Statistical exponential families: A digest with flash cards*

Frank Nielsen† and Vincent Garcia‡

November 25, 2009 (v1.0)

Abstract

This document describes concisely the ubiquitous class of exponential family distributions met in statistics. The first part recalls definitions and summarizes main properties and duality with Bregman divergences (all proofs are skipped). The second part lists decompositions and related formula of common exponential family distributions. We recall the Fisher-Rao-Riemannian geometries and the dual affine connection information geometries of statistical manifolds. It is intended to maintain and update this document and catalog by adding new distribution items.

*See the `jMEF` library, a Java package for processing mixture of exponential families. Available for download at `http://www.lix.polytechnique.fr/~nielsen/MEF/`

†École Polytechnique (France) and Sony Computer Science Laboratories Inc. (Japan).

‡École Polytechnique (France).

Part I
A digest of exponential families

1 Essentials of exponential families

1.1 Sufficient statistics

A fundamental problem in statistics is to recover the model parameters λ from a given set of observations $x_1, ...$, etc. Those samples are assumed to be randomly drawn from an independent and identically-distributed random vector with associated density $p(x; \lambda)$. Since the sample set is finite, statisticians estimate a close approximation $\hat{\lambda}$ of the true parameter. However, a surprising fact is that one can collect and concentrate from a random sample all necessary information for recovering/estimating the parameters. The information is collected into a few elementary statistics of the random vector, called the sufficient statistic.[1] Figure 1 illustrates the notions of statistics and sufficiency.

It is challenging to find sufficient statistics for a given parametric probability distribution function $p(x; \lambda)$. The Fisher-Neyman factorization theorem allows one to easily identify those sufficient statistics from the decomposition characteristics of the probability distribution function. A statistic $t(x)$ is sufficient if and only if the density can be decomposed as

$$p(x; \lambda) = a(x) b_\lambda(t(x)), \tag{1}$$

where $a(x) \geq 0$ is a non-negative function independent of the distribution parameters. The class of exponential families encompass most common statistical distributions and are provably the only ones (under mild conditions) that allow one for data reduction.

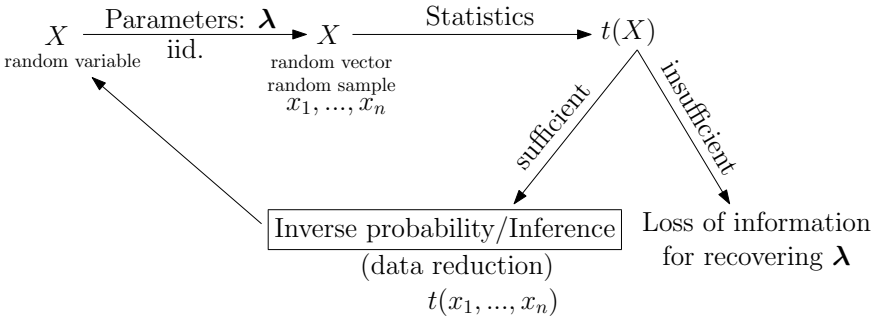

Figure 1: Collecting statistics of a parametric random vector allows one to perform data reduction for inference problem if and only if those statistics are sufficient. Otherwise, loss of information occurs and the parameters of the family of distributions cannot be fully recovered from the (insufficient) statistics.

[1] First coined by statistician Sir Ronald Fisher in 1922.

1.2 Exponential families: Definition and properties

An exponential family is a set of probability distributions admitting the following canonical decomposition:

$$p(x;\theta) = \exp\left(\langle t(x), \theta\rangle - F(\theta) + k(x)\right) \tag{2}$$

where

- $t(x)$ is the sufficient statistic,
- θ are the natural parameters,
- $\langle ., . \rangle$ is the inner product (commonly called dot product),
- $F(\cdot)$ is the log-normalizer,
- $k(x)$ is the carrier measure.

The exponential distribution is said univariate if the dimension of the observation space $\mathcal{X}$ is 1D, otherwise it is said multivariate. The order D of the family is the dimension of the natural parameter space $\mathcal{P}_\Theta$. Part II reports the canonical decompositions of common exponential families. Note that handling probability measures allows one to consider both probability densities and probability mass functions in a common framework. Consider (X, a, μ) a measurable space (with a a σ-algebra) and f a measurable map, the probability measure is defined as $P_\theta(\mathrm{d}x) = p_F(x;\theta)\mu(\mathrm{d}x)$. a is often a σ-algebra on the Borel sets with μ the Lebesgue measure restricted to X.

For example,

- Poisson distributions are univariate exponential distributions of order 1 (e.g., $\dim\mathcal{X} = 1$ and $\dim\mathcal{P} = 1$) with associated probability mass function:

$$\Pr(x = k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{3}$$

  for $k \in \mathbb{N}$.

  The canonical exponential family decomposition yields:

  - $t(x) = x$ is the sufficient statistic,
  - $\theta = \log\lambda$ are the natural parameters,
  - $F(\theta) = \exp\theta$ is the log-normalizer,
  - $k(x) = -\log x!$ is the carrier measure.

- 1D Gaussian distributions are univariate distributions of order 2 (e.g., $\dim\mathcal{X} = 1$ and $\dim\mathcal{P} = 2$), characterized by two parameters (μ, σ) with associated density

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \tag{4}$$

  for $x \in \mathbb{R}$.

  The canonical exponential family decomposition yields:

- $t(x) = (x, x^2)$ is the sufficient statistic,
- $\theta = (\theta_1, \theta_2) = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$ are the natural parameters,
- $F(\theta) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log\left(-\frac{\pi}{\theta_2}\right)$ is the log-normalizer,
- $k(x) = 0$ is the carrier measure.

Exponential families [Bro86] are characterized by their strictly convex and differentiable functions F, called log-normalizer (or cumulant/partition function). The sufficient statistics $t(x) : \mathcal{X} \mapsto \mathcal{P}_\Theta$ is said minimal if they are affinely independent. The carrier measure is usually the Lebesgue (e.g., Gaussian, Rayleigh, etc.) or counting measures (e.g., Poisson, binomial, etc.). Note that

$$F(\theta) = \log \int_x \exp(< t(x), \theta > + k(x))\mathrm{d}x. \tag{5}$$

It is thus easy to build an exponential family: fix $k(x) = 0$ and let us choose for $t(x)$ an arbitrary function for a given domain $x \in [x_{\min}, x_{\max}]$. For example, consider $t(x) = x$ for $x \in [-\infty, 1]$, then $F(\theta) = \int_x \exp \theta x \mathrm{d}x = [\frac{e^{\theta x}}{\theta}]_{x=-\infty}^1 = \frac{1}{\theta}(e^\theta - 1)$.

By remapping $t(x)$ to y, we can consider without loss of generality regular exponential family where the dimension of the observation space matches the parameter space. The regular canonical decomposition of the density simplifies to

$$p(y; \theta) = \exp\left(\langle y, \theta \rangle - F(\theta)\right) \tag{6}$$

with respect to the base measure $h(x) = \exp(k(x))$.
Exponential families include many familiar distributions [Bro86], characterized by their log-normaliser functions:
Gaussian or normal (generic, isotropic Gaussian, diagonal Gaussian, rectified Gaussian or Wald distributions, log-normal), Poisson, Bernoulli, binomial, multinomial (trinomial, Hardy-Weinberg distribution), Laplacian, Gamma (including the chi-squared), Beta, exponential, Wishart, Dirichlet, Rayleigh, probability simplex, negative binomial distribution, Weibull, Fisher-von Mises, Pareto distributions, skew logistic, hyperbolic secant, negative binomial, etc.

However, note that the uniform distribution does not belong to the exponential families.

The observation/sample space $\mathcal{X}$ can be of different types like integer (e.g., Poisson), scalar (e.g., normal), categorical (e.g., multinomial), symmetric positive definite matrix (e.g., Wishart), etc. For the latter case, the inner product of symmetric positive definite matrices is defined as the matrix trace of the product $< X, Y >= \mathrm{Tr}(XY)$, the sum of the eigenvalues of the matrix product $X \times Y$.

The k-order non-centered moment is defined as the expectation $E[X^k]$. The first moment is called the mean μ and the second centered moment $E[(X - \mu)^T(X - \mu)]$ is called the variance-covariance or dispersion matrix.

For exponential families, we have

$$E[X] = \mu = \nabla F(\theta) \tag{7}$$

$$E[(X - \mu)^T(X - \mu)] = \nabla^2 F(\theta) \tag{8}$$

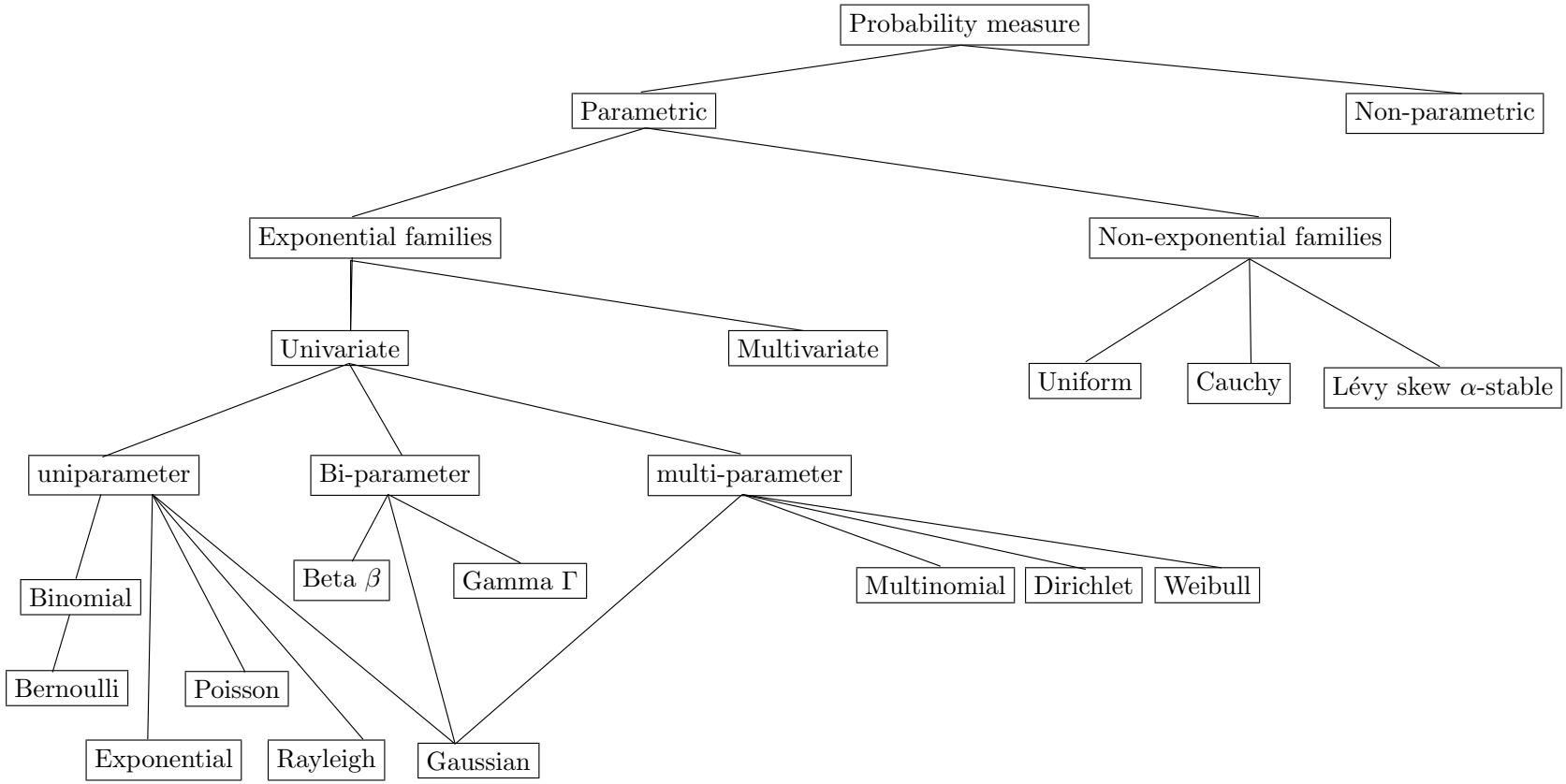

Table 1: Partial taxonomy of statistical distributions.

Notation $\nabla^2 F$ denotes the Hessian of the log-normalizer. It is a positive definite matrix since F is strictly convex and differentiable function. In fact, exponential families have all finite moments, and F is C^∞ differentiable. Thus Cauchy distributions are provably not an exponential family, since it has no defined mean. Another widely used family of distributions that are not exponential families are the Lévy skew α-stable distributions.

In practice, we need to consider minimal exponential family with the sufficient statistics $t(x)$. Since they are several ways to decompose a density/probability mass according to the terms $\theta, t(x), F(\theta)$ and $k(x)$, we adopt the following basic conventions:

The sufficient statistic $t(x)$ should be elementary functions (often polynomial functions) with leading unit coefficient. The carrier measure should not have constant term, so that we prefer to absorb the constant in the log-normalizer. Finally, we denote by Λ the traditional source parameters, and by $\mathcal{P}_\Lambda$ the source parameter space (canonical parameter space).

The natural parameter space

$$\mathcal{P}_\Theta = \{\Theta \mid |F(\Theta)| < +\infty\} \tag{9}$$

is necessarily an open convex set. Although the product of exponential families is an (unnormalized) exponential family, the mixture of exponential families is not an exponential family.

The moment generating function of an exponential family $\{p_F(x;\theta) \mid \theta \in \mathcal{P}_\Theta\}$ is:

$$m_\theta(x) = \exp\left(F(\theta + x) - F(\theta)\right) \tag{10}$$

Function F is thus sometimes called logarithmic moment generating function [iAN00] (p. 69).

The cumulant generating function is defined as

$$\kappa_\theta(x) = \log m_\theta(x) = F(\theta + x) - F(\theta) \tag{11}$$

Exponential families can be generated from Laplace transforms [Ban07]. Let $H(x)$ be a bounded non-negative measure with density $h(x) = \exp k(x)$. Consider the Laplace transform:

$$L(\theta) = \int \exp(< x, \theta >) h(x) \mathrm{d}x \tag{12}$$

Since $L(\theta) > 0$, we can rewrite the former equation to show that

$$p(x;\theta) = \exp\left(< x, \theta > - \log L(\theta)\right) \tag{13}$$

is a probability density with respect to the measure $H(x)$. That is, the log-normalizer of the exponential family is the logarithm of the Laplace transform of the measure $H(x)$.

1.3 Dual parameterizations: Natural and expectation parameters

A fundamental duality of convex analysis is the Legendre-Fenchel transform. Informally, it says that strictly convex and differentiable functions come by pairs.

For the log-normalize F, consider its Legendre dual $G = F^*$ defined by the "slope" transform

$$G(\eta) = \sup_{\theta \in \mathcal{P}_\Theta} < \theta, \eta > -F(\theta). \tag{14}$$

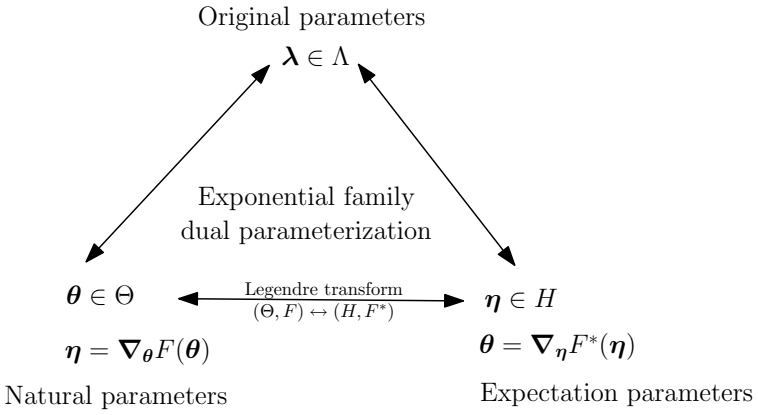

Figure 2: Dual parameterizations of exponential families from Legendre transformation.

The extremum is obtained for $\eta = \nabla F(\theta)$. η is called the moment parameter, since for exponential families we have $\eta = E[X] = \mu = \nabla F(\theta)$. Gradients of conjugate pairs are inversely reciprocal $\nabla F^* = (\nabla F)^{-1}$, and therefore $F^* = \int \nabla F^* = \int (\nabla F)^{-1}$.

Thus to describe a member of an exponential family, we can either use the source or canonical natural/expectation parameters. Figure 2 illustrates the conversions procedures.

1.4 Geometry of exponential families: Riemannian and information geometries

A family of parametric distributions $\{p(x;\theta)\}$ (exponential or not) may be thought as a smooth manifold that can be studied using the framework of differential geometry [Lau87]. We review two main types of geometries: (1) Riemannian geometry defined by a bilinear tensor with an induced Levi-Cevita connection, and non-metric geometry induced by a symmetric affine connection.

Cencov[2] proved [Čen72] (see also [Leb05] and [GS01] for an equivalent in quantum information geometry) that the only Riemannian metric that "makes sense" for statistical manifolds is the Fisher information metric:

$$I(\theta) = \left[\int \frac{\partial \log p(x;\theta)}{\partial \theta_i} \frac{\partial \log p(x;\theta)}{\partial \theta_j} p(x;\theta) \mathrm{d}x \right] = [g_{ij}] \tag{15}$$

The infinitesimal length element is given by

$$\mathrm{d}s^2 = \sum_{i=1}^{d} \sum_{i=1}^{d} \mathrm{d}\theta_i^T \nabla^2 F(\theta) \mathrm{d}\theta_j \tag{16}$$

Cencov proved that for a non-singular transformation of the parameters $\lambda = f(\theta)$, the information matrix

$$I(\lambda) = \left[\frac{\partial \theta_i}{\partial \lambda_j} \right] I(\theta) \left[\frac{\partial \theta_i}{\partial \lambda_j} \right], \tag{17}$$

is such that $\mathrm{d}s^2(\lambda) = \mathrm{d}s^2(\theta)$. Equipped with the tensor $I(\theta)$, the metric distance between two distributions on a statistical manifold can be computed from the geodesic length (e.g., shortest path):

[2] also written as Chentsov

$$D(p(x;\theta_1),p(x;\theta_2)) = \min_{\theta(t) \mid \theta(0)=\theta_1,\theta(1)=\theta_2} \int_0^1 \sqrt{\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^T I(\theta)\frac{\mathrm{d}\theta}{\mathrm{d}t}}\mathrm{d}t \tag{18}$$

Rao's geodesic distance is invariant by non-singular transformations. The multinomial Fisher-Rao-Riemannian geometry yields a spherical geometry, and the normal Fisher-Rao-Riemannian geometry yields a hyperbolic geometry [KV97, CSS05]. Indeed, the Fisher information matrix for univariate normal distributions is

$$I(\theta) = \frac{1}{\sigma^2}\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}. \tag{19}$$

The Fisher information matrix can be interpreted as the Hessian of the Shannon entropy:

$$g_{ij}(\theta) = -E\left[\frac{\partial^2 \log p(x;\theta)}{\partial\theta_i\partial\theta_j}\right] = \frac{\partial^2 H(p)}{\partial\theta_i\partial\theta_j}, \tag{20}$$

with $H(p) = -\int p(x;\theta)\log p(x;\theta)\mathrm{d}x$.

For an exponential family, the Kullback-Leibler divergence is a Bregman divergence on the natural parameters. Using Taylor approximation with exact remainder, we get $\mathrm{KL}(\theta||\theta+\mathrm{d}\theta) = \frac{1}{2}\mathrm{d}\theta^T\nabla^2 F(\theta)\mathrm{d}\theta$. Moreover, the infinitesimal Rao distance is $\sqrt{\mathrm{d}\theta^T I(\theta)\mathrm{d}\theta}$ for $I(\theta) = \nabla^2 F(\theta)$. We deduce that $D(\theta,\theta+\mathrm{d}\theta) = \sqrt{2\mathrm{KL}(\theta||\theta+\mathrm{d}\theta)}$.

The inner product[3] of two vectors x and y at a tangent space of point p is

$$< x, y >_p = x^T g_p y. \tag{21}$$

The length of a vector v in the tangent space at T_p at p is defined by $\|v\|_p = \sqrt{< v, v >_p}$.

For exponential families, the logarithm of the density is concave (since F is convex), and we have

$$I(\theta) = \left[\frac{\partial\eta}{\partial\theta}\right] = \nabla^2 F(\theta) = I^{-1}(\eta) = \left[\frac{\partial\theta}{\partial\eta}\right] \tag{22}$$

That is, the Fisher information is the Hessian of the log-normalizer $\nabla^2 F(\theta)$.

To a given Riemannian manifold $\mathcal{M}$ with tensor G, we may associate[4] a probability measure as follows: Let $p(\theta) = \frac{1}{V}\sqrt{\det G(\theta)}$ with overall volume $V = \int_{\theta\in\Theta}\sqrt{G(\theta)}\mathrm{d}\theta$. These distributions are built from infinitesimal volume element and bear the names of Jeffreys priors. They are commonly used in Bayesian estimation [iAN00, KV97].

Furthermore, in an exponential family manifold, the geometry is flat and θ/η are dual coordinate systems.

Amari [iAN00] focused on a pair of dual affine mixture/exponential connections ∇^m and ∇^e induced by a contrast function F (also called potential function).

A connection ∇ yields a function $\prod_{p,q}$ that maps vectors in any pair of tangent spaces T_p and T_q. An affine connection is defined by d^3 coefficients. Amari's investigated thoroughly the

[3]Technically speaking, it is a bilinear symmetric positive definite operator. A tensor $[T_p]_2^2$ of covariant degree 2 and of contravariant degree 0: $< \cdot,\cdot >_p$: $T_p \times T_p \to \mathbb{R}$.

[4]Or view/interpret the manifold as a statistical manifold.

α-connections and showed that $\nabla^{(0)}$, the metric Levi-Civita connection is obtained from the dual $\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$ α-connections:

$$\nabla^{(0)} = \frac{\nabla^{(\alpha)} + \nabla^{(-\alpha)}}{2}. \tag{23}$$

In particular, the $\nabla^{(1)}$ connection is called the exponential connection and the $\nabla^{(-1)}$ connection is called the mixture connection. The mixture/exponential geodesics for two distributions $p(x)$ and $q(x)$ induced by these dual connections are defined by

$$\gamma_\lambda(x) = (1-\lambda)p(x) + \lambda q(x) \tag{24}$$

$$\log \gamma_\lambda(x) = (1-\lambda)\log p(x) + \lambda \log q(x) - \log Z_\lambda, \tag{25}$$

with Z_λ the normalization coefficient $Z_\lambda = \int f^{(1-\lambda)}(x) g^\lambda(x) \mathrm{d}x$. The exponential connection can be equivalently rewritten as:

$$\gamma_\lambda(x) = \frac{1}{Z_\lambda} p(x)^{(1-\lambda)} + a q(x)^\lambda \tag{26}$$

The dual connections are also called conjugate connections.

The canonical divergence on exponential family distributions of these dually flat statistical manifolds is shown to be:

$$D(p(x;\theta_1)||p(x;\theta_2) = F(\theta_1) + F^*(\theta_2) - < \theta_1, \eta_2 > \tag{27}$$

This canonical divergence can be rewritten as a Bregman divergence on the natural parameter space:

$$D(p(x;\theta_1)||p(x;\theta_2)) = B_F(\theta_1||\theta_2) = F(\theta_1) - F(\theta_2) - < \theta_1 - \theta_2, \nabla F(\theta_2) > \tag{28}$$

Furthermore, we have $B_F(\theta_1||\theta_2) = B_{F^*}(\eta_2||\eta_1)$, where F^* is the Legendre conjugate of F, and $\eta_i = \nabla F(\theta_i)$.

The Kullback-Leibler divergence on two members of the same exponential family is equivalent to the Bregman divergence of the associated log-normalizer on swapped natural parameters:

$$\mathrm{KL}(p(x;\theta_1)||p(x;\theta_2)) = \int p(x;\theta_1) \log \frac{p(x;\theta_1)}{p(x;\theta_2)} \mathrm{d}x = B_F(\theta_2||\theta_1) \tag{29}$$

Note that the Bregman divergence may also be interpreted as a generalized relative entropy. Indeed,

$$KL(p||q) = H^\times(p||q) - H(p), \tag{30}$$

with

$$H(p) = \int p(x) \log \frac{1}{p(x)} \mathrm{d}x \tag{31}$$

the Shannon entropy, and the cross-entropy $H^\times(p||q) \geq H(p)$:

$$H^\times(p||q) = \int p(x) \log \frac{1}{q(x)} \mathrm{d}x \tag{32}$$

Indeed, let $H_F(p) = -F(p)$ denote the generalized entropy, and

$$H_F^{\times}(p||q) = -F(q) - < p-q, \nabla F(q) > \tag{33}$$

the generalized cross-entropy. The Bregman divergence can be considered as a generalized relative entropy:

$$B_F(p||q) = H_F^{\times}(p||q) - H_F(p) \tag{34}$$

Bregman divergences are the canonical divergences of dually flat Riemannian manifolds [iAN00]. Bregman divergences extend naturally the quadratic distances (squared Euclidean and squared Mahalanobis distances), and generalize the notions of orthogonality, projection, and Pythagoras' theorem.

The Bregman projection $q^{\perp}$ of a point q onto a subspace $\mathcal{W}$ is the unique minimizer of $B_F(w||q)$ for $w \in \mathcal{W}$:

$$q^{\perp} = \arg\min B_F(w||q) \tag{35}$$

This is a right-side projection. The left-side projection is simply a right-side projection for the Legendre conjugate generator F^*.

The following remarkable 3-point property hold for any arbitrary Bregman divergence:

$$B_F(p||q) + B_F(q||r) = B_F(p||r) + < p-q, \nabla F(r) - \nabla F(q) > . \tag{36}$$

This formula may be interpreted as the law of generalized cosines.

The proof follows by mathematical rewriting:

$$\begin{aligned} B_F(p||q) + B_F(q||r) &= F(p) - F(q) - < p-q, \nabla F(q) > + F(q) - F(r) - < q-r, \nabla F(r) > \\ &= B_F(p||r) + < p-r, \nabla F(r) > + < r, \nabla F(r) > - < p, \nabla F(q) > + < q, \nabla F(q) - \nabla F(r) > \\ &= B_F(p||r) + < p-q, \nabla F(r) - \nabla F(q) > \end{aligned}$$

Geodesic (pq) is orthogonal to dual geodesic (rq).

Indeed, choosing $r = q^{\perp}$, we end-up with a generalized Pythagoras' theorem:

$$B_F(p||q) + B_F(q||r) = B_F(p||r) \tag{37}$$

That is, triangle p, q, r is a "right-angle" triangle. The ∇-geodesic (pq) is orthogonal to the dual ∇^*-geodesic (qr). The inner product $< p-q, \nabla F(r) - \nabla F(q) >$ vanishes. Note that the notion of orthogonality is not commutative. Euclidean geometry is the special case of self-dual flat spaces with commutative orthogonality obtained for $F(x) = \frac{1}{2}x^T x$.

Banerjee et al. [BMDG05] formally proved the duality between exponential families and Bregman divergences for regular exponential families using Legendre transform:

$$\log p_F(x;\theta) = < t(x), \theta > - F(\theta) + k(x) = -B_{F*}(t(x)||\nabla F(\theta)) + \underbrace{F^*(x) + k(x)}_{=k_F(x)} \tag{38}$$

This duality reveals key for designing an expectation-maximization algorithm using soft Bregman clustering. The proof further reveals that

$$\mathcal{X}_F \subseteq \mathrm{dom}F^*. \tag{39}$$

That is, the space of observations for the exponential families with log normalizer F is included in the domain of the Legendre conjugate function.

For two very close points p and $q \to p$, the Kullback-Leibler divergence is related to the Fisher metric by

$$\mathrm{KL}(p||q) \simeq \frac{1}{2} I^2(p,q) \tag{40}$$

For exponential families, we can thus easily recover the Fisher information metric from the corresponding derivatives of the Bregman divergence.

However, it is difficult to compute Rao's distance $\int \sqrt{\mathrm{d}\theta^T \nabla^2 F(\theta) \mathrm{d}\theta}$ since it requires to compute the anti-derivative of $\sqrt{\nabla^2 F(\theta)}$. See for example, the work [RO03] that computes numerically an approximation of the Rao's distance for gamma distributions.

1.5 Statistical inference

1.5.1 Maximum likelihood estimator

Given n i.i.d. observations $x_1, ..., x_n$ sampled from a given exponential family $p_F(x;\theta)$, the maximum likelihood estimator recover the parameter of the distribution by maximizing

$$\hat{\theta} = \mathrm{argmax}_\theta \prod_{i=1}^n p_F(x;\theta). \tag{41}$$

It follows that the maximum likelihood estimator can be obtained from the center of mass of the sufficient statistics (the observed point):

$$\hat{\theta} = \nabla F^* \left(\frac{1}{n} \sum_{i=1}^n t(x_i) \right) \tag{42}$$

The variance of any unbiased estimator cannot beat the Cramér-Rao bound:

$$\mathrm{var}(\hat{\theta}) \geq I^{-1}(\theta) \tag{43}$$

1.5.2 Bayesian inference and conjugate priors

The family of conjugate priors of an exponential family with likelihood function:

$$L(\theta; x_1, ..., x_n) = \exp\left(< \theta, \sum_{i=1}^n t(x_i) > -nF(\theta) + \sum_{i=1}^n k(x_i) \right) \tag{44}$$

is

$$\exp\left(< \theta, g > -vF(\theta) \right). \tag{45}$$

1.5.3 Mixture of exponential families

Consider a mixture of exponential families of k-components

$$p_F(x;\theta_1,...,\theta_k) = \sum_{i=1}^{k} w_i p_F(x;\theta_i), \tag{46}$$

with $\sum_{i=1}^{k} w_i = 1$ and all $w_i \geq 0$. Mixture of exponential families include the Gaussian mixture models (GMMs), mixtures of Gamma distributions, mixture of zero-mean Laplacians, etc. To get a random sample from a mixture, we first draw randomly a number in $[0,1]$ to select the component, and then draw a random sample from the selected exponential family member.

To fit a mixture model to a set of n independently and identically distributed (i.i.d.) observations $x_1,...,x_n$, we use the general expectation-maximization procedure [DLR77].

Initialization. We first compute a Bregman hard clustering on the n observations $x_1,...,x_n$ to get a collection of k clusters. Let n_i denote the number of points in the ith cluster, and let $x_{i(j)}$ denote the points of the cluster for $j = 1,...,n_i$. For each cluster, we initialize the $w_i = \frac{n_i}{n}$ to the proportion of points inside the cluster, and estimate the expectation/natural parameters η_i/θ_i using the observed point. That is, $\eta_i = \frac{1}{n_i}\sum_{j=1}^{n_i} t(x_{i(j)})$ and in the dual coordinate system: $\theta_i = \nabla F^{-1}(\frac{1}{n_i}\sum_{j=1}^{n_i} t(x_{i(j)}))$.

Expectation step.

$$p(i,j) = \frac{w_j \exp(-\mathrm{D_G}(t(x_i)||\eta_j)))\exp(k(x_i))}{\sum_{l=1}^{n} w_l \exp -\mathrm{D_G}(t(x_i)||\eta_l))\exp(k(x_i))} \tag{47}$$

with the Bregman divergence for the Legendre conjugate $G = F^*$ defined as:

$$D_G(p||q) = G(p) - G(q) - \langle p-q, \nabla G(q)\rangle \tag{48}$$

We simplify[5] the terms $G(t(x_i))$ in the numerator/denominator to get

$$p(i,j) = \frac{w_j \exp(G(\eta_j) + \langle t(x_i)-\eta_j, \nabla G(\eta_j)\rangle)}{\sum_{l=1}^{n} w_l \exp(G(\eta_l) + \langle t(x_i)-\eta_l, \nabla G(\eta_l)\rangle)} \tag{49}$$

Maximization step.

$$w_j = \frac{1}{N}\sum_{i=1}^{N} p(i,j) \tag{50}$$

$$\eta_j = \frac{\sum_{i=1}^{N} p(i,j) t(x_i)}{\sum_{i=1}^{N} p(i,j)} \tag{51}$$

Given two mixtures of exponential families, we can bound the relative entropy of these distributions using Jensen's inequality on the convex Kullback-Leibler divergence as follows:

[5] This step is crucial since $G(t(x))$ may not be defined. Consider for example, the univariate Gaussian distribution. We have $G(\eta) = -\frac{1}{2}\log(\eta_1^2 - \eta_2)$ with $t_1(x) = x$ and $t_2(x) = x^2$. Thus $G(t(x)) = \log(x^2 - x^2)$ is not defined.

$$\mathrm{KL}(\sum_{i=1}^{k} w_i p_F(x;\theta_i)||\sum_{i=1}^{k'} w'_i p_F(x;\theta'_i)) \leq \sum_{i=1}^{k}\sum_{j=1}^{k'} w_i w'_j \mathrm{KL}(p_F(x;\theta_i)||p_F(x;\theta'_j) = \sum_{i=1}^{k}\sum_{j=1}^{k'} w_i w'_j B_F(\theta'_j||\theta_i) \tag{52}$$

This bound is far too crude to be useful in practice.

We may consider approximating the relative entropy by matching components of the mixture, and get the following approximation:

$$\mathrm{KL}(f||g) = \sum_{i=1}^{k} w_i \min_j B_F(\theta'_j||\theta_i) + \log \frac{w_i}{w'_i} \tag{53}$$

In practice, the unscented transform yields a better approximation to the Kullback-Leibler divergence. We consider $2dk$ sigma points as follows: For each component of the first mixture f, we decompose the Hessian $\Sigma_\theta = \nabla^2 F(\theta)$ into $2d$ points $\sqrt{(d\Sigma_\theta)_k}$, such that $\sqrt{(d\Sigma_\theta)_k}$ denote k-th column of the matrix square root of Σ_θ. Then the Kullback-Leibler divergence is approximated at the $2dk$ sigma points by

$$\frac{1}{2d}\sum_{i=1}^{k} w_i \sum_{j=1}^{2d} \log g(x_{i,j}), \tag{54}$$

where g is the probability density of the second mixture.

Even if it is widely known that GMMs can approximate arbitrarily finely any smooth probability density function, it is in fact possible to model any smooth density with a single member of an exponential family by defining the scalar product $< \cdot, \cdot >_{\mathcal{H}}$ in a reproducing kernel Hilbert space [CS06] (RKHS) $\mathcal{H}$. Thus exponential families in RKHSs are universal density estimators [ASH04].

Software library

The jMEF is a Java library implementing the hard/soft/hierarchical techniques for exponential families with respect to sided and symmetrized Bregman divergences [NN09]:

`http://www.lix.polytechnique.fr/~nielsen/MEF/`

Part II

Exponential families: Flash cards

2 Univariate Gaussian distribution

PDF expression	$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ for $x \in \mathbb{R}$
Kullback-Leibler divergence	$D_{\mathrm{KL}}(f_P \| f_Q) = \frac{1}{2}\left(2\log\frac{\sigma_Q}{\sigma_P} + \frac{\sigma_P^2}{\sigma_Q^2} + \frac{(\mu_Q-\mu_P)^2}{\sigma_Q^2} - 1\right)$
MLE	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i \qquad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \hat{\mu})^2}$
Source parameters	$\mathbf{\Lambda} = (\mu, \sigma^2) \ \in \mathbb{R} \times \mathbb{R}^+$
Natural parameters	$\mathbf{\Theta} = (\theta_1, \theta_2) \ \in \mathbb{R} \times \mathbb{R}^-$
Expectation parameters	$\mathbf{H} = (\eta_1, \eta_2) \ \in \mathbb{R} \times \mathbb{R}^+$
$\mathbf{\Lambda} \to \mathbf{\Theta}$	$\mathbf{\Theta} = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$
$\mathbf{\Theta} \to \mathbf{\Lambda}$	$\mathbf{\Lambda} = \left(-\frac{\theta_1}{2\theta_2}, -\frac{1}{2\theta_2}\right)$
$\mathbf{\Lambda} \to \mathbf{H}$	$\mathbf{H} = \left(\mu, \sigma^2 + \mu^2\right)$
$\mathbf{H} \to \mathbf{\Lambda}$	$\mathbf{\Lambda} = \left(\eta_1, \eta_2 - \eta_1^2\right)$
$\mathbf{\Theta} \to \mathbf{H}$	$\mathbf{H} = \nabla F(\mathbf{\Theta})$
$\mathbf{H} \to \mathbf{\Theta}$	$\mathbf{\Theta} = \nabla G(\mathbf{H})$
Log normalizer	$F(\mathbf{\Theta}) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log\left(-\frac{\pi}{\theta_2}\right)$
Gradient log normalizer	$\nabla F(\mathbf{\Theta}) = \left(-\frac{\theta_1}{2\theta_2}, -\frac{1}{2\theta_2} + \frac{\theta_1^2}{4\theta_2^2}\right)$
G	$G(\mathbf{H}) = -\frac{1}{2}\log\left(\eta_1^2 - \eta_2\right) + C$
Gradient G	$\nabla G(\mathbf{H}) = \left(-\frac{\eta_1}{\eta_1^2-\eta_2}, \frac{1}{2(\eta_1^2-\eta_2)}\right)$
Sufficient statistics	$t(x) = (x, x^2)$
Carrier measure	$k(x) = 0$

3 Univariate Gaussian distribution, σ^2 fixed

PDF expression	$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ for $x \in \mathbb{R}$
Kullback-Leibler divergence	$D_{\mathrm{KL}}(f_P \| f_Q) = \frac{(\mu_Q-\mu_P)^2}{2\sigma^2}$
MLE	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Source parameters	$\mathbf{\Lambda} = \mu \ \in \mathbb{R}$
Natural parameters	$\mathbf{\Theta} = \theta \ \in \mathbb{R}$
Expectation parameters	$\mathbf{H} = \eta \ \in \mathbb{R}$
$\mathbf{\Lambda} \rightarrow \mathbf{\Theta}$	$\mathbf{\Theta} = \frac{\mu}{\sigma^2}$
$\mathbf{\Theta} \rightarrow \mathbf{\Lambda}$	$\mathbf{\Lambda} = \theta\sigma^2$
$\mathbf{\Lambda} \rightarrow \mathbf{H}$	$\mathbf{H} = \mu$
$\mathbf{H} \rightarrow \mathbf{\Lambda}$	$\mathbf{\Lambda} = \eta$
$\mathbf{\Theta} \rightarrow \mathbf{H}$	$\mathbf{H} = \nabla F(\mathbf{\Theta})$
$\mathbf{H} \rightarrow \mathbf{\Theta}$	$\mathbf{\Theta} = \nabla G(\mathbf{H})$
Log normalizer	$F(\mathbf{\Theta}) = \frac{\sigma^2\theta^2 + \log(2\pi\sigma^2)}{2}$
Gradient log normalizer	$\nabla F(\mathbf{\Theta}) = \sigma^2\theta$
G	$G(\mathbf{H}) = \frac{\eta^2}{2\sigma^2} + C$
Gradient G	$\nabla G(\mathbf{H}) = \frac{\eta}{\sigma^2}$
Sufficient statistics	$t(x) = x$
Carrier measure	$k(x) = -\frac{x^2}{2\sigma^2}$

4 Multivariate Gaussian distribution

PDF expression	$f(x;\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}\lvert\Sigma\rvert^{1/2}} \exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right)$ for $x \in \mathbb{R}^d$
Kullback-Leibler divergence	$D_{\mathrm{KL}}(f_P \Vert f_Q) = \frac{1}{2}\left(\log\left(\frac{\det\Sigma_Q}{\det\Sigma_P}\right) + \mathrm{tr}\left(\Sigma_Q^{-1}\Sigma_P\right)\right)$ $+\frac{1}{2}\left((\mu_Q-\mu_P)^\top\Sigma_Q^{-1}(\mu_Q-\mu_P) - d\right)$
MLE	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i \qquad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (x_i-\hat{\mu})(x_i-\hat{\mu})^T$
Source parameters	$\mathbf{\Lambda} = (\mu, \Sigma)$ with $\mu \in \mathbb{R}^d$ and $\Sigma \succ 0$
Natural parameters	$\mathbf{\Theta} = (\theta, \Theta)$
Expectation parameters	$\mathbf{H} = (\eta, H)$
$\mathbf{\Lambda} \to \mathbf{\Theta}$	$\mathbf{\Theta} = \left(\Sigma^{-1}\mu, \frac{1}{2}\Sigma^{-1}\right)$
$\mathbf{\Theta} \to \mathbf{\Lambda}$	$\mathbf{\Lambda} = \left(\frac{1}{2}\Theta^{-1}\theta, \frac{1}{2}\Theta^{-1}\right)$
$\mathbf{\Lambda} \to \mathbf{H}$	$\mathbf{H} = \left(\mu, -(\Sigma + \mu\mu^T)\right)$
$\mathbf{H} \to \mathbf{\Lambda}$	$\mathbf{\Lambda} = \left(\eta, -(H + \eta\eta^T)\right)$
$\mathbf{\Theta} \to \mathbf{H}$	$\mathbf{H} = \nabla F(\mathbf{\Theta})$
$\mathbf{H} \to \mathbf{\Theta}$	$\mathbf{\Theta} = \nabla G(\mathbf{H})$
Log normalizer	$F(\mathbf{\Theta}) = \frac{1}{4}\mathrm{tr}(\Theta^{-1}\theta\theta^T) - \frac{1}{2}\log\det\Theta + \frac{d}{2}log\pi$
Gradient log normalizer	$\nabla F(\mathbf{\Theta}) = \left(\frac{1}{2}\Theta^{-1}\theta, -\frac{1}{2}\Theta^{-1} - \frac{1}{4}(\Theta^{-1}\theta)(\Theta^{-1}\theta)^T\right)$
G	$G(\mathbf{H}) = -\frac{1}{2}\log\left(1 + \eta^T H^{-1}\eta\right) - \frac{1}{2}\log\det(-H) - \frac{d}{2}\log(2\pi e)$
Gradient G	$\nabla G(\mathbf{H}) = \left(-(H+\eta\eta^T)^{-1}\eta, -\frac{1}{2}(H+\eta\eta^T)^{-1}\right)$
Sufficient statistics	$t(x) = (x, -xx^T)$
Carrier measure	$k(x) = 0$

5 Multivariate isotropic Gaussian distribution

We assume identity variance-covariance matrix $\Sigma = \mathrm{Id}$.

PDF expression	$f(x;\mu) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{(x-\mu)^T(x-\mu)}{2}\right)$ for $x \in \mathbb{R}^d$
Kullback-Leibler divergence	$D_{\mathrm{KL}}(f_P \| f_Q) = \frac{1}{2}(\mu_Q - \mu_P)^\top (\mu_Q - \mu_P)$
MLE	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i$
Source parameters	$\mathbf{\Lambda} = \mu$ with $\mu \in \mathbb{R}^d$
Natural parameters	$\mathbf{\Theta} = \theta$
Expectation parameters	$\mathbf{H} = \eta$
$\mathbf{\Lambda} \to \mathbf{\Theta}$	$\theta = \mu$
$\mathbf{\Theta} \to \mathbf{\Lambda}$	$\mathbf{\Lambda} = \theta$
$\mathbf{\Lambda} \to \mathbf{H}$	$\mathbf{H} = \mu$
$\mathbf{H} \to \mathbf{\Lambda}$	$\mathbf{\Lambda} = \eta$
$\mathbf{\Theta} \to \mathbf{H}$	$\mathbf{H} = \nabla F(\mathbf{\Theta})$
$\mathbf{H} \to \mathbf{\Theta}$	$\mathbf{\Theta} = \nabla G(\mathbf{H})$
Log normalizer	$F(\theta) = \frac{1}{2}\theta^\top\theta + \frac{d}{2}\log 2\pi$
Gradient log normalizer	$\nabla F(\theta) = \theta$
G	$G(\eta) = F(\theta) = \frac{1}{2}\eta^\top\eta + \frac{d}{2}\log 2\pi$
Gradient G	$\nabla G(\eta) = \eta$
Sufficient statistics	$t(x) = x$
Carrier measure	$k(x) = -\frac{1}{2}x^\top x$

6 Poisson distribution

PDF expression	$f(x;\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$ for $x \in \mathbb{N}^+$
Kullback-Leibler divergence	$D_{\mathrm{KL}}(f_P \| f_Q) = \lambda_Q - \lambda_P \left(1 + \log\left(\frac{\lambda_Q}{\lambda_P}\right)\right)$
MLE	$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Source parameters	$\mathbf{\Lambda} = \lambda \ \in \mathbb{R}^+$
Natural parameters	$\mathbf{\Theta} = \theta \ \in \mathbb{R}$
Expectation parameters	$\mathbf{H} = \eta \ \in \mathbb{R}^+$
$\mathbf{\Lambda} \rightarrow \mathbf{\Theta}$	$\mathbf{\Theta} = \log \lambda$
$\mathbf{\Theta} \rightarrow \mathbf{\Lambda}$	$\mathbf{\Lambda} = \exp \theta$
$\mathbf{\Lambda} \rightarrow \mathbf{H}$	$\mathbf{H} = \lambda$
$\mathbf{H} \rightarrow \mathbf{\Lambda}$	$\mathbf{\Lambda} = \eta$
$\mathbf{\Theta} \rightarrow \mathbf{H}$	$\mathbf{H} = \nabla F(\mathbf{\Theta})$
$\mathbf{H} \rightarrow \mathbf{\Theta}$	$\mathbf{\Theta} = \nabla G(\mathbf{H})$
Log normalizer	$F(\mathbf{\Theta}) = \exp \theta$
Gradient log normalizer	$\nabla F(\mathbf{\Theta}) = \exp \theta$
G	$G(\mathbf{H}) = \eta \log \eta - \eta + C$
Gradient G	$\nabla G(\mathbf{H}) = \log \eta$
Sufficient statistics	$t(x) = x$
Carrier measure	$k(x) = -\log(x!)$

7 Centered Laplacian distribution, $\mu = 0$

PDF expression	$f(x;\sigma) = \frac{1}{2\sigma}\exp\left(-\frac{\lvert x\rvert}{\sigma}\right)$ for $x \in \mathbb{R}$
Kullback-Leibler divergence	$D_{\mathrm{KL}}(f_P \Vert f_Q) = \log\left(\frac{\sigma_Q}{\sigma_P}\right) + \frac{\sigma_P - \sigma_Q}{\sigma_Q}$
MLE	$\hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n} \lvert x_i\rvert$
Source parameters	$\mathbf{\Lambda} = \sigma \ \in \mathbb{R}^+$
Natural parameters	$\mathbf{\Theta} = \theta \ \in \mathbb{R}^-$
Expectation parameters	$\mathbf{H} = \eta \ \in \mathbb{R}^+$
$\mathbf{\Lambda} \rightarrow \mathbf{\Theta}$	$\mathbf{\Theta} = -\frac{1}{\sigma}$
$\mathbf{\Theta} \rightarrow \mathbf{\Lambda}$	$\mathbf{\Lambda} = -\frac{1}{\theta}$
$\mathbf{\Lambda} \rightarrow \mathbf{H}$	$\mathbf{H} = \sigma$
$\mathbf{H} \rightarrow \mathbf{\Lambda}$	$\mathbf{\Lambda} = \eta$
$\mathbf{\Theta} \rightarrow \mathbf{H}$	$\mathbf{H} = \nabla F(\mathbf{\Theta})$
$\mathbf{H} \rightarrow \mathbf{\Theta}$	$\mathbf{\Theta} = \nabla G(\mathbf{H})$
Log normalizer	$F(\mathbf{\Theta}) = \log\left(-\frac{2}{\theta}\right)$
Gradient log normalizer	$\nabla F(\mathbf{\Theta}) = -\frac{1}{\theta}$
G	$G(\mathbf{H}) = -\log\eta + C$
Gradient G	$\nabla G(\mathbf{H}) = -\frac{1}{\eta}$
Sufficient statistics	$t(x) = \lvert x\rvert$
Carrier measure	$k(x) = 0$

8 Bernoulli distribution

PDF expression	$f(x;p) = p^x(1-p)^{1-x}$ for $x \in \{0,1\}$
Kullback-Leibler divergence	$D_{\mathrm{KL}}(f_1 \| f_2) = \log\left(\frac{1-p_1}{1-p_2}\right) - p_1 \log\left(\frac{p_2(1-p_1)}{p_1(1-p_2)}\right)$
MLE	$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Source parameters	$\mathbf{\Lambda} = p \in [0,1]$
Natural parameters	$\mathbf{\Theta} = \theta \in \mathbb{R}^+$
Expectation parameters	$\mathbf{H} = \eta \in [0,1]$
$\mathbf{\Lambda} \to \mathbf{\Theta}$	$\mathbf{\Theta} = \log\left(\frac{p}{1-p}\right)$
$\mathbf{\Theta} \to \mathbf{\Lambda}$	$\mathbf{\Lambda} = \frac{\exp\theta}{1+\exp\theta}$
$\mathbf{\Lambda} \to \mathbf{H}$	$\mathbf{H} = p$
$\mathbf{H} \to \mathbf{\Lambda}$	$\mathbf{\Lambda} = \eta$
$\mathbf{\Theta} \to \mathbf{H}$	$\mathbf{H} = \nabla F(\mathbf{\Theta})$
$\mathbf{H} \to \mathbf{\Theta}$	$\mathbf{\Theta} = \nabla G(\mathbf{H})$
Log normalizer	$F(\mathbf{\Theta}) = \log(1+\exp\theta)$
Gradient log normalizer	$\nabla F(\mathbf{\Theta}) = \frac{\exp\theta}{1+\exp\theta}$
G	$G(\mathbf{H}) = \log\left(\frac{\eta}{1-\eta}\right)\eta - \log\left(\frac{1}{1-\eta}\right) + C$
Gradient G	$\nabla G(\mathbf{H}) = \log\left(\frac{\eta}{1-\eta}\right)$
Sufficient statistics	$t(x) = x$
Carrier measure	$k(x) = 0$

9 Binomial distribution, n fixed $\in \mathbb{N}^+$

PDF expression	$f(x;n,p) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$ where $x \in \mathbb{N}^+$
Kullback-Leibler divergence	$D_{\mathrm{KL}}(f_1 \| f_2) = n(1-p_1)\log\left(\frac{1-p_1}{1-p_2}\right) + np_1\log\left(\frac{p_1}{p_2}\right)$
MLE	$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Source parameters	$\mathbf{\Lambda} = p \ \in [0,1]$
Natural parameters	$\mathbf{\Theta} = \theta \ \in \mathbb{R}$
Expectation parameters	$\mathbf{H} = \eta \ \in \mathbb{R}^+$
$\mathbf{\Lambda} \to \mathbf{\Theta}$	$\mathbf{\Theta} = \log\left(\frac{p}{1-p}\right)$
$\mathbf{\Theta} \to \mathbf{\Lambda}$	$\mathbf{\Lambda} = \frac{\exp\theta}{1+\exp\theta}$
$\mathbf{\Lambda} \to \mathbf{H}$	$\mathbf{H} = np$
$\mathbf{H} \to \mathbf{\Lambda}$	$\mathbf{\Lambda} = \frac{\eta}{n}$
$\mathbf{\Theta} \to \mathbf{H}$	$\mathbf{H} = \nabla F(\mathbf{\Theta})$
$\mathbf{H} \to \mathbf{\Theta}$	$\mathbf{\Theta} = \nabla G(\mathbf{H})$
Log normalizer	$F(\mathbf{\Theta}) = n\log(1+\exp\theta) - \log(n!)$
Gradient log normalizer	$\nabla F(\mathbf{\Theta}) = \frac{n\exp\theta}{1+\exp\theta}$
G	$G(\mathbf{H}) = \eta\log\left(\frac{\eta}{n-\eta}\right) - n\log\left(\frac{n}{n-\eta}\right) + C$
Gradient G	$\nabla G(\mathbf{H}) = \log\left(\frac{\eta}{n-\eta}\right)$
Sufficient statistics	$t(x) = x$
Carrier measure	$k(x) = -\log(x!(n-x)!)$

10 Multinomial distribution, n fixed

PDF expression	$f(x_1, \cdots, x_k; p_1, \cdots, p_k, n) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$ for $x_i \in \mathbb{N}^+$
Kullback-Leibler divergence	$D_{\mathrm{KL}}(f_\alpha \| f_\beta) = n p_{\alpha,k} \log \frac{p_{\alpha,k}}{p_{\beta,k}} - n \sum_{i=1}^{k-1} p_{\alpha,i} \log \frac{p_{\beta,i}}{p_{\alpha,i}}$
MLE	$\hat{p}_i = \frac{n_i}{n}$
Source parameters	$\mathbf{\Lambda} = (p_1, \cdots, p_k) \in [0,1]^k$ with $\sum_i p_i = 1$
Natural parameters	$\mathbf{\Theta} = (\theta_1, \cdots, \theta_{k-1}) \in \mathbb{R}^{k-1}$
Expectation parameters	$\mathbf{H} = (\eta_1, \cdots, \eta_{k-1}) \in [0,n]^{k-1}$
$\mathbf{\Lambda} \to \mathbf{\Theta}$	$\mathbf{\Theta} = \left(\log \left(\frac{p_i}{p_k} \right) \right)_i$
$\mathbf{\Theta} \to \mathbf{\Lambda}$	$\mathbf{\Lambda} = \begin{cases} p_i = \frac{\exp \theta_i}{1 + \sum_{j=1}^{k-1} \exp \theta_j} & \text{if } i < k \\ p_k = \frac{1}{1 + \sum_{j=1}^{k-1} \exp \theta_j} \end{cases}$
$\mathbf{\Lambda} \to \mathbf{H}$	$\mathbf{H} = (n p_i)_i$
$\mathbf{H} \to \mathbf{\Lambda}$	$\mathbf{\Lambda} = \begin{cases} p_i = \frac{\eta_i}{n} & \text{if } i < k \\ p_k = \frac{n - \sum_{j=1}^{k-1} \eta_j}{n} \end{cases}$
$\mathbf{\Theta} \to \mathbf{H}$	$\mathbf{H} = \nabla F(\mathbf{\Theta})$
$\mathbf{H} \to \mathbf{\Theta}$	$\mathbf{\Theta} = \nabla G(\mathbf{H})$
Log normalizer	$F(\mathbf{\Theta}) = n \log \left(1 + \sum_{i=1}^{k-1} \exp \theta_i \right) - \log n!$
Gradient log normalizer	$\nabla F(\mathbf{\Theta}) = \left(\frac{n \exp \theta_i}{1 + \sum_{j=1}^{k-1} \exp \theta_j} \right)_i$
G	$G(\mathbf{H}) = \left(\sum_{i=1}^{k-1} \eta_i \log \eta_i \right) + \left(n - \sum_{i=1}^{k-1} \eta_i \right) \log \left(n - \sum_{i=1}^{k-1} \eta_i \right) + C$
Gradient G	$\nabla G(\mathbf{H}) = \left(\log \left(\frac{\eta_i}{n - \sum_{j=1}^{k-1} \eta_j} \right) \right)_i$
Sufficient statistics	$t(x) = (x_1, \cdots, x_{k-1})$
Carrier measure	$k(x) = -\sum_{i=1}^{k} \log x_i!$

11 Rayleigh distribution

PDF expression	$f(x;\sigma^2) = \frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right)$
Kullback-Leibler divergence	$D_{\mathrm{KL}}(f_P \| f_Q) = \log\left(\frac{\sigma_Q^2}{\sigma_P^2}\right) + \frac{\sigma_P^2-\sigma_Q^2}{\sigma_Q^2}$
MLE	$\hat{\sigma} = \sqrt{\frac{1}{2n}\sum_{i=1}^{n} x_i^2}$
Source parameters	$\mathbf{\Lambda} = \sigma^2 \in \mathbb{R}^+$
Natural parameters	$\mathbf{\Theta} = \theta \in \mathbb{R}^-$
Expectation parameters	$\mathbf{H} = \eta \in \mathbb{R}^+$
$\mathbf{\Lambda} \to \mathbf{\Theta}$	$\mathbf{\Theta} = -\frac{1}{2\sigma^2}$
$\mathbf{\Theta} \to \mathbf{\Lambda}$	$\mathbf{\Lambda} = -\frac{1}{2\theta}$
$\mathbf{\Lambda} \to \mathbf{H}$	$\mathbf{H} = 2\sigma^2$
$\mathbf{H} \to \mathbf{\Lambda}$	$\mathbf{\Lambda} = \frac{\eta}{2}$
$\mathbf{\Theta} \to \mathbf{H}$	$\mathbf{H} = \nabla F(\mathbf{\Theta})$
$\mathbf{H} \to \mathbf{\Theta}$	$\mathbf{\Theta} = \nabla G(\mathbf{H})$
Log normalizer	$F(\theta) = -\log(-2\theta)$
Gradient log normalizer	$\nabla F(\theta) = -\frac{1}{\theta}$
G	$G(\eta) = -\log\eta$
Gradient G	$\nabla G(\eta) = -\frac{1}{\eta}$
Sufficient statistics	$t(x) = x^2$
Carrier measure	$k(x) = \log x$

12 Gamma distribution

$\Gamma(k) = (k-1)!$ and $\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$

PDF expression	$f(x;\lambda,k) = x^{k-1}\frac{e^{-\frac{x}{\lambda}}}{\lambda^k\Gamma(k)}$
Kullback-Leibler divergence	$D_{\mathrm{KL}}(f_P\|f_Q) = \log\frac{\Gamma(\lambda_Q)}{\lambda_P} + (k_P-k_Q)(\Psi(k_P)-\log\lambda_P) + k_P\frac{\lambda_Q-\lambda_P}{\lambda_P}$
MLE	$\hat{k}\hat{\lambda} = \sum_{i=1}^n x_i$ $\log\hat{k} - \Psi(k) = \log(\frac{1}{n}\sum_{i=1}^n x_i) - \frac{1}{n}\sum_{i=1}^n \log x_i$
Source parameters	$\mathbf{\Lambda} = (\lambda, k)$
Natural parameters	$\mathbf{\Theta} = (k-1, -\frac{1}{\lambda})$
Expectation parameters	$\mathbf{H} =$
$\mathbf{\Lambda} \to \mathbf{\Theta}$	$\mathbf{\Theta} =$
$\mathbf{\Theta} \to \mathbf{\Lambda}$	$\mathbf{\Lambda} =$
$\mathbf{\Lambda} \to \mathbf{H}$	$\mathbf{H} =$
$\mathbf{H} \to \mathbf{\Lambda}$	$\mathbf{\Lambda} =$
$\mathbf{\Theta} \to \mathbf{H}$	$\mathbf{H} =$
$\mathbf{H} \to \mathbf{\Theta}$	$\mathbf{\Theta} =$
Log normalizer	$F(\mathbf{\Theta}) = \log\Gamma(\theta_1+1) + (\theta_1+1)\log\frac{-1}{\theta_2}$
Gradient log normalizer	$\nabla F(\mathbf{\Theta}) = (\Psi(\theta_1+1) + \log\frac{-1}{\theta_2}, -\frac{\theta_1+1}{\theta_2})$
G	$G(\mathbf{H})$ non-closed form
Gradient G	$\nabla G(\mathbf{H})$ non-closed form
Sufficient statistics	$t(x) = (x, \log x)$
Carrier measure	$k(x) =$

13 Beta distributions

$\Gamma(k) = (k-1)!$, $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and $\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$

PDF expression	$f(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta}x^{\alpha-1}(1-x)^{\beta-1}$
Kullback-Leibler divergence	$D_{\mathrm{KL}}(f_P \Vert f_Q) = \log\frac{B(\alpha_Q,\beta_Q)}{B(\alpha_P,\beta_P)} - (\beta_Q - \beta_P)\Psi(\alpha_P) - (\beta_Q - \beta_P)\Psi(\beta_P) + (\alpha_Q - \alpha_P + \beta_Q - \beta_P)\Psi(\alpha_P + \alpha_Q)$
MLE	$\begin{cases} \frac{\partial B(\alpha,\beta)}{\partial\alpha} = B(\alpha,\beta)\sum_{i=1}^n \log x_i \\ \frac{\partial B(\alpha,\beta)}{\partial\beta} = B(\alpha,\beta)\sum_{i=1}^n \log(1-x_i) \end{cases}$
Source parameters	$\boldsymbol{\Lambda} = (\alpha,\beta)$
Natural parameters	$\boldsymbol{\Theta} =$
Expectation parameters	$\mathbf{H} =$
$\boldsymbol{\Lambda} \rightarrow \boldsymbol{\Theta}$	$\boldsymbol{\Theta} =$
$\boldsymbol{\Theta} \rightarrow \boldsymbol{\Lambda}$	$\boldsymbol{\Lambda} =$
$\boldsymbol{\Lambda} \rightarrow \mathbf{H}$	$\mathbf{H} =$
$\mathbf{H} \rightarrow \boldsymbol{\Lambda}$	$\boldsymbol{\Lambda} =$
$\boldsymbol{\Theta} \rightarrow \mathbf{H}$	$\mathbf{H} =$
$\mathbf{H} \rightarrow \boldsymbol{\Theta}$	$\boldsymbol{\Theta} =$
Log normalizer	$F(\boldsymbol{\Theta}) = \log B(\theta_1+1,\theta_2+1)$
Gradient log normalizer	$\nabla F(\boldsymbol{\Theta}) = (\Psi(\theta_1+1) - \Psi(\theta_1+\theta_2+2), \Psi(\theta_2+1) - \Psi(\theta_1+\theta_2+2))$
G	$G(\mathbf{H})$ non-closed form
Gradient G	$\nabla G(\mathbf{H})$ non-closed form
Sufficient statistics	$t(x) = (\log x, \log(1-x))$
Carrier measure	$k(x) =$

References

[ASH04] Yasemin Altun, Alex J. Smola, and Thomas Hofmann. Exponential families for conditional random fields. In *in Uncertainty in Artificial Intelligence (UAI)*, pages 2–9. AUAI Press, 2004.

[Ban07] Arindam Banerjee. An analysis of logistic models: Exponential family connections and online performance. In *Siam Data Mining*, 2007.

[BMDG05] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *J. Mach. Learn. Res.*, 6:1705–1749, 2005.

[Bro86] L. D. Brown. *Fundamentals of statistical exponential families: with applications in statistical decision theory*. Institute of Mathematical Statistics, Hayworth, CA, USA, 1986. PDF freely available from `http://projecteuclid.org/`.

[Čen72] N. N. Čencov. *Statistical decision rules and optimal inference.* Nauka, Moscow, 1972. In russian, translation in "Translations of Mathematical Monographs", 53.American Mathematical Society, 1982.

[CS06] Stéphane Canu and Alexander J. Smola. Kernel methods and the exponential family. *Neurocomputing*, 69(7-9):714–720, 2006.

[CSS05] S. I. R. Costa, S. A. Santos, and J. E. Strapasson. Fisher information matrix and hyperbolic geometry. pages 28–30, 2005.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[GS01] M. R. Grasselli and R. F. Streater. On the uniqueness of the chentsov metric in quantum information geometry. *QUANTUM PROB*, 4:173, 2001.

[iAN00] Shun ichi Amari and H. Nagaoka. *Methods of Information Geometry*. Oxford University Press, 2000.

[KV97] R. E. Kass and P. W. Vos. *Geometrical Foundations of Asymptotic Inference*. Wiley, 1997.

[Lau87] Stefan L. Lauritzen. Statistical manifolds. In S. I. Amari, O. E. Barndorff-Nielsen, R. E. Kass, S. L. Lauritzen, and C. R. Rao, editors, *Differential Geometry in Statistical Inference*, pages 163–216. Institute of Mathematical Statistics, Hayward, CA, 1987.

[Leb05] Guy Lebanon. Axiomatic geometry of conditional models. *IEEE Transactions on Information Theory*, 51(4):1283–1294, 2005.

[MR93] Michael Murray and John Rice. *Differential Geometry and Statistics*. Number 48 in Monographs on Statistics and Applied Probability. Chapman and Hall, 1st edition, 1993.

[NN09] Frank Nielsen and Richard Nock. Sided and symmetrized Bregman centroids. *IEEE Transactions on Information Theory*, 55(6):2048–2059, June 2009.

[RO03] F. Reverter and J. M. Oller. Computing the Rao distance for gamma distributions. *J. Comput. Appl. Math.*, 157(1):155–167, 2003.