Bregman divergences, dual information geometry, and generalized comparative convexity

Frank Nielsen
Sony Computer Science Laboratories Inc.
Tokyo, Japan

https://franknielsen.github.io/
Outline

• Bregman divergences

• Dual information geometry & Bregman manifolds

• Generalized convexity and designing divergences from convexity gaps
Part I.

Bregman divergences:
- Legendre-Fenchel transformation (dual parameterization)
- Fenchel-Young divergences (mixed parameterization)
- Statistical divergences, statistical models & Bregman divergences
Bregman divergences

• $F(\theta)$: strictly convex and differentiable convex function on an open convex domain $\Theta$

• Design the **Bregman divergence** as the vertical gap between $F(\theta_1)$ and the linear approximation of $F(\theta)$ at $\theta_2$ evaluated at $\theta_1$:

$$B_F(\theta_1 : \theta_2) = F(\theta_1) - \left( F(\theta_2) + (\theta_2 - \theta_1)^T \nabla F(\theta_2) \right)$$

$$= F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2)^T \nabla F(\theta_2)$$

[Bregman 1967]
Bregman divergences: Some properties

- **Positive-definite:**
  - $B_F(\theta_1 : \theta_2) > 0$ when $\theta_1 \neq \theta_2$
  
  - $B_F(\theta_1 : \theta_2) = 0$ if and only if $\theta_1 = \theta_2$

- **Symmetric** only for generalized squared Euclidean/Mahalanobis distance, **asymmetric** otherwise [N+ 2007]

  \[
  Q > 0 \quad D^2_Q(\theta_1, \theta_2) = B_{F_Q}(\theta_1, \theta_2) = (\theta_2 - \theta_1)^\top Q (\theta_2 - \theta_1), \quad F_Q(x) = x^\top Q x
  \]

  \[
  D^2_E(\theta_1, \theta_2)^2 = ||\theta_1 - \theta_2||_2^2 = D^2_I(\theta_1, \theta_2)
  \]

  \[
  M^2_{\Sigma} [\mathcal{N}(\mu_1, \Sigma), \mathcal{N}(\mu_2, \Sigma)] = D^2_{\Sigma^{-1}}(\mu_1, \mu_2) = \Delta \mu^\top \Sigma^{-1} \Delta \mu
  \]

- **Does not** satisfy the triangle inequality of metric distances
- **Smooth/differentiable** w.r.t. parameters $\Rightarrow$ **divergences** (contrast functions)

• Bregman divergence (BD) can be interpreted as the mean-value remainder of a first-order Taylor expansion of $F(\theta)$ at $\theta_2$:

$$F(\theta_1) = \underbrace{F(\theta_2) + (\theta_1 - \theta_2)^\top \nabla F(\theta_2)}_{\text{first-order Taylor expansion}} + \underbrace{R_F(\theta_1 : \theta_2)}_{\text{Taylor remainder}}$$

$$R_F(\theta_1 : \theta_2) = \frac{1}{2}(\theta_2 - \theta_1)^\top \nabla^2 F(\xi)(\theta_2 - \theta_1), \quad \xi \in [\theta_1 \theta_2]$$

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

• Since $F$ is strictly convex, the Hessian is positive-definite:

$$\nabla^2 F(\theta) \succ 0 \iff \forall x \neq 0, x^\top \nabla^2 F(\theta)x > 0$$

• and this proves that BDs are positive-definite:

$$R_F(\theta_1 : \theta_2) = \frac{1}{2}(\theta_2 - \theta_1)^\top \nabla^2 F(\xi)(\theta_2 - \theta_1) \geq 0$$
Scalar and separable Bregman divergences

- D-variate Bregman divergence w.r.t. parameter $\theta = (\theta^1, \ldots, \theta^D)$

- **Separable**: Bregman generator is sum of univariate/scalar Bregman generators:

$$F(\theta) := \sum_{i=1}^{D} F_i(\theta^i)$$

$$B_F(\theta_1 : \theta_2) = \sum_{i=1}^{D} B_{F_i}(\theta^i_1 : \theta^i_2)$$

- For example, generalized square Euclidean distance with diagonal matrix $Q$

$$Q = \text{diag}(q_1, \ldots, q_d)$$

$$B_{F_Q}(\theta_1, \theta_2) = \sum_{i=1}^{D} q_i (\theta^i_2 - \theta^i_1)^2 = \sum_{i=1}^{D} B_{F_{q_i}}(\theta^i_1 : \theta^i_2)$$

$$F_{q_i}(x) = q_i x^2$$
Extended Kullback-Leibler divergence

- **Extended Kullback-Leibler divergence** (eKL) is a D-dimensional separable Bregman divergence induced by the **Shannon negentropy** Bregman generator:

\[
D_{eKL}[p_{\theta_1} : p_{\theta_2}] := \sum_{i=1}^{D} \theta_1^i \log \frac{\theta_1^i}{\theta_2^i} + \theta_2^i - \theta_1^i =: B_{FeKL}(\theta_1 : \theta_2)
\]

\[
F_{eKL}(\theta) := \sum_{i=1}^{D} \theta_i \log \theta_i - \theta_i
\]

- \(p^+\) means a **positive measure** (not necessarily normalized to a probability)
- When \(p^+\) is normalized:

\[
D_{eKL}[p_{\lambda_1} : p_{\lambda_2}] := \sum_{i=1}^{D} \theta_1^i \log \frac{\theta_1^i}{\theta_2^i}
\]

- eKL divergence also called **extended relative entropy** in information theory

Lafferty, Lebanon, Boosting and maximum likelihood for exponential models, NeurIPS 14 (2002)
Discrete Kullback-Leibler divergence: A non-separable Bregman divergence

- The KLD between two **categorical distributions** a.k.a. *multinoulli* amounts to a **non-separable Bregman divergence** on the **natural parameters** of the multinoulli distributions interpreted as an **exponential family**.

\[ p_\lambda = (p_{\lambda_1}^1, \ldots, p_{\lambda_d}^d) \in \Delta_{d-1}^\circ, \quad \sum_{i=1}^d p_{\lambda_i}^i = 1 \]

\[ D_{KL}[p_{\lambda_1} : p_{\lambda_2}] := \sum_{i=1}^D \lambda_1^i \log \frac{\lambda_1^i}{\lambda_2^i} =: B_{FKL}(\theta_1 : \theta_2) \]

\[ \theta^i = \log \frac{\lambda_i^i}{\lambda_D}, \quad i \in \{1, \ldots, D = d - 1\} \]

\[ F_{KL}(\theta) = \log(1 + \sum_{i=1}^D \exp(\theta_i)) =: \text{LogSumExp}_{+}(\theta_1, \ldots, \theta_D) \]

LogSumExp is only convex but LogSumExp_{+} is strictly convex  

Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities, Entropy, 18(12), 2016
Legendre-Fenchel transformation

- Consider a Bregman generator of **Legendre-type** (proper, lower semi-continuous). Then its **convex conjugate** obtained from the **Legendre-Fenchel transformation** is a Bregman generator of Legendre type.

\[
F^*(\eta) = \sup_{\theta \in \Theta} \{\theta^T \eta - F(\theta)\} = -\inf_{\theta \in \Theta} \{F(\theta) - \theta^T \eta\}
\]

- Legendre-Fenchel transformation applies to any multivariate function
- Analogy of the Halfspace/Vertex representation of the **epigraph** of \(F\)
- Fenchel-Moreau’s **biconjugation theorem** for \(F\) of Legendre-type: \(F = (F^*)^*\)

Concave programming:

\[
F^*(\eta) = \sup_{\theta \in \Theta} \{\theta^T \eta - F(\theta)\} = \sup_{\theta \in \Theta} \{E(\theta)\}
\]

\[
\nabla E(\theta) = \eta - \nabla F(\theta) = 0 \Rightarrow \eta = \nabla F(\theta)
\]

[Touchette 2005] Legendre-Fenchel transforms in a nutshell  
[N 2010] Legendre transformation and information geometry
Reading the Legendre-Fenchel transformation

- Legendre-Fenchel transformation also called the **slope transform**

\[
F^*(\eta) = \sup_{\theta \in \Theta} \{ \theta^T \eta - F(\theta) \}
\]

\[
F(\theta) = \exp(\theta) \\
\eta = F'(\theta) = \exp(\theta) \\
\theta = F'^{-1}(\eta) = \log \eta = F^*(\eta) \\
F^*(\eta) = \theta \eta - F(\theta) = \eta \log \eta - \eta
\]

(Here, F was chosen as the cumulant function of the Poisson distributions)
Legendre-Fenchel transform:
Mixed coordinates and Fenchel-Young divergence

- **Dual parameterizations** of epigraph: \( \theta = \nabla F^*(\eta) \) and \( \eta = \nabla F(\theta) \)
- Convex conjugate expressed as: \( F^*(\eta) = \eta^\top \nabla F^*(\eta) - F(\nabla F^*(\eta)) \)
- To get in closed form the convex conjugate \( F^* \), we need \( \nabla F^*(\eta) \), i.e., invert \( \nabla F(\theta) \)

- **Fenchel-Young inequality**: \( F(\theta_1) + F^*(\eta_2) \geq \theta_1^\top \eta_2 \)
  with equality if and only if \( \eta_2 = \nabla F(\theta_1) \)

- **Fenchel-Young divergence** use mixed parameterization \( \theta/\eta \):
  \[
  Y_{F,F^*}(\theta_1 : \eta_2) := F(\theta_1) + F^*(\eta_2) - \theta_1^\top \eta_2 = Y_{F*,F}(\eta_2, \theta_1)
  \]
Dual Bregman and dual Fenchel divergences

• Identity of dual Bregman divergences: \( B_F(\theta_1 : \theta_2) = B_{F^*}(\eta_2 : \eta_1) \)

• In general, dual or reverse divergence: \( D^*(\theta_1 : \theta_2) := D(\theta_2 : \theta_1) \)

• Primal, dual or mixed parameterizations of Bregman divergences:

\[
B_F(\theta_1 : \theta_2) = Y_{F,F^*}(\theta_1 : \eta_2) = Y_{F^*,F}(\eta_2, \theta_1) = B_{F^*}(\eta_2 : \eta_1)
\]
3-parameter identity of Bregman divergences

- Generalize the **law of cosines** for the squared Euclidean distance

\[ B_F(\theta_1 : \theta_2) = B_F(\theta_1 : \theta_3) + B_F(\theta_3 : \theta_2) - (\theta_1 - \theta_3)^\top (\nabla F(\theta_2) - \nabla F(\theta_3)) \geq 0 \]

- Yields a **generalization of the Pythagorean theorem** when

\[ (\theta_1 - \theta_3)^\top (\nabla F(\theta_2) - \nabla F(\theta_3)) = 0 \]

On geodesic triangles with right angles in a dually flat space, Progress in Information Geometry: Theory and Applications, 2021
4-parameter identity of Bregman divergences

- Parallelogram identity

\[ B_F(\theta_1 : \theta) + B_F(\theta_2 : \theta) = B_F \left( \frac{\theta_1 + \theta_2}{2} \right) + B_F \left( \frac{\theta_2 + \theta_1}{2} \right) + 2B_F \left( \frac{\theta_1 + \theta_2}{2} : \theta \right) \]

- In Euclidean geometry:

\[ 2AB^2 + 2BC^2 = AC^2 + BD^2 \]
Symmetrized Bregman divergence: Geometric reading

\[ \eta = \nabla F(\theta) \]

\[ \nabla F(\theta) = \nabla F^*^{-1}(\theta) \]

\[ B_F(\theta_1 : \theta_2) = \int_{\theta_2}^{\theta_1} (F'(\theta) - F'(\theta_2))d\theta \]

\[ S_F(\theta_1, \theta_2) = B_F(\theta_1 : \theta_2) + B_F(\theta_2 : \theta_1) \]

\[ B_{F^*}(\eta_2 : \theta_1) = \int_{\eta_1}^{\eta_2} (F^*'(\eta) - F^*'(\eta_1))d\eta \]

\[ = B_F(\theta_1 : \theta_2) + B_{F^*}(\eta_1 : \eta_2) \]

\[ = (\theta_1 - \theta_2) \mathbf{^T} (\eta_1 - \eta_2) \]

[arXiv:2107.05901]
Statistical divergences between parametric models = parameter divergences

Statistical divergences between densities of a parametric model $\mathcal{F} = \{f_\theta(x)\}_\theta$ amount equivalently to (parameter) divergences between corresponding parameters:

$$\mathcal{D}[f_{\theta_1} : f_{\theta_2}] =: D_M(\theta_1 : \theta_2)$$

For which statistical models and statistical divergences, do we obtain $D_M(\theta_1 : \theta_2)$ as a Bregman divergence?
Example 1: Natural exponential family models

- Parametric model \( \mathcal{E} = \{ e_{\theta}(x) \}_\theta \) with densities \( e_{\theta}(x) = \exp \left( \sum_{i=1}^{D} t_i(x) \theta_i - F(\theta) + k(x) \right) \)

- Examples of **natural exponential families**:
  - Exponential distributions (continuous): p.d.f. \( \lambda e^{-\lambda x} \quad x \geq 0 \)
  - Poisson distributions (discrete): p.m.f. \( \Pr(X=k) = \frac{\lambda^k e^{-\lambda}}{k!} \)

- Examples of **exponential families** with density \( e_{\lambda}(x) = \exp \left( \sum_{i=1}^{D} t_i(x) \theta_i(\lambda) - F(\theta) + k(x) \right) \)
  - Gaussian distributions once reparameterized with natural parameters \( \theta(\lambda) = \theta(\mu, \sigma^2) \)
- We have \( D_{\text{KL}}[e_{\theta_1} : e_{\theta_2}] = B_F^*(\theta_1 : \theta_2) = B_F(\theta_2 : \theta_1) \) with Bregman generator:

  \[
  F_{\mathcal{E}}(\theta) = \log \left( \int \exp \left( \sum_{i=1}^{D} t_i(x) \theta_i + k(x) \right) d\mu(x) \right)
  \]
Example 2: Mixture family models

- Let $1, p_0(x), \ldots, p_D(x)$ be $(D+2)$ linearly independent densities

- Mixture family $\mathcal{M} = \{m_\theta(x)\}_\theta$ with densities:
  
  $$m_\theta(x) = \sum_{i=1}^{D} w_i p_i(x) + \left(1 - \sum_{i=1}^{D} w_i\right)p_0(x)$$

- We have:
  
  $$D_{KL}[m_{\theta_1} : m_{\theta_2}] = B_{\mathcal{M}}(\theta_1 : \theta_2)$$

  $$\theta = (w_1, \ldots, w_D)$$

  Information geometry/reconstruction

- with the Bregman generator = Shannon negentropy:

  $$F_{\mathcal{M}}(\theta) = \int m_\theta(x) \log m_\theta(x) d\mu(x)$$

  Usually $F_{\mathcal{M}}(\theta)$ not in closed-form...

  But 2-mixture family of Cauchy distributions has closed-form!

The dually flat information geometry of the mixture family of two prescribed Cauchy components, arXiv:2104.13801
Example 3: q-Gaussians and statistical divergence

• The set of **Cauchy distributions**
  \[ C := \left\{ p_{\lambda}(x) := \frac{s}{\pi (s^2 + (x-1)^2)}, \quad \lambda := (l, s) \in H := \mathbb{R} \times \mathbb{R}_+ \right\} \]
  form a **q-Gaussian exponential family** for q=2

• Deformed exponential family generalize exponential family with deformed log/exp functions

• Cumulant function of the Cauchy 2-Gaussian family:
  \[ F(\theta) = -\frac{\pi^2}{\theta_2} - \frac{\theta_1^2}{4\theta_2} - 1. \]
  \[ (\theta_1, \theta_2) = \left( 2\pi \frac{1}{s}, -\frac{\pi}{s} \right) \]

• The following statistical divergence between 2 Cauchy distributions amount to a Bregman divergence:
  \[ D_{\text{flat}}[p_{\lambda_1} : p_{\lambda_2}] := \frac{1}{\int p_{\lambda_2}^2(x) dx} \left( \int \frac{p_{\lambda_2}^2(x)}{p_{\lambda_1}(x)} dx - 1 \right) = B_F(\theta_1 : \theta_2) \]

• Bregman generator is the **q-free energy** for q=2

*On Voronoi diagrams on the information-geometric Cauchy manifolds, Entropy 22.7 (2020)*
Information geometry & Bregman divergences

• Bregman divergences are **canonical divergences** of dually flat spaces (Bregman manifolds)

• Information geometry gives a principle to **reconstruct** the statistical divergence corresponding to a Bregman divergence for a Bregman generator $F(f_{\theta})$, and not the converse

\[
F_F(\theta) = \log \left( \int \exp \left( \sum_{i=1}^{D} t_i(x)\theta_i + k(x) \right) d\mu(x) \right)
\]

$B_F(\theta_1 : \theta_2)$

$D_{KL}^*[p_1 : p_2] = D_{KL}[p_2 : p_1]

\[
D_F^*[e_{\theta_1} : e_{\theta_2}] = D_{KL}[e_{\theta_2} : e_{\theta_1}]
\]

*An elementary introduction to information geometry.* Entropy 22.10 (2020)
Class of Bregman generators modulo affine terms
& KLD between exponential family densities expressed as log-ratio

• Bregman generators are strictly convex and differentiable convex functions defined modulo affine terms: \( B_F = B_G \) iff. \( F(\theta) = G(\theta) + A\theta + b \)

• Choose for any \( \omega \) in the support of the exponential family the Bregman generator:

\[
F_\omega(\theta) := -\log(p_\theta(\omega)) = F(\theta) - (\theta^T t(\omega) + k(\omega)) \]

affine term in \( \theta \)

• We get: \( D_{KL}[p_{\lambda_1} : p_{\lambda_2}] = \log \left( \frac{p_{\lambda_1}(\omega)}{p_{\lambda_2}(\omega)} \right) + (\theta(\lambda_2) - \theta(\lambda_1))^T (t(\omega) - \nabla F(\theta(\lambda_1))) \), \( \forall \omega \in \mathcal{X} \)

• By choosing \( s \) points: \( D_{KL}[p_{\lambda_1} : p_{\lambda_2}] = \frac{1}{s} \sum_{i=1}^{s} \log \left( \frac{p_{\lambda_1}(\omega_i)}{p_{\lambda_2}(\omega_i)} \right) \) such that \( \frac{1}{s} \sum_{i=1}^{s} t(\omega_i) = E_{p_{\lambda_1}}[t(x)] \)

Computing Statistical Divergences with Sigma Points. GSI 2021
Cumulant-free closed-form formulas for some common (dis)similarities between densities of an exponential family,
Part II.
Information geometry & Bregman manifolds

The fabric of information geometry and the untangling of its geometry, divergence, statistical models.
Motivation & history of information geometry

• **Information geometry** studies the *geometric structures* and statistical *invariance principles* (*sufficient statistics, Markov kernels*) of a family of probability distributions (=statistical model) and demonstrate their use in information sciences (statistics, ML).

• The newly revealed geometric structures (e.g., dually flat space) can *also* be used in *non-statistical contexts* (e.g., mathematical programming)

• Born as a mathematical curiosity! Use **Fisher information matrix** as a Riemannian metric = Fisher metric [Hotelling 1930] [Rao 1945]

• Decouple metric tensor with Levi-Civita connection, consider a family of affine connections [Chenstov 1960-1970’s] : **Geometrostatistics**

• **Statistical curvature**, Efron’s e-connection, Dawid’s m-connection [Efron 1975]

• Consider **dual torsion-free affine connections coupled to the metric**, explicit *α*-structures [Amari 1980’s] (Amari-Chentsov totally symmetric cubic tensor)

• Non-parametric information geometry [Pistone 1990’s], quantum information geometry, algebraic statistics, geometric science of information, etc.
Part II.A
- Fisher-Riemannian geometry
Fisher information matrix (FIM)

- A parametric family of distributions \( \mathcal{P} = \{p_\theta\}_{\theta \in \Theta} \)
- **Fisher information matrix** is positive-semidefinite matrix:

  \[
  s(\theta) := \nabla_\theta \log p_\theta(x) \quad I_X(\theta) = \text{Cov}(s_\theta) \quad X = (x_1, \ldots, x_D)^\top \sim p_\theta
  \]

  Score:

- Under **independence**, Fisher information is **additive**:

  \[
  Y = (Y_1, \ldots, Y_n)_{\sim \text{iid} p_\theta} \quad \Rightarrow \quad I_Y(\theta) = n \ I_X(\theta)
  \]

- Under **regularity conditions I (FIM type 1)**:

  \[
  I_1(\theta) = E_{p_\theta} \left[ (\nabla_\theta \log p_\theta)(\nabla_\theta \log p_\theta)^\top \right]
  \]

- Under **regularity conditions II (FIM type 2)**:

  \[
  I_2(\theta) = -E_{p_\theta} \left[ \nabla^2_\theta \log p_\theta \right]
  \]

- **FIM can be singular** (hierarchical models like mixtures, neural networks in ML)
- **FIM can be infinite** (irregular models, e.g., support depend on parameters)

N., Cramér-Rao lower bound and information geometry, Connected at Infinity II, 2013
Soen and Sun, On the Variance of the Fisher Information for Deep Learning, NeurIPS 2021
Key concept: Sufficient statistics

• A **statistic** is a function of a random vector (e.g., mean, variance)

• A **sufficient statistic** collect and concentrate from a random sample all necessary information for estimating the parameters.

  Informally, a statistical lossless compression scheme...

• **Definition:** conditional distribution of $X$ given $t$ **does not depend** on $\theta$

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

• **Fisher-Neyman factorization theorem:** Statistic $t(x)$ sufficient iff. the density can be decomposed as:

  $$p(x; \lambda) = a(x)b_\lambda(t(x))$$

Natural exponential families (NEF)

- Consider a positive measure $\mu$ (usually counting or Lebesgue)
- A natural exponential family is a parametric family of densities that write as

$$p(x; \theta) = \exp(\theta x - F(\theta))$$

where $F$ is real-analytic, strictly convex and differentiable:

$$F(\theta) = \log \int \exp(\theta x) d\mu(x)$$

Natural parameter space $\Theta = \{ \theta : \int \exp(\theta x) d\mu(x) < \infty \}$

F: Log-normalizer (also known as partition function, cumulant function, etc.)
Exponential families (from Natural EFs to EFs)

- Consider a **(sufficient) statistic** \( t(x) \)
- Consider an **additional carrier measure term** \( k(x) \)
- Consider an **inner product** between \( t(x) \) and \( \theta \)

  (usual scalar/dot product)

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

**Properties:**

\[
E[t(X)] = \nabla F(\theta)
\]
\[
Cov[t(X)] = \nabla^2 F(\theta) = I(\theta)
\]

(Hessian of \(-\log p_\theta(x))

(FIM type 2)

**Exponential families have finite moments of any order**
Many common distributions are exponential families in disguise

Tojo and Yoshino, On a method to construct exponential families by representation theory, GSI 2019 (Springer)
Bhattacharyya arc: Likelihood Ratio Exponential Family

- **Bhattacharyya arc** or **Hellinger arc** induced by two mutually absolutely continuous distributions \( p \) and \( q \) (same support \( \mathcal{X} \)):

\[
\mathcal{E}(p, q) := \left\{ p_{\lambda}(x) := \frac{p^{1-\lambda}(x)q^{\lambda}(x)}{Z^G_{\lambda}(p, q)}, \quad \lambda \in (0, 1) \right\}
\]

\[
Z^G_{\lambda}(p, q) := \int_{\mathcal{X}} p^{1-\lambda}(x)q^{\lambda}(x)d\mu(x)
\]

- Log-normalizer \( F(\lambda) \) (aka cumulant generating function, log partition function):

\[
p_{\lambda}(x) = \frac{p_0^{1-\lambda}(x)p_1^{\lambda}(x)}{Z^G_{\lambda}(p, q)} = p_0(x) \exp \left( \lambda \log \left( \frac{p_1(x)}{p_0(x)} \right) - \log Z^G_{\lambda}(p, q) \right) = \exp(\lambda t(x) - F(\lambda) + k(x))
\]

\[
F(\lambda) := \log(Z^G_{\lambda}(p, q)) = \log \left( \int_{\mathcal{X}} p^{1-\lambda}(x)q^{\lambda}(x)d\mu(x) \right) =: -D^\text{Bhat}_{\lambda}[p : q]
\]

- **Bhattacharyya arc** (geometric mixtures) = **1D exponential family**:

Generalizing the Geometric Annealing Path using Power Means, UAI 2021
Likelihood Ratio Exponential Families, NeurIPS Workshop on Deep Learning through Information Geometry 2020
Rao’s length distance (Riemannian distance)

\[(M,g)\text{ Riemannian manifold: Parameter space equipped with the Fisher information metric}\]

\[d(\theta^1, \theta^2) = \min_{\theta(t)} \int_{t_1}^{t_2} \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{p} g_{ij}(\theta(t)) \frac{d\theta_i(t)}{dt} \frac{d\theta_j(t)}{dt}} dt.\]

**Invariant** under smooth & bijective reparameterization

E.g., normal family: \((\mu,\sigma), (\mu,\sigma^2), (\mu,\log \sigma)\)

FIM is **covariant** under reparameterization

Rao distance in the probability simplex:

\[\rho_{FHR}(p, q) = 2 \arccos \left( \sum_{i=0}^{d} \sqrt{\lambda_p^i \lambda_q^i} \right)\]
Rao’s distance between 1D normal distributions

Fisher information metric becomes the Poincare upper plane metric after scale change of variable

FIM of normals

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

\[ ds_F^2 = \frac{d\mu^2 + 2d\sigma^2}{\sigma^2} \]

\[ d_F((\mu_1, \sigma_1), (\mu_2, \sigma_2)) = \sqrt{2d_H \left( \left( \frac{\mu_1}{\sqrt{2}}, \sigma_1 \right), \left( \frac{\mu_2}{\sqrt{2}}, \sigma_2 \right) \right)} \]

\[ \text{dist}(\langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle) = \arccosh \left( 1 + \frac{(x_2 - x_1)^2 + (y_2 - y_1)^2}{2y_1y_2} \right) \]

On Voronoi diagrams on the information-geometric Cauchy manifolds, Entropy 22.7 (2020)
In practice, calculating Rao’s distance may be difficult!

E.g., no closed form of Rao’s distance between multivariate normals

1. Need to solve the Ordinary Differential Equation (ODE) for find the geodesic:

\[ d(\theta^1, \theta^2) = \min_{\theta(t)} \int_{t_1}^{t_2} \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{p} g_{ij}(\theta(t)) \left( \frac{d\theta_i(t)}{dt} \frac{d\theta_j(t)}{dt} \right) dt}. \]

\[
\frac{d^2 \theta_k}{dt^2} + \sum_{i=1}^{p} \sum_{j=1}^{p} \Gamma_{ij}^{k} \frac{d\theta_i}{dt} \frac{d\theta_j}{dt} = 0, \quad k = 1, \ldots, p,
\]

\[
\Gamma_{ij}^{k} = \frac{1}{2} \sum_{m=1}^{p} \left( \frac{\partial g_{im}(\theta)}{\partial \theta_j} + \frac{\partial g_{jm}(\theta)}{\partial \theta_i} - \frac{\partial g_{ij}(\theta)}{\partial \theta_m} \right) g^{mk}(\theta), \quad i, j, k = 1, \ldots, p,
\]

2. Need to **integrate** the infinitesimal length elements ds along the geodesics
Approximating geodesics for MVNs: geodesic shooting

Algorithm 1: Shooting method for minimal geodesics on $\mathcal{N}(n)$

Given: Initial point $P_0 = (\mu_0, \Sigma_0)$, final point $P_1 = (\mu_1, \Sigma_1)$.
Output: Minimal geodesic $P(t) = (\mu(t), \Sigma(t))$, $t \in [0, 1]$, such that $P(1) = (\mu_1, \Sigma_1)$.
Initialization: Choose initial velocities $V(0) = (\dot{\mu}(0), \dot{\Sigma}(0))$ (e.g., zeros), initial values for $\epsilon$ ($10^{-5}$), error = $10^6$.

while error $\geq \epsilon$ do

 Numerically integrate the geodesic equations (13), (14) for given initial conditions $(\mu_0, \Sigma_0, \dot{\mu}_0, \dot{\Sigma}_0)$ from $t = 0$ to $t = 1$.
 Denote the solution by $(\mu(t), \Sigma(t))$;
 Set $W(1) = (W_\mu(1), W_\Sigma(1)) = (\mu(1) - \mu(1), \Sigma(1) - \Sigma(1))$;
 Calculate error = $\|W(1)\|_{P_1} = \sqrt{W_\mu(1)^T \Sigma_1^{-1} W_\mu(1) + \frac{1}{2} \text{tr}(\Sigma_1^{-1} W_\Sigma(1)^2)}$;
 Numerically integrate the parallel transport equations (18) and (19) for given trajectory $(\mu(t), \Sigma(t))$ and final velocities $W(1)$, backward in time from $t = 1$ to $t = 0$;
 Numerically calculate Jacobi field $J(1)$ from (22),
 $J(1) = \frac{\exp_{P_0}(V(0) + \alpha W(0)) - \exp_{P_0}(V(0))}{\alpha}$, where $\alpha$ is sufficiently small value and we use $s = \frac{J(1)}{\|W(0)\|_{P_0}}$.

 Determine proper update size $s$:
 $s_1 = \frac{\langle J(1) \rangle_{P_1}}{\|J(1)\|_{P_1}}$
 if $\|W(1)\|_{P_1} > 0.05$ then
 $s = 0.05/\|W(1)\|_{P_1} s_1$;
 else
 $s = s_1$;
 end if
 $V(0) \leftarrow V(0) + s W(0)$;
end while

ODE with boundary value conditions

Minyeon Han · F.C. Park, DTI Segmentation and Fiber Tracking Using Metrics on Multivariate Normal Distributions, 2014
Part II.B
- Dual information geometry
Another look at Riemannian geodesics: Connections

• Riemannian geodesics are **locally minimizing length curves**

• The *general definition* of geodesics is wrt. to an **affine connection**:
  For Riemannian geodesics, the default connection = **Levi-Civita connection**.
  This special Levi-Civita connection is derived from the metric tensor g.

• A geodesic $\gamma(t)$ with respect to a connection $\nabla$ is an $\nabla$-**autoparallel curve**
  (straight free fall particle in physics):
  $$\nabla_{\dot{\gamma}}\dot{\gamma} = 0,$$
  $$\dot{\gamma} = \frac{d}{dt}\gamma(t)$$

  where $\nabla_X T$ is the **covariant derivative** of a tensor $T$ wrt. a vector field $X$
What makes the Levi-Civita connection so special?

• A connection is described by **Christoffel symbols** (functions $\Gamma$), and the geodesics is described by this ODE:

$$\ddot{\gamma}(t) + \Gamma^k_{ij} \dot{\gamma}(t) \dot{\gamma}(t) = 0, \quad \gamma^l(t) = x^l \circ \gamma(t),$$

An affine connection defines how to **parallel transport** a vector from one tangent plane to another tangent plane.

• **Fundamental theorem of Riemann geometry:**

Levi-Civita connection is the **unique torsion-free metric connection** induced by the metric tensor $g$:

$$LC \Gamma^k_{ij} \Sigma = \frac{1}{2} g^{kl} (\partial_i g_{il} + \partial_j g_{il} - \partial_l g_{ij})$$

$$\left< u, v \right>_c(t) = \left< \prod_{c(0) \rightarrow c(t)} u, \prod_{c(0) \rightarrow c(t)} v \right>_c(t) \quad \forall t.$$
A connection is flat if there exists locally a coordinate system such that the Christoffel symbols are all zero: Geodesics plotted in that coordinate system are line segments.

Torsion tensor \( T(X, Y) := \nabla_X Y - \nabla_Y X - [X, Y] \)

Connections that differ only on torsions yield the same geodesics.
Dualistic information geometry: \( (M, g, \nabla, \nabla^*) \)

- Given an affine torsion-free connection \( \nabla \) and a metric \( g \), we can build a unique dual affine torsion-free connection: the dual connection \( \nabla^* \) such that the metric (inner product) is preserved by the primal and dual parallel transports:

\[
\langle u, v \rangle_{c(0)} = \left\langle \prod_{c(0) \to c(t)} \nabla u, \prod_{c(0) \to c(t)} \nabla^* v \right\rangle_{c(t)}.
\]

- This amounts to say that \( \nabla^* \) is defined uniquely by

\[
Xg(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z),
\]

meaning

\[
X_p g_p (Y_p, Z_p) = g_p ((\nabla_X Y)_p, Z_p) + g_p (Y_p, (\nabla^*_X Z)_p).
\]

- The dual of a dual connection is the primal connection: \( (\nabla^*)^* = \nabla \).

An elementary introduction to information geometry, Entropy 22.10 (2020)
Amari/Chentsov’s $\alpha$-structures

\[ \{(\mathcal{P}, p g, p \nabla^{-\alpha}, p \nabla^{+\alpha})\}_{\alpha \in \mathbb{R}} \]

- Regular statistical parametric models (identifiable and finite positive-definite FIM) \( \mathcal{P} := \{p_\theta(x)\}_{\theta \in \Theta} \)
- Amari’s $\alpha$-connections \( p \Gamma_{ij,k}^\alpha(\theta) := E_\theta \left[ \left( \partial_i \partial_j l + \frac{1 - \alpha}{2} \partial_i l \partial_j l \right) (\partial_k l) \right] \)
- \( l(\theta; x):= \log L(\theta; x) = \log p_\theta(x) \)
- 0-connection is **Fisher Levi-Civita connection**
- 1-connection is **exponential connection** (flat for exponential families)
- -1 connection is **mixture connection** (flat for mixture families)

Lauritzen’ statistical manifolds: Cubic tensor

Beware: Apply also to non-statistical contexts too! \((M, g, C)\)

Dualistic structure with metric tensor \(g\) and cubic tensor \(C\)

\[
C(X, Y, Z) := \langle \nabla_X Y - \nabla^*_X Y, Z \rangle
\]

\[
C_{ijk} = C(\partial_i, \partial_j, \partial_k) = \langle \nabla_{\partial_i} \partial_j - \nabla^*_{\partial_i} \partial_j, \partial_k \rangle
\]

C is totally symmetric (= components invariant by index permutation)

In a local basis:

\[
C_{ijk} := \Gamma^k_{ij} - \Gamma^*_{ij}
\]

\(LC \nabla g = 0\)  \(\rightarrow\) Levi-Civita connection is self-dual with respect to the metric!

Lauritzen, Statistical manifolds, Differential geometry in statistical inference 10 (1987)
Eguchi’s Information geometry of divergences

• **Reverse/dual parameter divergence** (reference duality)

\[ D^* (\theta : \theta') := D(\theta' : \theta) \quad (D^*)^* = D \]

• Statistical manifold structures:

\[ (M, Dg, D\nabla, D^* \nabla) \quad (M, Dg, D\nabla^\alpha, (D\nabla^\alpha)^* = D\nabla^\alpha) \]

\[ Dg := -\partial_{i,j} D(\theta : \theta')|_{\theta=\theta'} = D^* g, \]
\[ D\Gamma_{ijk} := -\partial_{i,j,k} D(\theta : \theta')|_{\theta=\theta'}, \]
\[ D^* \Gamma_{ijk} := -\partial_{k,i,j} D(\theta : \theta')|_{\theta=\theta'}. \]

\[ DC_{ijk} = D^* \Gamma_{ijk} - D\Gamma_{ijk} \]
\[ D\nabla^* = D^* \nabla \]

Eguchi, Geometry of minimum contrast, Hiroshima Mathematical Journal 22.3 (1992)
Part II.C
- Bregman manifolds: Dually flat spaces
Dually flat geometry from a convex function

Not necessarily related to statistical models, but can always be realized by a regular statistical model.

Metric tensor using covariant/contravariant notations

**2-covariant metric tensor** in local coordinates:

\[ g_{ij}(\theta) = \nabla^2 F(\theta) \]

**Dual metric tensor** in local coordinates:

\[ g^{ij}(\eta) = g^{*ij}(\eta) = \nabla^2 F^*(\eta) \]

**Crouzeix’s identity**: \( x \) of Hessians of convex conjugates = \( \text{Id} \):

\[ \nabla^2 F(\theta) \nabla^2 F^*(\eta) = I \]

An elementary introduction to information geometry. Entropy 22.10 (2020)
Bregman information geometry: Bregman manifolds

- Start from a potential function $F(\theta)$
  \[ Fg = \nabla^2 F(\theta) \]

- Get the dual potential function $F^*(\eta)$
  \[ Fg^* = \nabla^2 F^*(\eta) \]

- Define the primal flat connection:
  \[ F\Gamma_{ijk}(\theta) = 0 \]

- Define the dual flat connection:
  \[ F\Gamma^*_{ijk}(\eta) = 0 \]

- Get the dual Bregman divergences or dual Fenchel-Young divergences

The many faces of information geometry, Notices of the AMS, January 2022
Bregman manifolds vs Hessian manifolds

- **Hessian metric** wrt. a flat connection $\nabla$, function is 0-form on $M$:
  
  $g = \nabla^2 F_M$

- **Hessian operator**:
  
  $\left( \nabla^2 F_M \right) (X, Y) := (\nabla_X d) (F_M(Y)) = X(dF_M(Y)) - dF_M(\nabla_X Y)$

  
  $\nabla^2 F_M \left( \frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j} \right) = \frac{\partial^2 F_M}{\partial x^i \partial x^j} - \Gamma^k_{ij} \frac{\partial F_M}{\partial x^k}$

  \[ \nabla^2 F_M \left( \frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j} \right) = \frac{\partial^2 F_M}{\partial x^i \partial x^j} \]

- **Bregman manifold**:
  
  geometry on an open convex domain:

  Here, $\nabla = \text{gradient}$

  Here, $\nabla, \nabla^* = \text{affine flat connections}$

  $g(\theta) = \nabla^2 F(\theta)$

  $g^*(\eta) = \nabla^2 F^*(\eta)$

  $\nabla : \Gamma_{ijk}(\theta) = 0$

  $\nabla^* : \Gamma^{*ijk}(\eta) = 0$

Part III
Generalized convexity and divergences from convexity gaps
Chordal slope lemma & Jensen/Bregman divergences

Jensen Divergence (JD)

BD as a limit of a scaled JD: \[ B_F(\theta_1 : \theta_2) = \lim_{\alpha \to 1^-} \frac{1}{\alpha(1-\alpha)} J_{F,\alpha}(\theta_1 : \theta_2) \]

Bregman Divergences (BDs):

\[ F'(\theta_1) \leq \frac{F(\theta_2) - F(\theta_1)}{\theta_2 - \theta_1} \leq F'(\theta_2) \]

\[ \frac{F(\theta) - F(\theta_1)}{\alpha(\theta_2 - \theta_1)} < \frac{F(\theta_2) - F(\theta_1)}{\theta_2 - \theta_1} \]

\[ F(\theta) - F(\theta_1) < \alpha(F(\theta_2) - F(\theta_1)) \]

\[ \alpha(F(\theta_2) - F(\theta_1)) - F(\theta) + F(\theta_1) > 0, \]

\[ J_{F,\alpha}(\theta_1 : \theta_2) := (1 - \alpha)F(\theta_1) + \alpha F(\theta_2) - F((1 - \alpha)\theta_1 + \alpha \theta_2) > 0. \]

\[ B_F(\theta_2 : \theta_1) \geq 0, \]

\[ B_F(\theta_1 : \theta_2) \geq 0. \]
Bregman divergences w.r.t comparative convexity

- Two **abstract means** $M$ and $N$, i.e.
  \[ \min\{p, q\} \leq M(p, q) \leq \max\{p, q\}. \]

- Define a function $F$ **(M,N) convex** if
  \[ F(M(p, q)) \leq N(F(p), F(q)), \quad \forall p, q \in \mathcal{X}, \]

- Consider the means **regular**: homogeneous, symmetric, continuous, and increasing in each variable

- Define **skew (M,N)-Jensen divergence** for a strictly convex (M,N)-function for regular means $M$ and $N$:
  \[ J_{F,\alpha}^{M,N}(p : q) = N_\alpha(F(p), F(q)) - F(M_\alpha(p, q)). \]

- By analogy of ordinary Bregman divergences obtained as limit of scaled skew Jensen divergences, define **(M,N)-Bregman divergences**:
  \[ B_{F,\alpha}^{M,N}(p : q) = \lim_{\alpha \to 1^-} \frac{1}{\alpha(1-\alpha)} J_{F,\alpha}^{M,N}(p : q) = \lim_{\alpha \to 1^-} \frac{1}{\alpha(1-\alpha)} (N_\alpha(F(p), F(q))) - F(M_\alpha(p, q))). \]

Quasi-arithmetic (rho-tau)-Bregman divergences

• For a strictly continuously monotone function $\gamma$, define the weighted quasi-arithmetic means

$$M_{\gamma,\alpha}(x, y) = \gamma^{-1}(((1 - \alpha)\gamma(x) + \alpha\gamma(y))$$

• Quasi-arithmetic Bregman divergence:

$$B_{F}^{\rho,\tau}(q : p) = \lim_{\alpha \to 0} \frac{1}{\alpha(1 - \alpha)} (M_{\tau,\alpha}(F(p), F(q))) - F(M_{\rho,\alpha}(p, q))$$

• Consider the ordinary convex function:

$$G(x) = \tau(F(\rho^{-1}(x)))$$

• Quasi-arithmetic (rho-tau)-Bregman divergences is a conformal regular Bregman divergence:

$$B_{F}^{\rho,\tau}(p : q) = \frac{1}{\tau'(F(q))} B_{G}(\rho(p) : \rho(q))$$

Quasi-convex Jensen and Bregman divergences

• Strictly quasiconvex function:

\[ Q((\theta'\theta)_{\alpha}) < \max\{Q(\theta), Q(\theta')\}, \quad \theta \neq \theta' \in \Theta \]

\[(\theta'\theta)_{\alpha} := (1 - \alpha)\theta + \alpha\theta' \]

• Quasiconvex Jensen divergence:

\[ qcvx_J^\alpha_Q(\theta : \theta') := \max\{Q(\theta), Q(\theta')\} - Q((\theta'\theta)_{\alpha}) \geq 0, \]

\[ = \max\{Q(\theta), Q(\theta')\} - Q((1 - \alpha)\theta + \alpha\theta'). \]

• Quasic-convex Jensen divergence is a \((\text{Max},A)\)-Jensen divergence!

Multivariate Bregman divergence as a family of univariate Bregman divergences

Proposition: A multivariate Bregman divergence $B_F(\theta_1 : \theta_2)$ can be written equivalently as a univariate Bregman divergence $B_{F_{\theta_1, \theta_2}}(0:1)$:

$$\forall \theta_1, \theta_2 \in \Theta, \quad B_F(\theta_1 : \theta_2) = B_{F_{\theta_1, \theta_2}}(0:1),$$

where

$$F_{\theta_1, \theta_2}(u) := F(\theta_1 + u(\theta_2 - \theta_1))$$

is a univariate Bregman divergence.

Proof: The univariate functions $F_{\theta_1, \theta_2}$ are proper 1D Bregman generators. We have the directional derivative:

$$\nabla_{\theta_2-\theta_1} F_{\theta_1, \theta_2}(u) = \lim_{\epsilon \to 0} \frac{F(\theta_1 + (\epsilon + u)(\theta_2 - \theta_1)) - F(\theta_1 + u(\theta_2 - \theta_1))}{\epsilon},$$

$$= (\theta_2 - \theta_1) \nabla F(\theta_1 + u(\theta_2 - \theta_1)).$$

Since $F_{\theta_1, \theta_2}(0) = F(\theta_1), F_{\theta_1, \theta_2}(1) = F(\theta_2), \text{ and } F'_{\theta_1, \theta_2}(u) = \nabla_{\theta_2-\theta_1} F_{\theta_1, \theta_2}(u)$, it follows that

$$B_{F_{\theta_1, \theta_2}}(0:1) = F_{\theta_1, \theta_2}(0) - F_{\theta_1, \theta_2}(1) - (0 - 1) \nabla_{\theta_2-\theta_1} F_{\theta_1, \theta_2}(1),$$

$$= F(\theta_1) - F(\theta_2) + (\theta_2 - \theta_1) \nabla F(\theta_2) = B_F(\theta_1 : \theta_2).$$
Designing divergences by measuring convexity gaps

The Bregman chord divergence, GSI, Springer, 2019
Thank you!

“The only constant in life is change”  -Heraclitus

My motto: "Invariance is the only constant in change!"

https://franknielsen.github.io/
Adaptive computational geometry (PhD, 1996)

Computational geometry:
Output-sensitive algorithms

Output-sensitive 2D lower envelopes
convex hull of objects

Output-sensitive peeling of k convex or maximal layers
(Pareto front)

Piercing/stabbing
d-dimensional isothetic boxes
Klee’s measure problem

Convex geometry:
Helly and Hellinger numbers for piercing

- Algorithmes géométriques adaptatifs (PhD), Université Nice Sophia Antipolis, 1996
- On piercing sets of objects, Proceedings of the twelfth annual symposium on Computational geometry. 1996.
Figure 13.2. Voronoi diagram and convex polyhedron.

[Voronoi by mapping to the paraboloid] [Bregman Voronoi by mapping to Bregman potential functions]

The fabric of information geometry and the untangling of its geometry, divergence, statistical models.