Revisiting Chernoff information with Likelihood Ratio Exponential Families

Frank Nielsen

Sony Computer Science Laboratories Inc.

Talk 24th November 2022 at Inria

Entropy 2022 [2207.03745]

Chernoff information: Definition & Background
A symmetric statistical divergence

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

Definition:

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

Herman Chernoff
(1923-)

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

(via Hölder inequality)

- **skewed Bhattacharyya coefficient ρ_α** (similarity coefficient)

- Synonyms: Chernoff divergence, Chernoff information number, Chernoff index...
- Found later many applications in information fusion, radar target detection, generative adversarial networks (GANs), 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)

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

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

Bhattacharyya distance when α=1/2

$$D_{B,\alpha}[p:q] = -\log \int p^\alpha q^{1-\alpha} \mathrm{d}\mu = D_{B,1-\alpha}[q:p]$$

$$\Downarrow \quad \alpha=1/2$$

$$D_B[p:q] = -\log \int \sqrt{pq}\,\mathrm{d}\mu = D_{B,\frac{1}{2}}[p:q]$$

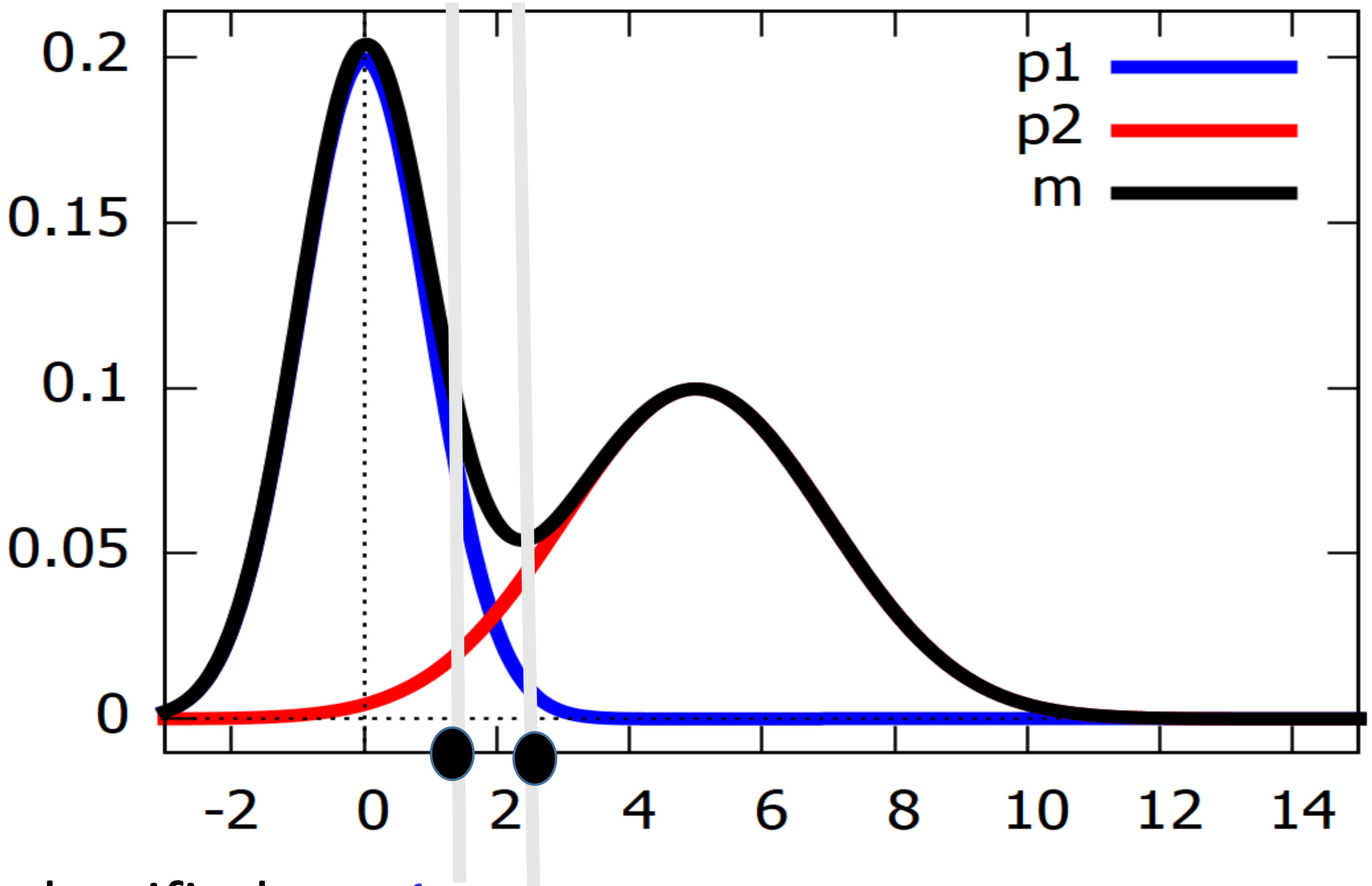
- **Bhattacharyya distance** does not satisfy the triangle inequality: not a metric
- Chernoff information tunes/learns the skewed Bhattacharyya distance
- Information = variational divergence (computed from an optimization procedure)
- Limit scaled skewed Bhattacharyya distance = Kullback-Leibler divergence

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

when α → 1, get KLD

Bhattacharyya, Anil. "On a measure of divergence between two statistical populations defined by their probability distributions." *Bull. Calcutta Math. Soc.* 35 (1943): 99-109.

Rationale for CI: Statistical hypothesis testing



x classified as p1

x classified as p2

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:

$$\boxed{P_e = \int \min(p_1(x), p_2(x))\mathrm{d}x}$$

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**:

$$P_e = \int \min(p_1(x), p_2(x))\mathrm{d}x \Rightarrow \boxed{P_e = \frac{1}{2}(1 - D_{\mathrm{TV}}[p_1, p_2])}$$

$$D_{\mathrm{TV}}[p_1, p_2] = \frac{1}{2}\int (p_1(x) - p_2(x))\mathrm{d}x$$

- Use a **generic (weighted) mean** which necessarily falls inbetween its extrema (e.g., **geometric mean**):

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

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

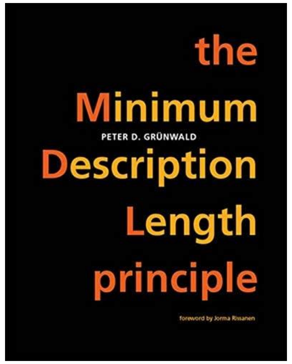
"Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means." ***Pattern Recognition Letters*** **42 (2014): 25-34.**

Outline of the contributions of this talk

- (Background: done!)
- **Interplay of non-parametric with parametric study** of Chernoff information via the concept of **lilelihood-ratio exponential families** (LREFS)

- Derive various **optimality conditions** for the Chernoff exponent α^*
- Give some **geometric interpretations** on Bregman manifolds which yield *fast approximation 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)

  between two densities p, q of Lebesgue Banach space $L_1(\mu)$

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

- Set of **geometric mixtures**:

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

  with **normalization factor**:

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

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

Sufficient statistics: log likelihood ratio

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

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

k(x)=log q(x)

$- D_{B,\alpha}[p:q]$

Natural parameter space:

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

LREFs: EF cumulant function is always analytic C^ω

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

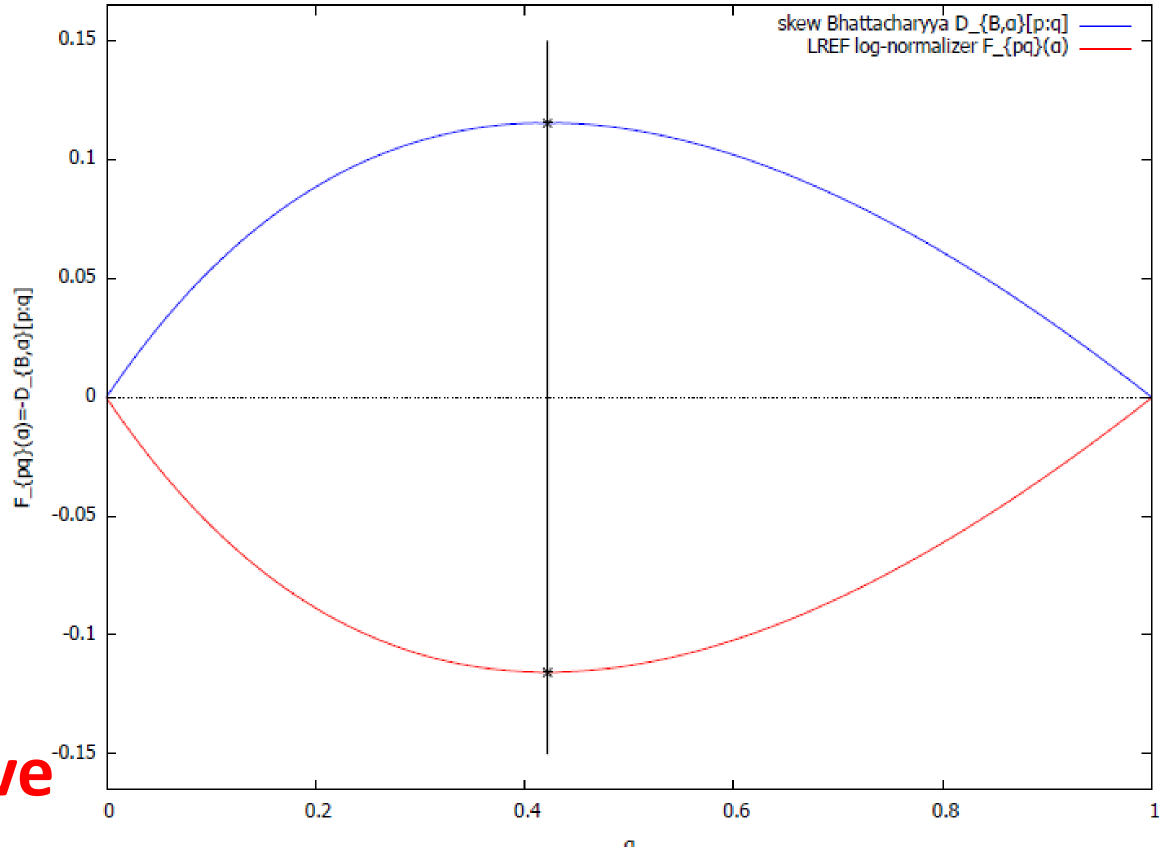
- Cumulant function is neg-Bhattacharyya distance:

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

$\Rightarrow$ Bhattacharyya. distance is **strictly concave**

- <u>Theorem</u>:

Chernoff exponent exists and is <u>unique</u>



p=N(0,1) q=N(1,2)

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

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

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

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 (depends on support):

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

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

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

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

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

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

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

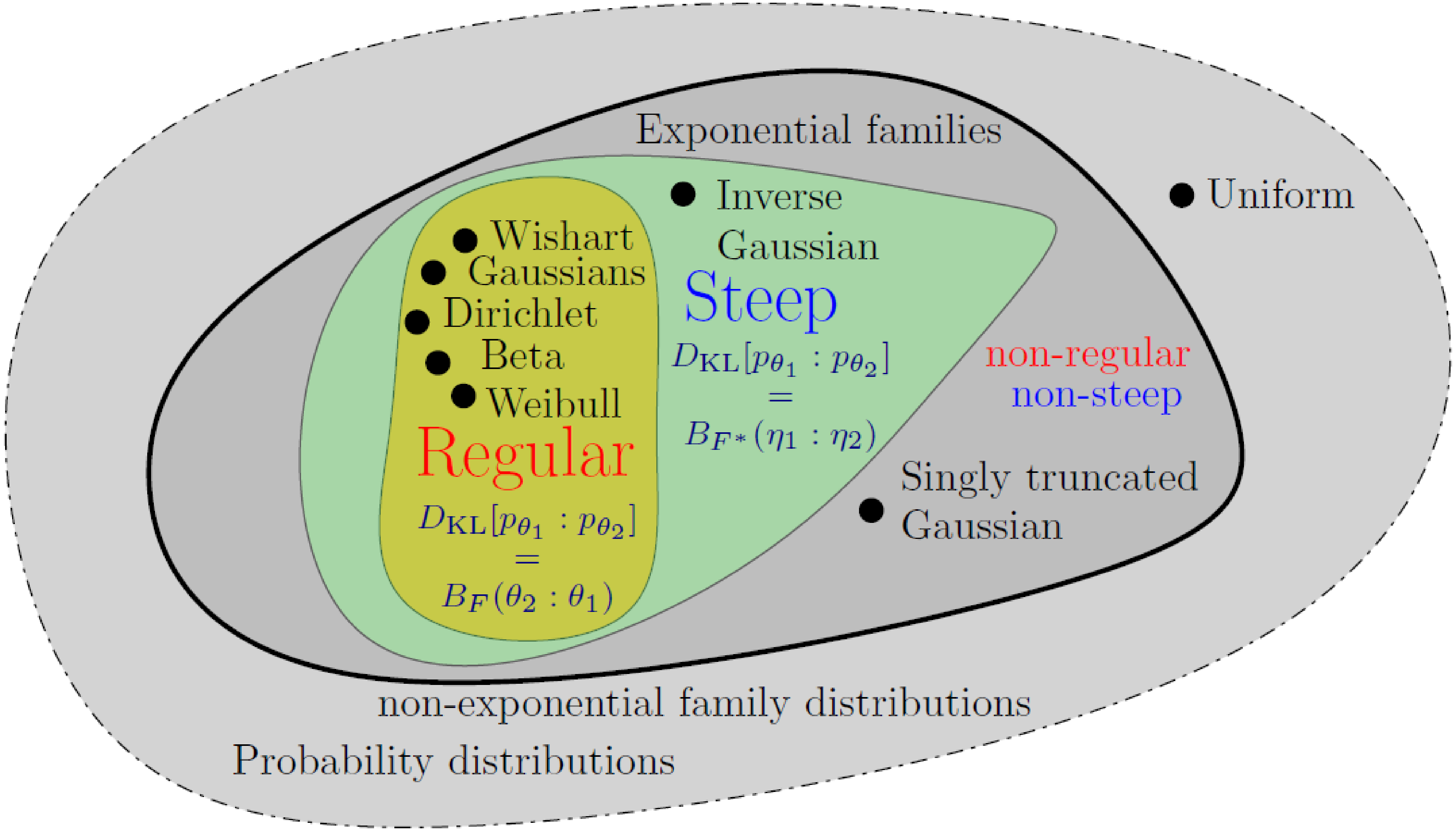
steep $\Rightarrow$ $E_{p_{\theta_1}}[t(x)] = \nabla F(\theta_1)$

regular EF $\Rightarrow$ steep EF

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

Venn diagram: Regular & steepness of (LR)EFs

- Steepness implies **duality between natural** θ **and moment** η **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 $D_{\mathrm{KL}}[p:q]$ and the reverse Kullback-Leibler divergence $D_{\mathrm{KL}}[q:p]$ are finite.*

PROOF

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

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

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

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

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

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

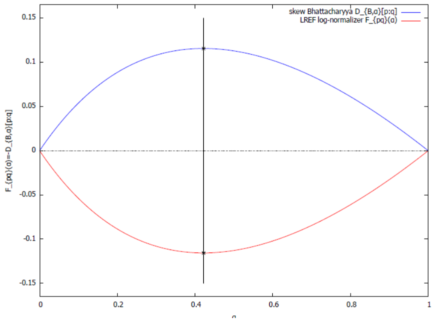
idem for $\boxed{D_{\mathrm{KL}}[q:p] = B_{F_{pq}}(\alpha_p : \alpha_q)} = B_{F_{pq}}(1:0) = -F'_{pq}(0) < \infty$

Chernoff information (for densities of a LREF)

- Proposition: $\boxed{D_C[p:q] = D_{\mathrm{KL}}[(pq)^G_{\alpha^*} : p] = D_{\mathrm{KL}}[(pq)^G_{\alpha^*} : q] = D_{B,\alpha^*}[p:q]}$

PROOF

First, **skew Bhattacharyya distance = skew Jensen divergence**



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

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

Thus we have:
$$\begin{aligned} D_{B,\alpha}((pq)^G_{\alpha_1} : (pq)^G_{\alpha_2}) &= J_{F_{pq},\alpha}(\alpha_1 : \alpha_2), \\ &= \alpha F_{pq}(\alpha_1) + (1-\alpha)F_{pq}(\alpha_2) - F_{pq}(\alpha\alpha_1 + (1-\alpha)\alpha_2) \end{aligned}$$

At the optimal value α*, we have $F'_{pq}(\alpha^*) = 0$

① $D_{\mathrm{KL}}[(pq)^G_{\alpha^*} : p] = B_{F_{pq}}(1 : \alpha^*) = -F(\alpha^*)$

② $D_{\mathrm{KL}}[(pq)^G_{\alpha^*} : q] = B_{F_{pq}}(0 : \alpha^*) = -F(\alpha^*)$

③ $D_C[p:q] = -\log \rho_{\alpha^*}(p:q) = J_{F_{pq},\alpha^*}(1:0) = -F_{pq}(\alpha^*)$

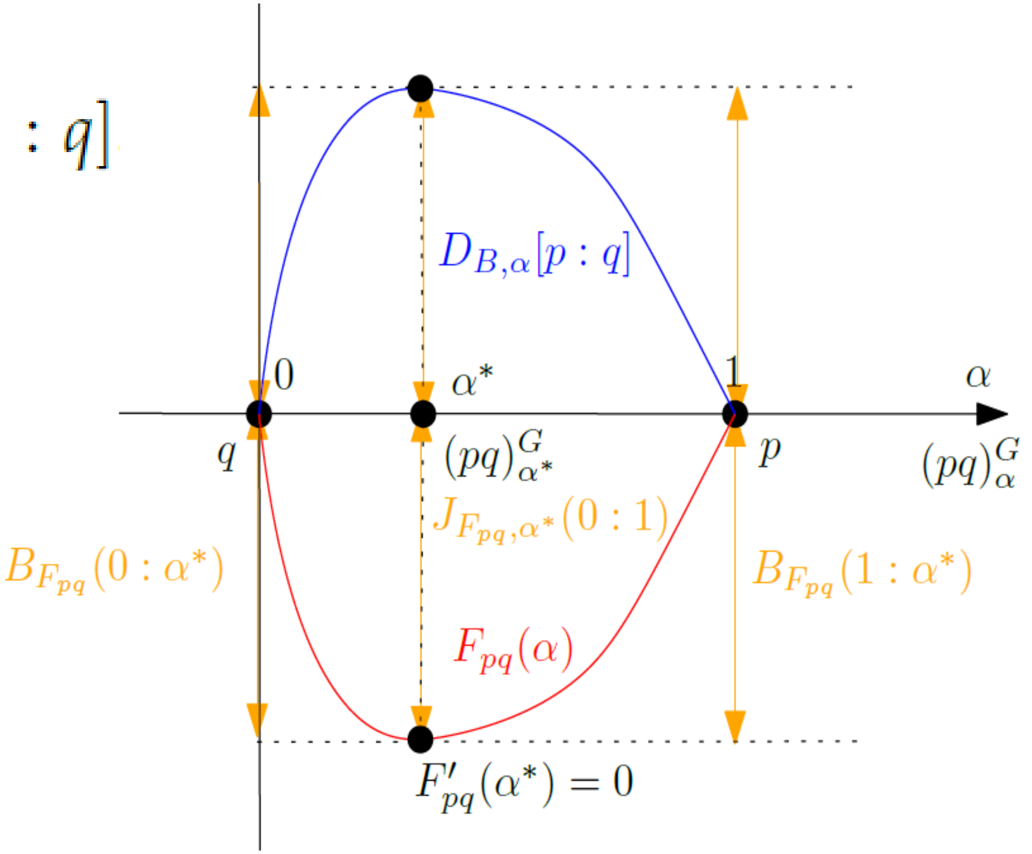
Jensen-Chernoff divergence

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

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**

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

Geometric interpretation for densities p, q on $L_1(\mu)$

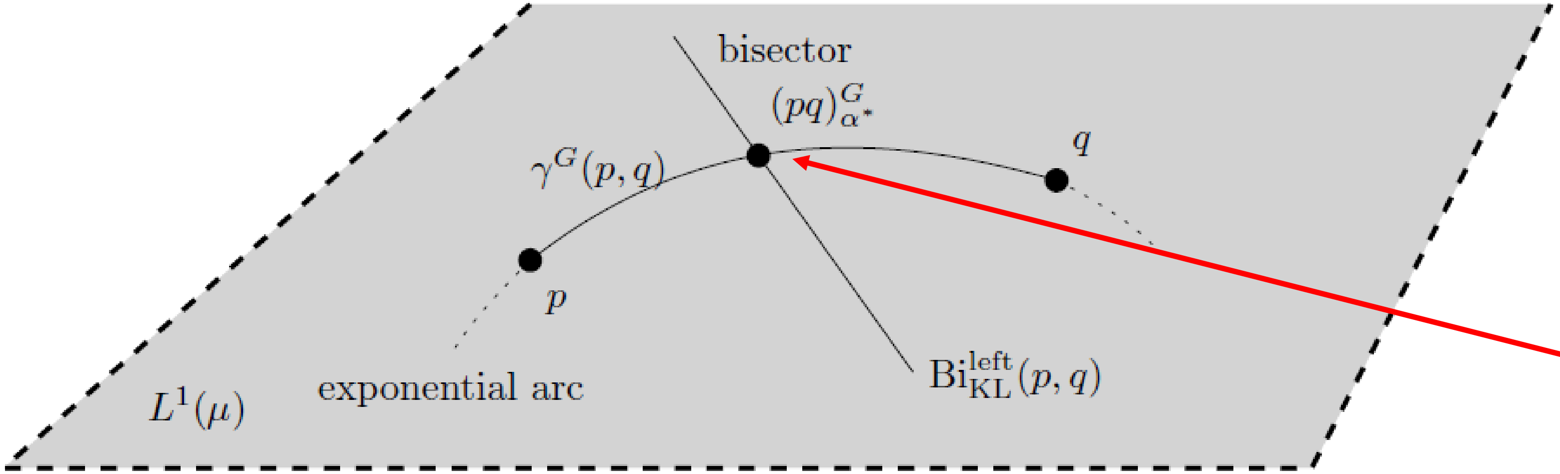
Proposition (Geometric characterization of the Chernoff information). *On the vector space $L^1(\mu)$, the Chernoff information distribution is the unique distribution*

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

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

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

2209.07481

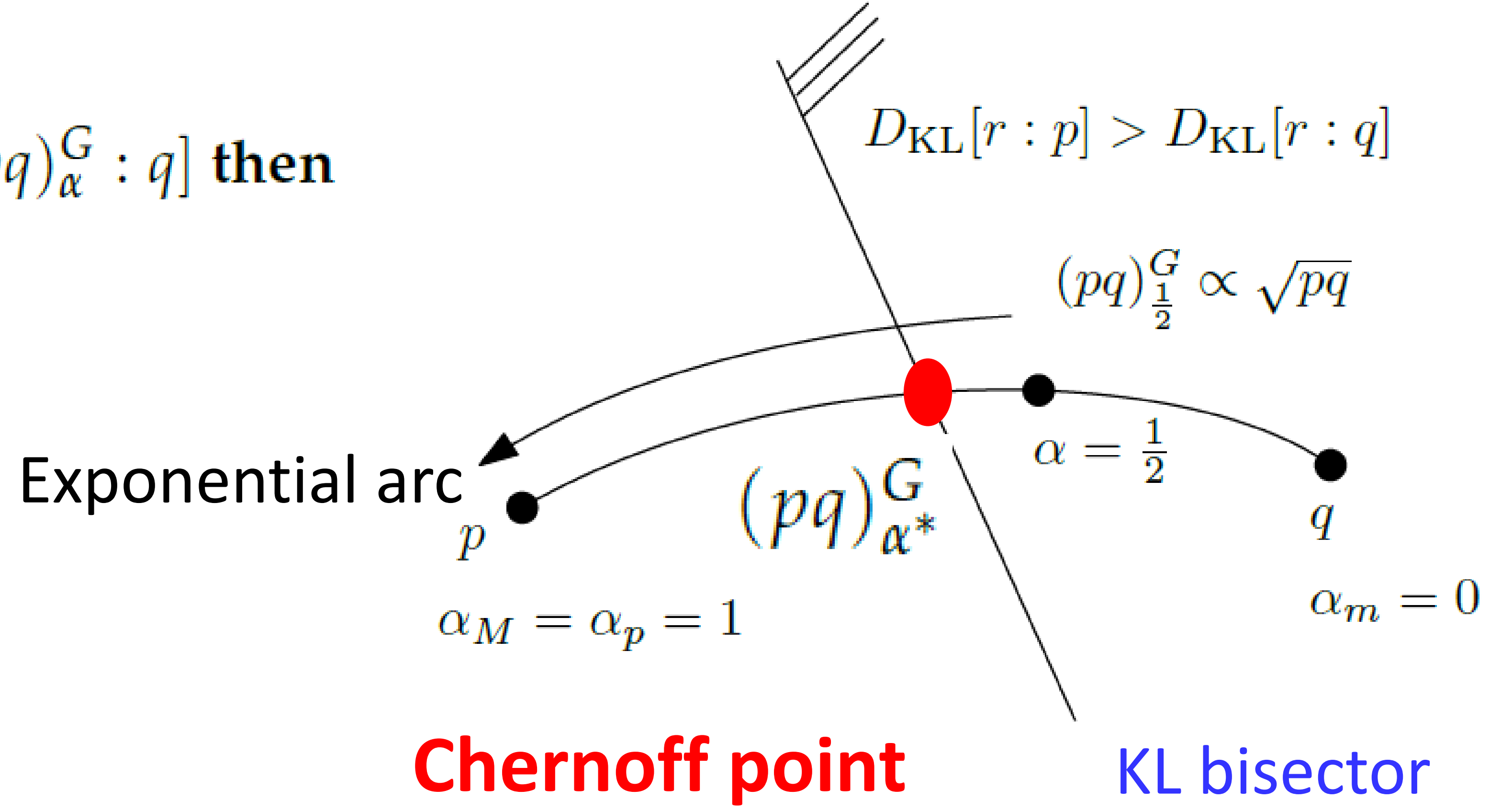


Chernoff point: $(pq)^G_{\alpha^*}$

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 $D_{\mathrm{KL}}[(pq)^G_\alpha : p] > D_{\mathrm{KL}}[(pq)^G_\alpha : q]$ **then**
    $\alpha_m = \alpha$;
    `// See Figure`
  end
  else
    $\alpha_M = \alpha$;
  end
end
return $D_{\mathrm{KL}}[(pq)^G_\alpha : p]$;

Bisection algorithm:



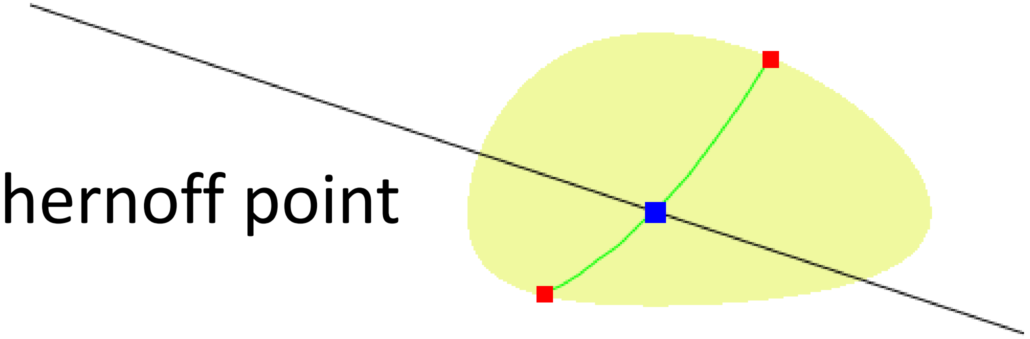
Chernoff information viewed as a symmetrization of KLD

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

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

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

circumcenter = Chernoff point



Chernoff point r*
(eKLD)

Dual moment parameterizations β of LREFs

Dual moment parameter:

α = natural primal parameter

$$\beta = \beta(\alpha) := E_{(pq)^G_\alpha}[t(x)] = E_{(pq)^G_\alpha}\left[\log \frac{p(x)}{q(x)}\right]$$

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

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

Moment parameter

Natural parameter

$$\beta(\alpha) = F'_{pq}(\alpha) \quad \xleftrightarrow{\text{Legendre transform}} \quad \alpha = {F^*_{pq}}'(\beta).$$

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

Dual parameterizations of LREFs and optimality condition for finding Chernoff exponent

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

$$\Downarrow$$

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

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

Primal optimality condition $\Leftrightarrow$ **Dual optimality condition**

Special case of LREF: p,q are densities of a same EF!

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

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

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

$$\Rightarrow \quad (p_{\theta_1} p_{\theta_2})_\alpha^G = p_{\alpha\theta_1+(1-\alpha)\theta_2} \qquad D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}] = B_F(\theta_2 : \theta_1)$$

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

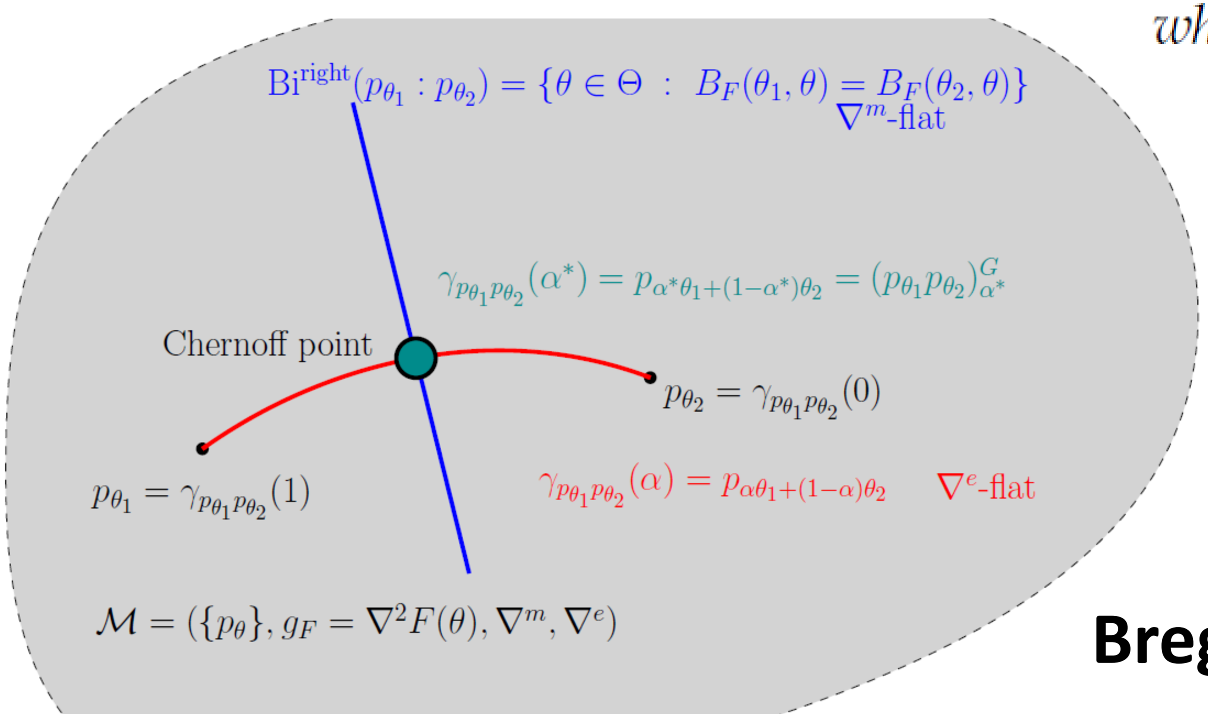
Proposition *Let p_{λ_1} and p_{λ_2} be two densities of a regular exponential family $\mathcal{E}$ with natural parameter $\theta(\lambda)$ and log-normalizer $F(\theta)$. Then the Chernoff information is*

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

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:

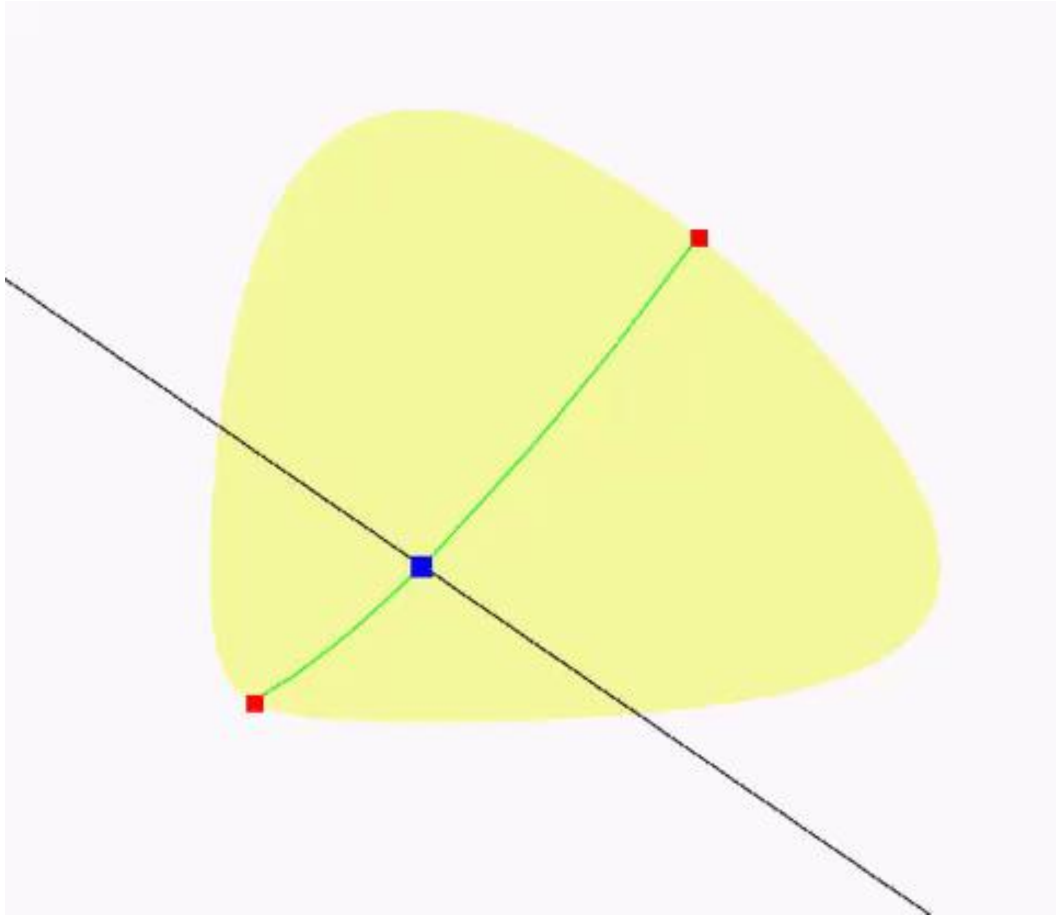
$$\boxed{\mathrm{OC}_{\mathrm{EF}} : \quad (\theta_2 - \theta_1)^\top \eta_{\alpha^*} = F(\theta_2) - F(\theta_1),}$$

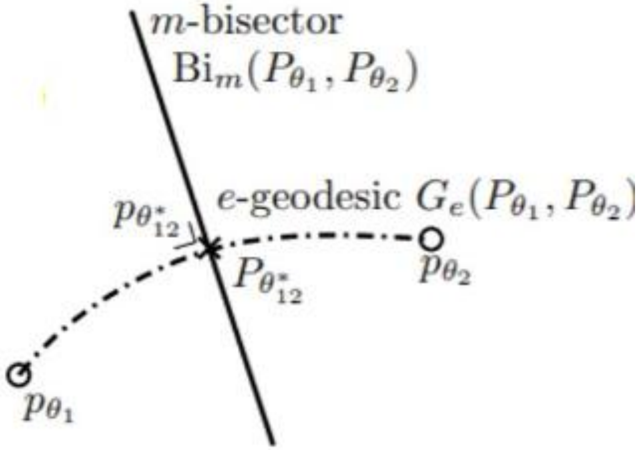
where $\eta_{\alpha^*} = \nabla F(\alpha^* \theta_1 + (1 - \alpha^*)\theta_2) = E_{p_{\alpha^*\theta_1 + (1-\alpha^*)\theta_2}}[t(x)]$.



Bregman manifold (= global Hessian manifold)

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 pure 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 $\gamma^e_{pq}(\alpha)$ and $\gamma^m_{pq}(\alpha)$ be a e-geodesic and m-geodesic passing through the points p and q of $\mathcal{M}$, respectively. Let $\mathrm{Bi}^m(p,q)$ and $\mathrm{Bi}^e(p,q)$ be the right-sided ∇^m-flat and left-sided ∇^e-flat Bregman bisectors, respectively. Then the intersection of $\gamma^e_{pq}(\alpha)$ with $\mathrm{Bi}^m(p,q)$ and the intersection of $\gamma^m_{pq}(\alpha)$ with $\mathrm{Bi}^e(p,q)$ are unique. The point $\gamma^e_{pq}(\alpha) \cap \mathrm{Bi}^m(p,q)$ is called the Chernoff point and the point $\gamma^m_{pq}(\alpha) \cap \mathrm{Bi}^e(p,q)$ 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_{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:

Unidimensional case

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

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

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

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

Closed
-form

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

1D LREF p,q $\mathcal{E}_{pq} := \left\{ (pq)^G_\alpha(x) := \frac{p(x)^\alpha q(x)^{1-\alpha}}{Z_{pq}(\alpha)} \ : \ \alpha \in \Theta \right\}$

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

p,q in (multivariate) EF

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

closed form for unidim. EF $\alpha^* = \dfrac{F'^{-1}\left(\frac{F(\alpha_2) - F(\alpha_1)}{\alpha_2 - \alpha_1}\right) - \alpha_2}{\alpha_1 - \alpha_2}$

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:

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

where

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

Harmonic weighted SPD matrix mean

Invariance under the action of the affine group

Affine group action:

$$(l_1, A_1).(l_2, A_2) = (l_1 + A_1 l_2, A_1 A_2)$$

Matrix group element representation:

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

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

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

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

where

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

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

Wlog., we can assume this is the **canonical case**

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

- Trivial case, α*=1/2

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

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

squared Mahalanobis distance:
(= Malahanobis divergence)

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

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

- Optimality condition amounts to solve a quadratic equation for α

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

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

- Use symbolic computing to get very long closed-form formula by solving a quadratic equation

Maxima

A Computer Algebra System

Symbolic computing

```
(%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]);
```

(%o11) $$0.5\log(2\,v2)+\left(\frac{1}{2\,v2}-\frac{1}{2\,v1}\right)\left(\frac{\left(alpha\,mu1\,v2+(1-alpha)\,mu2\,v1\right)^2}{\left(alpha\,v2+(1-alpha)\,v1\right)^2}+\frac{v1\,v2}{alpha\,v2+(1-alpha)\,v1}\right)+\frac{\left(\frac{mu1}{v1}-\frac{mu2}{v2}\right)\left(alpha\,mu1\,v2+(1-alpha)\,mu2\,v1\right)}{alpha\,v2+(1-alpha)\,v1}+\frac{mu2^2}{2\,v2}-0.5\log(2\,v1)-\frac{mu1^2}{2\,v1}$$

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

(%o13) $$(\mathrm{sqrt}(\,((4\,mu2^2-8\,mu1\,mu2+4\,mu1^2)\,v1\,v2^2+(-4\,mu2^2+8\,mu1\,mu2-4\,mu1^2)\,v1^2\,v2)\log(2\,v2)+v2^4-4\,v1\,v2^3+((-4\,mu2^2+8\,mu1\,mu2-4\,mu1^2)\,v1\log(2\,v1)+6\,v1^2)\,v2^2+ ((4\,mu2^2-8\,mu1\,mu2+4\,mu1^2)\,v1^2\log(2\,v1)-4\,v1^3+(4\,mu2^4-16\,mu1\,mu2^3+24\,mu1^2\,mu2^2-16\,mu1^3\,mu2+4\,mu1^4)\,v1)\,v2+v1^4)+(2\,v1^2-2\,v1\,v2)\log(2\,v2)+v2^2+ (2\,v1\log(2\,v1)-2\,v1)\,v2-2\,v1^2\log(2\,v1)+v1^2+(-2\,mu2^2+4\,mu1\,mu2-2\,mu1^2)\,v1)/ ((2\,v2^2-4\,v1\,v2+2\,v1^2)\log(2\,v2)-2\log(2\,v1)\,v2^2+(4\,v1\log(2\,v1)+2\,mu2^2-4\,mu1\,mu2+2\,mu1^2)\,v2-2\,v1^2\log(2\,v1)+(-2\,mu2^2+4\,mu1\,mu2-2\,mu1^2)\,v1)$$

General multivariate Gaussian case: Approximation

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

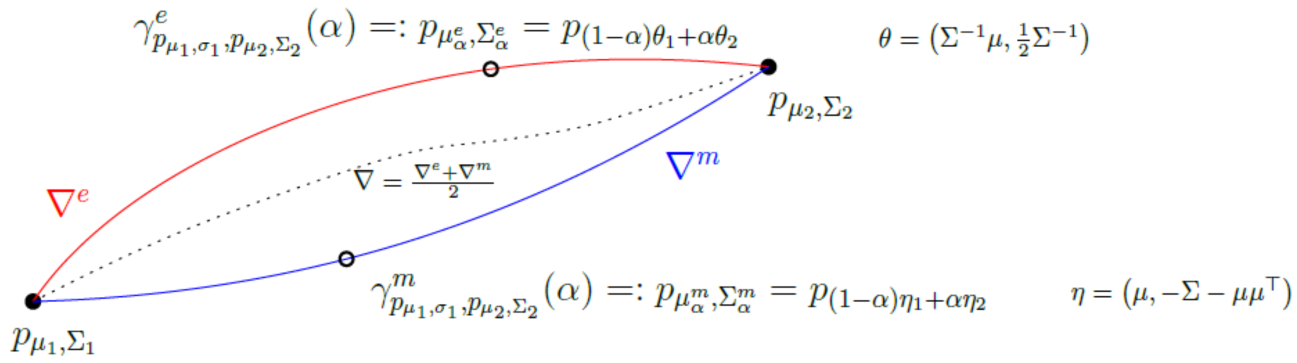
$\alpha_m = 0$;
$\alpha_M = 1$;
while $|\alpha_M - \alpha_m| > \epsilon$ **do**
  $\alpha = \frac{\alpha_m + \alpha_M}{2}$;
  $\Sigma_\alpha^e = \left((1-\alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1}\right)^{-1}$;
  $\mu_\alpha^e = \Sigma_\alpha^e\left((1-\alpha)\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2\right)$;
  //
  if $D_{\mathrm{KL}}[p_{\mu_\alpha^e,\Sigma_\alpha^e} : p_{\mu_1,\Sigma_1}] > D_{\mathrm{KL}}[p_{\mu_\alpha^e,\Sigma_\alpha^e} : p_{\mu_2,\Sigma_2}]$ **then**
    $\alpha_m = \alpha$;
    //
  end
  else
    $\alpha_M = \alpha$;
  end
end
return $D_{\mathrm{KL}}[p_{\mu_\alpha^e,\Sigma_\alpha^e} : p_{\mu_1,\Sigma_1}]$;

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

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

Efficient algorithm

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



$$\begin{aligned}\mu_\alpha^m &= (1-\alpha)\mu_1 + \alpha\mu_2 =: \bar\mu_\alpha\\ \Sigma_\alpha^m &= (1-\alpha)\Sigma_1 + \alpha\Sigma_2 + (1-\alpha)\mu_1\mu_1^\top + \alpha\mu_2\mu_2^\top - \bar\mu_\alpha\bar\mu_\alpha^\top\end{aligned}$$

New result: centered scaled multivariate Gaussian case

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

$$\boxed{\operatorname{tr}\left((\Sigma_2^{-1}-\Sigma_1^{-1})\,(\Sigma_1^{-1}+\alpha^*(\Sigma_2^{-1}-\Sigma_1^{-1}))^{-1}\right) \;=\; \log\frac{\det(\Sigma_1)}{\det(\Sigma_2)} = \log\det\left(\Sigma_1\Sigma_2^{-1}\right)}$$

- **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:*

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

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 $D_{B,\alpha}[p:q]$

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

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

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

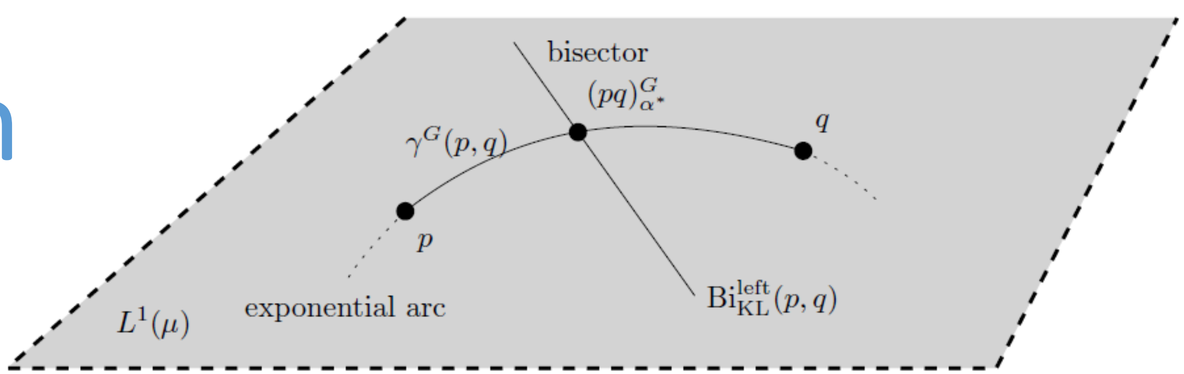
flatter minimum at Chernoff exponent

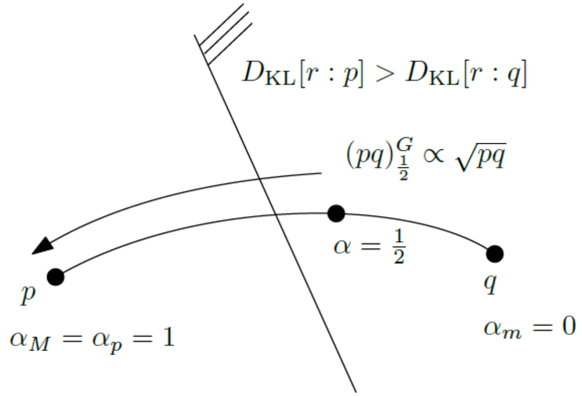
$$|J'_{F,\alpha}(\theta_1:\theta_2) - J'_{F,\alpha^*}(\theta_1:\theta_2)| = \underline{|F'(\alpha\theta_1+(1-\alpha)\theta_2)| > 0 = F'(\alpha^*\theta_1+(1-\alpha^*)\theta_2).}$$

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**:
  - NEW Chernoff exponent is unique (from the convexity of log-normalizer)
  - 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 for your attention!

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

Optimality conditions (OCs) for Chernoff exponenti

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

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

Codes

Entropy 2022 [2207.03745]

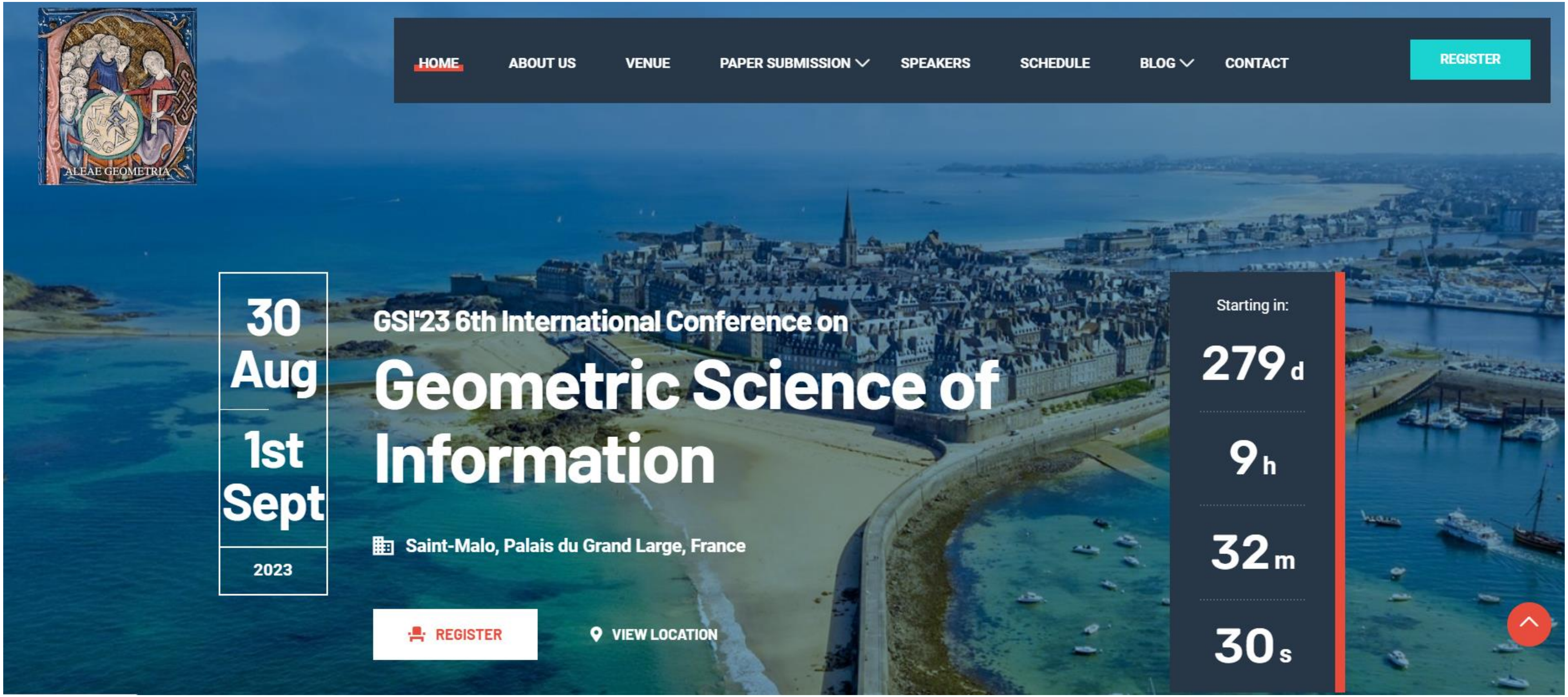
Some references

Entropy 2022 [2207.03745]

- 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.

gsi2023.org



ALEAE GEOMETRIA
HOME
ABOUT US
VENUE
PAPER SUBMISSION
SPEAKERS
SCHEDULE
BLOG
CONTACT
REGISTER
30 Aug
1st Sept
2023
GSI'23 6th International Conference on
Geometric Science of Information
Saint-Malo, Palais du Grand Large, France
REGISTER
VIEW LOCATION
Starting in:
279 d
9 h
32 m
30 s

https://franknielsen.github.io/IG/index.html

Information geometry and divergences

Foundations, Applications, and Software APIs

Historically, **Information Geometry** (IG, tutorials, textbooks and monographs) aimed at unravelling the geometric structures of families of probability distributions called the **statistical models**.
A statistical model can either be

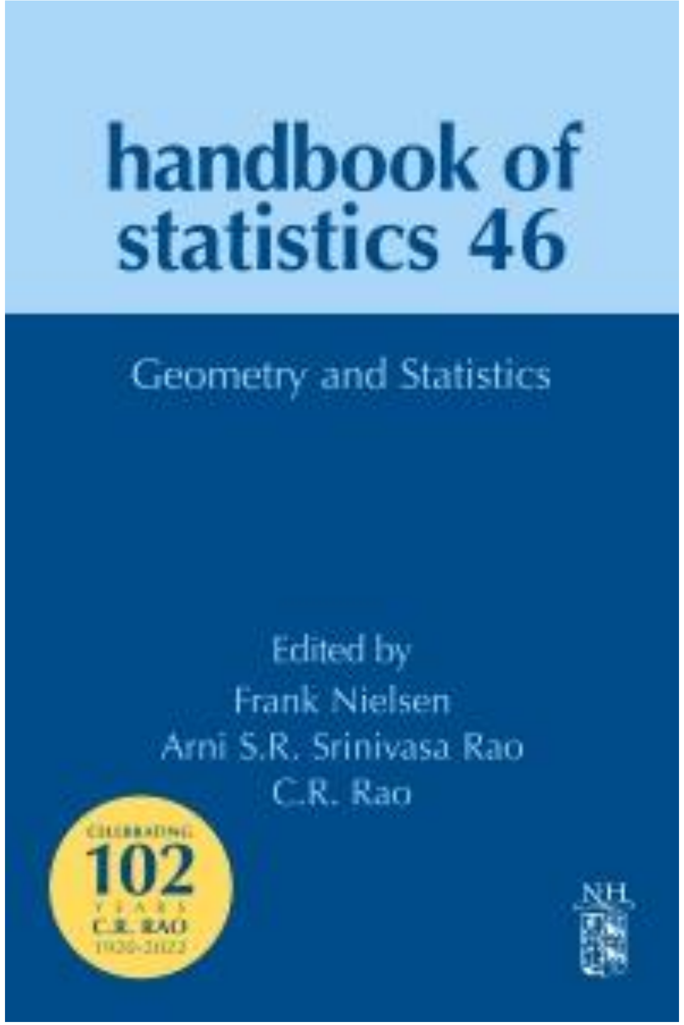
- parametric (eg., family of normal distributions),
- semi-parametric (eg., family of Gaussian mixture models) or
- non-parametric (family of mutually absolutely continuous smooth densities).

A parametric statistical model is said **regular** when the Fisher information matrix is positive-definite (and well-defined). Otherwise, the statistical model is irregular (eg., infinite Fisher information and semi-positive definite Fisher information when the model is not identifiable).

The **Fisher-Rao manifold** of a statistical parametric model is a Riemannian manifold equipped with the **Fisher information metric**. The geodesic length on a Fisher-Rao manifold is called **Rao's distance** [Hotelling 1930][Rao 1945]. More generally, Amari proposed the **dualistic structure** of IG which consists of a pair of torsion-free affine connections coupled to the Fisher metric [Amari 1980's]. Given a dualistic structure, we can build generically a one-parameter family of dualistic information-geometric structures, called the **α-geometry**. When both connections are flat, the information-geometric space is said **dually flat**: For example, the Amari's ±1-structures of **exponential families** and **mixture families** are famous examples of dually flat spaces in information geometry. In differential geometry, **geodesics** are defined as autoparallel curves with respect to a connection. When using the default Levi-Civita metric connection derived from the Fisher metric on Fisher-Rao manifolds, we get Rao's distance which are locally minimizing geodesics. Eguchi showed how to build from any smooth distortion (originally called a contrast function) a dualistic structure: The **information geometry of divergences** [Eguchi 1982]. The information geometry of **Bregman divergences** yields dually flat spaces: It is a special cases of **Hessian manifolds** which are differentiable manifolds equipped with a metric tensor being a Hessian metric and a flat connection [Shima 2007]. Since geometric structures scaffold spaces independently of any applications, these pure information-geometric Fisher-Rao structure and α-structures of statistical models can also be used in non-statistical contexts too: For example, for analyzing interior point methods with barrier functions in optimization, or for studying time-series models, etc.

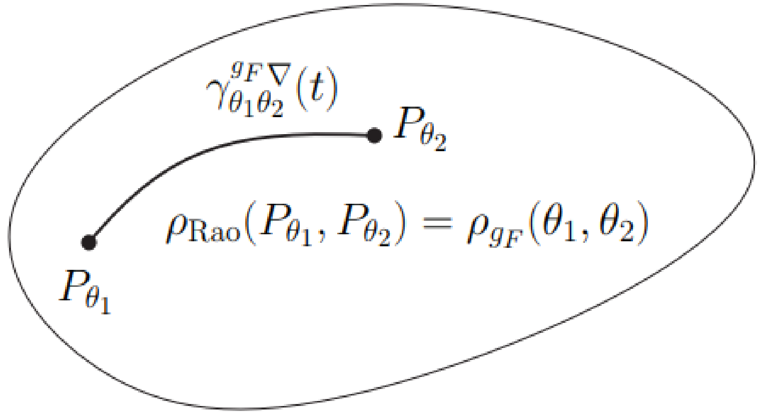
Statistical divergences between parametric statistical models amount to **parameter divergences** on which we can use the Eguchi's divergence information geometry to get a dualistic structure. A **projective divergence** is a divergence which is invariant by independent rescaling of its parameters. A statistical projective divergence is thus useful for estimating computationally intractable statistical models (eg., gamma divergences, Cauchy-Schwarz divergence and Hölder divergences, or singly-sided projective Hyvärinen divergence). A **conformal divergence** is a divergence scaled by a conformal factor which may depend on one or two of its arguments. The metric tensor obtained from Eguchi's information divergence of a conformal divergence is a **conformal metric** of the metric obtained from the divergence, hence its name. By analogy to total least squares vs least squares, a **total divergence** is a divergence which is invariant wrt. to rotations (eg., total Bregman divergences). An important property of divergences on the probability simplex is to be **monotone** by coarse-graining. That is, merging bins and considering reduced histograms should give a distance less or equal than the distance on the full resolution histograms. This **information monotonicity** property holds for f-divergences (called invariant divergences in information geometry), Hilbert log cross-ratio distance, or Aitchison distance for example. Some statistical divergences are upper **bounded** (eg., Jensen-Shannon divergence) while others are not (eg., Jeffreys' divergence). **Optimal transport** distances require a **ground base distance** on the sample space. A **diversity index** generalizes a **two-point distance** to a family of parameters/distributions. It usually measures the dispersion around a center point (eg., like variance measures the dispersion around the centroid).

https://www.elsevier.com/books/geometry-and-statistics/nielsen/978-0-323-91345-4

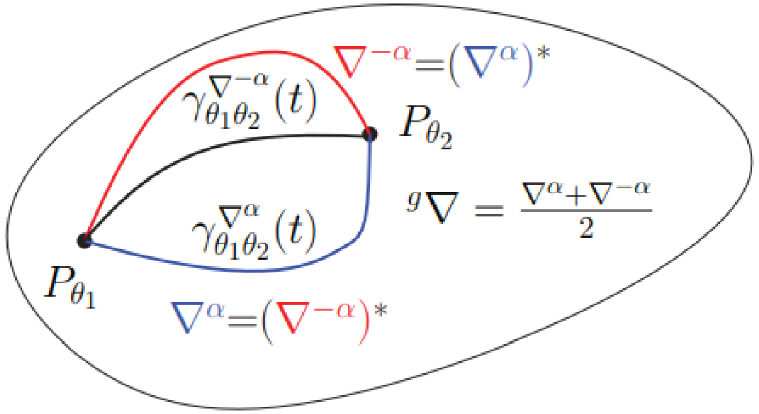


The Many Faces of Information Geometry

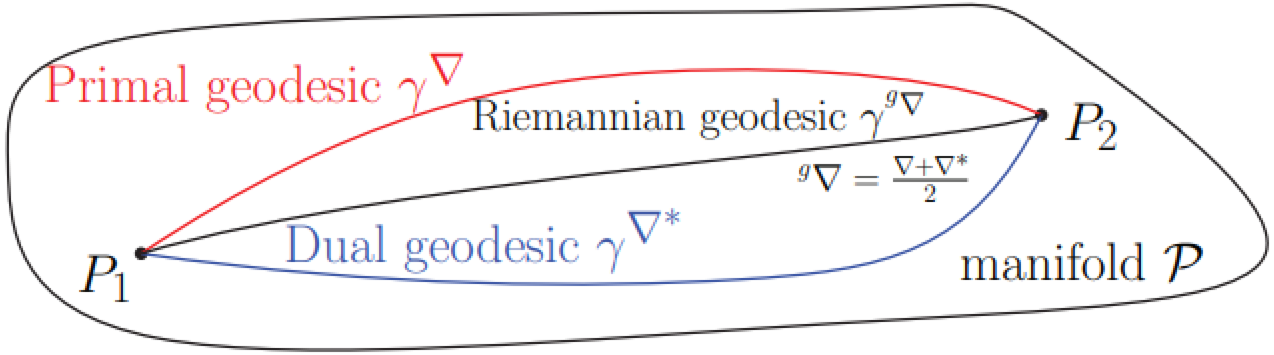
Fisher-Rao geometry
→ Fisher-Rao geodesic distance

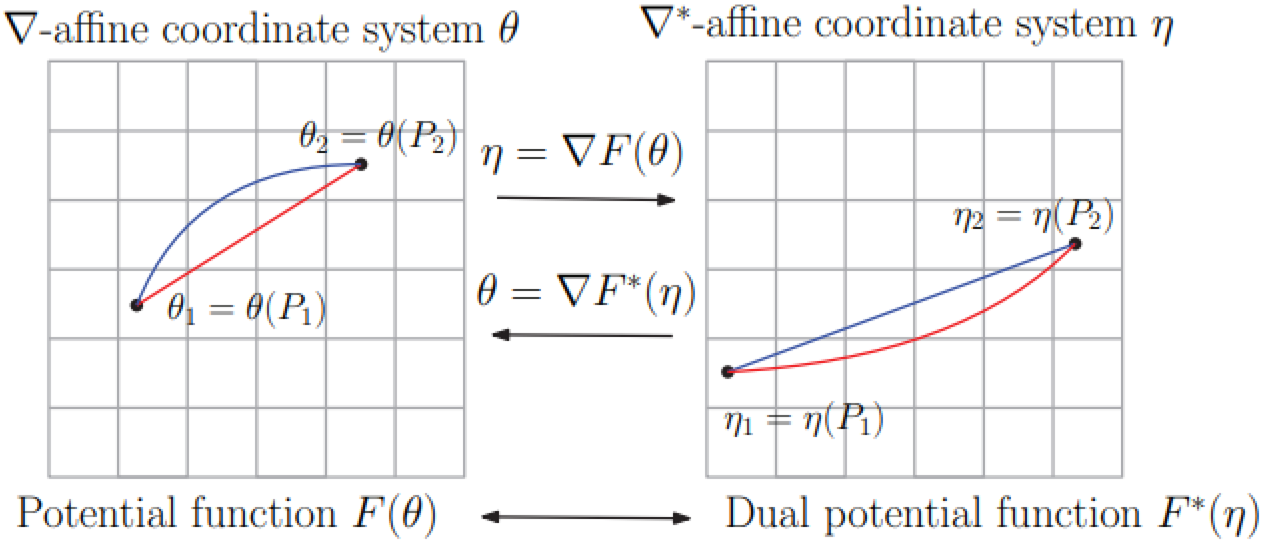


versus



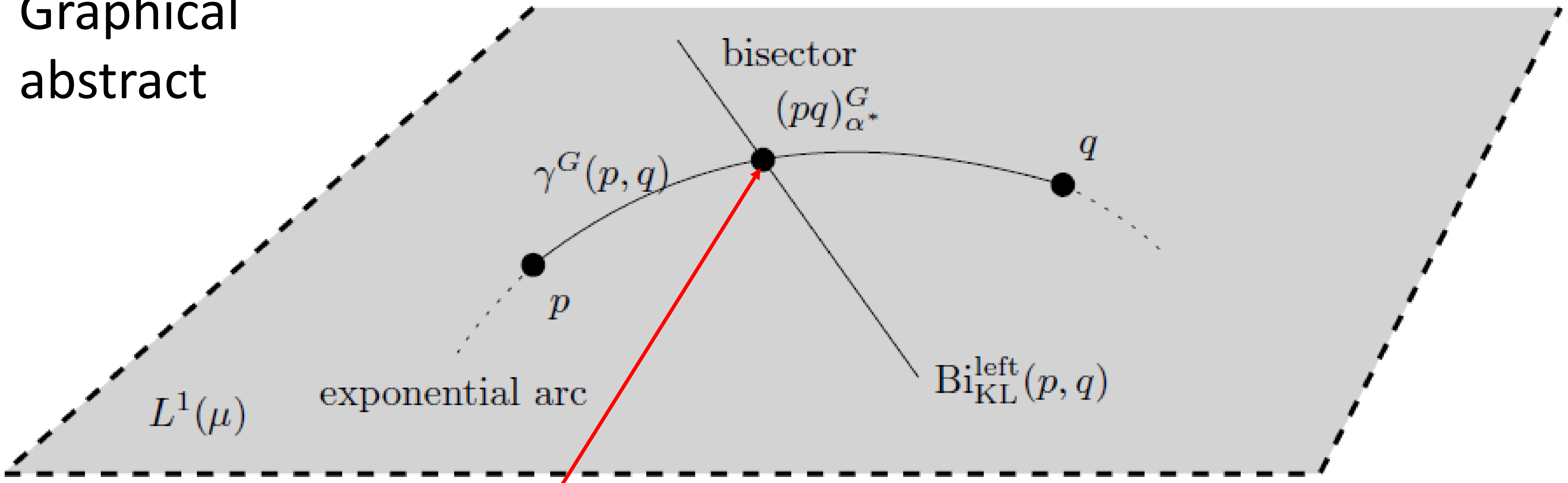
Dual α-geometry
→ No default divergence





Legendre-Fenchel transform

Graphical abstract



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

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

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

Entropy 2022
[2207.03745]