ENTROPIES AND CROSS-ENTROPIES OF EXPONENTIAL FAMILIES

Frank Nielsen

École Polytechnique
Sony Computer Science Laboratories Inc.

Richard Nock

Université des Antilles et de la Guyane
CEREGMIA

ABSTRACT

Statistical modeling of images plays a crucial role in modern image processing tasks like segmentation, object detection and restoration. Although Gaussian distributions are conveniently handled mathematically, the role of many other types of distributions has been revealed and emphasized by natural image statistics. In this paper, we consider a versatile class of distributions called exponential families that encompasses many well-known distributions, such as Gaussian, Poisson, multinomial, Gamma/Beta and Dirichlet distributions, just to name a few. For those families, we derive mathematical expressions for their Shannon entropy and cross-entropy, give a geometric interpretation, and show that they admit closed-form formula up to some entropic normalizing constant depending on the carrier measure but independent of the member of the family. This allows one to design algorithms that can compare *exactly* entropies and cross-entropies of exponential family distributions although some of them have *strictus sensus* no known closed forms (eg., Poisson). We discuss about maximum entropy and touch upon the entropy of mixtures of exponential families for which we provide a relative entropy upper bound.

Index Terms— Entropy, Cross-entropy, Relative entropy, Bregman divergence, Mixtures, Maximum entropy, Legendre transformation.

1. INTRODUCTION

A growing body of work in image processing considers statistical distributions inferred from raw pixel data to perform image tasks such as segmentation [1], object detection, inpainting, deblurring and restoration. The use of statistical mixture models such as Gaussian mixture models (GMMs) to fit pixel data becomes commonplace. Although very convenient GMMs are not always the best mixtures as attested by statistics empirically revealed in natural image collections. Alternative choices to the Gaussian distribution and mixtures thereof are thus gaining more and more attention. For example, Bougila [2] considered Dirichlet distributions to carry out image restoration. A key algorithmic component for manipulating statistical data is to get a notion of distance between distributions. The relative entropy has been consistently used over the years as it is theoretically well sounded. However, computing relative entropy and interpreting this measure in terms of entropy and cross-entropy is often performed case by case, according to the distribution family. See for example [3] that reports entropies and their estimators for a few multivariate distributions. We consider a large class of distributions, called exponential families that include many of the usual distributions (Gaussian, multinomial, Poisson, Beta/Gamma, Dirichlet) and give mathematical expressions of their entropy, cross-entropy and relative entropy. Although the relative entropy is always in closed-form solution for those distributions, it turns out that it may not be the case for the entropy nor the cross-entropy. Nevertheless, we show that this is *not crucial* as there exists for each exponential family an entropic normalization constant that is shared by all members of that class. Thus even if this constant is not in closed-form, it is not necessary to take it into account for comparing entropies and cross-entropies of members of the same class of exponential families.

The paper is organized as follows: Section 2 introduces the concepts of Shannon entropy, cross-entropy and relative entropy as a measure for statistical distributions. Section 3 describes concisely exponential families. We prove in Section 4 that the relative entropies of exponential families amounts to compute a Bregman divergence. Section 6 discusses on the maximum entropy principle and the role of the Legendre transform to solve easily MAXENT problems. It is followed in Section 7 by an analysis of an upper-bound on the relative entropy of mixtures of exponential families. Finally, Section 8 wraps up the paper and hint at further perspectives.

2. ENTROPY, CROSS-ENTROPY, AND RELATIVE ENTROPY

The Shannon *entropy* of a random variable P measures the amount of *uncertainty*, and is defined by means of its underlying distribution $p(x)$ as $H(P) = \int p(x) \log \frac{1}{p(x)} \mathrm{d}x = -\int p(x) \log p(x) \mathrm{d}x = E_P[-\log p(x)]$. Loosely speaking, the entropy characterizes quantitatively the degree of fuzziness of an uncertain variable. The maximum entropy community investigates properties of random variables linked to entropies; For example, a random variable that has maximum entropy follows necessarily the uniform distribution. A random variable with a given prescribed variance of maximum entropy follows a Gaussian distribution, etc.

In coding theory, one seeks for low entropy codes that uses the underlying structure of the language, at its best. Efficient codes stick close to the assumed real-world model, and as such entropy measures the quality of the code compared to the model. Since the true distribution P is unknown to the observer (hidden by nature), we rather define the *cross-entropy* between two random variables as $H^\times(P,Q) = E_P[-\log q(x)] = -\int p(x)\log q(x)\mathrm{d}x$ to measure the efficiency or accuracy of codes. The cross-entropy measures the average number of bits that are wasted by encoding events from an unknown distribution P with a code designed on a model distribution $Q = \tilde{P}$. The Kullback-Leibler divergence or relative entropy[1] between two distributions P and Q with respective densities $p(x)$ and $q(x)$ is defined according to their likelihood ratio $\frac{p(x)}{q(x)}$ by

$$\mathrm{KL}(p(x)||q(x)) = \int_x p(x)\log\frac{p(x)}{q(x)}\mathrm{d}x = E_P\left[\log\frac{p(x)}{q(x)}\right]$$

The Kullback-Leibler can be rewritten as

$$\mathrm{KL}(p(x)||q(x)) = H^\times(p(x)||q(x)) - H(p(x)) \geq 0$$

Namely, the relative entropy is the difference of the cross-entropy (inaccuracy) minus the entropy, with $H(p(x)) = \int_x p(x)\log\frac{1}{p(x)}\mathrm{d}x$ and $H^\times(p(x)||q(x)) = \int_x p(x)\log\frac{1}{q(x)}\mathrm{d}x$.

3. STATISTICAL EXPONENTIAL FAMILIES

In statistics, many common distribution families such as Poisson, Gaussian or binomial/multinomial distributions are class members of a generic super-family called *exponential families*. A random variable $X \sim E_F(\theta)$ is said to belong to the exponential families if and only if it admits the following canonical rewriting of its underlying distribution:

$$p_F(x;\theta) = \exp(\langle t(x),\theta\rangle - F(\theta) + k(x)), \qquad (1)$$

where $\langle x,y\rangle = x^Ty$ denotes the inner product, $t(x)$ the sufficient statistics, θ the natural parameters belonging to an open convex space Θ, $F(\theta)$ a C^∞ differentiable real-valued convex function defined on Θ, and $k(x)$ a carrier measure.

Since $F(\theta) = \log\int_x \exp\{\langle t(x),\theta\rangle + k(x)\}\mathrm{d}x$ (because $\int p_F(x;\theta)\mathrm{d}x = 1$), function F is called the log-normalizer[2] and fully characterizes the family, with the natural parameter θ denoting the member of the family E_F. A statistic is a function of the observations (like the sample mean or sample variance) that collects information about the distribution with the goal to concentrate information for later inference. A statistic is said sufficient if it allows one to concentrate information obtained from random observations *without loosing* information, in a sense that working directly on the observation sets or its compact sufficient statistics yields exactly the same results. It can be shown from the Neyman-Pearson theorem, under mild regularity conditions, that the class of distributions admitting sufficient statistics are the exponential families [4], the log-linear models. $k(x)$ is a term related to the carrier measure (using either Lebesgue or counting measure).

The exponential family is said univariate if $x \in \mathcal{X}$ is unidimensional or multivariate, otherwise. The order of the exponential family is the dimension of the natural parameter space. The family of Gaussian distributions is univariate of order 2.

Dealing with exponential families instead of particular family members (like Gaussians or multinomials) allows one to design *generic* solutions. For example, the maximum likelihood estimator for a i.i.d. sequence $x_1,...,x_n$ is given by $\hat{\theta} = \nabla F^{-1}(\frac{1}{n}\sum_{i=1}^n t(x_i))$, where ∇F^{-1} denote the reciprocal gradient of F: $\nabla F^{-1}\circ\nabla F = \nabla F\circ\nabla F^{-1} = \mathrm{Id}$.

Usually, exponential family parameters are handled not in the natural parameters. For example, the Gaussian distribution is commonly parameterized by $\lambda = (\mu,\sigma^2)$, and not $\theta = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$ (see [4]). However, there is a one-to-one map between these source parameters λ and the natural parameters θ.

4. BREGMAN DIVERGENCES AND RELATIVE ENTROPY OF EXPONENTIAL FAMILIES

A broad class of dissimilarity measures can be defined according to a generator function: Bregman divergences. The Bregman divergence associated to a real-valued strictly convex and differentiable function F is defined by:

$$B_F(p||q) = F(p) - F(q) - \langle p-q, \nabla F(q)\rangle.$$

Choosing $F(x) = \sum_i x_i^2$ yields the squared Euclidean distance. The relative entropy is obtained for $F(x) = \sum_i x_i\log x_i$. Note that Bregman generators F are defined modulo affine terms: $B_F(p||q) = B_G(p||q)$ for $G(x) = F(x) + \langle a,x\rangle + b$.

It can be shown that the relative entropy of two distributions $E_F(\theta)$ and $E_F(\theta')$ belonging to the *same* exponential family E_F is always available in closed-form. Namely, the relative entropy is equal to the Bregman divergence defined by the log-normalizer on the swapped natural parameters: $\mathrm{KL}(E_F(\theta)||E_F(\theta')) = B_F(\theta'||\theta)$.

Let us first prove that $\nabla F(\theta) = E[t(X)]$. That is, that the expectation of the sufficient statistics is equal to the Jacobian of the log-normalizer calculated at the natural parameter. We

[1] Relative entropy bears also frequently the name of information divergence, information deviance and relative entropy, etc.

[2] Or cumulant function, log-partition $Z(\theta)$ in statistical physics with $F(\theta) = \log Z(\theta)$. The log-normalizer is related to the Laplace transform.

have $F(\theta) = \log \int_x \exp\{\langle t(x), \theta\rangle + k(x)\}\mathrm{d}x$. It follows that

$$\nabla F(\theta) = \frac{[\int_x t(x) \exp\{\langle t(x), \theta\rangle + k(x)\}\mathrm{d}x]_j}{\int_x \exp\{\langle t(x), \theta\rangle + k(x)\}\mathrm{d}x}$$

Since $e^{F(\theta)} = \int_x \exp\{\langle t(x), \theta\rangle + k(x)\}\mathrm{d}x$, we get $\nabla F(\theta) = \int_x t(x) \exp\{\langle t(x), \theta\rangle - F(\theta) + k(x)\}\mathrm{d}x$. That is, $\nabla F(\theta) = \int_x t(x) p_F(x;\theta)\mathrm{d}x = E_\theta[t(x)]$. (Thus if x belongs to the sufficient statistics, we easily get closed-form expression for the mean $E[p_F(x;\theta)]$ of the distribution).

Similarly, the variance-covariance matrix of the sufficient statistics is the Hessian of the log-normalizer calculated at the natural parameter: $\mathrm{Cov}[t(X)] = \nabla^2 F(\theta)$, etc. In fact, all moments of exponential families are finite, which explains why the Cauchy distribution (of undefined mean) does not belong to the exponential families. The proof of the Kullback-Leibler/Bregman divergence equivalence for exponential families follows:

$$\mathrm{KL}(\theta_p||\theta_q) = \int_x p(x|\theta_p) \log \frac{p(x|\theta_p)}{p(x|\theta_q)}\mathrm{d}x$$

$\mathrm{KL}(\theta_p||\theta_q) = \int_x p(x|\theta_p)(F(\theta_q) - F(\theta_p) + \langle \theta_p - \theta_q, t(x)\rangle)\mathrm{d}x$ $= \int_x p(x|\theta_p)(B_F(\theta_q||\theta_p) +$ $\langle \theta_q - \theta_p, \nabla F(\theta_p)\rangle + \langle \theta_p - \theta_q, t(x)\rangle)\mathrm{d}x = B_F(\theta_q||\theta_p) + \int_x p(x|\theta_p)\langle \theta_q - \theta_p, \nabla F(\theta_p) - t(x)\rangle)\mathrm{d}x = B_F(\theta_q||\theta_p) - \int_x p(x|\theta_p)\langle \theta_q - \theta_p, t(x)\rangle \mathrm{d}x + \langle \theta_q - \theta_p, \nabla F(\theta_p)\rangle = B_F(\theta_q||\theta_p)$ since $\nabla F(\theta) = \left[\int_x t(x) \exp\{\langle \theta, t(x)\rangle - F(\theta) + k(x)\}\mathrm{d}x\right]$.
Note that finite discrete distributions (say, of d events) are exponential families in disguise: Those distributions are precisely multinomials with $d-1$ degrees of freedom.

5. ENTROPY AND CROSS-ENTROPY OF EXPONENTIAL FAMILIES

Let us write the relative entropy as the difference of the cross-entropy minus the entropy: $\mathrm{KL}(p||q) = H^\times(p||q) - H(p)$. Furthermore, consider the equivalence with Bregman divergences for exponential families. Separating, the terms independent of q from the other ones, we get:

$$\begin{aligned} \mathrm{KL}(p||q) &= B_F(\theta_q||\theta_p) \\ &= F(\theta_q) - F(\theta_p) - \langle \theta_q - \theta_p, \nabla F(\theta_p)\rangle \end{aligned}$$

$$\mathrm{KL}(p||q) = \underbrace{F(\theta_q) - \langle \theta_q, \nabla F(\theta_p)\rangle}_{\sim H_F^\times(\theta_p||\theta_q)} - \underbrace{(F(\theta_p) - \langle \theta_p, \nabla F(\theta_p)\rangle)}_{\sim H_F(\theta_p)}$$

Since F is defined modulo affine terms $ax + b$ in the Bregman divergence, and since the factor a leaves independent the entropy/cross-entropy terms, we deduce that

$$H(P) = H_F(\theta_p) = F(\theta_p) - \langle \theta_p, \nabla F(\theta_p)\rangle + b$$

Since $H(p) \geq 0$, we necessarily have $b \geq \langle \theta_p, \nabla F(\theta_p)\rangle - F(\theta_p)$. Thus we can compare exactly the entropy of two members of the same exponential family since the constant term b vanishes.
To determine explicitly the entropic normalization additive constant b, we proceed as follows:
$H_F(\theta) = -\int p_F(x;\theta) \log p_F(x;\theta)\mathrm{d}x$. This is equal to $-\int p_F(x;\theta)\langle t(x), \theta\rangle \mathrm{d}x - \int k(x) p_F(x;\theta)\mathrm{d}x$.
We get $b = -\int k(x) p_F(x;\theta)\mathrm{d}x = -E[k(x)]$, the mean of the carrier measure for the exponential distribution. For standard carrier measure $k(x) = 0$, we thus get $b = 0$.

Theorem 1 *The entropy of an exponential family $\mathcal{E}_F$ is given by $H(P) = H_F(\theta_p) = F(\theta_p) - \langle \theta_p, \nabla F(\theta_p)\rangle - E_P[k(x)]$. In particular, for standard zero carrier measure, we have $E_P[k(x)] = 0$, and the entropy in closed-form solution: $H(P) = F(\theta_p) - \langle \theta_p, \nabla F(\theta_p)\rangle$.*

Note that Gaussian distributions have zero carrier measure ($k(x) = 0$), and thus their entropy is in closed-form. To give yet another example, let us consider Rayleigh distribution $p(x;\sigma^2) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right)$ that belongs to the exponential families for the log-normalizer $F(\theta) = -\log(-2\theta)$, natural parameter $\theta = -\frac{1}{2\sigma^2}$, sufficient statistic $t(x) = x^2$, gradient $F'(\theta) = -\frac{1}{\theta}$ and carrier measure $k(x) = \log x$. Let $X \sim \mathrm{Rayleigh}(\sigma^2)$, we have: $H(X) = 1 + \ln \frac{\sigma}{\sqrt{2}} + \frac{\gamma}{2}$, where $\gamma = 0.57721566...$ stands for the Euler-Mascheroni constant. This is the term related to the carrier measure $\log x$ integrated over the distribution. Consider yet another univariate exponential family: the Poisson distribution with probability mass function $p(x;\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$. The entropy is $\lambda(1 - \log\lambda) - E[k(x)]$ Since $k(x) = -\log x!$ (see [4]), we have:
$-E[k(x)] = \sum_{k=0}^{\infty} p_F(x;\lambda) \log k! = e^{-\lambda} \sum \frac{\lambda^k \log k!}{k!}$. Although the entropy is not in closed form, we nevertheless end-up with the following closed-form relative entropy: $\mathrm{KL}(P||P') = \lambda'_P - \lambda_P \left(1 + \log\left(\frac{\lambda_{P'}}{\lambda_P}\right)\right)$
Once again, we insist on the fact that even if the entropy itself has non-closed form solution, we can always make exact entropy/cross-entropy comparisons for distributions of the same family members. Figure 1 illustrates graphically the entropy, cross-entropy and relative entropy quantities for exponential families.

6. MAXIMUM ENTROPY

The maximum entropy probability distribution is a probability distribution whose entropy is maximized over all members of a prescribed class of distributions. In physics, the principle of maximum entropy states that the distribution with largest entropy of the class should be chosen as it reflects the limit configuration of the system over time. For exponential families, the maximum entropy optimization problem becomes:

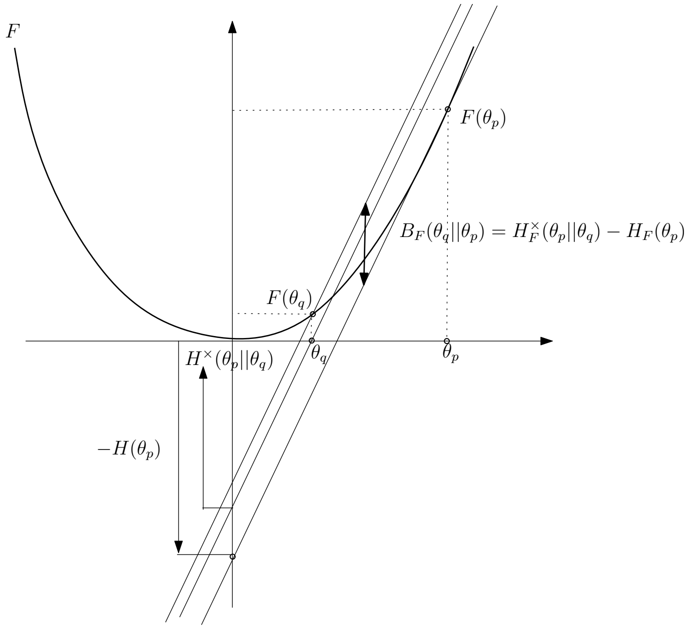
F
$F(\theta_p)$
$B_F(\theta_q||\theta_p) = H_F^\times(\theta_p||\theta_q) - H_F(\theta_p)$
$F(\theta_q)$
$H^\times(\theta_p||\theta_q)$
θ_q
θ_p
$-H(\theta_p)$

Fig. 1. Visualizing the entropy, cross-entropy and relative entropy of exponential families (for $k(x) = 0$).

$\max_{\theta \in \Theta} H_F(\theta)$. That is, choose the parameter θ for the class of exponential family E_F that maximizes its entropy. Consider $\eta_p = \nabla F(\theta_p)$ the *dual* expectation parameter of the exponential family. The natural/expectation parameters are related by the Legendre transformation. The Legendre transformation of a convex function F yields a dual convex function F^* such that $F^*(\eta) = \max_\theta \langle \theta, \eta \rangle - F(\theta)$. Maximum is achieved for $\eta = \nabla F(\theta)$. Legendre convex conjugates satisfy reciprocal gradient constraint: $\nabla F^* = (\nabla F)^{-1}$. Thus the maximum entropy optimization problem becomes $\max_\theta H_F(\theta) = \max_\theta F(\theta) - \langle \theta, \nabla F(\theta) \rangle - E_\theta[k(x)]$. Assuming $k(x) = 0$, we get $\max_\eta -F^*(\eta) \equiv \min_\eta F^*(\eta)$, a minimization problem for the dual Legendre convex conjugate. Provided the Legendre conjugate is in explicit closed form, we retrieve easily the solution for the minimum of the convex function F^*. Otherwise, we apply any numerical root-finding algorithm. (The Legendre dual F^* of the log-normalizer F for the expectation parameter η yields the Shannon negative entropy for the $\theta = \eta^* = \nabla F^{-1}(\eta)$ member: $F^*(\eta) = -H(p_F(x; \theta))$.)

7. MIXTURE MODELS

Consider a mixture of exponential families with k components expressed using the natural parameters: $p_F(x; \theta_1, ..., \theta_k) = \sum_{i=1}^k w_i p_F(x; \theta_i)$, 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. 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:

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

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

8. CONCLUSION

In this paper, we have explained how to derive a generic formula for calculating the entropy, cross-entropy, and relative entropy of statistical exponential families. Interestingly, even if those entropies and cross-entropies do not admit closed-form formula *strictu senso*, we have nevertheless shown how to compare exactly those quantities as the normalization entropic constant of an exponential family does depend only on the expectation of the carrier measure, and is independent of the family member for a given class. For the relative entropy, that normalization constant vanishes, yielding always a closed-form formula. We interpreted graphically those entropy and cross-entropy quantities, and further showed that maximum entropy problems solved well using Legendre transform. Finally, we reported some upper-bound on the relative entropy of mixtures of exponential families. Those results extend to matrix distributions of the exponential families as well (eg., Weibull, Wishart, etc.) We hope that those results will spur further interests in the image processing community to consider generic exponential families instead of the widely acknowledged Gaussian distribution. These results can further be extended to the classes of Rényi, Tsallis and α-entropies following [5].

9. REFERENCES

[1] F. Nielsen, "Visual Computing: Geometry, Graphics and Vision," Charles River Media, ISBN 1584504277, 2005.

[2] N. Bouguila, "Non-Gaussian mixture image models prediction," in *International Conference on Image Processing (ICIP)*, 2008, pp. 2580–2583.

[3] N. A. Ahmed and D. V. Gokhale, "Entropy expressions and their estimators for multivariate distributions," *IEEE Transactions on Information Theory*, vol. 35, no. 3, pp. 688–692, 1989.

[4] F. Nielsen and V. Garcia, "Statistical exponential families: A digest with flash cards," 2009, arXiv, 0911.4863.

[5] F. Nielsen and S. Boltz, "The Burbea-Rao and Bhattacharyya centroids," 2010, arXiv:1004.5049.