Computational Information Geometry on Matrix Manifolds

Frank Nielsen
Frank.Nielsen@acm.org
www.informationgeometry.org

Sony Computer Science Laboratories, Inc.

July 2013, ICTP, Trieste, IT

Geometry of matrix manifolds...

- Euclidean geometry, Fröbenius norm $\rightarrow$ distance:

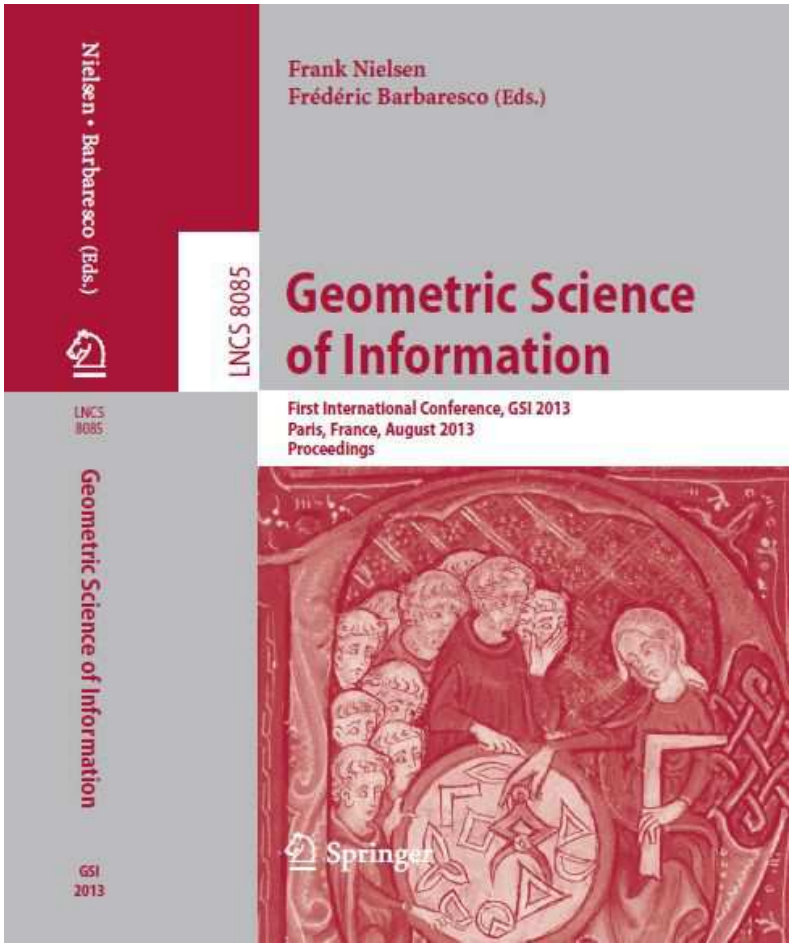
$$\|M\|_F^2 = \sum_{i,j} m_{ij}^2 = \sum_i \|M_{i*}\|_2^2 = \sum_j \|M_{*j}\|_2^2 = \mathrm{tr}(M^\top M)$$

- Riemannian geometry of symmetric positive definite (SPD) matrices [9, 2]
- Riemannian geometry of rank-deficient positive semi-definite (SPSD) matrices
  Stiefel/Grassman manifolds [3]
- Quantum geometry: SPD matrices with unit trace

"One geometry cannot be more true than another;
it can only be more convenient",
— Jules Henri Poincaré (1902)

Forthcoming conference (GSI)

28th-30th August, Paris.



What is Computational Information Geometry?

- What is Information? = *Essence* of data (datum="thing")
  (make it tangible $\rightarrow$ e.g., parameters of generative models)

- Can we do Intrinsic computing?
  (unbiased by any particular "data representation" $\rightarrow$ same results after recoding data)

- Geometry $\stackrel{?!}{\longrightarrow}$ Science of invariance
  (mother of Science, compass & ruler, Descartes analytic=coordinate/Cartesian, imaginaries, ...).
  ...the open-ended poetic mathematics!

Rationale for Computational Information Geometry

- Information is *...never void!* $\rightarrow$ lower bounds
  - Fisher information and Cramér-Rao lower bound (estimation)
  - Bayes error and Chernoff information (classification)
  - Coding and Shannon entropy (communication)
  - Program and Kolmogorov complexity (compression). (Unfortunately not computable!)
- Geometry:
  - Language (point, line, ball, dimension, orthogonal, projection, geodesic, immersion, etc.)
  - Power of characterization (eg., intersection of two pseudo-segments not admitting closed-form expression)
- Computing: Information computing. Seeking for mathematical *convenience* and mathematical *tricks* (RKHS in ML). How to manipulate "space of functions" ?!?

Example I: Matrix manifold

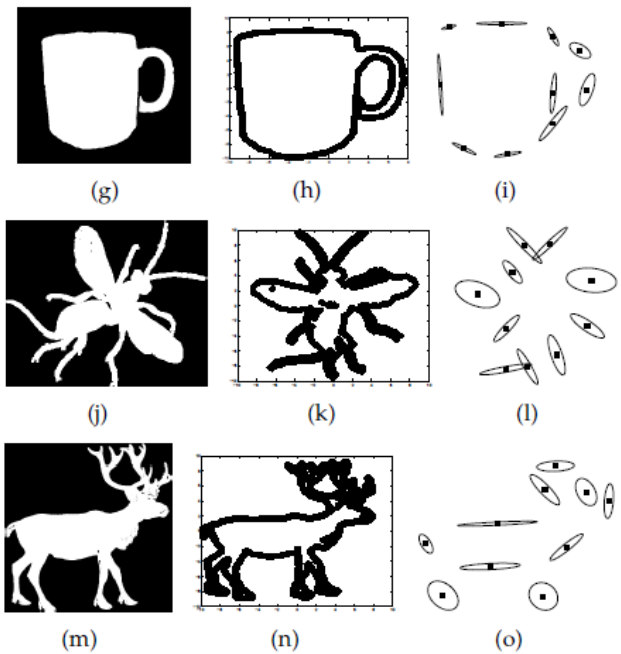
Pattern = Gaussian mixture models (universal class)

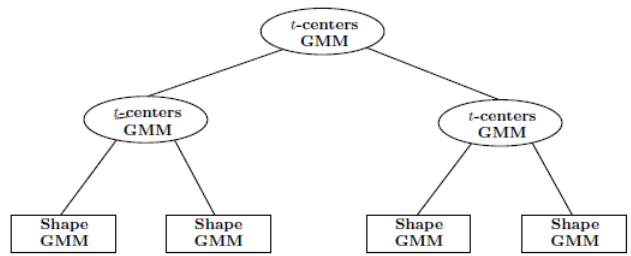
Statistical (dis)similarity/distance: total Bregman divergence (tBD, tKL).

Invariance: ..., $x_i \sim N(\mu_i, \Sigma_i)$, $y = A(x) = Lx + t$,

$y_i \sim N(L\mu_i + t, L\Sigma_i L^\top)$, $D(X_1 : X_2) = D(Y_1 : Y_2)$

(L: any invertible affine transformation, t a translation)





Shape Retrieval using Hierarchical Total Bregman Soft Clustering [7], IEEE PAMI, 2012.

Example II: Matrix manifolds

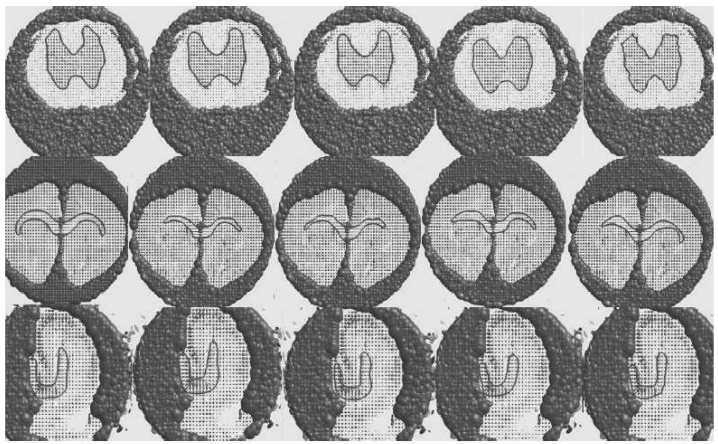
DTI: diffusion ellipsoids, tensor interpolation.

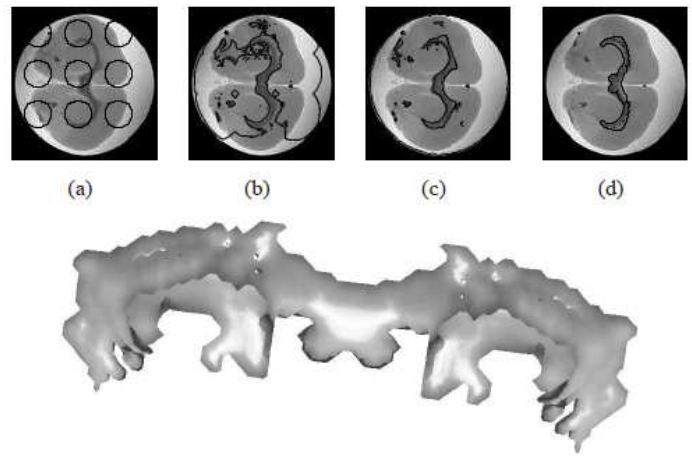
Pattern = zero-centered "Gaussians"

Statistical (dis)similarity/distance: total Bregman divergence (tBD, tKL).

Invariance: ..., $D(A^\top PA : A^\top QA) = D(P : Q)$, $A \in \mathrm{SL}(d)$: orthogonal matrix (volume/orientation preserving)

total Bregman divergence (tBD).



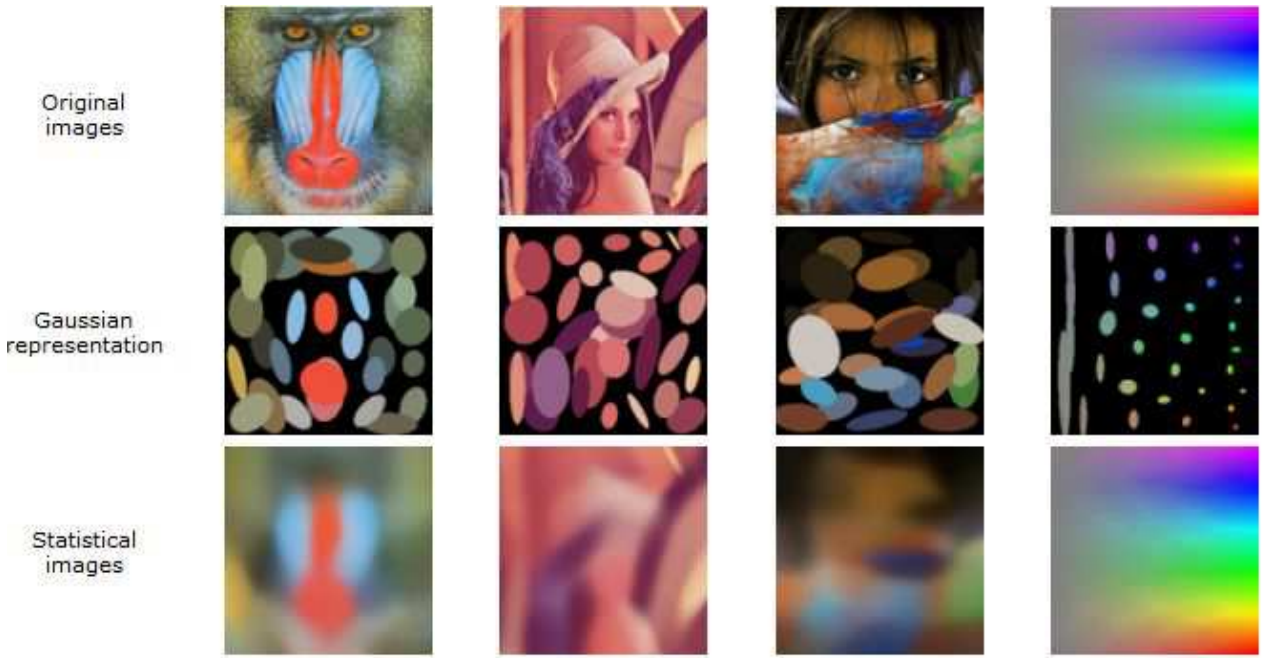


(3D rat corpus callosum)

Total Bregman Divergence and its Applications to DTI Analysis [20], IEEE TMI, 2011.

Example III: Gaussian manifolds

Consider 5D Gaussian Mixture Models (GMMs) of color images (image=RGBxy point set)



A Gaussian mixture model $\sum w_i N(\mu_i, \Sigma_i)$ is interpreted as a weighted point set $\{\theta_i = (\mu_i, \Sigma_i)\}$.

Matrix center points & clustering

Aggregation (matrix quantization for codebooks):

Given a data-set of matrices $\mathcal{M} = \{M_1, ..., M_n\} \subset \mathbb{M}$, compute a center matrix C.
Centering as a variational minimization problem:

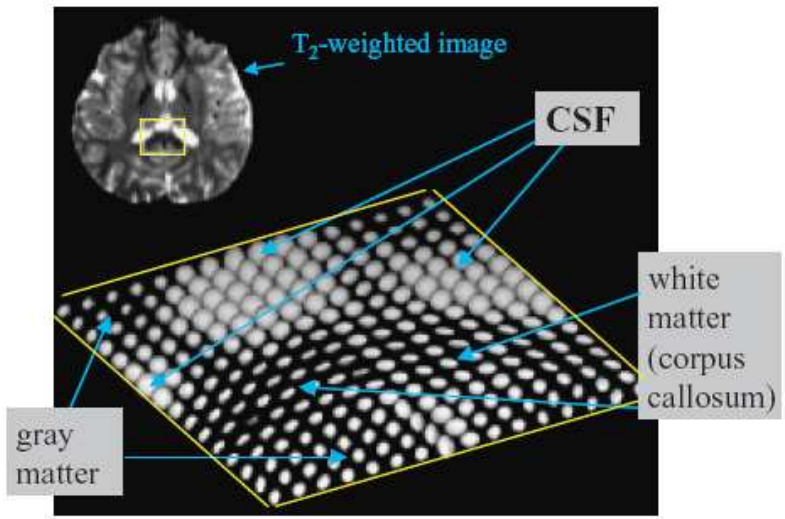
$$(OPT) : C_p = \arg\min_{C \in \mathbb{M}} \sum_i w_i \mathrm{distance}^p(C, M_i)$$

Notion of centrality, robustness to outliers?
For diagonal matrices, with "Euclidean" distance, usual geometric center points:

- median ($p = 1$): robust to outliers (Fermat-Weber point, no closed form),
- centroid ($p = 2$): breakdown point of 1 ($\rightarrow$ tBD)),
- circumcenter ($\lim p \rightarrow \infty$): minimize farthest point (minimax [1]).

Diffusion Tensor Magnetic Resonance Imaging

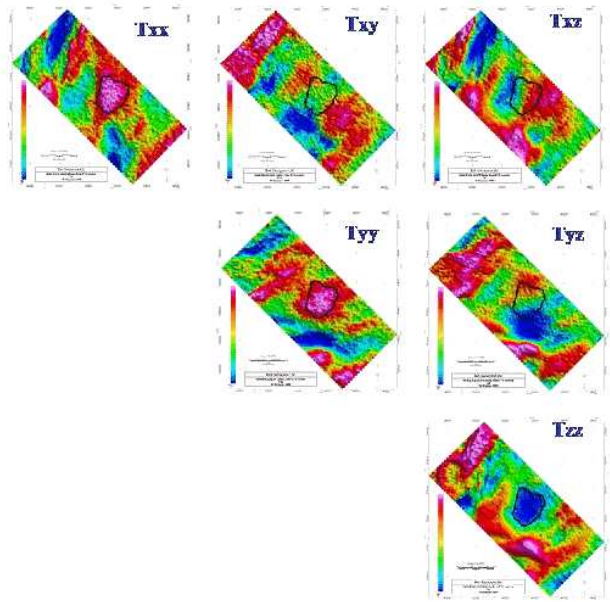
DT-MRI: Measures anisotropic diffusion of water molecules in a 3×3 tensor assigned to each voxel position (1990˜).
Used to analyze in-vivo connectivity patterns of brain tissues: gray matter, white matter (corpus callosum) and cerebrospinal fluid (CSF)



© Image courtesy Peter J. Basser
(Magnetic resonance imaging of the brain and spine, Chapter 31)

Gradiometry tensor: 3×3 SPSD matrices

Beyond the "constant" $g \simeq 9.81 m/s^2$. Gravity field measuring anisotropy.



$\rightarrow$ Oil & gas industry.
Courtesy of BellGeo.
`http://www.bellgeo.com/tech/technology_theory_of_FTG.html`

Structure tensors in computer vision

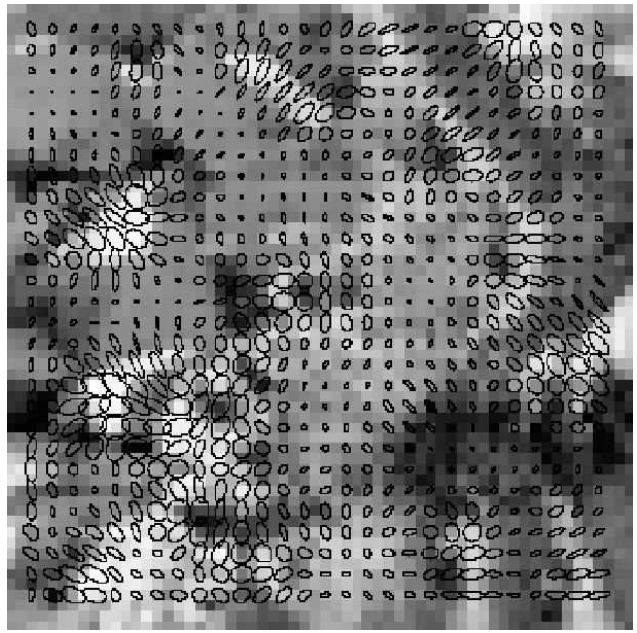
$\rightarrow$ Pioneered in image processing: tensor descriptor of a region at a pixel. (Harris-Stephens [6]).
Consider a kernel, and compute the tensor descriptor

$$\begin{aligned} T(p=(x,y)) &= K * \begin{bmatrix} I'^2(x) & I'(x)I'(y) \\ I'(y)I'(x) & I'(y)^2 \end{bmatrix}, \\ &= \sum_{u,v} w(u,v)\nabla I(u,v)(\nabla I(u,v))^T \end{aligned}$$

K: uniform, Gaussian kernel (eg., $s \times s$ window W centered at the pixel p)
$I'(x), I'(y)$: gradient, derivatives of the image.
Versatile method: corner detection, optical flow estimation, segmentation, stereo matching, etc.
$\rightarrow$ Tensor image processing

Harris-Stephens structure tensor (1988)

Deformation tensor field



Harris-Stephens combined corner-edge detector:

$$\boxed{R = \det\ T - k(\mathrm{tr}\ T)^2}$$

$\rightarrow$ Measures of tensor anisotropy.

Structure tensor represents local orientation (eigenvectors/eigenvalues).

Harris-Stephens' combined corner/edge detector (note)

Matrix with Fröbenius metric distance

Matrix space $\mathcal{M}$ with vectorial structure

$$d_E(P,Q) = \|P-Q\|_F \quad (1)$$
$$= \sqrt{\mathrm{tr}(P-Q)^T(P-Q)} \quad (2)$$

Centroid of tensors:

$$C_E = \frac{1}{n}\sum_{i=1}^{n} w_i T_i$$

$\rightarrow$ scalar average of each element of the tensor.
Tensor Field Segmentation Using Region Based Active Contour Model [21], ECCV, 2004.

Matrix vectorization & computational geometry

Computational geometry on $w \times h$-dimensional matrix spaces wrt Fröbenius distance amounts to computational geometry on Euclidean vector space for $D = w \times h$.

$\rightarrow$ Voronoi diagrams, smallest enclosing ball, minimum spanning tree, etc.

For symmetric matrices, we have $D = \frac{d(d+1)}{2}$ degrees of freedom, and vectorize as follows:

$$\begin{aligned} \|M\|_F &= \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d} m_{ij}^2} \\ &= \sqrt{\sum_{i=1}^{d} m_{ii}^2 + 2\sum_{i=1}^{d-1}\sum_{j=i+1}^{d} m_{ij}^2} \\ &= \|m\|_2 \end{aligned}$$

with $m = [m_{11} \ldots m_{dd} \sqrt{2} m_{12} \sqrt{2} m_{1d} \ldots \sqrt{2} m_{d-1,d}]^T = \vec{M}$.

Matrix functions

From the spectral decomposition:

$$M = UDU^\top$$

with $D = \lambda(M) = \mathrm{diag}(\lambda_1, ..., \lambda_d)$ the diagonal matrix of eigenvalues, consider real-valued function $x \mapsto f(x)$ to extend to matrices as

$$\boxed{f(M) = U\ \mathrm{diag}(f(\lambda_1), ..., f(\lambda_d))\ U^T}$$

Examples: $\log x, \exp x, |x|, x^{\frac{1}{2}}, x^2, etc.$
$O(d^3)$ SVD factorization complexity.

Riemannian cone of SPD matrices

Exponential maps from tangent planes (symmetric matrices Sym) to the manifold cone $\mathcal{C}$:

$$\exp_P \quad : \quad T_P\mathcal{C} = \mathrm{Sym} \to \mathcal{C}$$

Logarithmic maps from manifold cone $\mathcal{C}$ to tangent planes:

$$\log_P \quad : \quad \mathcal{C} \to T_P\mathcal{C} = \mathrm{Sym}$$
$$\log_P(Q) = P^{\frac{1}{2}} \log(P^{-\frac{1}{2}} Q P^{-\frac{1}{2}}) P^{\frac{1}{2}}$$

Map any point $Q \in \mathrm{Sym}_{++}$ to unique tangent vector at P such that $\gamma_0 = P$ and $\gamma_1 = Q$.
Geodesic equation:

$$\boxed{\gamma_t(P,Q) = P^{\frac{1}{2}} \left(P^{-\frac{1}{2}} Q P^{-\frac{1}{2}}\right)^t P^{\frac{1}{2}}}$$

Geodesic (metric length) distance:

$$\boxed{d_{\mathrm{R}}(P,Q) = \| \log P^{-\frac{1}{2}} Q P^{-\frac{1}{2}} \|}$$

Riemannian Karcher centroid

$$d_{\mathrm{R}}(P,Q) = \sqrt{\mathrm{tr}\log^2(P^{-1}Q)} = \sqrt{\sum_{i=1}^{d}\log^2\lambda_i}$$
$$= \|\log P^{-\frac{1}{2}}QP^{-\frac{1}{2}}\|$$

, where the λ_i's are the eigenvalues of $P^{-1}Q$.
($P^{-1}Q = Q^{\frac{1}{2}}P^{-1}Q^{\frac{1}{2}}$)
Unique mean characterized by $\sum_{i=1}^{n}\log(T_i^{-1}C_R) = 0$
Closed-form solution *only* for $n = 2$:
$C_R(P,Q) = P^{\frac{1}{2}}\left(P^{-\frac{1}{2}}QP^{-\frac{1}{2}}\right)^{\frac{1}{2}}P^{\frac{1}{2}}$ otherwise iterative approximation ($C_R = \lim_{t\to\infty} C_t$):

$$\boxed{C_{t+1} = C_t \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log C_t^{-1}T_i\right).}$$

Riemannian minimax SPD center (circumcenter [1])

Case of $p = \infty$, center that minimizes the maximum distance.

GEO-ALG:
Starts with $c_1 \in P$ and iteratively update the current circumcenter as follows: $c_{i+1} = \mathrm{Geodesic}(c_i, f_i, \frac{1}{i+1})$, where f_i denotes the farthest point of P to c_i, and $\mathrm{Geodesic}(p, q, t)$ denotes the intermediate point m on the geodesic passing through p and q such that $\rho(p, m) = t \times \rho(p, q)$.

Geodesic:

$$\gamma_t(P, Q) = P^{\frac{1}{2}} \left(P^{-\frac{1}{2}} Q P^{-\frac{1}{2}} \right)^t P^{\frac{1}{2}}$$

Find t such that $\sum_{i=1}^d \log^2 \lambda_i^t = t^2 \sum_{i=1}^d \log^2 \lambda_i = r^2 \sum_{i=1}^d \log^2 \lambda_i$.
That is $t = r$.
Prove core-set and guaranteed convergence.

Matrices as parameters in probability distributions

Exponential families: Gaussian, Wishart, etc.:

$$\boxed{p(x;\lambda) = p_F(x;\theta) = \exp\left(\langle t(x), \theta\rangle - F(\theta) + k(x)\right).}$$

Example: Poisson distribution

$$p(x;\lambda) = \frac{\lambda^x}{x!}\exp(-\lambda),$$

- the sufficient statistic $t(x) = x$,
- $\theta = \log\lambda$, the natural parameter,
- $F(\theta) = \exp\theta$, the log-normalizer $\rightarrow$ CONVEX,
- and $k(x) = -\log x!$ the carrier measure (with respect to the counting measure).

Gaussians as an exponential family

$$p(x;\lambda) = p(x;\mu,\Sigma) = \frac{1}{2\pi\sqrt{\det\Sigma}} \exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu))}{2}\right)$$

- $\theta = (\Sigma^{-1}\mu, \frac{1}{2}\Sigma^{-1}) \in \Theta = \mathbb{R}^d \times \mathbb{K}_{d\times d}$, with $\mathbb{K}_{d\times d}$ cone of positive definite matrices,
- $F(\theta) = \frac{1}{4}\mathrm{tr}\theta_2^{-1}\theta_1\theta_1^T - \frac{1}{2}\log\det\theta_2 + \frac{d}{2}\log\pi \rightarrow$ CONVEX
- $t(x) = (x, -x^Tx)$,
- $k(x) = 0$.

Inner product : composite, sum of a dot product and a matrix trace :

$$\langle\theta,\theta'\rangle = \theta_1^T\theta_1' + \mathrm{tr}\theta_2^T\theta_2'.$$

The coordinate transformation $\tau : \Lambda \rightarrow \Theta$ is given for $\lambda = (\mu,\Sigma)$ by

$$\tau(\lambda) = \left(\lambda_2^{-1}\lambda_1, \frac{1}{2}\lambda_2^{-1}\right), \qquad \tau^{-1}(\theta) = \left(\frac{1}{2}\theta_2^{-1}\theta_1, \frac{1}{2}\theta_2^{-1}\right)$$

Convex duality: Legendre transformation

- For a strictly convex and differentiable function $F : \mathcal{X} \to \mathbb{R}$:

$$F^*(y) = \sup_{x \in \mathcal{X}} \{ \underbrace{\langle y, x \rangle - F(x)}_{I_F(y;x);} \}$$

- Maximum obtained for $y = \nabla F(x)$:

$$\nabla_x I_F(y; x) = y - \nabla F(x) = 0 \Rightarrow y = \nabla F(x)$$

- Maximum *unique* from convexity of F ($\nabla^2 F \succ 0$):

$$\nabla_x^2 I_F(y; x) = -\nabla^2 F(x) \prec 0$$

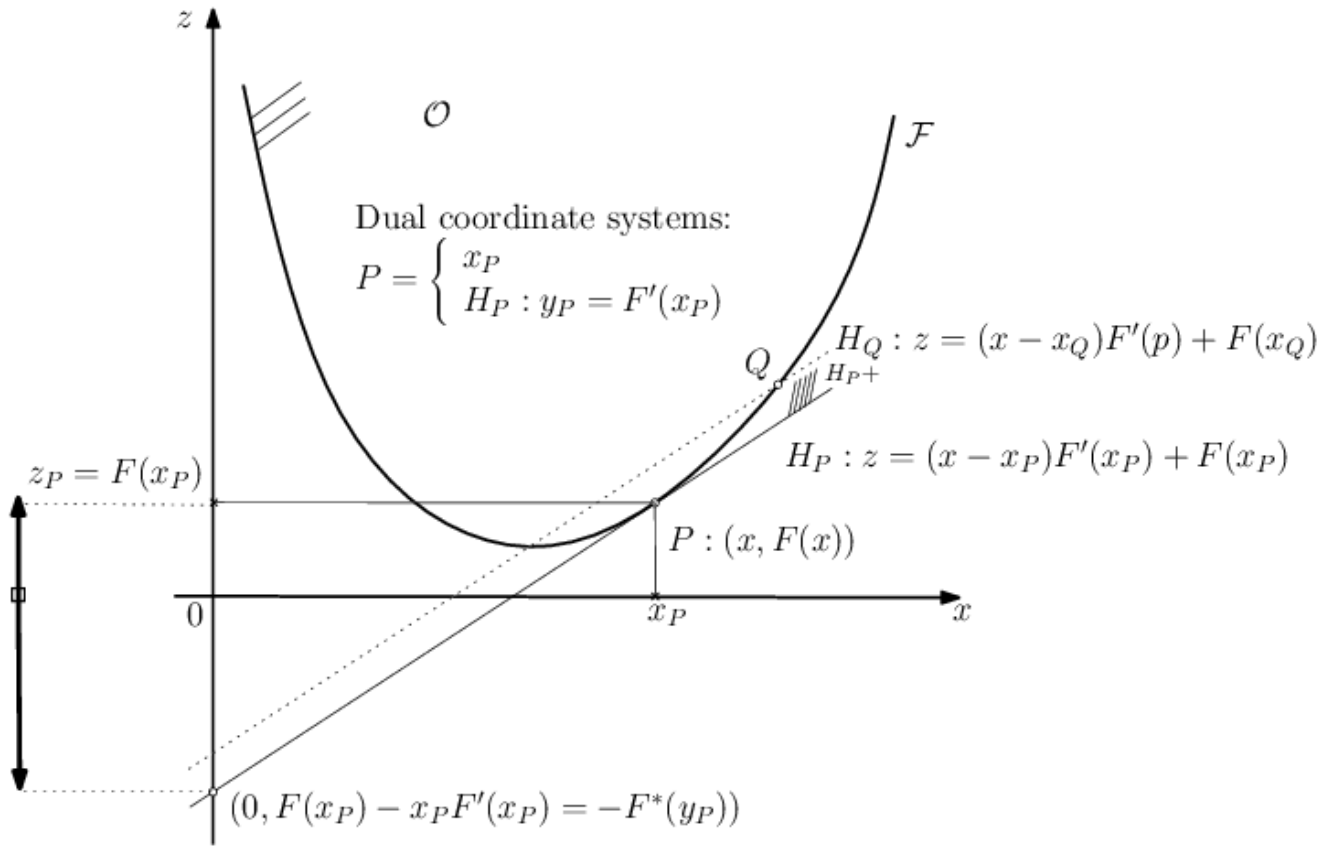
- Convex conjugates:

$$(F, \mathcal{X}) \Leftrightarrow (F^*, \mathcal{Y}), \qquad \mathcal{Y} = \{\nabla F(x) \mid x \in \mathcal{X}\}$$

Legendre duality: Geometric interpretation

Consider the epigraph of F as a convex object:

- convex hull (V-representation), versus
- half-space (H-representation).



Legendre transform also called *"slope" transform*.

Legendre duality & Canonical divergence

- Convex conjugates have *functional inverse* gradients
  $\nabla F^{-1} = \nabla F^*$
  ∇F^* may require numerical approximation
  (not always available in analytical closed-form)
- Involution: $(F^*)^* = F$ with $\nabla F^* = (\nabla F)^{-1}$.
- Convex conjugate F^* expressed using $(\nabla F)^{-1}$:

$$\begin{aligned} F^*(y) &= \langle x, y\rangle - F(x), x = \nabla_y F^*(y) \\ &= \langle (\nabla F)^{-1}(y), y\rangle - F((\nabla F)^{-1}(y)) \end{aligned}$$

- Fenchel-Young inequality at the heart of canonical divergence:

$$F(x) + F^*(y) \geq \langle x, y\rangle$$

$$\boxed{A_F(x : y) = A_{F^*}(y : x) = F(x) + F^*(y) - \langle x, y\rangle \geq 0}$$

Dual Bregman divergences & canonical divergence [14]

$$\begin{aligned}
\mathrm{KL}(P:Q) &= E_P\left[\log \frac{p(x)}{q(x)}\right] \geq 0 \\
&= B_F(\theta_Q:\theta_P) = B_{F^*}(\eta_P:\eta_Q) \\
&= F(\theta_Q) + F^*(\eta_P) - \langle \theta_Q, \eta_P \rangle \\
&= A_F(\theta_Q:\eta_P) = A_{F^*}(\eta_P:\theta_Q)
\end{aligned}$$

with θ_Q (natural parameterization) and $\eta_P = E_P[t(X)] = \nabla F(\theta_P)$ (moment parameterization).

$$\mathrm{KL}(P:Q) = \underbrace{\int p(x)\log\frac{1}{q(x)}\mathrm{d}x}_{H^\times(P:Q)} - \underbrace{\int p(x)\log\frac{1}{p(x)}\mathrm{d}x}_{H(p)=H^\times(P:P)}$$

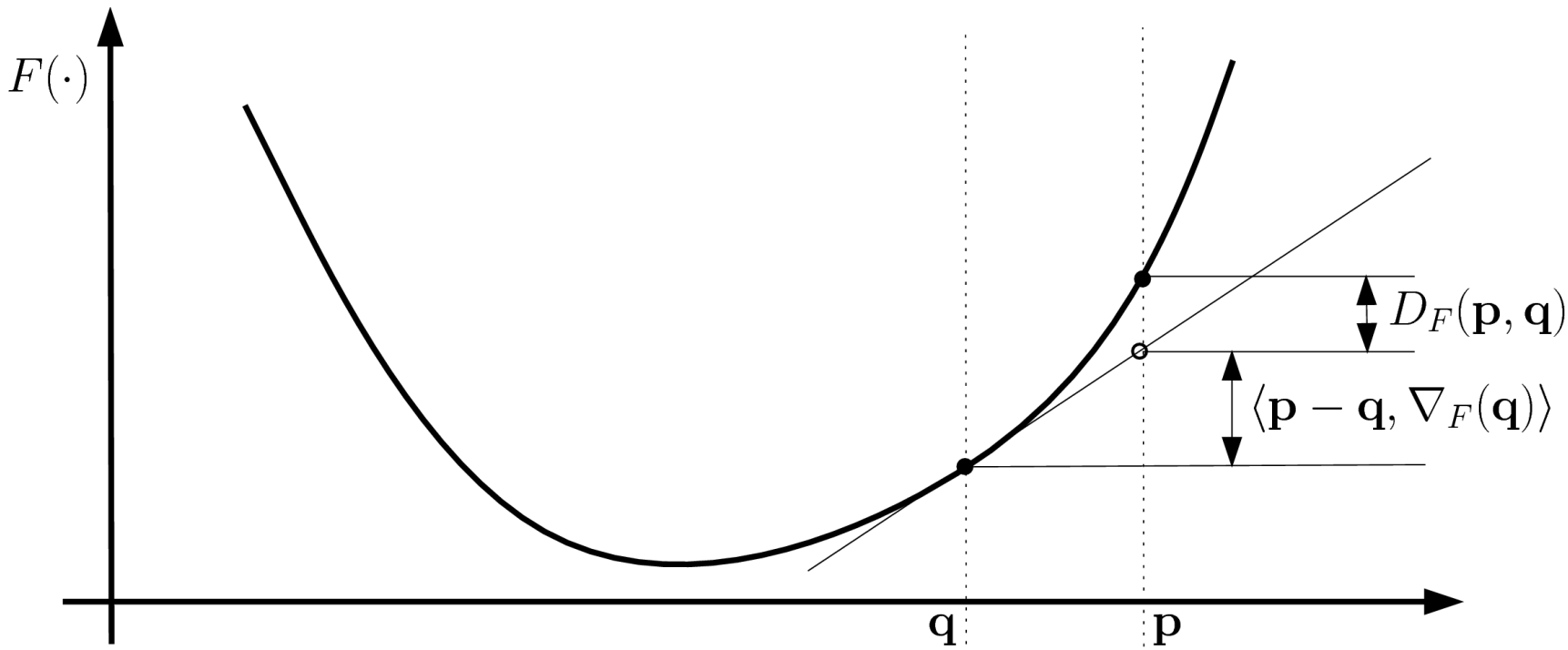
Shannon cross-entropy and entropy of EF [14]:

$$\begin{aligned}
H^\times(P:Q) &= F(\theta_Q) - \langle \theta_Q, \nabla F(\theta_P)\rangle - E_P[k(x)] \\
H(P) &= F(\theta_P) - \langle \theta_P, \nabla F(\theta_P)\rangle - E_P[k(x)] \\
H(P) &= -F^*(\eta_P) - E_P[k(x)]
\end{aligned}$$

Bregman divergence: Geometric interpretation (I)

Potential function F, graph plot $\mathcal{F} : (x, F(x))$.

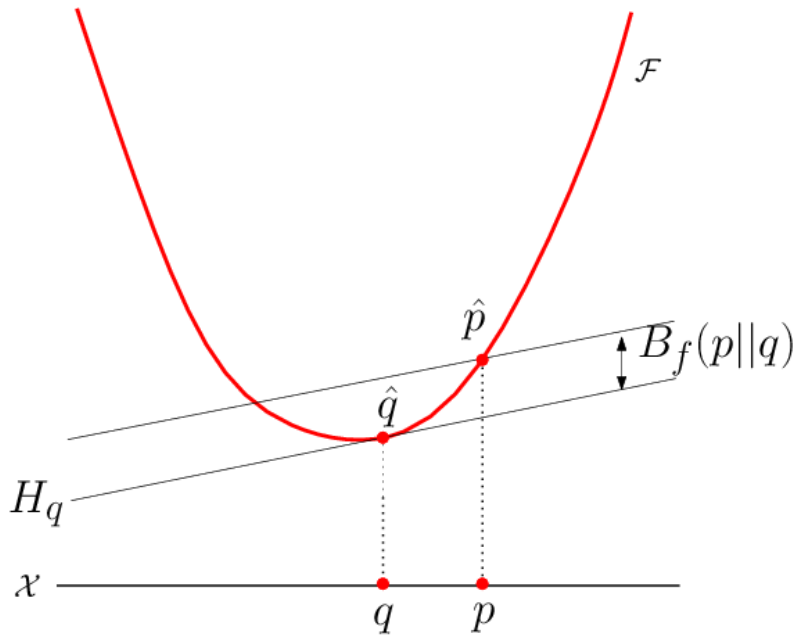
$$\boxed{D_F(p : q) = F(p) - F(q) - \langle p - q, \nabla F(q) \rangle}$$



Bregman divergence: Geometric interpretation (II)

Potential function f, graph plot $\mathcal{F} : (x, f(x))$.

$$B_f(p||q) = f(p) - f(q) - (p - q)f'(q)$$

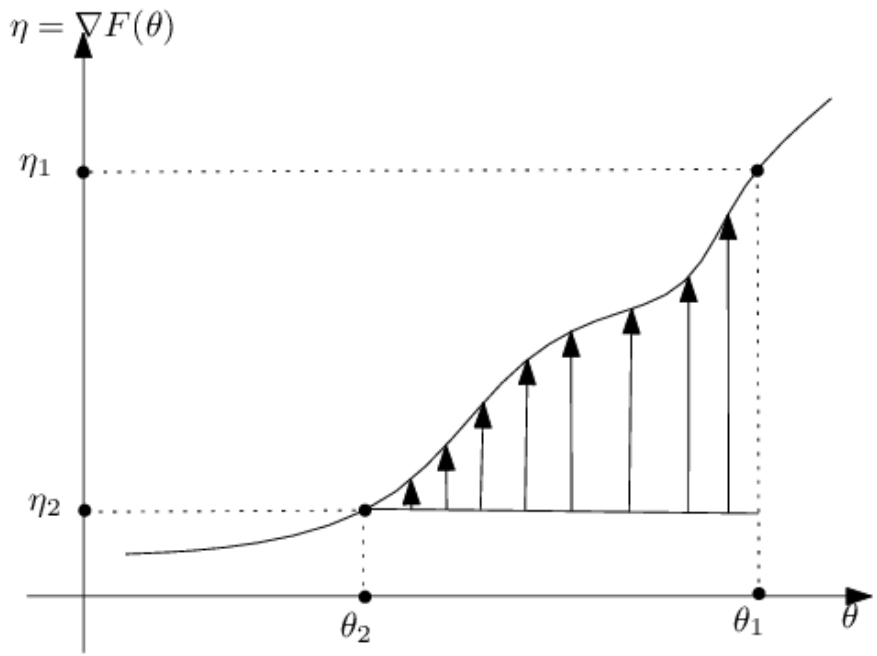


$B_f(.||q)$: vertical distance between the hyperplane H_q tangent to $\mathcal{F}$ at lifted point $\hat{q}$, and the translated hyperplane at $\hat{p}$.

Bregman divergence: Geometric interpretation (III)

Bregman divergence and path integrals

$$
\begin{aligned}
B(\theta_1 : \theta_2) &= F(\theta_1) - F(\theta_2) - \langle \theta_1 - \theta_2, \nabla F(\theta_2) \rangle, && (3)\\
&= \int_{\theta_2}^{\theta_1} \langle \nabla F(t) - \nabla F(\theta_2), \mathrm{d}t \rangle, && (4)\\
&= \int_{\eta_1}^{\eta_2} \langle \nabla F^*(t) - \nabla F^*(\eta_1), \mathrm{d}t \rangle, && (5)\\
&= B^*(\eta_2 : \eta_1) && (6)
\end{aligned}
$$



Matrix Bregman divergences [4, 16]

Choose F a real-valued functional generator and extend F to matrices:

$$F(X) = \mathrm{tr}(\Psi(X))$$

$$\Psi(X) = \sum_{k \geq 0} t_{F,k} N^k$$

($t_{F,k}$ from the Taylor expansion of real-valued F)

$$B_F(P : Q) = F(P) - F(Q) - \mathrm{tr}((P - Q)^\top \nabla F(Q)),$$

$$\nabla F(X) = \sum_{k \geq 0} t'_{F,k} N^k$$

($t'_{F,k}$ from the Taylor expansion of real-valued F')

Matrix Bregman divergences [16]

Table 15.1 Examples of Bregman matrix divergences. $\boldsymbol{\Sigma}$ is positive definite, $\cdot$ is the Hadamard product, $l, n \in \mathbb{R}^d$ and $\mathbf{1}$ is the all-1 vector

ψ	$D_\psi(\mathbf{L}\|\mathbf{N})$	Comments
$x \log x - x$	$\operatorname{Tr}(\mathbf{L}(\log \mathbf{L} - \log \mathbf{N}) - \mathbf{L} + \mathbf{N})$	von Neumann divergence
id.	id. + constraint $\operatorname{Tr}(\mathbf{L}) = \operatorname{Tr}(\mathbf{N})$	Umegaki's relative entropy [22]
$-\log x$	$\operatorname{Tr}\left(-\log \mathbf{L} + \log \mathbf{N} + \mathbf{L}\mathbf{N}^{-1}\right) - d$	logdet divergence [25]
$x \log x + (1-x)\log(1-x)$	$\operatorname{Tr}(\mathbf{L}(\log \mathbf{L} - \log \mathbf{N}) + (\mathbf{I} - \mathbf{L})(\log(\mathbf{I} - \mathbf{L}) - \log(\mathbf{I} - \mathbf{N})))$	binary quantum relative entropy
x^p $(p > 1)$	$\operatorname{Tr}\left(\mathbf{L}^p - p\mathbf{L}\mathbf{N}^{p-1} + (p-1)\mathbf{N}^p\right)$	
if $p = 2$	$\operatorname{Tr}\left(\mathbf{L}^2 - 2\mathbf{L}\mathbf{N} + \mathbf{N}^2\right)$	Mahalanobis divergence
	$= (l-n)^\top \boldsymbol{\Sigma}^{-1}(l-n)$ if $\mathbf{L} \doteq (\boldsymbol{\Sigma}^{-1/2} l)\mathbf{1}^\top \cdot \mathbf{I}, \mathbf{N} \doteq (\boldsymbol{\Sigma}^{-1/2} n)\mathbf{1}^\top \cdot \mathbf{I}$	
$\log(1 + \exp(x))$	$\operatorname{Tr}\left(\log(\mathbf{I} + \exp(\mathbf{L})) - \log(\mathbf{I} + \exp(\mathbf{N})) - (\mathbf{L} - \mathbf{N})(\mathbf{I} + \exp(\mathbf{N}))^{-1} \exp(\mathbf{N})\right)$	Dual bit entropy
$-\sqrt{1 - x^2}$	$\operatorname{Tr}\left((\mathbf{I} - \mathbf{L}\mathbf{N})(\mathbf{I} - \mathbf{N}^2)^{-1/2} - (\mathbf{I} - \mathbf{L}^2)^{1/2}\right)$	
$\exp(x)$	$\operatorname{Tr}(\exp(\mathbf{L}) - (\mathbf{L} - \mathbf{N} + \mathbf{I})\exp(\mathbf{N}))$	
$\phi_p \circ \psi_p$ $(p > 1$, Eq. (15.3))	$\frac{1}{2}\|\mathbf{L}\|_p^2 - \frac{1}{2}\|\mathbf{N}\|_p^2 - \frac{1}{\|\mathbf{N}\|_p^{p-2}} \operatorname{Tr}\left((\mathbf{L} - \mathbf{N})\mathbf{N}\|\mathbf{N}\|^{p-2}\right)$	Bregman–Schatten p-divergence

Particular case: Bregman Schatten p-divergences [5, 16]

Schatten p-norm of real symmetric matrix X:
(unitarily invariant matrix norms)

$$\|X\|_p = \|\lambda(X)\|_p$$

Bregman generator:

$$F(X) = \frac{1}{2}\|X\|_p^2$$

Used in regularized convex optimization [5], matrix data mining [16].

Matrix Legendre transformation

Extends classical Legendre-Fenchel transformation:

$$F^*(\eta) = \sup_{\mathrm{spec}(\theta) \subseteq \mathrm{dom}(F)} \mathrm{tr}(\theta\eta^\top) - F(\theta)$$

$$D_F(\theta_P : \theta_Q) = D_{F^*}(\eta_Q : \eta_P) = F(\theta) + F^*(\eta) - \mathrm{tr}(\theta\eta^\top)$$

θ and η are dual matrix coordinate systems on the matrix manifold.

Non-metric differential structure with dual coordinate systems.

Bregman matrix means

$$B_F(X,P) = F(X) - F(P) - \mathrm{tr}((X-P)^T \nabla F(P)),$$

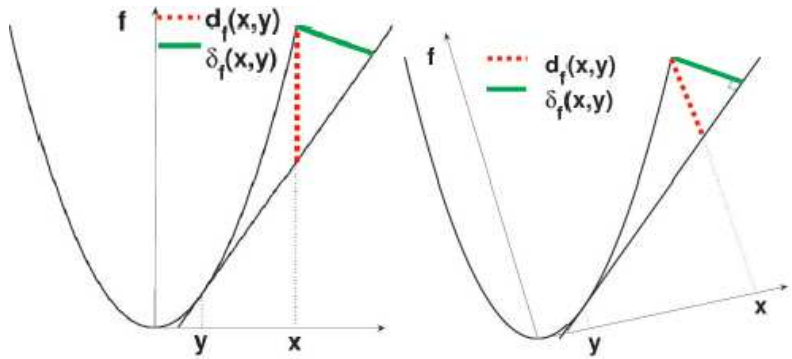
$F(\cdot)$: strictly convex and differentiable function on an open convex space.

$$\boxed{C = \nabla F^{-1}\left(\sum_{i=1}^{n} w_i \nabla F(T_i)\right)}$$

quasi-arithmetic mean for ∇F.
Since $B_F(X,P) \neq B_F(P,X)$, define a *right-sided* centroid M':
Find the center of mass [13] (independent of generator F)
$F(X) = \mathrm{tr}(X^T X)$: the quadratic matrix entropy,
$F(X) = -\log\det X$: the matrix Burg entropy, and
$F(X) = \mathrm{tr}(X\log X - X)$: the von Neumann entropy [19, 18, 15] (Umegaki quantum relative entropy).

Total Bregman divergences (tBD)

Instead of "vertical" projection in Bregman divergence, consider perpendicular projection.
(Analogy with least squares and total least squares regression.)



$$\mathrm{tB}_F(P,Q) = \frac{B_F(P,Q)}{\sqrt{1+\|\nabla F(Q)\|^2}}$$

$\rightarrow$ proven statistically robust.
Applications to robust DT-MRI segmentation [8].

Matrix Jensen/Burbea-Rao divergences [10]

Convexity gap defines a divergence

$$\boxed{\mathrm{BR}_F(P,Q) = \frac{F(P)+F(Q)}{2} - F\left(\frac{P+Q}{2}\right) \geq 0}$$

- $F(X) = \mathrm{tr}(X^T X)$: the quadratic matrix entropy,
- $F(X) = -\log\det X$: the matrix Burg entropy, and
- $F(X) = \mathrm{tr}(X\log X - X)$: the von Neumann entropy.
- etc.

Smooth family of convex generators [12, 17]

1-parameter family of generators:

$$F_\alpha(X) = \frac{1}{\alpha(1-\alpha)}\mathrm{tr}(\alpha X - X^\alpha + (1-\alpha)I), \alpha \neq \{0,1\}$$

$$B_\alpha(P:Q) = \frac{1}{\alpha(1-\alpha)}\mathrm{tr}(Q^\alpha - P^\alpha + \alpha Q^{\alpha-1}(P-Q))$$

$$\nabla F_\alpha(X) = \frac{1}{\alpha-1}(I - X^{\alpha-1}) \qquad \nabla F_\alpha^{-1}(X) = (I-(\alpha-1)X)^{\frac{1}{\alpha-1}}$$

When $\alpha \to 1$, $\nabla F_\alpha(X) = \nabla F_1(X) = \log X$. When $\alpha \to 0$, $\nabla F_\alpha(X) = \nabla F_0(X) = X^{-1} - I$.

- $\alpha = 2$: Quadratic matrix information
- $\alpha \to 1$: von Neumann information
- $\alpha \to 0$: Burg log-det information

Jensen (Burbea-Rao) divergences

Based on Jensen's inequality for a convex function F:

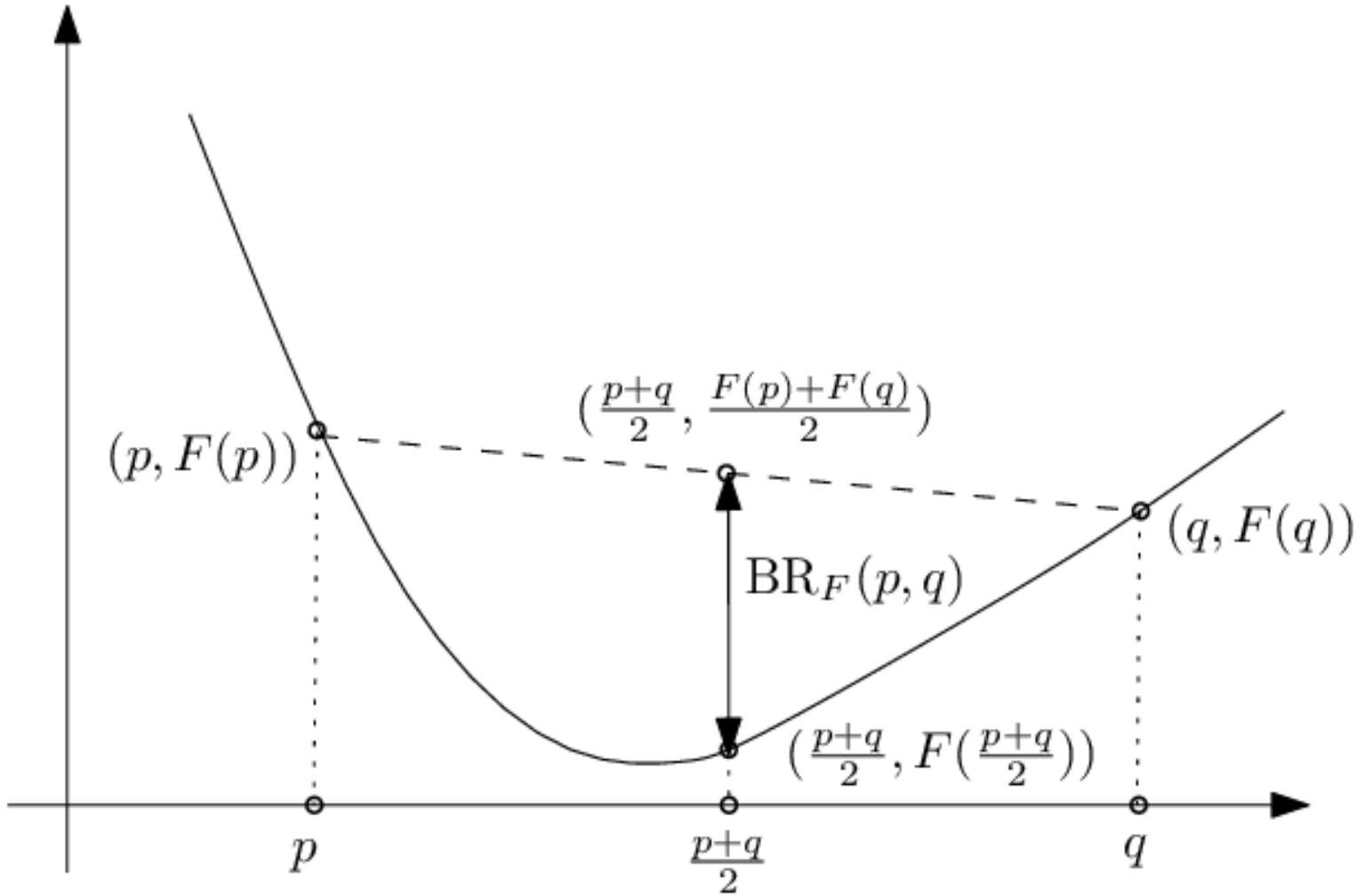
$$\mathrm{BR}_F(X,P) = \frac{F(X)+F(P)}{2} - F\left(\frac{X+P}{2}\right) \stackrel{\mathrm{def}}{=} \geq 0.$$

strictly convex function $F(\cdot)$.
Includes the special case of Jensen-Shannon divergence:

$$\mathrm{JS}(p,q) = H\left(\frac{p+q}{2}\right) - \frac{H(p)+H(q)}{2}$$

$F(x) = -H(x)$, the negative Shannon entropy $H(x) = -x\log x$.
$\rightarrow$ generators are convex and entropies are concave (negative generators)

Visualizing Burbea-Rao divergences



include Squared Mahalanobis distance.

Burbea-Rao from Symmetrizing Bregman divergences [13]

- **Jeffreys-Bregman divergences.**

$$\begin{aligned} S_F(p;q) &= \frac{B_F(p,q)+B_F(q,p)}{2} \\ &= \frac{1}{2}\langle p-q, \nabla F(p)-\nabla F(q)\rangle, \end{aligned}$$

- **Jensen-Bregman divergences (diversity index).**

$$\begin{aligned} J_F(p;q) &= \frac{B_F(p,\frac{p+q}{2})+B_F(q,\frac{p+q}{2})}{2} \\ &= \frac{F(p)+F(q)}{2} - F\left(\frac{p+q}{2}\right) = \mathrm{BR}_F(p,q) \end{aligned}$$

Skew Burbea-Rao divergences

$$\begin{aligned} \mathrm{BR}_F^{(\alpha)} &: \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+ \\ \mathrm{BR}_F^{(\alpha)}(p,q) &= \alpha F(p) + (1-\alpha)F(q) - F(\alpha p + (1-\alpha)q) \end{aligned}$$

$$\begin{aligned} \mathrm{BR}_F^{(\alpha)}(p,q) &= \alpha F(p) + (1-\alpha)F(q) - F(\alpha p + (1-\alpha)q) \\ &= \mathrm{BR}_F^{(1-\alpha)}(q,p) \end{aligned}$$

Skew symmetrization of Bregman divergences:

$$\alpha B_F(p, \alpha p + (1-\alpha)q) + (1-\alpha) B_F(q, \alpha p + (1-\alpha)q) \stackrel{\mathrm{def}}{=} \mathrm{BR}_F^{(\alpha)}(p,q)$$

$=$ skew Jensen-Bregman divergences.

Bregman divergences = asymptotic skewed Jensen divergences

$$
\begin{aligned}
B_F(p,q) &= \lim_{\alpha\to 1} \frac{1}{1-\alpha} \mathrm{BR}_F^{(\alpha)}(p,q) \\
B_F(q,p) &= \lim_{\alpha\to 0} \frac{1}{\alpha} \mathrm{BR}_F^{(\alpha)}(p,q)
\end{aligned}
$$

Burbea-Rao/Jensen centroids

$(p = 1)$

$$\mathrm{OPT} : C_F = \arg\min_X \sum_{i=1}^n w_i \mathrm{BR}_F^{(\alpha_i)}(X, T_i) = \arg\min_x L(x)$$

Wlog., equivalent to minimize

$$\boxed{E(c) = (\sum_{i=1}^n w_i \alpha_i) F(C) - \sum_{i=1}^n w_i F(\alpha_i C + (1 - \alpha_i) T_i)}$$

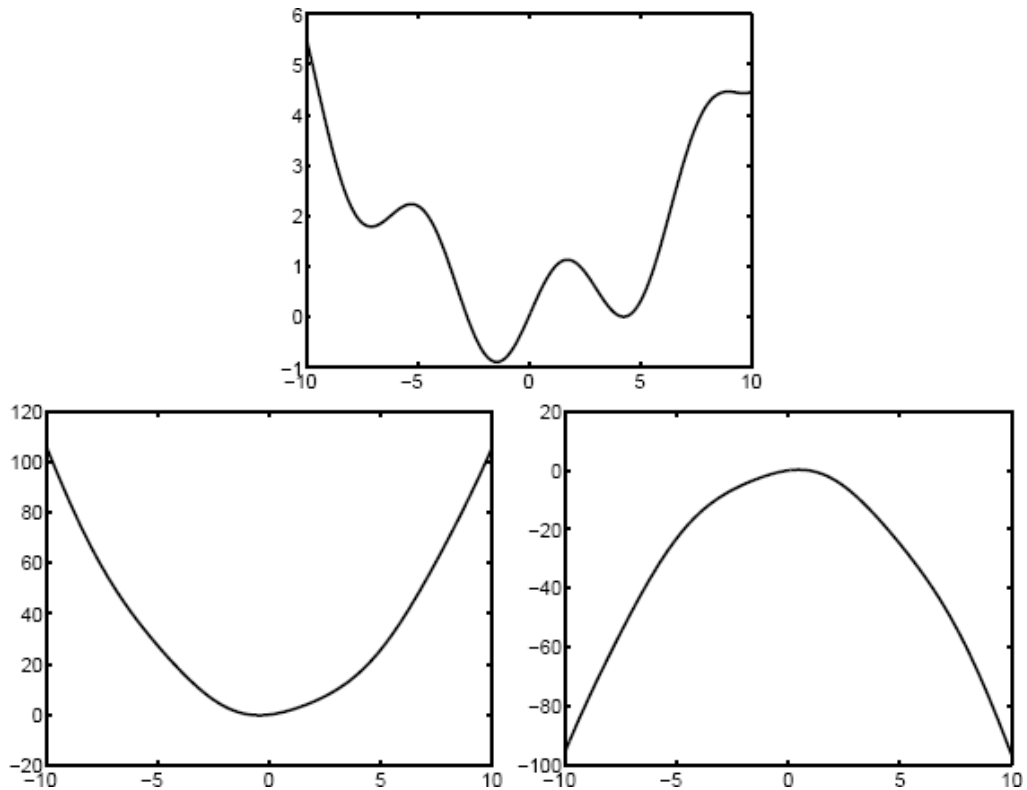
Sum $E = F + G$ of convex F + concave G function $\Rightarrow$
Convex-ConCave Procedure (CCCP, NIPS*01)
Start from arbitrary c_0, and iteratively update as:

$$\nabla F(C_{t+1}) = -\nabla G(C_t)$$

$\Rightarrow$ guaranteed convergence to a (local) minimum.

ConCave Convex Procedure (CCCP)

$$\min_x E(x) = F(x) + G(x)$$
$$\nabla F(c_{t+1}) = -\nabla G(c_t)$$



Decomposition may not be unique...

Iterative algorithm for Burbea-Rao centroids

Apply CCCP scheme

$$\nabla F(C_{t+1}) = \frac{1}{\sum_{i=1}^{n} w_i \alpha_i} \sum_{i=1}^{n} w_i \alpha_i \nabla F\left(\alpha_i C_t + (1-\alpha_i) T_i\right)$$

$$\boxed{C_{t+1} = \nabla F^{-1}\left(\frac{1}{\sum_{i=1}^{n} w_i \alpha_i} \sum_{i=1}^{n} w_i \alpha_i \nabla F\left(\alpha_i C_t + (1-\alpha_i) T_i\right)\right)}$$

Get arbitrarily fine approximations of the (skew) Burbea-Rao matrix centroids and barycenters.

Special case: α-log det divergence [15, 11]

Cone of Hermitian positive definite matrices (self-adjoint matrices $M^H = \bar{M}^T = M$).

$F(X) = -\log detX$, $\nabla F(X) = \nabla F^{-1}(X) = -X^{-1}$

Burbea-Rao α-log det divergences:

$$D_{\mathrm{ld}}^{(\alpha)}(P,Q) = \begin{cases} \mathrm{tr}(Q^{-1}P - I) - \log\det(Q^{-1}P)) & \alpha = 1 \\ \frac{4}{1-\alpha^2}\log\frac{\det(\frac{1-\alpha}{2}P + \frac{1+\alpha}{2}Q)}{(\det P)^{\frac{1-\alpha}{2}}(\det Q)^{\frac{1+\alpha}{2}}} & \alpha \in \mathbb{R}\backslash\{-1,1\} \\ \mathrm{tr}(P^{-1}Q - I) - \log\det(P^{-1}Q) & \alpha = -1 \end{cases}$$

Start with $C_1 = \frac{1}{n}\sum_{i=1}^n T_i$,

$$\boxed{C_{t+1} = n\left(\sum_{i=1}^n \left(\frac{1-\alpha}{2}T_i + \frac{1+\alpha}{2}C_t\right)^{-1}\right)^{-1}}$$

$\rightarrow$ unique global mean (obtained from CCCP).

Bhattacharyya coefficients/distances

Bhattacharyya coefficient and non-metric distance:

$$C_(p,q) = \int \sqrt{p(x)q(x)}\mathrm{d}x, 0 < C(p,q) \leq 1, B(p,q) = -\ln C(p,q).$$

(coefficient is always strictly positive). Hellinger metric

$$H(p,q) = \sqrt{\frac{1}{2}\int (\sqrt{p(x)} - \sqrt{q(x)})^2 \mathrm{d}x},$$

such that $0 \leq H(p,q) \leq 1$.

$$\begin{aligned} H(p,q) &= \sqrt{\frac{1}{2}\left(\int p(x)\mathrm{d}x + \int q(x)\mathrm{d}x - 2\int \sqrt{p(x)}\sqrt{q(x)}\mathrm{d}x\right)} \\ &= \sqrt{1 - C(p,q)}. \end{aligned}$$

Chernoff coefficients/α-divergences

Skew Bhattacharrya divergences based on Chernoff α-coefficients.

$$\begin{aligned} B_\alpha(p,q) &= -\ln \int_x p^\alpha(x) q^{1-\alpha}(x) \mathrm{d}x = -\ln C_\alpha(p,q) \\ &= -\ln \int_x q(x) \left(\frac{p(x)}{q(x)}\right)^\alpha \mathrm{d}x \\ &= -\ln E_q[L^\alpha(x)] \end{aligned}$$

Amari α-divergence:

$$D_\alpha(p||q) = \begin{cases} \frac{4}{1-\alpha^2}\left(1 - \int p(x)^{\frac{1-\alpha}{2}} q(x)^{\frac{1+\alpha}{2}} \mathrm{d}x\right), & \alpha \neq \pm 1, \\ \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x = \mathrm{KL}(p,q), & \alpha = -1, \\ \int q(x) \log \frac{q(x)}{p(x)} \mathrm{d}x = \mathrm{KL}(q,p), & \alpha = 1, \end{cases}$$

$$D_\alpha(p||q) = D_{-\alpha}(q||p)$$

Remapping $\alpha' = \frac{1-\alpha}{2}$ ($\alpha = 1 - 2\alpha'$) to get Chernoff α'-divergences

Bhattacharyya/Chernoff of exponential families [10]

Equivalence with skew Burbea-Rao distances:

$$B_\alpha(p_F(x;\theta_p), p_F(x;\theta_q)) = \mathrm{BR}_F^{(\alpha)}(\theta_p, \theta_q), \quad (7)$$

$$= \alpha F(\theta_p) + (1-\alpha)F(\theta_q) - F(\alpha\theta_p + (1-\alpha)\theta_q)$$

Bhat. divergence on probability distributions amounts to compute a Jensen divergence on its parameters

Closed-form Bhattacharyya distances for exp. fam.

Generic formula that instantiates in those well-known formula in statistical pattern recognition.

Exp. fam.	$F(\theta)$ (up to a constant)	Bhattacharyya/Burbea-Rao $\mathrm{BR}_F(\lambda_p, \lambda_q) = \mathrm{BR}_F(\tau(\lambda_p), \tau(\lambda_q))$
Multinomial	$\log(1 + \sum_{i=1}^{d-1} \exp\theta_i)$	$-\ln \sum_{i=1}^{d} \sqrt{p_i q_i}$
Poisson	$\exp\theta$	$\frac{1}{2}(\sqrt{\mu_p} - \sqrt{\mu_q})^2$
Gaussian	$-\frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log(-\frac{\pi}{\theta_2})$	$\frac{1}{4}\frac{(\mu_p-\mu_q)^2}{\sigma_p^2+\sigma_q^2} + \frac{1}{2}\ln\frac{\sigma_p^2+\sigma_q^2}{2\sigma_p\sigma_q}$
Gaussian	$\frac{1}{4}\mathrm{tr}\Theta^{-1}\theta\theta^T - \frac{1}{2}\log\det\Theta$	$\frac{1}{8}(\mu_p - \mu_q)^T \left(\frac{\Sigma_p+\Sigma_q}{2}\right)^{-1}(\mu_p - \mu_q) + \frac{1}{2}\ln\frac{\det\frac{\Sigma_p+\Sigma_q}{2}}{\det\Sigma_p\det\Sigma_q}$

Wrapping up

- Besides Euclidean, log-Euclidean and Riemannian metric-based means, proposed divergence-based matrix centroids,
- Total Bregman divergences and robustness (conformal geometry),
- Riemannian minimax center,
- skew Burbea-Rao/Jensen divergences extending Bregman divergences,
- Bhattacharrya means of densities = Burbea-Rao means on (matrix) parameters

Which mean you do you mean or need?

Non-metric matrix manifolds with dually affine connections

In a nutshell:

- asymmetric (Bregman) non-metric divergence,
- Legendre transform, convex conjugates & dual divergences
- Dual $\theta-$ or η- or mixed coordinate systems
- dual closed-form affine geodesics (convenient computationally)
- Pythagorean theorem

Thank you.

www.informationgeometry.org

**"One geometry cannot be more true than another;
it can only be more convenient"**,
— Jules Henri Poincaré (1902)

Bibliographic references I

Marc Arnaudon and Frank Nielsen.

On approximating the Riemannian 1-center.

Comput. Geom., 46(1):93–104, 2013.

Rajendra Bhatia.

The Riemannian mean of positive matrices.

In Frank Nielsen and Rajendra Bhatia, editors, *Matrix Information Geometry*, pages 35–51, 2012.

Silvere Bonnabel and Rodolphe Sepulchre.

Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank.

SIAM J. Matrix Analysis Applications, 31(3):1055–1070, 2009.

Inderjit S. Dhillon and Joel A. Tropp.

Matrix nearness problems with bregman divergences.

SIAM J. Matrix Anal. Appl., 29(4):1120–1146, November 2007.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari.

Composite objective mirror descent.

In Adam Tauman Kalai and Mehryar Mohri, editors, *COLT*, pages 14–26. Omnipress, 2010.

C. Harris and M. Stephens.

A Combined Corner and Edge Detection.

In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.

Bibliographic references II

Meizhu Liu, Baba C. Vemuri, Shun-ichi Amari, and Frank Nielsen.

Shape retrieval using hierarchical total Bregman soft clustering.

Transactions on Pattern Analysis and Machine Intelligence, 34(12):2407–2419, 2012.

Meizhu Liu, Baba C. Vemuri, Shun ichi Amari, and Frank Nielsen.

Shape retrieval using hierarchical total bregman soft clustering.

IEEE Trans. Pattern Anal. Mach. Intell., 34(12):2407–2419, 2012.

Maher Moakher.

A differential geometric approach to the geometric mean of symmetric positive-definite matrices.

SIAM Journal on Matrix Analysis and Applications, 26(3):735–747, 2005.

Frank Nielsen and Sylvain Boltz.

The Burbea-Rao and Bhattacharyya centroids.

IEEE Transactions on Information Theory, 57(8):5455–5466, 2011.

Frank Nielsen, Meizhu Liu, Xiaojing Ye, and Baba C. Vemuri.

Jensen divergence based SPD matrix means and applications.

In *International Conference on Pattern Recognition (ICPR)*, 2012.

Frank Nielsen and Richard Nock.

Quantum Voronoi diagrams and Holevo channel capacity for 1-qubit quantum states.

In *IEEE International Symposium on Information Theory (ISIT)*, pages 96–100, 2008.

Bibliographic references III

Frank Nielsen and Richard Nock.

Sided and symmetrized Bregman centroids.

IEEE Trans. Inf. Theor., 55(6):2882–2904, June 2009.

Frank Nielsen and Richard Nock.

Entropies and cross-entropies of exponential families.

In *International Conference on Image Processing (ICIP)*, pages 3621–3624, 2010.

R. Nock, B. Magdalou, E. Briys, and F. Nielsen.

On tracking portfolios with certainty equivalents on a generalization of Markowitz model: the fool, the wise and the adaptive.

In Thorsten Joachims, editor, *International Conference on Machine Learning (ICML)*. Omnipress, 2011.

Richard Nock, Brice Magdalou, Eric Briys, and Frank Nielsen.

Mining matrix data with Bregman matrix divergences for portfolio selection.

In Frank Nielsen and Rajendra Bhatia, editors, *Matrix Information Geometry*, pages 373–402, 2012.

Masanori Ohya and Dénes Petz.

Quantum Entropy and Its Use.

1st ed. 1993. Corr 2nd printing, 2004.

Koji Tsuda, Gunnar Rätsch, and Manfred K. Warmuth.

Matrix exponentiated gradient updates for on-line learning and Bregman projection.

J. Mach. Learn. Res., 6:995–1018, December 2005.

Bibliographic references IV

Hisaharu Umegaki.

Conditional expectation in an operator algebra. IV. Entropy and information.

KodaiMathSemRep, 14(2):59, 1962.

Baba Vemuri, Meizhu Liu, Shun ichi Amari, and Frank Nielsen.

Total Bregman divergence and its applications to DTI analysis.

IEEE Transactions on Medical Imaging, 2011.

Zhizhou Wang and Baba C. Vemuri.

An affine invariant tensor dissimilarity measure and its applications to tensor-valued image segmentation.

In *CVPR (1)*, pages 228–233, 2004.