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

K. V. Harsha and K. S. Subrahamanian Moosath

Department of Mathematics, Indian Institute of Space Science and Technology, Valiamala. P. O, Thiruvananthapuram-695547, Kerala, India

Abstract. We consider a family of probability distributions called F-exponential family which has got a dually flat structure obtained by the conformal flattening of the (F,G)-geometry. Geometry of F-likelihood estimator is discussed and the F-version of the maximum entropy theorem is proved.

Keywords: F-exponential family, (F,G)-connections, F-likelihood estimator, F-Max-ent Theorem
PACS: PACS 2010; 02.50.Cw, 02.50.Tt, 02.40.Ky

INTRODUCTION

The notion of exponential family is generalized by deforming the exponential function appearing in it which led to the q-exponential family of probability distributions having their entropic base in Tsalli's entropy [1], [2]. Naudts [3], [4], [5] studied them extensively and generalized to a large class of families of probability distributions. An information geometric foundation for the deformed exponential family was given by Amari et al. [6]. Also in [2], Amari and Ohara discussed the geometry of q-exponential family and the q-version of the Max-Ent theorem.

We also consider the generalized family of probability distributions based on the idea of (F,G)-geometry, called the F-exponential family. The F-exponential family has got a dually flat structure which is obtained by the conformal flattening of the (F,G)-geometry. Using a generalized notion of independence called the F-independence, we define F-likelihood function and the F-likelihood estimator. Further the geometry of F-likelihood estimator is discussed. Finally we give an analytic proof of the F-version of the maximum entropy theorem.

$F-$EXPONENTIAL FAMILY

In [6] Amari et al. considered a χ-family of probability distributions and studied the dually flat structure of it. Here from the context of (F,G)-geometry [7], we consider a generalized notion of exponential family called the F-exponential family and its dually flat structure is discussed. Moreover it is shown that the dually flat structure is obtained by the conformal flattening of the (F,G)-geometry.

The geometry of F-exponential family

Definition 0.1 *Let $F : (0,\infty) \longrightarrow \mathbb{R}$ be any smooth increasing concave function. Let Z be the inverse function of F. Then we define the standard form of an n-dimensional $F-$exponential family of distributions as*

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

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

Let $\mathscr{S} = \{p(x;\theta) \ / \ \theta \in E \subseteq \mathbb{R}^n\}$ be a $F-$exponential family. Now we analyze the geometric properties of $\mathscr{S}$ in detail.
Define a functional $h_F(\theta)$ as $h_F(\theta) = \int \frac{1}{F'(p(x;\theta))}dx$.

Theorem 0.2 *$F-$potential function $\psi_F(\theta)$ is a convex function of θ and*

$$\partial_i\partial_j\psi_F(\theta) = \frac{1}{h_F(\theta)}\int \frac{-F''(p)}{(F'(p))^3}\,\partial_i F\,\partial_j F dx = \frac{1}{h_F(\theta)}\int \frac{-pF''(p)}{F'(p)}\,\partial_i p\,\partial_j p\,\frac{1}{p}dx$$

Definition 0.3 *Let us define a Riemannian metric called $F-$metric g^F by*

$$g^F_{ij}(\theta) = \partial_i\partial_j\psi_F(\theta). \qquad (1)$$

Note that (g^F_{ij}) is positive definite since ψ_F is a convex function of θ.

Definition 0.4 *Define a divergence of Bregman-type using $\psi_F(\theta)$, called the $F-$**divergence**, as follows*

$$D_F[p(x;\theta_1) : p(x;\theta_2)] = \psi_F(\theta_2) - \psi_F(\theta_1) - \nabla\psi_F(\theta_1).(\theta_2 - \theta_1). \qquad (2)$$

The two distributions p and r which are parametrized by θ_1 and θ_2 respectively. Then we can rewrite the $F-$divergence as

$$D_F[p : r] = \frac{1}{h_F(\theta_1)}\int (F(p) - F(r))\frac{1}{F'(p)}dx. \qquad (3)$$

Definition 0.5 *For a distribution function p parametrized by θ, define a probability distribution called the $F-$**escort probability distribution** related to p as*

$$\hat{p}_F(x) = \frac{1}{h_F(\theta)F'(p)} \ , \ \text{if } h_F(\theta) = \int \frac{1}{F'(p)}dx \quad \text{exist.} \qquad (4)$$

Definition 0.6 *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 $F-$divergence can be written as $D_F[p:r] = E_{\hat{p}}(F(p) - F(r))$.
The geometric structures induced by the $F-$divergence D_F can be obtained as

Lemma 0.7 *The metric $g_{ij}^{D_F}$ and the affine connection ∇^{D_F} induced by the $F-$divergence D_F are given by*

$$g_{ij}^{D_F}(\theta) = g_{ij}^F(\theta) = \partial_i\partial_j\psi_F(\theta); \quad \Gamma_{ijk}^{D_F} = \partial_i\partial_j\partial_k\psi_F(\theta).$$

The dual D_F^* of D_F induces an affine connection $\nabla^{D_F^*}$ with components $\Gamma_{ijk}^{D_F^*} = 0$.

Note 0.8 *The Legendre transformation of the convex function $\psi_F(\theta)$ is given by $\eta_i = \partial_i\psi_F(\theta)$. Since there is a one to one correspondence between η and θ, we can take η as another co-ordinate system for $\mathscr{S}$. The dual potential function is given by*

$$\phi_F(\eta) = \max_\theta \{\theta.\eta - \psi_F(\theta)\}.$$

We have $\theta^i = \partial^i\phi_F(\eta)$; $\partial^i = \frac{\partial}{\partial\eta_i}$, so that η and θ are in dual correspondence.

$$\eta_i = \partial_i\psi_F(\theta) = E_{\hat{p}}(x_i); \quad \partial_i\eta_j = \partial_i\partial_j\psi_F(\theta) = g_{ij}^F(\theta).$$

Now with respect to the dual co-ordinate system (η_j) of (θ^i), the metric and the dual connections are given by

$$\tilde{g}_{ij}^{D_F}(\eta) = \partial^i\partial^j\phi_F(\eta); \quad \tilde{\Gamma}_{ijk}^{D_F}(\eta) = 0; \quad \tilde{\Gamma}_{ijk}^{D_F^*}(\eta) = \partial^i\partial^j\partial^k\phi_F(\eta). \tag{5}$$

Lemma 0.9 *The dual potential function $\phi_F(\eta)$ is given by*

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

Remark 0.10 *On the F-exponential family $\mathscr{S}$, the F-divergence D_F induces a dually flat structure $(g^{D_F}, \nabla^{D_F}, \nabla^{D_F^*})$. In this dually flat space, using the canonical divergence, we can have the Pythagorean theorem and the projection theorem [12]. The potential function ψ_F of the canonical parameter (θ^i) can be called as the $F-$**free energy**. Since $F-$potential function $\phi_F(\eta)$ is the Legendre dual of the $F-$free energy $\psi_F(\theta)$, we can call it as negative $F-$**entropy**.*

Conformal flattening of (F,G)-geometry

In [7], we introduced $(F,G)-$geometry with metric g^G and dual connections $\nabla^{F,G}$, $\nabla^{H,G}$. Now we show that the geometry induced by the F-divergence D_F is obtained by the conformal flattening([8], [9]) of (F,G)-geometry.

Lemma 0.11 *The metric $g_{ij}^{D_F}$ induced by the F-divergence D_F is 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
$$g_{ij}^{D_F}(\theta) = \partial_i\partial_j\psi_F(\theta) = \frac{1}{h_F(\theta)}\int \frac{-pF''(p)}{F'(p)}\,\partial_i p\,\partial_j p\,\frac{1}{p}dx \tag{6}$$
$$= K(\theta)g_{ij}^G \tag{7}$$

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 $G(p) = \frac{-pF''(p)}{F'(p)}$. Thus the new metric is obtained as a conformal transformation of the $G-$metric by a gauge function $K(\theta)$. ■

Theorem 0.12 *The affine connection ∇^{D_F} induced by the $F-$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.*

Proof The $G-$dual embedding H of F is defined by $H'(P) = \frac{G(p)}{pF'(p)}$.
When $G(p) = \frac{-pF''(p)}{F'(p)}$, then $1+\frac{pH''(p)}{H'(p)} = 1-\frac{2pF''(p)}{F'(p)}+\frac{pF'''(p)}{F'(p)}$.
Then we can rewrite the components of the connection $\nabla^{(H,G)}$ as

$$\Gamma_{ijk}^{(H,G)}(\theta) = \int\left[\partial_i\partial_j\ell\,\partial_k\ell\,+(1+\frac{pH''(p)}{H'(p)})\partial_i\ell\,\partial_j\ell\,\partial_k\ell\right]G(p)\,p\,dx \tag{8}$$
$$= \frac{1}{h_F(\theta)}\int\left(\frac{-pF''(p)}{F'(p)}-\frac{p^2F'''(p)}{F'(p)}+\frac{2p^2(F''(p))^2}{(F'(p))^2}\right)\partial_i\ell\,\partial_j\ell\,\partial_k\ell\,pdx$$
$$+\ \frac{1}{h_F(\theta)}\int(\frac{-pF''(p)}{F'(p)})\partial_i\partial_j\ell\,\partial_k\ell\,pdx \tag{9}$$

Now when $K(\theta) = \frac{1}{h_F(\theta)}$ and $G(p) = \frac{-pF''(p)}{F'(p)}$, we have

$$\partial_i K(\theta)g_{jk}^G(\theta) = \frac{-1}{(h_F(\theta))^2}\left(\int\frac{pF''(p)}{(F'(p))^2}\,\partial_i\ell\,dx\right)\int\frac{pF''(p)}{F'(p)}\partial_j\ell\,\partial_k\ell\,p\,dx \tag{10}$$

The components of the connection ∇^{D_F} are given by

$$\Gamma_{ijk}^{D_F} = \frac{1}{h_F(\theta)}\int\left(\frac{-pF''(p)}{F'(p)}-\frac{p^2F'''(p)}{F'(p)}+\frac{2p^2(F''(p))^2}{(F'(p))^2}\right)\partial_i\ell\,\partial_j\ell\,\partial_k\ell\,pdx$$
$$+\ \frac{1}{h_F(\theta)}\int(\frac{-pF''(p)}{F'(p)})\partial_i\partial_j\ell\,\partial_k\ell\,pdx$$
$$+\ \frac{1}{h_F(\theta)}\int\partial_j\partial_k\psi_F(\theta)\frac{pF''(p)}{(F'(p))^2}\,\partial_i\ell\,dx$$
$$+\ \frac{1}{h_F(\theta)}\int\partial_i\partial_k\psi_F(\theta)\frac{pF''(p)}{(F'(p))^2}\,\partial_j\ell\,dx \tag{11}$$
$$= K(\theta)\Gamma_{ijk}^{H,G}+\partial_j K(\theta)g_{ik}^G(\theta)+\partial_i K(\theta)g_{jk}^G(\theta) \tag{12}$$

using (9) and (10) 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 (H,G)-

connection $\nabla^{H,G}$ by a gauge function $K(\theta)$. ■
Similarly one can prove that

Theorem 0.13 *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 $G(p) = \frac{-pF''(p)}{F'(p)}$.*

Thus the dually flat structure on F-exponential family is the conformal flattening of the $(F,G)-$geometry. For $F(p) = \ln_q(p)$ and $G(p) = constant$, $(F,G)-$geometry is the Amari's α-geometry (upto a constant factor). Then the $F-$exponential family reduces to q-exponential family and the q-geometry is the conformal flattening of the α-geometry [2], [10].

$F-$LIKELIHOOD ESTIMATOR

Fujimoto and Murata [11] introduced a generalized notion of independence called U-independence. Here the generalized independence is defined using a smooth increasing function F and its inverse as in [1]. Using this, generalized likelihood function and likelihood estimators are defined. Further the geometry of the likelihood estimators is discussed.

$F-$independence

Let F be an increasing concave function and Z be its inverse function. Define the F-product of two numbers x, y as

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

The $F-$product satisfies the following properties

$$Z(x) \otimes_F Z(y) = Z(x+y); \quad F(x \otimes_F y) = F(x) + F(y)$$

Definition 0.14 *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 $p_2(y)$,*

$$p_F(x,y) = \frac{p_1(x) \otimes_F p_2(y)}{Z_{p_1,p_2}} \tag{13}$$

where Z_{p_1,p_2} is the normalization defined by $Z_{p_1,p_2} = \int\int_{\Omega_1\Omega_2} p_1(x) \otimes_F p_2(y) dxdy$

The geometry of $F-$likelihood estimators

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

$p(x;\theta) \in \mathscr{S}$. Let us define a $F-$likelihood function $L_F(\theta)$ as

$$L_F(\theta) = p(x^1;\theta) \otimes_F \ldots. \otimes_F p(x^N;\theta) \tag{14}$$

When $F(p) = \log_q p$, then $L_F(\theta)$ reduces to $q-$likelihood function $L_q(\theta)$ defined by Matsuzoe and Ohara [1].
Since F is an increasing function, it is equivalent to consider $F(L_F(\theta))$ as well.

$$F(L_F(\theta)) = F(p(x^1;\theta)) \otimes_F \ldots. \otimes_F F(p(x^N;\theta)) = \sum_{i=1}^{N} F(p(x^i;\theta)) \tag{15}$$

Definition 0.15 *A maximum $F-$likelihood estimator $\hat{\theta}$ is defined as*

$$\hat{\theta} = \arg\max_{\theta \in E} L_F(\theta) = \arg\max_{\theta \in E} F(L_F(\theta)) \tag{16}$$

Now let us look at the geometry of $F-$likelihood estimators for the F-exponential family. Let $\mathscr{S} = \{p(x;\theta) \; / \; \theta \in E \subseteq \mathbb{R}^n\}$ be a $F-$exponential family and let M be a curved F-exponential family in $\mathscr{S}$. 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 0.16 *The $F-$likelihood estimator for M is the orthogonal projection of that of $\mathscr{S}$ to the submanifold M with respect to the connection $\nabla^{D_F^*}$.*

Proof The $F-$likelihood function is given by

$$F(L_F(u)) = \sum_{j=1}^{N} F(p(x^j;u)) \quad = \quad \sum_{j=1}^{N} \left[\sum_{i=1}^{n} \theta^i(u) x_i^j - \psi_F(\theta(u)) \right] \tag{17}$$

$$= \quad \sum_{i=1}^{n} \theta^i(u) \sum_{j=1}^{N} x_i^j - N\psi_F(\theta(u)) \tag{18}$$

$$\partial_i F(L_F(u)) = \sum_{j=1}^{N} x_i^j - N\partial_i \psi_F(\theta(u)) \tag{19}$$

Thus the maximum $F-$likelihood estimator for $\mathscr{S}$ is given by $\hat{\eta}_i = \frac{1}{N}\sum_{j=1}^{N} x_i^j$.
The canonical divergence D_F^* for $\mathscr{S}$ can be calculated as

$$D_F^*[p(\theta(u)); p(\hat{\eta})] = D_F[p(\hat{\eta}); p(\theta(u))] \tag{20}$$

$$= \psi_F(\theta(u)) + \phi_F(\hat{\eta}) - \sum_{i=1}^{n} \theta^i(u)\hat{\eta}_i \tag{21}$$

$$= \phi_F(\hat{\eta}) - \frac{1}{N} F(L_F(u)) \tag{22}$$

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 $\nabla^{D_F^*}$. ∎

F-MAX-ENT THEOREM

In [6], Amari et al. gave a geometric proof of the χ- version of the Max-ent theorem. Here we define the F-entropy and give an analytic proof of the F-version of the Max-ent theorem.

Definition 0.17 *For any probability density function $p(x)$, the $F-$entropy is defined as*

$$H_F(p) = -E_{\hat{p}}(F(p)) = \frac{1}{h_F(p)} \int \frac{-F(p)}{F'(p)} dx \tag{23}$$

if $\int \frac{-F(p)}{F'(p)} dx$ and $h_F(p) = \int \frac{1}{F'(p)} dx$ exist.

When $F(p) = \ln_q p$, 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 0.18 ***(F-Max-ent theorem)***
Probability distributions maximizing the F-entropy H_F under the F-linear constraints

$$E_{\hat{p_F}}[c_k(x)] = a_k;\ \ k = 1, ..., m \tag{24}$$

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 Here, we use the method of Lagrange multipliers and the calculus of variation principle.
To maximize $H_F(p) = \frac{1}{h_F(p)} \int \frac{-F(p)}{F'(p)} dx$ subject to the m constraints

$$E_{\hat{p_F}}[c_k(x)] = \frac{1}{h_F(p)} \int \frac{c_k(x)}{F'(p)} dx = a_k;\ \ k = 1, ..., m \tag{25}$$

Consider,
$$\begin{aligned}\mathscr{L}(p, \lambda_0, \lambda_1, .., \lambda_m) = & \frac{1}{h_F(p)} \int_0^\infty \frac{-F(p)}{F'(p)} dx + \lambda_0 \int_0^\infty p dx \\ & + \sum_{i=1}^{m} \lambda_i \frac{1}{h_F(p)} \int_0^\infty \frac{c_k(x)}{F'(p)} dx - \lambda_0 - \sum_{i=1}^{m} \lambda_i a_i \end{aligned} \tag{26}$$

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

$$F(p) = \sum_{i=1}^{m} \lambda_i (c_i(x) - a_i) + \frac{1}{h_F(p)} \int_0^\infty \frac{F(p)}{F'(p)} dx \tag{27}$$

$$= \sum_{i=1}^{m} \lambda_i (c_i(x) - a_i) - H_F(p) \tag{28}$$

$$= \sum_{i=1}^{m} \lambda_i c_i(x) - \sigma(\lambda_i, a_i) \tag{29}$$

where $\sigma(\lambda_i, a_i) = \sum_{i=1}^{m} \lambda_i a_i + H_F(p)$.
Now using the m constraints we can solve for λ_i to get $\lambda_i = -\frac{dH_F(p)}{da_i}$ and λ_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 λ_i. Using the dual co-ordinates λ_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

Dually flat structure of the F-exponential family is obtained by the conformal flattening of the (F,G)-geometry. The geometry of the F-likelihood estimators and the F-version of the max-ent theorem are discussed. Further one can explore the asymptotic behavior of the mle of the F-escort probability density function and its various applications.

ACKNOWLEDGMENTS

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

REFERENCES

1. H. Matsuzoe, and A. Ohara, "Geometry for $q-$exponential families", Proceedings of the 2nd International Colloquium on Differential Geometry and its Related Fields, Veliko Tarnovo, September 6-10, 2010.
2. S.I. Amari, and A. Ohara, "Geometry of q-Exponential Family of Probability Distributions", *Entropy*, **13**, 1170–1185 (2011).
3. J. Naudts, "Deformed exponentials and logarithms in generalized thermostatistics", *Physica A*, **316** 323-334(2002).
4. J. Naudts, "Estimators, escort probabilities, and phi-exponential families in statistical physics", *J. Ineq. Pure Appl. Math.*, **5**, 102(2004).
5. J. Naudts, *Generalised Thermostatistics*; Springer: London, UK, 2011.
6. S.I. Amari,; A. Ohara, ; H. Matsuzoe, "Geometry of deformed exponential families: Invariant, dually flat and conformal geometries", *Physica A:Statistical Mechanics and its Applications*, textbf391, 4308-4319 (2012).
7. K.V. Harsha, and K.S. Subrahamanian Moosath, "F-geometry and Amari's α-geometry on a statistical manifold", *Entropy*, **16(5)**, 2472-2487 (2014).
8. T. Kurose, "On the Divergence of 1-conformally Flat Statistical Manifolds", *TĔĘohoku Math. J.*, **46**, 427âĂŞ433 (1994).
9. T. Kurose, "Conformal-projective geometry of statistical manifolds", *Interdisciplinary Information Sciences*, **8**, 89âĂŞ100(2002).
10. A. Ohara, ; H. Matsuzoe, and S.I. Amari, "A dually flat structure on the space of escort distributions", Journal of Physics: Conference Series, 201(01), 2010.
11. Y. Fujimoto, and N. Murata, " A generalization of Independence in Naive Bayes Model", *Lecture Notes in Computer Science*, 153-161(2010).
12. S.I. Amari, and H. Nagaoka, *Methods of Information Geometry, Translations of Mathematical Monographs*, Oxford University Press, Oxford, UK, 2000.