An information-geometric characterization of Chernoff information

Frank Nielsen, *Senior Member, IEEE*

Sony Computer Science Laboratories, Inc.

3-14-13 Higashi Gotanda

141-0022 Shinagawa-ku, Tokyo, Japan

nielsen@csl.sony.co.jp

Abstract

The Chernoff information was originally introduced for bounding the probability of error of the Bayesian decision rule in binary hypothesis testing. Nowadays, it is often used as a notion of symmetric distance in statistical signal processing or as a way to define a middle distribution in information fusion. Computing the Chernoff information requires to solve an optimization problem that is numerically approximated in practice. We consider the Chernoff distance for distributions belonging to the same exponential family including the Gaussian and multinomial families. By considering the geometry of the underlying statistical manifold, we define exactly the solution of the optimization problem as the unique intersection of a geodesic with a dual hyperplane. Furthermore, we prove analytically that the Chernoff distance amounts to calculate an equivalent but simpler Bregman divergence defined on the distribution parameters. It follows a closed-form formula for the singly-parametric distributions, or an efficient geodesic bisection search for multi-parametric distributions. Finally, based on this information-geometric characterization, we propose three novel information-theoretic symmetric distances and middle distributions, from which two of them admit always closed-form expressions.

Index Terms

Nielsen, F.; "An Information-Geometric Characterization of Chernoff Information," IEEE Signal Processing Letters (SPL), vol. 20, no. 3, pp. 269-272, March 2013,. doi: 10.1109/LSP.2013.2243726

Chernoff information, exponential families, information geometry, Bregman divergence, information fusion.

I. INTRODUCTION

Let $(\mathcal{X}, \mathcal{E})$ be a measurable space with $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{E}$ a σ-algebra on the set $\mathcal{X}$. The Chernoff information $C(P, Q)$ between two probability measures P and Q, with p and q denoting their Radon-Nikodym densities with respect to a dominating measure[1] ν, is defined as [2], [3]:

$$C(P, Q) = -\log \min_{\alpha \in (0,1)} \int p^{\alpha}(x) q^{1-\alpha}(x) \mathrm{d}\nu(x). \tag{1}$$

This notion of information was first introduced by Chernoff [2] (1952) for bounding the probability of error of a binary classification task. Namely, the Chernoff information is well-known in information theory as the *best achievable exponent* for a Bayesian probability of error in binary hypothesis testing (see [3], Chapter 11). Nowadays, the Chernoff information is often used as a *statistical distance* for various applications of signal processing ranging from sensor networks [4] to image processing tasks like image segmentation [5] or edge detection [6]. In fact, this notion of Chernoff distance can be understood as a generalization of the former Bhattacharrya distance [7], [8] (1943): Let $c_{\alpha}(P : Q) = \int p^{\alpha}(x) q^{1-\alpha}(x) \mathrm{d}\nu(x) \in [0, 1)$ denote the α-Chernoff coefficient of similarity generalizing the Bhattacharrya coefficient (obtained for $\alpha = \frac{1}{2}$). The α-Chernoff divergence[2]:

$$C_{\alpha}(P : Q) = -\log c_{\alpha}(P : Q) \tag{2}$$

generalizes the symmetric Bhattacharrya distance ($\alpha = \frac{1}{2}$). Thus we can interpret the Chernoff information as a *maximization* of the α-Chernoff divergence over the range $\alpha \in (0, 1)$: $C(P, Q) = \max_{\alpha \in (0,1)} C_{\alpha}(P : Q)$. By construction, the Chernoff distance is *symmetric*: $C(P, Q) = \max_{\alpha \in (0,1)} C_{\alpha}(P : Q) = \max_{\alpha \in (0,1)} C_{1-\alpha}(Q : P) = \max_{\beta \in (0,1)} C_{\beta}(Q : P) = C(Q, P)$ making it attractive for information retrieval (IR). In information fusion [4], the Chernoff

[1] We use the measure-theoretic framework [1] to handle both continuous distributions (eg., Gaussians, Beta, etc.) and discrete distributions (eg., Bernoulli, Poisson, multinomial, etc.).

[2] In information geometry [9], the α-Chernoff divergence is related also to Amari α-divergence: $A_{\alpha}(P : Q) = \frac{4}{\alpha(1-\alpha)}(1 - \int p^{\alpha}(x) q^{1-\alpha}(x) \mathrm{d}\nu(x)) = \frac{4}{\alpha(1-\alpha)}(1 - c_{\alpha}(P : Q))$ or Rényi divergences.

information $C(P,Q) = C_{\alpha^*}(P:Q)$ (where α^* denotes the optimal value) is used to define a middle distribution m^* with density $m^*(x) = \frac{p^{\alpha^*}(x)q^{1-\alpha^*}(x)}{c_\alpha(P:Q)}$. Merging probability distributions allows one to efficiently "compress" statistical models (e.g., simplify mixtures [10]).

This letter is organized as follows: Section II considers distributions belonging to the same exponential family, reports a closed-form formula for the α-Chernoff divergences, and shows that Chernoff information amounts to compute an equivalent Bregman divergence. Section III gives a geometric interpretation of the Chernoff distribution (achieving the Chernoff information) as the intersection of a primal geodesic with a dual hyperplane. Section IV presents three other types of Chernoff information and Chernoff middle distributions, with two of them admitting closed-form expressions. Finally, Section V concludes this work.

II. Chernoff information as a Bregman divergence

A. Basics of exponential families

Let $\langle x, y\rangle$ denote the inner product for $x, y \in \mathcal{X}$ that is taken as the scalar product for vector spaces $\mathcal{X}$: $\langle x, y\rangle = x^\top y$. An exponential family [1] $\mathcal{F}_F$ is a set of probability measures $\mathcal{F}_F = \{P_\theta\}_\theta$ dominated by a measure ν having their Radon-Nikodym densities p_θ expressed canonically as:

$$p_\theta(x) = \exp(\langle t(x), \theta\rangle - F(\theta) + k(x)), \tag{3}$$

for θ belonging to the natural parameter space: $\Theta = \left\{\theta \in \mathbb{R}^D \middle| \int p_\theta(x)\mathrm{d}\nu(x) = 1\right\}$. Since $\log \int_{x\in\mathcal{X}} p_\theta(x)\mathrm{d}\nu(x) = \log 1 = 0$, it follows that:

$$F(\theta) = -\log \int \exp(\langle t(x), \theta\rangle + k(x))\mathrm{d}\nu(x). \tag{4}$$

For full regular families [1], it can be proved that function F is strictly convex and differentiable over the open 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 $\mathcal{F}_F$. Let $D = \dim(\Theta)$ denote the dimension of Θ, the order of the family. The map $k(x) : \mathcal{X} \to \mathbb{R}$ 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 [1]. Many usual families of distributions $\{P_\lambda \mid \lambda \in \Lambda\}$ are exponential families [1] in disguise once an invertible mapping $\theta(\lambda) : \Lambda \rightarrow \Theta$ is elucidated and the density written in the canonical form of Eq. 3. We refer to [1] for such decompositions for the Poisson, Gaussian, multinomial, ... distributions. Besides those well-known distributions, exponential families provide a *generic framework* in statistics. Indeed, any smooth density can be arbitrary approximated by a member of an exponential family [11], although the cumulant function F may be defined implicitly only (using Eq. 4).

B. Chernoff α-distance for exponential family members

For distributions P_1 and P_2 of the same exponential family $\mathcal{F}_F$, indexed with respective natural parameter θ_1 and θ_2, the α-Chernoff coefficient can be expressed analytically [12] as:

$$c_\alpha(P_1 : P_2) = \int p_1^\alpha(x) p_2^{1-\alpha}(x) \mathrm{d}\nu(x) = \exp(-J_F^{(\alpha)}(\theta_1 : \theta_2)), \tag{5}$$

where $J_F^{(\alpha)}(\theta_1 : \theta_2)$ is a skew Jensen divergence defined for F on the natural parameter space as:

$$J_F^{(\alpha)}(\theta_1 : \theta_2) = \alpha F(\theta_1) + (1-\alpha) F(\theta_2) - F(\theta_{12}^{(\alpha)}), \tag{6}$$

where $\theta_{12}^{(\alpha)} = \alpha\theta_1 + (1-\alpha)\theta_2 = \theta_2 - \alpha\Delta\theta$, with $\Delta\theta = \theta_2 - \theta_1$.

C. Chernoff distance for exponential family members

It follows that maximizing the α-Chernoff divergence amounts equivalently to maximizing the skew Jensen divergence with respect to α. The directional derivative of F at x with direction u is defined (see [13], page 213) as $\mathrm{d}F(x;u) = \lim_{\tau \rightarrow 0} \frac{F(x+\tau u) - F(x)}{\tau}$. Since by definition $F(\theta) < \infty$ for all $\theta \in \Theta$, the limit always exist and F is Gâteaux differentiable with:

$$\mathrm{d}F(x;u) = \langle \nabla F(x), u \rangle. \tag{7}$$

Therefore, we have:

$$\begin{aligned} \frac{\mathrm{d}J_F^{(\alpha)}(\theta_1 : \theta_2)}{\mathrm{d}\alpha} &= F(\theta_1) - F(\theta_2) - \mathrm{d}F(\theta_{12}^{(\alpha)}; \theta_1 - \theta_2), \\ &= F(\theta_1) - F(\theta_2) - \langle \nabla F(\theta_{12}^{(\alpha)}), \theta_1 - \theta_2 \rangle \end{aligned}$$

Thus we need to find α^* such that:

$$F(\theta_1) - F(\theta_2) - \langle \nabla F(\theta_{12}^{(\alpha^*)}), \theta_1 - \theta_2 \rangle = 0 \tag{8}$$

Since the Hessian of the cumulant function is positive definite [1] ($\nabla^2 F \succ 0$), it follows that the second derivative of the skew Jensen divergence $-\langle \Delta\theta^\top \nabla^2 F(\theta_{12}^{(\alpha)}), \Delta\theta \rangle$ is always negative for $\theta_1 \neq \theta_2$. Therefore there is a *unique* solution for α^* provided members are distinct (if not, the Chernoff distance is obviously 0).

D. Chernoff distance as a Bregman divergence

Our first result states that the Chernoff information between any two distributions belonging to the same exponential family amounts to calculate equivalently a Bregman divergence defined on the natural parameter space, where the Bregman divergence [14] between θ and θ' is defined by setting the generator F to the log-normalizer of the exponential family as:

$$B_F(\theta : \theta') = F(\theta) - F(\theta') - \langle \theta - \theta', \nabla F(\theta') \rangle \tag{9}$$

Theorem 1: The Chernoff distance between two distinct distributions P_1 and P_2 of the same exponential family, with respective natural parameters θ_1 and θ_2, amounts to calculate a Bregman divergence: $C(P_1, P_2) = B_F(\theta_1 : \theta_{12}^{(\alpha^*)})$, where α^* is the unique value satisfying $\langle \nabla F(\theta_{12}^{(\alpha)}), \theta_1 - \theta_2 \rangle = F(\theta_1) - F(\theta_2)$, and $\theta_{12}^{(\alpha)} = \alpha\theta_1 + (1-\alpha)\theta_2$.

a) Proof:: Once the optimal value α^* has been computed, we calculate the Chernoff distance using Eq. 2 that reduces for exponential families to a skew Jensen divergence $C(P_1, P_2) = -\log \int c_{\alpha^*}(P_1 : P_2) = J_F^{(\alpha^*)}(\theta_1 : \theta_2)$. This skew Jensen divergence for the optimal value of α^* yields, in turn, a Bregman divergence:

$$J_F^{(\alpha^*)}(\theta_1 : \theta_2) = B_F(\theta_1 : \theta_{12}^{(\alpha^*)}) = B_F(\theta_2 : \theta_{12}^{(\alpha^*)}), \tag{10}$$

Indeed, from the definition of the Bregman divergence and the fact that $\theta_1 - \theta_{12}^{(\alpha^*)} = (1-\alpha^*)(\theta_1 - \theta_2)$, it follows that $B_F(\theta_1 : \theta_{12}^{(\alpha^*)}) = F(\theta_1) - F(\theta_{12}^{(\alpha^*)}) - (1-\alpha^*)\langle \theta_1 - \theta_2, \nabla F(\theta_{12}^{(\alpha^*)}) \rangle$. Furthermore, since $\langle \nabla F(\theta_{12}^{(\alpha^*)}), \theta_1 - \theta_2 \rangle = F(\theta_1) - F(\theta_2)$, it follows that $B_F(\theta_1 : \theta_{12}^{(\alpha^*)}) = F(\theta_1) - F(\theta_{12}^{(\alpha^*)}) - (1-\alpha^*)F(\theta_1) + (1-\alpha^*)F(\theta_2) = \alpha^* F(\theta_1) + (1-\alpha^*)F(\theta_2) - F(\theta_{12}^{(\alpha^*)}) = J_F^{(\alpha^*)}(\theta_1 : \theta_2)$. □

Note that for singly-parametric distributions, we get a closed-form expression of the Chernoff distance since $\alpha^* = \frac{(F')^{-1}\left(\frac{F(\theta_1)-F(\theta_2)}{\theta_1-\theta_2}\right)-\theta_2}{\theta_1-\theta_2}$. To illustrate the formula, consider the Poisson

exponential family with probability mass functions $p_\lambda(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ that can be decomposed canonically following Eq. 3 with $\theta(\lambda) = \log \lambda$, $F(\theta) = e^\theta = \lambda$, $F'(\theta) = e^\theta$, and $F'^{-1}(\theta) = \log \theta$ (and $t(x) = x$, $k(x) = -\log x!$ and $\nu = \nu_C$ the counting measure). The generic closed-form formula agrees[3] with the specific Poisson Chernoff information reported in [15]:

$$C(P_1, P_2) = B_F(\theta_1 : \theta_{12}^{(\alpha^*)}) = \lambda_1 \frac{(r-1)(\log \frac{r-1}{\log r} - 1) + \log r}{\log r},$$

where $r = \frac{\lambda_2}{\lambda_1}$.

Although we do not have an analytic expression of the Chernoff distance for higher-order exponential families ($D > 1$), we can nevertheless characterize it exactly using information geometry [9], as described in the following section.

III. Chernoff distribution and Chernoff point

Consider the parametric family of probability distributions $\mathcal{F}_F$ as a smooth manifold $\mathcal{M} = \{P_\theta \mid \theta \in \Theta\}$. This section concisely reviews the *dually flat geometry* of the statistical manifold induced by an exponential family. We refer to the textbook [9] for further details. A point $P \in \mathcal{M}$ denotes a distribution with parameter $\theta(P)$ in the *natural* coordinate system. It follows from the Legendre transformation[4] $F^*(\eta) = \max_{\theta \in \Theta} \langle \eta, \theta \rangle - F(\theta)$ that point P can also be indexed as $\eta(P)$ using a dual coordinate system, called the *expectation* parameter, with $\eta = \nabla F(\theta)$ (and $\theta = \nabla F^*(\eta)$). Let $H = \{\eta = \nabla F(\theta) \mid \theta \in \Theta\}$ denote the expectation parameter space. Thus, $P \in \mathcal{M} = P_{\theta(P)} = P_{\eta(P)}$. In the θ-coordinate system, we have $\int p_\theta \mathrm{d}\nu(x) = 1$, and in the dual η-coordinate system, we have for a random variable $X \sim P_\theta$, $E[t(X)] = \eta = \nabla F(\theta)$ (with $\int t_i(x) p_\theta \mathrm{d}\nu(x) = \eta_i$ for $1 \leq i \leq D$), hence its name (expectation parameter). Two points $P_1, P_2 \in \mathcal{M}$ can be connected using two kinds of geodesics: The *linear mixture geodesic* (or m-geodesic) yielding the mixture family:

$$G_m(P_1, P_2) = \{M_{12}^{(\lambda)} \mid \eta(M_{12}^{(\lambda)}) = (1-\lambda)\eta_1 + \lambda \eta_2, \lambda \in [0,1]\}, \tag{11}$$

[3] http://www.informationgeometry.org/ChernoffInformation/ for Java codes.

[4] In convex analysis [13], each strictly convex and differentiable function F is associated with a dual convex conjugate F^* by the Legendre-Fenchel transformation: $F^*(\eta) = \max_{\theta \in \Theta} \langle \eta, \theta \rangle - F(\theta)$. The maximum is obtained for $\eta = \nabla F(\theta)$ (and is unique since $\nabla^2 F \succ 0$). The transformation is an involution $(F^*)^* = F$, and the gradients are reciprocally inverse: $\nabla F^* = (\nabla F)^{-1}$.

(linear interpolation in the expectation parameter), and the *exponential mixture geodesic* (or e-geodesic) yielding:

$$G_e(P_1, P_2) = \{E_{12}^{(\lambda)} \mid \theta(E_{12}^{(\lambda)}) = (1-\lambda)\theta_1 + \lambda\theta_2, \lambda \in [0,1]\}, \quad (12)$$

(linear interpolation in the natural parameters), with distribution of the form $e_{12}^{(\lambda)} = \frac{p_1^\lambda p_2^{1-\lambda}}{c_\lambda(P_1:P_2)}$, where $c_\lambda(P_1 : P_2)$ plays the role of the normalizing coefficient so that $\int_{x\in\mathcal{X}} e_{12}^{(\lambda)} \mathrm{d}\nu(x) = 1$.

The *Chernoff distribution* is the distribution $E_{12}^{(\alpha^*)}$ (with density $e_{12}^{(\alpha^*)}$) belonging to the e-geodesic for $\lambda = 1-\alpha^*$. This distribution corresponds on the statistical manifold to the *Chernoff point* with coordinates $\theta_{12}^* = \alpha^*\theta_1 + (1-\alpha^*)\theta_2$. Since the Kullback-Leibler (KL) divergence $\mathrm{KL}(P_1 : P_2) = \int_{x\in\mathcal{X}} p_1(x) \log \frac{p_1(x)}{p_2(x)} \mathrm{d}\nu(x)$ for members P_1 and P_2 of the same exponential family amounts to compute a Bregman divergence $B_F(\theta_2 : \theta_1)$ on the swapped natural parameters [14], it follows from Eq. 10 that we have:

$$B_F(\theta_1 : \theta_{12}^{(\alpha^*)}) = B_F(\theta_2 : \theta_{12}^{(\alpha^*)}), \quad (13)$$

$$\mathrm{KL}(E_{12}^{(\alpha^*)} : P_1) = \mathrm{KL}(E_{12}^{(\alpha^*)} : P_2). \quad (14)$$

This shows that Chernoff distribution belongs to a bisector. The Chernoff distribution is commonly used in information fusion [4] for defining an average (or mean) distributions.

A. Geometric characterization of the Chernoff distribution

We prove that although the Chernoff distribution P^* may not be available analytically, it can always be exactly characterized geometrically as a unique intersection point:

Theorem 2: The Chernoff distribution P^* of two distributions P_1 and P_2 belonging to the same exponential family is the unique point on the exponential family manifold that belongs to both the e-geodesic and the m-bisector: $P^* = G_e(P_1, P_2) \cap \mathrm{Bi}_m(P_1, P_2)$.

b) Proof:: Since maximizing the α-Chernoff coefficient amounts to maximize the equivalent skew Jensen divergence defined on the natural parameters using linear interpolation $\theta_{12}^{(\alpha)}$, we deduce that the Chernoff distribution P^* belongs to the exponential geodesic $G_e(P_1, P_2)$. Furthermore, the Bregman equi-divergence constraint of Eq. 13 indicates that the Chernoff point should also belong to a Bregman bisector $\mathrm{Bi}_m(P_1, P_2)$ (that was implicitly revealed in Eq. 8), where

$\mathrm{Bi}_m(P_1, P_2)$ is defined as: $\mathrm{Bi}_m(P_1, P_2) : \{P \mid F(\theta_1) - F(\theta_2) + \langle \eta(P), \Delta\theta \rangle = 0\}$, or equivalently using the θ-coordinate system as $\mathrm{Bi}_m(P_1, P_2) : \{P| \ F(\theta_1) - F(\theta_2) + \langle \nabla F(\theta(P)), \Delta\theta \rangle = 0\}$. This bisector is a hyperplane in the $\eta = \nabla F(\theta)$ coordinate system [16] (but a hypersurface in the θ-coordinate system), hence its name m-bisector $\mathrm{Bi}_m(P_1, P_2)$. It follows that $P^* = G_e(P_1, P_2) \cap \mathrm{Bi}_m(P_1, P_2)$. □

Recall that in information fusion [4], the Chernoff distribution P^* defines the middle distribution obtained after merging the two distributions P_1 and P_2.

B. A simple geodesic bisection search

To approximate the Chernoff distribution $P^* = E_{12}^{(\alpha^*)}$, we bisect the exponential mixture geodesic $G_e(P, Q)$. Using the θ-coordinate system, let initially $\alpha \in [\alpha_-, \alpha_+]$ with $\alpha_- = 0$ and $\alpha_+ = 1$. Compute the θ-midpoint $\theta = \theta_1 + \alpha'(\theta_2 - \theta_1)$ with $\alpha' = \frac{\alpha_- + \alpha_+}{2}$. If $B_F(\theta_1 : \theta) < B_F(\theta_2 : \theta)$ recurse on interval $[\alpha', \alpha_+]$, otherwise recurse on interval $[\alpha_-, \alpha']$. At each stage we split the α-range in the θ-coordinate system thus yielding convergence to α^*. The bisection search can also be implemented using the dual η-coordinate system. Let initially $\beta \in [\beta_-, \beta_+]$ with $\beta_- = 0$ and $\beta_+ = 1$. We compute the η-midpoint $\beta' = \frac{\beta_- + \beta_+}{2}$ and let $\theta = \nabla F^*((1 - \beta)\eta_1 + \beta\eta_2)$. If $B_F(\theta_1 : \theta) < B_F(\theta_2 : \theta)$ recurse on interval $[\beta', \beta_+]$, otherwise recurse on interval $[\beta_-, \beta']$. We can also alternate between those dual coordinate systems, yielding a primal-dual-coordinate exponential mixture geodesic bisection search.

IV. Three novel points and divergences

The Chernoff point P^* (or Chernoff distribution) can also be interpreted as defining the "middle" of the e-geodesic:

$$\max_{\theta_{12}^{(\alpha^*)} \in \Theta} \min\{\mathrm{KL}(P_{\theta_{12}^{(\alpha^*)}} : P_1), \mathrm{KL}(P_{\theta_{12}^{(\alpha^*)}} : P_2)\}, \tag{15}$$

where the notion of middle is defined as the point that realizes the equi-divergence from the midpoint to the extremities. A different notion of half-way can be obtained by taking the equi-divergence from the extremities to the midpoint:

$$\max_{\theta_{12}^{(\beta)} \in \Theta} \min\{\mathrm{KL}(P_1 : P_{\theta_{12}^{(\beta)}}), \mathrm{KL}(P_2 : P_{\theta_{12}^{(\beta)}})\}. \tag{16}$$

This half-way distribution $P_{\theta_{12}^{(\beta^*)}}$ is geometrically interpreted as the unique intersection point $P_2^* = G_e(P_1, P_2) \cap \mathrm{Bi}_e(P_1, P_2)$ of the e-geodesic with the e-bisector:

$$\mathrm{Bi}_e(P_1, P_2) : \{P \in \mathcal{M} |\ \mathrm{KL}(P_1 : P) = \mathrm{KL}(P_2 : P)\}, \tag{17}$$

that is expressed in the θ-coordinate system as:

$$\{\theta\ |\langle \theta, \eta_2 - \eta_1 \rangle + F(\theta_2) - F(\theta_1) + \langle \eta_1, \theta_1 \rangle - \langle \eta_2, \theta_2 \rangle = 0\} \tag{18}$$

In the θ-coordinate system, this Chernoff point $P_2^* = P_{\theta_{12}^{(\beta^*)}}$ of type II is the intersection of a line segment with a hyperplane, and can therefore be computed exactly. Similarly, we can also cut the m-geodesic with the equi-divergence principle, yielding thus a total of *four* particular points:

$$P_1^* = G_e(P_1, P_2) \cap \mathrm{Bi}_m(P_1, P_2) = P_{\theta_{12}^{(\alpha^*)}}, \tag{19}$$

$$P_2^* = G_e(P_1, P_2) \cap \mathrm{Bi}_e(P_1, P_2) = P_{\theta_{12}^{(\beta^*)}}, \tag{20}$$

$$P_3^* = G_m(P_1, P_2) \cap \mathrm{Bi}_m(P_1, P_2) = P_{\eta_{12}^{(\gamma^*)}}, \tag{21}$$

$$P_4^* = G_m(P_1, P_2) \cap \mathrm{Bi}_e(P_1, P_2) = P_{\eta_{12}^{(\delta^*)}}. \tag{22}$$

The following theorem states that two of those points (and associated symmetric distance) can always be calculated in closed-form:

Theorem 3: Let $P_1 = P_{\theta_1}$ and $P_2 = P_{\theta_2}$ be two distributions of the same exponential family. Chernoff distributions $P_2^* = P_{\theta_{12}^{(\beta^*)}}$ (type II) and $P_3^* = P_{\eta_{12}^{(\gamma^*)}}$ (type III) can be exactly computed, with $\beta^* = \frac{B_F(\theta_2 : \theta_1)}{\langle \Delta\theta, \Delta\eta \rangle}$ and $\gamma^* = \frac{B_F(\theta_1 : \theta_2)}{\langle \Delta\theta, \Delta\eta \rangle}$, where $\Delta\theta = \theta_2 - \theta_1$ and $\Delta\eta = \eta_2 - \eta_1$.

c) Proof:: Points P_2^* and P_3^* (Chernoff middle distributions of type II and III) are intersection (of a straight line geodesic with a hyperplane either in the θ-coordinate or η-coordinate systems) , and thus admit closed-form expressions.[5] Wlog., consider $P_2^* \in G_e(P_1, P_2)$ parameterized by $\theta_{12}^{(\beta)} = \theta_2 - \beta\Delta\theta$. Plugging $\langle \theta, \Delta\eta \rangle = \langle \theta_2, \Delta\eta \rangle - \beta \langle \Delta\theta, \Delta\eta \rangle$ in bisector equation 18, we find that $-\beta^* \langle \Delta\theta, \Delta\eta \rangle + F(\theta_2) - F(\theta_1) - \langle \theta_2 - \theta_1, \eta_1 \rangle = 0$. That is, $\beta^* = \frac{B_F(\theta_2 : \theta_1)}{\langle \Delta\theta, \Delta\eta \rangle}$. Note that $\beta^* \leq 1$ since $B_F(\theta_2 : \theta_1) \leq B_F(\theta_1 : \theta_2) + B_F(\theta_2 : \theta_1) = \langle \Delta\theta, \Delta\eta \rangle$. $\square$

[5] http://www.informationgeometry.org/ChernoffInformation/ for Java codes.

Chernoff distributions of type I and IV can be arbitrarily approximated using geodesic bisection searches (Section III-B). For 1D exponential families, since geodesics G_e and G_m coincide, we have only two distinct Chernoff points ($P_1^* = P_3^*$ and $P_2^* = P_4^*$). Note that for the special case of the isotropic Gaussian family (i.e., fixed covariance matrix with $F(x) = \frac{1}{2}\langle x, x\rangle$), those four Chernoff points coincide since the θ-coordinate and η-coordinate systems are equivalent.

V. Conclusion

We characterized geometrically the optimal Chernoff distribution (inducing the Chernoff distance between two members of the same exponential family) in the statistical manifold as the *unique intersection point* of the exponential mixture geodesic with the mixture bisector. It follows an exact analytic expression for the Chernoff distance for singly-parametric distributions, or an efficient geodesic bisection algorithm for higher-order exponential families. Furthermore, we defined three novel "Chernoff points" as the intersection of exponential/mixture geodesics with exponential/mixture bisectors. Interestingly, two of those points can always be exactly calculated using closed-form formula.

References

[1] L. D. Brown, *Fundamentals of statistical exponential families: with applications in statistical decision theory*. Hayworth, CA, USA: Institute of Mathematical Statistics, 1986.

[2] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Annals of Mathematical Statistics*, vol. 23, pp. 493–507, 1952.

[3] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York, NY, USA: Wiley-Interscience, 1991.

[4] S. J. Julier, "An empirical study into the use of Chernoff information for robust, distributed fusion of Gaussian mixture models," 2006, pp. 1–8.

[5] F. Calderero and F. Marques, "Region merging techniques using information theory statistical measures," *Transactions on Image Processing*, vol. 19, no. 6, pp. 1567–1586, 2010.

[6] S. Konishi, A. L. Yuille, J. M. Coughlan, and S. C. Zhu, "Statistical edge detection: Learning and evaluating edge cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 57–74, 2003.

[7] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of Calcutta Mathematical Society*, vol. 35, pp. 99–110, 1943.

[8] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Transactions on Communications*, vol. 15, no. 1, pp. 52–60, 1967.

[9] S. Amari and H. Nagaoka, *Methods of Information Geometry*, A. M. Society, Ed. Oxford University Press, 2000.

[10] V. Garcia, F. Nielsen, and R. Nock, "Hierarchical Gaussian mixture model," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 4070–4073.

[11] D. C. Brody, "A note on exponential families of distributions," *Journal of Physics A Mathematical General*, vol. 40, p. 691, 2007.

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

[13] R. T. Rockafeller, *Convex Analysis*. Princeton University Press, 1970.

[14] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.

[15] D. H. Johnson and S. Sinanovic, "Symmetrizing the Kullback-Leibler distance," *Technical report*, 2001.

[16] J.-D. Boissonnat, F. Nielsen, and R. Nock, "Bregman Voronoi diagrams," *Discrete & Computational Geometry*, vol. 44, no. 2, pp. 281–307, 2010.