Article

Revisiting Chernoff Information with Likelihood Ratio Exponential Families

Frank Nielsen [ID]

Sony Computer Science Laboratories, Tokyo 141-0022, Japan; frank.nielsen.x@gmail.com

Abstract: The Chernoff information between two probability measures is a statistical divergence measuring their deviation defined as their maximally skewed Bhattacharyya distance. Although the Chernoff information was originally introduced for bounding the Bayes error in statistical hypothesis testing, the divergence found many other applications due to its empirical robustness property found in applications ranging from information fusion to quantum information. From the viewpoint of information theory, the Chernoff information can also be interpreted as a minmax symmetrization of the Kullback–Leibler divergence. In this paper, we first revisit the Chernoff information between two densities of a measurable Lebesgue space by considering the exponential families induced by their geometric mixtures: The so-called likelihood ratio exponential families. Second, we show how to (i) solve exactly the Chernoff information between any two univariate Gaussian distributions or get a closed-form formula using symbolic computing, (ii) report a closed-form formula of the Chernoff information of centered Gaussians with scaled covariance matrices and (iii) use a fast numerical scheme to approximate the Chernoff information between any two multivariate Gaussian distributions.

Keywords: Chernoff information; Chernoff–Bregman divergence; Chernoff–Jensen divergence; Chernoff information distribution; Kullback–Leibler divergence; Bhattacharyya distance; Rényi α-divergences; regular/steep exponential family; Gaussian measures; exponential arc; information geometry; L^1 measurable space; Bregman divergence; affine group

1. Introduction

1.1. Chernoff Information: Definition and Related Statistical Divergences

Let $(\mathcal{X}, \mathcal{A})$ denote a measurable space [1] with sample space $\mathcal{X}$ and finite σ-algebra $\mathcal{A}$ of events. A measure P is absolutely continuous with respect to another measure Q if $P(A) = 0$ whenever $Q(A) = 0$: P is said dominated by Q and written notationally for short as $P \ll Q$. We shall write $P \not\ll Q$ when P is not dominated by Q. When $P \ll Q$, we denote by $\frac{dP}{dQ}$ the Radon–Nikodym density [1] of P with respect to Q.

The Chernoff information [2], also called Chernoff information number [3,4] or the Chernoff divergence [5,6], is the following symmetric measure of dissimilarity (see Appendix A for some background on statistical divergences) between any two comparable probability measures P and Q dominated by μ:

$$D_C[P, Q] := \max_{\alpha \in (0,1)} -\log \rho_\alpha[P : Q] = D_C[Q, P], \tag{1}$$

$$\rho_\alpha[P : Q] := \int p^\alpha q^{1-\alpha} d\mu = \rho_{1-\alpha}[Q : P], \tag{2}$$

is the α-skewed Bhattacharyya affinity coefficient [7] (a coefficient measuring the similarity of two densities). In the remainder, we shall use the following conventions: When a (dis)similarity is asymmetric (e.g., $\rho_\alpha[P : Q]$), we use the colon notation ":" to separate its arguments. When the (dis)similarity is symmetric (e.g., $D_C[P, Q]$), we use the comma notation "," to separate its arguments.

The α-skewed Bhattacharyya coefficients are always upper bounded by 1 and are strictly greater than zero for non-empty intersecting support (non-singular PMs):

$$0 < \rho_\alpha[P : Q] \leq 1.$$

A proof can be obtained by applying Hölder's inequality (see also Appendix A for an alternative proof).

Since the affinity coefficient $\rho_\alpha[P : Q]$ does not depend on the underlying dominating measure μ [4], we shall write $D_C[p, q]$ instead of $D_C[P, Q]$ in the reminder.

Let $D_{B,\alpha}[p : q]$ denote the α-skewed Bhattacharyya distance [7,8]:

$$D_{B,\alpha}[p : q] := -\log \rho_\alpha[P : Q] = D_{B,1-\alpha}[q : p], \tag{3}$$

The α-skewed Bhattacharyya distances are not metric distances since they can be asymmetric and do not satisfy the triangle inequality even when $\alpha = \frac{1}{2}$.

Thus, the Chernoff information is defined as the maximal skewed Bhattacharyya distance:

$$D_C[p, q] = \max_{\alpha \in (0,1)} D_{B,\alpha}[p : q]. \tag{4}$$

Grünwald [9,10] called the skewed Bhattacharyya coefficients and distances the α-Rényi affinity and the unnormalized Rényi divergence, respectively, (see Section 19.6 of [9]) since the Rényi divergence [11,12] is defined by

$$D_{R,\alpha}[P : Q] = \frac{1}{\alpha - 1} \log \int p^\alpha q^{1-\alpha} \mathrm{d}\mu = \frac{1}{1-\alpha} D_{B,\alpha}[P : Q]. \tag{5}$$

Thus $D_{B,\alpha}[P : Q] = (1 - \alpha) D_{R,\alpha}[P : Q]$ can be interpreted as the unnormalized Rényi divergence in [9]. However, let us notice that the Rényi α-divergences are defined in general for a wider range $\alpha \in [0, \infty]\backslash\{1\}$ with $\lim_{\alpha \to 1} D_{R,\alpha}[P : Q] = D_{\mathrm{KL}}[P : Q]$ but the skew Bhattacharyya distances are defined for $\alpha \in (0, 1)$ in general.

The Chernoff information was originally introduced to upper bound the probability error of misclassification in Bayesian binary hypothesis testing [2] where the optimal skewing parameter α^* such that $D_C[p, q] = D_{B,\alpha^*}[p : q]$ is referred to in the statistical literature as the Chernoff error exponent [13–15] or Chernoff exponent [16,17] for short. The Chernoff information has found many other fruitful applications beyond its original statistical hypothesis testing scope like in computer vision [18], information fusion [19], time-series clustering [20], and more generally in machine learning [21] (just to cite a few use cases). It has been observed empirically that the Chernoff information exhibits superior robustness [22] compared to the Kullback–Leibler divergence in distributed fusion of Gaussian Mixtures Models [19] (GMMs) or in target detection in radar sensor network [23]. The Chernoff information has also been used for analysis deepfake detection performance of Generative Adversarial Networks [22] (GANs).

1.2. Prior Work and Contributions

The Chernoff information between any two categorical distributions (multinomial distributions with one trial also called "multinoulli" since they are extensions of the Bernoulli distributions) has been very well-studied and described in many reference textbooks of information theory or statistics (e.g., see Section 12.9 of [13]). The Chernoff information between two probability distributions of an exponential family was considered from the viewpoint of information geometry in [24], and in the general case from the viewpoint of unnormalized Rényi divergences in [11] (Theorem 32). By replacing the weighted geometric mean in the definition of the Bhattacharyya coefficient ρ_α of Equation (2) by an arbitrary weighted mean, the generalized Bhattacharyya coefficient and its associated divergences including the Chernoff information was generalized in [25]. The geometry of the Chernoff error exponent was studied in [26,27] when dealing with a finite set of mutually absolutely

probability distributions $P_1, \ldots, P_n$. In this case, the Chernoff information amounts to the minimum pairwise Chernoff information of the probability distributions [28]:

$$D_C[P_1, \ldots, P_n] := \min_{i \in \{1,\ldots,n\} \neq j \in \{1,\ldots,n\}} D_C[P_i, P_j].$$

We summarize our contributions as follows: In Section 2, we study the Chernoff information between two given mutually non-singular probability measures P and Q by considering their "exponential arc" [29] as a special 1D exponential family termed a Likelihood Ratio Exponential Family (LREF) in [10]. We show that the optimal skewing value (Chernoff exponent) defining their Chernoff information is unique (Proposition 1) and can be characterized geometrically on the Banach vector space $L^1(\mu)$ of equivalence classes of measurable functions (i.e., two functions f_1 and f_2 are said equivalent in $L^1(\mu)$ if they are equal μ-almost everywhere, abbreviated as μ-a.e. in the remainder) for which their absolute value is Lebesgue integrable (Proposition 4). This geometric characterization allows us to design a generic dichotomic search algorithm (Algorithm 1) to approximate the Chernoff optimal skewing parameter, generalizing the prior work [24]. When P and Q belong to a same exponential family, we recover in Section 3 the results of [24]. This geometric characterization also allows us to reinterpret the Chernoff information as a minmax symmetrization of the Kullback–Leibler divergence, and we define by analogy the forward and reverse Chernoff–Bregman divergences in Section 4 (Definition 2). In Section 5, we consider the Chernoff information between Gaussian distributions: We show that the optimality condition for the Chernoff information between univariate Gaussian distributions can be solved exactly and report a closed-form formula for the Chernoff information between any two univariate Gaussian distributions (Proposition 10). For multivariate Gaussian distributions, we show how to implement the dichotomic search algorithms to approximate the Chernoff information, and report a closed-form formula for the Chernoff information between two centered multivariate Gaussian distributions with scaled covariance matrices (Proposition 11). Finally, we conclude in Section 7.

2. Chernoff Information from the Viewpoint of Likelihood Ratio Exponential Families
2.1. LREFs and the Chernoff Information

Recall that $L^1(\mu)$ denotes the Lebesgue vector space of measurable functions f such that $\int_{\mathcal{X}} |f| \mathrm{d}\mu < \infty$. Given two prescribed densities p and q of $L^1(\mu)$, consider building a uniparametric exponential family [30] $\mathcal{E}_{pq}$ which consists of the weighted geometric mixtures of p and q:

$$\mathcal{E}_{pq} := \left\{ (pq)_\alpha^G(x) := \frac{p(x)^\alpha q(x)^{1-\alpha}}{Z_{pq}(\alpha)} \; : \; \alpha \in \Theta \right\}, \tag{6}$$

where

$$Z_{pq}(\alpha) = \int_{\mathcal{X}} p(x)^\alpha q(x)^{1-\alpha} \mathrm{d}\mu(x) = \rho_\alpha[p:q] \tag{7}$$

denotes the normalizer (or partition function) of the geometric mixture

$$(pq)_\alpha^G(x) \propto p(x)^\alpha q(x)^{1-\alpha}$$

so that $\int_{\mathcal{X}} (pq)_\alpha^G \mathrm{d}\mu = 1$. Parameter space Θ is defined as the set of α values which yieds convergence of the definite integral $Z_{pq}(\alpha)$:

$$\Theta := \{ \alpha \in \mathbb{R} \; : \; Z_{pq}(\alpha) < \infty \}. \tag{8}$$

Let us express the density $(pq)_\alpha^G$ in the canonical form $(*)$ of exponential families [30]:

$$(pq)_\alpha^G(x) = \exp\left(\alpha \log \frac{p(x)}{q(x)} - \log Z_{pq}(\alpha)\right) q(x), \tag{9}$$

$$\overset{*}{=:} \exp\left(\alpha t(x) - F_{pq}(\alpha) + k(x)\right). \tag{10}$$

It follows from this decomposition that $\alpha \in \Theta \subset \mathbb{R}$ is the scalar natural parameter, $t(x) = \log \frac{p(x)}{q(x)}$ denotes the sufficient statistic (minimal when $p(x) \neq q(x)$ μ-a.e.), $k(x) = \log q(x)$ is an auxiliary carrier term wrt. measure μ (i.e., measure $d\nu(x) = q(x)d\mu(x)$), and

$$F_{pq}(\alpha) = \log Z_{pq}(\alpha) = -D_{B,\alpha}[p:q] < 0 \tag{11}$$

is the log-normalizer (or log-partition or cumulant function). Since the sufficient statistic is the logarithm of the likelihood ratio of $p(x)$ and $q(x)$, Grünwald [9] (Section 19.6) termed $\mathcal{E}_{pq}$ a Likelihood Ratio Exponential Family (LREF). See also [31] for applications of LREFs to Markov chain Monte Carlo (McMC) methods.

We have $p = (pq)_1^G$ and $q = (pq)_0^G$. Thus, let $\alpha_p = 1$ and $\alpha_q = 0$, and let us interpret geometrically $\{(pq)_\alpha^G, \alpha \in \Theta\}$ as a maximal exponential arc [29,32,33] where $\Theta \subseteq \mathbb{R}$ is an interval. We denote by $\mathcal{E}_{\overline{pq}}$ the open exponential arc with extremities p and q.

Since the log-normalizers $F(\theta)$ of exponential families are always strictly convex and real analytic [30] (i.e., $F(\theta) \in C^\omega(\mathbb{R})$), we deduce that $D_{B,\alpha}[p:q] = -F_{pq}(\alpha)$ is strictly concave and real analytic. Moreover, we have $D_{B,0}[p:q] = D_{B,1}[p:q] = 0$. Hence, the Chernoff optimal skewing parameter α^* is unique when $p \neq q$ μ-a.e., and we get the Chernoff information calculated as

$$D_C[p:q] = D_{B,\alpha^*}(p,q).$$

See Figure 1 for a plot of the strictly concave function $D_{B,\alpha}[p:q]$ and the strictly convex function $F_{pq}(\alpha) = -D_{B,\alpha}[p:q]$ when $p = p_{0,1}$ is the standard normal density and $q = p_{1,2}$ is a normal density of mean 1 and variance 2.
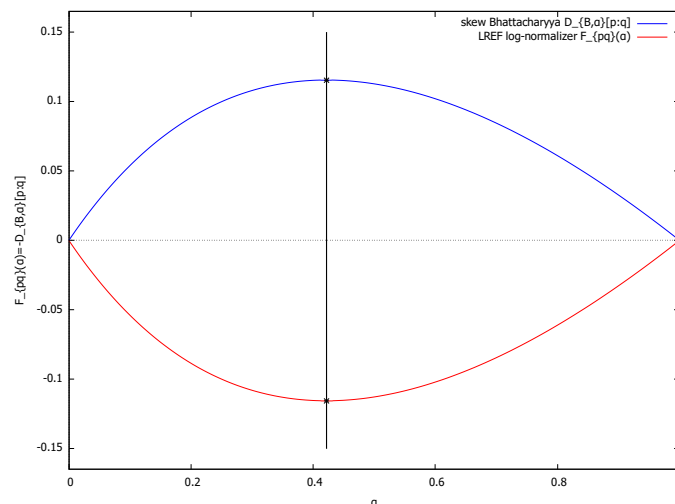

Figure 1. Plot of the Bhattacharryya distance $D_{B,\alpha}(p:q)$ (strictly concave, displayed in blue) and the log-normalizer $F_{pq}(\alpha)$ of the induced LREF $\mathcal{E}_{pq}$ (strictly convex, displayed in red) for two univariate normal densities $p = p_{0,1}$ (standard normal) and $q = p_{1,2}$: The curves $D_{B,\alpha}(p:q) = -F_{pq}(\alpha)$ are horizontally mirror symmetric to each others. The Chernoff information optimal skewing value α^* between these two univariate normal distributions can be calculated exactly in closed-form, see Section 5.2 (approximated numerically here for plotting the vertical grey line by $\alpha^* \approx 0.4215580558605244$).

Consider the *full* natural parameter space Θ_{pq} of $\mathcal{E}_{pq}$:

$$\Theta_{pq} = \{\alpha \in \mathbb{R} \,:\, \rho_\alpha(p:q) < +\infty\}. \tag{12}$$

The natural parameter space Θ_{pq} is always convex [30] and since $\rho_0(p:q) = \rho_1(p:q) = 1$, we necessarily have $(0,1) \in \Theta_{pq}$ but not necessarily $[0,1] \in \Theta_{pq}$ as detailed in the following remark:

Remark 1. *In order to be an exponential family, the densities $(pq)_\alpha^G$ shall have the same coinciding support for all values of α belonging to the natural parameter space. The support of the geometric mixture density $(pq)_\alpha^G$ is*

$$\mathrm{supp}\left((pq)_\alpha^G\right) = \begin{cases} \mathrm{supp}(p) \cap \mathrm{supp}(q), & \alpha \in \Theta_{pq} \backslash \{0,1\} \\ \mathrm{supp}(p), & \alpha = 1 \\ \mathrm{supp}(q), & \alpha = 0. \end{cases}$$

This condition is trivially satisfied when the supports of p and q coincide, and therefore $[0,1] \subset \Theta_{pq}$ in that case. Otherwise, we may consider the common support $\mathcal{X}_{pq} = \mathrm{supp}(p) \cap \mathrm{supp}(q)$ for $\alpha \in (0,1)$. In this latter case, we are poised to restrict the natural parameter space to $\Theta_{pq} = (0,1)$ even if $\rho_\alpha(p:q) < \infty$ for some α outside that range.

To emphasize that α^* depends on p and q, we shall use the notation $\alpha^*(p:q)$ whenever necessary. We have $\alpha^*(q:p) = 1 - \alpha^*(p:q)$, and since $D_{B,\alpha}(p:q) = D_{B,1-\alpha}(q:p)$, and we check that

$$D_C[p,q] = D_{B,\alpha^*(p:q)}(p:q) = D_{B,\alpha^*(q:p)}(q:p) = D_C[q,p].$$

Thus the skewing value $\alpha^*(q:p)$ may be called the conjugate Chernoff exponent (i.e., depends on the convention chosen for interpolating on the exponential arc).

However, since the Chernoff information does not satisfy the triangle inequality, it is not a metric distance and the Chernoff information is called a quasi-distance.

Proposition 1 (Uniqueness of the Chernoff information optimal skewing parameter [11,12])**.** *Let P and Q be two probability measures dominated by a positive measure μ with corresponding Radon–Nikodym densities p and q, respectively. The Chernoff information optimal skewing parameter $\alpha^*(p:q)$ is unique when $p \neq q$ μ-almost everywhere, and*

$$D_C[p,q] = D_{B,\alpha^*(p:q)}(p:q) = D_{B,\alpha^*(q:p)}(q:p) = D_C[q,p].$$

When $p = q$ μ-a.e., we have $D_C[p:q] = 0$ and α^* is undefined since it can range in $[0,1]$.

Definition 1. *An exponential family is called regular [30] when the natural parameter space Θ is open, i.e., $\Theta = \Theta^\circ$ where Θ° denotes the interior of Θ (i.e., an open interval).*

Proposition 2 (Finite sided Kullback–Leibler divergences)**.** *When the LREF $\mathcal{E}_{pq}$ is a regular exponential family with natural parameter space $\Theta \supsetneq [0,1]$, both the forward Kullback–Leibler divergence $D_{KL}[p:q]$ and the reverse Kullback–Leibler divergence $D_{KL}[q:p]$ are finite.*

Proof. A reverse parameter divergence $D^*(\theta_1 : \theta_2)$ is a parameter divergence on the swapped parameter order: $D^*(\theta_1 : \theta_2) := D(\theta_2 : \theta_1)$. Similarly, a reverse statistical divergence $D^*[p:q]$ is a statistical divergence on the swapped parameter order: $D^*[p:q] := D[q:p]$. We shall use the result pioneered in [34,35] that the KLD between two densities p_{θ_1} and p_{θ_2} of a regular exponential family $\mathcal{E} = \{p_\theta \,:\, \theta \in \Theta\}$ amounts to

a reverse Bregman divergence $(B_F)^*$ (i.e., a Bregman divergence on swapped parameter order) induced by the log-normalizer of the family:

$$D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] = (B_F)^*(\theta_1 : \theta_2) = B_F(\theta_2 : \theta_1), \tag{13}$$

where B_F is the Bregman divergence defined on domain $D = \mathrm{dom}(F)$ (see Definition 1 of [36]):

$$\begin{aligned} B_F : \quad & D \times \mathrm{ri}(D) \to [0, \infty) \\ & (\theta_1, \theta_2) \mapsto B_F(\theta_1 : \theta_2) = F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2)^\top \nabla F(\theta_2) < +\infty, \end{aligned}$$

where $\mathrm{ri}(D)$ denotes the relative interior of domain D. Bregman divergences are always finite and the only symmetric Bregman divergences are squared Mahalanobis distances [37] (i.e., with corresponding Bregman generators defining quadratic forms).

For completeness, we recall the proof as follows: We have

$$\log \frac{p_{\theta_1}(x)}{p_{\theta_2}(x)} = (\theta_1 - \theta_2)^\top t(x) - F(\theta_1) + F(\theta_2).$$

Thus we get

$$\begin{aligned} D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] &= E_{p_{\theta_1}}\left[\log \frac{p_{\theta_1}}{p_{\theta_2}}\right], \\ &= F(\theta_2) - F(\theta_1) - (\theta_1 - \theta_2)^\top E_{p_{\theta_1}}[t(x)], \end{aligned}$$

using the linearity property of the expectation operator. When $\mathcal{E}$ is regular, we also have $E_{p_{\theta_1}}[t(x)] = \nabla F(\theta_1)$ (see [38]), and therefore we get

$$D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] = F(\theta_2) - F(\theta_1) - (\theta_1 - \theta_2)^\top \nabla F(\theta_1) =: B_F(\theta_2 : \theta_1) = (B_F)^*(\theta_1 : \theta_2).$$

In our LREF setting, we thus have:

$$D_{\mathrm{KL}}[p : q] = (B_F)^*(\alpha_p : \alpha_q) = B_{F_{pq}}(\alpha_q : \alpha_p) = B_{F_{pq}}(0 : 1),$$

and $D_{\mathrm{KL}}[q : p] = B_{F_{pq}}(\alpha_p : \alpha_q) = B_{F_{pq}}(1 : 0)$ where $B_{F_{pq}}(\alpha_1 : \alpha_2)$ denotes the following scalar Bregman divergence:

$$B_{F_{pq}}(\alpha_1 : \alpha_2) = F_{pq}(\alpha_1) - F_{pq}(\alpha_2) - (\alpha_1 - \alpha_2)F'_{pq}(\alpha_2).$$

Since $F_{pq}(0) = F_{pq}(1) = 0$ and $B_{F_{pq}} : \Theta \times \mathrm{ri}(\Theta) \to [0, \infty)$, we have

$$D_{\mathrm{KL}}[p : q] = B_{F_{pq}}(\alpha_q : \alpha_p) = B_{F_{pq}}(0 : 1) = F'_{pq}(1) < \infty.$$

Similarly

$$D_{\mathrm{KL}}[q : p] = B_{F_{pq}}(\alpha_p : \alpha_q) = B_{F_{pq}}(1 : 0) = -F'_{pq}(0) < \infty.$$

Notice that since $B_{F_{pq}}(\alpha_1 : \alpha_2) > 0$, we have $F'_{pq}(1) > 0$ and $F'_{pq}(0) < 0$ when $p \neq q$ μ-almost everywhere (a.e.). Moreover, since $F_{pq}(\alpha)$ is strictly convex, $F'_{pq}(\alpha)$ is strictly monotonically increasing, and therefore there exists a unique $\alpha^* \in (0, 1)$ such that $F'_{pq}(\alpha^*) = 0$. $\quad\square$

Example 1. *When p and q belongs to a same regular exponential family $\mathcal{E}$ (e.g., p and q are two normal densities), their sided KLDs [37] are both finite. The LREF induced by two Cauchy distributions p_{l_1, s_1} and p_{l_2, s_2} is such that $[0, 1] \subset \Theta$ since the skewed Bhattacharyya distance is defined and finite for $\alpha \in \mathbb{R}$ [39]. Therefore the KLDs between two Cauchy distributions are always finite [39], see the closed-form formula in [40].*

Remark 2. *If $0 \in \Theta^\circ$, then $B_{F_{pq}}(1:0) < \infty$ and therefore $D_{\mathrm{KL}}[q:p] < \infty$. Since the KLD between a standard Cauchy distribution p and a standard normal distribution q is $+\infty$, we deduce that $D_{\mathrm{KL}}[p:q] \neq B_{F_{pq}}(0:1)$, and therefore $1 \notin \Theta^\circ$. Similarly, when $1 \in \Theta^\circ$, we have $B_{F_{pq}}(0:1) < \infty$ and therefore $D_{\mathrm{KL}}[p:q] < \infty$.*

Proposition 3 (Chernoff information expressed as KLDs). *(see also Theorem 32 of [11]) We have at the Chernoff information optimal skewing value $\alpha^* \in (0,1)$ the following identities:*

$$D_{\mathrm{C}}[p:q] = D_{\mathrm{KL}}[(pq)_{\alpha^*}^G : p] = D_{\mathrm{KL}}[(pq)_{\alpha^*}^G : q].$$

Proof. Since the skewed Bhattacharyya distance between two densities p_{θ_1} and p_{θ_2} of an exponential family with log-normalizer F amounts to a skew Jensen divergence for the log-normalizer [8,41], we have:

$$D_{B,\alpha}(p_{\theta_1} : p_{\theta_2}) = J_{F,\alpha}(\theta_1 : \theta_2),$$

where the skew Jensen divergence [42] is given by

$$J_{F,\alpha}(\theta_1 : \theta_2) = \alpha F(\theta_1) + (1-\alpha)F(\theta_2) - F(\alpha\theta_1 + (1-\alpha)\theta_2).$$

In the setting of the LREF, we have

$$\begin{aligned}
D_{B,\alpha}((pq)_{\alpha_1}^G : (pq)_{\alpha_2}^G) &= J_{F_{pq},\alpha}(\alpha_1 : \alpha_2), \\
&= \alpha F_{pq}(\alpha_1) + (1-\alpha)F_{pq}(\alpha_2) - F_{pq}(\alpha\alpha_1 + (1-\alpha)\alpha_2).
\end{aligned}$$

At the optimal value α^*, we have $F'_{pq}(\alpha^*) = 0$. Since $D_{\mathrm{KL}}[(pq)_{\alpha^*}^G : p] = B_{F_{pq}}(1:\alpha^*) = -F(\alpha^*)$ and $D_{\mathrm{KL}}[(pq)_{\alpha^*}^G : q] = B_{F_{pq}}(0:\alpha^*) = -F(\alpha^*)$ and $D_{\mathrm{C}}[p:q] = -\log\rho_{\alpha^*}(p:q) = J_{F_{pq},\alpha^*}(1:0) = -F_{pq}(\alpha^*)$, we get

$$D_{\mathrm{C}}[p:q] = D_{\mathrm{KL}}[(pq)_{\alpha^*}^G : p] = D_{\mathrm{KL}}[(pq)_{\alpha^*}^G : q].$$

Figure 2 illustrates the proposition on the plot of the scalar function $F_{pq}(\alpha)$. □
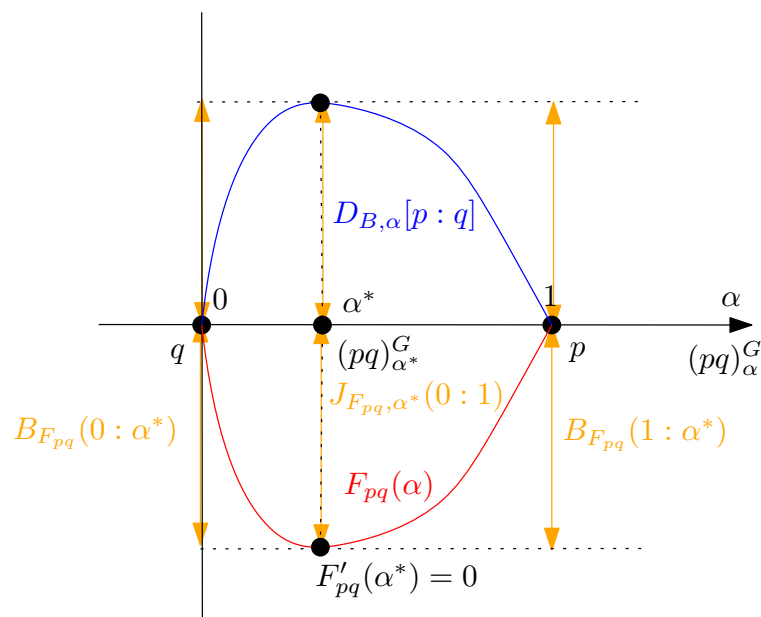

Figure 2. The best unique parameter α^* defining the Chernoff information optimal skewing parameter is found by setting the derivative of the strictly convex function $F_{pq}(\alpha)$ to zero. At the optimal value α^*, we have $D_{\mathrm{C}}[p:q] = D_{\mathrm{KL}}[(pq)_{\alpha^*}^G : p] = D_{\mathrm{KL}}[(pq)_{\alpha^*}^G : q] = -F(\alpha^*) > 0$.

Corollary 1. *The Chernoff information optimal skewing value $\alpha^*(p : q) \in (0, 1)$ can be used to calculate the Chernoff information $D_C[p, q]$ as a Bregman divergence induced by the LREF:*

$$D_C[p : q] = B_{F_{pq}}[1 : \alpha^*] = B_{F_{pq}}[0 : \alpha^*] = J_{F_{pq}, \alpha^*}(1 : 0).$$

In general, the divergence $J_F^C(\theta_1, \theta_2) = \max_{\alpha \in (0,1)} J_{F,\alpha}(\theta_1 : \theta_2)$ is called a Jensen–Chernoff divergence.

Proposition 3 let us interpret the Chernoff information as a special symmetrization of the Kullback–Leibler divergence [43], different from the Jeffreys divergence or the Jensen–Shannon divergence [44]. Indeed, the Chernoff information can be rewritten as

$$D_C[p : q] = \min_{r \in \mathcal{E}_{\overline{pq}}} \{D_{\mathrm{KL}}[r : p], D_{\mathrm{KL}}[r : q]\}. \tag{14}$$

As such, we can interpret the Chernoff information as the radius of a minimum enclosing left-sided Kullback–Leibler ball on the space $L^1(\mu)$. A related concept is the radius [12] of two densities p and q with respect to Rényi divergences of order α (see Equation (2) of [12]):

$$r_\alpha(p, q) := \inf_c \max\{D_{R,\alpha}[p : c], D_{R,\alpha}[q : c]\}.$$

When $\alpha = 1$, the radius is called the Shannon radius [12] since the Rényi divergences of order 1 corresponds to the Kullback–Leibler divergence (relative entropy).

2.2. Geometric Characterization of the Chernoff Information and the Chernoff Information Distribution

Let us term the probability distribution $(PQ)_{\alpha^*}^G \ll \mu$ with corresponding density $(pq)_{\alpha^*}^G$ the Chernoff information distribution to avoid confusion with another concept of Chernoff distributions [45] used in statistics. We can characterize geometrically the Chernoff information distribution $(pq)_{\alpha^*}^G$ on $L^1(\mu)$ as the intersection of a left-sided Kullback–Leibler divergence bisector:

$$\mathrm{Bi}_{\mathrm{KL}}^{\mathrm{left}}(p, q) := \left\{r \in L^1(\mu) \ : \ D_{\mathrm{KL}}[r : p] = D_{\mathrm{KL}}[r : q]\right\}, \tag{15}$$

with an exponential arc [29]

$$\gamma^G(p, q) := \left\{(pq)_\alpha^G \ : \ \alpha \in [0, 1]\right\}. \tag{16}$$

We thus interpret Proposition 3 geometrically by the following proposition (see Figure 3):
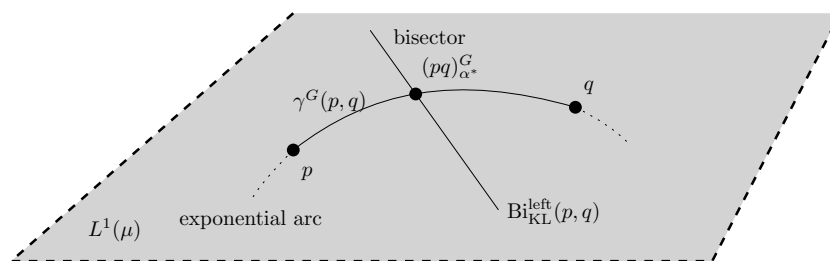

Figure 3. The Chernoff information distribution $(PQ)_{\alpha^*}^G$ with density $(pq)_{\alpha^*}^G$ is obtained as the unique intersection of the exponential arc $\gamma^G(p, q)$ linking density p to density q of $L^1(\mu)$ with the left-sided Kullback–Leibler divergence bisector $\mathrm{Bi}_{\mathrm{KL}}^{\mathrm{left}}(p, q)$ of p and q: $(pq)_{\alpha^*}^G = \gamma^G(p, q) \cap \mathrm{Bi}_{\mathrm{KL}}^{\mathrm{left}}(p, q)$.

Proposition 4 (Geometric characterization of the Chernoff information). *On the vector space $L^1(\mu)$, the Chernoff information distribution is the unique distribution*

$$(pq)_{\alpha^*}^G = \gamma^G(p, q) \cap \mathrm{Bi}_{\mathrm{KL}}^{\mathrm{left}}(p, q).$$

The point $(pq)_{\alpha^*}^G$ has been called the Chernoff point in [24].

Proposition 4 allows us to design a dichotomic search to numerically approximate α^* as reported in pseudo-code in Algorithm 1 (see also the illustration in Figure 4).

Algorithm 1 Dichotomic search for approximating the Chernoff information by approximating the optimal skewing parameter value $\tilde{\alpha} \approx \alpha^*$ and reporting $D_C[p:q] \approx D_{KL}[(pq)_{\tilde{\alpha}}^G : p]$. The search requires $\lceil \log_2 \frac{1}{\epsilon} \rceil$ iterations to guarantee $|\alpha^* - \tilde{\alpha}| \le \epsilon$.

input : Two densities p, q of $L^1(\mu)$, and a numerical precision threshold $\epsilon > 0$
$\alpha_m = 0;$
$\alpha_M = 1;$
while $|\alpha_M - \alpha_m| > \epsilon$ **do**
$\quad \alpha = \frac{\alpha_m + \alpha_M}{2};$
$\quad$ **if** $D_{KL}[(pq)_\alpha^G : p] > D_{KL}[(pq)_\alpha^G : q]$ **then**
$\quad\quad \alpha_m = \alpha;$
$\quad\quad$ `// See Figure 4 for an illustration and Proposition 4`
$\quad$ **end**
$\quad$ **else**
$\quad\quad \alpha_M = \alpha;$
$\quad$ **end**
end
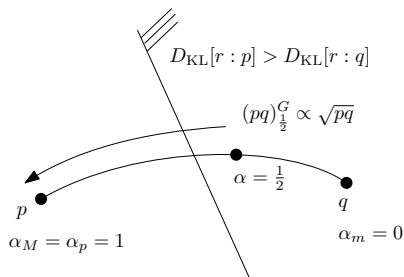**return** $D_{KL}[(pq)_\alpha^G : p];$

Figure 4. Illustration of the dichotomic search for approximating the optimal skewing parameter α^* to within some prescribed numerical precision $\epsilon > 0$.

Remark 3. *We do not need to necessarily handle normalized densities p and q since we have for* $\alpha \in \mathbb{R} \backslash \{0, 1\}$:

$$(pq)_\alpha^G = (\tilde{p}\tilde{q})_\alpha^G,$$

where $p(x) = \frac{\tilde{p}(x)}{Z_p}$ *and* $q(x) = \frac{\tilde{q}(x)}{Z_q}$ *with* $\tilde{p}$ *and* $\tilde{q}$ *denoting the computationally-friendly unnormalized positive densities. This property of geometric mixtures is used in Annealed Importance Sampling [46,47] (AIS), and for designing an asymptotically efficient estimator for computationally-intractable parametric densities [48]* $\tilde{q}_\theta$ *(e.g., distributions learned by Boltzmann machines).*

2.3. Dual Parameterization of LREFs

The densities $(pq)_\alpha^G$ of a LREF can also be parameterized by their dual moment parameter [30] (or mean parameter):

$$\beta = \beta(\alpha) := E_{(pq)_\alpha^G}[t(x)] = E_{(pq)_\alpha^G}\left[\log \frac{p(x)}{q(x)}\right]. \tag{17}$$

When the LREF is regular (and therefore steep [38]), we have $\beta(\alpha) = F_{pq}'(\alpha)$ and $\alpha = F_{pq}^{*'}(\beta)$, where F_{pq}^* denotes the Legendre transform of F_{pq}. At the optimal value α^*, we have $F_{pq}'(\alpha^*) = 0$. Therefore an equivalent condition of optimality is

$$\beta(\alpha^*) = F'_{pq}(\alpha^*) = 0 = E_{(pq)^G_{\alpha^*}}\left[\log\frac{p(x)}{q(x)}\right].$$

Notice that when $[0,1] \subset \Theta^\circ$, we have finite forward and reverse Kullback–Leibler divergences:

- $\alpha = 1$, we have $(pq)^G_1 = p$ and

$$\beta(1) = E_p\left[\log\frac{p(x)}{q(x)}\right] = D_{KL}[p:q] = F'_{pq}(1) > 0.$$

- $\alpha = 0$, we have $(pq)^G_0 = q$ and

$$\beta(0) = E_q\left[\log\frac{p(x)}{q(x)}\right] = -D_{KL}[q:p] = F'_{pq}(0) < 0.$$

Since $F_{pq}(\alpha)$ is strictly convex, we have $F''_{pq}(\alpha) > 0$ and F'_{pq} is strictly increasing with $F'_{pq}(0) = -D_{KL}[q:p] < 0$ and $F'_{pq}(1) = D_{KL}[p:q] > 0$. The value α^* is thus the unique value such that $F'_{pq}(\alpha^*) = 0$.

Proposition 5 (Dual optimality condition for the Chernoff information). *The unique Chernoff information optimal skewing parameter α^* is such that*

$$OC_\alpha : D_{KL}[(pq)^G_{\alpha^*} : p] = D_{KL}[(pq)^G_{\alpha^*} : q] \Leftrightarrow OC_\beta : \beta(\alpha^*) = E_{(pq)^G_{\alpha^*}}\left[\log\frac{p(x)}{q(x)}\right] = 0.$$

One can understand that the Chernoff information is more robust or stable than a skewed Bhattacharrya distance by considering the derivative of the corresponding skewed Bhattacharrya distance. Consider without loss of generality densities p_{θ_1} and p_{θ_2} of a 1D exponential family. Their skewed Bhattacharrya distances amounts to skew Jensen divergences, and we have:

$$J'_{F,\alpha}(\theta_1 : \theta_2) := \frac{d}{d\alpha} J_{F,\alpha}(\theta_1 : \theta_2) = F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2) F'(\alpha\theta_1 + (1-\alpha)\theta_2).$$

Since J_{F,α^*} is by definition maximal, we have $F'(\alpha^*\theta_1 + (1-\alpha^*)\theta_2) = 0$, and therefore $|J'_{F,\alpha}(\theta_1 : \theta_2) - J'_{F,\alpha^*}(\theta_1 : \theta_2)| = |(\theta_1 - \theta_2) F'(\alpha\theta_1 + (1-\alpha)\theta_2)| > 0$. Further assuming without loss of generality that $\theta_2 - \theta_1 = 1$, we get $|J'_{F,\alpha}(\theta_1 : \theta_2) - J'_{F,\alpha^*}(\theta_1 : \theta_2)| = |F'(\alpha\theta_1 + (1-\alpha)\theta_2)| > 0 = F'(\alpha^*\theta_1 + (1-\alpha^*)\theta_2)$.

As a side remark, let us notice that the Fisher information of a likelihood ratio exponential family $\mathcal{E}_{pq}$ is

$$I_{pq}(\alpha) = -E_{(pq)^G_\alpha}[(\log(pq)^G_\alpha)''] = F''_{pq}(\alpha) > 0,$$

and $= F''_{pq}(\alpha) F^{*''}_{pq}(\beta) = 1$.

3. Chernoff Information between Densities of an Exponential Family

3.1. General Case

We shall now consider that the densities p and q (with respect to measure μ) belong to a same exponential family [30]:

$$\mathcal{E} = \left\{ P_\lambda \ : \ \frac{dP_\lambda}{d\mu} = p_\lambda(x) = \exp(\theta(\lambda)^\top t(x) - F(\theta(\lambda))), \quad \lambda \in \Lambda \right\},$$

where $\theta(\lambda)$ denotes the natural parameter associated with the ordinary parameter λ, $t(x)$ the sufficient statistic vector and $F(\theta(\lambda))$ the log-normalizer. When $\theta(\lambda) = \lambda$ and $t(x) = x$,

the exponential family is called a natural exponential family (NEF). The exponential family $\mathcal{E}$ is defined by μ and $t(x)$, hence we may write when necessary $\mathcal{E} = \mathcal{E}_{\mu,t}$.

Example 2. *The set of univariate Gaussian distributions*

$$\mathcal{N} = \{p_{\mu,\sigma^2}(x) \; : \; \lambda = (\mu,\sigma^2) \in \Lambda = \mathbb{R} \times \mathbb{R}_{++}\}$$

forms an exponential family with the following decomposition terms:

$$
\begin{aligned}
\lambda &= (\mu,\sigma^2) \in \Lambda = \mathbb{R} \times \mathbb{R}_{++}, \\
\theta(\lambda) &= \left(\theta_1 = \frac{\mu}{\sigma^2}, \theta_2 = -\frac{1}{2\sigma^2}\right) \in \Theta = \mathbb{R} \times \mathbb{R}_{--}, \\
t(x) &= (x,x^2), \\
F(\theta) &= -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log\left(-\frac{\pi}{\theta_2}\right),
\end{aligned}
$$

where $\mathbb{R}_{++} = \{x \in \mathbb{R} \; : \; x > 0\}$ and $\mathbb{R}_{--} = \{x \in \mathbb{R} \; : \; x < 0\}$ denotes the set of positive real numbers and negative real numbers, respectively. Letting $v = \sigma^2$ be the variance parameter, we get the equivalent natural parameters $\left(\frac{\mu}{v}, -\frac{1}{2v}\right)$. The log-normalizer can be written using the (μ,v)-parameterization as $F(\mu,v) = \frac{1}{2}\log(2\pi v) + \frac{\mu^2}{2v}$ and $\theta = \left(\theta_1 = \frac{\mu}{v}, -\frac{1}{2v}\right)$. See Appendix B for further details concerning this normal exponential family.

Notice that we can check easily that the LREF between two densities of an exponential family forms a 1D sub-exponential family of the exponential family:

$$
\begin{aligned}
p_{\theta_1}(x)^\alpha \, p_{\theta_2}(x)^{1-\alpha} \;\; &\propto \;\; \exp(\langle \alpha\theta_1 + (1-\alpha)\theta_2, t(x)\rangle - \alpha F(\theta_1) - (1-\alpha)F(\theta_2)), \\
&= \;\; p_{\alpha\theta_1 + (1-\alpha)\theta_2}(x)\exp(F(\alpha\theta_1 + (1-\alpha)\theta_2) - \alpha F(\theta_1) - (1-\alpha)F(\theta_2))), \\
&= \;\; p_{\alpha\theta_1 + (1-\alpha)\theta_2}(x)\exp(-J_{F,\alpha}(\theta_1 : \theta_2)),
\end{aligned}
$$

where J_F denote the Jensen divergence induced by F.

The optimal skewing value condition of the Chernoff information between two categorical distributions [13] was extended to densities p_{θ_1} and p_{θ_2} of an exponential family in [24]. The family of categorical distributions with d choices forms an exponential family with natural parameter of dimension $d-1$. Thus, Proposition 7 generalizes the analysis in [13].

Let $p = p_{\theta_1}$ and $q = p_{\theta_2}$. Then we have the property that exponential families are closed under geometric mixtures:

$$(p_{\theta_1} p_{\theta_2})_\alpha^G = p_{\alpha\theta_1 + (1-\alpha)\theta_2}.$$

Since the natural parameter space Θ is convex, we have $\alpha\theta_1 + (1-\alpha)\theta_2 \in \Theta$.

The KLD between two densities p_{θ_1} and p_{θ_2} of a regular exponential family $\mathcal{E}$ amounts to a reverse Bregman divergence for the log-normalizer of $\mathcal{E}$:

$$D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] = B_F(\theta_2 : \theta_1),$$

where $B_F(\theta_2 : \theta_1)$ denotes the Bregman divergence:

$$B_F(\theta_2 : \theta_1) = F(\theta_2) - F(\theta_1) - (\theta_2 - \theta_1)^\top \nabla F(\theta_1).$$

Thus, when the exponential family $\mathcal{E}$ is regular, both the forward and reverse KLD are finite, and we can rewrite Proposition 3 to characterize α^* as follows:

$$\mathrm{OC}_{\mathrm{EF}}: \quad B_F(\theta_1 : \theta_{\alpha^*}) = B_F(\theta_2 : \theta_{\alpha^*}), \tag{18}$$

where $\theta_{\alpha^*} = \alpha^* \theta_1 + (1 - \alpha^*)\theta_2$.

The Legendre–Fenchel transform of $F(\theta)$ yields the convex conjugate

$$F^*(\eta) = \sup_{\theta \in \Theta}\{\theta^\top \eta - F(\theta)\} \tag{19}$$

with $\eta(\theta) = \nabla F(\theta)$. Let $H = \{\eta(\theta) : \theta \in \Theta\}$ denote the dual moment parameter space also called domain of means. The Legendre transform associates to $(\Theta, F(\theta))$ the convex conjugate $(H, F^*(\eta))$. In order for $(H, F^*(\eta))$ to be of the same well-behaved type of $(\Theta, F(\theta))$, we shall consider convex functions $F(\theta)$ which are steep, meaning that their gradient diverges when nearing the boundary bd(Θ) [49] and thus ensures that domain H is also convex. Steep convex functions are said of Legendre-type, and $((\Theta, F(\theta))^*)^* = (\Theta, F(\theta))$ (Moreau biconjugation theorem which shows that the Legendre transform is involutive). For Legendre-type functions, there is a one-to-one mapping between parameters $\theta(\eta)$ and parameters $\eta(\theta)$ as follows:

$$\theta(\eta) = \nabla F^*(\eta) = (\nabla F)^{-1}(\eta), \tag{20}$$

and

$$\eta(\theta) = \nabla F(\theta) = (\nabla F^*)^{-1}(\theta). \tag{21}$$

Exponential families with log-normalizers of Legendre-type are called steep exponential families [30]. All regular exponential families are steep, and the maximum likelihood estimator in steep exponential families exists and is unique [38] (with the likelihood equations corresponding to the method of moments for the sufficient statistics). The set of inverse Gaussian distributions form a non-regular but steep exponential family, and the set of singly truncated normal distributions form a non-regular and non-steep exponential family [50] (but the exponential family of doubly truncated normal distributions is regular and hence steep).

For Legende-type convex generators $F(\theta)$, we can express the Bregman divergence $B_F(\theta_1 : \theta_2)$ using the dual Bregman divergence: $B_F(\theta_1 : \theta_2) = B_{F^*}(\eta_2 : \eta_1)$ since there is a one-to-one correspondence between $\eta = \nabla F(\theta)$ and $\theta = \nabla F^*(\eta)$.

For Legendre-type generators $F(\theta)$, the Bregman divergence $B_F(\theta_1 : \theta_2)$ can be rewritten as the following Fenchel–Young divergence:

$$B_F(\theta_1 : \theta_2) = F(\theta_1) + F^*(\eta_2) - \theta_1^\top \eta_2 := Y_{F,F^*}(\theta_1 : \eta_2).$$

Proposition 6 (KLD between densities of a regular (and steep) exponential family). *The KLD between two densities p_{θ_1} and p_{θ_2} of a regular and steep exponential family can be obtained equivalently as*

$$D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] = B_F(\theta_2 : \theta_1) = Y_{F,F^*}(\theta_2 : \eta_1) = Y_{F^*,F}(\eta_1 : \theta_2) = B_{F^*}(\eta_1 : \eta_2),$$

where $F(\theta)$ and its convex conjugate $F^*(\eta)$ are Legendre-type functions.

Figure 5 illustrates the taxonomy of regularity and steepness of exponential families by a Venn diagram.

It follows that the optimal condition of Equation (18) can be restated as

$$\mathrm{OC_{YF}}: \quad Y_{F,F^*}(\theta_1 : \eta_{\alpha^*}) = Y_{F,F^*}(\theta_2 : \eta_{\alpha^*}), \tag{22}$$

where $\eta_{\alpha^*} = \nabla F^{-1}(\alpha^* \theta_1 + (1 - \alpha^*)\theta_2)$. From the equality of Equation (22), we get the following simplified optimality condition:

$$\mathrm{OC_{SEF}}: \quad (\theta_2 - \theta_1)^\top \eta_{\alpha^*} = F(\theta_2) - F(\theta_1), \tag{23}$$

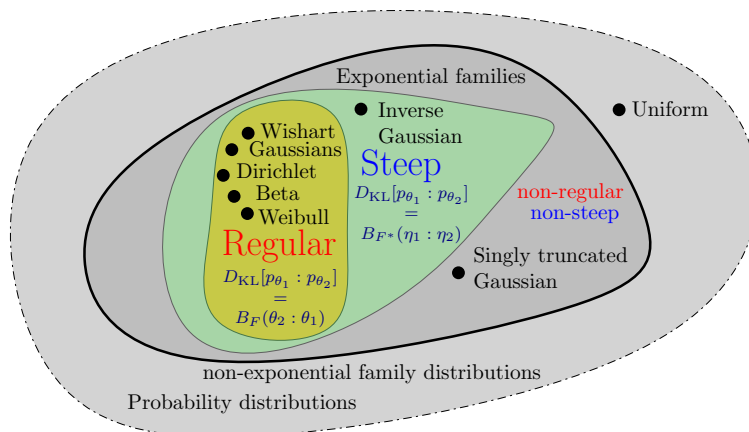where $\eta_{\alpha^*} = \nabla F(\alpha^* \theta_1 + (1 - \alpha^*)\theta_2)$.

Figure 5. Taxonomy of exponential families: Regular (and always steep) or steep (but not necessarily regular). The Kullback–Leibler divergence between two densities of a regular exponential family amounts to dual Bregman divergences.

Remark 4. *We can recover (OC$_{\text{SEF}}$) by instantiating the equivalent condition* $E_{p_{\bar{\theta}_{\alpha*}}}\left[\log \frac{p_{\theta_1}}{p_{\theta_2}}\right] = 0$. *Indeed, since* $\log \frac{p_{\theta_1}}{p_{\theta_2}} = (\theta_1 - \theta_2)^\top t(x) - F(\theta_1) + F(\theta_2)$, *we get*

$$
\begin{aligned}
E_{p_{\bar{\theta}_{\alpha*}}}[(\theta_1 - \theta_2)^\top t(x) - F(\theta_1) + F(\theta_2)] &= 0, \\
(\theta_1 - \theta_2)^\top \bar{\eta}_{\alpha*} &= F(\theta_1) - F(\theta_2).
\end{aligned}
$$

Since the α-skewed Bhattacharyya distance amounts to a α-skewed Jensen divergence [8], we get the Chernoff information as

$$
\begin{aligned}
D_C[p_{\lambda_1} : p_{\lambda_2}] &= J_{F,\alpha*}(\theta(\lambda_1) : \theta(\lambda_2)), \\
&= B_F(\theta_1 : \theta_{\alpha*}) = B_F(\theta_2 : \theta_{\alpha*}),
\end{aligned}
$$

where $J_{F,\alpha}(\theta_1 : \theta_2)$ is the Jensen divergence:

$$
J_{F,\alpha}(\theta_1 : \theta_2) = \alpha F(\theta_1) + (1 - \alpha)F(\theta_2) - F(\alpha \theta_1 + (1 - \alpha)\theta_2).
$$

Notice that we have the induced LREF with log-normalizer expressed as the negative Jensen divergence induced the log-normalizer of $\mathcal{E}$:

$$
F_{p_{\theta_1} p_{\theta_2}}(\alpha) = -\log \rho_\alpha[p_{\theta_1} : p_{\theta_2}] = -J_{F,\alpha}(\theta_1 : \theta_2).
$$

We summarize the result in the following proposition:

Proposition 7. *Let* p_{λ_1} *and* p_{λ_2} *be two densities of a regular exponential family* $\mathcal{E}$ *with natural parameter* $\theta(\lambda)$ *and log-normalizer* $F(\theta)$. *Then the Chernoff information is*

$$
D_C[p_{\lambda_1} : p_{\lambda_2}] = J_{F,\alpha*}(\theta(\lambda_1) : \theta(\lambda_2)) = B_F(\theta_1 : \theta_{\alpha*}) = B_F(\theta_2 : \theta_{\alpha*}),
$$

where $\theta_1 = \theta(\lambda_1)$, $\theta_2 = \theta(\lambda_2)$, *and the optimal skewing parameter* α^* *is unique and satisfies the following optimality condition:*

$$
\text{OC}_{\text{EF}}: \quad (\theta_2 - \theta_1)^\top \eta_{\alpha*} = F(\theta_2) - F(\theta_1), \tag{24}
$$

where $\eta_{\alpha*} = \nabla F(\alpha^* \theta_1 + (1 - \alpha^*)\theta_2) = E_{p_{\alpha*\theta_1 + (1-\alpha*)\theta_2}}[t(x)]$.

Figure 6 illustrates geometrically the Chernoff point [24] which is the geometric mixture $(p_{\theta_1} p_{\theta_2})_{\alpha*}$ induced by two comparable probability measures $P_{\theta_1}, P_{\theta_2} \ll \mu$.
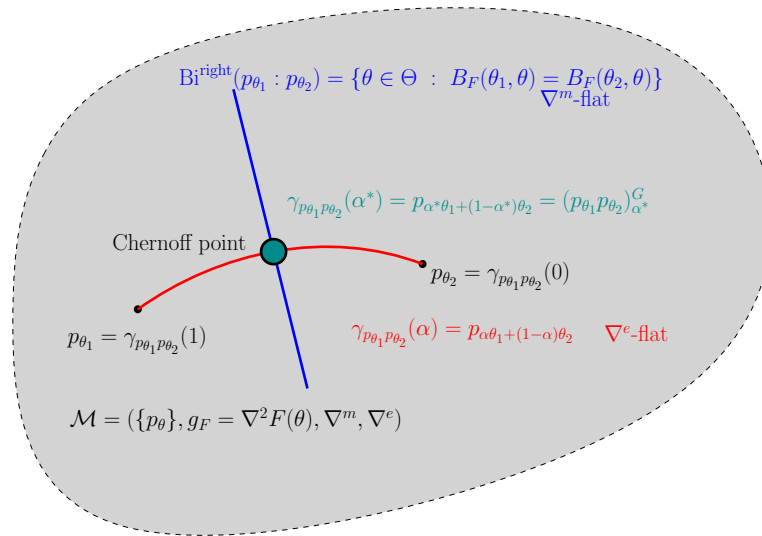
Figure 6. The Chernoff information optimal skewing parameter α^* for two densities p_{θ_1} and p_{θ_2} of some regular exponential family $\mathcal{E}$ inducing an exponential family dually flat manifold $\mathcal{M} = (\{p_\theta\}, g_F = \nabla^2 F(\theta), \nabla^m, \nabla^e)$ is characterized by the intersection of their ∇^e-flat exponential geodesic with their mixture bisector a ∇^m-flat right-sided Bregman bisector.

In information geometry [51], the manifold of densities $M = \{p_\theta : \theta \in \Theta\}$ of this exponential family is a dually flat space [51] $\mathcal{M} = (\{p_\theta\}, g_F(\theta) = \nabla^2 F(\theta), \nabla^m, \nabla^e)$ with respect to the exponential connection ∇^e and the mixture connection ∇^m, where $g_F(\theta)$ is the Fisher information metric expressed in the θ-coordinate system as $\nabla^2 F(\theta)$ (and in the dual moment parameter η as $g_F(\eta) = \nabla^2 F^*(\eta)$). Then the exponential geodesic ∇^e is flat and corresponds to the exponential arc of geometric mixtures when parameterized with the ∇^e-affine coordinate system θ.

The left-sided Kullback–Voronoi bisector:

$$\text{Bi}_{\text{KL}}^{\text{left}}(p_{\theta_1}, p_{\theta_2}) = \{p_\theta : D_{\text{KL}}[p_\theta : p_{\theta_1}] = D_{\text{KL}}[p_\theta : p_{\theta_1}]\}$$

corresponds to a Bregman right-sided bisector [52] and is ∇^m flat (i.e., an affine subspace in the η-coordinate system):

$$\text{Bi}_F^{\text{right}}(\theta_1, \theta_2) = \{\theta \in \Theta : B_F(\theta_1, \theta) = B_F(\theta_2, \theta)\}.$$

The Chernoff information distribution $(p_{\theta_1} p_{\theta_2})_{\alpha^*}^G$ is called the Chernoff point on this exponential family manifold (see Figure 6). Since the Chernoff point is unique and since in general statistical manifolds $(\mathcal{M}, g, \nabla, \nabla^*)$ can be realized by statistical models [53], we deduce the following proposition of interest for information geometry [51]:

Proposition 8. *Let $(\mathcal{M}, g, \nabla, \nabla^*)$ be a dually flat space with corresponding canonical divergence a Bregman divergence B_F. Let $\gamma_{pq}^e(\alpha)$ and $\gamma_{pq}^m(\alpha)$ be a e-geodesic and m-geodesic passing through the points p and q of $\mathcal{M}$, respectively. Let $\text{Bi}^m(p, q)$ and $\text{Bi}^e(p, q)$ be the right-sided ∇^m-flat and left-sided ∇^e-flat Bregman bisectors, respectively. Then the intersection of $\gamma_{pq}^e(\alpha)$ with $\text{Bi}^m(p, q)$ and the intersection of $\gamma_{pq}^m(\alpha)$ with $\text{Bi}^e(p, q)$ are unique. The point $\gamma_{pq}^e(\alpha) \cap \text{Bi}^m(p, q)$ is called the Chernoff point and the point $\gamma_{pq}^m(\alpha) \cap \text{Bi}^e(p, q)$ is termed the reverse or dual Chernoff point.*

3.2. Case of One-Dimensional Parameters

When the exponential family has one-dimensional natural parameter $\alpha \in \Theta \subset \mathbb{R}$, we thus get from OC_{SEF}:

$$\eta_{\alpha^*} = \frac{F(\alpha_2) - F(\alpha_1)}{\alpha_2 - \alpha_1}.$$

That is, α^* can be obtained as the following closed-form formula:

$$\alpha^* = \frac{F'^{-1}\left(\frac{F(\alpha_2)-F(\alpha_1)}{\alpha_2-\alpha_1}\right) - \alpha_2}{\alpha_1 - \alpha_2}. \tag{25}$$

Example 3. *Consider the exponential family* $\{p_v(x) \; : \; v > 0\}$ *of 0-centered scale univariate normal distributions with variance* $v = \sigma^2$ *and density*

$$p_v(x) = \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{1}{2}\frac{x^2}{v}\right).$$

The natural parameter corresponding to the sufficient statistic $t(x) = x^2$ *is* $\theta = -\frac{1}{2v}$. *The log-normalizer is* $F(\theta) = \frac{1}{2}\log\frac{\pi}{-\theta}$. *We have* $\eta = F'(\theta) = -\frac{1}{2\theta}$ *and* $F'^{-1}(\eta) = -\frac{1}{2\eta}$. *It follows that*

$$\alpha^*(p_{v_1} : p_{v_2}) = \frac{v_1 \log \frac{v_1}{v_2} - v_1 + v_2}{(v_2 - v_1) \log \frac{v_2}{v_1}}.$$

Let $s = \frac{v_2}{v_1}$. *Then we can rewrite* α^* *as*

$$\alpha^*(p_{v_1} : p_{v_2}) = \frac{s - 1 - \log s}{(s - 1) \log s}.$$

The Chernoff information is $D_C[p_{v_1}, p_{v_2}] = -\log \rho_{\alpha^*}[p_{v_1}, p_{v_2}]$, *with*

$$\rho_\alpha[p_{v_1}, p_{v_2}] = \frac{\sigma_1^{1-\alpha}\sigma_2^\alpha}{\sqrt{(1-\alpha)\sigma_1^2 + \alpha\sigma_2^2}}.$$

This result will be generalized in Proposition 11 to multivariate centered Gaussians with scaled covariance matrices.

For multi-dimensional parameters θ, we may consider the one-dimensional LREF $\mathcal{E}_{p_{\theta_1} p_{\theta_2}}$ induced by p_{θ_1} and p_{θ_2} with $F_{\theta_1,\theta_2}(\alpha) = F((1-\alpha)\theta_1 + \alpha\theta_2)$, and write $F'_{pq}(\alpha)$ as the following directional derivative:

$$\nabla_{\theta_2-\theta_1} F_{\theta_1,\theta_2}(\alpha) \quad := \quad \lim_{\epsilon\to 0} \frac{F(\theta_1 + (\epsilon+\alpha)(\theta_2-\theta_1)) - F(\theta_1 + \alpha(\theta_2-\theta_1))}{\epsilon}, \tag{26}$$

$$= \quad (\theta_2-\theta_1)^\top \nabla F(\theta_1 + \alpha(\theta_2-\theta_1)), \tag{27}$$

using a first-order Taylor expansion. Thus, the optimality condition

$$\mathrm{OC}_{\mathrm{SEF}'}: \quad F'_{\theta_1,\theta_2}(\alpha) = 0$$

amounts to

$$\mathrm{OC}_{\mathrm{SEF}}: \quad (\theta_2-\theta_1)^\top \nabla F(\theta_1 + \alpha^*(\theta_2-\theta_1)) = F(\theta_2) - F(\theta_1). \tag{28}$$

This is equivalent to Equation (8) of [24].

Remark 5. *In general, we may consider multivariate Bregman divergences as univariate Bregman divergences: We have*

$$B_F(\theta_1 : \theta_2) = B_{F_{\theta_1,\theta_2}}(0 : 1), \quad \forall \theta_1, \theta_2 \in \Theta \tag{29}$$

where

$$F_{\theta_1,\theta_2}(u) := F(\theta_1 + u(\theta_2 - \theta_1)). \tag{30}$$

The functions F_{θ_1,θ_2} are 1D Bregman generators (i.e., strictly convex and C^1), and we have the directional derivative

$$
\begin{aligned}
\nabla_{\theta_2-\theta_1} F_{\theta_1,\theta_2}(u) &= \lim_{\epsilon \to 0} \frac{F(\theta_1 + (\epsilon + u)(\theta_2 - \theta_1)) - F(\theta_1 + u(\theta_2 - \theta_1))}{\epsilon}, \\
&= (\theta_2 - \theta_1)^\top \nabla F(\theta_1 + u(\theta_2 - \theta_1)),
\end{aligned}
$$

Since $F_{\theta_1,\theta_2}(0) = F(\theta_1)$, $F_{\theta_1,\theta_2}(1) = F(\theta_2)$, and $F'_{\theta_1,\theta_2}(u) = \nabla_{\theta_2-\theta_1} F_{\theta_1,\theta_2}(u)$, it follows that

$$
\begin{aligned}
B_{F_{\theta_1,\theta_2}}(0:1) &= F_{\theta_1,\theta_2}(0) - F_{\theta_1,\theta_2}(1) - (0-1)\nabla_{\theta_2-\theta_1} F_{\theta_1,\theta_2}(1), \\
&= F(\theta_1) - F(\theta_2) + (\theta_2 - \theta_1)^\top \nabla F(\theta_2) = B_F(\theta_1 : \theta_2).
\end{aligned}
$$

Similarly, we can reparameterize Bregman divergences on a k-dimensional simplex by k-dimensional Bregman divergences.

Remark 6. *Closing the loop: The Chernoff information although obtained from the one-dimensional likelihood ratio exponential family yields as a corollary the general multi-parametric exponential families which as a special instance includes the one-dimensional exponential families (e.g., LREFs!).*

4. Forward and Reverse Chernoff–Bregman Divergences

In this section, we shall define Chernoff-type symmetrizations of Bregman divergences inspired by the study of Chernoff information, and briefly mention applications of these Chernoff–Bregman divergences in information theory.

4.1. Chernoff–Bregman Divergence

Let us define a Chernoff-like symmetrization of Bregman divergences [43] different from the traditional Jeffreys–Bregman symmetrization:

$$
\begin{aligned}
B_F^J(\theta_1 : \theta_2) &= B_F(\theta_1 : \theta_2) + B_F(\theta_2 : \theta_1), \\
&= (\theta_1 - \theta_2)^\top (\nabla F(\theta_1) - \nabla F(\theta_2)),
\end{aligned}
$$

or Jensen–Shannon-type symmetrization [44,54] which yields a Jensen divergence [42]:

$$
\begin{aligned}
B_F^{JS}(\theta_1 : \theta_2) &= \frac{1}{2}\left(B_F\left(\theta_1 : \frac{\theta_1 + \theta_2}{2}\right) + B_F\left(\theta_2 : \frac{\theta_1 + \theta_2}{2}\right) \right), \\
&= \frac{F(\theta_1) + F(\theta_2)}{2} - F\left(\frac{\theta_1 + \theta_2}{2}\right) =: J_F(\theta_1, \theta_2).
\end{aligned}
$$

Definition 2 (Chernoff–Bregman divergence). *Let the Chernoff symmetrization of Bregman divergence $B_F(\theta_1; \theta_2)$ be the forward Chernoff–Bregman divergence $C_F(\theta_1, \theta_2)$ defined by*

$$
C_F(\theta_1, \theta_2) = \max_{\alpha \in (0,1)} J_{F,\alpha}(\theta_1 : \theta_2), \tag{31}
$$

where $J_{F,\alpha}$ is the α-skewed Jensen divergence.

The optimization problem in Equation (31) may be equivalently rewritten [43] as $\min_\theta R$ such that both $B_F(\theta_1 : \theta) \le R$ and $B_F(\theta_2 : \theta) \le R$. Thus, the optimal value of α defines the circumcenter $\theta^* = \alpha\theta_1 + (1-\alpha)\theta_2$ of the minimum enclosing right-sided Bregman sphere [55,56] and the Chernoff–Bregman divergence:

$$
C_F(\theta_1, \theta_2) = \min_\theta \{ B_F(\theta_1 : \theta), B_F(\theta_2 : \theta) \}, \tag{32}
$$

corresponds to the radius of a minimum enclosing Bregman ball. To summarize, this Chernoff symmetrization is a min-max symmetrization, and we have the following identities:

$$
\begin{aligned}
C_F(\theta_1, \theta_2) &= \min_{\theta}\{B_F(\theta_1 : \theta), B_F(\theta_2 : \theta)\}, \\
&= \min_{\theta \in \Theta}\{\alpha B_F(\theta_1 : \theta) + (1 - \alpha)B_F(\theta_2 : \theta)\}, \\
&= \max_{\alpha \in (0,1)}\{\alpha B_F(\theta_1 : \alpha\theta_1 + (1 - \alpha)\theta_2) + (1 - \alpha)B_F(\theta_2 : \alpha\theta_1 + (1 - \alpha)\theta_2)\}, \\
&= \max_{\alpha \in (0,1)} J_{F,\alpha}(\theta_1 : \theta_2).
\end{aligned}
$$

The second identity shows that the Chernoff symmetrization can be interpreted as a variational Jensen–Shannon-type divergence [54].

Notice that in general $C_F(\theta_1, \theta_2) \neq C_{F^*}(\eta_1, \eta_2)$ because the primal and dual geodesics do not coincide. Those geodesics coincide only for symmetric Bregman divergences which are squared Mahalanobis divergences [52].

When $F(\theta) = F_{\text{Shannon}}(\theta) = \sum_{i=1}^{D} \theta_i \log \theta_i$ (discrete Shannon negentropy), the Chernoff–Bregman divergence is related to the capacity of a discrete memoryless channel in information theory [13,43].

Conditions for which $C_F(\theta_1, \theta_2)^a$ (with $a > 0$) becomes a metric have been studied in [43]: For example, $C_{F_{\text{Shannon}}}^{\frac{1}{e}}$ is a metric distance [43] (i.e., $a = \frac{1}{e} \simeq 0.36787944117$). It is also known that the square root of the Chernoff distance between two univariate normal distributions is a metric distance [57].

We can thus use the Bregman generalization of the Badoiu–Clarkson (BC) algorithm [55] (Algorithm 2) to compute an approximation of the smallest enclosing Bregman ball which in turn yields an approximation of the Chernoff–Bregman divergence:

Algorithm 2 Approximating the circumcenter of the Bregman smallest enclosing ball of two parameters θ_1 and θ_2.

---

input : Two parameters θ_1 and θ_2 of Θ, and a number of iterations $T \in \mathbb{N}$
// Initialize circumcenter
$i \leftarrow 0$;
$\theta^{(i)} \leftarrow \frac{\theta_1 + \theta_2}{2}$;
// Repeat T times
while $i < T$ **do**
    // Find which of θ_1 or θ_2 is the farthest parameter to $\theta^{(i)}$
    if $B_F(\theta_1 : \theta^{(i)}) > B_F(\theta_2 : \theta^{(i)})$ **then**
        | $f_i = 1$
    end
    else
        | $f_i = 2$
    end
    // Update the circumcenter by walking on θ-geodesic
    // This update corresponds to walking on the exponential arc $(\theta^{(i-1)}\theta_{f_i})^G$
    $i \leftarrow i + 1$;
    $\theta^{(i)} \leftarrow \frac{i}{i+1}\theta^{(i-1)} + \frac{1}{i+1}\theta_{f_i}$;
end
// Report the approximation of the circumcenter
return $\theta^{(T)}$

Notice that when there are only two points to compute their smallest enclosing Bregman ball, all the arcs $(\theta^{(i-1)}\theta_{f_i})^G$ are sub-arcs of the exponential arc $(\theta_1\theta_2)^G$. See [55] for convergence results of this iterative algorithm. Let us notice that Algorithm 1 approximates α^* while the Bregman BC Algorithm 2 approximates in spirit $D_C(\theta_1, \theta_2)$ (and as a byproduct α^*).

Remark 7. *To compute the farthest point to the current circumcenter with respect to Bregman divergence, we need to find the sign of*

$$B_F(\theta_2 : \theta) - B_F(\theta_1 : \theta) = F(\theta_2) - F(\theta_1) - (\theta_2 - \theta_1)\nabla F(\theta).$$

Thus we need to pre-calculate only once $F(\theta_1)$ and $F(\theta_2)$ which can be costly (e.g., $-\log\det(\Sigma)$ functions need to be calculated only once when approximating the Chernoff information between Gaussians).

4.2. Reverse Chernoff–Bregman Divergence and Universal Coding

Similarly, we may define the reverse Chernoff–Bregman divergence by considering the minimum enclosing left-sided Bregman ball:

$$C_F^R(\theta_1, \theta_2) = \min_\theta \{B_F(\theta : \theta_1), B_F(\theta : \theta_2)\}. \tag{33}$$

Thus the reverse Bregman Chernoff divergence $D_C^R[\theta_1, \theta_2] = R^*$ is the radius of a minimum enclosing left-sided Bregman ball.

This reverse Chernoff–Bregman divergence finds application in universal coding in information theory (chapter 13 of [13], pp. 428–433): Let $\mathcal{X} = \{A_1, \ldots, A_d\}$ be a finite discrete alphabet of d letters, and X be a random variable with probability mass function p on $\mathcal{X}$. Let $p_\lambda(x)$ denote the categorical distribution corresponding to X so that $\Pr(X = A_i) = p_\lambda(A_i)$ with $\lambda = (\lambda^1, \ldots, \lambda^d) \in \mathbb{R}_{++}^d$ and $\sum_{i=1}^d \lambda^i = 1$. The Huffman codeword for $x \in \mathcal{X}$ is of length $l(x) = -\log p(x)$ (ignoring integer ceil rounding), and the expected codeword length of X is thus given by Shannon's entropy $H(X) = -\sum_x p(x)\log p(x)$.

If we code according to a distribution $p_{\lambda'}$ instead of the true distribution p_λ, the code is not optimal, and the redundancy $R(p_\lambda, p_{\lambda'})$ is defined as the difference between the expected lengths of the codewords for $p_{\lambda'}$ and p_λ:

$$R(p_\lambda, p_{\lambda'}) = (-E_{p_\lambda}[\log p_{\lambda'}(x)] - (-E_{p_\lambda}[\log p_\lambda(x)]) = D_{\mathrm{KL}}[p_\lambda : p_{\lambda'}] \geq 0,$$

where D_{KL} is the Kullback–Leibler divergence.

Now, suppose that the true distribution p_λ belong to one of two prescribed distributions that we do not know: $p_\lambda \in \mathcal{P} = \{p_{\lambda_1}, p_{\lambda_2}\}$. Then we seek for the minimax redundancy:

$$R^* = \min_{p_\lambda} \max_{i \in \{1,2\}} D_{\mathrm{KL}}[p_{\lambda_i} : p_\lambda]. \tag{34}$$

The distribution p_{λ^*} achieving the minimax redundancy is the circumcenter of the right-centered KL ball enclosing the distributions $\mathcal{P}$.

Using the natural coordinates $\theta = (\theta_1, \ldots, \theta_D) \in \mathbb{R}^D$ with $\theta_i = \log\frac{\lambda^i}{\lambda^d}$ of the log-normalizer of the categorical distributions (an exponential family of order $D = d - 1$), we end up with calculating the smallest left-sided Bregman enclosing ball for the Bregman generator [58]: $F_{\mathrm{Categorical}}(\theta) = \log(1 + \sum_{i=1}^D \exp\theta_i)$:

$$R^* = \min_{\theta \in \in \mathbb{R}^D} \max_{i \in \{1,2\}} B_{F_{\mathrm{categorical}}}(\theta : \theta_i).$$

This latter minimax problem is unconstrained since $\theta \in \mathbb{R}^D = \mathbb{R}^{d-1}$.

5. Chernoff Information between Gaussian Distributions

5.1. Invariance of Chernoff Information under the Action of the Affine Group

The d-variate Gaussian density $p_\lambda(x)$ with parameter $\lambda = (\lambda_v = \mu, \lambda_M = \Sigma)$ where $\mu \in \mathbb{R}^d$ denotes the mean ($\mu = E_{p_\lambda}[x]$) and Σ is a positive-definite covariance matrix ($\Sigma = \mathrm{Cov}_{p_\lambda}[X]$ for $X \sim p_\lambda$) is given by

$$p_\lambda(x; \lambda) = \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{|\lambda_M|}} \exp\left(-\frac{1}{2}(x - \lambda_v)^\top \lambda_M^{-1}(x - \lambda_v)\right),$$

where $|\cdot|$ denotes the matrix determinant. The set of d-variate Gaussian distributions form a regular (and hence steep) exponential family with natural parameters $\theta(\lambda) = \left(\lambda_M^{-1}\lambda_v, \frac{1}{2}\lambda_M^{-1}\right)$ and sufficient statistics $t(x) = (x, xx^\top)$.

The Bhattacharrya distance between two multivariate Gaussians distributions p_{μ_1, Σ_1} and p_{μ_2, Σ_2} is

$$D_{B,\alpha}[p_{\mu_1, \Sigma_1}, p_{\mu_2, \Sigma_2}] = \frac{1}{2}\left(\alpha\mu_1^\top\Sigma_1^{-1}\mu_1 + (1-\alpha)\mu_2^\top\Sigma_2^{-1}\mu_2 - \mu_\alpha^\top\Sigma_\alpha^{-1}\mu_\alpha + \log\frac{|\Sigma_1|^\alpha|\Sigma_2|^{1-\alpha}}{|\Sigma_\alpha|}\right),$$

where

$$\begin{aligned}
\Sigma_\alpha &= (\alpha\Sigma_1^{-1} + (1-\alpha)\Sigma_2^{-1})^{-1}, \\
\mu_\alpha &= \Sigma_\alpha(\alpha\Sigma_1^{-1}\mu_1 + (1-\alpha)\Sigma_2^{-1}).
\end{aligned}$$

The Gaussian density can be rewritten as a multivariate location-scale family:

$$p_\lambda(x; \lambda) = |\lambda_M|^{-\frac{1}{2}} p_{\text{std}}(\lambda_M^{\frac{1}{2}}(x - \lambda_v)),$$

where

$$p_{\text{std}}(x) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2}x^\top x\right) = p_{(0,I)}$$

denotes the standard multivariate Gaussian distribution. The matrix $\lambda_M^{\frac{1}{2}}$ is the unique symmetric square-root matrix which is positive-definite when λ_M is positive-definite.

Remark 8. *Notice that the product of two symmetric positive-definite matrices P_1 and P_2 may not be symmetric but $P_1^{\frac{1}{2}} P_2 P_1^{\frac{1}{2}}$ is always symmetric positive-definite, and the eigenvalues of $P_1^{\frac{1}{2}} P_2 P_1^{\frac{1}{2}}$ coincides with the eigenvalues of $P_1 P_2$. Hence, we have $\lambda_{\text{sp}}(P_1^{-\frac{1}{2}} P_2 P_1^{-\frac{1}{2}}) = \lambda_{\text{sp}}(P_1^{-1} P_2)$ where $\lambda_{\text{sp}}(M)$ denotes the eigenspectrum of matrix M.*

We may interpret the Gaussian family as obtained by the action of the affine group $\text{Aff}(\mathbb{R}^d) = \mathbb{R}^d \rtimes \text{GL}_d(\mathbb{R})$ on the standard density p_{std}: Let the dot symbol "." denotes the group action. The affine group is equipped with the following (outer) semidirect product:

$$(l_1, A_1).(l_2, A_2) = (l_1 + A_1 l_2, A_1 A_2), \tag{35}$$

and this group can be handled as a matrix group with the following mapping of its elements to matrices:

$$(l, A) \equiv \begin{bmatrix} A & l \\ 0 & 1 \end{bmatrix}.$$

Then we have

$$p_{(\mu, \Sigma)}(x) = (\mu, \Sigma^{-\frac{1}{2}}).p_{\text{std}}(x) = (\mu, \Sigma^{-\frac{1}{2}}).p_{0,I}(x) = p_{(\mu, \Sigma^{-\frac{1}{2}}).(0,I)}(x).$$

We can show the following invariance of the skewed Bhattacharyya divergences:

Proposition 9 (Invariance of the Bhattacharyya divergence and f-divergences under the action of the affine group (Equation (35)))**.** *We have*

$$
\begin{aligned}
D_{B,\alpha}\left[(\mu,\Sigma^{-\frac{1}{2}}).p_{\mu_1,\Sigma_1} : (\mu,\Sigma^{-\frac{1}{2}}).p_{\mu_2,\Sigma_2}\right] \quad &:= \quad D_{B,\alpha}\left[p_{(\mu,\Sigma^{-\frac{1}{2}}).(\mu_1,\Sigma_1)} : p_{(\mu,\Sigma^{-\frac{1}{2}}).(\mu_2,\Sigma_2)}\right], \\
&= \quad D_{B,\alpha}\left[p_{\Sigma^{-\frac{1}{2}}(\mu_1-\mu):\Sigma^{-\frac{1}{2}}\Sigma_1\Sigma^{-\frac{1}{2}}}, p_{\Sigma^{-\frac{1}{2}}(\mu_2-\mu),\Sigma^{-\frac{1}{2}}\Sigma_2\Sigma^{-\frac{1}{2}}}\right], \\
&= \quad D_{B,\alpha}[p_{\mu_1,\Sigma_1} : p_{\mu_2,\Sigma_2}].
\end{aligned}
$$

Proof. The proof follows from the (f,g)-form of Ali and Silvey's divergences [59]. We can express $D_{B,\alpha}[p : q] = g(I_{h_\alpha}[p : q])$ where $h_\alpha(u) = -u^\alpha$ (convex for $\alpha \in (0,1)$) and $g(v) = -\log -v$. Then we rely on the proof of invariance of f-divergences under the action of the affine group (see Proposition 3 of [60] relying on a change of variable in the integral):

$$
I_f[p_{\mu_1,\Sigma_1} : p_{\mu_2,\Sigma_2}] = I_f\left[p_{0,I}, p_{\Sigma_1^{-\frac{1}{2}}(\mu_2-\mu_1):\Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}}\right] = I_f\left[p_{\Sigma_2^{-\frac{1}{2}}(\mu_1-\mu_2),\Sigma_2^{-\frac{1}{2}}\Sigma_1\Sigma_2^{-\frac{1}{2}}} : p_{0,I}\right],
$$

where I denotes the identity matrix. $\square$

Thus, by choosing $(\mu,\Sigma) = (\mu_1,\Sigma_1)$ and $(\mu,\Sigma) = (\mu_2,\Sigma_2)$, we obtain the following corollary:

Corollary 2 (Bhattacharyya divergence from canonical Bhattacharyya divergences). *We have*

$$
D_{B,\alpha}[p_{\mu_1,\Sigma_1} : p_{\mu_2,\Sigma_2}] = D_{B,\alpha}\left[p_{0,I}, p_{\Sigma_1^{-\frac{1}{2}}(\mu_2-\mu_1):\Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}}\right] = D_{B,\alpha}\left[p_{\Sigma_2^{-\frac{1}{2}}(\mu_1-\mu_2):\Sigma_2^{-\frac{1}{2}}\Sigma_1\Sigma_2^{-\frac{1}{2}}}, p_{0,I}\right].
$$

It follows that the Chernoff optimal skewing parameter enjoys the same invariance property:

$$
\alpha^*(p_{\mu_1,\Sigma_1} : p_{\mu_2,\Sigma_2}) = \alpha^*\left(p_{0,I}, p_{\Sigma_1^{-\frac{1}{2}}(\mu_2-\mu_1),\Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}}\right) = \alpha^*\left(p_{\Sigma_2^{-\frac{1}{2}}(\mu_1-\mu_2):\Sigma_2^{-\frac{1}{2}}\Sigma_1\Sigma_2^{-\frac{1}{2}}}, p_{0,I}\right).
$$

As a byproduct, we get the invariance of the Chernoff information under the action of the affine group:

Corollary 3 (Invariance of the Chernoff information under the action of the affine group). *We have:*

$$
D_C[p_{\mu_1,\Sigma_1}, p_{\mu_2,\Sigma_2}] = D_C\left[p_{0,I}, p_{\Sigma_1^{-\frac{1}{2}}(\mu_2-\mu_1),\Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}}\right] = D_C\left[p_{\Sigma_2^{-\frac{1}{2}}(\mu_1-\mu_2),\Sigma_2^{-\frac{1}{2}}\Sigma_1\Sigma_2^{-\frac{1}{2}}}, p_{0,I}\right].
$$

Thus, the formula for the Chernoff information between two Gaussians

$$
D_C(\mu_1,\Sigma_1,\mu_2,\Sigma_2) := D_C[p_{\mu_1,\Sigma_1}, p_{\mu_2,\Sigma_2}] = D_C(\mu_{12},\Sigma_{12})
$$

can be written as a function of two terms $\mu_{12} = \Sigma_1^{-\frac{1}{2}}(\mu_2-\mu_1)$ and $\Sigma_{12} = \Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}$.

5.2. Closed-Form Formula for the Chernoff Information between Univariate Gaussian Distributions

We shall report the exact solution for the Chernoff information between univariate Gaussian distributions by solving a quadratic equation. We can also report a complex closed-form formula by using symbolic computing because the calculations are lengthy and thus prone to human error.

Instantiating Equation (24) for the case of univariate Gaussian distributions paramterized by (μ, σ^2), we get the following equation for the optimality condition of α^*:

$$\langle \theta_2 - \theta_1, \eta_{\alpha^*} \rangle = F(\theta_2) - F(\theta_1), \tag{36}$$

$$\left\langle \left(\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}, \frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_2^2} \right), (m_\alpha, v_\alpha) \right\rangle = \frac{1}{2} \log \frac{\sigma_2^2}{\sigma_1^2} + \frac{\mu_2^2}{2\sigma_2^2} - \frac{\mu_1^2}{2\sigma_1^2}, \tag{37}$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product and with the interpolated mean and variance along an exponential arc $\{(m_\alpha, v_\alpha)\}_{\alpha \in (0,1)}$ passing through (μ_1, σ_1^2) when $\alpha = 1$ and (μ_2, σ_2^2) when $\alpha = 0$ given by

$$m_\alpha = \frac{\alpha \mu_1 \sigma_2^2 + (1-\alpha)\mu_2 \sigma_1^2}{(1-\alpha)\sigma_1^2 + \alpha \sigma_2^2} = \frac{\alpha(\mu_1 \sigma_2^2 - \mu_2 \sigma_1^2) + \mu_2 \sigma_1^2}{\sigma_1^2 + \alpha(\sigma_2^2 - \sigma_1^2)}, \tag{38}$$

$$v_\alpha = \frac{\sigma_1^2 \sigma_2^2}{(1-\alpha)\sigma_1^2 + \alpha \sigma_2^2} = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \alpha(\sigma_2^2 - \sigma_1^2)}. \tag{39}$$

That is, for $p = p_{\mu_1, \sigma_1^2}$ and $q = p_{\mu_2, \sigma_2^2}$, we have the weighted geometric mixture $(pq)_\alpha^G = p_{m_\alpha, v_\alpha}$.

Thus, the optimality condition of the Chernoff optimal skewing parameter is given by:

$$\text{OC}_{\text{Gaussian}} : \quad \left(\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2} \right) m_\alpha - \left(\frac{1}{2\sigma_2^2} - \frac{1}{2\sigma_1^2} \right) v_\alpha = \frac{1}{2} \log \frac{\sigma_2^2}{\sigma_1^2} + \frac{\mu_2^2}{2\sigma_2^2} - \frac{\mu_1^2}{2\sigma_1^2}. \tag{40}$$

Let us rewrite compactly Equation (40) as

$$\text{OC}_{\text{Gaussian}} : \quad a_{12} m_\alpha + b_{12} v_\alpha + c_{12} = 0, \tag{41}$$

with the following coefficients:

$$a_{12} = \frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}, \tag{42}$$

$$b_{12} = \frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_2^2}, \tag{43}$$

$$c_{12} = \frac{1}{2} \log \frac{\sigma_1^2}{\sigma_2^2} + \frac{\mu_1^2}{2\sigma_1^2} - \frac{\mu_2^2}{2\sigma_2^2}. \tag{44}$$

By multiplying both sides of Equation (41) by $\sigma_1^2 + \alpha \Delta_v$ where $\Delta_v := \sigma_2^2 - \sigma_1^2$ and rearranging terms, we get a quadratic equation with positive root being α^*.

Using the computer algebra system (CAS) Maxima, we can also solve exactly this quadratic equation in α as a function of μ_1, σ_1^2, μ_2, and σ_2^2: See listing in Appendix C.

Once we get the optimal value of $\alpha^* = \alpha^*(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$, we get the Chernoff information as

$$D_C[p_{\mu_1, \sigma_1^2}, p_{\mu_2, \sigma_2^2}] = D_{\text{KL}}[p_{m_{\alpha^*}, v_{\alpha^*}} : p_{\mu_1, \sigma_1^2}]$$

with the Kullback–Leibler divergence between two univariate Gaussians distributions p_{μ_1, σ_1^2} and p_{μ_2, σ_2^2} given by

$$D_{\text{KL}}[p_{\mu_1, \sigma_1^2} : p_{\mu_2, \sigma_2^2}] = \frac{1}{2} \left(\frac{(\mu_2 - \mu_1)^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - \log \frac{\sigma_1^2}{\sigma_2^2} - 1 \right).$$

Notice that from the invariance of Proposition 9, we have for any $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{++}$:

$$D_{\text{KL}}[p_{\mu_1, \sigma_1^2} : p_{\mu_2, \sigma_2^2}] = D_{\text{KL}} \left[p_{\frac{\mu_1 - \mu}{\sigma}, \frac{\sigma_1^2}{\sigma^2}} : p_{\frac{\mu_2 - \mu}{\sigma}, \frac{\sigma_2^2}{\sigma^2}} \right],$$

and therefore by choosing $(\mu, \sigma^2) = (\mu_1, \sigma_1^2)$, we have

$$D_{\mathrm{KL}}[p_{\mu_1,\sigma_1^2} : p_{\mu_2,\sigma_2^2}] = D_{\mathrm{KL}}\left[p_{0,1}, p_{\frac{\mu_2-\mu_1}{\sigma_1} : \frac{\sigma_2^2}{\sigma_1^2}} \right].$$

Proposition 10. *The Chernoff information between two univariate Gaussian distributions can be calculated exactly in closed form.*

One can also program these closed-form solutions in Python using the `SymPy` package (https://www.sympy.org/en/index.html (accessed on 30 July 2022)) for performing symbolic computations.

Let us report special cases with some illustrating examples.

- First, let us consider the Gaussian subfamily with prescribed variance. When $\sigma_1^2 = \sigma_2^2 = \sigma^2$, we always have $\alpha^* = \frac{1}{2}$, and the Chernoff information is

$$D_{\mathrm{C}}[p_{\mu_1,\sigma^2} : p_{\mu_2,\sigma^2}] = \frac{(\mu_2 - \mu_1)^2}{8\sigma^2}. \tag{45}$$

  Notice that it amounts to one eight of the squared Mahalanobis distance (see [60] for a detailed explanation).

- Second, let us consier Gaussian subfamily with prescribed mean. When $\mu_1 = \mu_2 = \mu$, we get the optimal skewing value independent of the mean μ:

$$\alpha^* = -\frac{v_1 \log(2\,v_2) - v_2 - v_1 \log(2\,v_1) + v_1}{(v_2 - v_1)\log(2\,v_2) - \log(2\,v_1)\,v_2 + v_1 \log(2\,v_1)}$$

  where $v_1 = \sigma_1^2$ and $v_2 = \sigma_2^2$. The Chernoff information is

$$D_{\mathrm{C}}[p_{\mu_1,v_1} : p_{\mu_2,v_2}] - \frac{(v_2 - v_1)\log\left(\frac{v_2 \log(2\,v_2) - \log(2\,v_1)\,v_2}{v_2 - v_1}\right) - v_2 \log(2\,v_2) + (\log(2\,v_1) + 1)\,v_2 - v_1}{2\,v_2 - 2\,v_1}. \tag{46}$$

- Third, consider the Chernoff information between the standard normal distribution and another normal distribution. When $(\mu_1, \sigma_1^2) = (0, 1)$ and $(\mu_2, \sigma_2^2) = (\mu, v)$, we get

$$\alpha^* = \frac{\sqrt{(4\,\mu^2 v^2 - 4\,\mu^2 v)\log(2\,v) + v^4 - 4\,v^3 + (6 - 4\log 2\,\mu^2)\,v^2 + (4\,\mu^4 + 4\log 2\,\mu^2 - 4)\,v + 1 + (2 - 2\,v)\log(2\,v) + v^2 + (2\log 2 - 2)\,v - 2\,\mu^2 - 2\log 2 + 1}}{(2\,v^2 - 4\,v + 2)\log(2\,v) - 2\log 2\,v^2 + (2\,\mu^2 + 4\log 2)\,v - 2\,\mu^2 - 2\log 2}$$

Example 4. *Let us consider $N(\mu_1 = 0, \sigma_1^2 = 1)$ and $N(\mu_2 = 1, \sigma_2^2 = 2)$. The Chernoff exponent is*

$$\alpha^* = \frac{\sqrt{8\log 4 - 8\log 2 + 9} - 2\log 4 + 2\log 2 - 1}{2\log 4 - 2\log 2 + 2} \approx 0.4215580558605244,$$

and the Chernoff information is (zoom in for the formula):

$$-\frac{\sqrt{8\log 4 - 8\log 2 + 9}\left((2\log 4 - 2\log 2 + 3)\log\left(\frac{4\log 4 - 4\log 2 + 4}{\sqrt{8\log 4 - 8\log 2 + 9} + 1}\right) - 4(\log 4)^2 + (8\log 2 - 6)\log 4 - 4(\log 2)^2 + 6\log 2 - 2\right) + (6\log 4 - 6\log 2 + 7)\log\left(\frac{4\log 4 - 4\log 2 + 4}{\sqrt{8\log 4 - 8\log 2 + 9} + 1}\right) + 4(\log 4)^2 + (10 - 8\log 2)\log 4 + 4(\log 2)^2 - 10\log 2 + 6}{(4\log 4 - 4\log 2 + 6)\sqrt{8\log 4 - 8\log 2 + 9} + 12\log 4 - 12\log 2 + 14}$$

$$\approx 0.1155433222682347$$

Using the bisection search of [24] with $\epsilon = 10^{-8}$ takes 28 iterations, and we get

$$\alpha^* \approx 0.42155805602669716,$$

and the Chernoff information is approximately 0.11554332226823472. Now, if we swap $p_{\mu_1,\sigma_1^2} \leftrightarrow p_{\mu_2,\sigma_2^2}$, we find $\alpha^* = 0.5784419439733028$ (and $0.5784419439733028 + 0.42155805602669716 \approx 1$).

Notice that in general, we may evaluate how good is the approximation $\tilde{\alpha}$ of α^* by evaluating the deficiency of the optimal condition:

$$\left| (\theta_2 - \theta_1)^\top \eta_{\tilde{\alpha}} - F(\theta_2) + F(\theta_1) \right|.$$

Example 5. *Let us consider $\mu_1 = 1$, $\sigma_1^2 = 3$ and $\mu_2 = 5$ and $\sigma_2^2 = 5$. We get*

$$\alpha^* = \frac{\sqrt{120 \log 10 - 120 \log 6 + 961} - 3 \log 10 + 3 \log 6 - 23}{2 \log 10 - 2 \log 6 + 16} \approx 0.4371453168322306$$

and the Chernoff information is reported in closed form and evaluated numerically as

$$0.5242883659200144.$$

In comparison, the bisection algorithm of [24] with $\epsilon = 10^{-8}$ takes 28 iterations, and reports $\alpha^* \approx 0.43714531883597374$ and the Chernoff information about

$$0.5242883659200137.$$

Corollary 4. *The smallest enclosing left-sided Kullback–Leibler disk of n univariate Gaussian distributions can be calculated exactly in randomized linear time [56].*

5.3. Fast Approximation of the Chernoff Information of Multivariate Gaussian Distributions

In general, the Chernoff information between d-variate Gaussians distributions is not known in closed-form formula when $d > 1$, see for example [61–63]. We shall consider below some special cases:

- When the Gaussians have the same covariance matrix Σ, the Chernoff information optimal skewing parameter is $\alpha = \frac{1}{2}$ and the Chernoff information is

$$D_C[p_{\mu_1, \Sigma}, p_{\mu_2, \Sigma}] = \frac{1}{8} \Delta_\Sigma^2(\mu_1, \mu_2),$$

  where $\Delta_\Sigma^2(\mu_1, \mu_2) = (\mu_2 - \mu_1)^\top \Sigma^{-1}(\mu_2 - \mu_1)$ is the squared Mahalanobis distance. The Mahalanobis distance enjoys the following property by congruence transformation:

$$\Delta_\Sigma(\mu_1, \mu_2) = \Delta_{A\Sigma A^\top}(A\mu_1, A\mu_2), \forall A \in \mathrm{GL}(d). \tag{47}$$

  Notice that we can rewrite the (squared) Mahalanobis distance as

$$\Delta_\Sigma^2(\mu_1, \mu_2) = \mathrm{tr}\left(\Sigma^{-1}(\mu_2 - \mu_1)(\mu_2 - \mu_1)^\top\right)$$

  using the matrix trace cyclic property. Then we check that

$$\begin{aligned} \Delta_{A\Sigma A^\top}^2(A\mu_1, A\mu_2) &= \mathrm{tr}\left(A^{-\top}\Sigma^{-1}A^{-1}A(\mu_2 - \mu_1)(\mu_2 - \mu_1)^\top A^\top\right), \\ &= \mathrm{tr}(\Sigma^{-1}(\mu_2 - \mu_1)(\mu_2 - \mu_1)^\top) = \Delta_\Sigma^2(\mu_1, \mu_2). \end{aligned}$$

- The Chernoff information for the special case of centered multivariate Gaussians distributions was studied in [62]. The KLD between two centered Gaussians p_{μ, Σ_1} and p_{μ, Σ_2} is half of the matrix Burg distance:

$$D_{\mathrm{KL}}[p_{\mu, \Sigma_1} : p_{\mu, \Sigma_2}] = \frac{1}{2}\left(\log \frac{\det(\Sigma_2)}{\det(\Sigma_1)} + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - d\right) =: \frac{1}{2} D_{\mathrm{Burg}}[\Sigma_1 : \Sigma_2]. \tag{48}$$

  When $d = 1$, the Burg distance corresponds to the well-known Itakura–Saito divergence. The matrix Burg distance is a matrix spectral distance [62]:

$$D_{\mathrm{Burg}}[\Sigma_1 : \Sigma_2] = \left(\sum_{i=1}^{d} \lambda_i - \log \lambda_i - 1 \right),$$

where the λ_i's are the eigenvalues of $\Sigma_2 \Sigma_1^{-1}$. The reverse KLD divergence $D_{\mathrm{KL}}[p_{\mu,\Sigma_2} : p_{\mu,\Sigma_1}] = \frac{1}{2} D_{\mathrm{Burg}}[\Sigma_2 : \Sigma_1]$ is obtained by replacing $\lambda_i \leftrightarrow \frac{1}{\lambda_i}$:

$$D_{\mathrm{KL}}[p_{\mu,\Sigma_2} : p_{\mu,\Sigma_1}] = \frac{1}{2} \left(\sum_{i=1}^{d} \frac{1}{\lambda_i} + \log \lambda_i - 1 \right).$$

More generally, the f-divergences between centered Gaussian distributions are always matrix spectral divergences [60].

Otherwise, for the general multivariate case, we implement the dichotomic search of Algorithm 1 in Algorithm 3 with the KLD between two multivariate Gaussian distributions expressed as

$$
\begin{aligned}
D_{\mathrm{KL}}[p_{\mu_1,\Sigma_1} : p_{\mu_2,\Sigma_2}] &= \frac{1}{2} \Delta_{\Sigma}^2(\mu_1, \mu_2) + \frac{1}{2} D_{\mathrm{Burg}}[\Sigma_1 : \Sigma_2], && (49) \\
&= \frac{1}{2} \left(\mathrm{tr}(\Sigma_2^{-1} \Sigma_1) - \log \frac{\det(\Sigma_2)}{\det(\Sigma_1)} - d + (\mu_2 - \mu_1)^{\top} \Sigma_2^{-1} (\mu_2 - \mu_1) \right). && (50)
\end{aligned}
$$

Algorithm 3 Dichotomic search for approximating the Chernoff information between two multivariate normal distributions p_{μ_1,Σ_1} and p_{μ_2,Σ_2} by approximating the optimal skewing parameter value $\alpha \approx \alpha^*$.

---

input : Two normal densities p_{μ_1,Σ_1} and p_{μ_2,Σ_2}, and a numerical precision threshold $\epsilon > 0$

$\alpha_m = 0$;
$\alpha_M = 1$;
while $|\alpha_M - \alpha_m| > \epsilon$ **do**

$\quad \alpha = \frac{\alpha_m + \alpha_M}{2}$;

$\quad \Sigma_{\alpha}^e = \left((1-\alpha)\Sigma_1^{-1} + \alpha \Sigma_2^{-1} \right)^{-1}$;

$\quad \mu_{\alpha}^e = \Sigma_{\alpha}^e \left((1-\alpha)\Sigma_1^{-1}\mu_1 + \alpha \Sigma_2^{-1}\mu_2 \right)$;

$\quad$ // Formula of the KLD between two normal distributions in Equation (50)

$\quad$ **if** $D_{\mathrm{KL}}[p_{\mu_{\alpha}^e,\Sigma_{\alpha}^e} : p_{\mu_1,\Sigma_1}] > D_{\mathrm{KL}}[p_{\mu_{\alpha}^e,\Sigma_{\alpha}^e} : p_{\mu_2,\Sigma_2}]$ **then**

$\quad\quad \alpha_m = \alpha$;

$\quad\quad$ // See Figure 4 for an illustration and Proposition 4

$\quad$ **end**

$\quad$ **else**

$\quad\quad \alpha_M = \alpha$;

$\quad$ **end**

end
return $D_{\mathrm{KL}}[p_{\mu_{\alpha}^e,\Sigma_{\alpha}^e} : p_{\mu_1,\Sigma_1}]$;

Example 6. *Let $d = 2$, $p_{\mu_1,\Sigma_1} = p_{0,I}$ be the standard bivariate Gaussian distribution and p_{μ_2,Σ_2} be the bivariate Gaussian distribution with mean $\mu_2 = [1\ 2]^{\top}$ and covariance matrix $\Sigma_2 = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}$. Setting the numerical precision threshold ϵ to $\epsilon = 10^{-8}$, the dichotomic search performs 28 split iterations, and approximate α^* by*

$$\alpha^* \approx 0.5825489424169064.$$

The Chernoff information $D_C[p_{0,I}, p_{\mu_2,\Sigma_2}]$ is approximated by 0.8827640697808525.

The *m*-interpolation of multivariate Gaussian distributions p_{μ_1,Σ_1} and p_{μ_2,Σ_2} with respect to the mixture connection ∇^m is given by

$$\gamma^m_{p_{\mu_1,\Sigma_1},p_{\mu_2,\Sigma_2}}(\alpha) = p_{\mu^m_\alpha,\Sigma^m_\alpha},$$

where

$$\begin{aligned}
\mu^m_\alpha &= (1-\alpha)\mu_1 + \alpha\mu_2 =: \bar{\mu}_\alpha, \\
\Sigma^m_\alpha &= (1-\alpha)\Sigma_1 + \alpha\Sigma_2 + (1-\alpha)\mu_1\mu_1^\top + \alpha\mu_2\mu_2^\top - \bar{\mu}_\alpha\bar{\mu}_\alpha^\top.
\end{aligned}$$

To *e*-interpolation of multivariate Gaussian distributions p_{μ_1,Σ_1} and p_{μ_2,Σ_2} with respect to the exponential connection ∇^e is given by

$$\gamma^e_{p_{\mu_1,\Sigma_1},p_{\mu_2,\Sigma_2}}(\alpha) = p_{\mu^e_\alpha,\Sigma^e_\alpha},$$

where

$$\begin{aligned}
\mu^e_\alpha &= \Sigma^e_\alpha\left((1-\alpha)\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2\right), \\
\Sigma^e_\alpha &= \left((1-\alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1}\right)^{-1}.
\end{aligned}$$

In information geometry, both these *e*- and *m*-connections defined with respect to an exponential family are shown to be flat. These geodesics correspond to linear interpolations in the ∇^e-affine coordinate system θ and in the dual ∇^m coordinate system η, respectively.

Figure 7 displays these two *e*-geodesic and *m*-geodesic between two multivariate normal distributions. Notice that the Riemannian geodesic with the Levi–Civita metric connection $\frac{\nabla^e + \nabla^m}{2}$ is not known in closed form for boundary value conditions. The expression of the Riemannian geodesic is known only for initial value conditions [64] (i.e., starting point with a given vector direction).
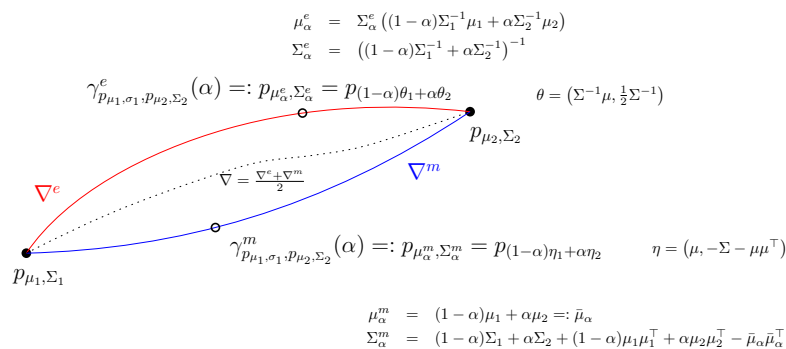

Figure 7. Interpolation along the *e*-geodesic and the *m*-geodesic passing through two given multivariate normal distributions. No closed-form is known for Riemannian geodesic with respect to the metric Levi–Civita connection (shown in dashed style).

5.4. Chernoff Information between Centered Multivariate Normal Distributions

The set

$$\mathcal{N}_0 = \left\{ p_\Sigma(x) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2}x^\top\Sigma^{-1}x\right) : \Sigma \succ 0 \right\}$$

of centered multivariate normal distributions is a regular exponential family with natural parameter $\theta = \Sigma^{-1}$, sufficient statistic $t(x) = -\frac{1}{2}xx^\top$, log-normalizer $F(\theta) = -\frac{1}{2}\log\det(\theta)$ and auxiliary carrier term $k(x) = -\frac{d}{2}\log(2\pi)$. Family $\mathcal{N}_0$ is also a multivariate scale family with scale matrices $\Sigma^{\frac{1}{2}}$ (standard deviation σ in 1D).

Let $\langle A, B \rangle = \text{tr}(A^\top B)$ defines the inner product between two symmetric matrices A and B. Then we can write the centered Gaussian distribution $p_\Sigma(x)$ in the canonical form of exponential families:

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

The function log det of a positive-definite matrix is strictly concave [65], and hence we check that $F(\theta)$ is strictly convex. Furthermore, we have $\nabla_X \log \det(X) = X^{-\top}$ so that $\nabla_\theta F(\theta) = -\frac{1}{2}\theta^{-\top}$.

The optimality condition equation of Chernoff best skewing parameter α^* becomes:

$$\langle \theta_2 - \theta_1, \nabla F(\theta_1 + \alpha^*(\theta_2 - \theta_1)) \rangle = F(\theta_2) - F(\theta_1), \tag{51}$$

$$-\frac{1}{2}\text{tr}((\theta_2 - \theta_1)^\top (\theta_1 + \alpha^*(\theta_2 - \theta_1))^{-1}) = -\frac{1}{2}\log\frac{\det(\theta_2)}{\det(\theta_1)}, \tag{52}$$

$$\text{tr}((\theta_2 - \theta_1)^\top (\theta_1 + \alpha^*(\theta_2 - \theta_1))^{-1}) = \log\frac{\det(\theta_2)}{\det(\theta_1)}, \tag{53}$$

$$\text{tr}((\Sigma_2^{-1} - \Sigma_1^{-1})(\Sigma_1^{-1} + \alpha^*(\Sigma_2^{-1} - \Sigma_1^{-1}))^{-1}) = \log\frac{\det(\Sigma_1)}{\det(\Sigma_2)} = \log\det\left(\Sigma_1\Sigma_2^{-1}\right). \tag{54}$$

When $\Sigma_2 = s\Sigma_1$ (and $\Sigma_2^{-1} = \frac{1}{s}\Sigma_1^{-1}$) for $s > 0$ and $s \neq 1$, we get a closed-form for α^* using the fact that $\det\left(\frac{I}{s}\right) = \frac{1}{s^d}$ and $\text{tr}(I) = d$ for d-dimensional identity matrix I. Solving Equation (54) yields

$$\alpha^*(s) = \frac{s - 1 - \log s}{(s-1)\log s} \in (0,1). \tag{55}$$

Therefore the Chernoff information between two scaled centered Gaussian distributions $p_{\mu,\Sigma}$ and $p_{\mu,s\Sigma}$ is available in closed form.

Proposition 11. *The Chernoff information between two scaled d-dimensional centered Gaussian distributions $p_{\mu,\Sigma}$ and $p_{\mu,s\Sigma}$ of $\mathcal{N}_\mu$ (for $s > 0$) is available in closed form:*

$$D_C[p_{\mu,\Sigma}, p_{\mu,s\Sigma}] = D_{B,\alpha^*}[p_{\mu,\Sigma}, p_{\mu,s\Sigma}] = d\frac{(s-1)\log\left(\frac{s}{s-1}\log s\right) - s\log s + s - 1}{2(1-s)}, \tag{56}$$

where $\alpha^* = \frac{s - 1 - \log s}{(s-1)\log s} \in (0,1)$.

Notice that $\alpha^*(p_{\mu,\Sigma} : p_{\mu,s\Sigma}) = \alpha^*(p_{\mu,\Sigma}, p_{\mu,\frac{1}{s}\Sigma})$ and $D_C[p_{\mu,\Sigma}, p_{\mu,s\Sigma}] = D_C[p_{\mu,\Sigma}, p_{\mu,\frac{1}{s}\Sigma}]$.

Example 7. *Consider $\mu_1 = \mu_2 = 0$ and $\Sigma_1 = I$, $\Sigma_2 = \frac{1}{2}I$. We find that $\alpha^* = \frac{2\log 2 - 1}{\log 2}$, which is independent of the dimension of the matrices. The Chernoff information depends on the dimension:*

$$D_C[p_{0,I}, p_{0,\frac{1}{2}I}] = d\frac{\log 2 - \log\log 2 - 1}{2}.$$

Notice that when $d = 1$, we have $s = \frac{\sigma_2^2}{\sigma_1^2}$, and we recover a special case of the closed-form formula for the Chernoff information between univariate Gaussians.

In [62], the following equation is reported for finding α^* based on Equation (54):

$$\text{OC}_{\text{CenteredGaussians}} : \quad \sum_{i=1}^{d} \frac{1 - \lambda_i}{\alpha^* + (1 - \alpha^*)\lambda_i} + \log \lambda_i = 0 \tag{57}$$

where the λ_i's are generalized eigenvalues of $\Sigma_1\Sigma_2^{-1}$ (this excludes the case of all λ_i's equal to one). The value of α^* satisfying Equation (57) is unique. Let us notice that the product of two symmetric positive-definite matrices is not necessarily symmetric anymore.

We can derive Equation (57) by expressing Equation (54) using the identity matrix I and matrix $\Sigma_2^{-\frac{1}{2}}\Sigma_1\Sigma_2^{-\frac{1}{2}}$.

Remark 9. *We can get closed-form solutions for α^* and the corresponding Chernoff information in some particular cases. For example, when the dimension $d = 2$, we need to solve a quadratic equation to get α^*. Thus, for $d \leq 4$, we get a closed-form solution for α^* by solving a polynomial equation characterizing the optimal condition, and obtain the Chernoff information in closed-form as a byproduct.*

Example 8. *Consider the Chernoff information between $p_{0,I}$ and $p_{0,\Lambda}$ with $\Lambda = \mathrm{diag}(1, 2, 3, 4)$. We get the exact Chernoff exponent value α^* by taking the root of a quartic polynomial equation falling in $(0, 1)$. By evaluating numerically this root, we find that $\alpha^* \simeq 0.59694$ and that the Chernoff information is $D_C[p_{0,I}, p_{0,\Lambda}] \simeq 0.22076$. See Appendix C for some symbolic computation code.*

6. Chernoff Information between Densities of Different Exponential Families

Let
$$\mathcal{E}_1 = \{p_\theta = \exp(\langle \theta, t_1(x)\rangle - F_1(\theta)) \; : \; \theta \in \Theta\},$$
and
$$\mathcal{E}_2 = \{q_{\theta'} = \exp(\langle \theta', t_2(x)\rangle - F_2(\theta')) \; : \; \theta' \in \Theta'\},$$
be two distinct exponential families, and consider the Chernoff information between the densities p_{θ_1} and $q_{\theta'_2}$. The exponential arc induced by p_{θ_1} and $q_{\theta'_2}$ is
$$\{(p_{\theta_1} q_{\theta'_2})^G_\alpha \propto p^\alpha_{\theta_1} q^{1-\alpha}_{\theta'_2} \; : \; \alpha \in (0, 1)\}.$$

Let $\mathcal{E}_{12}$ denote the exponential family with sufficient statistics $(t_1(x), t_2(x))$, log-normalizer $F_{12}(\theta, \theta')$, and denote by Θ_{12} its natural parameter space. Family $\mathcal{E}_{12}$ can be interpreted as a product exponential family which yields an exponential family. We have
$$(p_{\theta_1} q_{\theta'_2})^G_\alpha = \exp\big(\langle(t_1(x), t_2(x)), (\alpha\theta_1, (1-\alpha)\theta'_2)\rangle - F_{12}(\alpha\theta_1, (1-\alpha)\theta'_2)\big).$$

Thus the induced LREF $\mathcal{E}_{p_{\theta_1} q_{\theta'_2}}$ with natural parameter space $\Theta_{p_{\theta_1} q_{\theta'_2}}$ can be interpreted as a 1D curved exponential family of the product exponential family $\mathcal{E}_{12}$.

The optimal skewing parameter α^* is found by setting the derivative of $F_{12}(\alpha\theta_1, (1-\alpha)\theta'_2)$ with respect α to zero:
$$\frac{d}{d\alpha} F_{12}(\alpha\theta_1, (1-\alpha)\theta'_2) = 0.$$

Example 9. *Let $\mathcal{E}_1$ can be chosen as the exponential family of exponential distributions*
$$\mathcal{E}_1 = \{e_\lambda(x) = \lambda \exp(-\lambda x), \lambda \in (0, +\infty)\}$$
defined on the support $\mathcal{X}_1 = (0, \infty)$ and $\mathcal{E}_2$ can be chosen as the exponential family of half-normal distributions
$$\mathcal{E}_2 = \left\{h_\sigma(x) = \sqrt{\frac{2}{\pi\sigma^2}} \exp(-\frac{x^2}{2\sigma^2}) \; : \; \sigma^2 > 0\right\}$$
with support $\mathcal{X}_2 = (0, \infty)$.

The product exponential family corresponds to the singly truncated normal family [50] which is a non-regular (i.e., parameter space is not topologically an open set):
$$\Theta_{12} = (\mathbb{R} \times \mathbb{R}_{++}) \cup \Theta_0,$$
with $\Theta_0 = \{(\theta, 0) \; : \; \theta < 0\}$ (the part corresponding to the exponential family of exponential distributions). This exponential family $\mathcal{E}_{12} = \{p_{\theta_1, \theta_2}\}$ of singly truncated normal distributions is also non-steep [50]. The log-normalizer is

$$F_{12}(\theta_1, \theta_2) = \frac{1}{2} \log \frac{\pi}{\theta_2} + \log \Phi\left(\frac{\theta_1}{\sqrt{2\theta_2}}\right) + \frac{\theta_1^2}{4\theta_2},$$

where $\theta_1 = \frac{\mu}{\sigma^2}$ and $\theta_2 = \frac{1}{2\sigma^2}$, and Φ denotes the cumulative distribution function of the standard normal. Function F_{12} is proven of class C^1 on Θ_{12} (see Proposition 3.1 of [50]) with $F_{12}(\theta, 0) = -\log(-\theta)$ for $\theta < 0$.

Notice that the KLD between an exponential distribution and a half-normal distribution is $+\infty$ since the definite integral diverges (hence $D_{\mathrm{KL}}[e_\lambda : h_\sigma]$ is not equivalent to a Bregman divergence, and $\Theta_{e_{\theta_1} h_{\theta'_2}}$ is not open at 1) but the reverse KLD between a half-normal distribution and an exponential distribution is available in closed-form (using symbolic computing):

$$D_{\mathrm{KL}}[h_\sigma : e_\lambda] = \sqrt{\sqrt{8}\sigma\lambda - \sqrt{\pi}(1 + \log\frac{\pi\lambda^2\sigma^2}{2})2\sqrt{\pi}}.$$

Figure 8 illustrate the domain of the singly truncated normal distributions and displays an exponential arc between an exponential distribution and a half-normal distribution. Notice that we could have also considered a similar but different example by taking the exponential family of Rayleigh distributions which exhibit an additional extra carrier term $k(x)$.

The Bhattacharyya α-skewed coefficient calculated using symbolic computing (see Appendix C) is

$$\rho_\alpha[h_\sigma : e_\lambda] = \rho_{1-\alpha}[e_\lambda : h_\sigma] = \frac{\pi^{\frac{1}{2}-\frac{\alpha}{2}} e^{-\sigma^2\lambda^2}\left(2^{\frac{\alpha}{2}+\frac{1}{2}}\sigma\lambda\, e^{\frac{\alpha\sigma^2\lambda^2}{2}+\frac{\sigma^2\lambda^2}{2\alpha}}\,\mathrm{erf}\left(\frac{(\sqrt{2}\alpha-\sqrt{2})\sigma\lambda}{2\sqrt{\alpha}}\right) + 2^{\frac{\alpha}{2}+\frac{1}{2}}\sigma\lambda\, e^{\frac{\alpha\sigma^2\lambda^2}{2}+\frac{\sigma^2\lambda^2}{2\alpha}}\right)}{2\sqrt{\alpha}\,\sigma^\alpha\lambda^\alpha},$$

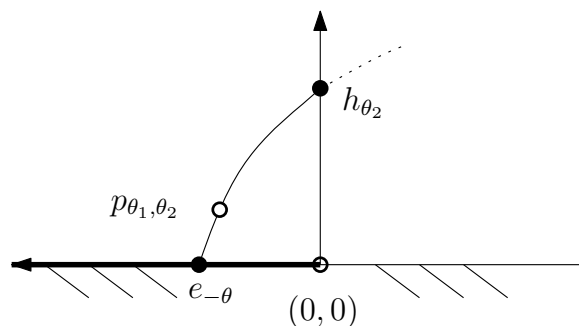*where erf denotes the error function.*

Figure 8. The natural parameter space of the non-regular full exponential family of singly truncated normal distributions is not regular (i.e., not open): The negative real axis corresponds to the exponential family of exponential distributions.

7. Conclusions

In this work, we revisited the Chernoff information [2] (1952) which was originally introduced to upper bound Bayes' error in binary hypothesis testing. A general characterization of Chernoff information between two arbitrary probability measures was given in [11] (Theorem 32) by considering Rényi divergences which can be interpreted as scaled skewed Bhattacharyya divergences. Since its inception, the Chernoff information has proven useful as a statistical divergence (Chernoff divergence) in many applications ranging from information fusion to quantum metrology due to its empirical robustness property [19]. Informally, we may observe empirically that in practice the skewed Bhattacharyya divergence is more stable around the Chernoff exponent α^* than in other part of the range $(0, 1)$. By considering the maximal extension of the exponential arc joining two densities p and q on a Lebesgue space $L^1(\mu)$, we built full likelihood ratio exponential families [10] $\mathcal{E}_{pq}$ (LREFs) in Section 2. When the LREF $\mathcal{E}_{pq}$ is a regular exponential family (with coinciding support of p and q), both the forward and reverse Kullback–Leibler divergence are finite and can be

rewritten as finite Bregman divergences induced by the log-normalizer F_{pq} of $\mathcal{E}_{pq}$ which amounts to minus skewed Bhattacharyya divergences. Since log-normalizers of exponential families are strictly convex, we deduced that the skewed Bhattacharyya divergences are strictly concave and their maximization yielding the Chernoff information is hence proven unique. As a byproduct, this geometric characterization in $L^1(\mu)$ allowed us to prove that the intersection of a e-geodesic with a m-bisector is unique in dually flat subspaces of $L^1(\mu)$, and similarly that the intersection of a m-geodesic with a e-bisector is unique (Proposition 8). We then considered the exponential families of univariate and multivariate normal distributions: We reported closed-form solutions for the Chernoff information of univariate normal distribution and centered normal distributions with scaled covariance matrices, and show how to implement efficiently a dichotomic search for approximating the Chernoff information between two multivariate normal distributions (Algorithm 3). Table 1 summarizes the various optimal condition studied characterizing the Chernoff exponent. Finally, inspired by the Chernoff information study, we defined in Section 4, the forward and reverse Bregman–Chernoff divergences [66], and show how these divergences are related to the capacity of a discrete memoryless channel and the minimax redundancy of universal coding in information theory [13].

Table 1. Summary of the optimal conditions characterizing the Chernoff exponent.

	Generic case
Primal LREF	$OC_\alpha : D_{KL}[(pq)^G_{\alpha^*} : p] = D_{KL}[(pq)^G_{\alpha^*} : q]$
Dual LREF	$OC_\beta : \beta(\alpha^*) = E_{(pq)^G_{\alpha^*}}\left[\log \frac{p(x)}{q(x)}\right] = 0$
Geometric OC	$(pq)^G_{\alpha^*} = \gamma^G(p,q) \cap \mathrm{Bi}^{\mathrm{left}}_{KL}(p,q)$
	Case of exponential families
Bregman	$OC_{EF} : \quad B_F(\theta_1 : \theta_{\alpha^*}) = B_F(\theta_2 : \theta_{\alpha^*})$
Fenchel–Young	$OC_{YF} : \quad Y_{F,F^*}(\theta_1 : \eta_{\alpha^*}) = Y_{F,F^*}(\theta_2 : \eta_{\alpha^*})$
Simplified	$OC_{SEF'} : \quad F'_{\theta_1,\theta_2}(\alpha) = 0$
	$OC_{SEF} : \quad (\theta_2 - \theta_1)^\top \nabla F(\theta_1 + \alpha^*(\theta_2 - \theta_1)) = F(\theta_2) - F(\theta_1)$
Geometric OC	$\gamma^e_{pq}(\alpha) \cap \mathrm{Bi}^m(p,q)$
1D EF	$\alpha^* = \dfrac{F'^{-1}\left(\frac{F(\theta_2)-F(\theta_1)}{\theta_2-\theta_1}\right)-\theta_2}{\theta_1-\theta_2}$
	Gaussian case
1D Gaussians	$OC_{\mathrm{Gaussian}} : \quad \left(\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}\right)m_\alpha - \left(\frac{1}{2\sigma_2^2} - \frac{1}{2\sigma_1^2}\right)v_\alpha = \frac{1}{2}\log\frac{\sigma_2^2}{\sigma_1^2} + \frac{\mu_2^2}{2\sigma_2^2} - \frac{\mu_1^2}{2\sigma_1^2}$
	α^* is root of quadratic polynomial in $(0,1)$
Centered Gaussians	$OC_{\mathrm{CenteredGaussians}} : \sum_{i=1}^d \frac{1-\lambda_i}{\alpha^*+(1-\alpha^*)\lambda_i} + \log\lambda_i = 0$
	where λ_i is the i-th eigenvalue of $\Sigma_1\Sigma_2^{-1}$
Centered Gaussians	$\alpha^* = \frac{s-1-\log s}{(s-1)\log s} \in (0,1)$
scaled covariances	when $\Sigma_2 = s\Sigma_1$

Additional material including MAXIMA and JAVA® snippet codes is available online at https://franknielsen.github.io/ChernoffInformation/index.html (accessed on 30 July 2022).

Appendix A. Background on Statistical Divergences

We introduce some statistical dissimilarities like the Kullback–Leibler divergence, the Bhattacharyya distance or the Hellinger divergence which have been proven among others useful is characterizing or bounding the probability of error in Bayesian statistical hypothesis testing [4,25,67].

The Kullback–Leibler divergence [13] (KLD) between two probability measures (PMs) P and Q is defined as

$$D_{\mathrm{KL}}[P:Q] = \begin{cases} \int_{\mathcal{X}} \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}P, & \text{if } P \ll Q \\ +\infty & \text{if } P \not\ll Q \end{cases}$$

Two PMs P and Q are mutually singular when there exists an event $A \in \mathcal{A}$ such that $P(A) = 0$ and $Q(\mathcal{X}\backslash A) = 0$. Mutually singular measures P and Q are notationally written as $P \perp Q$. Let P and Q be two non-singular probability measures on $(\mathcal{X}, \mathcal{A})$ dominated by a common σ-finite measure μ, and denote by $p = \frac{\mathrm{d}P}{\mathrm{d}\mu}$ and $q = \frac{\mathrm{d}Q}{\mathrm{d}\mu}$ their Radon–Nikodym densities with respect to μ. Then the KLD between P and Q can be calculated equivalently by the KLD between their densities as follows:

$$D_{\mathrm{KL}}[P:Q] = D_{\mathrm{KL}}[p:q] = \int_{\mathcal{X}} p \log\left(\frac{p}{q}\right)\mathrm{d}\mu. \tag{A1}$$

It can be shown that $D_{\mathrm{KL}}[p:q]$ is independent of the chosen dominating measure μ [4], and thus when $P, Q \ll \mu$, we write for short $D_{\mathrm{KL}}[P:Q] = D_{\mathrm{KL}}[p:q]$. Although the dominating measure μ can be set to $\mu = \frac{P+Q}{2}$ in general, it is either often chosen as μ_L the Lebesgue measure for continuous sample spaces $\mathbb{R}^d$ (with the σ-algebra $\mathcal{A} = \mathcal{B}(\mathbb{R}^d)$ of Borel sets) or as the counting measure $\mu_{\#}$ for discrete sample spaces (with the σ-algebra $\mathcal{A}$ of power sets). The KLD is not a metric distance because it is asymmetric and does not satisfy the triangle inequality.

Let $\mathrm{supp}(\mu) = \mathrm{cl}\{A \in \mathcal{A} : \mu(A) \neq 0\}$ denote the support of a Radon positive measure [1] μ where cl denotes the topological closure operation. Notice that $D_{\mathrm{KL}}[p : q] = +\infty$ when the definite integral of Equation (A1) divergences (e.g., the KLD between a standard Cauchy distribution and a standard normal distribution is $+\infty$ but the KLD between a standard normal distribution and a standard Cauchy distributions is finite), and $D_{\mathrm{KL}}[P:Q] = \infty$ when the probability measures have disjoint supports ($P \perp Q$). Thus, when the supports of P and Q are distinct but not nested, both the forward KLD $D_{\mathrm{KL}}[P:Q]$ and the reverse KLD $D_{\mathrm{KL}}[Q:P]$ are infinite.

Let $f_\alpha(u) = u^\alpha$ for $\alpha \in \mathbb{R}$. The functions $f_\alpha(u)$ are convex for $\alpha \in \mathbb{R}\backslash[0,1]$ and concave for $\alpha \in [0,1]$. Thus, we can define the f-divergences [59,68]

$$I_{f_\alpha}[p:q] = \int p f_\alpha(q/p)\mathrm{d}\mu = \int p^{1-\alpha} q^\alpha \mathrm{d}\mu, \tag{A2}$$

for $\alpha \in \mathbb{R}\backslash[0,1]$ and $I_{-f_\alpha}[p:q] = -\int p f_\alpha(q/p)\mathrm{d}\mu = -\int p^{1-\alpha} q^\alpha \mathrm{d}\mu$ for $\alpha \in (0,1)$ (or equivalently take the convex generator $h_\alpha(u) = -u^\alpha$ for $\alpha \in (0,1)$). Notice that the conjugate f-divergence is obtained for the generator $f_\alpha^*(u) = u f_\alpha(1/u) = u^{1-\alpha}$: $I_{f_\alpha}[q:p] = I_{f_\alpha^*}[p:q]$. By Jensen's inequality, we have that the f-divergences are lower bounded by $f(1)$. Thus, $I_{h_\alpha^*}[p:q] \geq h_\alpha(1) = -1$. Since f-divergences are upper bounded by $f(0) + f^*(0)$, we have that $I_{h_\alpha^*}[p:q] < 0$ for $\alpha \in (0,1)$. This gives another proof that the Bhattacharyya coefficient $\rho_\alpha[p:q] = -I_{h_{1-\alpha}}[p:q]$ is bounded between 0 and 1 since the $I_{h_{1-\alpha}}$ divergence is bounded between -1 and 0. Moreover, Ali and Silvey [59] further defined the (f, g)-divergences as $I_{f,g}[p:q] = g(I_f[p:q])$ for a strictly monotonically increasing function $g(v)$. Letting $g(v) = -\log(-v)$ (with $g'(v) = -\frac{1}{v} < 0$ when $v \in (0,1)$), we get that the $(h_{1-\alpha}, g)$-divergences are the Bhattacharyya distances for $\alpha \in (0,1)$. However, the Chernoff information is not a f-divergence despite the fact that Bhattacharyya distances are Ali–Silvey (f, g)-divergences because of the maximization criterion [59] of Equation (4).

We refer the reader to [69] (Chapter 14), [70] (Figure 1) and [71] (Figure 3) for other statistical distances and statistical dissimilarities with their connections.

Appendix B. Exponential Family of Univariate Gaussian Distributions

Consider the family of univariate normal distributions:

$$\mathcal{N} = \left\{ p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), \mu \in \mathbb{R}, \sigma^2 > 0 \right\}.$$

Let $\lambda = (\lambda_1 = \mu, \lambda_2 = \sigma^2)$ denote the mean-variance parameterization, and consider the sufficient statistic vector $t(x) = (x, x^2)$. Then the densities of $\mathcal{N}$ can be written in the canonical form of exponential families:

$$p_\lambda(x) = \exp(\langle \theta(\lambda), t(x) \rangle - F(\theta)),$$

where $\theta(\lambda) = \left(\frac{\lambda_1}{\lambda_2}, -\frac{1}{2\lambda_2}\right)$ and the log-normalizer is

$$F(\theta) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log\frac{\pi}{-\theta_2}.$$

The dual moment parameterization is $\eta(\lambda) = E_{p_\lambda}[t(x)] = (\lambda_1, \lambda_1^2 + \lambda_2)$, and the convex conjugate is:

$$F^*(\eta) = \sup_{\theta \in \Theta}\{\langle \theta, \eta \rangle - F(\eta)\} = -\frac{1}{2}(\log(2\pi e(\eta_2 - \eta_1^2)).$$

We check that the convex conjugate coincides with the negentropy [72]:

$$h[p_\lambda] = -F^*(\eta(\lambda)).$$

The conversion formulæ between the dual natural/moment parameters and the ordinary parameters are given by:

$$\theta(\lambda) = \left(\frac{\lambda_1}{\lambda_2}, -\frac{1}{2\lambda_2}\right), \tag{A3}$$

$$\lambda(\theta) = \left(-\frac{\theta_1}{2\theta_2}, -\frac{1}{2\theta_2}\right), \tag{A4}$$

$$\eta(\lambda) = \left(\lambda_1, \lambda_1^2 + \lambda_2\right), \tag{A5}$$

$$\lambda(\eta) = \left(\eta_1, \eta_2 - \eta_1^2\right), \tag{A6}$$

$$\eta(\theta) = (E[x], E[x^2]) = \nabla F(\theta) = \left(-\frac{\theta_1}{2\theta_2}, -\frac{1}{2\theta_2} + \frac{\theta_1^2}{4\theta_2^2}\right), \tag{A7}$$

$$\theta(\eta) = \nabla F^*(\eta) = \left(-\frac{\eta_1}{\eta_1^2 - \eta_2}, \frac{1}{2(\eta_1^2 - \eta_2)}\right) \tag{A8}$$

We check that

$$D_{\mathrm{KL}}[p_\lambda : p_{\lambda'}] = \frac{1}{2}\left(\frac{(\lambda_1 - \lambda_1')^2}{\lambda_2'^2} + \frac{\lambda_2^2}{\lambda_2'^2} - \log\frac{\lambda_2^2}{\lambda_2'^2} - 1\right),$$
$$= B_F(\theta(\lambda') : \theta(\lambda)) = B_{F^*}(\eta(\lambda) : \eta(\lambda')),$$
$$= Y_{F,F^*}(\theta(\lambda') : \eta(\lambda)) = Y_{F^*,F}(\eta(\lambda) : \theta(\lambda')),$$

where B_F and B_{F^*} are the dual Bregman divergences and Y_{F,F^*} and $Y_{F^*,F}$ are the dual Fenchel–Young divergences.

Appendix C. Code Snippets in MAXIMA

Code for plotting Figure 1.

Listing A1: Plot the cumulant function of a log ratio exponential family induced by two normal distributions.

```
varalpha(v1,v2,alpha):=(v1*v2)/((1-alpha)*v1+alpha*v2)$
mualpha(mu1,v1,mu2,v2,alpha):=(alpha*mu1*v2+(1-alpha)*mu2*v1)/((1-alpha)*v1+
    alpha*v2)$
assume(v1>0)$assume(v2>0)$
theta1(mu,v):=mu/v$
theta2(mu,v):=-1/(2*v)$;
F(theta1,theta2):=((-theta1**2)/(4*theta2))+(1/2)*log(-%pi/theta2)$
JF(alpha,theta1,theta2,theta1p,theta2p):=alpha*F(theta1,theta2)+(1-alpha)*F(
    theta1p,theta2p)-F(alpha*theta1+(1-alpha)*theta1p,alpha*theta2+(1-alpha)*
    theta2p);

m1:0;v1:1;m2:1;v2:2;

plot2d([JF(alpha,theta1(m1,v1),theta2(m1,v1),theta1(m2,v2),theta2(m2,v2)),
-JF(alpha,theta1(m1,v1),theta2(m1,v1),theta1(m2,v2),theta2(m2,v2)),
[discrete,[[0.4215580558605244,-0.15],[0.4215580558605244,0.15]]],
[discrete, [0.4215580558605244], [0.1154433222682347]],
[discrete, [0.4215580558605244], [-0.1155433222682347]]
],
[alpha,0,1], [xlabel,"alpha"], [ylabel,"F_{pq}(alpha)=-D_{B,alpha}[p:q]"],
[style,  [lines,1,1],[lines,1,2],
[lines,2,0], [points, 3,3],[points, 3,3]],[legend, "skew Bhattacharyya D_{B,
    alpha}[p:q]","LREF log-normalizer F_{pq}(alpha)","","","" ],
[color, blue, red,    black, black,black],[point_type,asterisk]);
```

Code for calculating the Chernoff information between two univariate Gaussian distributions (Proposition 10):

Listing A2: Calculate symbolically the exact Chernoff information between two univariate normal distributions.

```
varalpha(v1,v2,alpha):=(v1*v2)/((1-alpha)*v1+alpha*v2)$
mualpha(mu1,v1,mu2,v2,alpha):=(alpha*mu1*v2+(1-alpha)*mu2*v1)/((1-alpha)*v1+
    alpha*v2)$

/* Kullback--Leibler divergence */
KLD(mu1,v1,mu2,v2):=(1/2)*((((mu2-mu1)**2)/v2)+(v1/v2)-log(v1/v2)-1)$

assume(alpha>0)$assume(alpha<1)$
assume(v1>0)$assume(v2>0)$
theta1(mu,v):=mu/v$
theta2(mu,v):=-1/(2*v)$;
F(theta1,theta2):=-theta1**2/(4*theta2)+0.5*log(-1/theta2)$

eq: (theta1(mu1,v1)-theta1(mu2,v2))*mualpha(mu1,v1,mu2,v2,alpha)+(theta2(mu1,v1)
    -theta2(mu2,v2))*(mualpha(mu1,v1,mu2,v2,alpha)**2+varalpha(v1,v2,alpha))-F(
    theta1(mu1,v1),theta2(mu1,v1))+F(theta1(mu2,v2),theta2(mu2,v2));
solalpha: solve(eq,alpha)$
alphastar:rhs(solalpha[1]);

ChernoffInformation: KLD(mualpha(mu1,v1,mu2,v2,alphastar),varalpha(v1,v2,
    alphastar),mu1,v1)$
print("Chernoff information=")$ratsimp(ChernoffInformation);
```

Example of a plot of the α-Bhattacharyya distance for $\alpha \in [0,1]$ when p and q are two normal distributions.

Listing A3: Plot the skewed Bhattacharrya divergences between two normal distributions as an equivalent skewed Jensen divergence between two normal distributions.

```
varalpha(v1,v2,alpha):=(v1*v2)/((1-alpha)*v1+alpha*v2)$
mualpha(mu1,v1,mu2,v2,alpha):=(alpha*mu1*v2+(1-alpha)*mu2*v1)/((1-alpha)*v1+
    alpha*v2)$
assume(v1>0)$assume(v2>0)$
theta1(mu,v):=mu/v$
theta2(mu,v):=-1/(2*v)$;
F(theta1,theta2):=((-theta1**2)/(4*theta2))+(1/2)*log(-%pi/theta2)$
JF(alpha,theta1,theta2,theta1p,theta2p):=alpha*F(theta1,theta2)+(1-alpha)*F(
    theta1p,theta2p)-F(alpha*theta1+(1-alpha)*theta1p,alpha*theta2+(1-alpha)*
    theta2p);
m1:0;v1:1;m2:1;v2:2;
plot2d(JF(alpha,theta1(m1,v1),theta2(m1,v1),theta1(m2,v2),theta2(m2,v2)),[alpha
    ,0,1]);
```

Example which calculates exactly the Chernoff exponent between two centered 4D Gaussians by solving the polynomial roots of the Chernoff optimal condition:

Listing A4: Calculate the Chernoff information between two 4D centered normal distributions based on their eigenvalues.

```
assume(l1>0);assume(l2>0);assume(l3>0);assume(l4>0);
assume(alpha>0);assume(alpha<1);
l1:1;l2:2;l3:3;l4:4;
eq: (1-l1)/(alpha+(1-alpha)*l1)+ (1-l2)/(alpha+(1-alpha)*l2)+ (1-l3)/(alpha+(1-
    alpha)*l3)+ (1-l4)/(alpha+(1-alpha)*l4) + log(l1)+log(l2)+log(l3)+log(l4);
solve(eq,alpha);
sol:float(%);
realpart(sol);imagpart(sol);
/* alpha=0.5969427599369763 */
```

Example of choosing two different exponential families: The half-normal distributions and the exponential distributions:

Listing A5: Calculate symbolically the Kullback–Leibler divergence and the Bhattacharyya coefficient between a half normal distribution and an exponential distribution.

```
assume(sigma>0);
halfnormal(x,sigma):=(sqrt(2)/(sqrt(%pi*sigma**2)))*exp(-x**2/(2*sigma**2));
assume(lambda>0);
exponential(x,lambda):=lambda*exp(-lambda*x);
/* KLD diverges */
integrate(exponential(x,lambda)*log(exponential(x,lambda)/halfnormal(x,sigma)),x
    ,0,inf);
/* KLD converges */
integrate(halfnormal(x,sigma)*log(halfnormal(x,sigma)/exponential(x,lambda)),x
    ,0,inf);
/* Bhattacharyya coefficient */
assume(alpha>0);
assume(alpha<1);
integrate( (halfnormal(x,sigma)**alpha) * (exponential(x,lambda)**(1-alpha)),x
    ,0,inf);
```

References

1. Keener, R.W. *Theoretical Statistics: Topics for a Core Course*; Springer Science & Business Media: New York, NY, USA, 2010.
2. Chernoff, H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.* **1952**, *23*, 493–507. [CrossRef]
3. Csiszár, I. A class of measures of informativity of observation channels. *Period. Math. Hung.* **1972**, *2*, 191–213. [CrossRef]
4. Torgersen, E. *Comparison of Statistical Experiments*; Cambridge University Press: Cambridge, UK, 1991; Volume 36.
5. Audenaert, K.M.; Calsamiglia, J.; Munoz-Tapia, R.; Bagan, E.; Masanes, L.; Acin, A.; Verstraete, F. Discriminating states: The quantum Chernoff bound. *Phys. Rev. Lett.* **2007**, *98*, 160501. [PubMed]
6. Audenaert, K.M.; Nussbaum, M.; Szkoła, A.; Verstraete, F. Asymptotic error rates in quantum hypothesis testing. *Commun. Math. Phys.* **2008**, *279*, 251–283. [CrossRef]

7. Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.* **1943**, *35*, 99–109.

8. Nielsen, F.; Boltz, S. The Burbea-Rao and Bhattacharyya centroids. *IEEE Trans. Inf. Theory* **2011**, *57*, 5455–5466. [CrossRef]

9. Grünwald, P.D. *The Minimum Description Length Principle*; MIT Press: Cambridge, MA, USA, 2007.

10. Grünwald, P.D. Information-Theoretic Properties of Exponential Families. In *The Minimum Description Length Principle*; MIT Press: Cambridge, MA, USA, 2007; pp. 623–650.

11. Van Erven, T.; Harremos, P. Rényi divergence and Kullback–Leibler divergence. *IEEE Trans. Inf. Theory* **2014**, *60*, 3797–3820. [CrossRef]

12. Nakiboğlu, B. The Rényi capacity and center. *IEEE Trans. Inf. Theory* **2018**, *65*, 841–860. [CrossRef]

13. Cover, T.M. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 1999.

14. Borade, S.; Zheng, L. *I*-projection and the geometry of error exponents. In Proceedings of the Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 27–29 September 2006.

15. Boyer, R.; Nielsen, F. On the error exponent of a random tensor with orthonormal factor matrices. In *International Conference on Geometric Science of Information*; Springer: Cham, Switzerland, 2017; pp. 657–664.

16. D'Costa, A.; Ramachandran, V.; Sayeed, A.M. Distributed classification of Gaussian space-time sources in wireless sensor networks. *IEEE J. Sel. Areas Commun.* **2004**, *22*, 1026–1036. [CrossRef]

17. Yu, N.; Zhou, L. Comments on and Corrections to "When Is the Chernoff Exponent for Quantum Operations Finite?". *IEEE Trans. Inf. Theory* **2022**, *68*, 3989–3990. [CrossRef]

18. Konishi, S.; Yuille, A.L.; Coughlan, J.; Zhu, S.C. Fundamental bounds on edge detection: An information theoretic evaluation of different edge cues. In Proceedings of the 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), Fort Collins, CO, USA, 23–25 June 1999; IEEE: Piscataway, NJ, USA, 1999; Volume 1, pp. 573–579.

19. Julier, S.J. An empirical study into the use of Chernoff information for robust, distributed fusion of Gaussian mixture models. In Proceedings of the 2006 9th International Conference on Information Fusion, Florence, Italy, 10–13 July 2006; IEEE: Piscataway, NJ, USA, 2006; pp. 1–8.

20. Kakizawa, Y.; Shumway, R.H.; Taniguchi, M. Discrimination and clustering for multivariate time series. *J. Am. Stat. Assoc.* **1998**, *93*, 328–340. [CrossRef]

21. Dutta, S.; Wei, D.; Yueksel, H.; Chen, P.Y.; Liu, S.; Varshney, K. Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing. In Proceedings of the 37th International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 2803–2813.

22. Agarwal, S.; Varshney, L.R. Limits of deepfake detection: A robust estimation viewpoint. *arXiv* **2019**, arXiv:1905.03493.

23. Maherin, I.; Liang, Q. Radar sensor network for target detection using Chernoff information and relative entropy. *Phys. Commun.* **2014**, *13*, 244–252. [CrossRef]

24. Nielsen, F. An information-geometric characterization of Chernoff information. *IEEE Signal Process. Lett.* **2013**, *20*, 269–272. [CrossRef]

25. Nielsen, F. Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means. *Pattern Recognit. Lett.* **2014**, *42*, 25–34. [CrossRef]

26. Westover, M.B. Asymptotic geometry of multiple hypothesis testing. *IEEE Trans. Inf. Theory* **2008**, *54*, 3327–3329. [CrossRef]

27. Nielsen, F. Hypothesis testing, information divergence and computational geometry. In *International Conference on Geometric Science of Information*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 241–248.

28. Leang, C.C.; Johnson, D.H. On the asymptotics of *M*-hypothesis Bayesian detection. *IEEE Trans. Inf. Theory* **1997**, *43*, 280–282. [CrossRef]

29. Cena, A.; Pistone, G. Exponential statistical manifold. *Ann. Inst. Stat. Math.* **2007**, *59*, 27–56. [CrossRef]

30. Barndorff-Nielsen, O. *Information and Exponential Families: In Statistical Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2014.

31. Brekelmans, R.; Nielsen, F.; Makhzani, A.; Galstyan, A.; Steeg, G.V. Likelihood Ratio Exponential Families. *arXiv* **2020**, arXiv:2012.15480.

32. De Andrade, L.H.; Vieira, F.L.; Vigelis, R.F.; Cavalcante, C.C. Mixture and exponential arcs on generalized statistical manifold. *Entropy* **2018**, *20*, 147. [CrossRef]

33. Siri, P.; Trivellato, B. Minimization of the Kullback–Leibler Divergence over a Log-Normal Exponential Arc. In *International Conference on Geometric Science of Information*; Springer: Cham, Switzerland, 2019; pp. 453–461.

34. Azoury, K.S.; Warmuth, M.K. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Mach. Learn.* **2001**, *43*, 211–246. [CrossRef]

35. Collins, M.; Dasgupta, S.; Schapire, R.E. A generalization of principal components analysis to the exponential family. *Adv. Neural Inf. Process. Syst.* **2001**, *14*, 617–624

36. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman divergences. *J. Mach. Learn. Res.* **2005**, *6*. [CrossRef]

37. Nielsen, F.; Nock, R. Sided and symmetrized Bregman centroids. *IEEE Trans. Inf. Theory* **2009**, *55*, 2882–2904. [CrossRef]

38. Sundberg, R. *Statistical Modelling by Exponential Families*; Cambridge University Press: Cambridge, UK, 2019; Volume 12.

39. Nielsen, F.; Okamura, K. On *f*-divergences between Cauchy distributions. *arXiv* **2021**, arXiv:2101.12459.

40. Chyzak, F.; Nielsen, F. A closed-form formula for the Kullback–Leibler divergence between Cauchy distributions. *arXiv* **2019**, arXiv:1905.10965

41. Huzurbazar, V.S. Exact forms of some invariants for distributions admitting sufficient statistics. *Biometrika* **1955**, *42*, 533–537. [CrossRef]
42. Burbea, J.; Rao, C. On the convexity of some divergence measures based on entropy functions. *IEEE Trans. Inf. Theory* **1982**, *28*, 489–495. [CrossRef]
43. Chen, P.; Chen, Y.; Rao, M. Metrics defined by Bregman divergences: Part 2. *Commun. Math. Sci.* **2008**, *6*, 927–948. [CrossRef]
44. Nielsen, F. On the Jensen–Shannon symmetrization of distances relying on abstract means. *Entropy* **2019**, *21*, 485. [CrossRef]
45. Han, Q.; Kato, K. Berry–Esseen bounds for Chernoff-type nonstandard asymptotics in isotonic regression. *Ann. Appl. Probab.* **2022**, *32*, 1459–1498. [CrossRef]
46. Neal, R.M. Annealed importance sampling. *Stat. Comput.* **2001**, *11*, 125–139. [CrossRef]
47. Grosse, R.B.; Maddison, C.J.; Salakhutdinov, R. Annealing between distributions by averaging moments. In *Advances in Neural Information Processing Systems 26, Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, 5–10 December 2013*; Citeseer: La Jolla, CA, USA, 2013; pp. 2769–2777.
48. Takenouchi, T. Parameter Estimation with Generalized Empirical Localization. In *International Conference on Geometric Science of Information*; Springer: Cham, Switzerland, 2019; pp. 368–376.
49. Rockafellar, R.T. Conjugates and Legendre transforms of convex functions. *Can. J. Math.* **1967**, *19*, 200–205. [CrossRef]
50. Del Castillo, J. The singly truncated normal distribution: A non-steep exponential family. *Ann. Inst. Stat. Math.* **1994**, *46*, 57–66. [CrossRef]
51. Amari, S.I. *Information Geometry and Its Applications*; Springer: Tokyo, Japan, 2016; Volume 194.
52. Boissonnat, J.D.; Nielsen, F.; Nock, R. Bregman Voronoi diagrams. *Discret. Comput. Geom.* **2010**, *44*, 281–307. [CrossRef]
53. Lê, H.V. Statistical manifolds are statistical models. *J. Geom.* **2006**, *84*, 83–93. [CrossRef]
54. Nielsen, F. On a Variational Definition for the Jensen–Shannon Symmetrization of Distances Based on the Information Radius. *Entropy* **2021**, *23*, 464. [CrossRef]
55. Nock, R.; Nielsen, F. Fitting the smallest enclosing Bregman ball. In Proceedings of the European Conference on Machine Learning, Porto, Portugal, 3–7 October 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 649–656.
56. Nielsen, F.; Nock, R. On the smallest enclosing information disk. *Inf. Process. Lett.* **2008**, *105*, 93–97. [CrossRef]
57. Costa, R. Information Geometric Probability Models in Statistical Signal Processing. Ph.D. Thesis, University of Rhode Island, Kingston, RI, USA, 2016.
58. Nielsen, F.; Garcia, V. Statistical exponential families: A digest with flash cards. *arXiv* **2009**, arXiv:0911.4863.
59. Ali, S.M.; Silvey, S.D. A general class of coefficients of divergence of one distribution from another. *J. R. Stat. Soc. Ser. B (Methodol.)* **1966**, *28*, 131–142. [CrossRef]
60. Nielsen, F.; Okamura, K. A note on the f-divergences between multivariate location-scale families with either prescribed scale matrices or location parameters. *arXiv* **2022**, arXiv:2204.10952.
61. Athreya, A.; Fishkind, D.E.; Tang, M.; Priebe, C.E.; Park, Y.; Vogelstein, J.T.; Levin, K.; Lyzinski, V.; Qin, Y. Statistical inference on random dot product graphs: A survey. *J. Mach. Learn. Res.* **2017**, *18*, 8393–8484.
62. Li, B.; Wei, S.; Wang, Y.; Yuan, J. Topological and algebraic properties of Chernoff information between Gaussian graphs. In Proceedings of the 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 2–5 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 670–675.
63. Tang, M.; Priebe, C.E. Limit theorems for eigenvectors of the normalized Laplacian for random graphs. *Ann. Stat.* **2018**, *46*, 2360–2415. [CrossRef]
64. Calvo, M.; Oller, J.M. An explicit solution of information geodesic equations for the multivariate normal model. *Stat. Risk Model.* **1991**, *9*, 119–138. [CrossRef]
65. Boyd, S.P.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.
66. Chen, P.; Chen, Y.; Rao, M. Metrics defined by Bregman divergences. *Commun. Math. Sci.* **2008**, *6*, 915–926. [CrossRef]
67. Kailath, T. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. Commun. Technol.* **1967**, *15*, 52–60. [CrossRef]
68. Csiszar, I. Eine information's theoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoschen Ketten. *Publ. Math. Inst. Hung. Acad. Sc.* **1963**, *3*, 85–107.
69. Deza, M.M.; Deza, E. Encyclopedia of distances. In *Encyclopedia of Distances*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1–583.
70. Gibbs, A.L.; Su, F.E. On choosing and bounding probability metrics. *Int. Stat. Rev.* **2002**, *70*, 419–435. [CrossRef]
71. Jian, B.; Vemuri, B.C. Robust point set registration using Gaussian mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 1633–1645. [CrossRef]
72. Nielsen, F.; Nock, R. Entropies and cross-entropies of exponential families. In Proceedings of the 2010 IEEE International Conference on Image Processing, Hong Kong, China, 26–29 September 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 3621–3624.