

# Revisiting Chernoff information with Likelihood Ratio Exponential Families

Frank Nielsen

Sony Computer Science Laboratories Inc.

Tokyo, Japan

*<https://franknielsen.github.io/>*



**Sony CSL**

**Entropy 2022**  
**[2207.03745]**

# Chernoff information:

## A symmetric statistical divergence

- Originally introduced by Chernoff (1952) to *upper bound the probability of error* (Bayes' error) in statistical hypothesis testing.

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

$$\rho_\alpha[P : Q] := \int p^\alpha q^{1-\alpha} d\mu = \rho_{1-\alpha}[Q : P] \quad 0 < \rho_\alpha[P : Q] \leq 1.$$

(via Hölder inequality)



**Herman Chernoff**  
(1923-)

- skewed Bhattacharyya coefficient**  $\rho_\alpha$  (similarity)
- Synonyms: Chernoff divergence, Chernoff information number, Chernoff index...
- Found later many applications in information fusion, radar target detection, generative adversarial networks, etc. due to its **empirical robustness**

# Chernoff information = Maximally skewed Bhattacharyya distance

- **skewed Bhattacharyya distance** (a Ali-Silvey **f-divergence**):

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

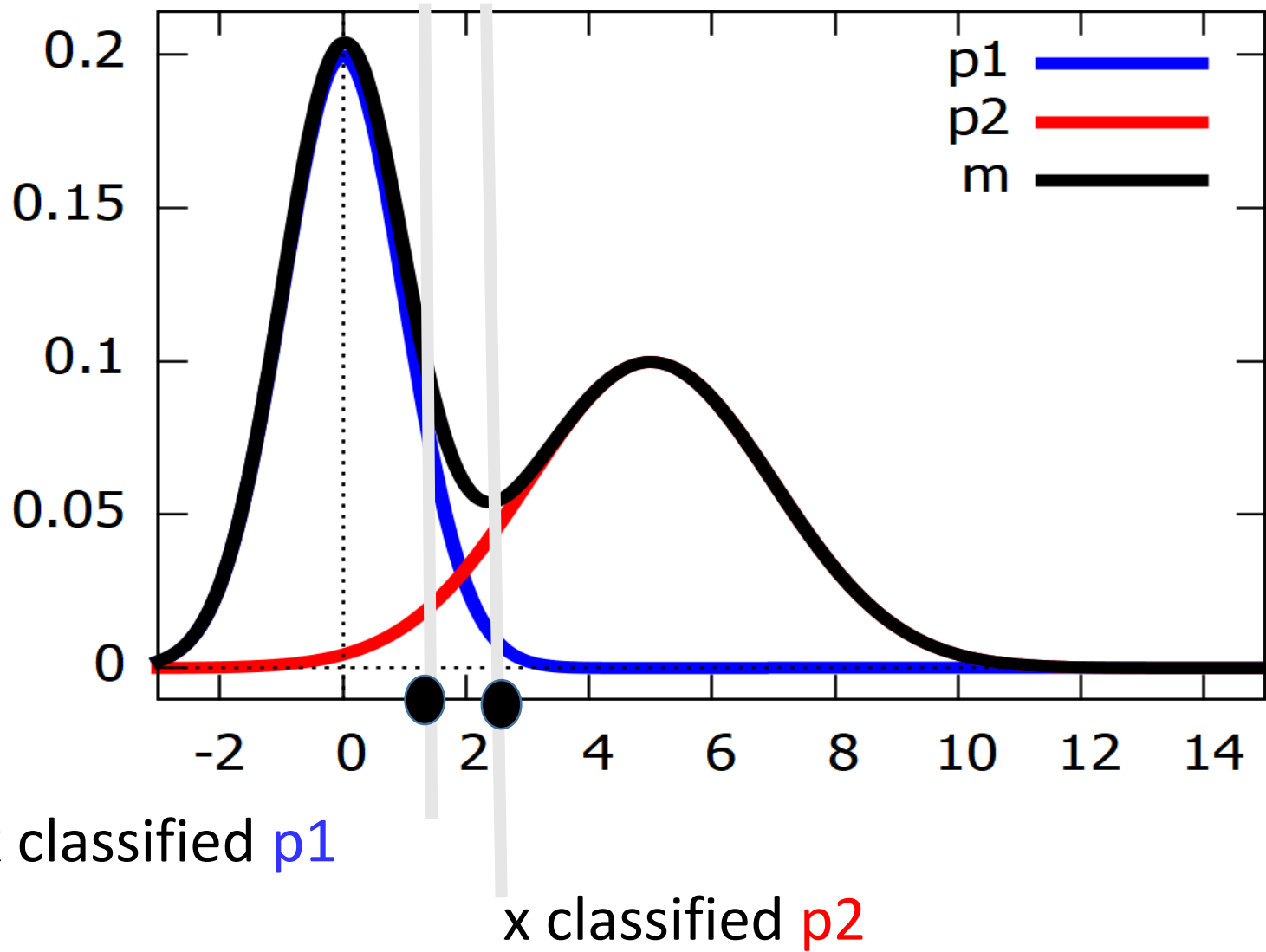
- Chernoff information:  $D_C[p, q] = \max_{\alpha \in (0,1)} D_{B,\alpha}[p : q].$

- **scaled skewed Bhattacharyya distance = Rényi divergence** (extends KLD)

$$D_{R,\alpha}[P : Q] = \frac{1}{\alpha - 1} \log \int p^\alpha q^{1-\alpha} d\mu = \frac{1}{1 - \alpha} D_{B,\alpha}[P : Q] \quad \alpha \in [0, \infty] \setminus \{1\}$$

- Optimal values of  $\alpha$  is called ``**Chernoff (error) exponent**'' (due to its seminal use in statistical hypothesis testing)

# Rationale: Statistical hypothesis testing



Statistical mixture:

$$m(x) = 0.5 * N(0,1) + 0.5 * N(5,2)$$

Hypothesis task:

Decides whether  $x$  emanates from  $p1$  or  $p2$ ?

Classification rule:

**Maximum a posteriori** (MAP)

if  $p1(x) > p2(x)$  classify as  $p1$   
else classify as  $p2$

Error at  $x$ :  $\min(p1(x), p2(x))$

Histogram intersection similarity:

$$P_e = \int \min(p_1(x), p_2(x)) dx$$

# Rewriting and bounding the probability of error

- Use **rewriting trick**  $\min(a,b)=(a+b)/2 + |b-a|/2$  for  $a,b>0$

express the probability of error using the **total variation distance**:

$$P_e = \int \min(p_1(x), p_2(x)) dx \quad \longrightarrow \quad P_e = \frac{1}{2} (1 - D_{\text{TV}}[p_1, p_2])$$
$$D_{\text{TV}}[p_1, p_2] = \frac{1}{2} \int (p_1(x) - p_2(x)) dx$$

- Use a **generic (weighted) mean** which necessarily falls inbetween its extrema:

$$\min(a, b) \leq M(a, b) \leq \max(a, b) \quad \longrightarrow \quad \min(a, b) \leq M_\alpha(a, b) \leq \max(a, b), \forall \alpha \in [0, 1]$$

$$P_e = \int \min(p_1(x), p_2(x)) dx \leq \min_{\alpha \in [0,1]} \int M_\alpha(p_1(x), p_2(x)) dx \quad \xrightarrow[\text{geometric weighted mean}]{M_\alpha(a, b) = a^\alpha b^{1-\alpha}} \quad P_e \leq \rho_\alpha(p_1, p_2)$$

"Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means." *Pattern Recognition Letters* 42 (2014): 25-34.

# Outline of the contributions of this talk

- (Background, done)
- **Interplay of non-parametric with parametric study** of Chernoff information via the concept of log-ratio exponential families (LREFs)
- Derive various **optimality conditions** for the chernoff exponent  $\alpha^*$
- Give some **geometric interpretations** on Bregman manifolds which yield fast approximations algorithms
- **novel closed-form solutions** for the Chernoff information between univariate Gaussians, centered scaled covariance matrices, etc.

# Likelihood ratio exponential families (LREFs)

- Geometric mixture (Bhattacharyya /exponential arc )  $(pq)_\alpha^G(x) \propto p(x)^\alpha q(x)^{1-\alpha}$   
between two densities  $p, q$  of Lebesgue Banach space  $L_1(\mu)$

- Set of **geometric mixtures**:  
with **normalization factor**: 
$$\mathcal{E}_{pq} := \left\{ (pq)_\alpha^G(x) := \frac{p(x)^\alpha q(x)^{1-\alpha}}{Z_{pq}(\alpha)} : \alpha \in \Theta \right\}$$
$$Z_{pq}(\alpha) = \int_{\mathcal{X}} p(x)^\alpha q(x)^{1-\alpha} d\mu(x) = \underline{\rho_\alpha[p : q]}$$

- geometric mixture interpreted as a **1D exponential family**: LREF

$$\begin{aligned} (pq)_\alpha^G(x) &= \exp\left(\alpha \log \frac{p(x)}{q(x)} - \log Z_{pq}(\alpha)\right) q(x), \\ &\stackrel{*}{=} \exp(\alpha t(x) - F_{pq}(\alpha) + k(x)). \end{aligned}$$

$\swarrow$   $\swarrow$   $\swarrow$   
 $t(x) = \log \frac{p(x)}{q(x)}$   $F_{pq}(\alpha) = -D_{B,\alpha}[p : q]$   $k(x) = \log q(x)$

Natural parameter space:

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

# LREFs: EF cumulant function is always analytic $C^\omega$

- Cumulant function of EF is **strictly convex** (and smooth for regular EFs)

- Cumulant function is neg-Bhattacharyya distance:

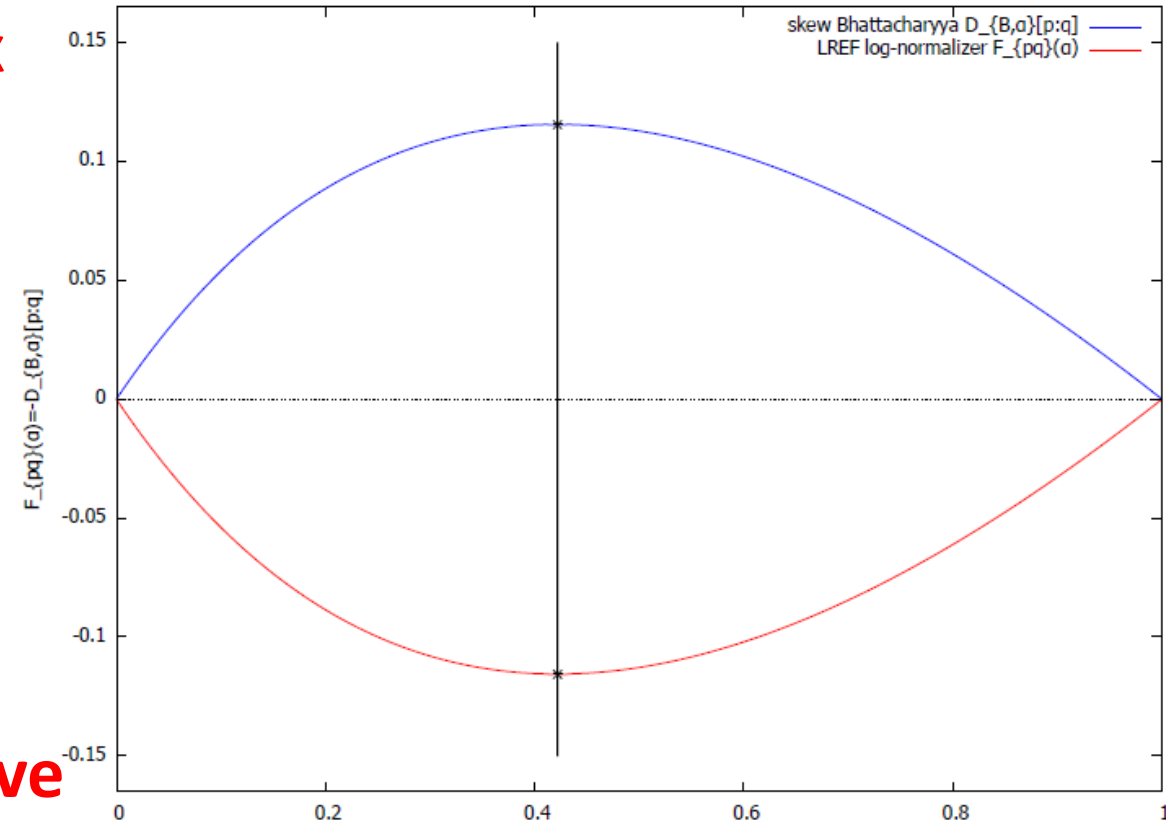
$$F_{pq}(\alpha) = \log Z_{pq}(\alpha) = -D_{B,\alpha}[p : q] < 0$$

⇒ Bhattacharyya. distance is **strictly concave**

- Theorem:

**Chernoff exponent exists and is unique**

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



$$p = N(0, 1)$$

$$q = N(1, 2)$$

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

$$\alpha^*(q : p) = 1 - \alpha^*(p : q)$$



# Geometric mixtures and LREFs: Regular EFs

- Natural parameter space:  $\Theta_{pq} = \{\alpha \in \mathbb{R} : \rho_\alpha(p : q) < +\infty\}$

**always contains (0,1)** since  $0 < \rho_\alpha[P : Q] \leq 1$ .

- What happens at extremities and when extrapolating (support):

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

- Exponential family is said **regular** when the natural parameter space  $\Theta$  is **open** (e.g., normal family, Dirichlet family, Wishart family, etc.)

Definition:      regular EF             $\Theta = \Theta^\circ$

# When (0,1) is strictly included in regular LREFs

**Proposition** (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  $D_{\text{KL}}[p : q]$  and the reverse Kullback-Leibler divergence  $D_{\text{KL}}[q : p]$  are finite.*

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

- **KLD between two densities of a regular EF = reverse Bregman divergence:**

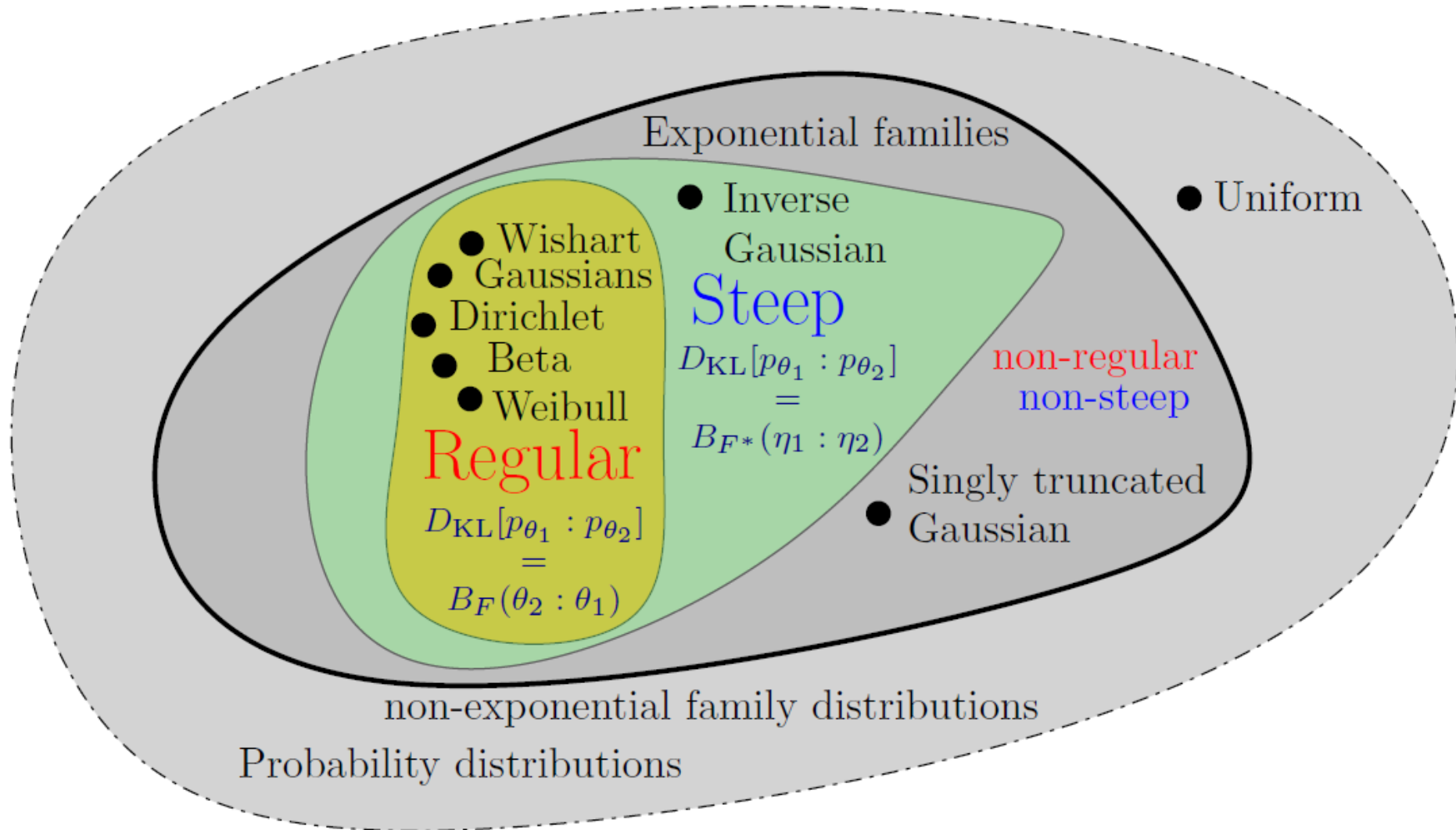
$$\begin{aligned} D_{\text{KL}}[p_{\theta_1} : p_{\theta_2}] &= E_{p_{\theta_1}} \left[ \log \frac{p_{\theta_1}}{p_{\theta_2}} \right], \\ &= F(\theta_2) - F(\theta_1) - (\theta_1 - \theta_2)^\top E_{p_{\theta_1}}[t(x)]. \end{aligned}$$

steep  $\Rightarrow E_{p_{\theta_1}}[t(x)] = \nabla F(\theta_1)$   
regular EF  $\Rightarrow$  steep EF

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

# Venn diagram: Regular & steepness of (LR)EFs

- Steepness implies **duality between natural  $\theta$  and moment  $\eta$  parameters**



**Proposition** (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  $D_{\text{KL}}[p : q]$  and the reverse Kullback-Leibler divergence  $D_{\text{KL}}[q : p]$  are finite.

## PROOF

Remember KLD=Bregman divergence between densities of a **regular (LR)EF**

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

Scalar Bregman divergence  $B_{F_{pq}} : \Theta \times \text{ri}(\Theta) \rightarrow [0, \infty)$

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

$$F_{pq}(0) = F_{pq}(1) = 0$$

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

idem for

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

# Chernoff information (for densities of a LREF)

- Proposition:  $D_C[p : q] = D_{\text{KL}}[(pq)_{\alpha^*}^G : p] = D_{\text{KL}}[(pq)_{\alpha^*}^G : q].$

## PROOF

First, **Bhattacharyya distance = Jensen divergence**

$$D_{B,\alpha}[p : q] := -\log \rho_\alpha[P : Q] \quad \longrightarrow \quad D_{B,\alpha}(p_{\theta_1} : p_{\theta_2}) = J_{F,\alpha}(\theta_1 : \theta_2).$$

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

Thus 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  $\alpha^*$ , we have  $F'_{pq}(\alpha^*) = 0$

$$\begin{aligned} \textcircled{1} \quad D_{\text{KL}}[(pq)_{\alpha^*}^G : p] &= B_{F_{pq}}(-F(\alpha^*)) & \textcircled{2} \quad D_{\text{KL}}[(pq)_{\alpha^*}^G : q] &= B_{F_{pq}}(0 : \alpha^*) = -F(\alpha^*) \\ \textcircled{3} \quad D_C[p : q] &= -\log \rho_{\alpha^*}(p : q) = J_{F_{pq},\alpha^*}(1 : 0) = -F_{pq}(\alpha^*) \end{aligned}$$

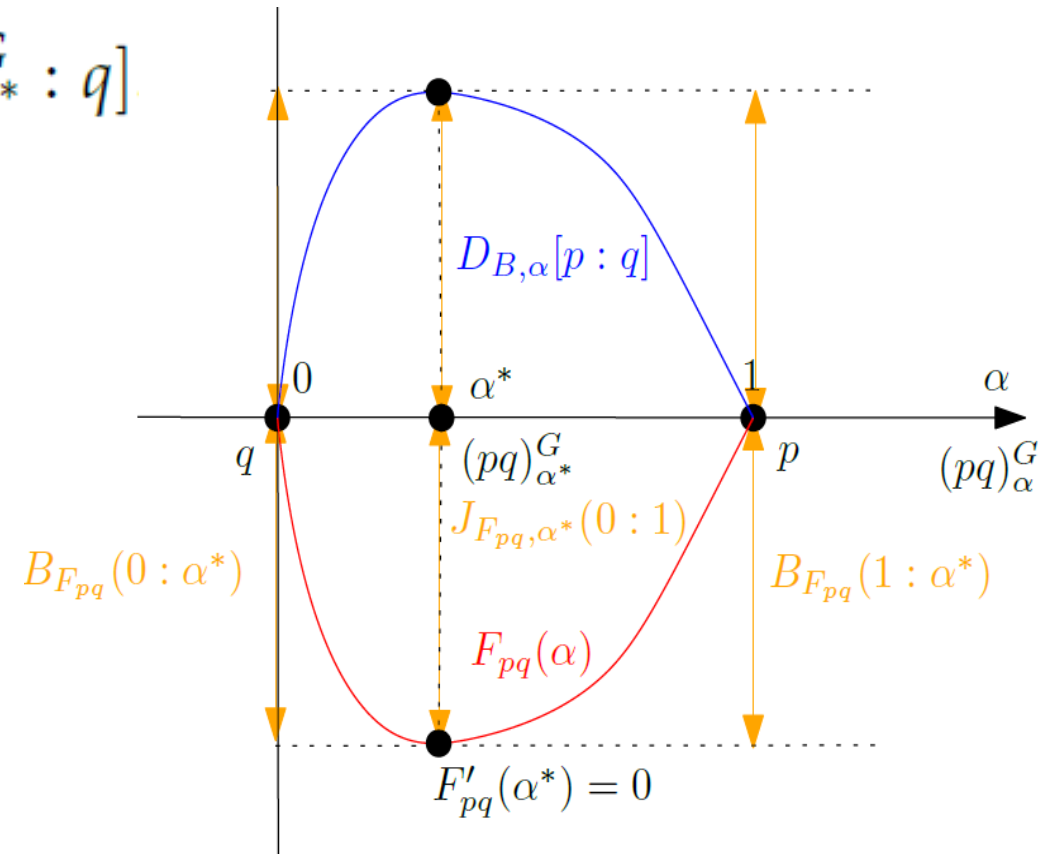
# Jensen-Chernoff divergence

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

non-parametric arguments

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

scalar parametric arguments



In general, define **Jensen-Chernoff divergence**

$$J_F^C(\theta_1 : \theta_2) := \max_{\alpha \in (0,1)} J_{F, \alpha}(\theta_1 : \theta_2)$$

# Geometric interpretation for densities on $L_1(\mu)$

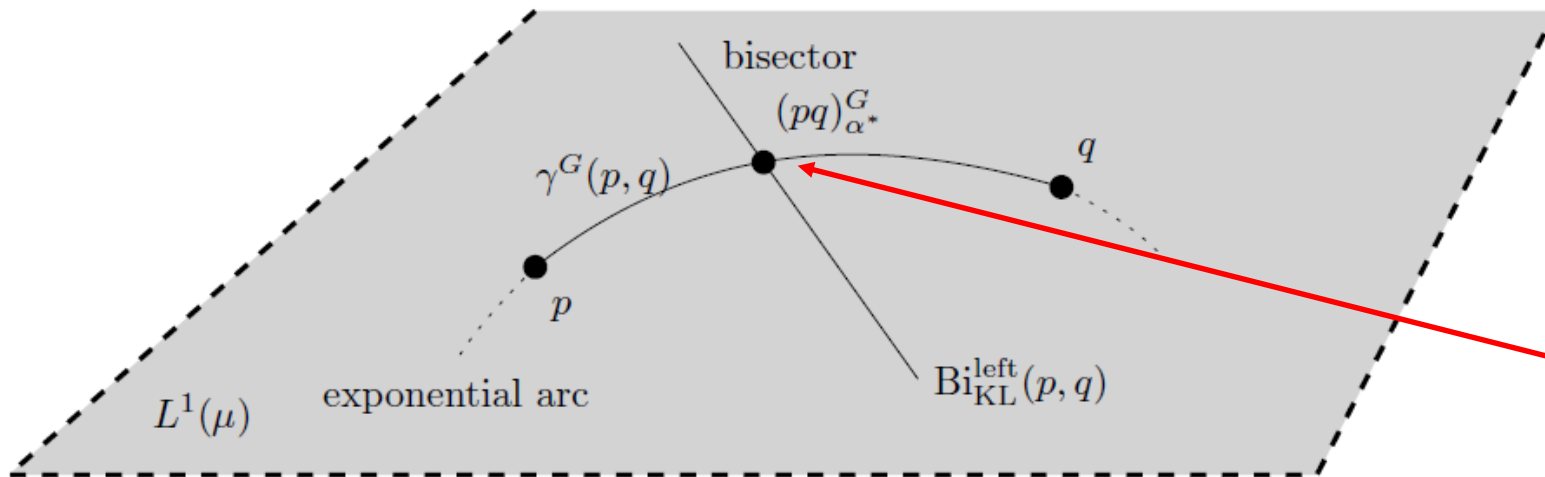
**Proposition** (Geometric characterization of the Chernoff information). *On the vector space  $L^1(\mu)$ , the Chernoff information distribution is the unique distribution*

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

**Left KL Voronoi bisector:**  $\text{Bi}_{\text{KL}}^{\text{left}}(p, q) := \left\{ r \in L^1(\mu) : D_{\text{KL}}[\underline{r} : p] = D_{\text{KL}}[\underline{r} : q] \right\}$

**Geodesic** = exponential arc:  $\gamma^G(p, q) := \left\{ (pq)_{\alpha}^G : \alpha \in [0, 1] \right\}$

2209.07481



**Chernoff point:**  $(pq)_{\alpha^*}^G$



# Fast dichotomic search for approximating the Chernoff point

**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**

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

**if**  $D_{\text{KL}}[(pq)_\alpha^G : p] > D_{\text{KL}}[(pq)_\alpha^G : q]$  **then**

$\alpha_m = \alpha;$

        // See Figure

**end**

**else**

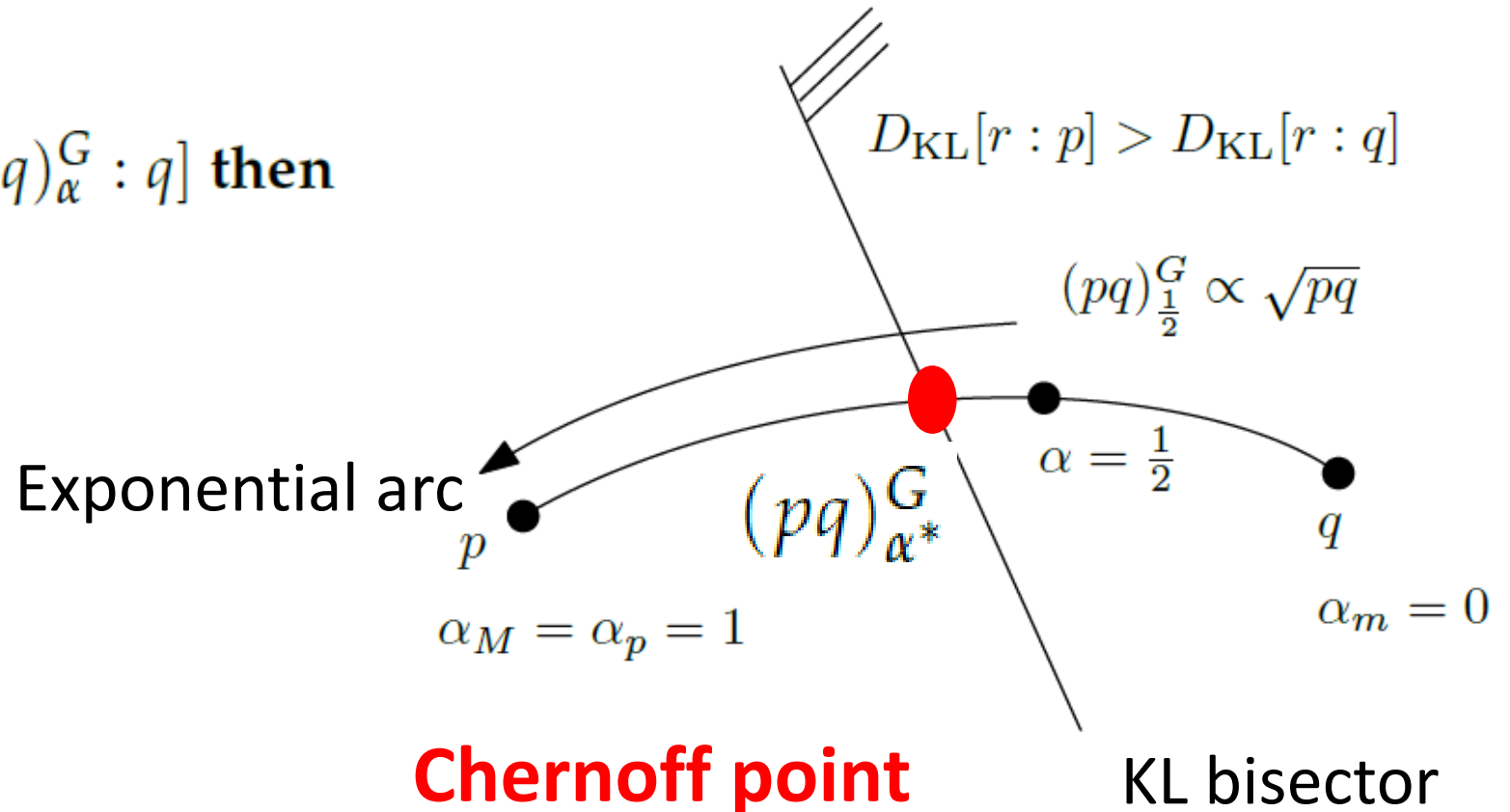
$\alpha_M = \alpha;$

**end**

**end**

**return**  $D_{\text{KL}}[(pq)_\alpha^G : p];$

Bisection algorithm:





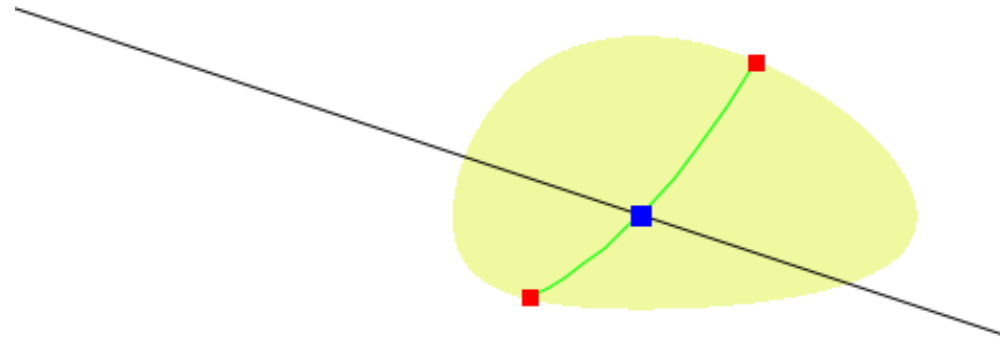
# Chernoff information as a symmetrization of KLD

Rewrite  $D_C[p : q] = D_{\text{KL}}[(pq)_{\alpha^*}^G : p] = D_{\text{KL}}[(pq)_{\alpha^*}^G : q]$

as  $D_C[p : q] = \min_{r \in \mathcal{E}_{pq}} \{D_{\text{KL}}[r : p], D_{\text{KL}}[r : q]\}.$

Chernoff information as the radius of  
a **minimum enclosing left-sided Kullback–Leibler ball**

circum center  $r$



Chernoff point  $r^*$   
(eKLD)

# Dual moment parameterizations $\beta$ of LREFs

**Dual moment parameter:**  $\beta = \beta(\alpha) := E_{(pq)_{\alpha}^G}[t(x)] = \underline{E_{(pq)_{\alpha}^G} \left[ \log \frac{p(x)}{q(x)} \right]}$   
 $\alpha$  = natural primal parameter

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

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

$$\beta(\alpha) = F'_{pq}(\alpha) \quad \longleftrightarrow \quad \text{Legendre transform} \quad \alpha = F_{pq}^{*'}(\beta),$$

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

# Dual parameterizations of LREFs and optimality condition for finding Chernoff exponent

$$F'_{pq}(\alpha^*) = 0$$



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

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

**Primal optimality condition**

**Dual optimality condition**

# Special case of LREF: $p, q$ are densities of an EF!

EF includes Gaussians, Beta, Dirichlet, Wishart, etc.

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

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

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

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

$$\text{OC}_{\text{EF}} : \quad B_F(\theta_1 : \theta_{\alpha^*}) = B_F(\theta_2 : \theta_{\alpha^*})$$

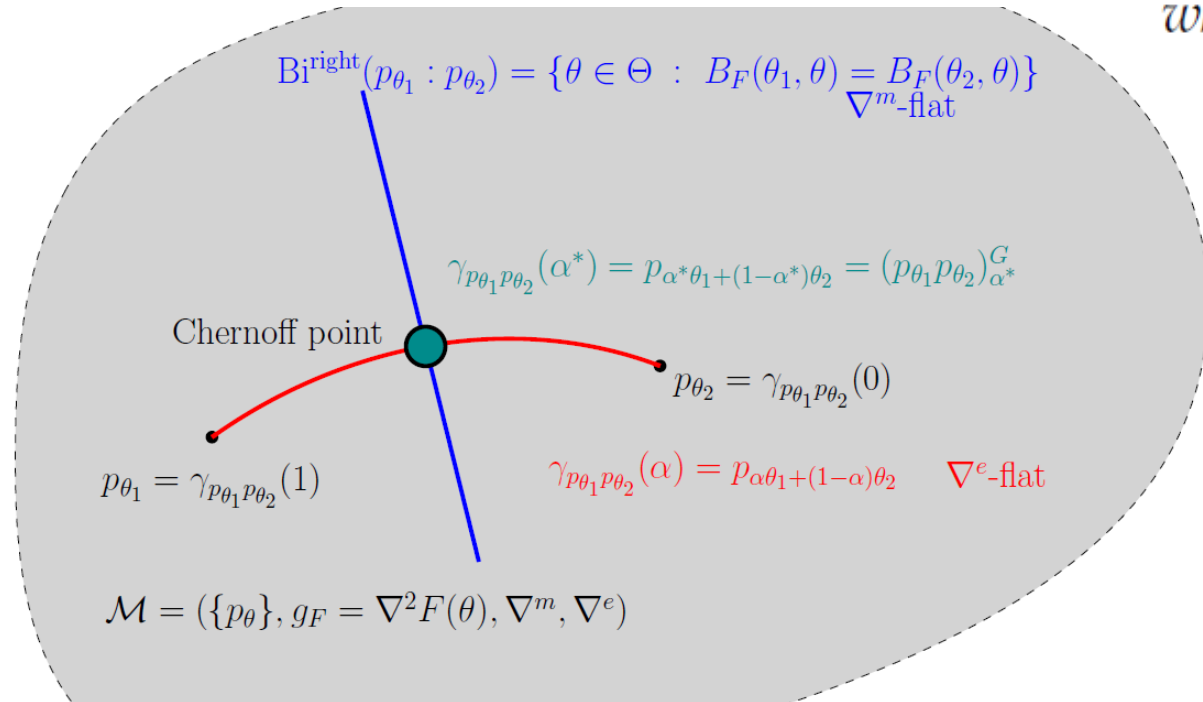
**Proposition** Let  $p_{\lambda_1}$  and  $p_{\lambda_2}$  be two densities of a regular exponential family  $\mathcal{E}$  with natural parameter  $\theta(\lambda)$  and log-normalizer  $F(\theta)$ . Then the Chernoff information is

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

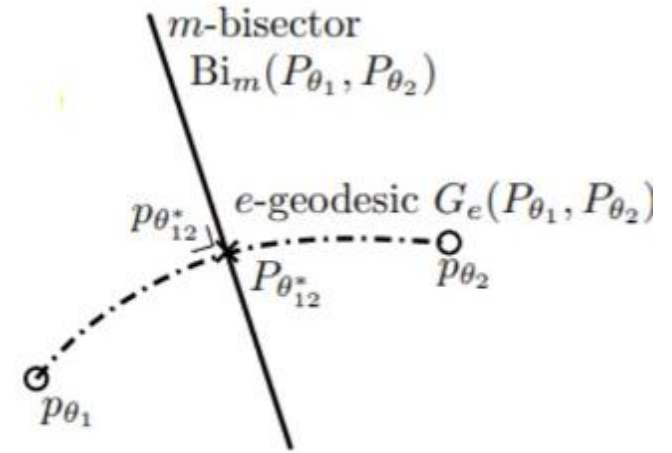
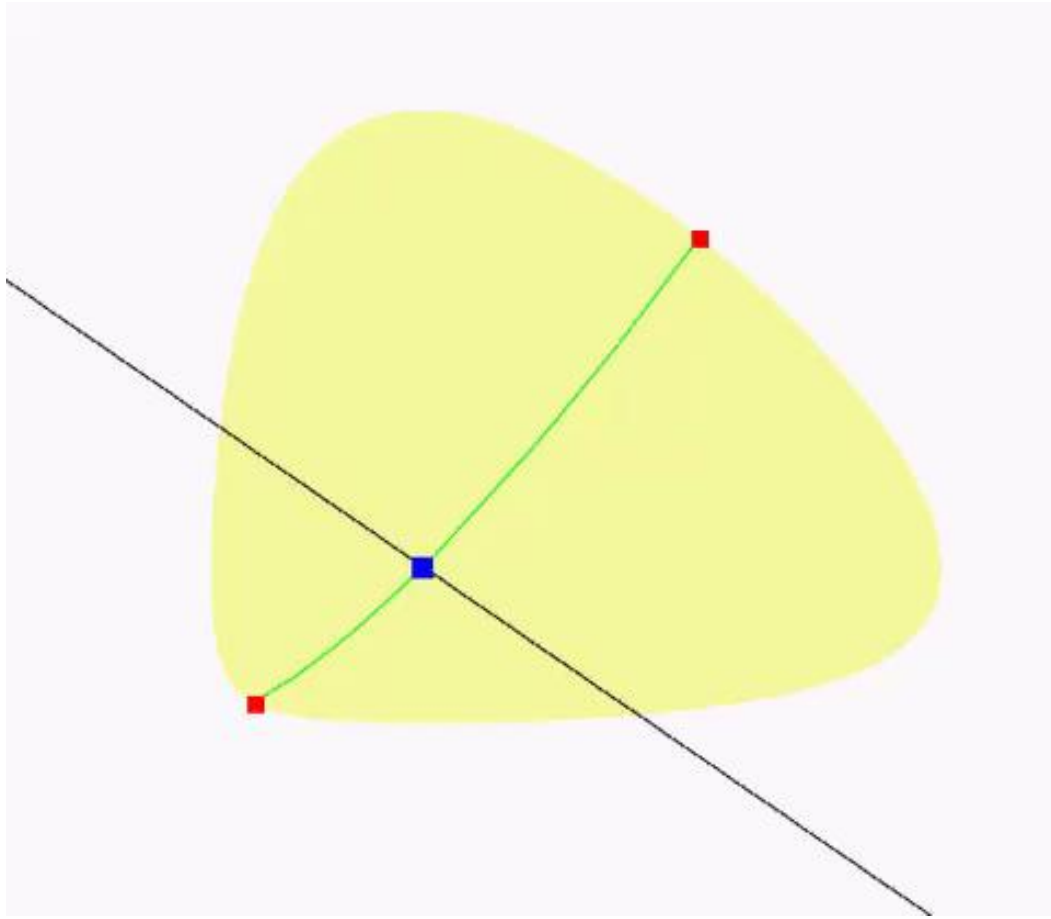
where  $\theta_1 = \theta(\lambda_1)$ ,  $\theta_2 = \theta(\lambda_2)$ , and the optimal skewing parameter  $\alpha^*$  is unique and satisfies the following optimality condition:

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

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



# Chernoff point (in blue) for extended KLD case



input distributions

Chernoff point

exponential arc (e-flat dim. 1)

Voronoi bisector (m-flat codim. 1)

# Interpreting the uniqueness of Chernoff exponent from information geometry point of view

- Since the Chernoff point is unique, we can also interpret more generally this property in a general dually flat space (not necessarily an EF) as known as a **Bregman manifold**

**Proposition** *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  $\nabla^m$ -flat and left-sided  $\nabla^e$ -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.*

**"On geodesic triangles with right angles in a dually flat space."  
Progress in Information Geometry. Springer, 2021. 153-190.**



**Proposition** Let  $p_{\lambda_1}$  and  $p_{\lambda_2}$  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  $\alpha^*$  is unique and satisfies the following optimality condition:

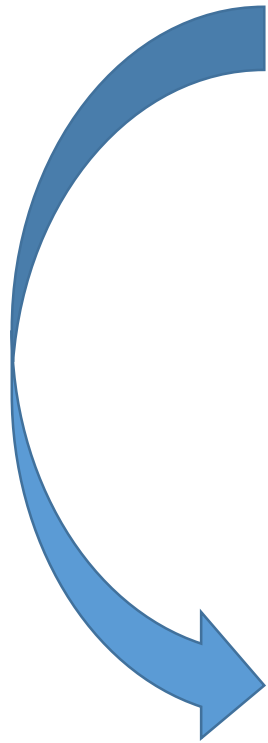
$$\text{OC}_{\text{EF}} : (\theta_2 - \theta_1)^\top \eta_{\alpha^*} = F(\theta_2) - F(\theta_1),$$

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

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

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

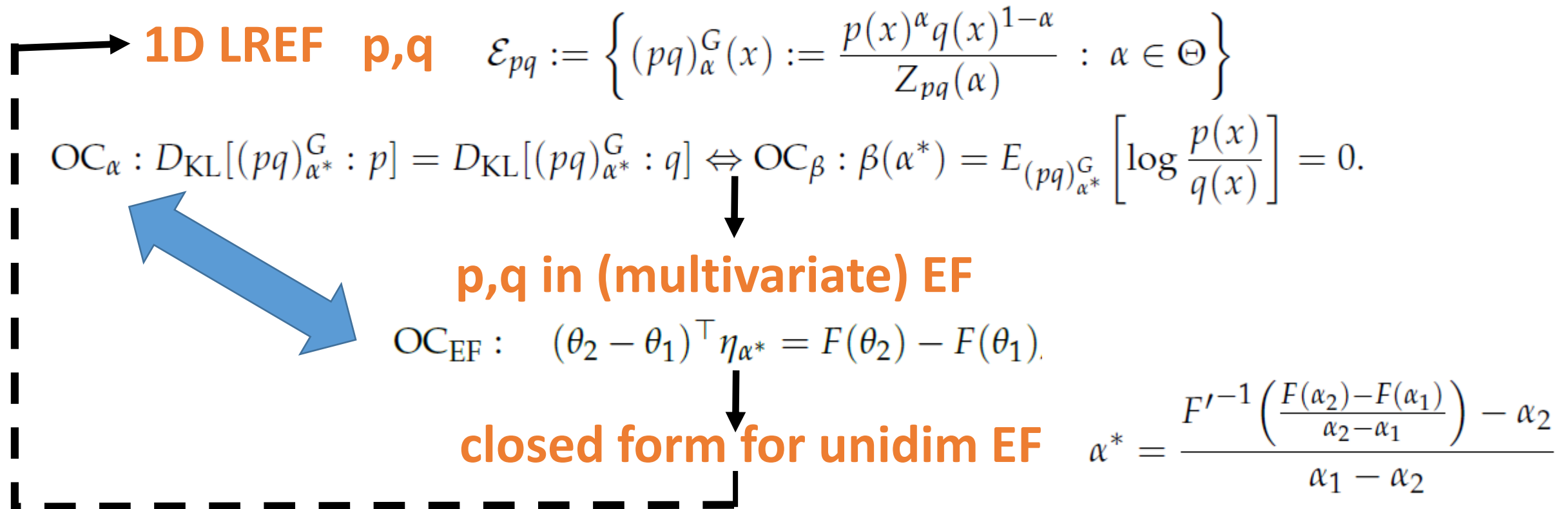
Unidimensional case





# Recap: Closing the loop!

- Started from two densities  $p, q$  of  $L_1(\mu)$ , built LREF and got 1D optimal condition
- Applied to case where  $p, q$  are densities of the same exponential family
- In particular get closed-form for univariate 1D EFs = LREFs



# Bhattacharyya distances between Gaussian densities

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

Bhattacharyya distance (non-metric) in closed form:

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

where

$$\begin{aligned} \Sigma_{\alpha} &= (\alpha \Sigma_1^{-1} + (1 - \alpha) \Sigma_2^{-1})^{-1}, & \text{Harmonic weighted mean} \\ \mu_{\alpha} &= \Sigma_{\alpha} (\alpha \Sigma_1^{-1} \mu_1 + (1 - \alpha) \Sigma_2^{-1} \mu_2). \end{aligned}$$

# Invariance under the action of the affine group

Affine group action:

$$(l_1, A_1) \cdot (l_2, A_2) = (l_1 + A_1 l_2, A_1 A_2)$$

Matrix group element representation:

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

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

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

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

where  $\mu_{12} = \Sigma_1^{-\frac{1}{2}}(\mu_2 - \mu_1)$

Wlog, we can assume this canonical case

$$\Sigma_{12} = \Sigma_1^{-\frac{1}{2}} \Sigma_2 \Sigma_1^{-\frac{1}{2}}.$$

# Exact closed form for Chernoff information between same-covariance matrix Gaussians

- Trivial case,  $\alpha^*=1/2$

$$D_C[p_{\mu_1, \sigma^2} : p_{\mu_2, \sigma^2}] = \frac{(\mu_2 - \mu_1)^2}{8\sigma^2}.$$

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

**square Mahalanobis distance:**  $\Delta_{\Sigma}^2(\mu_1, \mu_2) = (\mu_2 - \mu_1)^{\top} \Sigma^{-1} (\mu_2 - \mu_1)$

# New result: Exact closed-form for Chernoff between univariate Gaussian distributions

- Optimality condition amounts to solve a quadratic equation for  $\alpha$

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



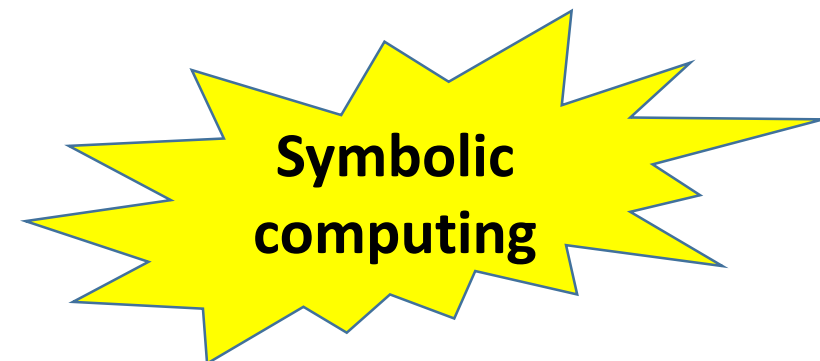
$$\text{OC}_{\text{Gaussian}} : \left( \frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2} \right) m_\alpha - \left( \frac{1}{2\sigma_2^2} - \frac{1}{2\sigma_1^2} \right) v_\alpha = \frac{1}{2} \log \frac{\sigma_2^2}{\sigma_1^2} + \frac{\mu_2^2}{2\sigma_2^2} - \frac{\mu_1^2}{2\sigma_1^2}.$$

- Use symbolic computing to get (very long) closed-form formula



# Maxima

A Computer Algebra System



```
(%i13) 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)$
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]);
```

$$0.5 \log(2 v_2) + \left( \frac{1}{2 v_2} - \frac{1}{2 v_1} \right) \left( \frac{(\alpha \mu_1 v_2 + (1 - \alpha) \mu_2 v_1)^2}{(\alpha v_2 + (1 - \alpha) v_1)^2} + \frac{v_1 v_2}{\alpha v_2 + (1 - \alpha) v_1} \right) + \left( \frac{\mu_1}{v_1} - \frac{\mu_2}{v_2} \right) \frac{(\alpha \mu_1 v_2 + (1 - \alpha) \mu_2 v_1)}{\alpha v_2 + (1 - \alpha) v_1} + \frac{\mu_2^2}{2 v_2} - 0.5 \log(2 v_1) - \frac{\mu_1^2}{2 v_1}$$

rat: replaced -0.5 by -1/2 = -0.5

rat: replaced 0.5 by 1/2 = 0.5

$$\begin{aligned} & (\text{sqrt}((4 \mu_2^2 - 8 \mu_1 \mu_2 + 4 \mu_1^2) v_1 v_2^2 + (-4 \mu_2^2 + 8 \mu_1 \mu_2 - 4 \mu_1^2) v_1^2 v_2) \log(2 v_2) + v_2^4 - 4 v_1 v_2^3 + ((-4 \mu_2^2 + 8 \mu_1 \mu_2 - 4 \mu_1^2) v_1 \log(2 v_1) + 6 v_1^2) v_2^2 + \\ & ((4 \mu_2^2 - 8 \mu_1 \mu_2 + 4 \mu_1^2) v_1^2 \log(2 v_1) - 4 v_1^3 + (4 \mu_2^4 - 16 \mu_1 \mu_2^3 + 24 \mu_1^2 \mu_2^2 - 16 \mu_1^3 \mu_2 + 4 \mu_1^4) v_1) v_2 + v_1^4) + (2 v_1^2 - 2 v_1 v_2) \log(2 v_2) + v_2^2 + \\ & (2 v_1 \log(2 v_1) - 2 v_1) v_2 - 2 v_1^2 \log(2 v_1) + v_1^2 + (-2 \mu_2^2 + 4 \mu_1 \mu_2 - 2 \mu_1^2) v_1) / \\ & ((2 v_2^2 - 4 v_1 v_2 + 2 v_1^2) \log(2 v_2) - 2 \log(2 v_1) v_2^2 + (4 v_1 \log(2 v_1) + 2 \mu_2^2 - 4 \mu_1 \mu_2 + 2 \mu_1^2) v_2 - 2 v_1^2 \log(2 v_1) + (-2 \mu_2^2 + 4 \mu_1 \mu_2 - 2 \mu_1^2) v_1) \end{aligned}$$

# General multivariate Gaussian case

**input** : Two normal densities  $p_{\mu_1, \Sigma_1}$  and  $p_{\mu_2, \Sigma_2}$ , and a numerical precision threshold  $\epsilon > 0$

$\alpha_m = 0;$

$\alpha_M = 1;$

**while**  $|\alpha_M - \alpha_m| > \epsilon$  **do**

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

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

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

    //

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

$\alpha_m = \alpha;$

        //

**end**

**else**

$\alpha_M = \alpha;$

**end**

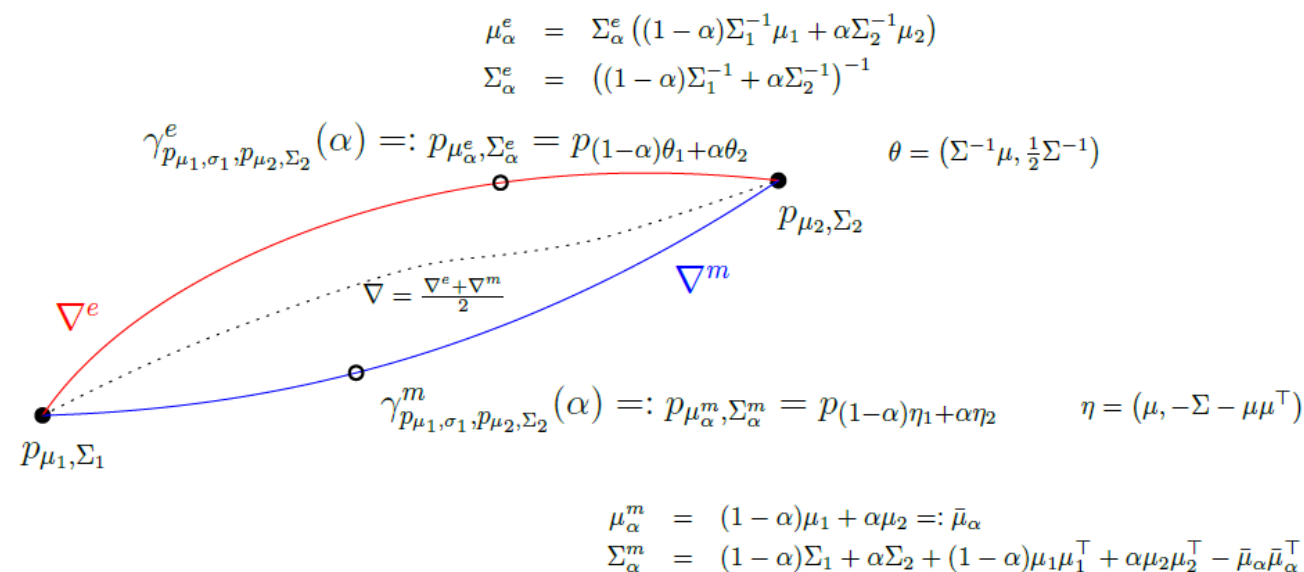
**end**

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

Kullback-Leibler divergence (= rev. Bregman div):

$$\frac{1}{2} \left( \text{tr}(\Sigma_2^{-1}\Sigma_1) - \log \frac{\det(\Sigma_2)}{\det(\Sigma_1)} - d + (\mu_2 - \mu_1)^\top \Sigma_2^{-1}(\mu_2 - \mu_1) \right)$$

**Efficient  
algorithm**



# New results: centered multivariate Gaussian case

- **Optimality condition** for centered multivariate Gaussian distributions:

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

- **Special case** of centered scaled covariance matrices  $\Sigma_1 = \Sigma$  and  $\Sigma_2 = s\Sigma$  in closed-form ( $s > 0$ ):

**Proposition** *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:*

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

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



# Robustness of Chernoff information (informal viewpoint)

- CI often used in **information fusion** community instead of a priori  $D_{B,\alpha}[p : q]$

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

For Chernoff information we have  $F'(\alpha^*\theta_1 + (1 - \alpha^*)\theta_2) = 0$

Wlog, assume  $\theta_2 - \theta_1 = 1$ ,

$$|J'_{F,\alpha}(\theta_1 : \theta_2) - J'_{F,\alpha^*}(\theta_1 : \theta_2)| = \underbrace{|F'(\alpha\theta_1 + (1 - \alpha)\theta_2)|}_{> 0} = F'(\alpha^*\theta_1 + (1 - \alpha^*)\theta_2).$$

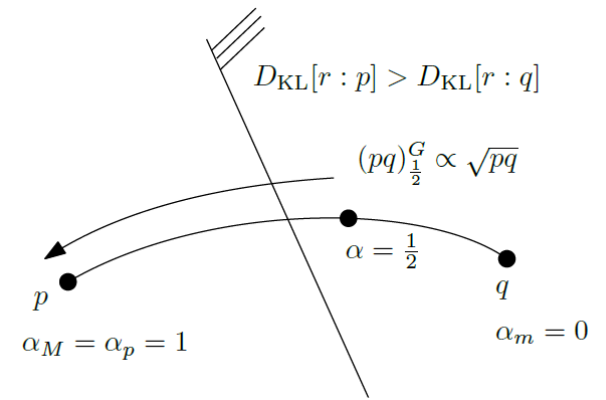
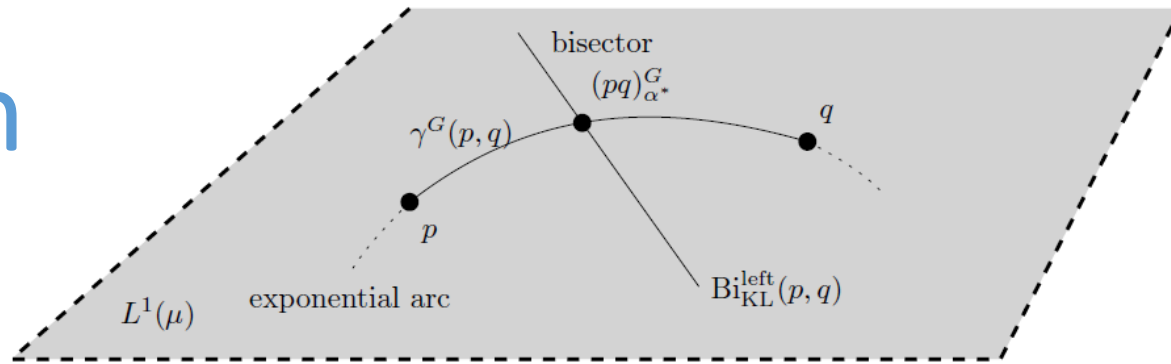


Chernoff information **more stable and robust**  
than any other Bhattacharyya distances  $D_{B,\alpha}[p : q]$

Julier, Simon J. "An empirical study into the use of Chernoff information for robust, distributed fusion of Gaussian mixture models."

*9th International Conference on Information Fusion. IEEE, 2006.*

# Conclusion



- We revisited **Chernoff information** of two probability distributions under the umbrella of special exponential families
- The geometric mixture is a 1D **log-ratio exponential family**:
  - Chernoff exponent is unique (from the convexity of log-normalizer) **NEW**
  - Geometrically, Chernoff point = intersection of a Voronoi bisector with a geodesic
  - Approximate the Chernoff information by **bisection search on exponential arc**
  - Express the **optimality condition** of Chernoff error exponent in various ways
- Consider Chernoff information between Gaussian densities:
  - NEW** • exact closed-form for **univariate Gaussians** (solve quadratic eq. using symbolic computing)
  - NEW** • exact closed-form for **centered scaled covariance matrices**
  - Practical bisection search on the exponential arc with numerical experiments

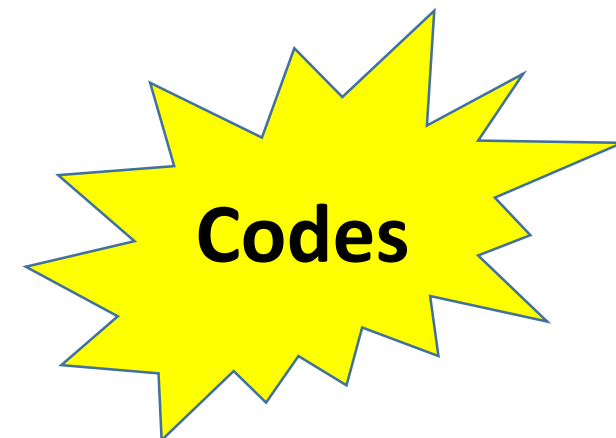
# Thank you!

<https://franknielsen.github.io/ChernoffInformation/index.html>

Optimality conditions (OCs)  
for Chernoff exponent

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

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



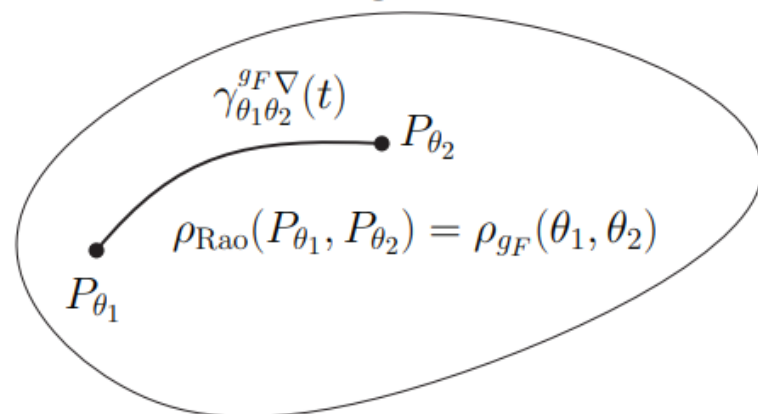
Entropy 2022  
[2207.03745]

## Some references

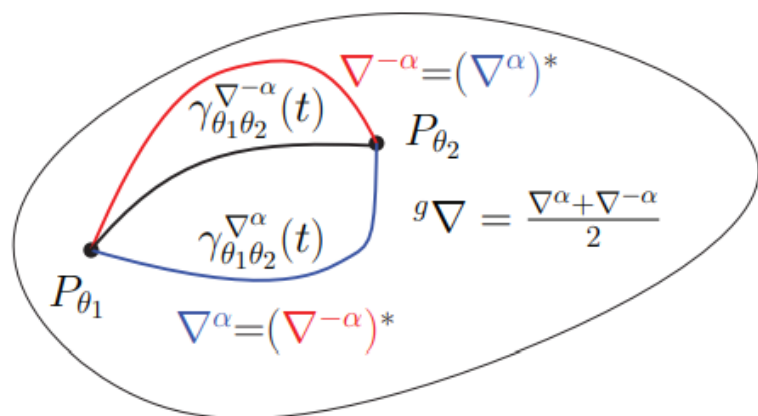
- Chernoff, Herman. "*A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations.*" *The Annals of Mathematical Statistics* (1952): 493-507.
- Nielsen, Frank. "*Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means.*" *Pattern Recognition Letters* 42 (2014): 25-34.
- Nielsen, Frank. "*An information-geometric characterization of Chernoff information.*" *IEEE Signal Processing Letters* 20.3 (2013): 269-272.
- Deformed log-ratio exponential families (deformed LREFs):  
Masrani, Vaden, et al. "*q-Paths: Generalizing the geometric annealing path using power means.*" *Uncertainty in Artificial Intelligence*. PMLR, 2021.

# The Many Faces of Information Geometry

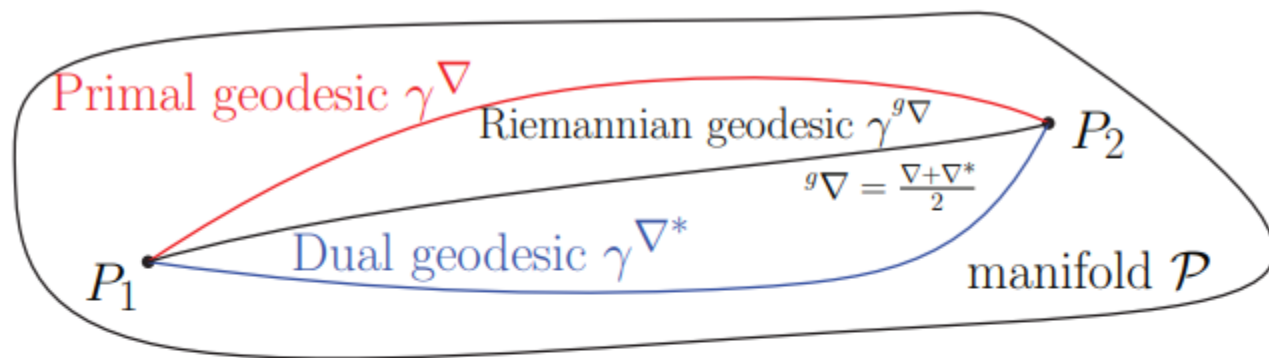
**Fisher-Rao geometry**  
→ **Fisher-Rao geodesic distance**



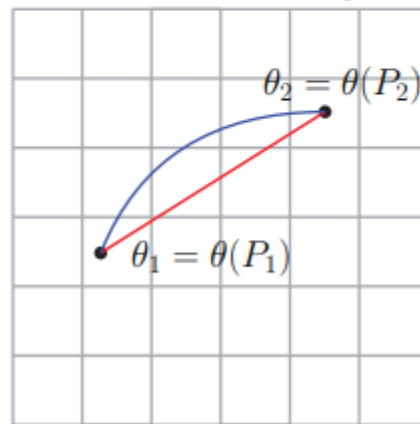
**versus**



**Dual  $\alpha$ -geometry**  
→ **No default divergence**

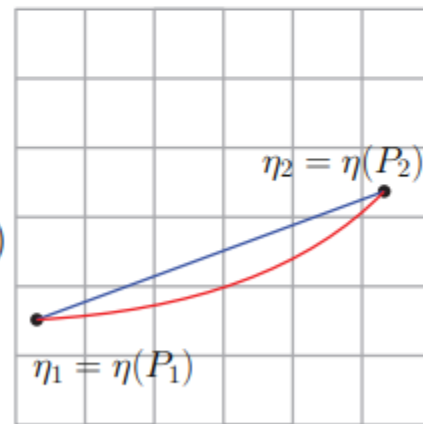


$\nabla$ -affine coordinate system  $\theta$



Potential function  $F(\theta)$

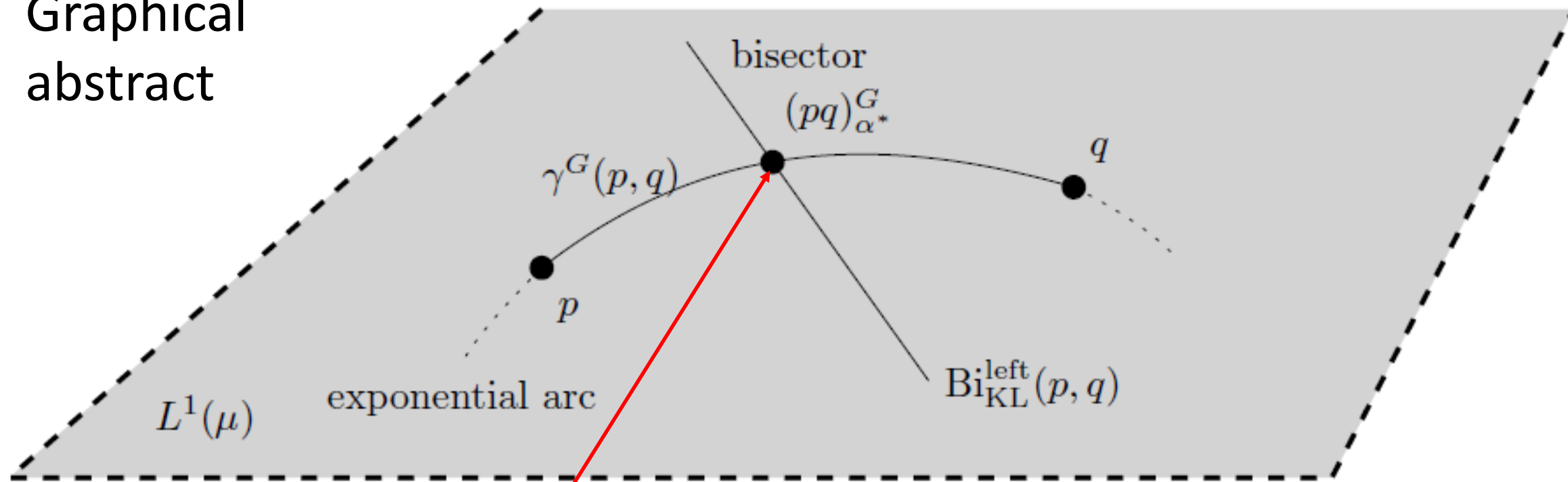
$\nabla^*$ -affine coordinate system  $\eta$



Dual potential function  $F^*(\eta)$

Legendre-Fenchel transform

Graphical  
abstract



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

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

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

Entropy 2022  
[2207.03745]