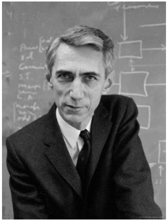
The dual geometry of Shannon information



Frank Nielsen[12]

@FrnkNlsn

[1]École Polytechnique [2]Sony CSL

Shannon centennial birth lecture

October 28th, 2016

Outline

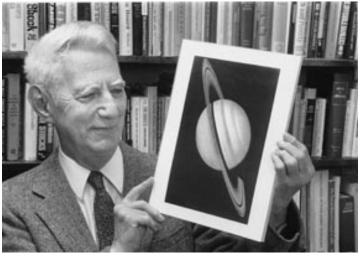
A storytelling...

- Getting started with the framework of information geometry:
  1. Shannon entropy and satellite concepts
  2. Invariance and information geometry
  3. Relative entropy minimization as information projections

- Recent work overview:
  4. Chernoff information and Voronoi information diagrams
  5. Some geometric clustering in information spaces
  6. Summary of statistical distances with their properties

- Closing: Information Theory onward

Chapter I.
Shannon entropy and satellite concepts



Shannon entropy (1940's): Big bang of IT!

- **Discrete entropy**: probability mass function (pmf) $p_i = P(X = x_i), x_i \in \mathcal{X}$ $(0 \log 0 = 0)$

$$H(X) = \sum_{i=1} p_i \log \frac{1}{p_i} = -\sum_{i=1} p_i \log p_i$$

- **Differential entropy**: probability density function (pdf) $X \sim p$ with support $\mathcal{X}$

$$h(X) = -\int_{\mathcal{X}} p(x) \log p(x) \mathrm{d}x$$

- Probability measure: random variable $X \sim P \ll \mu$

$$\boxed{H(X) = -\int_{\mathcal{X}} \log \frac{\mathrm{d}P}{\mathrm{d}\mu} \mathrm{d}P}$$

$$H(X) = -\int_{\mathcal{X}} p(x) \log p(x) \mathrm{d}\mu(x), \quad p = \frac{\mathrm{d}P}{\mathrm{d}\mu}$$

Lebesgue measure μ_L, counting measure μ_c,

Discrete vs differential Shannon entropy

Entropy: Measure the (expected) uncertainty of a random variable (rv)

$$H(X) = -\int_{\mathcal{X}} p(x) \log p(x) \mathrm{d}\mu(x) = \boxed{-E_X[\log X]}, \quad X \sim P$$

- Discrete entropy is **bounded**: $0 \leq H(X) \leq \log |\mathcal{X}|$ with support $\mathcal{X}$
- Differential entropy...
  - may be **negative**:

    $$H(X) = \frac{1}{2} \log(2\pi e \sigma^2), \quad X \sim N(\mu, \sigma)$$

    for Gaussians
  - may be **infinite** when integral diverges:

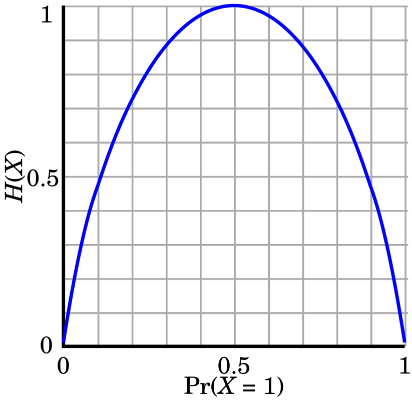
    $$H(X) = \infty$$

    $$X \sim p(x) = \frac{\log(2)}{x \log^2 x} \text{ for } x > 2, \text{ with support } \mathcal{X} = (2, \infty)$$

Key property: Shannon entropy is concave...

Graph plot of Shannon binary entropy (H of Bernoulli trial):
$X \sim \mathrm{Bernoulli}(p)$ with $p = \Pr(X = 1)$

$$H(X) = -(p \log p + (1-p) \log(1-p))$$



... and **Shannon information** $-H(X)$ (neg-entropy) is **convex**

Maximum entropy principle (Jaynes [12], 1957): Exponential families (Gibbs distribution)

- A finite set of D *moment (expectation) constraints* t_i:

$$\boxed{E_{p(x)}[t_i(X)] = \eta_i}$$

  for $i \in [D] = \{1, \ldots, D\}$

- Solution (Lagrangian multipliers): $=$
  Exponential Family [34]

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

  where $\langle a, b\rangle = a^\top b$: dot/scalar/inner product.

- **MaxEnt**: $\max_\theta H(p(x;\theta))$ such that $E_{p(x;\theta)}[t(X)] = \eta$, $t(x) = (t_1(x), \ldots, t_D(x))$ and $\eta = (\eta_1, \ldots, \eta_D)$
- Consider a parametric family $\{p(x;\theta)\}_{\theta\in\Theta}$, $\theta \in \mathbb{R}^D$, D: order

Exponential families (EFs) [34]

- Log-normalizer (cumulant, partition function, free energy):

$$\boxed{F(\theta) = \log\left(\int \exp(\langle\theta, t(x)\rangle)\right) \mathrm{d}\nu(x)} \leftarrow \int p(x;\theta)\mathrm{d}\nu(x) = 1$$

  Here, F **strictly convex**, here C^∞. $p(x;\theta) = e^{\langle\theta,t(x)\rangle - F(\theta)}$

- Natural parameter space:

$$\Theta = \{\theta \in \mathbb{R}^D \ : \ F(\theta) < \infty\}$$

- EFs have **all finite order moments** expressed using the Moment Generating Function (MGF):

$$\boxed{M(u) = E[\exp(\langle u, X\rangle)] = \exp(F(\theta + u) - F(\theta))}$$

$$\text{Geometric moments:} \quad E[t(X)^l] = M^{(l)}(0) \quad \text{for order } D = 1$$

$$E[t(X)] = \nabla F(\theta) = \eta, \quad V[t(X)] = \nabla^2 F(\theta) \succ 0$$

Example: MaxEnt distribution with fixed mean and fixed variance = Gaussian family

- $\max_p H(p(x)) = \max_\theta H(p(x;\theta))$ such that:

$$\begin{aligned} E_{p(x;\theta)}[X] &= \eta_1 (= \mu), \\ E_{p(x;\theta)}[X^2] &= \eta_2 (= \mu^2 + \sigma^2) \end{aligned}$$

  Indeed, $V_{p(x;\theta)}[X] = E[(X-\mu)^2] = E[X^2] - \mu^2 = \sigma^2$

- **Gaussian distribution** is maxent distribution:

$$\boxed{p(x;\theta(\mu,\sigma)) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)} = e^{\langle\theta,t(x)\rangle - F(\theta)}$$

  - sufficient statistic vector: $t(x) = (x, x^2)$
  - natural parameter vector: $\theta = (\theta_1, \theta_2) = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$
  - log-normalizer: $F(\theta) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log\left(-\frac{\pi}{\theta_2}\right)$

- By construction,
  $E[t(x) = (x, x^2)] = \nabla F(\theta) = \eta = (\mu, \mu^2 + \sigma^2)$

Entropy of an EF and convex conjugates

$$X \sim p(x;\theta) = \exp\left(\langle \theta, t(x)\rangle - F(\theta)\right), \quad E_{p(x;\theta)}[t(X)] = \eta$$

- Entropy of an EF:

$$H(X) = -\int p(x;\theta)\log p(x;\theta) = F(\theta) - \langle \theta, \eta\rangle$$

- Legendre **convex conjugates** [20]: $F^*(\eta) = -F(\theta) + \langle \theta, \eta\rangle$
- $H(X) = F(\theta) - \langle \theta, \eta\rangle = -F^*(\eta) < \infty$ (always finite here!)
- **A member of an exponential family can be canonically parameterized either by using its natural parameter $\theta = \nabla F^*(\eta)$ or by using its expectation parameter $\eta = \nabla F(\theta)$**, see [34]
- Converting η-to-θ parameters can be seen as a MaxEnt optimization problem. *Rarely in closed-form!*

MaxEnt and Kullback-Leibler divergence

- **Statistical distance**: Kullback-Leibler divergence
  Aka. relative entropy, $P, Q \ll \mu$, $p = \frac{\mathrm{d}P}{\mathrm{d}\mu}$, $q = \frac{\mathrm{d}Q}{\mathrm{d}\mu}$

$$\boxed{\mathrm{KL}(P : Q) = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}\mu(x)}$$

- KL is *not* a metric distance: asymmetric and does not satisfy triangle inequality
- $\mathrm{KL}(P : Q) \geq 0$ (Gibb's inequality) and KL **may be infinite**:

$p(x) = \frac{1}{\pi(1+x^2)}$ = Cauchy distribution
$q(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$ = standard normal distribution

$\mathrm{KL}(p : q) = +\infty$ diverges while $\mathrm{KL}(q : p) < \infty$ converges.

MaxEnt as a convex minimization program

- Maximizing concave entropy H under linear moment constraints
  $\equiv$ **minimizing convex information**

- MaxEnt $\equiv$ convex minimization with linear constraints (the $t_i(x_j)$ are prescribed constants)

$$\min_{p \in \Delta^{D+1}} \sum_j p_j \log p_j \quad \text{(CVX)}$$

constraints:

$$\sum_j p_j t_i(x_j) = \eta_j, \quad \forall i \in [D]$$

$$p_j \geq 0, \quad \forall i \in [|\mathcal{X}|]$$

$$\sum_j p_j = 1$$

Δ^{D+1}: D-dimensional probability simplex, embedded in $\mathbb{R}_+^{D+1}$

MaxEnt with prior and general canonical EF

MaxEnt $H(P) \equiv$ **left-sided** $\min_P \mathrm{KL}(\boxed{\text{P}} : U)$ wrt U

U: uniform distribution $H(U) = \log|\mathcal{X}|$.

$\max_P H(P) = \log|\mathcal{X}| - \min_P \mathrm{KL}(P : U)$

with KL amounting to "cross-entropy minus entropy":

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

- **Generalized MaxEnt problem**: Minimize KL distance to **prior distribution** h under constraints (MaxEnt is recovered when $h = U$, uniform distribution)

$$\min_p \mathrm{KL}(p : h)$$

constraints:

$$\sum_j p_j t_i(x_j) = \eta_j, \quad \forall i \in [D]$$

$$p_j \geq 0, \quad \forall i \in [|\mathcal{X}|], \quad \sum_j p_j = 1$$

Solution of MaxEnt with prior distribution

- General canonical form of exponential families (using Lagrange multipliers for constrained optimization)

$$\boxed{p(x;\theta) = \exp(\langle \theta, t(x)\rangle - F(\theta))h(x)}$$

- Since $h(x) > 0$, let $h(x) = \exp(k(x))$ for $k(x) = \log h(x)$
- Exponential families are **log-concave** (F is convex):

$$l(x;\theta) = \log p(x;\theta) = \langle \theta, t(x)\rangle - F(\theta) + k(x)$$

- Entropy of general EF [37]:

$$X \sim p(x;\theta), \quad \boxed{H(X) = -F^*(\eta) - E[k(x)]}$$

- many common distributions [34] $p(x;\lambda)$ are EFs with $\theta = \theta(\lambda)$ and **carrier distribution** $\mathrm{d}\nu(x) = e^{k(x)}\mathrm{d}\mu(x)$ (eg., Rayleigh)

Maximum Likelihood Estimator (MLE) for EFs

- Given observations $\mathcal{S} = \{s_1, \ldots, s_m\} \sim_{\mathrm{iid}} p(x; \theta_0)$, MLE:

$$\begin{aligned} \hat{\theta}_m &= \mathrm{argmax}_\theta L(\theta; \mathcal{S}) = \prod_i p(s_i; \theta) \\ &\equiv \mathrm{argmax}_\theta l(\theta; \mathcal{S}) = \frac{1}{m} \sum_i l(s_i; \theta) \end{aligned}$$

- "Normal equation" of MLE [34]:

$$\boxed{\hat{\eta}_m = \nabla F(\hat{\theta}_m) = \frac{1}{m} \sum_{i=1}^m t(s_i)}$$

- MLE problem is **linear in η** but **convex in θ**: $\min_\theta F(\theta) - \left\langle \frac{1}{m} \sum_i t(s_i), \theta \right\rangle$
- MLE is **consistent**: $\lim_{m \to \infty} \hat{\theta}_m = \theta_0$
- Average log-likelihood [23]: $l(\hat{\theta}_m; \mathcal{S}) = F^*(\hat{\eta}_m) + \frac{1}{m} \sum_i k(s_i)$

MLE as a right-sided KL minimization problem

- Empirical distribution: $p_e(x) = \frac{1}{m}\sum_{i=1}^m \delta_{s_i}(x)$.
  Powerful modeling: data and models <u>coexist</u> in the space of distributions
  $p_e \ll p(x;\theta)$ is absolutely continuous with respect to $p(x;\theta)$

$$
\begin{aligned}
\min \quad & \mathrm{KL}(p_e(x) : \boxed{p_\theta(x)}) \\
= \quad & \int p_e(x)\log p_e(x)\mathrm{d}x - \int p_e(x)\log p_\theta(x)\mathrm{d}x \\
= \quad & \min -H(p_e) - \underbrace{E_{p_e}[\log p_\theta(x)]} \\
& \equiv \max \frac{1}{n}\sum \delta(x-x_i)\log p_\theta(x) \\
= \quad & \max \frac{1}{n}\sum_i \log p_\theta(x_i) = \mathrm{MLE}
\end{aligned}
$$

- Since $\mathrm{KL}(p_e(x) : p_\theta(x)) = H^\times(p_e(x) : p_\theta(x)) - H(p_e(x))$, min $\mathrm{KL}(p_e(x) : p_\theta(x))$ amounts to **minimize the cross-entropy**

Fisher Information Matrix (FIM) and CRLB [24]

Notation: $\partial_i l(x;\theta) = \frac{\partial}{\partial \theta_i} l(x;\theta)$

- Fisher Information Matrix (FIM) :

$$\boxed{I = [I_{i,j}]_{ij}\,, I_{i,j}(\theta) = \mathrm{E}_\theta[\partial_i l(x;\theta)\partial_j l(x;\theta)]}, \quad I(\theta) \succeq 0$$

- Cramér-Rao/Fréchet lower bound (CRLB) for an *unbiased estimator* $\hat{\theta}_m$ with θ_0 optimal parameter (hidden by nature):

$$\boxed{V[\hat{\theta}_m] \succeq I^{-1}(\theta_0)}, \quad V[\hat{\theta}_m] - I^{-1}(\theta_0) \text{ is PSD}$$

- efficiency: unbiased estimator matching the CR lower bound
- asymptotic normality of MLE $\hat{\theta}$ (on random vectors):

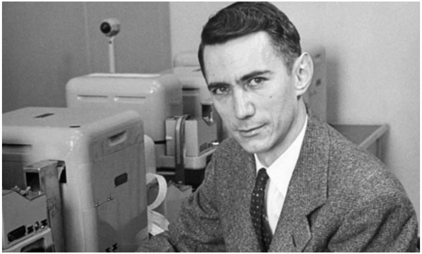
$$\boxed{\hat{\theta}_m \sim N\left(\theta_0, \frac{1}{m} I^{-1}(\theta_0)\right)}$$

Recap of Chapter I: Shannon cosmos

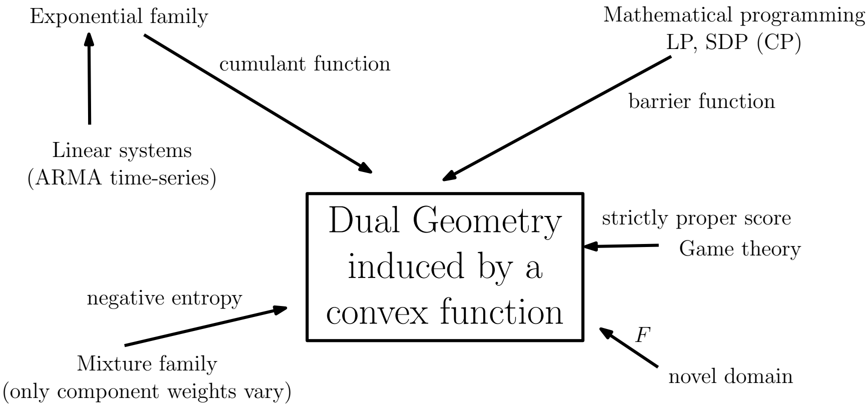
Shannon's Big Bang: The story so far has begun with ...

- Shannon entropy H is **concave**
- **MaxEnt** yields **exponential families**
- Entropy of EFs P can either be expressed using θ natural or η expectation parameterizations of EFs.
  Converting $\eta \rightarrow \theta$ by MaxEnt optimization
- Shannon information of EF $-H(P) = F^*(\eta)$ is **convex**
- MaxEnt amounts to **min KL on left argument**
  (right argument is prescribed prior distribution)
- MLE for EFs amounts to **min KL on right argument**
  (left argument is prescribed empirical distribution)
- Min variance of estimator is lower bounded by inverse of Fisher Information Matrix (FIM): Cramér-Rao lower bound
- MLE is consistent, Fisher efficient, with asymptotic normality

Chapter II.
Invariance and geometry



Differential geometry from a convex function



$\boxed{\text{Shannon information } F = -H \text{ is convex!}}$

Three remarkable properties of the KL divergence

- KL is a **separable divergence**:
  $\mathrm{KL}(P, Q) = \int_{\mathcal{X}} \mathrm{kl}(p(x) : q(x)) \mathrm{d}\mu(x)$, where
  $\mathrm{kl}(a : b) = a \log \frac{a}{b}$ is a **1D function** on scalars.
  Squared Euclidean distance is separable but not the Euclidean distance.
- KL satisfies the **information monotonicity**:

$$\boxed{\mathrm{KL}(P : Q) \geq \mathrm{KL}(P_{\mathcal{Y}} : Q_{\mathcal{Y}})}$$

  where $X_{\mathcal{Y}}$ is a **coarse-grained quantization** of X ($\mathcal{Y} = \uplus_j \mathcal{I}_j$: a partition of $\mathcal{X}$). $p_{\mathcal{Y}}(y) = \int_{\mathcal{I}_j} p(x) \mathrm{d}\mu(x)$ for $y \in \mathcal{I}_j$.
- KL is locally $\approx\propto$ **quadratic FIM form** for arbitrary smooth family distributions P, Q (not necessarily EFs):

$$\boxed{\mathrm{KL}(P_{\theta_1} : P_{\theta_2}) = \frac{1}{2} M^2_{I_{\theta_1}}(\theta_1, \theta_2) + o(\|\theta_1 - \theta_2\|^2)}$$

  $M_G(p, q) = \sqrt{(p-q)^\top G (p-q)}$ is a **Mahalanobis distance** for $G \succ 0$

Those 3 properties are satisfied by all f-divergences [41]

$$\boxed{I_f(X_1 : X_2) = \int x_1(x) f\left(\frac{x_2(x)}{x_1(x)}\right) \mathrm{d}\nu(x) \geq f(1) = 0}$$

where f is a **convex function**

$$f : (0, \infty) \subseteq \mathrm{dom}(f) \mapsto [0, \infty]$$

such that $f(1) = 0$.

Jensen inequality: $I_f(X_1 : X_2) \geq f(\int x_2(x) \mathrm{d}\nu(x)) = f(1) = 0$.

May consider $f'(1) = 0$ and fix the scale of divergence ($I_{\lambda f} = \lambda I_f$) by setting $f''(1) = 1$.

f-divergences can always be **symmetrized**:

$$S_f(X_1 : X_2) = I_f(X_1 : X_2) + I_{f^\diamond}(X_1 : X_2)$$

with $f^\diamond(u) = u f(1/u)$, and $I_{f^\diamond}(X_1 : X_2) = I_f(X_2 : X_1)$, $f^\diamond$ convex.

Some common examples of f-divergences [41]

Kullback-Leibler belongs to the broad class of f-divergences

Name of the f-divergence	Formula $I_f(P:Q)$	Generator $f(u)$ with $f(1)=0$
Total variation (metric)	$\frac{1}{2}\int \lvert p(x)-q(x)\rvert \mathrm{d}\nu(x)$	$\frac{1}{2}\lvert u-1\rvert$
Squared Hellinger	$\int(\sqrt{p(x)}-\sqrt{q(x)})^2\mathrm{d}\nu(x)$	$(\sqrt{u}-1)^2$
Pearson χ^2_P	$\int \frac{(q(x)-p(x))^2}{p(x)}\mathrm{d}\nu(x)$	$(u-1)^2$
Neyman χ^2_N	$\int \frac{(p(x)-q(x))^2}{q(x)}\mathrm{d}\nu(x)$	$\frac{(1-u)^2}{u}$
Pearson-Vajda χ^k_P	$\int \frac{(q(x)-\lambda p(x))^k}{p^{k-1}(x)}\mathrm{d}\nu(x)$	$(u-1)^k$
Pearson-Vajda $\lvert\chi\rvert^k_P$	$\int \frac{\lvert q(x)-\lambda p(x)\rvert^k}{p^{k-1}(x)}\mathrm{d}\nu(x)$	$\lvert u-1\rvert^k$
Kullback-Leibler	$\int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}\nu(x)$	$-\log u$
reverse Kullback-Leibler	$\int q(x)\log\frac{q(x)}{p(x)}\mathrm{d}\nu(x)$	$u\log u$
Triangular	$\frac{1}{2}\int \frac{(q(x)-p(x))^2}{p(x)+q(x)}\mathrm{d}\nu(x)$	$\frac{(u-1)^2}{2(1+u)}$
Squared triangular	$\int \frac{(p(x)-q(x))^2}{p(x)+q(x)}\mathrm{d}\nu(x)$	$\frac{(u-1)^2}{2(1+u)}$
Squared perimeter	$\int \sqrt{p^2(x)+q^2(x)}\mathrm{d}\nu(x)-\sqrt{2}$	$\sqrt{1+u^2}-\frac{1+u}{\sqrt{2}}$
α-divergence	$\frac{4}{1-\alpha^2}(1-\int p^{\frac{1-\alpha}{2}}(x)q^{1+\alpha}(x)\mathrm{d}\nu(x))$	$\frac{4}{1-\alpha^2}(1-u^{\frac{1+\alpha}{2}})$
Jensen-Shannon	$\frac{1}{2}\int(p(x)\log\frac{2p(x)}{p(x)+q(x)}+q(x)\log\frac{2q(x)}{p(x)+q(x)})\mathrm{d}\nu(x)$	$-(u+1)\log\frac{1+u}{2}+u\log u$

Invariance of f-divergences

- Diffeomorphism $h : \mathcal{X} \rightarrow \mathcal{Y}$, $y = h(x)$

$$\boxed{p_Y(y) = |J|^{-1} p_X(h^{-1}(x))} \quad \leftarrow \text{rewrite density}$$

with J the Jacobian matrix $\left(\frac{\partial y_i}{\partial x_j}\right)_{i,j}$

- f-divergences are **invariant** under differentiable and invertible h.

$$\boxed{D_f(x : x') = D_f(y : y')}$$

$\leftarrow$ More generally, technically invariant to "sufficiency of stochastic kernels" [50, 14].

- Conversely, integration measures invariant to diffeomorphisms are f-divergences [52].
(Exhaustivity property for deterministic transformation)

Covariance of Fisher Information Matrix

- Let $\theta = \theta(\eta)$ and $\eta = \eta(\theta)$ be two 1-to-1 parameterizations. From Legendre transformation: $\eta = \nabla F(\theta)$ and $\theta = \nabla F^*(\theta)$
- $J = [J_{i,j}]_{i,j}$: Jacobian matrix $J_{i,j} = \frac{\partial \theta_i}{\partial \eta_j}$.

$$\boxed{I_\eta(\eta) = J^\top \times I_\theta(\theta(\eta)) \times J}$$

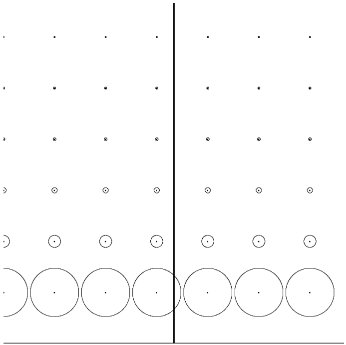
Fisher information matrix <u>depends</u> on the parameterization of the parameter space (covariant), but not the infinitesimal length elements $\mathrm{d}s^2(p) = \langle \cdot, \cdot \rangle_{I(p)}$: $\mathrm{d}s_\theta(\theta_p) = \mathrm{d}s_\eta(\eta_p)$
$\rightarrow$ Fisher-Riemannian geometry (Hotelling 1930, Rao 1945)

In 2D, we can always **diagonalize the FIM** [58] by (θ, η) **mixed reparameterization**. In general, cannot find a change of coordinates to have diagonal FIM.

Riemannian statistical manifolds with g =FIM

For univariate normal distributions (or location-scale families):
$\equiv$ Hyperbolic geometry [38]

$$\cosh \rho(p_1, p_2) = 1 + \frac{\|p_1 - p_2\|^2}{2 y_1 y_2}, \quad g(p) = \begin{bmatrix} \frac{1}{y^2} & 0 \\ 0 & \frac{1}{y^2} \end{bmatrix} = \frac{1}{y^2} I$$



conformal (upper space model): $g(p) = \frac{1}{y^2} I$

Statistical manifolds: Differential Geometry (DG)

- Geometric structure $\mathcal{M}$ of parametric family $\{p_\theta\}_{\theta\in\Theta}$ equipped with **metric tensor** $g = I$, the FIM:
  Scalar product at each **tangent plane** T_p:

$$\langle u, v\rangle_p = u^\top I(\theta(p))v$$

$$u \perp_p v \Leftrightarrow \langle u, v\rangle_p = 0 \quad \text{(Fisher orthogonality)}$$

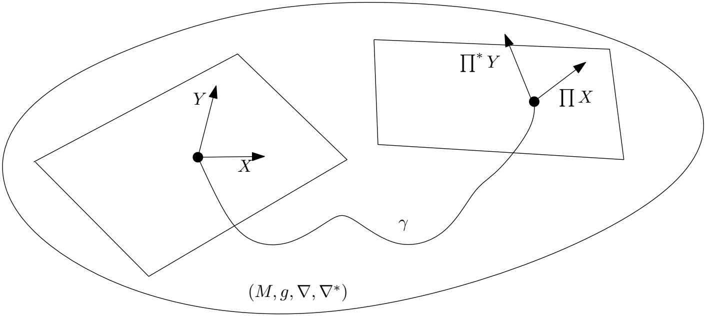
- **Riemannian geometry**: geodesics are shortest paths that parallel transport vectors using the Levi-Cevita metric connection ∇^0 induced by g.
  The Riemannian distance is a metric distance.
- **Affine differential geometry**: dual geodesics preserving dual parallel transports.
  Distance is a non-metric divergence
  (C^3 differentiable dissimilarity measure)

Affine Diff. Geometry: Dually affine connections

- Two coupled affine connections $\prod$ and $\prod^*$ (and covariant derivatives ∇ and ∇^*)
- Property of inner product (keeps angles by parallel transport):

$$\boxed{\langle X, Y\rangle_g = \langle \prod X, \overset{*}{\prod} Y\rangle_g}$$

- Riemannian geometry: $\prod = \prod^* = \prod_0$



$$\langle X, Y\rangle_g = \langle \prod X, \prod^* Y\rangle_g$$

Dual vector basis and covariance/contravariance

- Geometric objects (points, vectors, tensors) are parameterized by **coordinates** that "arithmetize space".
- Tangent planes T_p are **vector spaces** equipped with **local basis**
- Vector $v = \sum_i v^i e_i$ is expressed in a given basis $[e] = (e_1, \ldots, e_D)$ with coordinates $(v^1, \ldots, v^D)$. The coordinates of e_i are $e_i[e] = (0, \ldots, 0, 1, 0, \ldots, 0)$.
- Under **change of basis**, tensor components change but geometric tensor objects are invariant = "facts of universe"
- Aim at writing $v^i = \langle v, e_i \rangle$ but this works only for orthonormal coordinate systems: $\langle e_i, e_j \rangle = \delta_{ij}$.
- Fortunately, there always exist a **dual basis** with *reciprocal basis vectors* e^j such that $\langle e_i, e^j \rangle = \delta_i^j$ ($\delta_i^j = 1$ iff $i = j$, and 0 otherwise) so that:
$$v^i = \langle v, e^i \rangle$$
- A vector can be manipulated either using its contravariant components v^i or using its dual covariant components v_i

Dually flat manifolds from a convex function F

Canonical geometry induced by strictly convex and differentiable convex function F.

- Potential functions: F and Legendre convex conjugate $G = F^*$
- Dual affine coordinate systems: $\theta = \nabla F^*(\eta)$ and $\eta = \nabla F(\theta)$
- **Metric tensor** g: written equivalently using the two coordinate systems:

$$\boxed{g_{ij}(\theta) = \frac{\partial^2}{\partial\theta^i \partial\theta^j} F(\theta),\ g^{ij}(\eta) = \frac{\partial^2}{\partial\eta_i \partial\eta_j} G(\eta)},\ \nabla^2 F(\theta) \nabla^2 G(\eta) = I$$

- Divergence from Young's inequality of convex conjugates:

$$\boxed{D(P : Q) = F(\theta(P)) + F^*(\eta(Q)) - \langle \theta(P), \eta(Q) \rangle}$$

This **canonical divergence** is a Bregman divergence when we rewrite it using a single parameterization

Recap of Chapter 2: Invariance and geometry

- f-**divergence** are **separable divergences** that satisfy **information monotonicity** and locally proportional to **squared Fisher Mahalanobis distances**

- A smooth dually flat manifold $\mathcal{M} = (M, g, \nabla, \nabla^*)$ can be built from any strictly convex function F
  Parameterizations: $G = \nabla^2 F(\theta)$ or $G^* = \nabla^2 F^*(\eta)$ with $GG^* = I$
  Metric tensor g:
  contravariant components g_{ij} and covariant components g^{ij}

- This explains the dual structure of "exponential family manifold" or "mixture family manifold" met in information geometry, among others
- Euclidean geometry is self-dual for $F(x) = F^*(x) = \frac{1}{2}\langle x, x\rangle$. The geometry of multivariate normal families with identical covariance matrix.

Chapter III. Information Projections

Dually affine connections: e/m-connections and e/m-flats

- Exponential e-geodesics and mixture m-geodesics for probability densities:

$$
\begin{aligned}
\gamma_m(p,q,\alpha) &: \quad r(x,\alpha) = \alpha p(x) + (1-\alpha) q(x) \\
\gamma_e(p,q,\alpha) &: \quad \log r(x,\alpha) = \alpha p(x) + (1-\alpha) q(x) - F(t)
\end{aligned}
$$

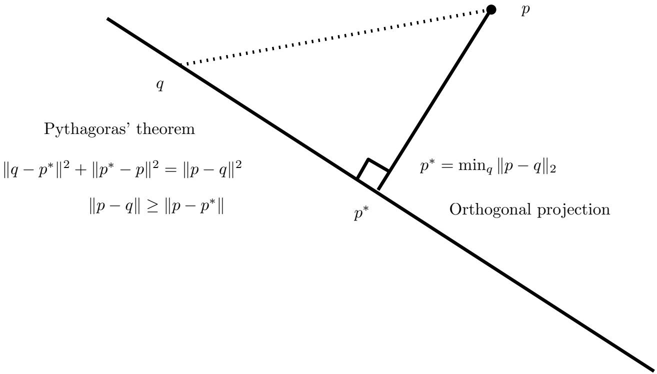
- In IG, e-connection corresponds to $\alpha = +1$-connection (θ), and m-connection corresponds to $\alpha = -1$-connection (η)

$$\boxed{\nabla^{(e)} = \nabla^{(1)}, \quad \nabla^{(m)} = \nabla^{(-1)}} \quad \alpha\text{-connections}$$

- Geodesics are **straight lines** in either θ or η parameterization
- e-flat is an affine subspace in θ-coordinate system
  m-flat is an affine subspace in η-coordinate system

Projection, orthogonality and Pythagoras' theorem

Recalling Euclidean geometry...



Information projections: *e*-projection and *m*-projection

- *e*-projection q_e^* is **unique** if $M \subseteq S$ is *m*-flat and minimizes the *m*-divergence $\mathrm{KL}(\boxed{q} : p)$ (left-sided argument):

$$e\text{-projection:} \quad \boxed{q_e^* = \arg\min_q \mathrm{KL}(\boxed{q} : p)}$$

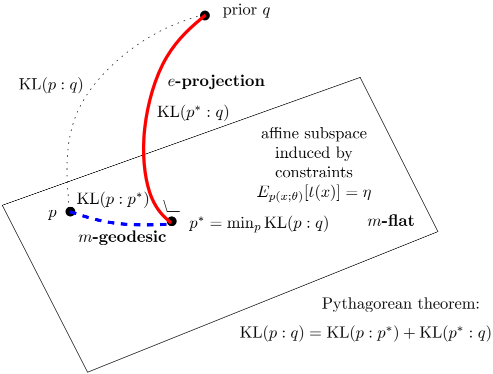
- *m*-projection q_m^* is **unique** if $M \subseteq S$ is *e*-flat and minimizes the *e*-divergence $\mathrm{KL}(p : \boxed{q})$ (right-sided argument):

$$m\text{-projection:} \quad \boxed{q_m^* = \arg\min_q \mathrm{KL}(p : \boxed{q})}$$

I-projection, rI-projection, KL-projection, etc.

MaxEnt with prior $q(x)$ as an information projection

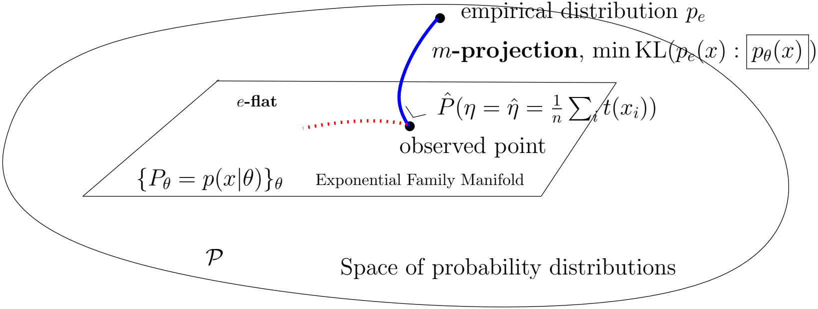
MaxEnt linear constraints define a *m*-flat



Pythagoras' theorem, $\gamma_m(p, p^*) \perp_{\mathrm{FIM}} \gamma_e(p^*, q)$ **(Fisher orthogonality)**

MLE $\equiv$ min KL: Information projection

Exponential Family Manifold (EFM) is *e*-flat



Observed point & sufficiency

- Remember MLE of EF is given in closed-form in η-coordinate system:

$$\boxed{\hat{\eta}_m = \frac{1}{m}\sum_{i=1}^{m} t(s_i) = \nabla F(\hat{\theta}_m)}$$

  ... but to get θ, we need to compute $\nabla F^{-1} = \nabla F^*$, or solve MaxEnt problem.
- The point with η-coordinate $\frac{1}{m}\sum_{i=1}^{m} t(s_i)$ is called the observed point in information geometry.
- $t(x)$ is called the sufficient statistics:

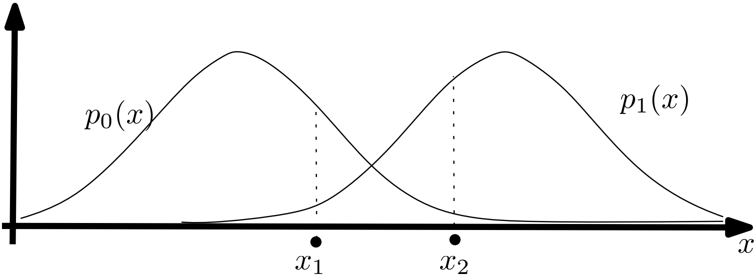
$$\Pr(x|t,\theta) = \Pr(x|t)$$

  All information about θ for inference is contained in t
  Exponential families have finite sufficient statistics
  = lossless statistical information compression

Chapter IV.
Chernoff information and Voronoi diagrams

The Hypothesis Testing (HT) problem

Given two distributions hypothesis P_0 and P_1, classify observation x (=decide) either as sampled from P_0 or from P_1?



P_0: signal, P_1: noise...

The Multiple Hypothesis Testing (MHT) problem

Given a random variable X with n hypothesis $H_1 : X \sim P_1$, ..., $H_n : X \sim P_n$, decide for a Identically and Independently Distributed (IID) sample $x_1, ..., x_m \sim X$ **which hypothesis holds true**?

$$P^m_{\mathrm{correct}} = 1 - P^m_{\mathrm{error}} = 1 - P^m_e$$

Seek the **asymptotic regime exponent** α:

$$\boxed{\alpha = -\frac{1}{m} \log P^m_e}, \quad m \to \infty$$

Bayesian hypothesis testing (preliminaries)

- **prior class probabilities**: $w_i = \Pr(X \sim P_i) > 0$ (with $\sum_{i=1}^n w_i = 1$)
- **conditional class probabilities**: $\Pr(X = x | X \sim P_i)$.
- **Total probability** (mixture of classes):

$$
\begin{aligned}
\Pr(X = x) &= \sum_{i=1}^{n} \Pr(X \sim P_i) \Pr(X = x | X \sim P_i) \\
&= \sum_{i=1}^{n} w_i \Pr(X | P_i)
\end{aligned}
$$

- Let $c_{i,j}$ = cost of deciding H_i when in fact H_j is true. Matrix $[c_{ij}]$= **cost *design* matrix**
- Let $p_{i,j}(u)$ = probability of making this decision using **rule** u.

Bayesian detector & Probability of Error

Minimize the *expected cost* for a rule r.
Special case: **Probability of error** P_e obtained for $c_{i,i} = 0$ (correct classification) and $c_{i,j} = 1$ for $i \neq j$ (misclassification):

$$\boxed{P_e = E_X\left[\sum_i \left(w_i \sum_{j\neq i} p_{i,j}(r(x))\right)\right]}$$

The **maximum *a posteriori* probability** (MAP) rule considers classifying x:

$$\boxed{\mathrm{MAP}(x) = \mathrm{argmax}_{i\in\{1,\ldots,n\}}\ w_i p_i(x)}$$

where $p_i(x) = \Pr(X = x | X \sim P_i)$ are the conditional probabilities.
$\rightarrow$ **MAP Bayesian detector minimizes P_e over all rules [13]**

Probability of error P_e and divergences

Without loss of generality, consider **equal priors** ($w_1 = w_2 = \frac{1}{2}$):

$$\boxed{P_e = \int_{x\in\mathcal{X}} p(x) \min(\Pr(H_1|x), \Pr(H_2|x)) \mathrm{d}\nu(x)}$$

($P_e > 0$ as soon as $\mathrm{supp}(p_1) \cap \mathrm{supp}(p_2) \neq \emptyset$)

From Bayes' rule $\Pr(H_i|X = x) = \frac{\Pr(H_i)\Pr(X=x|H_i)}{\Pr(X=x)} = w_i p_i(x)/p(x)$

$$\boxed{P_e = \frac{1}{2} \int_{x\in\mathcal{X}} \min(p_1(x), p_2(x)) \mathrm{d}\nu(x)}$$

Aka. "histogram intersection distance".

Bounding the Probability of error P_e

Trick: $\boxed{\min(a,b) \leq \min_{\alpha\in(0,1)} a^\alpha b^{1-\alpha}}$ for $a, b > 0$, *upper bound* P_e:

$$\begin{aligned} P_e &= \frac{1}{2}\int_{x\in\mathcal{X}} \min(p_1(x), p_2(x))\mathrm{d}\nu(x) \\ &\leq \frac{1}{2}\min_{\alpha\in(0,1)}\int_{x\in\mathcal{X}} p_1^\alpha(x)p_2^{1-\alpha}(x)\mathrm{d}\nu(x). \end{aligned}$$

Chernoff information:

$$C(P_1, P_2) = -\log \min_{\alpha\in(0,1)}\int_{x\in\mathcal{X}} p_1^\alpha(x)p_2^{1-\alpha}(x)\mathrm{d}\nu(x) \geq 0,$$

Best error exponent α^* [11] bounds proba. of error:

$$P_e \leq w_1^{\alpha^*} w_2^{1-\alpha^*} e^{-C(P_1,P_2)} \leq e^{-C(P_1,P_2)}$$

Bounding technique can be extended using any **quasi-arithmetic means** [28, 22] (f-means or Kolmogorov-Nagumo means)

MAP decision rule for EFs and additive Bregman Voronoi diagrams

$$\boxed{\mathrm{KL}(p_{\theta_1} : p_{\theta_2}) = B(\theta_2 : \theta_1) = A(\theta_2 : \eta_1) = A^*(\eta_1 : \theta_2) = B^*(\eta_1 : \eta_2)}$$

Canonical divergence (mixed primal/dual coordinates):

$$A(\theta_2 : \eta_1) = F(\theta_2) + F^*(\eta_1) - \theta_2^\top \eta_1 \geq 0$$

Bregman divergence (uni-coordinates, primal or dual):

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

Duality Bregman divergences with exponential families:

$$\boxed{\log p_{\theta_i}(x) = -B^*(t(x) : \eta_i) + F^*(t(x)) + k(x), \quad \eta_i = \nabla F(\theta_i) = \eta(P_{\theta_i})}$$

Optimal **MAP decision rule**: Additive Bregman Voronoi diagram

$$\begin{aligned} \mathrm{MAP}(x) &= \mathrm{argmax}_{i\in\{1,\ldots,n\}} w_i p_i(x) \\ &= \arg\min_{i\in\{1,\ldots,n\}} B^*(t(x) : \eta_i) - \log w_i \end{aligned}$$

→ *nearest neighbor classifier [3, 23, 47, 51]*

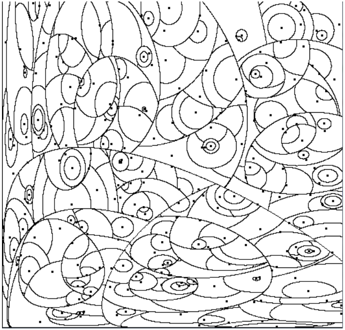
MAP of EFs & nearest neighbor classifier

Bregman Voronoi diagrams (with additive weights) are affine diagrams [3].

$$\boxed{\arg\min_{i\in\{1,\ldots,n\}} B^*(t(x):\eta_i) - \log w_i}$$

Need to answer fast Bregman proximity queries:

- point location in arrangement [4] (small dims),
- Divergence-based search trees [51],
- GPU brute force [8].



Geometry of the best error exponent: binary hypothesis

On the exponential family manifold, Chernoff α-coefficient [5]:

$$c_\alpha(P_{\theta_1} : P_{\theta_2}) = \int p_{\theta_1}^{\alpha}(x) p_{\theta_2}^{1-\alpha}(x) \mathrm{d}\mu(x) = \exp(-J_F^{(\alpha)}(\theta_1 : \theta_2)),$$

Skew Jensen divergence [32] on the natural parameters:

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

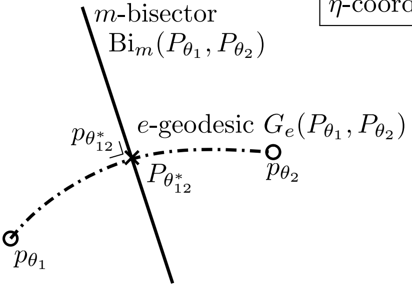
Theorem: Chernoff information = Bregman divergence for exponential families at the optimal exponent value:

$$\boxed{C(P_{\theta_1} : P_{\theta_2}) = B(\theta_1 : \theta_{12}^{(\alpha^*)}) = B(\theta_2 : \theta_{12}^{(\alpha^*)})}$$

Geometry of the best error exponent: binary hypothesis on the exponential family manifold

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

$\boxed{\eta\text{-coordinate system}}$



$$C(\theta_1 : \theta_2) = B(\theta_1 : \theta_{12}^*)$$

Synthetic information geometry ("Hellinger arc"):
Exact characterization but not necessarily closed-form formula

Geometry of the best error exponent: binary hypothesis

"Chernoff distribution" P^* [26]:

$$\boxed{P^* = P_{\theta^*_{12}} = G_e(P_1, P_2) \cap \mathrm{Bi}_m(P_1, P_2)}$$

e-**geodesic** (also sometimes called "Bhattacharrya arc"):

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

m-**bisector**:

$$\mathrm{Bi}_m(P_1, P_2) : \{P \mid F(\theta_1) - F(\theta_2) + \eta(P)^\top \Delta\theta = 0\},$$

Optimal natural parameter of P^*:

$$\theta^* = \theta_{12}^{(\alpha^*)} = \arg\min_{\theta\in\Theta} B(\theta_1 : \theta) = \arg\min_{\theta\in\Theta} B(\theta_2 : \theta).$$

$\rightarrow$ closed-form for order-1 family, or efficient bisection search [26].

Geometry of the best error exponent: multiple hypothesis

n-ary Multiply Hypothesis Testing (MHT) [13]: Bound P_e from **minimum pairwise Chernoff distance**:

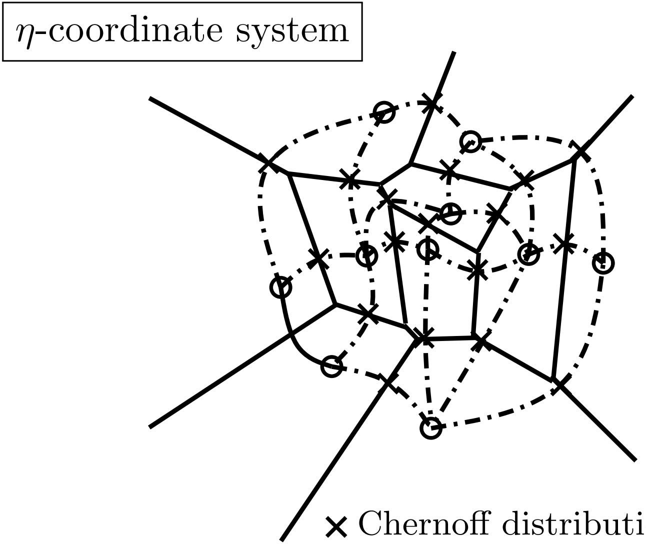
$$\boxed{C(P_1, \ldots, P_n) = \min_{i,j \neq i} C(P_i, P_j)}$$

$$P_e^m \leq e^{-mC(P_{i^*}, P_{j^*})}, \quad (i^*, j^*) = \arg\min_{i,j \neq i} C(P_i, P_j)$$

Compute for each pair of **natural neighbors** [4] P_{θ_i} and P_{θ_j}, the Chernoff distance $C(P_{\theta_i}, P_{\theta_j})$, and choose the pair with minimal distance.

$\rightarrow$ **Closest Bregman pair** problem for EFs
(Chernoff distance fails triangle inequality).

Multiple hypothesis testing: Illustration



× Chernoff distribution between natural neighbours

Recap of Chapter 4.

Bayesian multiple hypothesis testing [25] from the viewpoint of computational information geometry.

- Probability of error P_e & best MAP Bayesian rule
- P_e upper-bounded by the Chernoff distance
- MAP rule = Nearest Neighbor classifier (additive Bregman Voronoi diagram on the Exponential Family Manifold, EFM)
- Binary hypothesis: best error exponent from intersection primal geodesic/dual bisector (synthetic information geometry)
- Multiple hypothesis: best error exponent from closest Bregman pair for EFs

Chapter V. Geometric clustering in information spaces

Computing divergence-based centroids (survey)

$$\boxed{c^* = \arg\min_c \sum_{i=1}^n w_i D(p_i : c)} \leftarrow \text{weighted convex combination}$$

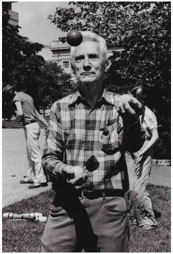
- D=Bregman divergence $\rightarrow$ closed-form [2, 36]
- D=Jeffreys divergence (symmetrized KL): Jeffreys centroid using Lambert W function [27]
- D=skew Jensen divergence $\rightarrow$ use Convex-ConCave Procedure (CCCP) [33]. Skew Bhattacharrya distances on EFs amounts to skew Jensen divergences on natural parameters
- Robust centroid: D=total Bregman $\rightarrow$ closed-form [15, 59, 16], total Jensen divergence [43]

Divergence-based Hard Clustering (survey)

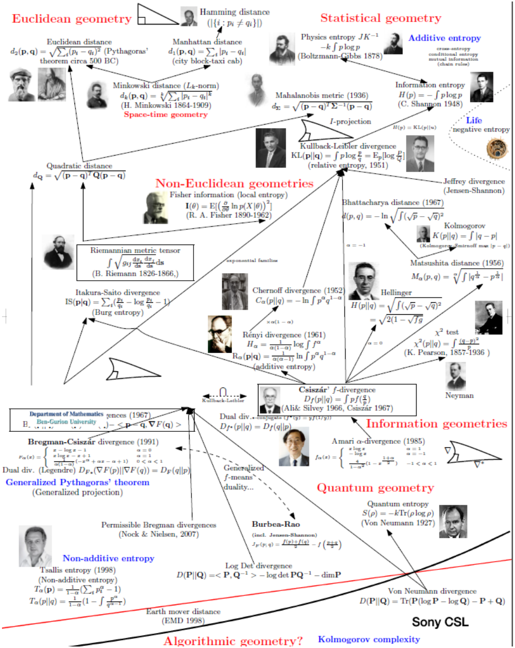
- **Baseline algorithm**: Bregman k-means hard clustering [2] with Bregman k-means++ initialization
  In 1D, exact using dynamic programming [42])

- Extend to **divergence-based centroid**: Minimize $\sum_i w_i D(p_i : c)$, and prove the arg min is unique...
- When divergence-based centroid not in closed-form (say, f-divergence centroids), use **variational k-means** [43]
- Introduce new classes of divergences to make clustering provably **robust**: total Bregman divergences [15, 59, 16], total Jensen divergences [43]. These are **conformal divergences** [49]: $\boxed{D(p:q) = \rho(p,q) D'(p:q)}$.
  $\rightarrow$ Applications to shape retrieval and biomedical imaging.
- To handle symmetrized divergences (SKL=Jeffreys), use mixed clustering [46] with **two dual centroids per cluster** (in closed form)

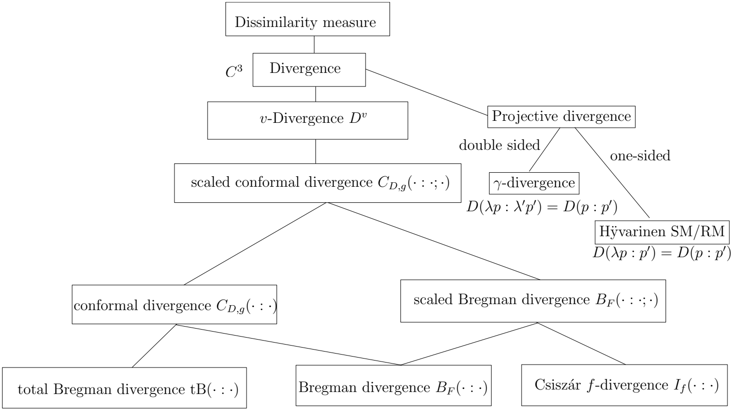
Chapter VI. Juggling with statistical distances and divergences



From a historical view of statistical distances...



... To a structural view of classes of distances



$D^v(P:Q) = D(v(P):v(Q))$

$I_f(P:Q) = \int p(x) f\left(\left(\frac{q(x)}{p(x)}\right)\right) \mathrm{d}\nu(x)$

$B_F(P:Q) = F(P) - F(Q) - \langle P-Q, \nabla F(Q)\rangle$

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

$C_{D,g}(P:Q) = g(Q) D(P:Q)$

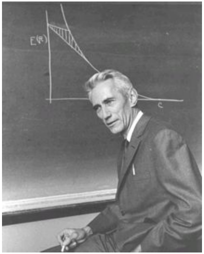
$B_{F,g}(P:Q;W) = W B_F\left(\frac{P}{Q}:\frac{Q}{W}\right)$

Axiomatic approach, exhausitivity characteristics

Calculating/estimating statistical distances $\int_{\mathcal{X}}$

- **Closed-form formula** for distributions of the same EF: Shannon [37], Rényi [40], Tsallis [40], Sharma-Mittal [39] (relative) entropies and relative entropies
- KL of mixtures is **not analytic**, but deterministic lower and upper bounds [48] using log-sum-exp inequalities
- **Unify** Jeffreys (SKL) with Jensen-Shannon (JS) divergences via a symmetric parametric family of divergences [19]
- **Design** tailored divergences for closed-form formula on mixtures: Cauchy-Schwarz divergence [21], Jensen-Rényi divergence [21], etc.
- Design **projective divergences** for inference of unnormalized models [7, 44] (like PEFs: Polynomial Exponential Families [45]): $D(\lambda p, \lambda' q) = D(p, q)$ for $\lambda, \lambda' > 0$.
  $\rightarrow$ Useful for handling unnormalized probability models.
- etc.

Conclusion:
Looking IT onward



Computational Information Geometry

In a nutshell...

- Computation...
  = science of transformations

- Information...
  = science of communication
  (between data and models)

- Geometry...
  = science of invariance

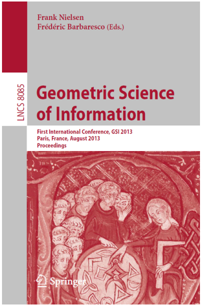
... nice interactions of C & I & G for future of IT!

IT onward: Computational Information Geometry

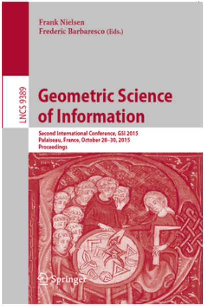
- Shannon information, the negative entropy, is convex, and thus it induces a **dually flat geometry**. Bring insights in MLE/MaxEnt as information projection.
- In many cases, the log-normalizer F of EFs is **computationally intractable** (Ising/Potts models, Restricted Boltzman Machines, etc.), and we need to consider non-MLE inference schemes (CDs, SMs, RMs, etc.)
- Furthermore, most statistical learning machines have **singularities** (FIM is degenerate $\rightarrow$ algebraic geometry [60])
- Alternative approach: **Optimal transport** (regularized) metric (Wasserstein centroid [1], Sinkhorn distance [6, 18]) but invariance is with respect to support geometry (not sufficient statistic)
- Deep Learning have gigantic FIM describing the **neuromanifold** that needs tailored inference strategies (eg, Krönecker factorization with natural gradient)
- Distances for correlated random variables: optimal copula transport for time-series datasets [17], etc.

Thank you I

Geometric Sciences of Information (GSI) biannual conferences:



2013



2015

3rd edition GSI'17: `www.gsi2017.org`

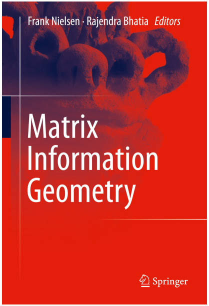
Geometric Sciences of Information, Paris, Fall 2017

GSI Portal:

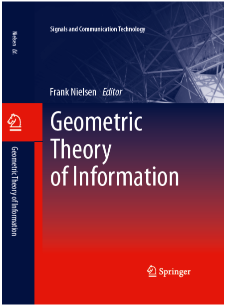
`http://forum.cs-dc.org/category/72/`

Thank you II

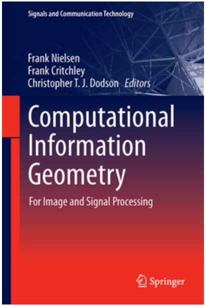
Edited books:



2012 [31]



2014 [29]



2016 [30]

Happy centennial birthday Claude E. Shannon!

claude shannon

1916 - 2016

References I

[1] Martial Agueh and Guillaume Carlier.
Barycenters in the Wasserstein space.
SIAM Journal on Mathematical Analysis, 43(2):904–924, 2011.

[2] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh.
Clustering with Bregman divergences.
Journal of Machine Learning Research, 6:1705–1749, 2005.

[3] Jean-Daniel Boissonnat, Frank Nielsen, and Richard Nock.
Bregman Voronoi diagrams.
Discrete & Computational Geometry, 44(2):281–307, 2010.

[4] Jean-Daniel Boissonnat and Mariette Yvinec.
Algorithmic Geometry.
Cambridge University Press, New York, NY, USA, 1998.

[5] Herman Chernoff.
A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations.
Annals of Mathematical Statistics, 23:493–507, 1952.

[6] Marco Cuturi.
Sinkhorn distances: Lightspeed computation of optimal transport.
In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

[7] Hironori Fujisawa and Shinto Eguchi.
Robust parameter estimation with a small bias against heavy contamination.
Journal of Multivariate Analysis, 99(9):2053–2081, 2008.

[8] Vincent Garcia, Eric Debreuve, Frank Nielsen, and Michel Barlaud.
k-nearest neighbor search: Fast GPU-based implementations and application to high-dimensional feature matching.
In *IEEE International Conference on Image Processing (ICIP)*, pages 3757–3760, 2010.

References II

[9] Vincent Garcia and Frank Nielsen.
Simplification and hierarchical representations of mixtures of exponential families.
Signal Processing, 90(12):3197–3212, 2010.

[10] Vincent Garcia, Frank Nielsen, and Richard Nock.
Hierarchical Gaussian mixture model.
In *ICASSP*, pages 4070–4073, 2010.

[11] Martin E. Hellman and Josef Raviv.
Probability of error, equivocation and the Chernoff bound.
IEEE Transactions on Information Theory, 16:368–372, 1970.

[12] Edwin Thompson Jaynes.
Information theory and statistical mechanics.
The Physical Review, 106(4):620–630, May 1957.

[13] C. C. Leang and D. H. Johnson.
On the asymptotics of M-hypothesis Bayesian detection.
IEEE Transactions on Information Theory, 43(1):280–282, January 1997.

[14] F. Liese and I. Vajda.
On divergences and informations in statistics and information theory.
Information Theory, IEEE Transactions on, 52(10):4394–4412, October 2006.

[15] Meizhu Liu, Baba C Vemuri, Shun-Ichi Amari, and Frank Nielsen.
Total Bregman divergence and its applications to shape retrieval.
In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3463–3468. IEEE, 2010.

[16] Meizhu Liu, Baba C Vemuri, Shun-ichi Amari, and Frank Nielsen.
Shape retrieval using hierarchical total Bregman soft clustering.
IEEE transactions on pattern analysis and machine intelligence, 34(12):2407–2419, 2012.

References III

[17] Gautier Marti, Frank Nielsen, and Philippe Donnat.
Optimal copula transport for clustering multivariate time series.
In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2379–2383. IEEE, 2016.

[18] B. Muzellec, R. Nock, G. Patrini, and F. Nielsen.
Tsallis Regularized Optimal Transport and Ecological Inference.
ArXiv e-prints, September 2016.

[19] Frank Nielsen.
A family of statistical symmetric divergences based on Jensen's inequality.
arXiv preprint arXiv:1009.4004, 2010.

[20] Frank Nielsen.
Legendre transformation and information geometry, 2010.
memo online.

[21] Frank Nielsen.
Closed-form information-theoretic divergences for statistical mixtures.
In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1723–1726. IEEE, 2012.

[22] Frank Nielsen.
Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means.
submitted, 2012.

[23] Frank Nielsen.
k-MLE: A fast algorithm for learning statistical mixture models.
In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 869–872. IEEE, 2012.

References IV

[24] Frank Nielsen.
Cramer-Rao lower bound and information geometry.
arXiv preprint arXiv:1301.3578, 2013.

[25] Frank Nielsen.
Hypothesis testing, information divergence and computational geometry.
In *Geometric Science of Information - First International Conference, GSI 2013, Paris, France, August 28-30, 2013. Proceedings*, pages 241–248, 2013.

[26] Frank Nielsen.
An information-geometric characterization of Chernoff information.
IEEE Signal Processing Letters (SPL), 20(3):269–272, March 2013.

[27] Frank Nielsen.
Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms.
IEEE Signal Processing Letters, 20(7):657–660, 2013.

[28] Frank Nielsen.
Pattern learning and recognition on statistical manifolds: An information-geometric review.
In Edwin Hancock and Marcello Pelillo, editors, *Similarity-Based Pattern Recognition*, volume 7953 of *Lecture Notes in Computer Science*, pages 1–25. Springer Berlin Heidelberg, 2013.

[29] Frank Nielsen.
Geometric Theory of Information.
Springer, 2014.

[30] Frank Nielsen.
Computational Information Geometry: For Signal and Image Processing.
Springer, 2016.

[31] Frank Nielsen and Rajendra Bhatia, editors.
Matrix Information Geometry (Revised Invited Papers). Springer, 2012.

References V

[32] Frank Nielsen and Sylvain Boltz.
The Burbea-Rao and Bhattacharyya centroids.
IEEE Transactions on Information Theory, 57(8):5455–5466, 2011.

[33] Frank Nielsen and Sylvain Boltz.
The Burbea-Rao and Bhattacharyya centroids.
IEEE Transactions on Information Theory, 57(8):5455–5466, 2011.

[34] Frank Nielsen and Vincent Garcia.
Statistical exponential families: A digest with flash cards.
arXiv preprint arXiv:0911.4863, 2009.

[35] Frank Nielsen and Richard Nock.
Clustering multivariate normal distributions.
In *Emerging Trends in Visual Computing*, pages 164–174. Springer Berlin Heidelberg, 2009.

[36] Frank Nielsen and Richard Nock.
Sided and symmetrized Bregman centroids.
IEEE transactions on Information Theory, 55(6):2882–2904, 2009.

[37] Frank Nielsen and Richard Nock.
Entropies and cross-entropies of exponential families.
In *2010 IEEE International Conference on Image Processing*, pages 3621–3624. IEEE, 2010.

[38] Frank Nielsen and Richard Nock.
Hyperbolic Voronoi diagrams made easy.
In *Computational Science and Its Applications (ICCSA), 2010 International Conference on*, pages 74–80. IEEE, 2010.

[39] Frank Nielsen and Richard Nock.
A closed-form expression for the Sharma-Mittal entropy of exponential families.
Journal of Physics A: Mathematical and Theoretical, 45(3):032003, 2011.

References VI

[40] Frank Nielsen and Richard Nock.
On Rényi and Tsallis entropies and divergences for exponential families.
arXiv preprint arXiv:1105.3259, 2011.

[41] Frank Nielsen and Richard Nock.
On the chi square and higher-order chi distances for approximating f-divergences.
IEEE Signal Process. Lett., 21(1):10–13, 2014.

[42] Frank Nielsen and Richard Nock.
Optimal interval clustering: Application to Bregman clustering and statistical mixture learning.
IEEE Signal Process. Lett., 21(10):1289–1292, 2014.

[43] Frank Nielsen and Richard Nock.
Total Jensen divergences: definition, properties and clustering.
In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2016–2020. IEEE, 2015.

[44] Frank Nielsen and Richard Nock.
Patch matching with polynomial exponential families and projective divergences.
In *Similarity Search and Applications - 9th International Conference, SISAP 2016, Tokyo, Japan, October 24-26, 2016. Proceedings*, pages 109–116, 2016.

[45] Frank Nielsen and Richard Nock.
Patch Matching with Polynomial Exponential Families and Projective Divergences, pages 109–116.
Springer International Publishing, Cham, 2016.

[46] Frank Nielsen, Richard Nock, and Shun-ichi Amari.
Sided, symmetrized and mixed α-clustering.
Entropy, 20:2, 2013.

References VII

[47] Frank Nielsen, Paolo Piro, and Michel Barlaud.
Bregman vantage point trees for efficient nearest neighbor queries.
In *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo (ICME)*, pages 878–881, 2009.

[48] Frank Nielsen and Ke Sun.
Guaranteed bounds on the Kullback-Leibler divergence of univariate mixtures using piecewise log-sum-exp inequalities.
arXiv preprint arXiv:1606.05850, 2016.

[49] Richard Nock, Frank Nielsen, and Shun-ichi Amari.
On conformal divergences and their population minimizers.
IEEE Transactions on Information Theory, 62(1):527–538, 2016.

[50] María del Carmen Pardo Llorente.
About distances of discrete distributions satisfying the data processing theorem of information theory.
IEEE transactions on information theory, 43(4):1288–1293, 1997.

[51] Paolo Piro, Frank Nielsen, and Michel Barlaud.
Tailored Bregman ball trees for effective nearest neighbors.
In *European Workshop on Computational Geometry (EuroCG)*, LORIA, Nancy, France, March 2009. IEEE.

[52] Yu Qiao and Nobuaki Minematsu.
A study on invariance of f-divergence and its application to speech recognition.
Transactions on Signal Processing, 58(7):3884–3890, July 2010.

[53] Christophe Saint-Jean and Frank Nielsen.
A new implementation of k-MLE for mixture modeling of Wishart distributions.
In *Geometric Science of Information - First International Conference, GSI 2013, Paris, France, August 28-30, 2013. Proceedings*, pages 249–256, 2013.

References VIII

[54] Olivier Schwander and Frank Nielsen.
Model centroids for the simplification of kernel density estimators.
In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 737–740. IEEE, 2012.

[55] Olivier Schwander and Frank Nielsen.
Learning mixtures by simplifying kernel density estimators.
In *Matrix Information Geometry*, pages 403–426. Springer, 2013.

[56] Olivier Schwander, Frank Nielsen, et al.
Comix: Joint estimation and lightspeed comparison of mixture models.
In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2449–2453. IEEE, 2016.

[57] Olivier Schwander, Aurélien J Schutz, Frank Nielsen, and Yannick Berthoumieu.
k-MLE for mixtures of generalized gaussians.
In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 2825–2828. IEEE, 2012.

[58] Ke Sun and Frank Nielsen.
Relative natural gradient for learning large complex models.
CoRR, abs/1606.06069, 2016.

[59] Baba C Vemuri, Meizhu Liu, Shun-Ichi Amari, and Frank Nielsen.
Total Bregman divergence and its applications to dti analysis.
IEEE Transactions on medical imaging, 30(2):475–483, 2011.

[60] Sumio Watanabe.
Algebraic information geometry for learning machines with singularities.
In *Advances in Neural Information Processing Systems 13*, pages 329–335. 2000.

Two common dually flat manifolds in statistics

Dual Geometry induced by a convex function $\xleftarrow{\mathrm{F}}$

Statistics:

- Exponential family: $F(\theta) = \log \int \exp(x^\top \theta) \mathrm{d}x$
- Mixture family: $F(\eta) = C_0(x) + \sum_i \eta_i F_i(x)$

KL of EF members $\equiv$ Bregman divergences

- Kullback-Leibler divergence = Cross-entropy - entropy

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

- KL between two distributions of the same EF:

$$\begin{aligned}\mathrm{KL}(P:Q) &= E_P\left[\log\frac{p(x)}{q(x)}\right] \geq 0\\ &= B_F(\theta_Q:\theta_P)\end{aligned}$$

- Bregman divergence:

$$B_F(\theta_1:\theta_2) = F(\theta_1) - F(\theta_2) - \langle\theta_1-\theta_2, \nabla F(\theta_2)\rangle$$

KL and dual Bregman divergences

For P and Q belonging to the <u>same</u> exponential families

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

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

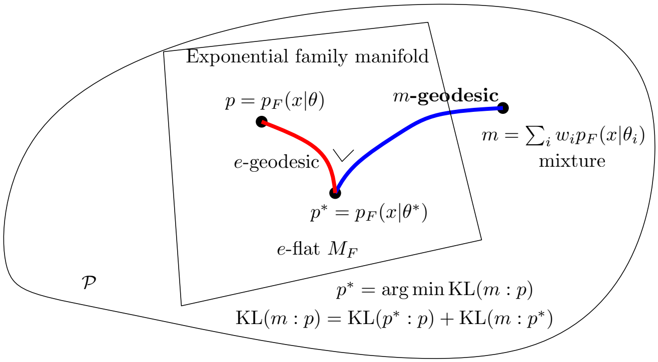
- Young inequality at the heart of the <u>canonical divergence</u>:

$$F(x) + F^*(y) \geq \langle x, y \rangle \quad \text{Young inequality}$$

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

Simplifying a mixture model into a single component [55]

m-projection of the mixture model m onto the e-flat (exponential family manifold): Best single distribution that approximates an exponential family mixture is found by taking the center of mass of the moment parameters: $\bar{\eta} = \sum_i w_i \eta_i$.



Mixture learning & mixture toolbox jMEF/PyMEF

Learning mixtures:

- Using the bijection of exponential families with Bregman divergences $\log p_F(x;\theta) = -B_{F^*}(t(x):\eta) + F^*(\eta) + k(x)$, Expectation Maximization for learning mixtures of EFs is equivalent to **soft Bregman k-means** [2] (locally consistent but global optimum difficult)
- k-MLE [23, 53] (hard EM, non consistent), add an extra stage where we can choose the exponential family component (= k-GMLE [57]). Monotonically converging.
- Learn a mixture by simplifying a Kernel Density Estimator (KDE) [54]
- Learn jointly a set of mixtures (comixs) [56]

Toolbox (software libraries `jMEF`/`PyMEF`):

- Simplify a mixture (like multivariate normal mixture) by entropic KL clustering [35] or by Fisher-Rao clustering [54]
- Hierarchical mixture models [10, 9] (level of details in CG)