Geometry of $F$-likelihood Estimators and $F$-Max-Ent Theorem

Harsha K V and Subrahmanian Moosath K S

Department of Mathematics
Indian Institute of Space Science and Technology

26 September 2014
1. Introduction

- On a statistical manifold $\mathcal{S}$, the Fisher information acts as a Riemannian metric called the Fisher information metric $g$.

- Amari defined a one parameter family of connections $\nabla^\alpha$ called $\alpha$–connections using $\alpha$–embeddings.

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

- The connections $\nabla^\alpha$ and $\nabla^{-\alpha}$ are dual with respect to the Fisher information metric $g$. 
1. Introduction

- On a statistical manifold $\mathcal{S}$, the Fisher information acts as a Riemannian metric called the Fisher information metric $g$.

- Amari defined a one parameter family of connections $\nabla^\alpha$ called $\alpha-$connections using $\alpha-$embeddings.

\[
L_\alpha(p) = \begin{cases} 
\frac{2}{1-\alpha} p^{\frac{1-\alpha}{2}} & \alpha \neq 1 \\
\log p & \alpha = 1 
\end{cases}
\]

- The connections $\nabla^\alpha$ and $\nabla^{-\alpha}$ are dual with respect to the Fisher information metric $g$.

- An exponential family is flat with respect to $\nabla^1$-connection (also known as exponential connection) and by duality it is also $(-1)$-flat. Hence $(g, \nabla^1, \nabla^{-1})$ is a dually flat structure on exponential family.

- Amari has defined a $\alpha-$family for any $\alpha \in \mathbb{R}$. But for $\alpha \neq 1$, $\alpha-$family is not flat with respect to the $\alpha-$connection.
A $q$-exponential family extends the notion of an exponential family.

A $q$–exponential family, which is an $\alpha$–family with $\alpha = 1 - 2q$, has a dually flat structure called $q$-structure.

This $q$-geometry is obtained by the conformal flattening of $\alpha$–geometry.
A $q$-exponential family extends the notion of an exponential family.

A $q$-exponential family, which is an $\alpha$-family with $\alpha = 1 - 2q$, has a dually flat structure called $q$-structure.

This $q$-geometry is obtained by the conformal flattening of $\alpha$-geometry.

Naudts generalized the notion of exponential family to a large class of families of probability distributions called $\phi$-exponential family and studied the dually flat structure.

An information geometric foundation for the deformed exponential family ($\chi$-family) is given by Amari et al.
A $q$-exponential family extends the notion of an exponential family.

A $q$–exponential family, which is an $\alpha$–family with $\alpha = 1 - 2q$, has a dually flat structure called $q$-structure.

This $q$-geometry is obtained by the conformal flattening of $\alpha$–geometry.

Naudts generalized the notion of exponential family to a large class of families of probability distributions called $\phi$-exponential family and studied the dually flat structure.

An information geometric foundation for the deformed exponential family ($\chi$-family) is given by Amari et al.

We extended Amari’s $\alpha$-geometry to a new geometry called $F$–geometry using an embedding function $F$ of $S$ into the space of random variables $\mathbb{R}_X$.

When $F$ is the $\alpha$–embeddings of Amari, the $F$–geometry reduces to $\alpha$-geometry.

In this general embedding case also a dualistic structure $(g, \nabla^F, \nabla^H)$ can be introduced on $S$, where $H$ is the dual embedding of $F$. 
Further we introduced \((F, G)\)-geometry using the embedding \(F\) and a positive smooth function \(G\).

From the idea of \((F, G)\)-geometry, we consider a \(F\)-exponential family which is an extension of \(q\)-exponential family which is not flat with respect to the \(F\)-connection \(\nabla^F\).

A dually flat structure can be defined on \(F\)-exponential family by the conformal flattening of \((F, G)\)-geometry.
Further we introduced \((F, G)\)-geometry using the embedding \(F\) and a positive smooth function \(G\).

From the idea of \((F, G)\)-geometry, we consider a \(F\)-exponential family which is an extension of \(q\)-exponential family which is not flat with respect to the \(F\)-connection \(\nabla^F\).

A dually flat structure can be defined on \(F\)-exponential family by the conformal flattening of \((F, G)\)-geometry.

Further using the function \(F\) and its inverse, we generalize the notion of independence called \(F\)-independence.

Then we define \(F\)-likelihood function and \(F\)-likelihood estimators and discusses its geometry.

An analytic proof of the \(F\)-version of the max-ent theorem is outlined.
2. \textit{F-Geometry}

Let $S$ be an $n$-dimensional statistical manifold. Let $F : (0, \infty) \rightarrow \mathbb{R}$ be an injective function which is at least twice differentiable. Then $F$ is an embedding of $S$ into $\mathbb{R}_x$ which takes each $p(x; \xi) \mapsto F(p(x; \xi))$.

The metric induced by the embedding $F$ is the \textbf{Fisher information metric} $g$ defined by

$$g_{ij}(\theta) = \int \partial_i \ell \partial_j \ell p(x; \theta) \, dx \quad \text{where} \quad \ell(x; \theta) = \log p(x; \theta).$$

The affine connection induced by the embedding $F$, the $F-$\textit{connection} is defined by

$$\Gamma^F_{ijk}(\xi) = \int \left[ (\partial_i \partial_j \ell + (1 + \frac{pF''(p)}{F'(p)})\partial_i \ell \partial_j \ell)(\partial_k \ell) \right] p \, dx.$$

\textbf{Remark 2.1}

Amari's $\alpha-$\textit{geometry} is a special case of the $F-$\textit{geometry}.
3. \((F, G)\)-Geometry

**Definition 3.1** Let \(G : (0, \infty) \rightarrow \mathbb{R}\) be a positive smooth function.

- Define the \(G\)-metric as
  \[
g_{ij}^G(\theta) = \int \partial_i \ell \partial_j \ell G(p) \ p \ dx.
  \]

- Using the embedding \(F\) and the function \(G\), Define \((F, G)\)-connection \(\nabla^{F, G}\) as
  \[
  \Gamma_{ijk}^{F, G} = \int \left( (\partial_i \partial_j \ell + \left(1 + \frac{pF''(p)}{F'(p)}\right) \partial_i \ell \partial_j \ell(\partial_k \ell) \right) G(p) \ p \ dx.
  \]

**Theorem 3.2**

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

\[
H'(p) = \frac{G(p)}{pF'(p)}.
\]

We call such an embedding \(H\) as a \(G\)-dual embedding of \(F\).
4. $F$–exponential family

**Definition 4.1**

- Let $F : (0, \infty) \rightarrow \mathbb{R}$ be any smooth increasing concave function. Let $Z$ be the inverse function of $F$.
- The standard form of an $n$-dimensional $F$–exponential family of distributions $S = \{ p(x; \theta) \mid \theta \in E \subseteq \mathbb{R}^n \}$ is written as

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

where $x = (x_1, \ldots, x_n)$ is a set of random variables, $\theta = (\theta^1, \ldots, \theta^n)$ are the canonical parameters.
- $\psi_F(\theta)$ is called the $F$-free energy or the $F$-potential and is determined from the normalization condition.

$$\int Z\left( \sum_{i=1}^{n} \theta^i x_i - \psi_F(\theta) \right) dx = 1$$

- Define a functional $h_F(\theta)$ as

$$h_F(\theta) = \int \frac{1}{F'(p(x; \theta))} dx$$
Theorem 4.2

The $F-$potential function $\psi_F(\theta)$ is a convex function of $\theta$ and

$$\partial_i \partial_j \psi_F(\theta) = \frac{1}{h_F(\theta)} \int \frac{-pF''(p)}{F'(p)} \partial_i p \partial_j p \frac{1}{p} dx.$$ 

Definition 4.3

- Define a Riemannian metric called $F-$metric $g^F$ by

$$g_{ij}^F(\theta) = \partial_i \partial_j \psi_F(\theta).$$

Note that $(g_{ij}^F)$ is positive definite since $\psi_F$ is a convex function of $\theta$.

- Define a divergence of Bregman-type using $\psi_F(\theta)$, called the $F-$divergence as

$$D_F[p(x; \theta_1) : p(x; \theta_2)] = \psi_F(\theta_2) - \psi_F(\theta_1) - \nabla \psi_F(\theta_1).(\theta_2 - \theta_1).$$

The two distributions $p$ and $r$ which are parametrized by $\theta_1$ and $\theta_2$ respectively. Then the $F-$divergence can be written as

$$D_F[p : r] = \frac{1}{h_F(\theta_1)} \int (F(p) - F(r)) \frac{1}{F'(p)} dx.$$
Definition 4.4

- For a density function $p$ parametrized by $\theta$, define the $F$-escort probability distribution of $p$ as
  \[
  \hat{p}_F(x) = \frac{1}{h_F(\theta)F'(p)}. \]

- Using $\hat{p}_F$, define the $\hat{F}$-expectation of a random variable as
  \[
  E_{\hat{p}}(f(x)) = \frac{1}{h_F(\theta)} \int \frac{1}{F'(p)} f(x) dx.
  \]

Then the $F$-divergence can be written as
\[
D_F[p : r] = E_{\hat{p}}(F(p) - F(r)).
\]

Lemma 4.5

The metric $g^{DF}_{ij}$ and the affine connection $\nabla^{DF}$ induced by the $F$-divergence $D_F$ are given by
\[
g^{DF}_{ij}(\theta) = g^{F}_{ij}(\theta) = \partial_i \partial_j \psi_F(\theta); \quad \Gamma^{DF}_{ijk} = \partial_i \partial_j \partial_k \psi_F(\theta).
\]

The dual $D_F^*$ of $D_F$ induces an affine connection $\nabla^{D_F^*}$ defined by $\Gamma^{D_F^*}_{ijk} = 0$. 
The Legendre transformation of the convex function $\psi_F(\theta)$ is given by

$$\eta_i = \partial_i \psi_F(\theta).$$

The dual potential function $\phi_F$ is called the negative $F$-entropy and is given by

$$\phi_F(\eta) = \max_{\theta} \{ \theta \cdot \eta - \psi_F(\theta) \} = E_{\hat{\beta}}(F(p)) = \frac{1}{h_F(\theta)} \int \frac{F(p)}{F'(p)} dx.$$

- We have, $\eta_i = \partial_i \psi_F(\theta) = E_{\hat{\beta}}(x_i); \ \partial_i \eta_j = \partial_i \partial_j \psi_F(\theta) = g_{ij}^F(\theta)$.
- With respect to the dual co-ordinate system $(\eta_j)$, the metric and the dual connections are given by

$$\tilde{g}_{ij}^D(\eta) = \partial^i \partial^j \phi_F(\eta); \ \tilde{\Gamma}_{ijk}^D(\eta) = 0; \ \tilde{\Gamma}_{ijk}^{D^*}(\eta) = \partial^i \partial^j \partial^k \phi_F(\eta).$$
6. Conformal flattening of \((F, G)\)-geometry

**Definition 6.1** Two statistical manifolds \((M, \nabla, h)\) and \((M, \tilde{\nabla}, \tilde{h})\) are said to be \(\beta\)-conformally equivalent if there exist a positive function \(K\) on \(M\) such that

\[
\tilde{h}(X, Y) = K \, h(X, Y) \\
\tilde{h}(\tilde{\nabla}_X Y, Z) = K \, h(\nabla_X Y, Z) - \frac{1 + \beta}{2} h(X, Y) dK(Z) + \frac{1 - \beta}{2} \{ h(Y, Z) dK(X) + h(X, Z) dK(Y) \}
\]

- In terms of the basis vectors, we can rewrite the above expression as

\[
\tilde{h}(\partial_i, \partial_j) = \tilde{h}_{ij} = K \, h(\partial_i, \partial_j) = K \, h_{ij}
\]

\[
\tilde{\Gamma}_{ijk}^\beta = K \, \Gamma_{ijk} - \frac{1 + \beta}{2} h_{ij} \partial_k K + \frac{1 - \beta}{2} \{ h_{jk} \partial_i K + h_{ik} \partial_j K \}
\]

- The dually flat structure on \(F\)-exponential family is obtained by the conformal flattening of \((F, G)\)-geometry.
Theorem 6.2

The metric \( g_{ij}^{DF} \) induced by the \( F \)-divergence \( D_F \) is obtained by the conformal flattening of the \( G \)-metric \( g_{ij}^{G} \) by a gauge function \( K(\theta) = \frac{1}{h_F(\theta)} \), with \( G(p) = \frac{-pF''(p)}{F'(p)} \).

**Proof:**

\[
g_{ij}^{DF}(\theta) = \partial_i \partial_j \psi_F(\theta) = \frac{1}{h_F(\theta)} \int \frac{-pF''(p)}{F'(p)} \partial_i p \partial_j p \frac{1}{p} dx
\]

\[
= K(\theta) g_{ij}^{G}
\]

where \( K(\theta) = \frac{1}{h_F(\theta)} \) and \( g_{ij}^{G} = \int \partial_i p \partial_j p \frac{G(p)}{p} dx \) is the \( G \)-metric with \( G(p) = \frac{-pF''(p)}{F'(p)} \).

Theorem 6.3

The affine connection \( \nabla^{DF} \) induced by \( D_F \) 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 \( F \).
**Proof:** The components of the connection $\nabla^{D_F}$ are given by

$$\Gamma^{D_F}_{ijk} = \frac{1}{h_F(\theta)} \int \left( \frac{-pF''(p)}{F'(p)} - \frac{p^2 F'''(p)}{F'(p)^2} + \frac{2p^2(F''(p))^2}{(F'(p))^2} \right) \partial_i \partial_j \partial_k p \, dx$$

$$+ \frac{1}{h_F(\theta)} \int \left( \frac{-pF''(p)}{F'(p)} \right) \partial_i \partial_j \partial_k p \, dx$$

$$+ \frac{1}{h_F(\theta)} \int \partial_j \partial_k \psi_F(\theta) \frac{pF''(p)}{(F'(p))^2} \partial_i \, dx$$

$$+ \frac{1}{h_F(\theta)} \int \partial_i \partial_k \psi_F(\theta) \frac{pF''(p)}{(F'(p))^2} \partial_j \, dx$$

$$= K(\theta) \Gamma^{H,G}_{ijk} + \partial_j K(\theta) g^{G}_{ik}(\theta) + \partial_i K(\theta) g^{G}_{jk}(\theta)$$

with $G(p) = \frac{-pF''(p)}{F'(p)}$ and $K(\theta) = \frac{1}{h_F(\theta)}$.

**Theorem 6.4**

The affine connection $\nabla^{D_F^*}$ induced by $D_F^*$ 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 $G(p) = \frac{-pF''(p)}{F'(p)}$. 

7. *F—likelihood estimator*

**Definition 7.1** Let $F$ be an increasing concave function and let $Z$ be its inverse function. Then the *$F$-product* of two numbers $x, y$ is defined as

$$x \otimes_F y = Z[F(x) + F(y)]$$

The *$F$-product* satisfies the following properties

- $Z(x) \otimes_F Z(y) = Z(x + y)$
- $F(x \otimes_F y) = F(x) + F(y)$

**Definition 7.2**

Two random variables $X$ and $Y$ are said to be *$F$—independent with normalization* if the joint probability density function $p_F(x, y)$ is given by the *$F$—product* of the marginal probability density functions $p_1(x)$ and $p_2(y)$.

$$p_F(x, y) = \frac{p_1(x) \otimes_F p_2(y)}{Z_{p_1, p_2}}$$

where $Z_{p_1, p_2}$ is the normalization defined by $Z_{p_1, p_2} = \int \int_{\Omega_1 \Omega_2} p_1(x) \otimes_F p_2(y) \, dx \, dy$
8. The geometry of $F$–likelihood estimators

Let $S = \{p(x; \theta) / \theta \in E \subseteq \mathbb{R}^n\}$ be an $n$–dimensional statistical manifold defined on a sample space $\Omega \subseteq \mathbb{R}$ and let $\{x^1, \ldots, x^N\}$ be $N$ independent observations from $p(x; \theta) \in S$.

**Definition 8.1**

The $F$–likelihood function $L_F(\theta)$ is defined as

$$L_F(\theta) = p(x^1; \theta) \otimes_F \cdots \otimes_F p(x^N; \theta)$$

Since $F$ is an increasing function, it is equivalent to consider $F(L_F(\theta))$ as well.

$$F(L_F(\theta)) = F(p(x^1; \theta)) \otimes_F \cdots \otimes_F F(p(x^N; \theta)) = \sum_{i=1}^N F(p(x^i; \theta))$$

**Definition 8.2**

A maximum $F$–likelihood estimator $\hat{\theta}$ is defined as

$$\hat{\theta} = \arg \max_{\theta \in E} L_F(\theta) = \arg \max_{\theta \in E} F(L_F(\theta))$$
Let \( S = \{ p(x; \theta) / \theta \in E \subseteq \mathbb{R}^n \} \) be a \( F \)-exponential family and let \( M \) be a curved \( F \)-exponential family in \( S \). Consider \( \{ x^1, \ldots, x^N \} \) be \( N \) independent observations from a probability density function \( p(x; u) = p(x; \theta(u)) \in M \).

**Theorem 8.3**

The \( F \)-likelihood estimator for \( M \) is the orthogonal projection of that of \( S \) to the submanifold \( M \) with respect to the connection \( \nabla^{DF^*} \).

**Proof:** The \( F \)-likelihood function is given by

\[
F(L_F(u)) = \sum_{j=1}^{N} F(p(x^j; u)) = \sum_{j=1}^{N} \left[ \sum_{i=1}^{n} \theta^i(u) x^j_i - \psi_F(\theta(u)) \right] = \sum_{i=1}^{n} \theta^i(u) \sum_{j=1}^{N} x^j_i - N \psi_F(\theta(u)).
\]

\[
\partial_i F(L_F(u)) = \sum_{j=1}^{N} x^j_i - N \partial_i \psi_F(\theta(u)).
\]
Thus the maximum $F$—likelihood estimator for $S$ is given by

$$\hat{\eta}_i = \frac{1}{N} \sum_{j=1}^{N} x^j_i.$$ 

The canonical divergence $D^*_F$ for $S$ can be calculated as

$$D^*_F[p(\theta(u)); p(\hat{\eta})] = D_F[p(\hat{\eta}); p(\theta(u))]$$

$$= \psi_F(\theta(u)) + \phi_F(\hat{\eta}) - \sum_{i=1}^{n} \theta^i(u)\hat{\eta}_i$$

$$= \phi_F(\hat{\eta}) - \frac{1}{N} F(L_F(u)).$$

- Hence the $F$—likelihood is maximum if the canonical divergence (or the dual of the $F$-divergence) is minimum.
- Equivalently, by the projection theorem, we can say that $F$—likelihood estimator for $M$ is the orthogonal projection of $\hat{\eta}$ to the submanifold $M$ with respect to the connection $\nabla^{D^*_F}$. 
9. $F$-Max-Ent Theorem

**Definition 9.1**

For any probability density function $p(x)$, the $F$-entropy is defined as

$$H_F(p) = -E_p(F(p)) = \frac{1}{h_F(p)} \int -F(p) \frac{F'(p)}{F'(p)} dx.$$  \hspace{1cm} (1)

- When $F(p) = \ln_q p$, the $q$-logarithm then $H_F(p)$ 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 9.2**

Probability distributions maximizing the $F$-entropy $H_F$ under the $F$-linear constraints

$$E_{\hat{p}_F}[c_k(x)] = a_k; \quad k = 1, \ldots, m$$  \hspace{1cm} (2)

for $m$ random variables $c_k(x)$ and various values of $a_k \in \mathbb{R}$ form an $m$-dimensional $F$-exponential family

$$F(p(x; \theta)) = \sum_{i=1}^{m} \theta^i c_i(x) - \psi(\theta)$$
Proof:

- We use the method of Lagrange multipliers and the calculus of variation principle.

Our aim is to maximize \( H_F(p) = \frac{1}{h_F(p)} \int \frac{-F(p)}{F'(p)} \, dx \) subject to the \( m \) constraints

\[
E_{\hat{\rho}_F} [c_k(x)] = \frac{1}{h_F(p)} \int \frac{c_k(x)}{F'(p)} \, dx = a_k; \quad k = 1, \ldots, m
\]

- Consider

\[
\mathcal{L}(p, \lambda_0, \lambda_1, \ldots, \lambda_m) = \frac{1}{h_F(p)} \int_0^\infty \frac{-F(p)}{F'(p)} \, dx + \lambda_0 \int_0^\infty p \, dx \\
+ \sum_{i=1}^m \lambda_i \frac{1}{h_F(p)} \int_0^\infty \frac{c_k(x)}{F'(p)} \, dx - \lambda_0 - \sum_{i=1}^m \lambda_i a_i
\]

- At maximum \( F \)-entropy distribution we have \( \frac{d\mathcal{L}}{dp} = 0 \).
Using this we get $\lambda_0 = \frac{1}{h_F(p)}$ and

$$F(p) = \sum_{i=1}^{m} \lambda_i (c_i(x) - a_i) + \frac{1}{h_F(p)} \int_0^{\infty} F(p) \frac{F'(p)}{F(p)} dx$$

$$= \sum_{i=1}^{m} \lambda_i (c_i(x) - a_i) - H_F(p)$$

$$= \sum_{i=1}^{m} \lambda_i c_i(x) - \sigma(\lambda_i, a_i)$$

where $\sigma(\lambda_i, a_i) = \sum_{i=1}^{m} \lambda_i a_i + H_F(p)$.

Using the $m$ constraints we can solve for $\lambda_i$ to get $\lambda_i = -\frac{dH_F(p)}{da_i}$ and $\lambda_i$’s are the canonical co-ordinates for the $F$-exponential family.

Hence $a_i$’s are the dual co-ordinate of the canonical co-ordinate $\lambda_i$.

Using the dual co-ordinates $\lambda_i, a_i$ and their potential functions, $F(p)$ takes the form of a $F$-exponential family. Thus

$$F(p(x; \theta)) = \sum_{i=1}^{m} \theta^i c_i(x) - \psi(\theta)$$
Conclusion

- Using a general embedding $F$ and a positive smooth function $G$, the $(F, G)$-geometry is introduced and the geometry of the $F$-exponential family is studied.

- The dually flat structure of the $F$-exponential family is obtained by the conformal flattening of the $(F, G)$-geometry.

- Using a generalized notion of independence called the $F$-independence, we defined the $F$-likelihood function.

- Further, the geometry of the $F$-likelihood estimators are discussed.

- A generalized notion of entropy called $F$-entropy is given and an analytic proof of the $F$-version of the max-ent theorem is outlined.

Future work

Further one can explore the asymptotic behavior of the mle of the $F$-escort probability density function and its various applications.
Acknowlegdement

We would like to express our sincere gratitude to Prof. Shun-ichi Amari for his valuable suggestions and fruitful discussions.


Thank You