Computational Information Geometry for Machine Learning

Part I: Geometry of statistical manifolds

Frank Nielsen

École Polytechnique
Sony Computer Science Laboratories, Inc
e-mail: Frank.Nielsen@acm.org

MLSS 2015
Computational Information Geometry (CIG) : Background

Computational Information Geometry (CIG) relies seamlessly on :

- statistics and probability (STAT & PR),
- information theory (IT),
- differential geometry (DG, including multilinear algebra of tensors),
- computation :
  Yes, we are computer scientists and programmers! How do we compute friendly? (make wide & wise use of dualities...)

Many application fields : computational statistics, machine learning (ML), information retrievals (IRs), computer vision (CV), medical imaging, radar signal processing, etc.

Motivations : Setting goals
Computational Information Geometry : Main goals

1. understand “distances” and group them axiomatically into classes and build generic meta-algorithms (unifying former algorithms) :
Bregman divergences $B_F$, Csiszár $f$-divergences $I_f$, proper scoring rules, etc.
Seek for “properties with exhaustivity”,

2. understand relationships between distances and geometries,

3. understand generalized entropies, cross-entropies, maximum entropy probability distributions, and their induced geometries (beyond Shannon/Boltzmann/Gibbs).

4. provide (coordinate-free) intrinsic computing using the language/affordances of geometry (for computational statistics, machine learning and predictive analytics)
Goal 1. Dissimilarities (distances) and meta-algorithms

- unify algorithms into **meta-algorithms** working on classes of distances (metrics, divergences):
  - parameter estimation (with goodness-of-fit),
  - clustering (with Bregman distances),
  - learning (with surrogate loss functions),
  - forecasting (with score functions),
  - etc.

- propose new **principled classes of distances**:
  total Bregman divergences [7], total Jensen divergences [15],
  conformal divergences [16], etc.

- understand axiomatically properties and relationships between distances (or multi-entity diversity indexes) and search for exhaustive characterizations.
Goal 2. Distances and geometries

Not 1-to-1 (because same geometry can be realized for different distances).
Geometry = meta-model
Embedding (isometrically) a geometry into another geometry:
= model interpreted into another larger model.

▶ Underlying geometries of distances/divergences:
  ▶ Riemannian geometry with metric distances (and metric Levi-Civita connection),
  ▶ Dually coupled affine differential geometry ($\pm \alpha$-geometry) and non-metric distances (aka. divergences),
  ▶ monotone embeddings into $(\rho, \tau)$-structure (extending $l_\alpha$-embedding),
  ▶ etc.

▶ geometries of probability distributions/positive measures and distances:
   How to define statistical manifolds?
Goal 3. Entropies, cross-entropies, relative entropies and MaxEnt distributions

- entropies $H(P)$ (Shannon-Boltzmann-Gibbs), cross-entropies $H^\times(P : Q)$ and relative entropies $I$
  $\left(\text{KL}(P : Q) = H^\times(P : Q) - H(P)\right)$
- generalized entropies (so called deformed “logarithms”), the concept of escort distributions,
- maximum entropy principle and equilibrium distributions
  (Boltzmann-Gibbs, Tsallis’s heavy tailed distributions, etc.)
- entropies, information ($=\text{neg-entropy}$) and complexity
  (Kolmogorov, non-computability)
Goal 4. Geometric computing for intrinsic computing

Propose a **paradigm for data science**: from “datum” (biased) processing to geometric “pointum” (non-biased) coordinate-free computing

- get **unbiased** processing: coordinate-free,
- use **affordances** of the geometric language for building/explaining algorithms: points, geodesics, balls, orthogonality, projection, Pythagoras, flat, submanifold, etc.
- **analytic and synthetic** geometries (closed-form or exact geometric characterization).
  Example: Two pseudo-segments always intersect in a common point... that many not be in closed-form.
- invariance (and **statistical invariance**) and geometry: group of invariance, invariance and sufficiency, statistical invariance, etc.

Geometrizing probability spaces yields **statistical manifolds**.
Outline of Part I (until 10:30am)

1. Fisher information (Cramér-Rao lower bound) & sufficiency (1922)
2. Structures from differential geometry of population spaces (1930, 1945, 1980’s)
3. Maximum entropy principle (exponential families) (1957, Jaynes)
4. Information projections (and Pythagoras’ theorem)
I. Fisher Information

$I(\theta)$
Old days - :) Discrete and Continuous random variables (RVs)

- **Discrete RV**: probability mass function (pmf) $X \sim p$, discrete support $\mathcal{X}$.

$$
\mathbb{E}[X] = \sum_{x \in \mathcal{X}} p(x)x = \langle X \rangle
$$

Distributions: Bernoulli, binomial, multinomial, Poisson, etc.-$\infty$.

- **Continuous RV**: probability density function (pdf) $X \sim p$, continuous support $\mathcal{X}$.

$$
\mathbb{E}[X] = \int_{\mathcal{X}} p(x)x\,dx = \langle X \rangle
$$

Distributions: exponential, normal, lognormal, gamma, beta, Dirichlet, Wishart, etc.-$\infty$.
From data sets to empirical (discrete) distributions

Given $X = \{x_1, ..., x_n\}$ observations...
...build the empirical distribution:

$$p_e(X) = \frac{1}{n} \sum_{i=1}^{n} \delta(X - X(i))$$

$$F_e(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{[x_i \leq x]} \text{ (cdf)}$$

$$p_e^i = \frac{1}{n} \#\{x = i\} \text{ (frequency)}$$

Support $\mathcal{X}$ is unknown a priori: not a multinomial nor a finite mixture!

Sample mean $\bar{\mu} = \frac{1}{n} \sum_i x_i = \langle X \rangle_{p_e} = \sum_{i \in \mathcal{X}} p_e^i i$.

Estimation $X \sim D(\theta)$ by the method of moments:

$$\langle X \rangle_{p_e} = \mathbb{E}[X] = \langle X \rangle.$$
Old days : Discrete and continuous random variables

- **Discrete RV. Shannon entropy** :
  \[
  H(X) = \sum_{x \in X} p(x) \log \frac{1}{p(x)} \geq 0
  \]
  always positive (notion of uncertainty! max uncertainty for uniform distribution \( \log n \))

- **Continuous RV. Differential entropy** :
  \[
  H(X) = \int_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \, dx
  \]
  can be negative (physical interpretation!) ...

For example, for multivariate normals (MVNs) \( N(\mu, \Sigma) \) :

\[
H(X) = \frac{1}{2} \log(2\pi e)^d |\Sigma|
\]
Mixture sampling: Example of a Gaussian Mixture Model (GMM)

To sample a variate $x$ from a GMM:

- Choose a component $l$ according to the weight distribution $w_1, \ldots, w_k$,
- Draw a variate $x$ according to $N(\mu_l, \Sigma_l)$.

→ Sampling is a **doubly stochastic process**:

- throw a biased dice with $k$ faces to choose the component:

  \[ l \sim \text{Multinomial}(w_1, \ldots, w_k) \]

  (Multinomial is normalized histogram without void bins)

- then draw at random a variate $x$ from the $l$-th component

  \[ x \sim N(\mu_l, \Sigma_l) \]

  \[ x = \mu + Cz \]

  with Cholesky: $\Sigma = CC^T$ and $z = [z_1 \ldots z_d]^T$

  standard normal random variate $z_i = \sqrt{-2 \log U_1 \cos(2\pi U_2)}$
Statistical mixtures: discrete, continuous or mixed!

Finite mixture models \((k \in \mathbb{N})\) have pmf/pdf:

\[
m(x) = \sum_{i=1}^{k} w_i p_i(x)
\]

(not sum of RVs, \(M \neq \sum_i w_i X_i\) that have convolutional densities)

- mixtures of Gaussians (universal representation for smooth densities)
- multinomial distribution is a mixture
  (and also an exponential family in information geometry...)

What about the mixture of a standard Gaussian with a binomial distribution? → Neither discrete nor continuous!
Measure theory (axiom system of Kolmogorov, 1933)

- unify discrete and continuous RVs as probability measures (pm) $\mu, \nu$, etc.
- can handle RVs that are neither continuous nor discrete (eg., a mixture of Poisson with a Gaussian)
- for probability measures, pmfs/pdfs are Radon-Nikodym derivatives
- expectation notation is unified as:

$$
\mathbb{E}[X] = \int_{x \in \mathcal{X}} x p(x) \, d\nu(x)
$$

- Two usual base measures:
  - counting measure: $\nu_C (\sum \rightarrow \sum)$
  - Lebesgue measure: $\nu_L$
Measure theory: Probability space (recalling terminology)

- $\mathcal{X}$ a set, the sample space
- $\sigma$-algebra $\mathcal{F}$ over $\mathcal{X}$: subsets of $\mathcal{X}$ closed under countable many intersections, unions, and complements.

- $(\mathcal{X}, \mathcal{F})$: measurable space
- measure $\mu: \mathcal{F} \to \mathbb{R} \cup \{\pm \infty\}$ with
  - $\mu(E) \geq 0, \forall E \in \mathcal{F}, \mu(\emptyset) = 0$
  - $\mu(\bigcup_{i \geq 1} E_i) = \sum_{i \geq 1} \mu(E_i)$ for pairwise disjoint sequence $\{E_i \in \mathcal{F}\}_i$

- $(\mathcal{X}, \mathcal{F}, \mu)$, a (positive) measure space
- $(\mathcal{X}, \mathcal{F}, \mu)$ with $\mu(\mathcal{X}) = 1$, a probability space, $F \in \mathcal{F}$ are events
Measurable functions and random variables

- **Measurable function** $f : \mathcal{X} \to \mathcal{Y}$ between two measurable spaces $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$:
  \[
  \forall G \in \mathcal{G}, \quad f^{-1}(G) \in \mathcal{F}
  \]

- **Random variable** $X = \text{measurable function } X : \mathcal{X} \to \mathbb{R}$. Therefore:
  \[
  \{x \in X \mid a < X(x) < b\} \in \mathcal{F}
  \]
  all sample states with $X$ taking values between $a$ and $b$ is an event (CDF)

- continuous $\text{RV} = \text{measures on Borel } \sigma\text{-algebra}
Dominance and Radon-Nikodym derivatives

- measure $\mu$ is dominated by measure $\nu$ ($\mu \ll \nu$) iff
  $$\nu(E) = 0 \Rightarrow \mu(E) = 0$$

- $\mu \ll \nu$ $\sigma$-finite ($\mathcal{X}=$countable union of measurable sets with finite measure) then $\mu$ admits a density $f$ wrt to $\nu$, the Radon-Nikodym derivative:

  $$f \overset{n.}{=} \frac{d\mu}{d\nu}$$

  $\forall \nu$ - measurable $E$, $\mu(E) \overset{n.}{=} \int_{e \in E} f d\nu(e)$

- $P \ll \nu$, Shannon entropy: $H(P) = - \int p(x) \log p(x) d\nu(x)$. 
Statistical estimation : parametric estimation $\hat{\theta}$

- Given iidd. $X = \{x_1, ..., x_n\} \sim p_{\theta_0}(x)$ (hidden by Nature), estimate $\theta$ in family $\{p_{\theta}(x)\}_\theta$?
  → from observation sets to random vectors
- **Maximum Likelihood Principle** (MLE):
  \[
  \hat{\theta}_n = \arg\max_\theta \prod_i p_{\theta}(x_i) = \arg\max_\theta l(X; \theta) = \sum_i \log p_{\theta}(x_i)
  \]
- **Consistency**: $\lim_{n \to \infty} \hat{\theta}_n = \theta_0$
- **score function**: $s(\theta, x) = \nabla_\theta \log p_{\theta}(x)$ with $\nabla_\theta = (\partial_i = \frac{\partial}{\partial \theta_i})_i$. score indicates the *sensitivity of the log-likelihood curve*.
- For strictly concave log-likelihood, unique $\hat{\theta}$ such that $s(\hat{\theta}, x) = 0$ (MVNs, Beta, Poisson, Dirichlet, etc).
Fisher information $I = \text{Variance of the score}$

Amount of information that an observable random variable $X$ carries about an unknown parameter $\theta$:
First moment of score: $0$, not discriminative!

$$I \left( \frac{\partial}{\partial \theta} \log p(X; \theta) \mid \theta \right) = \mathbb{E} \left[ \frac{\partial}{\partial \theta} p(X; \theta) \mid \theta \right] = \mathbb{E} \left[ \frac{\partial}{\partial \theta} \frac{p(X; \theta)}{p(X; \theta)} p(X; \theta) \mid \theta \right] = \int \frac{\partial}{\partial \theta} \frac{p(x; \theta)}{p(x; \theta)} p(x; \theta) \, dx$$

$$= \int \frac{\partial}{\partial \theta} p(x; \theta) \, dx = \frac{\partial}{\partial \theta} \int f(x; \theta) \, dx$$

$$= \frac{\partial}{\partial \theta} 1 = 0.$$

Second moment of score: $\partial i l(x; \theta) = \frac{\partial}{\partial \theta_i} l(x; \theta)$

$$I(\theta) = \mathbb{E} \left[ \left( \frac{\partial}{\partial \theta} \log f(X; \theta) \right)^2 \mid \theta \right] = \int \left( \frac{\partial}{\partial \theta} \log f(x; \theta) \right)^2 f(x; \theta) \, dx > 0$$

Multi-parameter: $I_{i,j}(\theta) = \mathbb{E}_\theta[\partial i l(x; \theta) \partial j l(x; \theta)]$, $I(\theta) \succeq 0$, PS(S)D
Fisher information and Cramér-Rao lower bound

How good is an estimator? how to measure goodness?

- **Mean Square Error (MSE)**: \( \text{MSE}(\theta) \overset{\text{eq}}{=} \mathbb{E}[\|\hat{\theta} - \theta_0\|^2] \)  
  (consistency: MSE → 0)

- **Cramér-Rao lower bound**: for an unbiased estimator \( \hat{\theta} \):
  \[
  \mathbb{V}[\hat{\theta}] \succeq I^{-1}(\theta_0)
  \]

- **efficiency**: unbiased estimator matching the CR lower bound

- **asymptotic normality** of \( \hat{\theta} \) (on random vectors):
  \[
  \hat{\theta} \sim \mathcal{N}\left(\theta_0, \frac{1}{n} I^{-1}(\theta_0)\right)
  \]
Fisher Information Matrix (FIM)

\[ I(\theta) = [l_{i,j}(\theta)]_{i,j}, \quad l_{i,j}(\theta) = \mathbb{E}_\theta[\partial_i l(x; \theta) \partial_j l(x; \theta)] \]

- For multinomials \((p_1, \ldots, p_d)\):

\[
I(\theta) = \begin{bmatrix}
p_1(1 - p_1) & -p_1p_2 & \cdots & -p_1p_k \\
-p_1p_2 & p_2(1 - p_2) & \cdots & -p_2p_k \\
\vdots & \vdots & \ddots & \vdots \\
-p_1p_k & -p_2p_k & \cdots & p_k(1 - p_k)
\end{bmatrix}
\]

- For multivariate normals (MVNs) \(N(\mu, \Sigma)\):

\[
l_{i,j}(\theta) = \frac{\partial \mu^\top}{\partial \theta_i} \Sigma^{-1} \frac{\partial \mu}{\partial \theta_j} + \frac{1}{2} \text{tr} \left( \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j} \right)
\]

matrix trace : \(\text{tr}\).
Reparameterization of the Fisher information matrix

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

\[
I_\eta(\eta) = J^\top \times I_\theta(\theta(\eta)) \times J
\]

Fisher information matrix depends on the parameterization of the parameter space (covariant)
Statistics : Information and sufficiency

- **sufficiency**: $\mathbb{P}(x|t, \theta) = \mathbb{P}(x|t)$
  $\Rightarrow$ all information about $\theta$ is contained inside $t$

- $I_s(x)(\theta) \leq I_x(\theta)$ for a statistic $s$, with equality iff. $s$ is **sufficient**

- **Fisher-Neyman’s factorization criterion**: $t(x)$ is sufficient then we have the following canonical factorization:
  
  $$p(x; \theta) = g(t(x); \theta)h(x)$$

- Ex. : $t(x) = (\sum_i x_i, \sum_i x_i^2)$ sufficient for univariate normals.
  
  - All information about $\theta$ in two quantities: data reduction without loss of statistical information
  - Sample mean $\bar{\mu} = \frac{1}{n} \sum_i x_i$, sample variance

  $$\bar{\nu} = \frac{1}{n} \sum_i (x_i - \bar{\mu})^2 = \frac{1}{n} \sum_i x_i^2 - \bar{\mu}^2 = \frac{1}{n} \sum_i x_i^2 - (\frac{1}{n} \sum_i x_i)^2$$

- Not all statistics carry information on $\theta$: **ancillary statistics**, statistics that does not depend on the parameter $\theta$. 
We are interested in finite-dimensional sufficient statistics...
(statistical lossless data reduction)
Exponential families and finite sufficiency

Probability measure

Parametric

Non-parametric

Exponential families

Univariate

Multivariate

Univariate

Bi-parameter

multi-parameter

uniparameter

Bi-parametric

multi-parameter

uniparameter

Bi-parametric

multi-parameter

Binomial

Bernoulli

Poisson

Exponential

Rayleigh

Gaussian

Univariate

Beta $\beta$

Gamma $\Gamma$

Bi-parametric

Multinomial

Dirichlet

Weibull

Non-exponential families

Uniform

Cauchy

Lévy skew $\alpha$-stable

Univariate

Multivariate

Beware: Exponential distribution belongs to the exponential families too.
Exponential families: families of parametric distributions

- Canonical decomposition \((t(x)\) sufficient statistics, \(k(x)\) auxiliary carrier term): 
  \[
  p(x; \theta) = \exp(\langle t(x), \theta \rangle) - F(\theta) + k(x)
  \]

- log-Laplace transform: \(F(\theta) = \log \int \exp(\langle t(x), \theta \rangle) + k(x)dx\)

- many distributions \(p(x; \lambda)\) (normal, gamma, beta, multinomial, Poisson) are exponential families with \(\theta(\lambda)\)

- \(F\) is strictly convex on convex natural parameter space 
  \[
  \Theta = \{\theta \in \mathbb{R}^D \mid F(\theta) < \infty\}
  \]

- Dual parameterizations: \(\theta(\lambda)\) or \(\eta(\lambda) = \nabla F(\theta(\lambda)) = \mathbb{E}[t(X)]\)

- Fisher information matrix: \(I(\theta) = \nabla^2 F(\theta) \succ 0\) (Hessian of strictly convex function)

- MLE: \(\hat{\eta} = \frac{1}{n} \sum_i t(x_i) = \nabla F(\theta)\) (condition on existence)
Convex duality: Legendre-Fenchel transformation [10, 8]

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

\[
F^*(y) = \sup_{x \in \mathcal{X}} \left\{ \langle y, x \rangle - F(x) \right\}
\]

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

\[
\nabla_x l_F(y; x) = y - \nabla F(x) = 0 \implies y = \nabla F(x)
\]

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

\[
\nabla_x^2 l_F(y; x) = -\nabla^2 F(x) \prec 0
\]

- Convex conjugates with domains:

\[
(F, \mathcal{X}) \leftrightarrow (F^*, \mathcal{Y}), \quad \mathcal{Y} = \{\nabla F(x) \mid x \in \mathcal{X}\}
\]
Legendre duality: Geometric interpretation

Consider the epigraph of $F$ as a convex object:

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

![Diagram showing Legendre duality]

Legendre transform also called “slope” transform.
Legendre duality & Canonical divergence

- Convex conjugates have functional inverse gradients
  \[ \nabla F^{-1} = \nabla F^* \]
  \[ \nabla F^* \text{ may require numerical approximation} \]
  (not always available in analytical closed-form)

- **Involution**: \((F^*)^* = F\) with \(\nabla F^* = (\nabla F)^{-1}\).

- Convex conjugate \(F^*\) expressed using \((\nabla F)^{-1}\):

  \[
  F^*(y) = \langle x, y \rangle - F(x), \quad x = \nabla_y F^*(y) \\
  F^*(y) = \langle ((\nabla F)^{-1}(y), y \rangle - F((\nabla F)^{-1}(y))
  \]

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

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

\[
A_F(x : y) = A_{F^*}(y : x) = F(x) + F^*(y) - \langle x, y \rangle \geq 0
\]
Parameters of exponential families

- $D$: order of the exponential family
- $d$: uni- ($d = 1$) or multi-variate family

Many parameterizations are possible but only two are canonical: natural parameters and expectation parameters.

Original parameters

\[
\lambda \in \Lambda
\]

Exponential family dual parameterization

\[
\begin{align*}
\theta &\in \Theta \\
\eta &\in H
\end{align*}
\]

Legendre transform

\[
(\Theta, F) \leftrightarrow (H, F^*)
\]

Natural parameters

\[
\eta = \nabla_{\theta} F(\theta)
\]

Expectation parameters

\[
\theta = \nabla_{\eta} F^*(\eta)
\]
Canonical decomposition of exponential families

\[ \langle \cdot, \cdot \rangle : \text{inner product on vectors (scalar product), matrices} \]
\[ (\text{ReTr}(AB^*)) \]
\[ t(x) \text{ sufficient statistics, } k(x) \text{ auxiliary carrier term} : \]

\[
p(x; \theta) = \exp(\langle t(x), \theta \rangle) - F(\theta) + k(x)
\]

Not unique decomposition because:

- natural parameter and sufficient statistic: \( t'(x) = At(x) \) and \( \theta' = A^{-1}\theta \) (for \( |A| \neq 0 \) affine transformation)
- constant in \( F'(\theta) = F(\theta) + c \) and \( k'(x) = k(x) - c \)

Let us give decomposition examples...

IntraVascular UltraSound (IVUS) imaging:

Rayleigh distribution:
\[ p(x; \lambda) = \frac{x}{\lambda^2} e^{-\frac{x^2}{2\lambda^2}} \]
\[ x \in \mathbb{R}^+ \]
\[ d = 1 \text{ (univariate)} \]
\[ D = 1 \text{ (order 1)} \]
\[ \theta = -\frac{1}{2\lambda^2} \]
\[ \Theta = (-\infty, 0) \]
\[ F(\theta) = -\log(-2\theta) \]
\[ t(x) = x^2 \]
\[ k(x) = \log x \]
(Weibull \( k = 2 \))

Coronary plaques: fibrotic tissues, calcified tissues, lipidic tissues

Rayleigh Mixture Models (RMMs):
for segmentation and classification tasks
Statistical mixtures: Gaussian MM [3, 11, 4]

Gaussian mixture models (GMMs): model low frequency.
Color image interpreted as a 5D xyRGB point set.

Gaussian distribution $p(x; \mu, \Sigma)$:

$$
\frac{1}{(2\pi)^{d/2} \sqrt{|\Sigma|}} e^{-\frac{1}{2} D_{\Sigma^{-1}}(x-\mu, x-\mu)}
$$

Squared Mahalanobis distance:

$$
D_{Q}(x, y) = (x - y)^T Q (x - y)
$$

$x \in \mathbb{R}^d$

$d$ (multivariate)

$$
D = \frac{d(d+3)}{2} \text{ (order)}
$$

$$
\theta = (\Sigma^{-1}\mu, \frac{1}{2} \Sigma^{-1}) = (\theta_v, \theta_M)
$$

$$
\Theta = \mathbb{R} \times S^d_{++}
$$

$$
F(\theta) = \frac{1}{4} \theta_v^T \theta_M^{-1} \theta_v - \frac{1}{2} \log |\theta_M| + \frac{d}{2} \log \pi
$$

$$
t(x) = (x, -xx^T)
$$

$$
k(x) = 0
$$
MLE of exponential families: Two coordinate systems

\[ \eta = \mathbb{E}[t(x)] = \nabla F(\theta), \quad \theta = (\nabla F)^{-1}(\eta) = \nabla F^*(\eta) \]

- **Closed-form in expectation parameter coordinate system** \( \eta \):
  \[ \hat{\eta} = \frac{1}{n} \sum_i t(x_i) \]
- **Convex optimization in the natural parameter coordinate system** \( \theta \):
  \[ \max_\theta I(\theta; x_1, ..., x_n) = \frac{1}{n} \sum_i (\langle t(x_i), \theta \rangle - F(\theta)) \equiv \min_\theta F(\theta) - \langle \theta, \bar{t} \rangle \] (that is, \( \nabla F(\hat{\theta}) = \bar{t} \))

Convex optimization

\[ \min_\theta F(\theta) - \langle \theta, \bar{t} \rangle \]

Trivial solution

\[ \hat{\eta} = \frac{1}{n} \sum_i t(x_i) = \bar{t} \]
Exponential families: Universal families!

Universal representations of “smooth” densities:

- **mixtures** of exponential families approximate any smooth density (mixtures of Gaussians)

- a single exponential family (possibly multimodal) approximates also any smooth density: Similar to approximations of functions by polynomials. We can choose the sufficient statistics in $(1, x, x^2, x^3, ...)$ and $(\log x, \log^2 x, \log^3 x, ...)$. But then $F(\theta)$ not in closed form:

\[
F(\theta) = \int_x \exp \left( \theta^\top t(x) + k(x) \right) \, d\nu(x)
\]

(common problem met in practice not to have closed-form expression of $F$)
Boltzmann-Gibbs distribution in statistical physics

Let $E(X; \theta)$ be an energy function.

$$p(X; \theta) = \frac{1}{Z(\theta)} \exp(-E(X; \theta))$$

$Z(\theta)$ normalization factor (aka. partition function):

$$Z(\theta) = \int_{x} \exp(-E(X; \theta)) d\nu(x)$$

$$F(\theta) = \log Z(\theta)$$
The observed point $\hat{P}$ in information geometry

- $\{P_\theta\}_\theta$: a parametric (exponential family) model, \textbf{identifiable}
- View $P_\theta$ as a point on a manifold (dual coordinates $\theta$ and $\eta$)
- Observed point $\hat{P}$ with $\eta$-coordinate $\overline{t(x)} = \frac{1}{n} \sum_i t(x_i)$ (MLE)

We shall see later that $\hat{P}$ is $m$-projection of the empirical distribution on the e-flat...
MLE of exponential families [9]

- \( \hat{\eta} = \overline{t(x)} \) but we would like \( \hat{\theta} = (\nabla F^{-1})(\hat{\eta}) \)
- maximum likelihood:
  
  \[
  l(\theta; x_1, \ldots, x_n) = F^*(\hat{\eta}) + \overline{k(x)}
  \]

  \[
  \overline{k(x)} = \frac{1}{n} \sum_{i=1}^{n} k(x_i)
  \]
  \( F^* \) is neg-entropy

- When \( F(\theta) \) not in closed-form: Contrastive Divergence (MCMC), score matching (Fisher divergence), etc.
II. Geometric structures of probability manifolds:

- \((M, g)\)

- \((M, g, \nabla, \nabla^*)\)
Population space & Parameter space


- $\mathcal{P} = \{ p(x|\theta) \mid \theta \in \Theta \}$ a parametric family of distributions, the population space,
- $\Theta$, the parameter space of dimension $D$
- immersion $i(\theta) = p(x|\theta)$ from the parameter space to the population space:
  - $i$ : one-to-one (model identifiability)
  - $i$ of rank $\dim(\Theta) = D$:
    $$\frac{\partial p(x|\theta)}{\partial \theta_1}, \ldots, \frac{\partial p(x|\theta)}{\partial \theta_D}$$
    ... are linearly independent
- Geometric structures of SPD matrices when we consider the particular space $\{ \mathcal{N}(0, \Sigma) \mid \Sigma \succ 0 \}$

© 2015 Frank Nielsen
Fisher information matrix (FIM)

- log-likelihood \( l(\theta|x) = \log p(x|\theta) \), \( \partial_i = \frac{\partial}{\partial \theta_i} \).
- Metric tensor, \( D \times D \) matrix: \( g = [g_{ij}] = \sum_{i,j} dx_i \otimes dx_j \) (tensor product)

\[
g_{ij} = \mathbb{E}_\theta[\partial_i l(\theta) \partial_j l(\theta)]
\]

- can be rewritten *equivalently* as:

\[
g_{ij} = 4 \int_x \partial_i \sqrt{p(x|\theta)} \partial_j \sqrt{p(x|\theta)} dx
\]

- \( g \) symmetric positive definite (SPD), non-degenerate when \( \{\partial_i p(x|\theta)\}_i \) are linear independent (problem with mixture models where \( \exists \theta, l(\theta) = 0 \))
Fisher information matrix & Hessian

Negative expectation of the Hessian of the log-likelihood function:

\[ g_{ij} = \mathbb{E}_\theta [ \partial_i l(\theta) \partial_j l(\theta) ] \]

\[ g_{ij} = 4 \int_x \partial_i \sqrt{p(x|\theta)} \partial_j \sqrt{p(x|\theta)} dx \]

\[ g_{ij} = -\mathbb{E}_\theta [ \partial_i \partial_j l(\theta) ] \]

For exponential families \( p(x|\theta) = \exp(\langle \theta, x \rangle - F(\theta)) \),

\[ I(\theta) = \nabla^2 F(\theta) \geq 0 \]
Fisher information: invariance and covariance

- **Invariant under reparameterization of the sample space**: $X$ RV. with $p(x|\theta)$ and $Y = f(X)$ for an invertible transformation $f(\cdot)$ with density $\bar{p}(y|\theta)$.

  $$g_{ij}(\theta) = \bar{g}_{ij}(\theta)$$

- **Covariant under reparameterization of the parameter space**: Let $\eta = \eta(\theta)$ be an invertible transformation with $\bar{p}_\eta(x) = p_{\eta(\theta)}(x)$

  $$\bar{g}_{ij}(\eta) = g_{kr} \bigg|_{\eta=\eta(\theta)} \frac{\partial \theta_k}{\partial \eta_i} \frac{\partial \theta_r}{\partial \eta_j}$$

- **Sufficient statistics**: $p(x|t, \theta) = p(x|t)$, non-deterministic Markov morphism transformations (statistical invariance).
Basics of Riemannian geometry

- $(M, g)$ : Riemannian manifold
- $\langle \cdot, \cdot \rangle$, Riemannian metric tensor $g$ : definite positive bilinear form on each tangent space $T_x M$ (depends smoothly on $x$)
- $\| \cdot \|_x : \| u \| = \langle u, u \rangle^{1/2}$ : Associated norm in $T_x M$
- $\rho(x, y)$ : metric distance between two points on the manifold $M$ (length space)

$$
\rho(x, y) = \inf \left\{ \int_0^1 \| \dot{\gamma}(t) \| \, dt, \quad \gamma \in C^1([0, 1], M), \quad \gamma(0) = x, \quad \gamma(1) = y \right\}
$$

- Shortest paths (length space)
- but technically parallel transport wrt. Levi-Civita metric connection $\nabla$ ($\nabla g = 0$).
Basics of Riemannian geometry: Exponential map

- Local map from the tangent space $T_x M$ to the manifold defined with geodesics (wrt $\nabla$).

\[ \forall x \in M, D(x) \subseteq T_x M : D(x) = \{ v \in T_x M : \gamma_v(1) \text{ is defined} \} \]

with $\gamma_v$ maximal (i.e., largest domain) geodesic with $\gamma_v(0) = x$ and $\gamma'_v(0) = v$.

- Exponential map:

\[ \exp_x(\cdot) : D(x) \subseteq T_x M \to M \]

\[ \exp_x(v) = \gamma_v(1) \]

$D$ is star-shaped.
Basics of Riemannian geometry : Geodesics

- **Geodesic**: smooth path which locally minimizes the distance between two points.

- Given a vector \( v \in T_x M \) with base point \( x \), there is a unique geodesic started at \( x \) with speed \( v \) at time 0: \( t \mapsto \exp_x(tv) \) or \( t \mapsto \gamma_t(v) \).

- Geodesic on \( [a, b] \) is *minimal* if its length is less or equal to others. For *complete* \( M \) (i.e., \( \exp_x(v) \)), taking \( x, y \in M \), there exists a *minimal* geodesic from \( x \) to \( y \) in time 1.

\[ \gamma_{(x, y)} : [0, 1] \to M, \ t \mapsto \gamma_t(x, y) \] with the conditions \( \gamma_0(x, y) = x \) and \( \gamma_1(x, y) = y \).

- \( U \subseteq M \) is *convex* if for any \( x, y \in U \), there exists a unique minimal geodesic \( \gamma_{(x, y)} \) in \( M \) from \( x \) to \( y \). Geodesic *fully lies* in \( U \) and depends smoothly on \( x, y, t \).
Basics of Riemannian geometry: Geodesics

- Geodesic $\gamma(x, y)$: locally minimizing curves linking $x$ to $y$

- Speed vector $\gamma'(t)$ parallel along $\gamma$:
  \[
  \frac{D\gamma'(t)}{dt} = \nabla_{\gamma'(t)}\gamma'(t) = 0
  \]

- When manifold $M$ embedded in $\mathbb{R}^d$, acceleration is normal to tangent plane:
  \[
  \gamma''(t) \perp T_{\gamma(t)}M
  \]

- $\|\gamma'(t)\| = c$, a constant (say, unit).

$\Rightarrow$ Parameterization of curves with constant speed (otherwise, you get the trace of the geodesic only...)
Basics of Riemannian geometry: Geodesics and means

Constant speed geodesic $\gamma(t)$ so that $\gamma(0) = x$ and $\gamma(\rho(x, y)) = y$ (constant speed 1, the unit of length).

$$x \#_t y = m = \gamma(t) : \rho(x, m) = t \times \rho(x, y)$$

For example, in the Euclidean space:

$$x \#_t y = (1 - t)x + ty = x + t(y - x) = m$$

$$\rho_E(x, m) = \| t(y - x) \| = t \| y - x \| = t \times \rho(x, y), \ t \in [0, 1]$$

$\Rightarrow m$ interpreted as a **mean** (barycenter) between $x$ and $y$. 
Basics of Riemannian geometry: Injectivity radius

Diffeomorphism from the tangent space to the manifold

- Injectivity radius $\text{inj}(M)$: largest $r > 0$ such that for all $x \in M$, the map $\exp_x(\cdot)$ restricted to the open ball in $T_x M$ with radius $r$ is an embedding.

- Global injectivity radius: infimum of the injectivity radius over all points of the manifold.

Important for navigating back and forth from $T_x M$ to $M$ (extrinsic/intrinsic computing)...
Riemannian geometry of population spaces

- Consider \((M, g)\) with \(g = I(\theta)\), Hotelling (1930), Rao (1945). Fisher information matrix is unique up to a constant (for statistical invariance).
- Geometry of multinomials is spherical (on the orthant)
- For univariate location-scale families, hyperbolic geometry or Euclidean geometry (location only)

\[
p(x|\mu, \sigma) = \frac{1}{\sigma} p_0 \left( \frac{x - \mu}{\sigma} \right), \quad X = \mu + \sigma X_0
\]

(Normal, Cauchy, Laplace, Student \(t\)-, etc.)
Tangent planes, tangent bundles, vector fields

- $T_p$ : tangent plane at $p$
- $TM$, tangent bundle
- vector field = global section of the tangent bundle
- Mahalanobis metric distance on tangent planes $T_x$:

$$M_Q(p, q) = \sqrt{(p - q)^\top Q(x)(p - q)}$$

axioms of the metric for $Q(x) = g(x) \succ 0$ (SPD).

- Rao’s distance between close points amounts to $\rho \simeq \sqrt{2KL} = \sqrt{SKL}$. For exponential families, $\rho \simeq$ Mahalanobis $= \sqrt{\Delta \theta^\top I(\theta) \Delta}$. 
Tangent plane : basis vectors

- $\left( \partial_i \right)_x = \left( \frac{\partial}{\partial \theta^i} \right)_x$
- $X_x = \sum_{i=1}^{D} X^i (\partial_i)_x$
- Define proper metric tensor : $g_{ij}(x) = g_x (\partial_i, \partial_j) > 0$
\(\alpha\)-representations and parameterizations of the tangent planes

\[
f_\alpha(u) = \begin{cases} 
\frac{2}{1-\alpha} u^{1-\alpha/2}, & \alpha \neq 1 \\
\log u, & \alpha = 1.
\end{cases}
\]

- \(\alpha = -1\) : \(p(x|\theta) \rightarrow f_{-1}(p(x|\theta)) = p(x|\theta)\): usual parameterization of the tangent plane \(T^{(-1)}_x M\) with basis \(\partial_i^{(-1)} = \partial_i\).

- \(\alpha = 0\) : square root representation:
  \(p(x|\theta) \rightarrow f_0(p(x|\theta)) = 2\sqrt{p(x|\theta)} \). \(\partial^{(0)}\) perpendicular to \(\theta\), identified with the tangent plane \(T^{(0)}_x M\).

- \(\alpha = 1\) : logarithmic representation:
  \(p(x|\theta) \rightarrow f_1(p(x|\theta)) = \log p(x|\theta)\).
  \(\partial^{(1)} = \partial_i f_1(p(x|\theta)) = \frac{1}{p(x|\theta)} \partial_i p(x|\theta)\)

Tangent planes are invariant objects: do not depend on the \(\alpha\)-representation.
Extrinsic Computational Geometry on tangent planes

- Tensor \( g = Q(x) > 0 \) defines smooth inner product
  \( \langle p, q \rangle_x = p^\top Q(x) q \) that induces a normed distance:
  \( d_x(p, q) = \|p - q\|_x = \sqrt{(p - q)^\top Q(x)(p - q)} \)

- Mahalanobis metric distance on tangent planes:

\[
\Delta_{\Sigma}(X_1, X_2) = \sqrt{(\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 - \mu_2)} = \sqrt{\Delta \mu^\top \Sigma^{-1} \Delta \mu}
\]

- Cholesky decomposition \( \Sigma = LL^\top \), lower triangular matrix \( L \):

\[
\Delta(X_1, X_2) = D_E(L^{-1} \mu_1, L^{-1} \mu_2)
\]

- Computing on tangent planes = Euclidean computing on transformed points \( x' \leftarrow L^{-1}x \).

Extrinsic vs intrinsic computations.
Riemannian Mahalanobis metric tensor ($\Sigma^{-1}$, PSD)

$$\rho(p_1, p_2) = \sqrt{(p_1 - p_2)^\top \Sigma^{-1} (p_1 - p_2)}, \quad g(p) = \Sigma^{-1} = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}$$

non-conformal geometry: $g(p) \neq f(p)I$

(Visualization with Tissot indicatrix)
Normal/Gaussian family and 2D location-scale families

- Fisher Information Matrix (FIM):

\[
I(\theta) = \left[ I_{i,j}(\theta) = \mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta_i} \log p(x|\theta) \frac{\partial}{\partial \theta_j} \log p(x|\theta) \right] \right] = \mathbb{E}_\theta[\partial_i I \partial_j I]
\]

- FIM for univariate normal/multivariate spherical distributions:

\[
I(\mu, \sigma) = \begin{bmatrix}
\frac{1}{\sigma^2} & 0 \\
0 & \frac{2}{\sigma^2}
\end{bmatrix} = \frac{1}{\sigma^2} \begin{bmatrix}
1 & 0 \\
0 & 2
\end{bmatrix}
\]

\[
I(\mu, \sigma) = \text{diag} \left( \frac{1}{\sigma^2}, ..., \frac{1}{\sigma^2}, \frac{2}{\sigma^2} \right)
\]

- \(\to\) amount to Poincaré metric \(\frac{dx^2 + dy^2}{y^2}\), hyperbolic geometry in upper half plane/space.
Riemannian Poincaré upper plane metric tensor (conformal)

\[
\cosh \rho(p_1, p_2) = 1 + \frac{\|p_1 - p_2\|^2}{2y_1y_2}, \quad g(p) = \begin{bmatrix} \frac{1}{y^2} & 0 \\ 0 & \frac{1}{y^2} \end{bmatrix} = \frac{1}{y^2} I
\]

\text{conformal : } g(p) = \frac{1}{y^2} I
Matrix SPD spaces and hyperbolic geometry

Symmetric Positive Definite matrices $M : \forall x \neq 0, x^T M x > 0$.

- 2D SPD(2) matrix space has dimension $d = 3$ : A positive cone.

$$\text{SPD}(2) \{ (a, b, c) \in \mathbb{R}^3 : a > 0, \ ab - c^2 > 0 \}$$

- Can be peeled into sheets of dimension 2, each sheet corresponding to a constant value of the determinant of the elements

$$\text{SPD}(2) = \text{SSPD}(2) \times \mathbb{R}^+$$

where $\text{SSPD}(2) = \{ a, b, c = \sqrt{1 - ab} : a > 0, ab - c^2 = 1 \}$

- Mapping $M(a, b, c) \to \mathbb{H}^2$ :
  - $(x_0 = \frac{a+b}{2} \geq 1, x_1 = \frac{a-b}{2}, x_2 = c)$ in hyperboloid model [14]
  - $z = \frac{a-b+2ic}{2+a+b}$ in Poincaré disk [14].
Riemannian Poincaré disk metric tensor (conformal)

→ often used in Human Computer Interfaces, network routing (embedding trees), etc.
Riemannian Klein disk metric tensor (non-conformal)

- recommended for “computing space” since geodesics are straight line segments
- Klein is also conformal at the origin (so we can perform translation from and back to the origin via Möbius transform.)
- Geodesics passing through O in the Poincaré disk are straight (so we can perform translation from and back to the origin)
Riemannian geometry: Optimization on the manifold with the natural gradient [1]

Numerical optimization on manifolds:

- defined on a manifold, generalize Euclidean gradient
  \[ \nabla_x f(x) = \left( \frac{\partial}{\partial x_1} f(x), \ldots, \frac{\partial}{\partial x_D} f(x) \right) \]
- natural gradient respects intrinsic geometry of the manifold:
  \[
  \tilde{\nabla}_\theta f(\theta) = (I(\theta))^{-1} \nabla_\theta f(\theta)
  \]

  (Euclidean geometry: \( I(\theta) = I \).)
- invariant under changes of the parameterization (natural gradient = contravariant form of the gradient)
- Information-geometric optimization (IGO), black-box optimization
Jeffrey’s prior from volume element

- Volume of the manifold:

\[ \nu(M) = \int \sqrt{|g(\theta)|} \, d\theta < \infty \]

- Consider the prior distribution:

\[ q(\theta) = \frac{1}{\nu(M) \sqrt{|g(\theta)|}} \]

- Invariant under reparameterization

- Bayesian statistics (and other \( \pm \alpha \)-volume element in IG:

\[ |g|^{\frac{1\pm\alpha}{2}} \]
Affine differential geometry: dual connections $\nabla$ and $\nabla^*$ coupled with a metric $g$
Connections $\Pi$ and covariant derivatives $\nabla$

- Connections $\Pi$ set correspondences between vectors in
tangent spaces $T_p$ and $T_q$. When manifold $M$ is embedded in
$\mathbb{R}^d$, there exists a natural correspondence. Otherwise,
connections $\Pi$ need to be formally defined.

- Covariant derivatives $\nabla$: differentiation of a vector field $Y$
in the direction of another vector field $X$, yielding a vector field
$Z = \nabla_X Y$.

- Connections and covariant derivatives induce the same geometric structure. Yield notions of geodesics,
flatness/curvature, parallelness, torsion.

- Riemannian structure $(M, g)$ has an induced metric connection
$\nabla_g = \nabla_{LC} = \nabla^{(0)}$, called the Levi-Civita connection.
Connections and parallel transport

- $\prod_{p,q}$ a connection from $T_p$ to $T_q$

$$\prod_{p,q} : T_p \to T_q$$

so that $v \in T_p$ yields $w = \prod_{p,q}(v) \in T_q$

- from linear isomorphism between tangent spaces of neighboring points to tangent points between arbitrary points by integrating along a curve $\gamma_{p,q}$ connecting $p$ with $q$.

- $d^3$ coefficients $\Gamma_{ijk}(p)$ required for defining $\prod$.

- Vector field $X$ along $\gamma$ with $X(t + dt) = \prod_{\gamma(t), \gamma(t+dt)} X(t)$. We say vector fields $\{X(t) \mid t\}$ along $\gamma$ are parallel with respect to the connection $\prod$. Parallel transport.
Covariant derivatives $\nabla$

$\nabla$: differentiation of a vector field $Y$ in the direction of another vector field $X$, yielding a vector field $Z = \nabla_X Y$.

$$\nabla : V(M) \times V(M) \to V(M)$$

Properties $\nabla$ should have:

$$\nabla_{f_1 X_1 + f_2 X_2} Y = f_1 \nabla_{X_1} Y + f_2 \nabla_{X_2} Y$$

$$\nabla_X (Y_1 + Y_2) = \nabla_X Y_1 + \nabla_X Y_2$$

$$\nabla_X (fY) = f \nabla_X Y + (Xf) Y$$

Linear combinations of covariant derivatives is a covariant derivative
Vector field parallel to a curve

Vector field $Y \in V(M)$ is $\nabla$-parallel to a curve $\gamma(t)$:

$$\forall t, \forall X \in V(M), \quad \nabla_{\dot{\gamma}(t)} Y = 0$$
Geodesics in differential geometry

Curves $\gamma$ on $(M, \nabla)$ such that

$$\forall t, \quad \nabla \dot{\gamma}(t) \dot{\gamma}(t) = 0$$
Affine coordinate system and flat connection

In general, specify a connection/covariant $\nabla$ by $D^3$ coefficients:

$$\nabla \partial_i \partial_j = \Gamma^k_{ij} \partial_k, \quad \forall i, j, k \in \{1, \ldots, D\}$$

$(M, \nabla)$, $\theta$ a coordinate system.

$\theta$ is an affine coordinate system iff:

- Vector fields $\{\partial_i = \frac{\partial}{\partial \theta_i}\}$ are parallel in $M$
- Equivalent to $\forall i, j, \nabla \partial_i \partial_j = 0$
- Equivalent to $\forall i, j, k, \quad \Gamma^k_{ij} = 0$ (Christoffel symbols)

When there exists an affine coordinate system for $(M, \nabla)$, we say that $M$ is flat.
Metric connection : Special case of Levi-Civita connection
\( \nabla_{LC} = \nabla^{(0)} \)

Given \((M, g)\), there exists a unique metric connection, the Levi-Civita connection:

\[ \Gamma^k_{ij} = \frac{\partial_i g_{jk} + \partial_j g_{ki} - \partial_k g_{ij}}{2} \]

\[ \text{and we have } g(\nabla_{\partial_i} \partial_j, \partial_k) = \Gamma^k_{ij}. \]

\[ \text{Parallel transport of tangent vectors preserves the inner product.} \]

\[ \text{Therefore angles are kept, henceforth “parallel transport”} \]
Autoparallel submanifold

\[ N \subset M \text{ of } (M, N) \text{ is autoparallel} : \]

- Property on the tangent bundle \( TN \)

\[
\forall X, Y \in TN, \quad \nabla_X Y \in TN
\]

- Parallel (\( \nabla \))-transport of tangent vectors for \( N \) are tangent vectors of \( N \).

- Notion of “hyperplanes” in differential geometry

- For an affine connection with coordinate system \( \theta \), equivalent to an affine subspace of \( \theta \in \mathbb{R}^D \).
Differential-geometric structures: Summary

Manifold $M$

- Riemannian manifold
  - metric tensor $g$ (inner product)
  - (angle, orthogonality)

$(M, g)$

Differential structure $(M, g, \nabla)$

- Levi-Civita connection
  - $\nabla_{LC} = \nabla(g)$ (coefficients $\Gamma^k_{ij}$)
  - geodesics preserves $\langle \cdot, \cdot \rangle$
  - $\rho(P, Q)$ metric distance
    - (shortest paths)

Connection $\Pi$, covariant derivatives $\nabla$

- $\Pi \Leftrightarrow \nabla$
  - parallel transport
    - (flatness, autoparallel)

$(M, \nabla)$

Dual connections $(M, g, \nabla, \nabla^*)$
Dually affine connections

- Two affine connections $\Pi$ and $\Pi^*$ (and covariant derivatives $\nabla$ and $\nabla^*$)

- Property of inner product:

\[
\langle X, Y \rangle_g = \langle \Pi X, \Pi^* Y \rangle_g
\]

- Riemannian geometry: $\Pi = \Pi^*$
Dually affine connections: $e$-connection and $m$-connection

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

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

\[
\nabla^{(e)}_{\dot{\gamma}_e} \dot{\gamma}_e(t) = 0, \quad \nabla^{(m)}_{\dot{\gamma}_m} \dot{\gamma}_m(t) = 0
\]

Flat but not Riemannian flat: $e$-flat and $m$-flat.
Dually $\alpha$-affine connections

$\alpha \in \mathbb{R}, \quad \nabla(\alpha) = \frac{1 + \alpha}{2} \nabla + \frac{1 - \alpha}{2} \nabla^*$

- $\nabla = \nabla^e$ or $\nabla^m$
- Dually-coupled affine connections: $\nabla(\alpha)$ and $\nabla(-\alpha)$
- $\alpha = 0 : \nabla^{(0)} = \frac{\nabla + \nabla^*}{2} = \nabla_{LC}$, Levi-Civita metric connection (self-dual $\nabla^{(0)} = \nabla^{(0)*}$)
- 0-geometry is Riemannian geometry (often curved but not for isotropic Gaussians)
Dually flat orthogonal coordinate systems

- $\theta$- and $\eta$-coordinate systems
- partial derivatives: $\partial_i = \frac{\partial}{\partial \theta_i}$, $\partial^i = \frac{\partial}{\partial \eta^i}$
- $\langle \partial_i, \partial^j \rangle = \delta_{ij}$ (biorthogonal coordinate systems)
- metric-coupled connection:

$$X \langle Y, Z \rangle = \langle \nabla_X Y, Z \rangle + \langle Y, \nabla^*_X Z \rangle$$

- $\Gamma_{ijk}(\theta) = \Gamma^*_{ijk}(\eta) = 0$

This is key advantage over the Riemannian ($\nabla_{LC}$) structure: Geodesics are known in closed form with the affine coordinate systems. Line segments in either the $\theta$- or $\eta$-coordinate systems.
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 coordinate systems**: $\theta = \nabla F^*(\eta)$ and $\eta = \nabla F(\theta)$. 
- **Metric tensor $g$**: written equivalently using the two coordinate systems:

\[
g_{ij}(\theta) = \frac{\partial^2}{\partial \theta_i \partial \theta_j} F(\theta), \quad g^{ij}(\eta) = \frac{\partial^2}{\partial \eta_i \partial \eta_j} G(\eta)
\]

- **Divergence from Young’s inequality of convex conjugates**:

\[
D(P : Q) = F(\theta(P)) + F^*(\eta(Q)) - \langle \theta(P), \eta(Q) \rangle
\]

This is a Bregman divergence in disguise - :) ...

- **Exponential family**: $p(x|\theta) = \exp(\langle \theta, x \rangle - F(\theta))$, $F$ : cumulant function, $G$ : negative entropy
Geometry induced from a potential function

\( F \) a strictly convex potential function

\[
g_{ij} = \frac{\partial^2 F}{\partial_i \partial_j}
\]

\[
\Gamma^{(\alpha)}_{ijk} = \frac{1}{2} \left( \frac{\partial^3 F}{\partial_i \partial_j \partial_k} \right)
\]

Dually coupled \( \pm \alpha \)-connections (affine torsion-free, Kurose [6], 1994):

\[
\forall X, Y, Z \in V(M), \quad Xg(Y, Z) = g(\nabla^{(\alpha)}_X Y, Z) + g(Y, \nabla^{(\alpha)}_X Z)
\]

Curvature: \( \kappa = \frac{1-\alpha^2}{4} \)
Bregman divergences: An old friend from the optimization community
Bregman divergences

\[
D_F(p : q) = F(p) - F(q) - \langle p - q, \nabla F(q) \rangle
\]

includes...

- squared Euclidean distance: \( F(x) = \langle x, x \rangle \), and squared Mahalanobis \( F(x) = x \top Q x \) (only symmetric divergences)
- (extended) Kullback-Leibler divergence:
  \[
  F(x) = \sum_i x_i \log x_i - x_i \text{ (Shannon information)},
  \]
  \[
  eKL(p : q) = \sum_i \left( p_i \log \frac{p_i}{q_i} + q_i - p_i \right)
  \]
- \( F(x) = -\sum_i \log x_i \) (Burg information), Itakura-Saito divergence:
  \[
  IS(p : q) = \sum_i \left( \frac{p_i}{q_i} - \log \frac{p_i}{q_i} - 1 \right)
  \]
- and many others!
Bregman divergence: Geometric interpretation (I)

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

$$D_F(p : q) = F(p) - F(q) - \langle p - q, \nabla F(q) \rangle$$
Bregman divergence: Geometric interpretation (II)

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

$$B_f(p \parallel q) = f(p) - f(q) - (p - q)f'(q)$$

$B_f(. \parallel q)$: 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

\[ B(\theta_1 : \theta_2) = F(\theta_1) - F(\theta_2) - \langle \theta_1 - \theta_2, \nabla F(\theta_2) \rangle, \quad (1) \]

\[ = \int_{\theta_2}^{\theta_1} \langle \nabla F(t) - \nabla F(\theta_2), dt \rangle, \quad (2) \]

\[ = \int_{\eta_1}^{\eta_2} \langle \nabla F^*(t) - \nabla F^*(\eta_1), dt \rangle, \quad (3) \]

\[ = B^*(\eta_2 : \eta_1) \quad (4) \]
Dual Bregman divergences & canonical divergence [12]

For $P$ and $Q$ belonging to the same exponential families

$$\text{KL}(P : Q) = E_P \left[ \log \frac{p(x)}{q(x)} \right] \geq 0$$

$$= B_F(\theta_Q : \theta_P) = B_{F^*}(\eta_P : \eta_Q)$$

$$= F(\theta_Q) + F^*(\eta_P) - \langle \theta_Q, \eta_P \rangle$$

$$= A_F(\theta_Q : \eta_P) = A_{F^*}(\eta_P : \theta_Q)$$

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

$$\text{KL}(P : Q) = \int p(x) \log \frac{1}{q(x)} \mathrm{d}x - \int p(x) \log \frac{1}{p(x)} \mathrm{d}x$$

$$H^\times(P : Q) \quad H^\times(P : P)$$

Shannon cross-entropy and entropy of EF [12] :

$$H^\times(P : Q) = F(\theta_Q) - \langle \theta_Q, \nabla F(\theta_P) \rangle - E_P[k(x)]$$

$$H(P) = F(\theta_P) - \langle \theta_P, \nabla F(\theta_P) \rangle - E_P[k(x)]$$

$$H(P) = -F^*(\eta_P) - E_P[k(x)]$$
III. Principle of Maximum Entropy (MaxEnt)
Maximum entropy (MaxEnt)

Underconstrained optimization problem (Jaynes’s principle for maximum ignorance):

\[
\max_P H(p) = \sum_x p(x) \log \frac{1}{p(x)}
\]

\[
\sum_x p(x) t_i(x) = m_i, \quad \forall i \in \{1, \ldots, D\}
\]

\[
p(x) \geq 0, \quad \forall x \in \{1, \ldots, n\}
\]

\[
\sum_x p(x) = 1
\]

- Maximizing a concave function \((H)\) subject to linear constraints
- Convex optimization problem.
A more general setting for MaxEnt

Given a prior \( q \), find the closest distribution which satisfies the linear constraints:

\[
\min KL(p : q) = \sum_x p(x) \log \frac{p(x)}{q(x)}
\]

\[
\sum_x p(x)t_i(x) = m_i, \quad \forall i \in \{1, \ldots, D\}
\]

\[
p(x) \geq 0, \quad \forall x \in \{1, \ldots, n\}
\]

\[
\sum_x p(x) = 1
\]

\[\rightarrow\] Maximum entropy when \( q = \frac{1}{n} \), the uniform prior
\[ p^* = \min_p \text{KL}(p : q) \quad m\text{-flat} \]
Analytic solution: exponential families!

Using Lagrange multipliers $\theta$ with $t(x) = (t_1(x), \ldots, t_D(x))$:

$$p(x) = \frac{1}{Z(\theta)} \exp(\langle \theta, t(x) \rangle) q(x)$$

...but Lagrange multipliers usually not in explicit form.

- Canonical exponential families: $\exp(\langle \theta, t(x) \rangle - F(\theta) + k(x))$
- Prior $q$ gives the carrier measure $q(x) = e^{k(x)}$
- $Z(\theta)$ is the normalizer
- called Gibbs distribution, Maxwell-Boltzmann distribution in statistical mechanics
A toy example for MaxEnt

- A distribution $p$ with support $\mathbb{R}$ has $\mathbb{E}[X] = 3$ and $\mathbb{E}[X^2] = 25$. Which distribution should we choose for $p$?
- $t(x) = (x, x^2)$ defines the univariate Gaussian family of distributions.
- So we choose $p \sim N(\mu = 3, \sigma = 5)$
Another insightful proof

Any other distribution $p \neq p^*$ satisfying the constraints is such that $\text{KL}(p : q) > \text{KL}(p^* : q)$.
Consider the difference $\text{KL}(p : q) - \text{KL}(p^* : q)$:

$$
\begin{align*}
&= \sum_x p(x) \log \frac{p(x)}{q(x)} - \sum_x p^*(x) \log \frac{p^*(x)}{q(x)} \\
&= \sum_x p(x) \log \frac{p(x)}{q(x)} - \sum_x p(x) \log \frac{p^*(x)}{q(x)} \\
&= \sum_x p(x) \log \frac{p(x)}{p^*(x)} = \text{KL}(p : p^*) > 0
\end{align*}
$$

Pythagorean relation: $\text{KL}(p : q) = \text{KL}(p : p^*) + \text{KL}(p^* : q)$
An illustration...

 Prior $q$

 $KL(p : q)$

 $e$-projection

 $KL(p^* : q)$

 Affine subspace induced by constraints

 $p$ $p^* = \min_p KL(p : q)$ $m$-flat

 $KL(p : q) = KL(p : p^*) + KL(p^* : q)$

 Pythagoras’ theorem...
Computing I projections easily

- Project the prior $q$ onto
  \[ A = \{ p \mid \mathbb{E}_p[t_i(x)] = m_i, \ \forall i \in \{1, \ldots, D\} \}. \]  
  Let
  \[ A_i = \{ p \mid \mathbb{E}_p[t_i(x)] = m_i \} \]
- Let $t = 0$ and $p_0 = q$
- Repeat until convergence (within a threshold) :
  \[ p_{t+1} = \text{l-projection of } p_t \text{ onto } L_t \mod D \]
- 1D projection easy : Find $\theta_i$ such that $F_{\neq i}(\theta_i) = m_i$ (for example, using line search)
Cyclic (line search) 1D information projections
IV. Information projection
MLE as min KL : Information projection

- Empirical distribution:  
  \[ p_e(x) = \frac{1}{n} \sum_i \delta(x - x_i). \]
- \( p_e \) is absolutely continuous with respect to \( p_\theta(x) \)

\[
\begin{align*}
\min \text{KL}(p_e(x) : p_\theta(x)) &= \int p_e(x) \log p_e(x) \, dx - \int p_e(x) \log p_\theta(x) \, dx \\
&= \min -H(p_e) - E_{p_e}[\log p_\theta(x)]
\end{align*}
\]

\[
\equiv \max \left\{ \frac{1}{n} \sum_i \delta(x - x_i) \log p_\theta(x) \right\}
\]

\[
= \max \left\{ \frac{1}{n} \sum_i \log p_\theta(x_i) \right\} = \boxed{\text{MLE}}
\]
Projections: $e$-projection and $m$-projection

\[
\nabla^{(e)} = \nabla^{(1)}, \quad \nabla^{(m)} = \nabla^{(-1)}
\]

- **$e$-projection** $q$ is **unique** if $M \subseteq S$ is $m$-flat and minimizes the $m$-divergence $KL(q : p)$.
- **$m$-projection** $q$ is **unique** if $M \subseteq S$ is $e$-flat and minimizes the $e$-divergence $KL(p : q)$.

$KL$ and reverse $KL$ are $\alpha$-divergences for $\alpha = \pm 1\ldots$
Log-likelihood function

\[ l(\theta; X) = \frac{1}{n} \sum_{i=1}^{n} \log p(x_i|\theta) = \langle \log p(x|\theta) \rangle_{p_e} \]

Empirical distribution: \( p_e(X) = \frac{1}{n} \sum_{i=1}^{n} \delta(X - X(i)) \)

MLE = \textit{m-projection from } p_e \textit{ to the model submanifold}
Nested and curved exponential families

\[ \mathcal{P}(\theta) \text{ an exponential family} \]

- **nested EFs**: Fix some parameters \( \theta = (\theta_{\text{fixed}}, \theta_{\text{variable}}) \). Then \( \mathcal{P}_{\theta_{\text{fixed}}}(\theta_{\text{variable}}) \) is a nested exponential family. Get stratified EFs with uni-order EF easy to handle algorithmically (Legendre)

- **curved EFs**: \( \mathcal{C}(\gamma) \subseteq \mathcal{P}(\theta) \) embedded in \( \mathcal{P}(\theta) \). Example: \( \{N(\mu, \mu^2) \mid \mu \in \mathbb{R}\} \) is embedded into \( \{N(\mu, \sigma^2)\} \).
MLE for curved exponential families

Entropy $H(\theta) = -E_\theta[\log p(x|\theta)] = F(\theta) - \langle \theta, \nabla F(\theta) \rangle = -F^*(\eta)$

(when $k(x) = 0$, otherwise add $-E[k(x)]$).

$$D(p(\hat{\eta}) : p(\gamma)) = -H(\hat{\eta}) - \frac{1}{n} \log L(\gamma)$$

$$\max_{\gamma} L(\gamma) \equiv \min_{\gamma} D(p(\hat{\eta}) : p(\gamma))$$

$\hat{\gamma}$ is the $m$-projection of the observation point (with $\eta$-coordinate $\hat{\eta}$)
Illustration: MLE for curved exponential families

MLE observed point
\( \hat{\eta} = \frac{1}{n} \sum_{i=1}^{n} t(x_i) \)

\( \hat{\gamma} = \min_{\gamma} KL(p(\hat{\eta}) : p(\gamma)) \)

Fisher orthogonal

m-projection

curved exponential family

information loss, statistical curvature.
Simplifying a mixture model into a single component [18]

$m$-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.

\[
p = p_F(x|\theta) \\
p^* = p_F(x|\theta^*) \\
m = \sum_i w_i p_F(x|\theta_i) \\
\]

\[
p^* = \arg \min KL(m : p) \\
KL(m : p) = KL(p^* : p) + KL(m : p^*)
\]
Kullback-Leibler divergence and Fisher information

\[ \text{KL}(\theta + \Delta \theta : \theta) \approx \frac{1}{2} \theta^\top I(\theta) \theta \]

... square Mahalanobis induced locally by half squared Mahalanobis distance for the Fisher information matrix.

\[ g_{ij}(\theta_0) = \left. \frac{\partial^2}{\partial \theta^i \partial \theta^j} \right|_{\theta = \theta_0} \text{KL}(P(\theta) \| P(\theta_0)) \]

This holds for \( f \)-divergences \( \int p(x)f\left( \frac{q(x)}{p(x)} \right) d\nu(x) \) (that includes Kullback-Leibler divergence) : divergence inducing a metric proportional to Fisher information (Part II).
Additive Shannon/Rényi versus non-additive Tsallis entropies

- additive (Shannon-Rényi)
  \[ H(P \times Q) = H(P) + H(Q) \]

- non-additive (Tsallis)
  \[ T_q(X) = \frac{1}{q-1} (1 - \sum_i p_i^q) \]
  \[ T_q(X \times Y) = T_q(X) + T_q(Y) + (1 - q) T_q(X) T_q(Y) \]

- Both can be unified with Sharma-Mittal [13] 2-parameter family of entropies
- Cross-entropies and relative entropies
Part I : Summary

- Motivations
- Fisher information (Cramér-Rao lower bound) & sufficiency (1922)
- Differential geometry of population spaces :
  - Fisher-Rao geometry (Hotelling, 1930) : $g(\theta) = I(\theta)$
  - Dually-coupled connection geometry (1970’s-1980’s, Cencov, Amari, Kurose) : $(M, g, \nabla^{(\alpha)}, \nabla^{(-\alpha)})$, or $(M, g, T)$
  - Dually-flat manifold (Bregman divergence) : potential function and canonical divergence.
- Bregman divergences & canonical divergences in dually flat spaces (exhaustivity)
- Maximum entropy principle (exponential families)
- Information-geometric projections : MLE (and MLE in curved exponential families), and in mixture simplification.
Bibliography I

**Shun-Ichi Amari.**
Natural gradient works efficiently in learning.
*Neural computation, 10(2) :251–276, 1998.*

**Shun-ichi Amari and Hiroshi Nagaoka.**
*Methods of Information Geometry.*

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

**Vincent Garcia, Frank Nielsen, and Richard Nock.**
Levels of details for Gaussian mixture models.

**Harold Hotelling.**
Spaces of statistical parameters.

**Takashi Kurose.**
On the divergences of 1-conformally flat statistical manifolds.

**Meizhu Liu, Baba C. Vemuri, Shun-ichi Amari, and Frank Nielsen.**
Shape retrieval using hierarchical total Bregman soft clustering.
Bibliography II

Frank Nielsen.
Legendre transformation and information geometry.

Frank Nielsen.
k-MLE : A fast algorithm for learning statistical mixture models.

Frank Nielsen.
Cramér-rao lower bound and information geometry.

Frank Nielsen and Vincent Garcia.
*arXiv.org :0911.4863*.

Frank Nielsen and Richard Nock.
Entropies and cross-entropies of exponential families.

Frank Nielsen and Richard Nock.
A closed-form expression for the Sharma-Mittal entropy of exponential families.

Frank Nielsen and Richard Nock.
Visualizing hyperbolic Voronoi diagrams.
Frank Nielsen and Richard Nock.
Total Jensen divergences : Definition, properties and clustering.

Richard Nock, Frank Nielsen, and Shun-ichi Amari.
On conformal divergences and their population minimizers.
*CoRR*, abs/1311.5125, 2013.

Calyampudi Radhakrishna Rao.
Information and the accuracy attainable in the estimation of statistical parameters.

Olivier Schwander and Frank Nielsen.
Learning mixtures by simplifying kernel density estimators.