On the Chi Square and Higher-Order Chi Distances for Approximating f-Divergences

Frank Nielsen, *Senior Member, IEEE*, and Richard Nock

***Abstract*—We report closed-form formula for calculating the Chi square and higher-order Chi distances between statistical distributions belonging to the same exponential family with affine natural space, and instantiate those formula for the Poisson and isotropic Gaussian families. We then describe an analytic formula for the f-divergences based on Taylor expansions and relying on an extended class of Chi-type distances.**

***Index Terms*—Chi square distance, exponential families, Kullback–Leibler divergence, statistical divergences, Taylor series.**

I. Introduction

A. Statistical Divergences: f-divergences

MEASURING the similarity or *dissimilarity* between two probability measures is met ubiquitously in signal processing. Some usual distances are the Pearson χ^2_P and Neyman chi square distances $\chi^2_N(X_1 : X_2) = \chi^2_P(X_2 : X_1)$, and the Kullback–Leibler divergence [1] defined respectively by:

$$\chi^2_P(X_1 : X_2) = \int \frac{(x_2(x) - x_1(x))^2}{x_1(x)} \mathrm{d}\nu(x), \quad (1)$$

$$\mathrm{KL}(X_1 : X_2) = \int x_1(x) \log \frac{x_1(x)}{x_2(x)} \mathrm{d}\nu(x), \quad (2)$$

where X_1 and X_2 are probability measures absolutely continuous with respect to a reference measure ν, and x_1 and x_2 denote their Radon-Nikodym densities, respectively. Those dissimilarity measures M are termed *divergences* to contrast with metric distances since they are oriented distances (*i.e.*, $M(X_1 : X_2) \neq M(X_2 : X_1)$) that do not satisfy the triangular inequality. In the 1960's, many of those divergences were unified using the generic framework of f-divergences [3], [2], I_f, defined for an arbitrary functional f:

$$I_f(X_1 : X_2) = \int x_1(x) f\left(\frac{x_2(x)}{x_1(x)}\right) \mathrm{d}\nu(x) \geq 0, \quad (3)$$

where f is a convex function $f : (0, \infty) \subseteq \mathrm{dom}(f) \mapsto [0, \infty]$ such that $f(1) = 0$). Indeed, it follows from Jensen inequality that $I_f(X_1 : X_2) \geq f(\int x_2(x)\mathrm{d}\nu(x)) = f(1) = 0$. Furthermore, wlog., we may consider $f'(1) = 0$ and fix the scale of divergence by setting $f''(1) = 1$, see [3]. Those f-divergences[1] can always be symmetrized by taking $S_f(X_1 : X_2) = I_f(X_1 : X_2) + I_{f^*}(X_1 : X_2)$, with $f^*(u) = uf(1/u)$, and $I_{f^*}(X_1 : X_2) = I_f(X_2 : X_1)$. See Table I for a list of common f-divergences with their corresponding generators f. In information theory, f-divergences are characterized as the *unique* family of convex separable [3] divergences that satisfies the *information monotonicity* property [4]. Note that f-divergences may evaluate to infinity (that is, *unbounded* $I_f = +\infty$) when the integral diverge.

B. Stochastic Approximations of f-Divergences

To bypass the integral evaluation of I_f of Eq. (3) (often mathematically intractable), we carry out a *stochastic integration*:

$$\widehat{I}_f(X_1 : X_2) \sim \frac{1}{2n} \sum_{i=1}^{n} \left(f\left(\frac{x_2(s_i)}{x_1(s_i)}\right) + \frac{x_1(t_i)}{x_2(t_i)} f \times \left(\frac{x_2(t_i)}{x_1(t_i)}\right)\right), \quad (4)$$

with $s_1, \ldots, s_n$ and $t_1, \ldots, t_n$ IID. sampled from X_1 and X_2, respectively. Those approximations, although converging to the true values when $n \to \infty$, are time consuming and yield poor results in practice, specially when the dimension of the observation space, $\mathcal{X}$, is large. In practice, f-divergences can be efficiently estimated from random samples emanating from X_1 and X_2 (the datasets) by estimating the density ratio [7] (without estimating the distribution parameters). In this letter, we concentrate on obtaining exact or arbitrarily fine approximation formula for f-divergences by considering a restricted class of exponential families with *given* distribution parameters.

C. Exponential Families

Let $\langle x, y\rangle$ denote the inner product for $x, y \in \mathcal{X}$: The inner product for vector spaces $\mathcal{X}$ is the scalar product $\langle x, y\rangle = x^\top y$. An exponential family [8] is a set of probability measures $\mathcal{E}_F = \{P_\theta\}_\theta$ dominated by a measure ν having their Radon-Nikodym densities p_θ expressed canonically as:

$$p_\theta(x) = \exp(\langle t(x), \theta\rangle - F(\theta) + k(x)), \quad (5)$$

for θ belonging to the *natural parameter space*: $\Theta = \{\theta \in \mathbb{R}^D | \int p_\theta(x)\mathrm{d}\nu(x) = 1\}$. Since $\log \int_{x\in\mathcal{X}} p_\theta(x)\mathrm{d}\nu(x) = \log 1 = 0$, it follows that $F(\theta) = -\log \int \exp(\langle t(x), \theta\rangle + k(x))\mathrm{d}\nu(x)$. For full regular families [8], it can be proved that function F is strictly convex and differentiable over the open

Manuscript received September 12, 2013; revised October 25, 2013; accepted October 25, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joseph Tabrikian.

F. Nielsen is with Sony Computer Science Laboratories, Inc., 141-0022 Shinagawa-ku, Tokyo, Japan (e-mail: nielsen@csl.sony.co.jp).

R. Nock is with UAG CEREGMIA, Martinique, France (e-mail: rnock@martinique.univ-ag.fr).

Digital Object Identifier 10.1109/LSP.2013.2288355

[1]Beware that sometimes the χ^2_N and χ^2_P definitions are inverted in the literature. This may stem from an alternative definition of f-divergences defined as $I'_f(X_1 : X_2) = \int x_2(x) f(\frac{x_1(x)}{x_2(x)})\mathrm{d}\nu(x) = I_f(X_2 : X_1)$.

TABLE I
SOME COMMON f-DIVERGENCES I_f WITH CORRESPONDING GENERATORS: EXCEPT THE TOTAL VARIATION, f-DIVERGENCES ARE NOT METRIC [5]

Name of the f-divergence	Formula $I_f(P:Q)$	Generator $f(u)$ with $f(1)=0$
Total variation (metric)	$\frac{1}{2}\int \lvert p(x)-q(x)\rvert \mathrm{d}\nu(x)$	$\frac{1}{2}\lvert u-1\rvert$
Pearson χ^2_P	$\int \frac{(q(x)-p(x))^2}{p(x)}\mathrm{d}\nu(x)$	$(u-1)^2$
Neyman χ^2_N	$\int \frac{(p(x)-q(x))^2}{q(x)}\mathrm{d}\nu(x)$	$\frac{(1-u)^2}{u}$
Pearson-Vajda χ^k_P	$\int \frac{(q(x)-p(x))^k}{p^{k-1}(x)}\mathrm{d}\nu(x)$	$(u-1)^k$
Pearson-Vajda $\lvert\chi\rvert^k_P$	$\int \frac{\lvert q(x)-p(x)\rvert^k}{p^{k-1}(x)}\mathrm{d}\nu(x)$	$\lvert u-1\rvert^k$
Kullback-Leibler	$\int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}\nu(x)$	$-\log u$
reverse Kullback-Leibler	$\int q(x)\log\frac{q(x)}{p(x)}\mathrm{d}\nu(x)$	$u\log u$
Jensen-Shannon	$\frac{1}{2}\int\left(p(x)\log\frac{2p(x)}{p(x)+q(x)}+q(x)\log\frac{2q(x)}{p(x)+q(x)}\right)\mathrm{d}\nu(x)$	$-(u+1)\log\frac{1+u}{2}+u\log u$

TABLE II
EXAMPLES OF EXPONENTIAL FAMILIES WITH AFFINE NATURAL SPACE Θ. ν_c DENOTES THE COUNTING MEASURE AND ν_L THE LEBESGUE MEASURE

$$\mathrm{Poi}(\lambda) \quad : \quad p(x|\lambda)=\frac{\lambda^x e^{-\lambda}}{x!}, \lambda>0, x\in\{0,1,\ldots\}$$
$$\mathrm{Nor}_I(\mu) \quad : \quad p(x|\mu)=(2\pi)^{-\frac{d}{2}}e^{-\frac{1}{2}(x-\mu)^\top(x-\mu)}, \quad \mu\in\mathbb{R}^d, x\in\mathbb{R}^d$$

Family	θ	Θ	$F(\theta)$	$k(x)$	$t(x)$	ν
Poisson	$\log\lambda$	$\mathbb{R}$	e^θ	$-\log x!$	x	ν_c
Iso.Gaussian	μ	$\mathbb{R}^d$	$\frac{1}{2}\theta^\top\theta$	$\frac{d}{2}\log 2\pi-\frac{1}{2}x^\top x$	x	ν_L

convex set Θ. Function F characterizes the family, and bears different names in the literature (partition function, log-normalizer or cumulant function) and parameter θ (natural parameter) defines the member P_θ of the family $\mathcal{E}_F$. Let $D=\dim(\Theta)$ denote the dimension of Θ, the order of the family. The map $k(x):\mathcal{X}\rightarrow\mathbb{R}$ is an auxiliary function defining a carrier measure ξ with $\mathrm{d}\xi(x)=e^{k(x)}\mathrm{d}\nu(x)$. In practice, we often consider the Lebesgue measure ν_L defined over the Borel σ-algebra $\mathcal{E}=B(\mathbb{R}^d)$ of $\mathbb{R}^d$ for continuous distributions (e.g., Gaussian), or the counting measure ν_c defined on the power set σ-algebra $\mathcal{E}=2^{\mathcal{X}}$ for discrete distributions (e.g., Poisson or multinomial families). The term $t(x)$ is a measure mapping called the sufficient statistic [8]. Table II shows the canonical decomposition for the Poisson and isotropic Gaussian families. Interestingly, any smooth distribution can be arbitrarily finely approximated by a single distribution of an exponential family [10]. Notice that the Kullback–Leibler divergence between members $X_1\sim\mathcal{E}_F(\theta_1)$ and $X_2\sim\mathcal{E}_F(\theta_2)$ of the same exponential family amount to compute a Bregman divergence on swapped natural parameters [11]: $\mathrm{KL}(X_1:X_2)=B_F(\theta_2:\theta_1)$, where $B_F(\theta:\theta')=F(\theta)-F(\theta')-(\theta-\theta')^\top\nabla F(\theta')$, where ∇F denotes the gradient.

II. χ^2 AND HIGHER-ORDER χ^k DISTANCES

A. A Closed-form Formula

When X_1 and X_2 belong to the same restricted exponential family $\mathcal{E}_F$, we obtain the following result:

Lemma 1: The Pearson/Neyman Chi square distance between $X_1\sim\mathcal{E}_F(\theta_1)$ and $X_2\sim\mathcal{E}_F(\theta_2)$ is given by:

$$\chi^2_P(X_1:X_2)=e^{F(2\theta_2-\theta_1)-(2F(\theta_2)-F(\theta_1))}-1, \tag{6}$$
$$\chi^2_N(X_1:X_2)=e^{F(2\theta_1-\theta_2)-(2F(\theta_1)-F(\theta_2))}-1, \tag{7}$$

provided that $2\theta_2-\theta_1$ and $2\theta_1-\theta_2$ belongs to the natural parameter space Θ.

In that case, this implies that the chi square distances are *always* bounded. The proof relies on the following lemma:

Lemma 2: The integral $I_{p,q}(X_1:X_2)=\int x_1(x)^p x_2(x)^q\mathrm{d}\nu(x)$ with $p+q=1$ for $X_1\sim\mathcal{E}_F(\theta_1)$ and $X_2\sim\mathcal{E}_F(\theta_2)$, $p\in\mathbb{R}$, $p+q=1$ converges and equals to $I_{p,q}=e^{F(p\theta_1+q\theta_2)-(pF(\theta_1)+qF(\theta_2))}$, provided the natural parameter space Θ is *affine*.

Proof: Let us calculate the integral $I_{p,q}$:

$$\begin{aligned}
&= \int \exp(p(\langle t(x),\theta_1\rangle-F(\theta_1)+k(x))) \\
&\quad\times\exp(q(\langle t(x),\theta_2\rangle-F(\theta_2)+k(x)))\mathrm{d}\nu(x), \\
&= \int e^{\langle t(x),p\theta_1+q\theta_2\rangle-(pF(\theta_1)+qF(\theta_2))+k(x)}\mathrm{d}\nu(x), \\
&= e^{F(p\theta_1+q\theta_2)-(pF(\theta_1)+qF(\theta_2))}\int p_F(x|p\theta_1+q\theta_2)\mathrm{d}\nu(x).
\end{aligned}$$

When $p\theta_1+q\theta_2\in\Theta$, we have $\int p_F(x|p\theta_1+q\theta 2)\mathrm{d}\nu(x)=1$, hence the result. ∎

To prove Lemma 1, we rewrite $\chi^2_P(X_1:X_2)=\int(\frac{x_2^2(x)}{x_1(x)}-2x_2(x)+x_1(x))\mathrm{d}\nu(x)=(\int x_1(x)^{-1}x_2(x)^2\mathrm{d}\nu(x))-1$, and apply Lemma 2 for $p=-1$ and $q=2$ (checking that $p+q=1$). The closed-form formula for the Neyman chi square follows from the fact that $\chi^2_N(X_1:X_2)=\chi^2_P(X_2:X_1)$. Thus when the natural parameter space Θ is affine, the Pearson/Neyman Chi square distances and its symmetrization $\chi^2_P+\chi^2_N$ between members of the same exponential family are available in closed-form. Examples of such families are the Poisson, binomial, multinomial, or isotropic Gaussian families to name a few. Let us call those families: *affine exponential families* for short. Note that we can rewrite $I_{p,q}(X_1:X_2)=e^{-J_{p,q}(\theta_1:\theta_2)}$ with $J_{p,q}(\theta_1:\theta_2)=pF(\theta_1)+qF(\theta_2)-F(p\theta_1+q\theta_2)$.

B. The Poisson and Isotropic Gaussian Cases

As reported in Table II, those Poisson and isotropic Gaussian exponential families have affine natural parameter spaces Θ.

- The Poisson family. For $P_1\sim\mathrm{Poi}(\lambda_1)$ and $P_2\sim\mathrm{Poi}(\lambda_2)$, we have:

$$\chi^2_P(\lambda_1:\lambda_2)=\exp\left(\frac{\lambda_2^2}{\lambda_1}-2\lambda_2+\lambda_1\right)-1. \tag{8}$$

To illustrate this formula with a numerical example, consider $X_1 \sim \mathrm{Poi}(1)$ and $X_2 \sim \mathrm{Poi}(2)$. Then, it comes that $\chi^2_P(P_1 : P_2) = e - 1 \simeq 1.718$.

- The isotropic Normal family. For $N_1 \sim \mathrm{Nor}_I(\mu_1)$ and $N_2 \sim \mathrm{Nor}_I(\mu_2)$, we have according to Table II: $\chi^2_P(\mu_1 : \mu_2) = e^{\frac{1}{2}(2\mu_2-\mu_1)^\top(2\mu_2-\mu_1)-(\mu_2^\top\mu_2-\frac{1}{2}\mu_1^\top\mu_1)} - 1$. In that case the χ^2 distance is symmetric:

$$\chi^2_P(\mu_1 : \mu_2) = e^{(\mu_2-\mu_1)^\top(\mu_2-\mu_1)} - 1 = \chi^2_N(\mu_1 : \mu_2) \quad (9)$$

C. Extensions to Higher-order Vajda χ^k Divergences

The higher-order Pearson-Vajda χ^k_P and $|\chi^k_P|$ distances [6] are defined by:

$$\chi^k_P(X_1 : X_2) = \int \frac{(x_2(x) - x_1(x))^k}{x_1(x)^{k-1}} \mathrm{d}\nu(x), \quad (10)$$

$$|\chi|^k_P(X_1 : X_2) = \int \frac{|x_2(x) - x_1(x)|^k}{x_1(x)^{k-1}} \mathrm{d}\nu(x), \quad (11)$$

are f-divergences for the generators $(u-1)^k$ and $|u-1|^k$ (with $|\chi|^k_P(X_1 : X_2) \geq \chi^k_P(X_1 : X_2)$). When $k = 1$, we have $\chi^1_P(X_1 : X_2) = \int (x_1(x) - x_2(x))\mathrm{d}\nu(x) = 0$ (*i.e.*, divergence is never discriminative), and $|\chi^1_P|(X_1, X_2)$ is twice the total variation distance (the only metric f-divergence). χ^0_P is the unit constant. Observe that the χ^k_P "distance" may be negative for odd k (signed distance), but not $|\chi|^k_P$. We can compute the χ^k_P term explicitly by performing the binomial expansion:

Lemma 3: The (signed) χ^k_P distance ($k \in \mathbb{N}$) between members $X_1 \sim \mathcal{E}_F(\theta_1)$ and $X_2 \sim \mathcal{E}_F(\theta_2)$ of the same affine exponential family is always bounded and equal to:

$$\chi^k_P(X_1 : X_2) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} \frac{e^{F((1-j)\theta_1 + j\theta_2)}}{e^{(1-j)F(\theta_1) + jF(\theta_2)}}. \quad (12)$$

Proof:

$$\chi^k_P(X_1 : X_2) = \int \frac{(x_2(x) - x_1(x))^k}{x_1(x)^{k-1}} \mathrm{d}\nu(x), \quad (13)$$

$$= \int \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} \frac{x_1(x)^{k-j} x_2(x)^j}{x_1(x)^{k-1}} \mathrm{d}\nu(x), \quad (14)$$

$$= \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} \int x_1(x)^{1-j} x_2(x)^j \mathrm{d}\nu(x). \quad (15)$$

Then the proof follows from Lemma 2 that shows that $I_{1-j,j}(X_1 : X_2) = \int x_1(x)^{1-j} x_2(x)^j \mathrm{d}\nu(x) = \frac{e^{F((1-j)\theta_1 + j\theta_2)}}{e^{(1-j)F(\theta_1) + jF(\theta_2)}}$. ∎

For Poisson/Normal distributions, we get:

$$\chi^k_P(\lambda_1 : \lambda_2) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} e^{\lambda_1^{1-j}\lambda_2^j - ((1-j)\lambda_1 + j\lambda_2)}, \quad (16)$$

$$\chi^k_P(\mu_1 : \mu_2) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} e^{\frac{1}{2}j(j-1)(\mu_1-\mu_2)^\top(\mu_1-\mu_2)}. \quad (17)$$

Observe that for $\lambda_1 = \lambda_2 = \lambda$, we have $\chi^k_P(\lambda_1 : \lambda_2) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} e^{\lambda-\lambda} = (1-1)^k = 0$ when $k \in \mathbb{N}$, as expected. The χ^k_P value is always bounded. For sanity check, consider the binomial expansion for $k = 2$, we have: $\chi^2_P(\lambda_1 : \lambda_2) = \binom{2}{0} e^{\lambda_1 - \lambda_1} - \binom{2}{1} e^{\lambda_2 - \lambda_2} + \binom{2}{2} e^{\frac{\lambda_2^2}{\lambda_1} - 2\lambda_2} = e^{\frac{\lambda_2^2}{\lambda_1} - 2\lambda_2} - 1$, in accordance with Eq. (8). Consider a numerical example: Let $\lambda_1 = 0.6$ and $\lambda_2 = 0.3$ ($\lambda_1 > \lambda_2$), then $\chi^2_P \sim 0.16$, $\chi^3_P \sim -0.03$, $\chi^4_P \sim 0.04$, $\chi^5_P \sim -0.02$, $\chi^6_P \sim 0.018$, $\chi^7_P \sim -0.013$, $\chi^8_P \sim 0.01$, $\chi^9_P \sim -0.0077$, $\chi^{10}_P \sim 0.006$, etc. This numerical example illustrates the alternating sign of those χ^k-type signed distances. The series of $(\chi^k_P)_k$ may diverge. Consider $\lambda_1 = \frac{1}{2}$ and $\lambda_2 = 0.7$ ($\lambda_1 < \lambda_2$). We have $\chi^2_P \sim 0.083$, $\chi^3_P \sim 0.063$, $\chi^4_P \sim 0.11$, $\chi^5_P \sim 0.28$, $\chi^6_P \sim 1.08$, $\chi^7_P \sim 6.96$, $\chi^8_P \sim 80.3$, $\chi^9_P \sim 1951.9$, and $\chi^{10}_P \sim 132503.9$.

III. f-Divergences From Taylor Series

Recall that the f-divergence defined for a generator f is $I_f(X_1 : X_2) = \int x_1(x) f\left(\frac{x_2(x)}{x_1(x)}\right) \mathrm{d}\nu(x)$. Assuming f analytic, we use the Taylor expansion about a point λ: $f(x) = f(\lambda) + f'(\lambda)(x - \lambda) + \frac{1}{2}f''(\lambda)(x-\lambda)^2 + \ldots = \sum_{i=0}^{\infty} \frac{1}{i!} f^{(i)}(\lambda)(x-\lambda)^i$, the power series expansion of f, for $\lambda \in \mathrm{int}(\mathrm{dom}(f^{(i)}))\forall i \geq 0$.

Lemma 4 (extends Theorem 1 of [6]): When bounded, the f-divergence I_f can be expressed as the power series of higher order Chi-type distances:

$$I_f(X_1 : X_2) = \int x_1(x) \sum_{i=0}^{\infty} \frac{1}{i!} f^{(i)}(\lambda) \left(\frac{x_2(x)}{x_1(x)} - \lambda\right)^i \mathrm{d}\nu(x),$$

$$\overset{*}{=} \sum_{i=0}^{\infty} \frac{1}{i!} f^{(i)}(\lambda)\ \chi^i_{\lambda,P}(X_1 : X_2), \quad (18)$$

In the $*$ equality, we swapped the integral and sum according to Fubini theorem since we assumed that $I_f < \infty$, and $\chi^i_{\lambda,P}(X_1 : X_2)$ is a generalization of the χ^i_P defined by:

$$\chi^i_{\lambda,P}(X_1 : X_2) = \int \frac{(x_2(x) - \lambda x_1(x))^i}{x_1(x)^{i-1}} \mathrm{d}\nu(x). \quad (19)$$

and $\chi^0_{\lambda,P}(X_1 : X_2) = 1$ by convention. Note that $\chi^i_{\lambda,P} \geq f(1) = 0$ is a f-divergence for $f(u) = (u-\lambda)^k - (1-\lambda)^k$ (convex for even k). Eq. (18) yields a meaningful numerical approximation scheme by truncating the series to the first s terms, provided that the Taylor remainder is bounded.

- Choosing $\lambda = 1 \in \mathrm{int}(\mathrm{dom}(f^{(i)}))$, we approximate the f-divergence as follows (Theorem 1 of [6]):

$$|I_f(X_1 : X_2) - \sum_{k=0}^{s} \frac{f^{(k)}(1)}{k!} \chi^k_P(X_1 : X_2)| \leq \frac{1}{(s+1)!} \|f^{(s+1)}\|_\infty (M-m)^s, \quad (20)$$

where $\|f^{(s+1)}\|_\infty = \sup_{t \in [m,M]} |f^{(s+)}(t)|$ and $m \leq \frac{p}{q} \leq M$. Notice that by assuming the "fatness" of $\frac{p}{q}$, we ensure that $I_f < \infty$.

- Choosing $\lambda = 0$ (whenever $0 \in \mathrm{int}(\mathrm{dom}(f^{(i)}))$) and affine exponential families, we get the f-divergence in a much simpler analytic expression:

$$I_f(X_1 : X_2) = \sum_{i=0}^{\infty} \frac{f^{(i)}(0)}{i!} I_{1-i,i}(\theta_1 : \theta_2), \tag{21}$$

$$I_{1-i,i}(\theta_1 : \theta_2) = \frac{e^{F(i\theta_2 + (1-i)\theta_1)}}{e^{iF(\theta_2)+(1-i)F(\theta_1)}}. \tag{22}$$

Lemma 5: The bounded f-divergences between members of the same affine exponential family can be computed as an equivalent power series whenever f is analytic.

Corollary 1: A second-order Taylor expansion yields $I_f(X_1 : X_2) \sim f(1) + f'(1)\chi_P^1(X_1 : X_2) + \frac{1}{2}f''(1)\chi_P^2(X_1 : X_2)$. Since $f(1) = 0$ (f can always be renormalized) and $\chi_P^1(X_1 : X_2) = 0$, it follows that

$$I_f(X_1 : X_2) \sim \frac{f''(1)}{2}\chi_P^2(X_1 : X_2), \tag{23}$$

and reciprocally $\chi_P^2(X_1 : X_2) \sim \frac{2}{f''(1)} I_f(X_1 : X_2)$ ($f''(1) > 0$ follows from the strict convexity of the generator). When $f(u) = u \log u$, this yields the well-known approximation [1]: $\chi_P^2(X_1 : X_2) \sim 2\ \mathrm{KL}(X_1 : X_2)$.

For affine exponential families, we then plug the closed-form formula of Lemma 1 to get a simple approximation formula of I_f. For example, consider the Jensen-Shannon divergence (Table I) with $f''(u) = \frac{1}{u} - \frac{1}{u+1}$ and $f''(1) = \frac{1}{2}$. It follows that $I_{\mathrm{JS}}(X_1 : X_2) \sim \frac{1}{4}\chi_P^2(X_1 : X_2)$. (For Poisson distributions $\lambda_1 = 5$ and $\lambda_2 = 5.1$, we get 1.15% relative error).

A. Example 1: χ^2 Revisited

Let us start with a sanity check for the χ^2 distance between Poisson distributions. The Pearson chi square distance is a f-divergence for $f(t) = t^2 - 1$ with $f'(t) = 2t$ and $f''(t) = 2$ and $f^{(i)}(t) = 0$ for $i > 2$. Thus, with $f^{(0)}(0) = -1$, $f^{(1)}(0) = 0$, $f^{(2)}(0) = 2$, and $f^{(i)}(0) = 0$ for $i > 2$. Recall that $I_{1-i,i}(\theta_1 : \theta_2) = e^{F(i\theta_2+(1-i)\theta_1) - (iF(\theta_2)+(1-i)F(\theta_1))} = \exp(\lambda_2^i \lambda_1^{1-i} - i\lambda_2 - (1-i)\lambda_1)$. Note that $I_{1-i,i}(\lambda, \lambda) = e^0 = 1$ for all i. Thus we get: $I_f(X_1 : X_2) = -I_{1,0} + I_{-1,2}$ with $I_{1,0} = e^{\lambda_1 - \lambda_1} = 1$ and $I_{-1,2} = e^{\frac{\lambda_2^2}{\lambda_1} - 2\lambda_2 + \lambda_1}$. Thus, we obtain $I_f(X_1 : X_2) = -1 + e^{\frac{\lambda_2^2}{\lambda_1} - 2\lambda_2 + \lambda_1}$, in accordance with Eq. (8).

B. Example 2: Kullback–Leibler Divergence

By choosing $f(u) = -\log u$, we obtain the Kullback–Leibler divergence (see Table I). We have $f^{(i)}(u) = (-1)^i(i-1)!u^{-i}$, and hence $\frac{f^{(i)}(1)}{i!} = \frac{(-1)^i}{i}$, for $i \geq 1$ (with $f(1) = 0$). Since $\chi_{1,P}^1 = 0$, it follows that:

$$\mathrm{KL}(X_1 : X_2) = \sum_{j=2}^{\infty} \frac{(-1)^i}{i}\ \chi_P^j(X_1 : X_2). \tag{24}$$

Note that for the case of KL divergence between members of the same exponential families, the divergence can be expressed in a simpler closed-form using a Bregman divergence [11] on the swapped natural parameters. For example, consider Poisson distributions with $\lambda_1 = 0.6$ and $\lambda_2 = 0.3$, the Kullback–Leibler divergence computed from the equivalent Bregman divergence yields KL ~ 0.1158, the stochastic evaluation of Eq. (4) with $n = 10^6$ yields $\widehat{KL} \sim 0.1156$ and the KL divergence obtained from the truncation of Eq. (24) to the first s terms yields the following sequence: $0.0809(s = 2)$, $0.0910(s = 3)$, $0.1017(s = 4)$, $0.1135(s = 10)$, $0.1150(s = 15)$, etc.

IV. Concluding Remarks

We investigated the calculation of statistical f-divergences between members of the same exponential family with affine natural space. We first reported a generic closed-form formula for the Pearson/Neyman χ^2 and Vajda χ^k-type distance (always bounded), and instantiated that formula for two affine exponential families: (1) Poisson and (2) isotropic Gaussian families. We then considered the Taylor expansion of the generator f at any given point λ to deduce an analytic expression of f-divergences using Pearson-Vajda-type distances (Eq. (20) and Eq. (21)). In practice, the f-divergences can be well-approximated by the truncated series when the Taylor exact remainder is *bounded*. The convergence rate of the f-divergence approximation depends on the values of the successive derivatives of $f^{(i)}(\lambda)$. A second-order Taylor approximation yielded a fast estimation of f-divergences. This framework shall find potential applications in signal processing and when designing inequality bounds between divergences.

A Java package that illustrates numerically the lemmata is provided at: www.informationgeometry.org/fDivergence/

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, 1991.

[2] F. Österreicher, "Distances based on the perimeter of the risk set of a testing problem," *Austrian J. Statist.*, vol. 42, no. 1, p. 3.19, 2013.

[3] S.-i. Amari, "Alpha-divergence is unique, belonging to both f-divergence and Bregman divergence classes," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4925–4931, 2009.

[4] S. Amari and H. Nagaoka, *Methods of Information Geometry*. Cambridge, U.K.: Oxford Univ. Press, 2000.

[5] M. Khosravifard, D. Fooladivanda, and T. A. Gulliver, "Confliction of the convexity and metric properties in f-divergences," *IEICE Trans. Fund. Electron. Commun. Comput. Sci.* , vol. 9, pp. 1848–1853, 2007.

[6] N. Barnett, P. Cerone, S. Dragomir, and A. Sofo, "Approximating Csiszár f-divergence by the use of taylor's formula with integral remainder," *Math. Inequal. Applicat.*, vol. 5, no. 3, pp. 417–434, 2002.

[7] T. Kanamori, T. Suzuki, and M. Sugiyama, "f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 708–720, 2012.

[8] L. D. Brown, *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*. Beachwood, OH, USA: Inst. Math. Statist., 1986.

[9] A. Cichocki and S.-i. Amari, "Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities," *Entropy*, vol. 12, no. 6, pp. 1532–1568, 2010.

[10] D. C. Brody, "A note on exponential families of distributions," *J. Phys. A: Math. Gen.*, vol. 40, pp. 691–695, 2007.

[11] F. Nielsen and S. Boltz, "The Burbea-Rao and Bhattacharyya centroids," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5455–5466, 2011.