Chapter 2
Exponential Families and Mixture Families of Probability Distributions

The present chapter studies the geometry of the exponential family of probability distributions. It is not only a typical statistical model, including many well-known families of probability distributions such as discrete probability distributions S_n, Gaussian distributions, multinomial distributions, gamma distributions, etc., but is associated with a convex function known as the cumulant generating function or free energy. The induced Bregman divergence is the KL-divergence. It defines a dually flat Riemannian structure. The derived Riemannian metric is the Fisher information matrix and the two affine coordinate systems are the natural (canonical) parameters and expectation parameters, well-known in statistics. An exponential family is a universal model of dually flat manifolds, because any Bregman divergence has a corresponding exponential family of probability distributions (Banerjee et al. 2005).

We also study the mixture family of probability distributions, which is the dual of the exponential family. Applications of the generalized Pythagorean theorem demonstrate how useful this is.

2.1 Exponential Family of Probability Distributions

The standard form of an exponential family is given by the probability density function

$$p(x, \boldsymbol{\theta}) = \exp\left\{\theta^i h_i(x) + k(x) - \psi(\boldsymbol{\theta})\right\}, \tag{2.1}$$

where x is a random variable, $\boldsymbol{\theta} = \left(\theta^1, \ldots, \theta^n\right)$ is an n-dimensional vector parameter to specify a distribution, $h_i(x)$ are n functions of x which are linearly independent, $k(x)$ is a function of x, ψ corresponds to the normalization factor and the Einstein summation convention is working. We introduce a new vector random variable $\boldsymbol{x} = (x_1, \ldots, x_n)$ by

$$x_i = h_i(x). \tag{2.2}$$

We further introduce a measure in the sample space $X = \{\boldsymbol{x}\}$ by

S. Amari, *Information Geometry and Its Applications*,
Applied Mathematical Sciences 194, DOI 10.1007/978-4-431-55978-8_2

$$d\mu(\boldsymbol{x}) = \exp\{k(x)\}\, dx. \tag{2.3}$$

Then, (2.1) is rewritten as

$$p(x, \boldsymbol{\theta})dx = \exp\{\boldsymbol{\theta} \cdot \boldsymbol{x} - \psi(\boldsymbol{\theta})\}\, d\mu(\boldsymbol{x}). \tag{2.4}$$

Hence, we may put

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\{\boldsymbol{\theta} \cdot \boldsymbol{x} - \psi(\boldsymbol{\theta})\}, \tag{2.5}$$

which is a probability density function of $\boldsymbol{x}$ with respect to measure $d\mu(\boldsymbol{x})$.

The family of distributions

$$M = \{p(\boldsymbol{x}, \boldsymbol{\theta})\} \tag{2.6}$$

forms an n-dimensional manifold, where $\boldsymbol{\theta}$ is a coordinate system. From the normalization condition

$$\int p(\boldsymbol{x}, \boldsymbol{\theta})d\mu(\boldsymbol{x}) = 1, \tag{2.7}$$

ψ is written as

$$\psi(\boldsymbol{\theta}) = \log \int \exp(\boldsymbol{\theta} \cdot \boldsymbol{x})d\mu(\boldsymbol{x}). \tag{2.8}$$

We proved in Chap. 1 that $\psi(\boldsymbol{\theta})$ is a convex function of $\boldsymbol{\theta}$, known as the cumulant generating function in statistics and free energy in physics. A dually flat Riemannian structure is introduced in M by using $\psi(\boldsymbol{\theta})$. The affine coordinate system is $\boldsymbol{\theta}$, which is called the natural or canonical parameter of an exponential family. The dual affine parameter is given by the Legendre transformation,

$$\boldsymbol{\theta}^* = \nabla\psi(\boldsymbol{\theta}), \tag{2.9}$$

which is the expectation of $\boldsymbol{x}$ denoted by $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$,

$$\boldsymbol{\eta} = \mathbf{E}[\boldsymbol{x}] = \int \boldsymbol{x}\, p(\boldsymbol{x}, \boldsymbol{\theta})d\mu(\boldsymbol{x}). \tag{2.10}$$

This $\boldsymbol{\eta}$ is called the expectation parameter in statistics. Since the dual affine parameter $\boldsymbol{\theta}^*$ is nothing other than $\boldsymbol{\eta}$, we hereafter use $\boldsymbol{\eta}$, instead of $\boldsymbol{\theta}^*$, to represent the dual affine parameter in an exponential family. This is a conventional notation used in Amari and Nagaoka (2000), avoiding the cumbersome $*$ notation. So we have

$$\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta}). \tag{2.11}$$

Hence, $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are two affine coordinate systems connected by the Legendre transformation.

We use $\varphi(\boldsymbol{\eta})$ to denote the dual convex function $\psi^*\left(\boldsymbol{\theta}^*\right)$, the Legendre dual of ψ, which is defined by

$$\varphi(\boldsymbol{\eta}) = \max_{\boldsymbol{\theta}} \{\boldsymbol{\theta} \cdot \boldsymbol{\eta} - \psi(\boldsymbol{\theta})\}. \tag{2.12}$$

In order to obtain $\varphi(\boldsymbol{\eta})$, we calculate the negative entropy of $p(\boldsymbol{x}, \boldsymbol{\theta})$, obtaining

$$\mathrm{E}\left[\log p(\boldsymbol{x}, \boldsymbol{\theta})\right] = \int p(\boldsymbol{x}, \boldsymbol{\theta}) \log p(\boldsymbol{x}, \boldsymbol{\theta}) d\mu(\boldsymbol{x}) = \boldsymbol{\theta} \cdot \boldsymbol{\eta} - \psi(\boldsymbol{\theta}). \tag{2.13}$$

Given $\boldsymbol{\eta}$, the $\boldsymbol{\theta}$ that maximizes the right-hand side of (2.12) is given by the solution of $\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta})$. Hence, the dual convex function ψ^* of ψ, which we hereafter denote as φ, is given by the negative entropy,

$$\varphi(\boldsymbol{\eta}) = \int p(\boldsymbol{x}, \boldsymbol{\theta}) \log p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x}, \tag{2.14}$$

where $\boldsymbol{\theta}$ is regarded as a function of $\boldsymbol{\eta}$ through $\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta})$. The inverse transformation is given by

$$\boldsymbol{\theta} = \nabla\varphi(\boldsymbol{\eta}). \tag{2.15}$$

The divergence from $p(\boldsymbol{x}, \boldsymbol{\theta}')$ to $p(\boldsymbol{x}, \boldsymbol{\theta})$ is written as

$$\begin{aligned} D_\psi\left[\boldsymbol{\theta}' : \boldsymbol{\theta}\right] &= \psi\left(\boldsymbol{\theta}'\right) - \psi(\boldsymbol{\theta}) - \boldsymbol{\eta} \cdot \left(\boldsymbol{\theta}' - \boldsymbol{\theta}\right) \\ &= \int p(\boldsymbol{x}, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{x}, \boldsymbol{\theta})}{p\left(\boldsymbol{x}, \boldsymbol{\theta}'\right)} d\mu(\boldsymbol{x}) = D_{KL}\left[\boldsymbol{\theta} : \boldsymbol{\theta}'\right]. \end{aligned} \tag{2.16}$$

The Riemannian metric is given by

$$g_{ij}(\boldsymbol{\theta}) = \partial_i \partial_j \psi(\boldsymbol{\theta}), \tag{2.17}$$

$$g^{ij}(\boldsymbol{\eta}) = \partial^i \partial^j \varphi(\boldsymbol{\eta}), \tag{2.18}$$

for which we hereafter use the abbreviation

$$\partial_i = \frac{\partial}{\partial\theta^i}, \quad \partial^i = \frac{\partial}{\partial\eta_i}. \tag{2.19}$$

Here, the position of the index i is important. If it is lower, as in ∂_i, the differentiation is with respect to θ^i, whereas, if it is upper as in ∂^i, the differentiation is with respect to η_i.

The Fisher information matrix plays a fundamental role in statistics. We prove the following theorem which connects geometry and statistics.

Theorem 2.1 *The Riemannian metric in an exponential family is the Fisher information matrix defined by*

$$g_{ij} = \mathrm{E}\left[\partial_i \log p(\boldsymbol{x}, \boldsymbol{\theta}) \partial_j \log p(\boldsymbol{x}, \boldsymbol{\theta})\right]. \tag{2.20}$$

Proof From

$$\partial_i \log p(\boldsymbol{x}, \boldsymbol{\theta}) = x_i - \partial_i \psi(\boldsymbol{\theta}) = x_i - \eta_i, \tag{2.21}$$

we have

$$\mathbf{E}\left[\partial_i \log p(\boldsymbol{x}, \boldsymbol{\theta}) \partial_j \log p(\boldsymbol{x}, \boldsymbol{\theta})\right] = \mathbf{E}\left[(x_i - \eta_i)\left(x_j - \eta_j\right)\right], \tag{2.22}$$

which is equal to $\nabla\nabla\psi(\boldsymbol{\theta})$. This is the Riemannian metric derived from $\psi(\boldsymbol{\theta})$, as is shown in (1.56). □

2.2 Examples of Exponential Family: Gaussian and Discrete Distributions

There are many statistical models belonging to the exponential family. Here, we show only two well-known, important distributions.

2.2.1 Gaussian Distribution

The Gaussian distribution with mean μ and variance σ^2 has the probability density function

$$p(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{2.23}$$

We introduce a new vector random variable $\boldsymbol{x} = (x_1, x_2)$,

$$x_1 = h_1(x) = x, \tag{2.24}$$

$$x_2 = h_2(x) = x^2. \tag{2.25}$$

Note that x and x^2 are dependent, but are linearly independent. We further introduce new parameters

$$\theta^1 = \frac{\mu}{\sigma^2}, \tag{2.26}$$

$$\theta^2 = -\frac{1}{2\sigma^2}. \tag{2.27}$$

Then, (2.23) is written in the standard form,

$$p(x, \boldsymbol{\theta}) = \exp\left\{\boldsymbol{\theta} \cdot \boldsymbol{x} - \psi(\boldsymbol{\theta})\right\}. \tag{2.28}$$

The convex function $\psi(\boldsymbol{\theta})$ is given by

$$\begin{aligned}\psi(\boldsymbol{\theta}) &= \frac{\mu^2}{2\sigma^2} + \log\left(\sqrt{2\pi}\sigma\right) \\ &= -\frac{\left(\theta^1\right)^2}{4\theta^2} - \frac{1}{2}\log\left(-\theta^2\right) + \frac{1}{2}\log\pi. \end{aligned} \tag{2.29}$$

Since x_1 and x_2 are not independent but satisfy the relation

$$x_2 = (x_1)^2, \tag{2.30}$$

we use the dominating measure of

$$d\mu(\boldsymbol{x}) = \delta\left(x_2 - x_1^2\right) dx, \tag{2.31}$$

where δ is the delta function.

The dual affine coordinates $\boldsymbol{\eta}$ are given from (2.10) as

$$\eta_1 = \mu, \quad \eta_2 = \mu^2 + \sigma^2. \tag{2.32}$$

2.2.2 Discrete Distribution

Distributions of discrete random variable x over $X = \{0, 1, \ldots, n\}$ form a probability simplex S_n. A distribution $\boldsymbol{p} = (p_0, p_1, \ldots, p_n)$ is represented by

$$p(x) = \sum_{i=0}^{n} p_i \delta_i(x). \tag{2.33}$$

We show that S_n is an exponential family. We have

$$\begin{aligned}\log p(x) &= \sum_{i=0}^{n} (\log p_i)\, \delta_i(x) = \sum_{i=1}^{n} (\log p_i)\, \delta_i(x) + (\log p_0)\, \delta_0(x) \\ &= \sum_{i=1}^{n} \left(\log \frac{p_i}{p_0}\right) \delta_i(x) + \log p_0, \end{aligned} \tag{2.34}$$

because of

$$\delta_0(x) = 1 - \sum_{i=1}^{n} \delta_i(x). \tag{2.35}$$

We introduce new random variables x_i,

$$x_i = h_i(x) = \delta_i(x), \quad i = 1, \ldots, n \tag{2.36}$$

and new parameters

$$\theta^i = \log \frac{p_i}{p_0}. \tag{2.37}$$

Then, a discrete distribution $\boldsymbol{p}$ is written from (2.34) as

$$p(x, \boldsymbol{\theta}) = \exp\left\{ \sum_{i=1}^{n} \theta^i x_i - \psi(\boldsymbol{\theta}) \right\}, \tag{2.38}$$

where the cumulant generating function is

$$\psi(\boldsymbol{\theta}) = -\log p_0 = \log\left\{ 1 + \sum_{i=1}^{n} \exp\left(\theta^i\right) \right\}. \tag{2.39}$$

The dual affine coordinates $\boldsymbol{\eta}$ are

$$\eta_i = E\left[h_i(x)\right] = p_i, \quad i = 1, \ldots, n. \tag{2.40}$$

The dual convex function is the negative entropy,

$$\varphi(\boldsymbol{\eta}) = \sum \eta_i \log \eta_i + \left(1 - \sum \eta_i\right) \log\left(1 - \sum \eta_i\right). \tag{2.41}$$

By differentiating it, we have $\boldsymbol{\theta} = \nabla\varphi(\boldsymbol{\eta})$.

$$\theta^i = \log \frac{\eta_i}{1 - \sum \eta_i}. \tag{2.42}$$

2.3 Mixture Family of Probability Distributions

A mixture family is in general different from an exponential family, but family S_n of discrete distributions is an exponential family and a mixture family at the same time. We show that the two families play a dual role.

Given $n + 1$ probability distributions $q_0(x), q_1(x), \ldots, q_n(x)$ which are linearly independent, we compose a family of probability distributions given by

$$p(x, \boldsymbol{\eta}) = \sum_{i=0}^{n} \eta_i q_i(x), \tag{2.43}$$

where

$$\sum_{i=0}^{n} \eta_i = 1, \quad \eta_i > 0. \tag{2.44}$$

This is a statistical model called a mixture family, where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ is a coordinate system and $\eta_0 = 1 - \sum \eta_i$. (We sometimes consider the closure of the above family, where $\eta_i \geq 0$.)

As is easily seen from (2.33), a discrete distribution $p(x) \in S_n$ is a mixture family, where

$$q_i(x) = \delta_i(x), \quad \eta_i = p_i, \quad i = 0, 1, \ldots, n. \tag{2.45}$$

Hence, $\boldsymbol{\eta}$ is a dual affine coordinate system of the exponential family S_n. We consider a general mixture family (2.43) which is not an exponential family. Even in this case, the negative entropy

$$\varphi(\boldsymbol{\eta}) = \int p(x, \boldsymbol{\eta}) \log p(x, \boldsymbol{\eta}) dx \tag{2.46}$$

is a convex function of $\boldsymbol{\eta}$. Therefore, we regard it as a dual convex function and introduce the dually flat structure to $M = \{p(x, \boldsymbol{\eta})\}$, having $\boldsymbol{\eta}$ as the dual affine coordinate system. Then, the primary affine coordinates are given by the gradient,

$$\boldsymbol{\theta} = \nabla \varphi(\boldsymbol{\eta}). \tag{2.47}$$

It defines the primal affine structure dually coupled with $\boldsymbol{\eta}$, although $\boldsymbol{\theta}$ is not the natural parameter of an exponential family, except for the case of S_n where $\boldsymbol{\theta}$ is the natural parameter.

The divergence given by $\varphi(\boldsymbol{\eta})$ is the KL-divergence

$$D_{\varphi}\left[\boldsymbol{\eta} : \boldsymbol{\eta}'\right] = \int p(x, \boldsymbol{\eta}) \log \frac{p(x, \boldsymbol{\eta})}{p\,(x, \boldsymbol{\eta}')} dx. \tag{2.48}$$

2.4 Flat Structure: *e*-flat and *m*-flat

The manifold M of exponential family is dually flat. The primal affine coordinates which define straightness or flatness are the natural parameter $\boldsymbol{\theta}$ in an exponential family. Let us consider the straight line, that is a geodesic, connecting two distributions $p(\boldsymbol{x}, \boldsymbol{\theta}_1)$ and $p(\boldsymbol{x}, \boldsymbol{\theta}_2)$. This is written in the $\boldsymbol{\theta}$ coordinate system as

$$\boldsymbol{\theta}(t) = (1 - t)\boldsymbol{\theta}_1 + t\boldsymbol{\theta}_2, \tag{2.49}$$

where t is the parameter. The probability distributions on the geodesic are

$$p(\boldsymbol{x}, t) = p\{\boldsymbol{x}, \boldsymbol{\theta}(t)\} = \exp\{t(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) \cdot \boldsymbol{x} + \boldsymbol{\theta}_1 \boldsymbol{x} - \psi(t)\}. \tag{2.50}$$

Hence, a geodesic itself is a one-dimensional exponential family, where t is the natural parameter.

By taking the logarithm, we have

$$\log p(\boldsymbol{x}, t) = (1-t) \log p(\boldsymbol{x}, \boldsymbol{\theta}_1) + t \log p(\boldsymbol{x}, \boldsymbol{\theta}_2) - \psi(t). \tag{2.51}$$

Therefore, a geodesic consists of a linear interpolation of the two distributions in the logarithmic scale. Since (2.51) is an exponential family, we call it an e-geodesic, e standing for "exponential". More generally, a submanifold which is defined by linear constraints in $\boldsymbol{\theta}$ is said to be e-flat. The affine parameter $\boldsymbol{\theta}$ is called the e-affine parameter.

The dual affine coordinates are $\boldsymbol{\eta}$, and define the dual flat structure. The dual geodesic connecting two distributions specified by $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ is given by

$$\boldsymbol{\eta}(t) = (1-t)\boldsymbol{\eta}_1 + t\boldsymbol{\eta}_2 \tag{2.52}$$

in terms of the dual coordinate system. Along the dual geodesic, the expectation of $\boldsymbol{x}$ is linearly interpolated,

$$\mathbf{E}_{\eta(t)}[\boldsymbol{x}] = (1-t)\mathbf{E}_{\eta_1}[\boldsymbol{x}] + t\mathbf{E}_{\eta_2}[\boldsymbol{x}]. \tag{2.53}$$

In the case of discrete probability distributions S_n, the dual geodesic connecting $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ is

$$\boldsymbol{p}(t) = (1-t)\boldsymbol{p}_1 + t\boldsymbol{p}_2, \tag{2.54}$$

which is a mixture of two distributions $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$. Hence, a dual geodesic is a mixture of two probability distributions. We call a dual geodesic an m-geodesic and, by this reasoning, $\boldsymbol{\eta}$ is called the m-affine parameter, where m stands for "mixture". A submanifold which is defined by linear constraints in $\boldsymbol{\eta}$ is said to be m-flat. The linear mixture

$$(1-t)p\left(\boldsymbol{x}, \boldsymbol{\eta}_1\right) + tp\left(\boldsymbol{x}, \boldsymbol{\eta}_2\right) \tag{2.55}$$

is not included in M in general, but $p\left(\boldsymbol{x}, (1-t)\boldsymbol{\eta}_1 + t\boldsymbol{\eta}_2\right)$ is in M, where we used the abuse of notation $p(\boldsymbol{x}, \boldsymbol{\eta})$ to specify the distribution of M of which dual coordinates are $\boldsymbol{\eta}$.

Remark An m-geodesic (2.52) is not a linear mixture of two distributions specified by $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ in the case of a general exponential family. However, we use the term m-geodesic even in this case.

2.5 On Infinite-Dimensional Manifold of Probability Distributions

We have shown that S_n of discrete probability distributions is an exponential family and a mixture family at the same time. It is a super-manifold, in which any statistical model of a discrete random variable is embedded as a submanifold. When x is a continuous random variable, we are apt to consider the geometry of the manifold F of all probability density functions $p(x)$ in a similar way. It is a super-manifold including all statistical models of a continuous random variable. It is considered to be an exponential family and a mixture family at the same time. However, the problem is not mathematically easy, since it is a function space of infinite dimensions. We show a naive idea of studying the geometry of F. This is not mathematically justified, although it works well in most cases, except for "pathological" situations.

Let $p(x)$ be a probability density function of real random variable $x \in \boldsymbol{R}$, which is mutually absolutely continuous with respect to the Lebesgue measure.[1] We put

$$F = \left\{ p(x) \,\middle|\, p(x) > 0, \ \int p(x)dx = 1 \right\}. \tag{2.56}$$

Then, F is a function space consisting of L_1 functions. For two distributions $p_1(x)$ and $p_2(x)$, the exponential family connecting them is written as

$$p_{\exp}(x, t) = \exp\{(1-t)\log p_1(x) + t \log p_2(x) - \psi(t)\}, \tag{2.57}$$

provided it exists in F. Also the mixture family connecting them

$$p_{\text{mix}}(x, t) = (1-t)p_1(x) + tp_2(x) \tag{2.58}$$

is assumed to belong to F. Then, F is regarded as an exponential and a mixture family at the same time as S_n is. Mathematically, there is a delicate problem concerning the topology of F. The L_1-topology and L_2-topology of the function space F are different. Also the topology induced by $p(x)$ is different from that induced by $\log p(x)$.

Disregarding such mathematical problems, we discretize the real line $\boldsymbol{R}$ into $n+1$ intervals, $I_0, I_1, \ldots, I_n$. Then, the discretized version of $p(x)$ is given by the discrete probability distribution $\boldsymbol{p} = (p_0, p_1, \ldots, p_n)$,

$$p_i = \int_{I_i} p(x)dx, \quad i = 0, 1, \ldots, n. \tag{2.59}$$

[1] It would be better to use density function $p(x)$ with respect to the Gaussian measure

$$d\mu(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx,$$

rather than the Lebsesque measure dx.

This gives a mapping from F to S_n, which approximates $p(x)$ by $\boldsymbol{p} \in S_n$. When the discretization is done in such a way that p_i in each interval converges to 0 as n tends to infinity, the approximation looks fine. Then, the geometry of F would be defined by the limit of S_n consisting of discretized $\boldsymbol{p}$. However, we have difficulty in this approach. The limit $n \to \infty$ of the geometry of S_n might not be unique, depending on the method of discretization. Moreover, an admissible discretization would be different for different $p(\boldsymbol{x})$.

Forgetting about the difficulty, by using the delta function $\delta(x)$, let us introduce a family of random variables $\delta(s - x)$ indexed by a real parameter s, which plays the role of index i in $\delta_i(x)$ of S_n. Then, we have

$$p(x) = \int p(s)\delta(x - s)ds, \tag{2.60}$$

which shows that F is a mixture family generated by the delta distributions $\delta(s - x)$, $s \in \boldsymbol{R}$. Here, $p(s)$ are mixing coefficients. Similarly, we have

$$p(x) = \exp\left\{\int \theta(s)\delta(s - x)dx - \psi\right\}, \tag{2.61}$$

where

$$\theta(s) = \log p(s) + \psi \tag{2.62}$$

and ψ is a functional of $\theta(s)$ formally given by

$$\psi[\theta(s)] = \log\left\{\int \exp\{\theta(s)\}\, ds\right\}. \tag{2.63}$$

Hence, F is an exponential family where $\theta(s) = \log p(s) + \psi$ is the $\boldsymbol{\theta}$ affine coordinates and $\eta(s) = p(s)$ is the dual affine coordinates $\boldsymbol{\eta}$. The dual convex function is

$$\varphi[\eta(s)] = \int \eta(s) \log \eta(s)ds. \tag{2.64}$$

Indeed the dual coordinates are given by

$$\eta(s) = \mathbf{E}_p[\delta(s - x)] = p(s) \tag{2.65}$$

and we have

$$\eta(s) = \nabla\psi[\theta(s)], \tag{2.66}$$

where ∇ is the Fréchet-derivative with respect to function $\theta(s)$. The e-geodesic connecting $p(x)$ and $q(x)$ is (2.57) and the m-geodesic (2.58). The tangent vector of an e-geodesic is

$$\frac{d}{dt}\log p(x,t) = \dot{l}(x,t) = \log q(x) - \log p(x) \tag{2.67}$$

in the e-coordinates, and that of an m-geodesic is

$$\dot{p}(x,t) = q(x) - p(x) \tag{2.68}$$

in the m-coordinates.

The KL-divergence is

$$D_{KL}\left[p(x) : q(x)\right] = \int p(x)\log\left\{\frac{p(x)}{q(x)}\right\}dx, \tag{2.69}$$

which is the Bregman divergence derived from $\psi[\theta]$ and it gives F a dually flat structure. The Pythagorean theorem is written, for three distributions $p(x)$, $q(x)$ and $r(x)$, as

$$D_{KL}\left[p(x) : r(x)\right] = D_{KL}\left[p(x) : q(x)\right] + D_{KL}\left[q(x) : r(x)\right], \tag{2.70}$$

when the mixture geodesic connecting p and q is orthogonal to the exponential-geodesic connecting q and r, that is, when

$$\int \{p(x) - q(x)\}\{\log r(x) - \log q(x)\}\,dx = 0. \tag{2.71}$$

It is easy to prove this directly. The projection theorem follows similarly.

The KL-divergence between two nearby distributions $p(x)$ and $p(x) + \delta p(x)$ is expanded as

$$\begin{aligned} D_{KL}\left[p(x) : p(x) + \delta p(x)\right] &= \int p(x)\log\left\{1 - \frac{\delta p(x)}{p(x)}\right\}dx \\ &= \frac{1}{2}\int \frac{\{\delta p(x)\}^2}{p(x)}dx. \end{aligned} \tag{2.72}$$

Hence, the squared distance of an infinitesimal deviation $\delta p(x)$ is

$$ds^2 = \int \frac{\{\delta p(x)\}^2}{p(x)}dx, \tag{2.73}$$

which defines the Riemannian metric given by the Fisher information.

Indeed, the Riemannian metric in $\boldsymbol{\theta}$-coordinates are given by

$$g(s,t) = \nabla\nabla\psi = p(s)\delta(s-t) \tag{2.74}$$

and its inverse is

$$g^{-1}(s,t) = \frac{1}{p(s)}\delta(s-t) \tag{2.75}$$

in $\boldsymbol{\eta}$-coordinates.

It appears that most of the results we have studied in S_n hold well even in the function space F with naive treatment. They are practically useful even though no mathematical justification is given. Unfortunately, we are not free from mathematical difficulties. We show some examples.

The pathological nature in the continuous case has long been known. The following fact was pointed out by Csiszár (1967). We define a quasi-ε-neighborhood of $p(x)$ based on the KL-divergence,

$$N_\varepsilon = \{q(x) \,|\, D_{KL}\,[p(x) : q(x)] < \varepsilon\}\,. \tag{2.76}$$

However, the set of the quasi-ε-neighborhoods does not satisfy the axiom of the topological subbase. Hence, we cannot use the KL-divergence to define the topology. More simply, it is demonstrated that the entropy functional

$$\varphi[p(x)] = \int p(x)\log p(x)dx \tag{2.77}$$

is not continuous in F, whereas it is continuous and differentiable in S_n (Ho and Yeung 2009).

G. Pistone and his co-workers studied the geometrical properties of F based on the theory of Orlicz space, where F is not a Hilbert space but a Banach space. See Pistone and Sempi (1995), Gibilisco and Pistone (1998), Pistone and Rogathin (1999), Cena and Pistone (2007). This was further developed by Grasselli (2010). See recent works by Pistone (2013) and Newton (2012), where trials for mathematical justification using innocent ideas have been developed.

2.6 Kernel Exponential Family

Fukumizu (2009) proposed a kernel exponential family, which is a model of probability distributions of function degrees of freedom. Let $k(x, y)$ be a kernel function satisfying positivity,

$$\int k(x,y)f(x)f(y)dxdy > 0 \tag{2.78}$$

for any $f(x)$ not equal to 0. A typical example is the Gaussian kernel

$$k_\sigma(x,y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-y)^2\right\}, \tag{2.79}$$

where σ is a free parameter.

A kernel exponential family is defined by

$$p(x,\theta) = \exp\left\{\int \theta(y)k(x,y)dx - \psi[\theta]\right\} \tag{2.80}$$

with respect to suitable measure $d\mu(x)$, e.g.,

$$d\mu(x) = \exp\left\{-\frac{x^2}{2\tau^2}\right\} dx. \tag{2.81}$$

The natural or canonical parameter is a function $\theta(y)$ indexed by y instead of θ^i and the dual parameter is

$$\eta(y) = \mathrm{E}[k(x,y)], \tag{2.82}$$

where expectation is taken by using $p(x,\theta)$. $\psi[\theta]$ is a convex functional of $\theta(y)$. This exponential family does not cover all $p(x)$ of probability density functions. So there are many such models, depending on $k(x,y)$ and $d\mu(x)$. The naive treatment in Sect. 2.5 may be regarded as the special case where the kernel $k(x,y)$ is put equal to the delta function $\delta(x-y)$.

2.7 Bregman Divergence and Exponential Family

An exponential family induces a Bregman divergence $D_\psi\left[\boldsymbol{\theta} : \boldsymbol{\theta}'\right]$ given in (2.16). Conversely, when a Bregman divergence $D_\psi\left[\boldsymbol{\theta} : \boldsymbol{\theta}'\right]$ is given, is it possible to find a corresponding exponential family $p(\boldsymbol{x},\boldsymbol{\theta})$? The problem is solved positively by Banerjee et al. (2005). Consider a random variable $\boldsymbol{x}$. It specifies a point $\boldsymbol{\eta}' = \boldsymbol{x}$ in the $\boldsymbol{\eta}$-coordinates of a dually flat manifold given by ψ. Let $\boldsymbol{\theta}'$ be its $\boldsymbol{\theta}$-coordinates. The ψ-divergence from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$, the latter of which is the $\boldsymbol{\theta}$-coordinates of $\boldsymbol{\eta}' = \boldsymbol{x}$, is written as

$$D_\psi\left[\boldsymbol{\theta} : \boldsymbol{\theta}'(\boldsymbol{x})\right] = \psi(\boldsymbol{\theta}) + \varphi(\boldsymbol{x}) - \boldsymbol{\theta}\cdot\boldsymbol{x}. \tag{2.83}$$

Using this, we define a probability density function written in terms of the divergence as

$$p(\boldsymbol{x},\boldsymbol{\theta}) = \exp\left\{-D_\psi\left[\boldsymbol{\theta} : \boldsymbol{\theta}'\right] + \varphi(\boldsymbol{x})\right\} = \exp\left\{\boldsymbol{\theta}\cdot\boldsymbol{x} - \psi(\boldsymbol{\theta})\right\}, \tag{2.84}$$

where $\boldsymbol{\theta}'$ is determined from $\boldsymbol{x}$ as the $\boldsymbol{\theta}$-coordinates of $\boldsymbol{\eta}' = \boldsymbol{x}$. Thus, we have an exponential family derived from D_ψ.

The problem is restated as follows: Given a convex function $\psi(\boldsymbol{\theta})$, find a measure $d\mu(\boldsymbol{x})$ such that (2.8), or equivalently

$$\exp\left\{\psi(\boldsymbol{\theta})\right\} = \int \exp\left\{\boldsymbol{\theta}\cdot\boldsymbol{x}\right\} d\mu(\boldsymbol{x}), \tag{2.85}$$

is satisfied. This is the inverse of the Laplace transform. A mathematical theory concerning the one-to-one correspondence between (regular) exponential families and (regular) Bregman divergences is established in Banerjee et al. (2005).

Theorem 2.2 *There is a bijection between regular exponential families and regular Bregman divergences.*

The theorem shows that a Bregman divergence has a probabilistic expression given by an exponential family of probability distributions. A Bregman divergence is always written in the form of the KL-divergence of the corresponding exponential family.

Remark A mixture family $M = \{p(x, \boldsymbol{\eta})\}$ has a dually flat structure, where the negative entropy $\varphi(\boldsymbol{\eta})$ is a convex function. We can define an exponential family of which the convex function is $\varphi(\boldsymbol{\theta})$. However, this is different from the original M. Hence, Theorem 2.2 does not imply that a mixture family is an exponential family, even though it is dually flat.

2.8 Applications of Pythagorean Theorem

A few applications of the generalized Pythagorean Theorem are shown here to illustrate its usefulness.

2.8.1 Maximum Entropy Principle

Let us consider discrete probability distributions $S_n = \{p(x)\}$, although the following arguments hold even when x is a continuous vector random variable. Let $c_1(x), \ldots, c_k(x)$ be k random variables, that is, k functions of x. Their expectations are

$$\mathrm{E}\,[c_i(x)] = \sum p(x)c_i(x), \quad i = 1, 2, \ldots, k. \tag{2.86}$$

We consider a probability distribution $p(x)$ for which the expectations of $c_i(x)$ take prescribed values $\boldsymbol{a} = (a_1, \ldots, a_k)$,

$$\mathrm{E}\,[c_i(x)] = a_i, \quad i = 1, 2, \ldots, k. \tag{2.87}$$

There are many such distributions and they form an $(n-k)$-dimensional submanifold $M_{n-k}(\boldsymbol{a}) \subset S_n$ specified by $\boldsymbol{a}$, because k restrictions given by (2.87) are imposed. This M_{n-k} is m-flat, because any mixtures of distributions in M_{n-k} belong to the same M_{n-k}.

When one needs to choose a distribution from $M_{n-k}(\boldsymbol{a})$, if there are no other considerations, one would choose the distribution that maximizes the entropy. This is called the maximum entropy principle.

Let P_0 be the uniform distribution that maximizes the entropy in S_n. The dual divergence between $P \in S_n$ and P_0 is written as

$$D_\psi [P_0 : P] = \psi(\boldsymbol{\theta}_0) + \varphi(\boldsymbol{\eta}) - \boldsymbol{\theta}_0 \cdot \boldsymbol{\eta}, \tag{2.88}$$

where the e-coordinates of P_0 are given by $\boldsymbol{\theta}_0$, $\boldsymbol{\eta}$ is the m-coordinates of P and $\varphi(\boldsymbol{\eta})$ is the negative entropy. This is the KL-divergence $D_{KL}[P : P_0]$ from P to P_0. Since P_0 is the uniform distribution, $\boldsymbol{\theta}_0 = 0$. Hence, maximizing the entropy $\varphi(\boldsymbol{\eta})$ is equivalent to minimizing the divergence. Let $\hat{P} \in M_{n-k}$ be the point that maximizes the entropy. Then, triangle $P\hat{P}P_0$ is orthogonal and the Pythagorean relation

$$D_{KL}[P : P_0] = D_{KL}\left[P : \hat{P}\right] + D_{KL}\left[\hat{P} : P_0\right] \tag{2.89}$$

holds (Fig. 2.1). This implies that the entropy maximizer $\hat{P}$ is given by the e-projection of P_0 to $M_{n-k}(\boldsymbol{a})$.

Each $M_{n-k}(\boldsymbol{a})$ includes the entropy maximizer $\hat{P}(\boldsymbol{a})$. By changing $\boldsymbol{a}$, all of these $\hat{P}(\boldsymbol{a})$ form a k-dimensional submanifold E_k which is an exponential family, where the natural coordinates are specified by $\boldsymbol{\theta} = \boldsymbol{a}$ (Fig. 2.1),

$$\hat{p}(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\{\boldsymbol{\theta} \cdot \boldsymbol{c}(x) - \psi(\boldsymbol{\theta})\}. \tag{2.90}$$

It is easy to obtain this result by the variational method that maximizes the entropy $\varphi(\boldsymbol{\eta})$ under constraints (2.87).

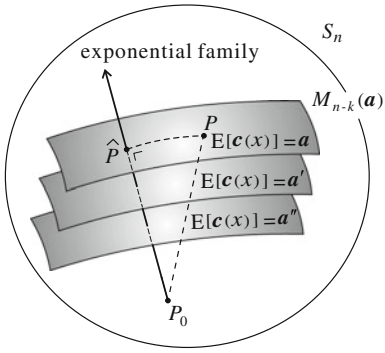

Fig. 2.1 The family maximizing entropy under linear constraints is an exponential family

2.8.2 Mutual Information

Let us consider two random variables x and y and the manifold M consisting of all $p(x, y)$. When x and y are independent, the probability can be written in the product form as

$$p(x, y) = p_X(x)p_Y(y), \tag{2.91}$$

where $p_X(x)$ and $p_Y(y)$ are respective marginal distributions.

Let the family of all the independent distributions be M_I. Since the exponential family connecting two independent distributions is again independent, the e-geodesic connecting them consists of independent distributions. Therefore, M_I is an e-flat submanifold.

Given a non-independent distribution $p(x, y)$, we search for the independent distribution which is closest to $p(x, y)$ in the sense of KL-divergence. This is given by the m-projection of $p(x, y)$ to M_I (Fig. 2.2). The projection is unique and given by the product of the marginal distributions

$$\hat{p}(x, y) = p_X(x)p_Y(y). \tag{2.92}$$

The divergence between $p(x, y)$ and its projection is

$$D_{KL}\left[p(x, y) : \hat{p}(x, y)\right] = \int p(x, y) \log \frac{p(x, y)}{\hat{p}(x, y)} dxdy, \tag{2.93}$$

which is the mutual information of two random variables x and y. Hence, the mutual information is a measure of discrepancy of $p(x, y)$ from independence.

The reverse problem is also interesting. Given an independent distribution (2.92), find the distribution $p(x, y)$ that maximizes $D_{KL}\left[p : \hat{p}\right]$ in the class of distributions having the same marginal distributions as $\hat{p}$. These distributions are the inverse image of the m-projection. This problem is studied by Ay and Knauf (2006) and Rauh (2011). See Ay (2002), Ay et al. (2011) for applications of information geometry to complex systems.

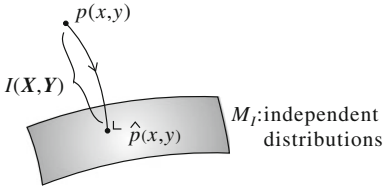

Fig. 2.2 Projection of $p(x, y)$ to the family M_I of independent distributions is the m-projection. The mutual information $I(\boldsymbol{X}, \boldsymbol{Y})$ is the KL-divergence $D_{KL}[p(x, y) : p_X(x)p_Y(y)]$

2.8.3 Repeated Observations and Maximum Likelihood Estimator

Statisticians use a number of independently observed data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ from the same probability distribution $p(\boldsymbol{x}, \boldsymbol{\theta})$ in an exponential family M for estimating $\boldsymbol{\theta}$. The joint probability density of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ is given by

$$p\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \boldsymbol{\theta}\right) = \prod_{i=1}^{N} p\left(\boldsymbol{x}_i, \boldsymbol{\theta}\right) \tag{2.94}$$

having the same parameter $\boldsymbol{\theta}$. We see how the geometry of M changes by multiple observations.

Let the arithmetic average of $\boldsymbol{x}_i$ be

$$\bar{\boldsymbol{x}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i. \tag{2.95}$$

Then, (2.94) is rewritten as

$$p_N\left(\bar{\boldsymbol{x}}, \boldsymbol{\theta}\right) = p\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \boldsymbol{\theta}\right) = \exp\left\{N\boldsymbol{\theta} \cdot \bar{\boldsymbol{x}} - N\psi(\boldsymbol{\theta})\right\}. \tag{2.96}$$

Therefore, the probability density of $\bar{\boldsymbol{x}}$ has the same form as $p(\boldsymbol{x}, \boldsymbol{\theta})$, except that $\boldsymbol{x}$ is replaced by $\bar{\boldsymbol{x}}$ and the term $\boldsymbol{\theta} \cdot \boldsymbol{x} - \psi(\boldsymbol{\theta})$ becomes N times larger.

This implies that the convex function becomes N times larger and hence the KL-divergence and Riemannian metric (Fisher information matrix) also become N times larger. The dual affine structure of M does not change. Hence, we may use the original M and the same coordinates $\boldsymbol{\theta}$ even when multiple observations take place for statistical inference. The binomial distributions and multinomial distributions are exponential families derived from S_2 and S_n by multiple observations.

Let M be an exponential family and consider a statistical model $S = \{p(\boldsymbol{x}, \boldsymbol{u})\}$ included in it as a submanifold, where S is specified by parameter $\boldsymbol{u} = (u_1, \ldots, u_k)$, $k < n$. Since it is included in M, the e-coordinates of $p(\boldsymbol{x}, \boldsymbol{u})$ in M are determined by $\boldsymbol{u}$ in the form of $\boldsymbol{\theta}(\boldsymbol{u})$. Given N independent observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, we estimate the parameter $\boldsymbol{u}$ based on them.

The observed data specifies a distribution in the entire M, such that its m-coordinates are

$$\bar{\boldsymbol{\eta}} = \frac{1}{N} \sum \boldsymbol{x}_i = \bar{\boldsymbol{x}}. \tag{2.97}$$

This is called an observed point. The KL-divergence from the observed $\bar{\boldsymbol{\eta}}$ to a distribution $p(\boldsymbol{x}, \boldsymbol{u})$ in S is written as $D_{KL}\left[\bar{\boldsymbol{\theta}} : \boldsymbol{\theta}(\boldsymbol{u})\right]$, where $\bar{\boldsymbol{\theta}}$ is the $\boldsymbol{\theta}$-coordinates of the observed point $\bar{\boldsymbol{\eta}}$. We consider a simple case of S_n, where the observed point is given by the histogram

$$\bar{p}(x) = \frac{1}{N} \sum_{i=1}^{N} \delta (x - x_i) . \tag{2.98}$$

Then, except for a constant term, minimizing $D_{KL} [\bar{p}(x) : p(x, \boldsymbol{u})]$ is equivalent to maximizing the log-likelihood

$$L = \sum_{i=1}^{N} \log p (x_i, \boldsymbol{u}) . \tag{2.99}$$

Hence, the maximum likelihood estimator that minimizes the divergence is given by the m-projection of $\bar{p}(x)$ to S. See Fig. 2.3. In other words, the maximum likelihood estimator is characterized by the m-projection.

Remarks

An exponential family is an ideal model to study the dually flat structure and also statistical inference. The Legendre duality between the natural and expectation parameter was pointed out by Barndorff-Nielsen (1978). It is good news that the family S_n of discrete distributions is an exponential family, because any statistical model having a discrete random variable is regarded as a submanifold of an exponential family. Therefore, it is wise to study the properties of the exponential family first and then see how they are transferred to curved subfamilies.

Unfortunately, this is not the case with continuous random variable x. There are many statistical models which are not subfamilies of exponential families, even though many are curved-exponential families, that is, submanifolds of exponential families. Again, the study of the exponential family is useful. In the case of a truly non-exponential model, we use its local approximation by using a larger exponential family. This gives an exponential fibre-bundle-like structure to statistical models. This is useful for studying the asymptotic theory of statistical inference. See Amari (1985).

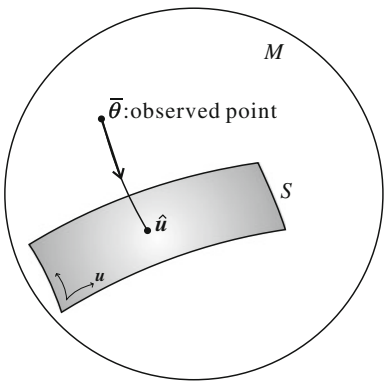

Fig. 2.3 The maximum likelihood estimator is the m-projection of observed point to S

It should be remarked that a generalized linear model provides a dually flat structure, although it is not an exponential family. See Vos (1991). A mixture model also has remarkable characteristics from the point of view of geometry. See Marriott (2002), Critchley et al. (1993).

http://www.springer.com/978-4-431-55977-1

Information Geometry and Its Applications

Amari, S.-i.

2016, XIII, 374 p. 98 illus., Hardcover
ISBN: 978-4-431-55977-1