escort probability distribution instead of the original distribution.

Let $\mathbf{x}_N = (x^1, \cdots, x^N)$ be N independent observations from $\mathcal{S} = \{p(x;\theta)\}$. Then the joint probability density function $p(\mathbf{x}_N;\theta)$ is

$$p(\mathbf{x}_N;\theta) = \prod_{i=1}^{N} p(x^i;\theta) \tag{6.37}$$

Consider the F-escort distribution of the joint probability density $p(\mathbf{x}_N;\theta)$

$$\hat{p}_F(\mathbf{x}_N;\theta) = \frac{1}{F'(p(\mathbf{x}_N;\theta))} \frac{1}{\int \frac{1}{F'(p(\mathbf{x}_N;\theta))} d\mathbf{x}_N} \tag{6.38}$$

Definition 6.2.1. *For N independent observations $\mathbf{x}_N = (x^1, \cdots, x^N)$ from $p(x;\theta) \in \mathcal{S}$ the $\mathbf{x}_N$* ***based F-escort MLE*** *$\hat{\theta}_F$ is defined as the maximizer of the F-escort distribution $\hat{p}_F(\mathbf{x}_N;\theta)$ of the joint probability density $p(\mathbf{x}_N;\theta)$. That is,*

$$\hat{\theta}_F = \arg\max_{\theta \in \mathbb{E}} \hat{p}_F(\mathbf{x}_N;\theta). \tag{6.39}$$

Remark 6.2.2. *It is clear that in general the F-escort MLE $\hat{\theta}_F$ is different from the ordinary MLE. When $F(p) = \log p$, the F-escort MLE is the MLE.*

Now for N independent observations $\mathbf{x}_N = (x^1, \cdots, x^N)$ from $p(x;\theta) \in \mathcal{S}$ we can consider the product of F-escort distribution of the marginal probability density of single observations x^i given by

$$\prod_{i=1}^{N} \hat{p}_F(x^i;\theta) \tag{6.40}$$

Note that in general

$$\hat{p}_F(\mathbf{x}_N;\theta) \neq \prod_{i=1}^{N} \hat{p}_F(x^i;\theta) \tag{6.41}$$

Thus one can consider two types of F-escort MLE's, the F-escort MLE based on $\mathbf{x}_N$ and the F-escort MLE based on the product of F-escort distribution of the marginal probability density of single observations x^i.

The $\mathbf{x}_N$-based F-escort MLE is the maximizer of

$$\hat{p}_F(\mathbf{x}_N;\theta). \tag{6.42}$$

The F-escort MLE based on the product of F-escort distribution of the marginal

probability density of single observations x^i is the maximizer of

$$\prod_{i=1}^{N} \hat{p}_F(x^i;\theta) = \prod_{i=1}^{N} \frac{1}{F'(p(x^i;\theta))(h_F(\theta))^N}. \tag{6.43}$$

Note that the geometries of the two types of F-escort MLEs are different. But when $F(p) = \log_q p$

$$\hat{p}_F(\mathbf{x}_N;\theta) = \prod_{i=1}^{N} \hat{p}_F(x^i;\theta) = \prod_{i=1}^{N} \frac{1}{F'(p(x^i;\theta))(h_F(\theta))^N}. \tag{6.44}$$

Thus the two geometries coincide in this case. It will be an interesting problem to find the relation between the two geometries for a general F and also to study the properties of the F-MLE in a deformed exponential family.

Amari and Ohara [27] gave an interpretation of the q-escort MLE as a Bayesian MAP with the prior distribution $p(\theta) = (h_q(\theta))^{\frac{-N}{q}}$.
Now we show that this property can be used as a characterization of the q-escort MLE among the F-escort MLE.

Theorem 6.2.3. *The $\mathbf{x}_N$-based F-escort MLE is a Bayesian MAP with a prior distribution $p(\theta)$ only when the F-escort MLE is the q-escort MLE.*

Proof. A Bayesian MAP $\hat{\theta}_{\mathrm{MAP}}$ satisfies

$$\begin{aligned} \hat{\theta}_{\mathrm{MAP}} &= \arg\max_{\theta\in\mathbb{E}} p(\theta \mid \mathbf{x}_N) = \arg\max_{\theta\in\mathbb{E}} p(\theta)p(\mathbf{x}_N \mid \theta) && (6.45)\\ &= \arg\max_{\theta\in\mathbb{E}} \frac{1}{F'(p(\theta)p(\mathbf{x}_N,\theta))} && (6.46) \end{aligned}$$

with prior $p(\theta)$.
(For proving Equation (6.46), let $g(u) = \frac{1}{F'(u)}$, then $g'(u) = \frac{-F''(u)}{(F'(u))^2} > 0$ since $F''(u) < 0$. That is, g is a monotonically increasing function.)
The F-escort MLE $\hat{\theta}_F$ is the maximizer of

$$\hat{p}_F(\mathbf{x}_N;\theta) = \frac{1}{F'(p(\mathbf{x}_N;\theta))} \frac{1}{\int \frac{1}{F'(p(\mathbf{x}_N;\theta))} d\mathbf{x}_N} \tag{6.47}$$

Then the two estimators are identical if

$$\frac{1}{F'(p(\mathbf{x}_N;\theta))}\frac{1}{\int \frac{1}{F'(p(\mathbf{x}_N;\theta))}d\mathbf{x}_N} = \frac{1}{F'(p(\theta)p(\mathbf{x}_N;\theta))}. \tag{6.48}$$

This implies that F' is a homogeneous function of some order k. Then $F'(p) = p^k$ and $h_F(\theta) = h_k(\theta) = \int (p(\mathbf{x}_N;\theta))^k \, d\mathbf{x}_N$. Thus from Equation (6.48), $p(\theta) = (h_k(\theta))^{\frac{-N}{k}}$. Hence the $\mathbf{x}_N$-based F-escort MLE is a Bayesian MAP with a prior distribution $p(\theta)$ only when it is the q-escort MLE. □

6.2.1 F-Maximum entropy theorem

The Shannon entropy or information entropy or Boltzmann-Gibbs-Shannon entropy plays a major role in the areas of information theory and statistical thermodynamics. According to Boltzmann theorem probability distributions maximizing the Shannon entropy under a finite number of moment constraints form a finite dimensional exponential family. There are many generalizations of the Shannon entropy in the existing literature [22], [26], [27], [37]. One among them is the well known Tsallis entropy [26]. The maximization of Tsallis entropy under appropriate constraints leads to Tsallis distribution or the q-exponential family. Amari et al. [27] defined a χ-entropy and gave a geometric proof of the χ-version of the maximum entropy theorem. Here we present an analytic proof of the same using the F-formulation of the deformed exponential family.

Definition 6.2.4. *For any probability density function $p(x)$ the F-entropy is defined as*

$$H_F(p) = -E_{\hat{p}_F}(F(p)) = \frac{1}{h_F(p)}\int \frac{-F(p)}{F'(p)}dx \tag{6.49}$$

if $\int \frac{-F(p)}{F'(p)}dx$ and $h_F(p) = \int \frac{1}{F'(p)}dx$ exist.

When $F(p) = \ln_q p$, $H_F(p)$ reduces to the q**-entropy** $H_q(p) = \frac{1}{1-q}\left(1 - \frac{1}{h_q(p)}\right)$ and when $F(p) = \ln p$, $H_F(p)$ reduces to the **Shannon entropy** $H(p) = -\int p(x)\ln p(x)\, dx$.

Theorem 6.2.5. *Probability distributions maximizing the F-entropy H_F under the F-linear constraints*

$$E_{\hat{p}_F}[c_k(x)] = a_k, \quad k = 1, \cdots, m \tag{6.50}$$

where $c_k(x)$ are m random variables and $a_k \in \mathbb{R}$, form an m-dimensional F-exponential *family*

$$F(p(x;\theta)) = \sum_{i=1}^{m} \theta^i c_i(x) - \psi(\theta) \tag{6.51}$$

where $\theta = (\theta^1, \cdots, \theta^m)$ is the canonical coordinate and $\psi(\theta)$ can be determined from the normalization condition.

Proof. We use the method of Lagrange multipliers and the calculus of variation principle.

To maximize the F-entropy $H_F(p) = \frac{1}{h_F(p)} \int \frac{-F(p)}{F'(p)} dx$ subject to the m constraints

$$E_{\hat{p}_F}[c_k(x)] = \frac{1}{h_F(p)} \int \frac{c_k(x)}{F'(p)} dx = a_k; \quad k = 1, \cdots, m \tag{6.52}$$

consider,

$$\begin{aligned} \mathcal{L}(p, \lambda_0, \lambda_1, \cdots, \lambda_m) &= \frac{1}{h_F(p)} \int_0^{\infty} \frac{-F(p)}{F'(p)} dx + \lambda_0 \left(\int_0^{\infty} p dx - 1 \right) \\ &+ \sum_{i=1}^{m} \lambda_i \left(\frac{1}{h_F(p)} \int_0^{\infty} \frac{c_k(x)}{F'(p)} dx - a_i \right) \qquad (6.53) \\ &= \frac{1}{h_F(p)} \int_0^{\infty} \frac{-F(p)}{F'(p)} dx + \lambda_0 \int_0^{\infty} p dx \\ &+ \sum_{i=1}^{m} \lambda_i \frac{1}{h_F(p)} \int_0^{\infty} \frac{c_k(x)}{F'(p)} dx - \lambda_0 - \sum_{i=1}^{m} \lambda_i a_i \qquad (6.54) \end{aligned}$$

Then

$$\begin{aligned} \frac{d\mathcal{L}}{dp} &= \frac{1}{h_F(p)} \left[\frac{F(p)F''(p)}{(F'(p))^2} - 1 \right] + \frac{1}{(h_F(p))^2} \frac{F''(p)}{(F'(p))^2} \int_0^{\infty} \frac{-F(p)}{F'(p)} dx \\ &+ \lambda_0 + \sum_{i=1}^{m} \lambda_i \frac{1}{h_F(p)} \frac{F''(p)}{(F'(p))^2} (a_i - c_i(x)) \qquad (6.55) \end{aligned}$$

So at maximum F-entropy distribution

$$\frac{d\mathcal{L}}{dp} = 0. \tag{6.56}$$

That is,

$$\begin{aligned} &\frac{1}{h_F(p)} \left[\frac{F(p)F''(p)}{(F'(p))^2} - 1 \right] + \frac{1}{(h_F(p))^2} \frac{F''(p)}{(F'(p))^2} \int_0^{\infty} \frac{-F(p)}{F'(p)} dx \\ &\qquad + \lambda_0 + \sum_{i=1}^{m} \lambda_i \frac{1}{h_F(p)} \frac{F''(p)}{(F'(p))^2} (a_i - c_i(x)) = 0 \qquad (6.57) \end{aligned}$$

Now dividing the Equation (6.57) by $\frac{F''(p)}{F'(p)}$ and integrating,

$$\lambda_0 = \frac{1}{h_F(p)} \tag{6.58}$$

Thus Equation (6.57) can be written as

$$\begin{aligned} F(p) + \frac{1}{h_F(p)} \int_0^{\infty} \frac{-F(p)}{F'(p)} dx \\ + \sum_{i=1}^{m} \lambda_i (a_i - c_i(x)) = 0 \end{aligned} \tag{6.59}$$

Then

$$\begin{aligned} F(p) &= \sum_{i=1}^{m} \lambda_i (c_i(x) - a_i) + \frac{1}{h_F(p)} \int_0^{\infty} \frac{F(p)}{F'(p)} dx & (6.60) \\ &= \sum_{i=1}^{m} \lambda_i (c_i(x) - a_i) - H_F(p) & (6.61) \end{aligned}$$

Now using the m constraints we can solve for λ_i. Note that from the m constraints, the probability distribution p is parametrized by a vector $(a_1, \cdots, a_m)$.
By differentiating Equation (6.61) with respect to a_i,

$$\frac{dF}{da_i} = -\lambda_i - \frac{dH_F(p)}{da_i} \tag{6.62}$$

By multiplying with $\frac{1}{F'(p)}$ and integrating,

$$\int \frac{1}{F'(p)} \frac{dF}{da_i} dx = -\lambda_i \int \frac{1}{F'(p)} dx - \int \frac{dH_F(p)}{da_i} \frac{1}{F'(p)} dx \tag{6.63}$$

Since

$$\int \frac{1}{F'(p)} \frac{dF}{da_i} dx = \int \frac{dp}{da_i} dx = 0 \tag{6.64}$$

from Equation (6.63),

$$\lambda_i = -\frac{dH_F(p)}{da_i} \tag{6.65}$$

Note that a_i's are in one to one correspondence with the coordinates λ_i's and λ_i's are the canonical coordinates. Thus $F(p)$ takes the form of a F-exponential family

$$F(p(x;\theta)) = \sum_{i=1}^{m} \theta^i c_i(x) - \psi(\theta) \tag{6.66}$$

with $\theta^i = \lambda_i, \ i = 1, \cdots, m$. □

6.3 Estimation in F-Exponential Family

Here first we describe the U-estimator in a F-exponential family [36]. Then we give a proof of the generalized Cramer-Rao bound defined by Naudts [28] and show that the U-estimator in a deformed exponential family is optimal with respect to this bound.

6.3.1 U-estimator in F-exponential family

Eguchi et al. [36] defined a generalized estimator called the U-estimator and discussed its properties in a deformed exponential family (named as the U-model). Let us briefly describe their work here.

Consider a statistical model $\mathcal{S} = \{p(x;\theta)\}$ and N independent observations $\mathbf{x}_N = (x^1, \cdots, x^N)$ from $p(x;\theta) \in \mathcal{S}$. Let $U : \mathbb{R} \to \mathbb{R}_+$ be an increasing convex function and let U^* be the convex conjugate of U given by $U^*(t) = t\xi(t) - U(\xi(t))$, where $\xi(t)$ is the inverse function of the derivative of $U(t)$. So $\frac{d}{dt}U^*(t) = \xi(t)$.
Eguchi et al. [36] defined a U-loss function $L_U(\theta)$ given by

$$L_U(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \xi(p(x^i;\theta)) + b_U(\theta) \tag{6.67}$$

where

$$b_U(\theta) = \int U(\xi(p(x;\theta)))dx \tag{6.68}$$

Then the U**-estimator** $\hat{\theta}_U$ is defined as

$$\hat{\theta}_U = \arg\min_{\theta\in\mathbb{E}} L_U(\theta) \tag{6.69}$$

They showed that the U-estimator is asymptotically consistent and also investigated the asymptotic normality for the U-estimator. The estimating function is given by

$$s_U(x;\theta) = \frac{\partial}{\partial\theta}\xi(p(x;\theta)) - E_p[\frac{\partial}{\partial\theta}\xi(p(x;\theta))] \tag{6.70}$$

They showed that $\sqrt{N}(\hat{\theta}_U - \theta)$ is asymptotically normal with zero mean and variance $J(\theta)^{-1}V(\theta)J(\theta)$ where

$$V(\theta) = \mathrm{Var}(s_U(X;\theta)), \quad J(\theta) = E_p[\frac{\partial}{\partial\theta}s_U(X;\theta)] \tag{6.71}$$

In general the U-estimator is not asymptotically efficient. When $U = \exp$, the U-estimator is the MLE and we get that the MLE is asymptotically efficient.
Further they considered a U-model $\mathcal{S} = \{p(x;\theta) = u(\sum_{i=1}^n \theta^i x_i - \kappa_U(\theta))\}$, where $u = U'$. Then the U-loss function on $\mathcal{S}$ is given by

$$L_U(\theta) \quad = \quad -\sum_{i=1}^{n}\theta^i \bar{x}_i + \kappa_U(\theta) + b_U(\theta) \tag{6.72}$$

Note that the U-model is a F-exponential family $\mathcal{S} = \{p(x;\theta) = Z(\sum_{i=1}^n \theta^i x_i - \psi_F(\theta)) \ / \ \theta \in \mathbb{E} \subseteq \mathbb{R}^n\}$ with $U' = u = Z$ and $\kappa_U(\theta) = \psi_F(\theta)$. Then $U^*(t) = \int_1^t F(u)\, du$ and $(U^*)'(p) = F(p) = \xi(p)$. On the F-exponential family $\mathcal{S}$ consider the U-estimator $\hat{\theta}_U$ determined from

$$\partial_i L_U(\theta) = -\bar{x}_i + \partial_i \psi_F(\theta) + \partial_i b_U(\theta). \tag{6.73}$$

Since $\partial_i L_U(\theta) \mid_{\hat{\theta}_U} = 0$

$$\bar{x}_i \quad = \quad \partial_i \psi_F(\hat{\theta}_U) + \partial_i b_U(\hat{\theta}_U) \tag{6.74}$$

In Chapter 4 we described the U-geometry of the F-exponential family. In U-geometry, the dual coordinate η is given by

$$\eta_i = E_p[x_i] = \partial_i \Psi(\theta) = \partial_i \psi_F(\theta) + \partial_i b_U(\theta) \tag{6.75}$$

Thus U-estimator is directly written in terms of the dual coordinate η. Hence for the F-exponential family consider the U-estimator $\hat{\eta}_U$ for the the coordinate $\eta = E_p[x]$. Then

$$\hat{\eta}_U = \bar{x}. \tag{6.76}$$

The estimator $\hat{\eta}_U$ is unbiased since $E_p[\hat{\eta}_U] = \eta$.

Remark 6.3.1. *It is easy to see that in general the estimator $\hat{\eta}_U$ is not efficient. That* *is, $\hat{\eta}_U$ is not optimal with respect to the Cramer-Rao bound. More precisely, variance of $\hat{\eta}_U$ does not attain Cramer-Rao equality for a general U-estimator $\hat{\eta}_U$. But when $U(s) = \exp(s)$ the U-estimator is the MLE and it is efficient.*

In the context of the statistical mechanics Naudts [28] defined a generalized Cramer-Rao bound using an escort probability distribution and gave sufficient conditions for an estimator in a statistical model to be optimal with respect to this bound. He showed that a deformed exponential family naturally has an estimator which satisfies the sufficient conditions for optimality.

Now we give a proof of the generalized Cramer-Rao bound defined by Naudts using a generalized score vector and an F-escort probability density function.

For the sake of computational convenience we consider a one dimensional statistical manifold parametrized by a real parameter θ.

Let $\mathcal{S} = \{p(x;\theta) \ / \ \theta \in \mathbb{E} \subseteq \mathbb{R}\}$ be a statistical manifold. We have the score function

$$\partial_\theta \ln p(x;\theta), \quad \text{where} \quad \partial_\theta = \frac{\partial}{\partial\theta}. \tag{6.77}$$

Note that the expectation of the score is zero.

$$E_p[\partial_\theta \ln p(x;\theta)] = \int \partial_\theta \ln p(x;\theta) p(x;\theta) dx = 0 \tag{6.78}$$

Let F be a smooth real valued function on $(0,\infty)$ satisfying $F'(x) > 0$ and $F''(x) < 0$. Define a generalized score function called F-**score** as

$$\partial_\theta F(p(x;\theta)) = F'(p)\partial_\theta p(x;\theta) \tag{6.79}$$

For $p \in \mathcal{S}$, the F-escort probability $\hat{p}_F$ of p

$$\hat{p}_F = \frac{1}{h_F(\theta)F'(p)}, \quad \text{where} \quad h_F(\theta) = \int \frac{1}{F'(p)} dx \tag{6.80}$$

Now we show that the expectation of the F-score function with respect to the F-escort distribution $\hat{p}_F$ is zero.

$$\begin{aligned} E_{\hat{p}_F}[\partial_\theta F(p(x;\theta))] &= \int \partial_\theta F(p(x;\theta))\hat{p}_F(x;\theta)dx && (6.81) \\ &= \frac{1}{h_F(\theta)} \int \partial_\theta F(p(x;\theta)) \frac{1}{F'(p)} dx && (6.82) \\ &= \frac{1}{h_F(\theta)} \int \partial_\theta p(x;\theta)dx = 0 && (6.83) \end{aligned}$$

The Fisher information metric $I(\theta)$ is given by

$$I(\theta) = \int \partial_\theta \ln p(x;\theta) \partial_\theta \ln p(x;\theta) p(x;\theta) dx \tag{6.84}$$

Using the F-score a generalized Fisher metric $I_F(\theta)$ can be defined as

$$I_F(\theta) = \int \partial_\theta F(p(x;\theta)) \partial_\theta F(p(x;\theta)) \hat{p}_F(x;\theta) dx \tag{6.85}$$

Naudts [28] defined a generalized metric using an escort distribution P_θ of the original distribution p_θ as

$$g^N(\theta) = \int \frac{1}{P_\theta} (\partial_\theta p)^2 dx \tag{6.86}$$

Rewriting this metric using F-escort distribution

$$\begin{aligned} g^N(\theta) &= \int \frac{1}{\hat{p}_F(\theta)} (\partial_\theta p)^2 dx && (6.87) \\ &= h_F(\theta) \int F'(p)(\partial_\theta p)^2 dx && (6.88) \\ &= h_F(\theta) g^G(\theta) && (6.89) \end{aligned}$$

where g^G is the G-metric with $G(p) = pF'(p)$. Also

$$I_F(\theta) = \frac{1}{h_F(\theta)} g^G = \frac{1}{(h_F(\theta))^2} g^N(\theta) \tag{6.90}$$

Using the F-escort probability distribution, a generalized variance Var_F called **F-variance** of a random variable X is defined as

$$\mathrm{Var}_F(X) = E_{\hat{p}_F}[(X - E_{\hat{p}_F}(X))^2] \tag{6.91}$$

Theorem 6.3.2. *Let X be a random variable with density $p(x;\theta) \in \mathcal{S}$. Let $T = t(X)$ be an unbiased estimator for $\psi(\theta)$ so that $E_p[t(X)] = \psi(\theta)$. Also let the F-expectation of $t(X)$ is $E_{\hat{p}_F}[t(X)] = \phi(\theta)$. Then the F-variance satisfies the lower bound*

$$\mathrm{Var}_F(T) \geq \frac{|\ \psi'(\theta)\ |^2}{g^N(\theta)} \tag{6.92}$$

where $g^N(\theta) = h_F(\theta) g^G(\theta)$ with $G(p) = pF'(p)$.

Proof. Since $E_{\hat{p}_F}[\partial_\theta F(p(x;\theta))] = 0$, the F-variance of the F-score is

$$\begin{aligned} \mathrm{Var}_F(\partial_\theta F(p(x;\theta))) &= E_{\hat{p}_F}[(\partial_\theta F(p(x;\theta)))^2] && (6.93) \\ &= I_F(\theta) && (6.94) \end{aligned}$$

Let $V = \partial_\theta F(p(x;\theta))$. Then the F-covariance $\mathrm{Cov}_F(V,T)$ is

$$\begin{aligned} \mathrm{Cov}_F(V,T) &= E_{\hat{p}_F}\left[\partial_\theta F(p(x;\theta))\,(t(X) - E_{\hat{p}_F}(t(X)))\right] && (6.95) \\ &= \frac{1}{h_F(\theta)} \int \partial_\theta F(p(x;\theta))\ (t(x) - \phi(\theta)) \frac{1}{F'(p)}\ dx && (6.96) \\ &= \frac{1}{h_F(\theta)} \int \partial_\theta p(x;\theta)\ (t(x) - \phi(\theta))\ dx && (6.97) \\ &= \frac{1}{h_F(\theta)}\ \partial_\theta \int t(x)\ p(x;\theta)\ dx && (6.98) \\ &= \frac{1}{h_F(\theta)} \psi'(\theta) && (6.99) \end{aligned}$$

By the Cauchy-Schwarz inequality,

$$\mathrm{Var}_F(V)\mathrm{Var}_F(T) \geq |\ \mathrm{Cov}_F(V,T)\ |^2 = \frac{|\ \psi'(\theta)\ |^2}{h_F(\theta)^2} \tag{6.100}$$

Thus

$$\mathrm{Var}_F(T) \geq \frac{|\ \psi'(\theta)\ |^2}{h_F(\theta)^2 Var(V)} = \frac{|\ \psi'(\theta)\ |^2}{h_F(\theta)^2 I_F(\theta)} \tag{6.101}$$

Substituting Equation (6.90) into Equation (6.101)

$$\mathrm{Var}_F(T) \geq \frac{|\ \psi'(\theta)\ |^2}{h_F(\theta)^2 I_F(\theta)} = \frac{|\ \psi'(\theta)\ |^2}{g^N(\theta)} \tag{6.102}$$

□

Eguchi et al. [36] showed the asymptotic normality of the U-estimator in a deformed exponential family. But the U-estimator in a deformed exponential family is not an efficient estimator in general. That is the U-estimator is not optimal with respect to the usual Cramer-Rao lower bound. Now we prove that in a F-exponential family the U-estimator for the dual coordinate η in the U-geometry is optimal with respect to the generalized Cramer-Rao bound defined by Naudts.

Theorem 6.3.3. *Let $\mathcal{S} = \{p(x;\theta) = Z(\theta x - \psi_F(\theta))\}$ be a F-exponential family and let $\eta = E_p[x]$ be the dual coordinate in the U-geometry. Then U-estimator $\hat{\eta}_U = \bar{x}$ for η is optimal with respect to the generalized Cramer-Rao bound defined by Naudts. That is,*

$$\mathrm{Var}_F(\hat{\eta}_U) = \frac{1}{g^N(\eta)}. \tag{6.103}$$

Proof. The U-estimator $\hat{\eta}_U = \bar{x}$ is unbiased so that $E_p[\hat{\eta}_U] = \eta$. Also from the definition of the F-exponential family

$$\partial_\theta F = \bar{x} - \partial_\theta \psi_F(\theta), \quad E_{\hat{p}_F}[x] = \partial_\theta \psi_F(\theta) \tag{6.104}$$

The F-variance of $\hat{\eta}_U$ is

$$\begin{aligned} \mathrm{Var}_F(\hat{\eta}_U) &= E_{\hat{p}_F}[(\bar{x} - E_{\hat{p}_F}[\bar{x}])^2] & (6.105)\\ &= E_{\hat{p}_F}[(\bar{x} - \partial_\theta \psi_F(\theta))^2] & (6.106)\\ &= E_{\hat{p}_F}[(\partial_\theta F)^2] & (6.107) \end{aligned}$$

From Equation (6.85), it follows that

$$\mathrm{Var}_F(\hat{\eta}_U) = I_F(\theta) = \frac{g^G(\theta)}{h_F(\theta)} \tag{6.108}$$

In U-geometry θ and η are dual coordinates. Also the metric $g^G(\eta) = \frac{\partial \theta}{\partial \eta}$ and $g^G(\eta) = (g^G(\theta))^{-1}$. Thus

$$\partial_\eta p = \frac{\partial \theta}{\partial \eta}\, \partial_\theta p = g^G(\eta)\, \partial_\theta p \tag{6.109}$$

Then

$$\begin{aligned} g^N(\eta) &= h_F(\eta) \int F'(p)(\partial_\eta p)^2 dx & (6.110) \\ &= h_F(\theta) g^G(\eta) = \frac{h_F(\theta)}{g^G(\theta)} & (6.111) \end{aligned}$$

Since $\hat{\eta}_U$ is unbiased, $\psi'(\eta) = 1$ and then

$$\frac{\mid \psi'(\eta) \mid^2}{g^N(\eta)} = \frac{1}{g^N(\eta)} = \frac{g^G(\theta)}{h_F(\theta)} \tag{6.112}$$

Thus from Equations (6.108) and (6.112),

$$\mathrm{Var}_F(\hat{\eta}_U) = \frac{1}{g^N(\eta)} \tag{6.113}$$

Hence $\hat{\eta}_U$ is optimal with respect to the generalized Cramer-Rao bound. □

6.3.2 F-MLE in a F-exponential family

In the previous chapters we discussed the geometrical and statistical properties of an exponential family. The standard exponential family is dually flat with respect to the (± 1)-connections. The exponential family naturally has a sufficient statistics which is also the MLE for the dual coordinate. Also the MLE is a finite sample efficient estimator. Thus, an exponential family has an estimator for the dual coordinate which attains equality in the Cramer-Rao lower bound. The lower bound is given by the Fisher information metric which is the Riemannian metric associated to the dually flat structure of the exponential family. Hence the maximum likelihood estimation in an exponential family is closely related to the dually flat structure of the exponential family.

In this context one may think of the estimation problem in a deformed exponential family. As in the case of the exponential family, does a deformed exponential family has an estimator which is closely related to its dually flat geometry? There is a theorem by Amari and Nagaoka [14] which states that an estimator for a statistical model $\{p(x;\theta)\}$ is finite sample efficient iff $\mathcal{S}$ is an exponential family and θ is a m-affine (flat with respect to (-1)-connection) coordinate system. So there does not exist a finite sample efficient estimator for a deformed exponential family except for the exponential family. Thus one may have to define some generalized notion of efficiency which in turn requires a generalized Cramer-Rao lower bound.

In the context of nonextensive thermostatistics Naudts [28] defined a generalized Cramer-Rao bound using an escort probability density function. Then in the deformed experiential family he defined a dually flat structure, the U-geometry, using a Bregman type divergence and showed that this bound is optimal. The divergence that Naudts considered is a U-divergence defined by Murata et al. [22]. Eguchi et al. [36] studied the geometry associated with the U-divergence and defined an estimator called the U-estimator. In the previous section we proved that in a deformed exponential family the U-estimator for the dual coordinate in the U-geometry is optimal with respect to the generalized Cramer-Rao bound by Naudts.

Deformed exponential family has two dually flat structures, the U-geometry and the χ-geometry. As the MLE in an exponential family is related to the dually flat structure of the exponential family, the U-estimator is related to the dually flat U-geometry of the deformed exponential family. Now is it possible to find an estimator which is closely related to the dually flat χ-geometry of the deformed exponential family? In Section 6.1 we defined the F-MLE which is a generalized notion of MLE. Let us consider the F-MLE for a deformed exponential family.

Let $\mathcal{S} = \{p(x;\theta)\}$ be a F-exponential family and $\mathbf{x}_N = (x^1, \cdots, x^N)$ be N independent observations from $p(x;\theta) \in \mathcal{S}$. The F-likelihood function is given by

$$F(L_F(\theta)) = \sum_{j=1}^{N} F(p(x^j;\theta)) = \sum_{j=1}^{N} \left[\sum_{i=1}^{n} \theta^i x_i^j - \psi_F(\theta) \right] \tag{6.114}$$

$$= \sum_{i=1}^{n} \theta^i \sum_{j=1}^{N} x_i^j - N\psi_F(\theta) \tag{6.115}$$

The F-MLE is

$$\bar{x}_i = \frac{x_i^1 + \cdots + x_i^N}{N} = \partial_i \psi_F(\hat{\theta}_F) \tag{6.116}$$

Since the dual coordinates η_i in χ-geometry is $\eta_i = \partial_i \psi_F(\theta) = E_{\hat{p}_F}[x_i]$, the F-MLE can be directly written in terms of the dual coordinate. Hence

$$\bar{x}_i = \partial_i \psi_F(\hat{\theta_F}) = \hat{\eta}_i \tag{6.117}$$

Thus the F-MLE is given in terms of the dual coordinate in the χ-geometry. Also note that the dual coordinate is defined in terms of the F-escort probability distribution. In Chapter 4 we showed that the χ-geometry is the (± 1)-conformal flattening of the (F, G)-geometry. That is, the manifold $\mathcal{S}'$ of the F-escort probability distributions is dually flat by conformally flattening the (F, G)-geometry on the original manifold $\mathcal{S}$. Thus to study the F-MLE in a deformed exponential family one has to consider both $\mathcal{S}$ and $\mathcal{S}'$. Does the generalized Cramer-Rao bound by Naudts work in the case of F-MLE or do we need to define the notions of consistency and efficiency suitably to analyze the properties of the F-MLE? This would be an interesting problem for further study.

6.4 Summary

In this chapter the F-product of two real numbers, the F-independence of two random variables and the F-MLE are defined. We showed that the F-MLE is a MAP estimator with a suitable prior. Further we defined the F-escort MLE which is also a generalized notion of MLE. Then a characterization of the q-escort MLE among the F-escort MLE is given. Also an analytic proof of the F-version of Maximum entropy theorem is given. Further we discussed the estimation problem in a deformed exponential family. We gave a proof of the generalized Cramer-Rao bound defined by Naudts and showed that in a deformed exponential family the U-estimator for the dual coordinate in the U-geometry is optimal with respect to this bound. Finally we posed an open problem regarding the consistency and efficiency of the F-MLE in a deformed exponential family.

Concluding Remarks

On a statistical manifold a generalized class of geometric structures called the (F, G)-geometry is defined in which the α-geometry is a special case. Invariance properties of various geometric structures are studied and classified them into invariant and non-invariant. The α-geometry is the only invariant geometry among the (F, G)-geometry. The role of the non-invariant (F, G)-geometry in the study of the dually flat structures of a deformed exponential family gives a clarifying picture of the state of the art. The geometric interpretation of the estimation problem based on a mismatched model in an exponential family is given in terms of the ancillary manifold. In an exponential family the maximum likelihood estimation is closely related to its dually flat structure. Certain attempts have been made in the case of estimation in a deformed exponential family with the dually flat U-geometry (Naudts, Eguchi, Komori, Ohara), see Section 6.3 in Chapter 6. We pose an open problem regarding the estimation in a deformed exponential family with the dually flat χ-geometry. In a deformed exponential family the generalized MLE, F-MLE, is given in terms of the dual coordinate in the χ-geometry. To analyze the properties of the F-MLE one has to look at some generalized notions of consistency and efficiency. Also the applications of the non-invariant (F, G)-geometry in various fields are to be investigated in detail.

REFERENCES

[1] Rao, C.R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37: 81-91.

[2] Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. In *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 1946, volume 186(1007), pp. 453-461.

[3] Jeffreys, H. (1948). *Theory of Probability*(Second ed.). Clarendon Press: Oxford.

[4] Chentsov, N. N. (1982). *Statistical Decision Rules and Optimal Inference*, Translations of the Mathematical Monographs, volume 53, American Mathematical Society: Providence, Rhode Island.

[5] Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Annals of Statistics*, 3: 1189-1242.

[6] Efron, B. (1978). The Geometry of Exponential Families. *Annals of Statistics*, 6: 362-376.

[7] Dawid, A.P. (1975). A Discussion to Efron's paper. *Annals of Statistics*, 3: 1231-1234.

[8] Reeds, J. (1975). Discussion to Efron's Paper. *Annals of Statistics*, 3: 1234-1238.

[9] Madsen, L.T. (1979). The geometry of statistical model-a generalization of curvature. *Research Report 79-1*, Statistical Research Unit., Danish Medical and Social Science Research Council.

[10] Amari, S. (1980). Theory of Information Space: A Differential-Geometrical Foundation of Statistics. *Post RAAG Reports No.106.*, pp. 53.

[11] Amari, S. (1982) Differential geometry of curved exponential families-curvature and information loss. *Annals of Statistics*, 10: 357-385.

[12] Amari, S. (1985). *Differential-Geometrical methods in Statistics*, Lecture Notes in Statistics, volume 28. Springer: New York.

[13] Murray, M.K. and Rice, R.W. (1995). *Differential Geometry and Statistics*. Chapman and Hall: London.

[14] Amari, S. and Nagaoka, H. (2000). *Methods of Information Geometry, Translations of Mathematical Monographs*. Oxford University Press: Oxford.

[15] Burbea, J. (1986). Informative geometry of probability spaces. *Expositiones Mathematicae*, 4: 347-378.

[16] Harsha, K.V. and Subrahamanian Moosath, K.S. (2014). F-geometry and Amari's α-geometry on a statistical manifold. *Entropy*, 16(5): 2472-2487.

[17] Eguchi, S. (1983). Second order efficiency of minimum contrast estimators in a curved exponential family. *Annals of Statistics*, 11: 793-803.

[18] Csiszar, I. (1967). Information-type Measures of Difference of Probability Distributions and Indirect Observation. *Studia Scientiarum Mathematicarum Hungarica*, 2: 229-318.

[19] Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B (Methodological)*, 28(1): 131-142.

[20] Bregman, L. M. (1967). The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3): 200-217.

[21] Zhang, J. (2004). Divergence function, duality and convex analysis. *Neural Computation*, 16: 159-195.

[22] Murata, N., Takenouchi, T., Kanamori, T. and Eguchi, S. (2004). Information geometry of U-Boost and Bregman Divergence. *Neural Computation*, 16: 1437-1481.

[23] Corcuera, J.M. and Giummole, F. (1998). A characterization of monotone and regular divergences. *Annals of the Institute of Statistical Mathematics*, 50(3): 433-450.

[24] Wagenaar, D.A. (1998). “Information Geometry for Neural Networks”, Centre for Neural Networks, King’s College London. Retrieved from http://www.danielwagenaar.net/res/papers/98-Wage2.pdf.

[25] Ay, N., Jost, J., Le H. V., and Schwachhofer, L. (2015). Information geometry and sufficient statistics. *Probability Theory and Related Fields*, 162(1-2): 327-364.

[26] Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52: 479-487.

[27] Amari, S. and Ohara, A. (2011). Geometry of q-Exponential Family of Probability Distributions. *Entropy*, 13: 1170-1185.

[28] Naudts, J. (2004). Estimators, escort probabilities, and ϕ-exponential families in statistical physics. *Journal of Inequalities in Pure and Applied Mathematics*, 5(4), Article 102.

[29] Naudts, J. (2008). Generalized exponential families and associated entropy functions. *Entropy*, 10: 131-149.

[30] Naudts, J. (2011). *Generalised Thermostatistics*. Springer: London.

[31] Kaniadakis, G., Lissa, M. and Scarfone, A. M. (2004). Deformed Logarithms and Entropies. *Physica A*, 340: 41-49.

[32] Pistone, G. (2009). κ-exponential models from the geometric viewpoint. *The European Physical Journal B*, 70: 29-37.

[33] Vigelis, R. and Cavalcante, C. (2011). On the φ-exponential family of probability distributions. *Journal of Theoretical Probability*, 21: 1-15.

[34] Matsuzoe, H. and Henmi, M. (2013). Hessian structures on deformed exponential families. In *Geometric Science of Information, Lecture Notes in Computer Science*, 2013, volume 8085, pp. 275-282.

[35] Matsuzoe, H. (2014). Hessian structures on deformed exponential families and their conformal structures. *Differential Geometry and its Applications*, 35: 323-333.

[36] Eguchi, S., Komori, o. and Ohara, A. (2014). Duality of Maximum Entropy and Minimum Divergence. *Entropy*, 16: 3552-3572.

[37] Amari, S., Ohara, A. and Matsuzoe, H. (2012). Geometry of deformed exponential families: Invariant, dually flat and conformal geometries. *Physica A: Statistical Mechanics and its Applications*, 391: 4308-4319.

[38] Harsha, K.V. and Subrahamanian Moosath, K.S. (2015). Dually Flat Geometries of the Deformed Exponential Family. Physica A: Statistical Mechanics and its Applications, 433: 136-147.

[39] Kass, A. M. (1980). "The Riemannian structure of model spaces: A geometrical approach to the inference", Ph.D. Thesis, University of Chicago.

[40] Atkinson, C. and Mitchell, A. F. (1981). Rao's distance measure. *Sankhya: The Indian Journal of Statistics, Series A*, 43(3): 345-365.

[41] Nagaoka, H. and Amari, S. (1982). Differential geometry of smooth families of probability distributions, *METR 82-7*, University of Tokyo.

[42] Amari, S. and Kumon, M. (1983). Differential geometry of Edgeworth expansions in curved exponential family. *Annals of the Institute of Statistical Mathematics*, 35(1): 1-24.

[43] Kumon, M. and Amari, S. (1983). Geometrical Theory of Higher-Order Asymptotics of Test, Interval Estimator and Conditional Inference. In *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 1983, volume 387(1793), pp. 429-458.

[44] Kass, A. M. (1984). Canonical parametrization and zero parameter effects curvature. *Journal of the Royal Statistical Society, Series B*, 46: 86-92.

[45] Skovgaard, L. T. (1984). A Riemannian geometry of the multivariate normal model. *Scandinavian Journal of Statistics*, 11: 211-223.

[46] Barndorff-Nielsen, O.E., Cox,D.R. and Reid, N. (1986). The role of differential geometry in statistical theory. *International Statistical Review*, 54: 83-96.

[47] Wu, S., Nakahara, H. and Amari, S. (2001). Population Coding with Correlation and an Unfaithful Model. *Neural Computation*, 13(4): 775-797.

[48] Oizumi, M. Okada, M. and Amari, S. (2011). Information loss associated with imperfect observation and mismatched decoding. *Frontiers in Computational Neuroscience*, 5(9): 2011.

[49] Umarov, S., Tsallis, C. and Steinberg, S. (2008). On a q-Central Limit Theorem Consistent with Nonextensive Statistical Mechanics. *Milan Journal of Mathematics*, 76(1): 307-328.

[50] Borges, E. P. (2004). A possible deformed algebra and calculus inspired in nonextensive thermostatistics. *Physica A: Statistical Mechanics and its Applications*, 340(1-3): 95-101.

[51] Umarov, S. and Tsallis, C. (2007). On multivariate generalizations of the q-central limit theorem consistent with nonextensive statistical mechanics, In *AIP Conference Proceedings*, Catania, Italy, July 1-5, 2007, 965(34).

[52] Ferrari. D and Yang, Y. (2010). Maximum L$_q$-likelihood estimation. *Annals of Statistics*, 38: 753-783.

[53] Matsuzoe, H. and Ohara, A. (2010). Geometry for q-exponential families, In *Proceedings of the* 2*nd International Colloquium on Differential Geometry and its Related Fields*, Veliko Tarnovo, September 6-10, 2010.

[54] Fujimoto, Y. and Murata, N. (2010). A generalization of Independence in Naive Bayes Model., In *Intelligent Data Engineering and Automated Learning-IDEAL 2010, Lecture Notes in Computer Science*, Paisley, UK, September 1-3, 2010, 6283, pp. 153-161.

[55] Amari, S. (2009). α-divergence is unique, belonging to both f-divergence and Bregman divergence classes. *IEEE Transactions on Information Theory*, 55: 4925-4931.

[56] Picard, D. B. (1992). Statistical Morphisms and Related Invariance Properties. *Annals of the Institute of Statistical Mathematics*, 44(1): 45-61.

[57] Zhang, J. (2015). On Monotone Embedding in Information Geometry. *Entropy*, 17: 4485-4499.

[58] Matumoto, T. (1993). Any statistical manifold has a contrast function-On the C^3-functions taking the minimum at the diagonal of the product manifold. *Hiroshima Mathematical Journal*, 23: 327-332.

[59] Kurose, T. (1994). On the Divergence of 1-conformally Flat Statistical Manifolds. *Tohoku Mathematical Journal*, 46: 427-433.

[60] Kurose, T. (2002). Conformal-projective geometry of statistical manifolds. *Interdisciplinary Information Sciences*, 8: 89-100.

[61] Harsha, K. V. and Subrahamanian Moosath, K.S. (2015). Geometry of F-likelihood Estimators and F-Max-Ent Theorem, In *AIP Conference Proceedings on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, Amboise, France, September 21-26, 2014, 1641, pp. 263-270.

[62] Ohara, A., Matsuzoe, H. and Amari, S. (2010). A dually flat structure on the space of escort distributions. *Journal of Physics: Conference series*. 201(1): 2010.

[63] Harsha, K.V. and Subrahamanian Moosath, K.S. A dually Flat Geometry of the Manifold of F-Escort Probability Distributions, Accepted in *Journal of Combinatorics, Information and System Sciences*.

LIST OF PAPERS BASED ON THESIS

Papers in Refereed International Journals

1. Harsha, K.V. and Subrahamanian Moosath, K.S. (2014). F-geometry and Amari's α-geometry on a Statistical Manifold. *Entropy*, 16(5): 2472-2487.

2. Harsha, K.V. and Subrahamanian Moosath, K.S. (2015). Dually Flat Geometries of the Deformed Exponential Family. *Physica A: Statistical Mechanics and its Applications*, 433: 136-147.

3. Harsha, K.V. and Subrahamanian Moosath, K.S. A dually Flat Geometry of the Manifold of F-Escort Probability Distributions. Accepted in the *Journal of Combinatorics, Information and System Sciences*.

Presentations in Conferences

1. Harsha, K.V. and Subrahamanian Moosath, K.S. (2014). Geometry of F-likelihood Estimators and F-Max-Ent Theorem, In *AIP Conference Proceedings on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, Amboise, France, September 21-26, 2014, 1641, pp. 263-270.

2. Harsha, K.V. and Subrahamanian Moosath, K.S. (2014). A dually Flat geometry of the manifold of F-escort probability distributions, In *23rd International Conference of Forum for Interdisciplinary Mathematics*, NIT-Karnataka, Karnataka, India, December 18-20, 2014 (the paper won second Best Paper Award- Professor R. S. Varma Memorial Award).

3. Harsha, K.V. and Subrahamanian Moosath, K.S. (2014). Hessian Structures and (F, G)-geometry on a Deformed Exponential Family. Lecture notes in computer science, pp 213-221, *Geometric Science of Information, Second International Conference- GSI 2015*, Paris-Saclay, France, 28-30 October 2015.