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 $\chi^2$ and Neyman chi square distances $\chi^2_\mu(X_1 : X_2) - \chi^2_\mu(X_2 : X_1)$, and the Kullback–Leibler divergence [1] defined respectively by:

$$\chi^2_\mu(X_1 : X_2) = \int \frac{(x_2(x) - x_1(x))^2}{x_1(x)} d\nu(x),$$

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

where $X_1$ and $X_2$ are probability measures absolutely continuous with respect to a reference measure $\nu$, and $x_1$ and $x_2$ denote their Radon-Nikodym densities, respectively. Those dissimilarity measures $\mathcal{M}$ are termed divergences to contrast with metric distances since they are oriented distances (i.e., $\mathcal{M}(X_1, X_2) \neq \mathcal{M}(X_2, X_1)$) 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], defined for an arbitrary functional $f$:

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

where $f$ is a convex function $f : (0, \infty) \subseteq \text{dom}(f) \mapsto [0, \infty]$ such that $f(1) = 0$. Indeed, it follows from Jensen inequality that $I_f(X_1 : X_2) \leq f(\int x_2(x) d\nu(x)) = f(1) = 0$. Furthermore, wlog., we may consider $f'(1) = 0$ and fix the scale of divergence by setting $f''(1) = 1$, see [3]. Those $f$-divergences 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:

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

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 $n \rightarrow \infty$, 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 $\langle x, y \rangle = x^T y$. An exponential family [8] is a set of probability measures $\mathcal{E}_F = \{ P_{\theta} \}$, dominated by a measure $\nu$ having their Radon-Nikodym densities $p_{\theta}$ expressed canonically as:

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

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

1Beware that sometimes the $\chi^2_\mu$ and $\chi^2_\mu$ definitions are inverted in the literature. This may stem from an alternative definition of $f$-divergences defined as $I_f(X_1 : X_2) = \int x_2(x) f\left(\frac{x_1(x)}{x_2(x)}\right) d\nu(x) = I_{f^*}(X_2 : X_1)$.
TABLE I
SOME COMMON f-DIVERGENCES f WITH CORRESPONDING GENERATORS: EXCEPT THE TOTAL VARIATION, f-DIVERGENCES ARE NOT METRIC [5]

<table>
<thead>
<tr>
<th>Name of the f-divergence</th>
<th>Formula ( I_f(P : Q) )</th>
<th>Generator ( f(u) ) with ( f(1) = 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total variation (metric)</td>
<td>( \frac{1}{2} \int \left</td>
<td>p(x) - q(x) \right</td>
</tr>
<tr>
<td>Pearson ( \chi^2_p )</td>
<td>( \int \frac{(p(x) - q(x))^2}{p(x)} , d\nu(x) )</td>
<td>( (u - 1)^2 )</td>
</tr>
<tr>
<td>Neyman ( \chi^2_N )</td>
<td>( \int \frac{(p(x) - q(x))^2}{q(x)} , d\nu(x) )</td>
<td>( \frac{1}{u} (u - 1) )</td>
</tr>
<tr>
<td>Pearson-Vajda ( \chi^k_p )</td>
<td>( \int \frac{(p(x) - q(x))^k}{p(x)} , d\nu(x) )</td>
<td>( (u - 1)^k )</td>
</tr>
<tr>
<td>Pearson-Vajda ( \chi^k_p )</td>
<td>( \int \frac{(p(x) - q(x))^k}{q(x)} , d\nu(x) )</td>
<td>( \frac{1}{u} (u - 1)^k )</td>
</tr>
<tr>
<td>Kullback-Leibler</td>
<td>( \int p(x) \log \frac{p(x)}{q(x)} , d\nu(x) )</td>
<td>( -\log u )</td>
</tr>
<tr>
<td>reverse Kullback-Leibler</td>
<td>( \int q(x) \log \frac{q(x)}{p(x)} , d\nu(x) )</td>
<td>( u \log u )</td>
</tr>
<tr>
<td>Jensen-Shannon</td>
<td>( \frac{1}{2} \int \left( p(x) \log \frac{p(x)}{p(x) + q(x)} + q(x) \log \frac{q(x)}{p(x) + q(x)} \right) , d\nu(x) )</td>
<td>( -(u + 1) \log \frac{1 + u}{2} + u \log u )</td>
</tr>
</tbody>
</table>

TABLE II
EXAMPLES OF EXPONENTIAL FAMILIES WITH AFFINE NATURAL SPACE \( \Theta, \nu \). DENOTES THE COUNTING MEASURE AND \( \nu_\beta \) THE LEBESGUE MEASURE

| Convex set \( \Theta \). Function \( F \) characterizes the family, and bears different names in the literature (partition function, log-normalizer or cumulant function) and parameter \( \theta \) (natural parameter) defines the member \( p_{\theta} \) of the family \( \mathcal{E} \). Let \( D = \dim(\Theta) \) denote the dimension of \( \Theta \), the order of the family. The map \( k(x) : \mathcal{E} \to \mathbb{R} \) is an auxiliary function defining a carrier measure \( \xi \) with \( d\xi(x) = e^{k(x)} \, d\nu(x) \). In practice, we often consider the Lebesgue measure \( \nu_{\mathbb{R}} \). The Borel \( \sigma \)-algebra \( \mathcal{E} = B(\mathbb{R}^d) \) for continuous distributions (e.g., Gaussian), or the counting measure \( \nu_{\mathbb{B}} \) defined on the power set \( \sigma \)-algebra \( \mathcal{E} = 2^\mathbb{B} \) for discrete distributions (e.g., Poisson or multinomial families). The term \( t(x) \) is a measurement mapping called the sufficient statistic [8]. Table II shows canonical decompositions 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 amounts to compute a Bregman divergence on swapped natural parameters [11]:

\[ \text{KL}(X_1 : X_2) = B_F(\theta_2 : \theta_1), \]

where \( B_F(\theta : \theta') = F(\theta) - F(\theta') - (\theta - \theta')^\top \nabla F(\theta') \), where \( \nabla F \) denotes the gradient.

II. \( \chi^2 \) AND HIGHER-ORDER \( \chi^k \) 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:

\[ \chi^2_p(X_1 : X_2) = \int \frac{(p(x) - q(x))^2}{p(x)} \, d\nu(x) \]

\[ = e^{F(\theta_2) - F(\theta_1)} \int \frac{(p(x) - q(x))^2}{p(x)} \, d\nu(x) - 1, \]

\[ \chi^2_N(X_1 : X_2) = e^{F(\theta_2) - F(\theta_1)} \int \frac{(p(x) - q(x))^2}{q(x)} \, d\nu(x) - 1, \]

provided that \( 2\theta_2 - \theta_1 \) and \( 2\theta_1 - \theta_2 \) belongs to the natural parameter space \( \Theta \).

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

**Lemma 2:** The integral \( \int p(x) \, d\nu(x) \) for \( p + q = 1 \) 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 ealualts to \( \text{KL}(p : q) = e^{F(\theta_1) + qF(\theta_2) - (pF(\theta_1) + qF(\theta_2))}, \) provided the natural parameter space \( \Theta \) is affine.

**Proof:** Let us calculate the integral \( \int p(x) \, d\nu(x) \):

\[ \int p(x) \, d\nu(x) = \int \exp(p(t(x) \theta_1 - F(\theta_1)) + k(x)) \times \exp(q(t(x) \theta_2 - F(\theta_2)) + k(x)) \, d\nu(x). \]

\[ = \int \exp(t(x) \theta_1 + q \theta_2) \exp(k(x)) \, d\nu(x). \]

\[ = e^{F(\theta_1) + qF(\theta_2) - (pF(\theta_1) + qF(\theta_2))}. \]

When \( \theta_1 + q \theta_2 \in \Theta \), we have \( \int p(x) \, d\nu(x) = 1 \), hence the result.

To prove Lemma 1, we rewrite \( \chi^2_p(X_1 : X_2) = \int \frac{(p(x) - q(x))^2}{p(x)} \, d\nu(x) = \int \frac{(p(x) - q(x))^2}{q(x)} \, d\nu(x) = 1 \) and apply Lemma 2 for \( p = -1 \) and \( q = 2 \) (checking that \( p + q = 1 \)). The closed-form formula for the Pearson/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 \( \Theta \) is affine, the Pearson/Neyman Chi square distances and its symmetrization \( \chi^2_N \) 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 \( \chi^2_p(X_1 : X_2) = e^{-\chi^2_p(\theta_1 : \theta_2)} \), with \( \chi^2_p(\theta_1 : \theta_2) = pF(\theta_1) + qF(\theta_2) - F(\theta_1 + q \theta_2) \).

B. The Poisson and Isotropic Gaussian Cases

As reported in Table II, those Poisson and isotropic Gaussian exponential families have affine natural parameter spaces \( \Theta \).

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

\[ \chi^2_p(\lambda_1 : \lambda_2) = \exp \left( \frac{\lambda_2^2}{\lambda_1} - 2 \lambda_2 + \lambda_1 \right) - 1. \]
To illustrate this formula with a numerical example, consider $X_1 \sim \text{Poi}(1)$ and $X_2 \sim \text{Poi}(2)$. Then, it comes that $\chi^2_P(P_1 : P_2) = e^{-1} \approx 1.718$.

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

$$\chi^2_P(\mu_1 : \mu_2) = e^{(\mu_2 - \mu_1)^2} - 1 - \chi^2_P(\mu_2 : \mu_1)$$  \hspace{1cm} (9)

C. Extensions to Higher-order Vajda $\chi^k$ Divergences

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

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

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

are $f$-divergences for the generators (w.e) and $|u - 1|$ (with $|\chi^k_P(X_1 : X_2) \geq \chi^k_P(X_1 : X_2)$). When $k = 1$, we have $\chi^1_P(X_1 : X_2) = \|x_1(x) - x_2(x)\|/d\nu(x) = 0$ (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). $\chi^0_P$ is the unit constant. Observe that the $\chi^k_P$-"distance" may be negative for odd $k$ (signed distance), but not $\chi^k_P$. We can compute the $\chi^k_P$ term explicitly by performing the binomial expansion:

**Lemma 3:** The (signed) $\chi^k_P$ 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:

$$\chi^k_P(X_1 : X_2) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} e^{\frac{1}{2}((1-j)\theta_1 + j\theta_2)}.$$  \hspace{1cm} (12)

**Proof:**

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

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

$$= \int \frac{1}{x_1(x)^{k+1}} \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} x_2(x)^j d\nu(x).$$  \hspace{1cm} (15)

Then the proof follows from Lemma 2 that shows that $I_{1-k,j}(X_1 : X_2) = \int x_1(x)^{1-j} x_2(x)^j d\nu(x) = \frac{1}{e^{\frac{1}{(1-j)x_1(x) + jx_2(x)}}}.$

For Poisson/Normal distributions, we get:

$$\chi^k_P(\lambda_1 : \lambda_2) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} e^{\frac{1}{2}((1-j)\lambda_1 + j\lambda_2)}.$$  \hspace{1cm} (16)

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

Observe that for $\lambda_1 = \lambda_2 = \lambda$, we have $\chi^k_P(\lambda_1 : \lambda_2) = \Sigma_{j=0}^{k} (-1)^{k-j} \binom{k}{j} e^{-j \lambda} = (1 - 1) = 0$ when $k \in \mathbb{N}$, as expected. The $\chi^k_P$ value is always bounded. For sanity check, consider the binomial expansion for $k = 2$, we have: $\chi^2_P(\lambda_1 : \lambda_2) = \frac{2(2)}{2} e^{\lambda_1 - \lambda_2} - \frac{2(1)}{1} e^{\lambda_2 - \lambda_1} + \frac{2(2)}{2} e^{\lambda_2 - 2\lambda_1} = e^{\lambda_1 - 2\lambda_2} - 1$, in accordance with Eq. (8). Consider a numerical example: Let $\lambda_1 = 0.6$ and $\lambda_2 = 0.3 (\lambda_1 > \lambda_2)$, then $\chi^2_P \approx 0.16$, $\chi^3_P \approx -0.03$, $\chi^4_P \approx 0.04$, $\chi^5_P \approx -0.02$, $\chi^6_P \approx 0.018$, $\chi^7_P \approx -0.013$, $\chi^8_P \approx 0.001$, $\chi^9_P \approx -0.0077$, etc. This numerical example illustrates the alternating sign of those $\chi^k$-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 $\chi^2_P \approx 0.083$, $\chi^3_P \approx 0.063$, $\chi^4_P \approx 0.11$, $\chi^5_P \approx 0.28$, $\chi^6_P \approx 1.08$, $\chi^7_P \approx 6.96$, $\chi^8_P \approx 80.3$, $\chi^9_P \approx 1951.9$, and $\chi^10_P \approx 132563.9$.

III. $f$-Divergences From Taylor Series

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

**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:

$$I_f(X_1 : X_2) = \int x_1(x) \sum_{i=0}^{\infty} \frac{1}{i!} f^{(i)}(\lambda) \left( \frac{x_2(x)}{x_1(x)} - \lambda \right)^i d\nu(x),$$  \hspace{1cm} (18)

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

$$\chi^k_{\lambda,f}(X_1 : X_2) = \int \frac{(x_2(x) - \lambda x_1(x))^k}{x_1(x)^{k+1}} d\nu(x),$$  \hspace{1cm} (19)

and $\chi^0_{\lambda,f}(X_1 : X_2) = 1$ by convention. Note that $\chi^k_{\lambda,f} \geq f(1) = 0$ 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 $\lambda = 1 \in \text{int(dom}(f^{(i)}))$, we approximate the $f$-divergence as follows (Theorem 1 of [6]):

$$I_f(X_1 : X_2) = \int \frac{(x_2(x) - x_1(x))^s}{x_1(x)^{s+1}} d\nu(x).$$  \hspace{1cm} (20)

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

---

This page contains a detailed mathematical analysis of $\chi$-divergences and their higher-order generalizations, specifically the $\chi^k$-divergence. It includes numerical examples and theoretical derivations to illustrate the concepts, focusing on the isotropic Normal family and extensions to higher-order Vajda divergences. The text also introduces $f$-divergences from Taylor series and discusses their approximation using numerical schemes.
Choosing $\lambda = 0$ (whenever $0 \in \text{int}(\text{dom}(f))$ and affine exponential families, we get the $f$-divergence in a much simpler analytic expression:

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

$$I_{1-i}(\theta_1 : \theta_2) = \frac{e^{F(\theta_1) - (1-i)F(\theta_2)}}{e^{F(\theta_1) - (1-i)F(\theta_2)}}.$$  \hspace{1cm} (21)

**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

$$I_f(X_1 : X_2) \sim f(1) + f'(1)\chi^2_f(X_1 : X_2) + \frac{1}{2} f''(1)\chi^2_f(X_1 : X_2).$$

Since $f(1) = 0$ ($f$ can always be renormalized) and $\chi^2_f(X_1 : X_2) = 0$, it follows that

$$I_f(X_1 : X_2) \sim \frac{f''(1)}{2} \chi^2_f(X_1 : X_2),$$

and reciprocally

$$\chi^2_f(X_1 : X_2) \sim \frac{f''(1)}{2} I_f(X_1 : X_2) f''(1) > 0 \text{ follows from the strict convexity of the generator}.$$  \hspace{1cm} (23)

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 $f''(u) = \frac{1}{u - \frac{1}{u + 2}}$ and $f''(1) = \frac{1}{2}$. It follows that $I_{JS}(X_1 : X_2) \sim \frac{1}{2} \chi^2_f(X_1 : X_2)$. (For Poisson distributions $\lambda_1 = 5$ and $\lambda_2 = 5.1$, we get a 1.15\% relative error).

**A. Example 1: $\chi^2$ Revisited**

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

$$KL(X_1 : X_2) = \sum_{i=2}^{\infty} \frac{(-1)^i}{i} \chi^2_f(X_1 : X_2).$$

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 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 $KL \sim 0.1158$, the stochastic evaluation of Eq. (4) with $n = 10^6$ yields $KL \sim 0.1156$ and the KL divergence obtained from the truncation of Eq. (24) to the first 1 terms yields the following sequence: $0.0809(s = 2), 0.0910(s = 3), 0.1017(s = 4), 0.1135(s = 10), 0.1155(s = 15), \text{ etc.}$

**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 $\chi^2$ and Vajda $\chi^4$-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 $\lambda$ 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 $f^{(i)}$. 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**


