Divergences and comparative convexity

Frank Nielsen

Sony Computer Science Laboratories Inc
Tokyo, Japan

Sony CSL

arXiv:2312.12849
2024

Rationale

- Need to define statistical dissimilarity measures D(p,q) between statistical models p and q in statistics and machine learning: for example, total variation distance, Kullback-Leibler divergence, Wasserstein, Maximum Mean Discrepancy, etc.

- Infer models from a statistical model $P=\{p_\theta\}$: estimate θ and measure goodness-of-fit from data (empirical distribution)

- Statistical dissimilarity measure between parametric models $P=\{p_\theta\}$ amount to dissimilarity between parameters:

$$D(p_{\theta_1}, p_{\theta_2}) =: D_{\mathcal{P}}(\theta_1, \theta_2)$$

- In this talk, we investigate some relationships between statistical dissimilarities, statistical models, model parameter dissimilarities, and their underlying geometries.

Outline

Background

- Kullback-Leibler divergence, exponential families, and Bregman divergences
- Dual information geometry of convex functions

Interplay of Models/distances

- Two normalizations of exponential families: **cumulant** or **partition** functions, and their relationships with parameter divergences:

1. well-known: Bhattacharyya distances/Rényi divergences and skewed Jensen divergences with respect to cumulant function
2. New: **α-divergences and skewed Jensen divergences with respect to partition function of exponential families**

Deforming convex functions

- Comparative convexity, quasi-arithmetic means, and convex deformations preserving convexity
- Some generalization of Bregman divergences yielding conformal Bregman divergences

Kullback-Leibler divergence: relative entropy

- The Kullback-Leibler divergence (KLD) is a *dissimilarity measure* between probability distributions or measures P and Q :

$$D_{\mathrm{KL}}(P:Q) = \begin{cases} \int p \log \frac{p}{q} \, \mathrm{d}\mu, & p = \frac{\mathrm{dP}}{\mathrm{d}\mu}, Q = \frac{\mathrm{dQ}}{\mathrm{d}\mu}, \quad P \ll Q \\ +\infty & P \not\ll Q \end{cases}$$

- KLD fails symmetry and triangle inequality of metrics but is always non-negative, a property known as Gibbs' inequality:

$$D_{\mathrm{KL}}(P:Q) \geq 0, \quad D_{\mathrm{KL}}(P:Q) \neq D_{\mathrm{KL}}(Q:P), \quad D_{\mathrm{KL}}(P:Q) \not\leq D_{\mathrm{KL}}(P:Q) + D_{\mathrm{KL}}(Q:R)$$

- KLD also called relative entropy because it is the difference between the cross-entropy and Shannon entropy:

$$D_{\mathrm{KL}}(P:Q) = H^{\times}(P:Q) - H(P), H^{\times}(P:Q) = -\int p \log q \, \mathrm{d}\mu, H(P) = H^{\times}(P:P) = -\int p \log p \, \mathrm{d}\mu$$

- Interpretation of KLD in information theory: expected difference of the number of bits required for Huffman encoding of P using a code optimized for Q rather than the Huffman code optimized for P.

Exponential families: Discrete/continuous/measures

- A parametric family of distributions $\{P_\lambda\}$ all dominated by a measure μ is an **exponential family** iff the densities wrt μ can be expressed canonically as

$$\begin{aligned} p_\lambda(x) &= \exp\left(\langle\theta(\lambda), t(x)\rangle - F(\theta(\lambda)) + k(x)\right) \\ &= \frac{1}{Z(\theta)} \exp\left(\langle\theta(\lambda), t(x)\rangle\right) h(x) \end{aligned}$$

- θ is **natural parameter**
- t(x) is **sufficient statistics,** and k(x) or h(x) is an auxiliary carrier term
- Inner product (e.g., scalar product for vectors, tr(AB) for symmetric matrices)

- Unnormalized density:
- *Subtractive normalization*:
- *Divisive normalization*:

$$\tilde{p}_\lambda(x) = \exp\left(\langle\theta(\lambda), t(x)\rangle\right) h(x), \quad \tilde{p}_\theta(x) = \exp\left(\langle\theta, t(x)\rangle\right) h(x)$$

$$F(\theta) = \log Z(\theta) = \log \int_{\mathcal{X}} \tilde{p}_\theta(x)\, \mathrm{d}\mu(x)$$

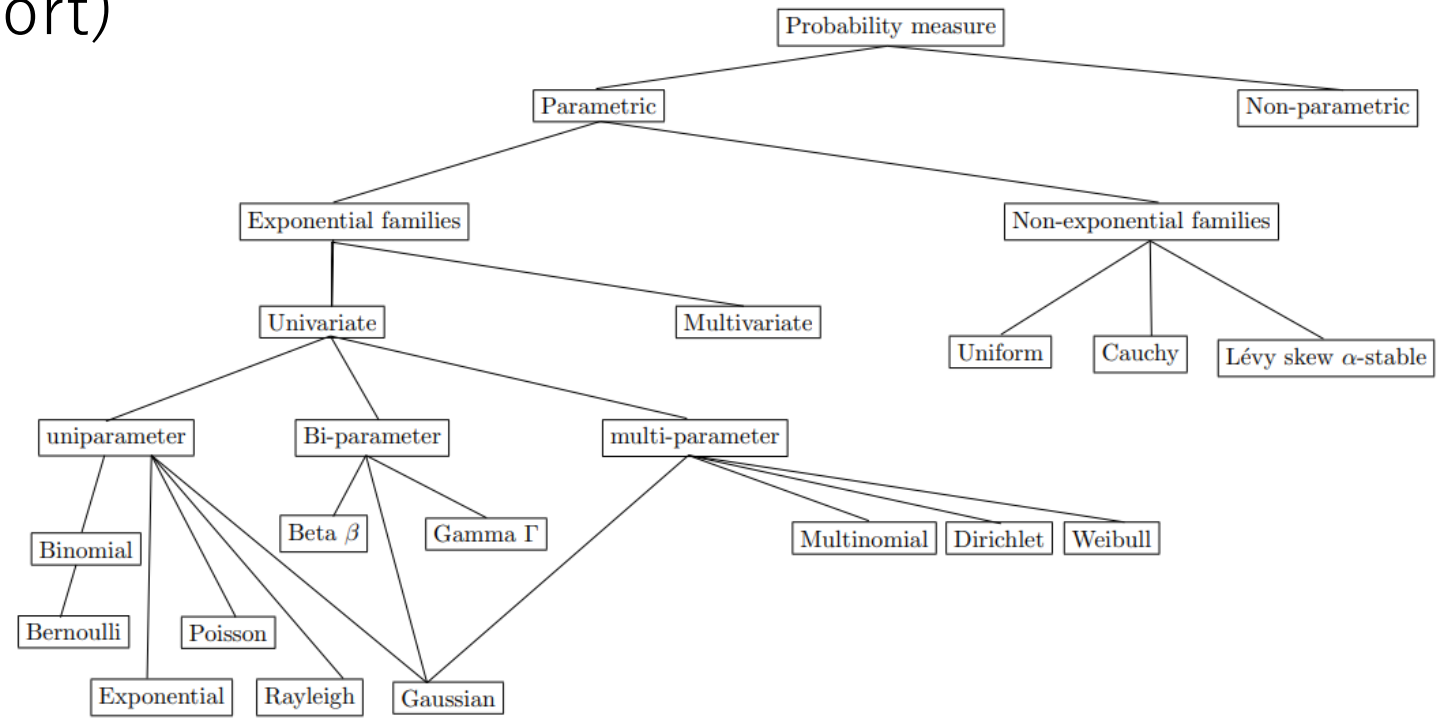
$$Z(\theta) = \exp(F(\theta)) = \int_{\mathcal{X}} \tilde{p}_\theta(x)\, \mathrm{d}\mu(x)$$

- F called **cumulant function** in statistics
- F called **free energy** and Z called **partition function** in thermodynamics

Exponential families (EFs): Some examples

- Many common distributions in statistics are exponential families in disguise (common support)

$$
\begin{aligned}
p_\lambda(x) &= \exp\left(\langle \theta(\lambda), t(x) \rangle - F(\theta(\lambda)) + k(x)\right) \\
&= \frac{1}{Z(\theta)} \exp\left(\langle \theta(\lambda), t(x) \rangle\right) h(x)
\end{aligned}
$$



- Many statistical models in machine learning are exponential families: undirected graphical models, energy-based models including Markov random fields and conditional random fields:

  ➤ Normalizers F or Z are often **computationally intractable**

KLD between two densities of an Exp. Fam.

- Bypass integral calculations of KLDs, and express the KLD as a divergence between parameters: **Bregman divergences**

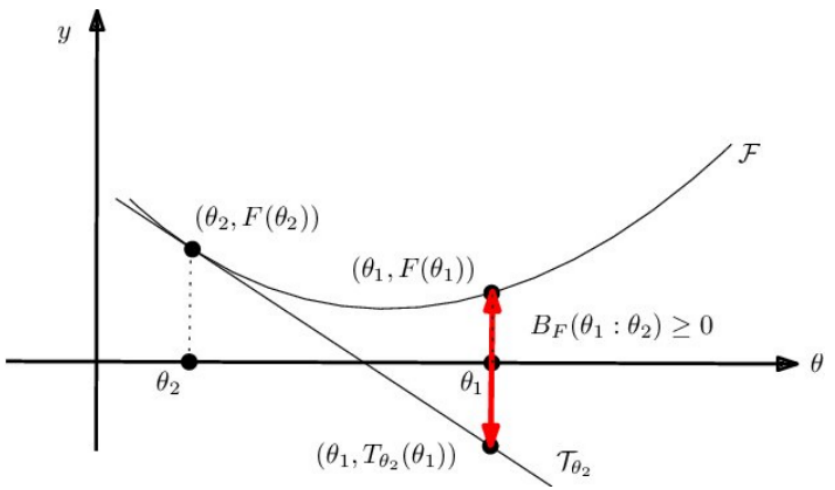
$$p_\lambda(x) = \exp\left(\langle \theta(\lambda), t(x) \rangle - F(\theta(\lambda)) + k(x)\right)$$

$$\begin{aligned} D_{\mathrm{KL}}(p_{\lambda_1} : p_{\lambda_2}) &= D_{\mathrm{KL}}(p_{\theta_1} : p_{\theta_2}) = \int p_{\theta_1} \log \frac{p_{\theta_1}}{p_{\theta_2}} \, \mathrm{d}\mu \\ &= B_F(\theta_2 : \theta_1) = B_F(\theta(\lambda_2) : \theta(\lambda_1)) \\ &= F(\theta_2) - F(\theta_1) - \langle \theta_2 - \theta_1, \nabla F(\theta_1) \rangle \end{aligned}$$

Geometric interpretation of BDs:

$$B_F(\theta_1 : \theta_2) = F(\theta_1) - T_{\theta_2}(\theta_1)$$

$$T_\theta(\omega) := F(\theta) + (\omega - \theta) F'(\theta)$$



- **Dual expectation/moment parameterization**: $\eta = \nabla F(\theta) = E_{p_\theta}[t(X)]$
- Many equivalent parameterizations of EFs: $p_\lambda \leftrightarrow p_{\theta(\lambda)} \leftrightarrow p_{\eta(\theta)}$

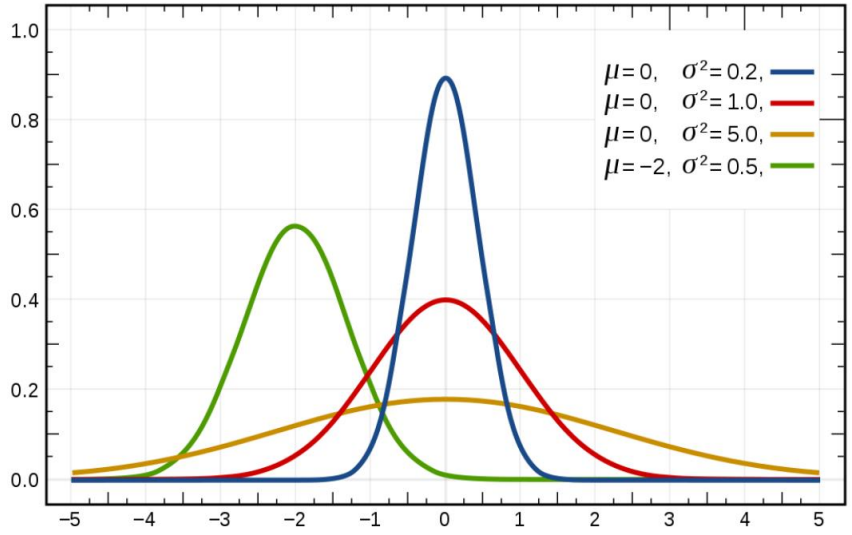
$\mathrm{N}(\mu, \sigma)$, $\mathrm{N}(\mu, \sigma^2)$, $\mathrm{N}(\mu, \log \sigma)$, etc

Exponential family of univariate normal distributions

Usual parametrization $\lambda = (\mu, \sigma)$ or (μ, σ^2) :

Density: $$p_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\begin{aligned} p_\lambda(x) &= \exp\left(\langle\theta(\lambda), t(x)\rangle - F(\theta(\lambda)) + k(x)\right) \\ &= \frac{1}{Z(\theta)} \exp\left(\langle\theta(\lambda), t(x)\rangle\right) h(x) \end{aligned}$$



Normal family $N=\{p_\theta\}$:

Auxiliary carrier term k(x)=0, h(x)=1 with respect to Lebesgue measure μ

KLD Equivalent BD $= \frac{1}{2}\left[\frac{(\mu_2-\mu_1)^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - \ln\frac{\sigma_1^2}{\sigma_2^2} - 1\right]$

Mahalanobis^2+Itakura-Saito

Sufficient statistic: $t(x)=(x,x^2)$

Natural parameters: $$\theta = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$$

natural parameters unique up to affine transformations

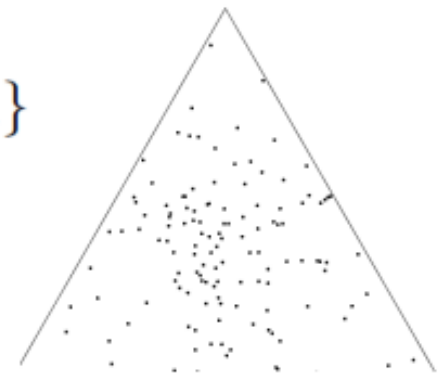
Log-normalizer F: $$F(\theta) = -\frac{(\theta_1)^2}{4\theta_2} - \frac{1}{2}\ln(-\theta_2) + \frac{1}{2}\log\pi$$

constant can be added/subtracted with k(x)

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**.

$$p_\lambda = (p_\lambda^1, \ldots, p_\lambda^d) \in \Delta_{d-1}^\circ, \quad \sum_{i=1}^{d} p_\lambda^i = 1 \qquad \theta^i = \log \frac{\lambda^i}{\lambda^D}, i \in \{1, \ldots, D = d-1\}$$



$$\mathcal{D}_{\mathrm{KL}}[p_{\lambda_1} : p_{\lambda_2}] := \sum_{i=1}^{D} \lambda_1^i \log \frac{\lambda_1^i}{\lambda_2^i} =: B_{F_{\mathrm{KL}}}(\theta_1 : \theta_2)$$

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

LogSumExp is only convex but **LogSumExp$_+$ is strictly convex** [NH 2019]

[NH 2019] Monte Carlo information-geometric structures, Geometric Structures of Information, 2019.
Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities, Entropy, 18(12), 2016

Convex duality: convex conjugate pairs (F,F*)

- **Legendre-Fenchel transformation** of a function:

  as known as slope transform: $$F^*(\eta) = \sup_{\theta \in \Theta} \langle \theta, \eta \rangle - F(\theta)$$

- Supremum reached for $\eta = \nabla F(\theta)$: defines the **gradient map**
- Moment parameter space: $H = \{\nabla F(\theta) \ : \ \theta \in \Theta\}$
- Restrict F to **Legendre-type function** $(F(\theta), \Theta)$ so that the convex conjugate is also of Legendre type: $(F^*(\eta), H)$

$$\theta = \nabla F^*(\eta) \Leftrightarrow \eta = \nabla F(\theta), \nabla F^*(\nabla F(\theta)) = \theta$$

- And we have: $\theta = \nabla F^*(\eta)$ and $F^{**} = F$ reciprocal gradient: $\nabla F^* = (\nabla F)^{-1}$
- Legendre transformation:

  only need to invert ∇F

$$F^*(\eta) = \langle \eta, (\nabla F)^{-1}(\eta) \rangle - F((\nabla F)^{-1}(\eta))$$

Dual Bregman divergence/Fenchel-Young divergence

- Bregman divergence can be expressed equivalently as a **Fenchel-Young divergence** using the convex pairs:

$$B_F(\theta_1 : \theta_2) = Y_{F,F^*}(\theta_1 : \eta_2) = F(\theta_1) + F^*(\eta_2) - \langle \theta_1, \eta_2 \rangle$$

- **Dual Bregman divergence**: $B_F(\theta_1 : \theta_2) = B_{F^*}(\eta_2 : \eta_1)$

- KLD between densities of an exponential family expressed equivalently as:

$$\begin{aligned} D_{\mathrm{KL}}(p_{\lambda_1} : p_{\lambda_2}) &= D_{\mathrm{KL}}(p_{\theta_1} : p_{\theta_2}) = \int p_{\theta_1} \log \frac{p_{\theta_1}}{p_{\theta_2}} \, \mathrm{d}\mu \\ &= B_F(\theta(\lambda_2) : \theta(\lambda_1)) = B_F(\theta_2 : \theta_1) = Y_{F,F^*}(\theta_2 : \eta_1) \\ &= B_{F^*}(\eta_1 : \eta_2) = Y_{F^*,F}(\eta_1 : \theta_2) \end{aligned}$$

① ② ③ ④

Information geometry: Dually structures

Differential geometry

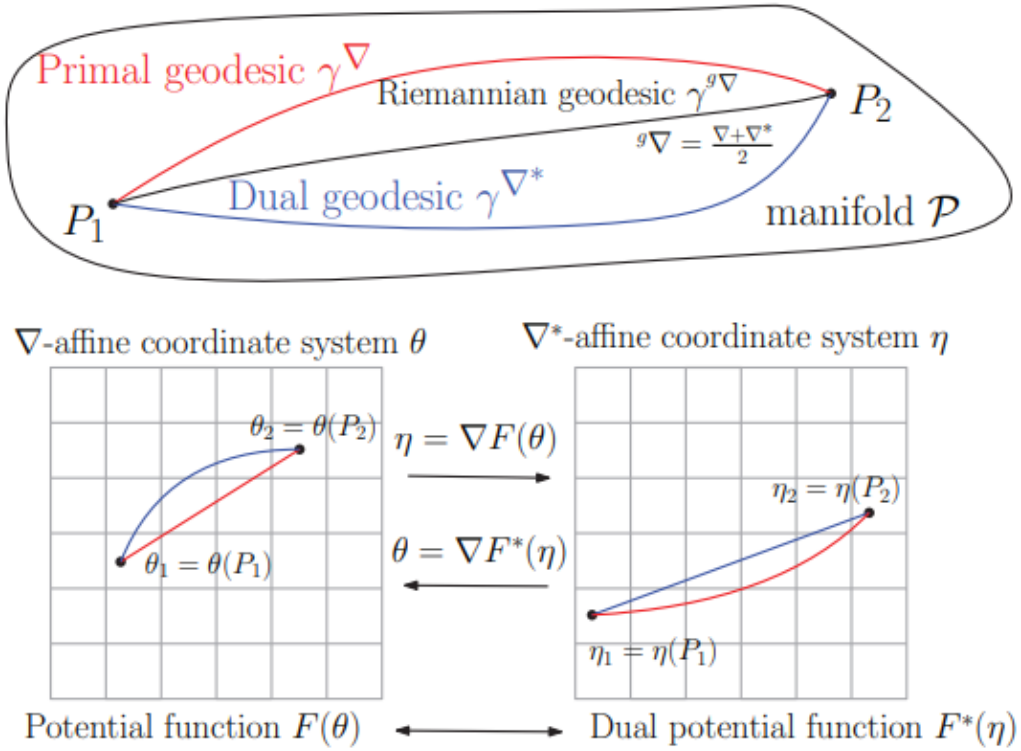
- Riemannian metric g is smooth inner product on a manifold which allows to measure vector lengths and angles between vectors in tangent spaces
- Affine connection ∇ defines how to connect vectors between infinitesimally close tangent spaces. Affine connection defines ∇-geodesic as autoparallel curves

- Information geometry considers dual structures: A manifold M equipped with a Riemannian metric tensor g and dual torsion-free affine connections ∇ and ∇^* coupled to the metric so that the Levi-Civita connection wrt g is $(\nabla+\nabla^*)/2$: Structure (M,g,∇,∇^*)

- Information geometry induced by ① statistical models $\{p_\theta\}$, ② information geometry induced by divergences, ③ information geometry induced by convex functions, ④ information geometry induced by regular cones, etc.

Information geometry of convex functions: Dually flat spaces, global Hessian manifolds

- An affine connection ∇ is flat if there exists a coordinate system θ called ∇-affine coordinate system such that the Christoffel symbols Γ vanish
- ∇-geodesics are straight lines in θ-chart
- Hessian metric tensor g expressed in θ-chart as $\nabla^2 F(\theta)$
- Legendre duality yields dual expression of Hessian metric $\nabla^2 F^*(\eta)$ and dual affine flat connection ∇^* with ∇^*-geodesics straight in η-chart
- Dually flat space DFS(F(θ), ⊖)=(M,g,∇,∇^*)

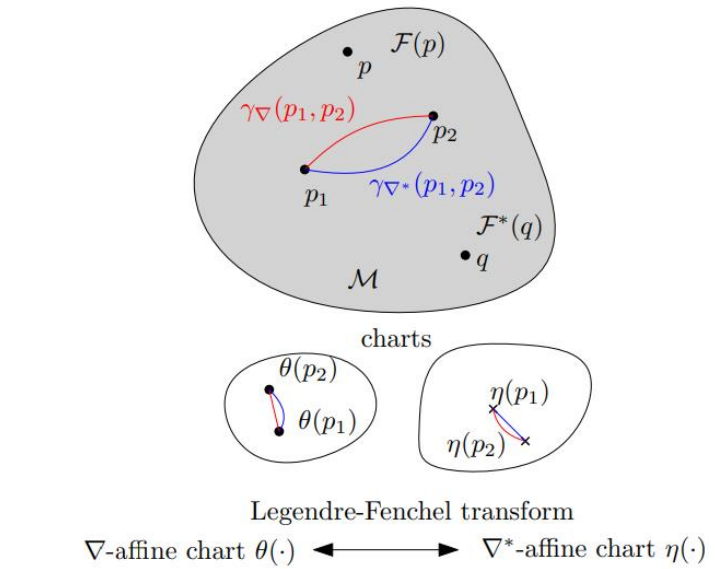


Canonical divergences of dually flat spaces: Dually flat divergences

- Given a dually flat space (M, g, ∇, ∇^*), we can reconstruct locally two potential functions $F(\theta)$ and $F^*(\eta)$ related by Legendre-Fenchel transformation
- The dually flat divergence $D_{\nabla, \nabla^*}(P:Q)$ can be expressed using the *mixed coordinate system* θ and η as a Fenchel-Young divergence or equivalently using dual Bregman divergences either in the θ- or η-charts

$$\begin{aligned} D_{\nabla,\nabla^*}(P:Q) &= Y_{F,F^*}(\theta(P):\eta(Q)) \\ &= B_F(\theta(P):\theta(Q)) \\ &= B_{F^*}(\eta(Q):\eta(P)) \\ &= Y_{F^*,F}(\eta(Q):\theta(P)) \end{aligned}$$

$$\mathrm{DFS}(\bar{F}(\bar{\theta}), \bar{\Theta}) = \mathrm{DFS}(F(\theta), \Theta), \quad \bar{F}(\bar{\theta}) = A\,F(\bar{\theta}) + a, \bar{\theta} = B\,\theta + b$$



Canonical divergence of cumulant functions amount to statistical reverse KLD: $B_F(\theta_1 : \theta_2) = D^*_{\mathrm{KL}}(p_{\theta_1} : p_{\theta_2})$

Usually, in Statistics/ML, we prove ⇒: $D_{\mathrm{KL}}(p_{\theta_1} : p_{\theta_2}) = {B_F}^*(\theta_1 : \theta_2) = B_F(\theta_2 : \theta_1)$

where D* is dual divergence: $D^*(p : q) = D(q : p)$

Let us prove ⇐ from information geometry of canonical divergence of DFS(F(θ), ϴ)

① $F^*(\eta) = E_{p_\theta}[\log p_\theta] = -H(p_\theta)$:

$$\begin{aligned} H(p_\theta) &= -E_{p_\theta}[\log p_\theta], \\ &= -E_{p_\theta}[\langle \theta, x\rangle - F(\theta)], \\ &= F(\theta) - \langle \theta, E_{p_\theta}[x]\rangle, \\ &= F(\theta) - \langle \theta, \eta\rangle, \\ &= -F^*(\eta). \end{aligned}$$

② $$\begin{aligned} \langle \theta_1, \eta_2\rangle &= \langle \theta_1, E_{p_{\theta_2}}[x]\rangle, \\ &= E_{p_{\theta_2}}[\langle \theta_1, x\rangle] = E_{p_{\theta_2}}[\log \tilde{p}_{\theta_1}(x)], \\ &= E_{p_{\theta_2}}[\log p_{\theta_1}(x) + F(\theta_1)], \\ &= E_{p_{\theta_2}}[\log p_{\theta_1}(x)] + F(\theta_1). \end{aligned}$$

③ $$\begin{aligned} D(p_{\theta_1} : p_{\theta_2}) = Y_{F,F^*}(\theta_1 : \eta_2) &= F(\theta_1) + F^*(\eta_2) - \langle \theta_1, \eta_2\rangle, \\ &= F(\theta_1) - H(p_{\theta_2}) - E_{p_{\theta_2}}[\log p_{\theta_1}(x)] - F(\theta_1) \\ &= H^\times(p_{\theta_2} : p_{\theta_1}) - H(p_{\theta_2}), \\ &= D_{\mathrm{KL}}(p_{\theta_2} : p_{\theta_1}), \\ &= D^*_{\mathrm{KL}}(p_{\theta_1} : p_{\theta_2}). \end{aligned}$$

Canonical divergence of cumulant functions amount to statistical reverse KLD:

$$D_{\mathrm{KL}}(p_{\theta_1} : p_{\theta_2}) = B_F{}^*(\theta_1 : \theta_2) = B_F(\theta_2 : \theta_1)$$

We reconstruct Kullback-Leibler divergence by relaxing to arbitrary densities

$B_F(\theta_1 : \theta_2) = D^*_{\mathrm{KL}}(p_{\theta_1} : p_{\theta_2})$ ⇒KLD

Interpretations:

- $F(\theta)$ is the cumulant function (also called free energy in thermodynamics),
- $\eta = \nabla F(\theta) = E_{p_\theta}[t(x)]$ is the moment of the sufficient statistic,
- $F^*(\eta) = -H(p_\theta)$ is the negentropy, and
- $\theta = \nabla F^*(\eta)$ are the Lagrangian multipliers in the maximum entropy problem

Natural parameter space Θ of Exp Fam is convex

Proof. Let Θ denote the natural parameter space:

$$\Theta = \left\{ \theta \ : \ Z(\theta) = \int \exp(\langle \theta, x \rangle) \mathrm{d}\mu < \infty \right\} = \left\{ \theta \ : \ F(\theta) = \log \int \exp(\langle \theta, x \rangle) \mathrm{d}\mu < \infty \right\}.$$

Let $\theta_0, \theta_1 \in \Theta$ and consider $\theta_\alpha = \theta_0 + \alpha(\theta_1 - \theta_0)$ for $\alpha \in (0,1)$. In order to show that Θ is convex, we need to prove that $\theta_\alpha \in \Theta$, i.e., $Z(\theta_\alpha) < \infty$. We have

$$\begin{aligned} \int \exp(\langle \theta_\alpha, x \rangle) \, \mathrm{d}\mu(x) &= \int \exp(\langle \alpha\theta_0, x \rangle) \exp(\langle (1-\alpha)\theta_1, x \rangle) \mathrm{d}\mu(x), \\ &= \int (\exp(\langle \theta_0, x \rangle))^\alpha \, (\exp(\langle \theta_1, x \rangle))^{(1-\alpha)} \, \mathrm{d}\mu(x). \end{aligned} \tag{31}$$

Now, recall Hölder inequality for positive functions $f(x)$ and $g(x)$ with conjugate exponents p and q in $[1, \infty)$ such that $\frac{1}{p} + \frac{1}{q} = 1$:

$$\int f(x) g(x) \mathrm{d}\mu(x) \leq \left(\int f^p(x) \mathrm{d}\mu(x) \right)^{\frac{1}{p}} \left(\int g^q(x) \mathrm{d}\mu(x) \right)^{\frac{1}{q}}.$$

Consider $f(x) = (\exp(\langle \theta_0, x \rangle))^\alpha$ and $p = \frac{1}{\alpha} > 1$ and $g(x) = (\exp(\langle \theta_1, x \rangle))^{1-\alpha}$ with $q = \frac{1}{1-\alpha} > 1$ (we check that $\frac{1}{p} + \frac{1}{q} = \alpha + 1 - \alpha = 1$). Thus we upper bound Eq. 31 using Hölder inequality as follows:

$$\int \exp(\langle \theta_\alpha, x \rangle) \, \mathrm{d}\mu(x) \leq \left(\int \exp(\langle \theta_0, x \rangle) \mathrm{d}\mu(x) \right)^\alpha \left(\int \exp(\langle \theta_1, x \rangle) \mathrm{d}\mu(x) \right)^{1-\alpha} < \infty, \tag{32}$$

since both $\int \exp(\langle \theta_0, x \rangle) \mathrm{d}\mu(x) < \infty$ and $\int \exp(\langle \theta_1, x \rangle) \mathrm{d}\mu(x) < \infty$ because θ_0 and θ_1 both belong to Θ. Hence, we have shown that Θ is convex. $\square$

Partition function $Z(\theta)=\exp(F(\theta))$ is strictly log-convex
Cumulant function $F(\theta)=\log Z(\theta)$ is strictly convex

When we proved that natural parameter space is convex, we had

$$\int \exp(\langle \theta_\alpha, x\rangle)\,\mathrm{d}\mu(x) \leq \left(\int \exp(\langle \theta_0, x\rangle)\mathrm{d}\mu(x)\right)^{\alpha} \left(\int \exp(\langle \theta_1, x\rangle)\mathrm{d}\mu(x)\right)^{1-\alpha} < \infty$$

That is for short: $Z(\theta_\alpha) \leq Z(\theta_0)^\alpha\, Z(\theta_1)^{1-\alpha}.$ (Z=partition function)

Take the logarithm on both sides:

$$\begin{aligned} \log Z(\theta_\alpha) &\leq \log\left(Z(\theta_0)^\alpha\, Z(\theta_1)^{1-\alpha}\right), \\ F(\alpha\theta_0 + (1-\alpha)\theta_1) &\leq \alpha F(\theta_0) + (1-\alpha)F(\theta_1) \end{aligned}$$

F is strictly convex since Eq. iff $\theta_1 = \theta_2$

<u>Definition</u>: A function Z is stricty log-convex is log Z is strictly convex

$\Rightarrow$ $Z(\theta)=\exp(F(\theta))$ is strictly convex because $F(\theta)$ strictly convex:

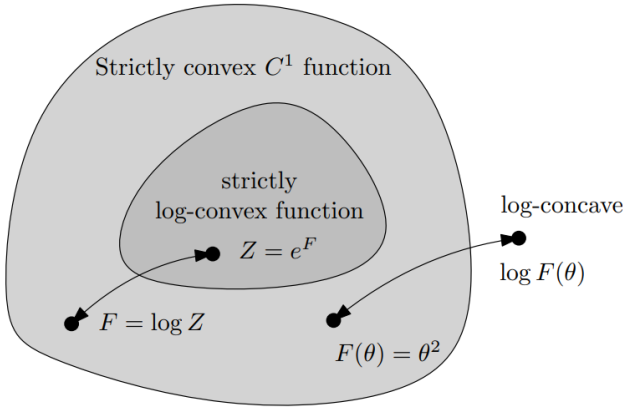
Property: A log-convex function is also convex (but not necessarily the converse)

Proof. By definition, function $Z(\theta)$ is strictly log-convex if and only if:

$$\forall \theta_0 \neq \theta_1, \quad Z(\alpha\theta_0 + (1-\alpha\theta_1)) < Z(\theta_0)^\alpha \, Z(\theta_1)^{1-\alpha}, \tag{3}$$

i.e., by taking the logarithm on both sides of the inequality, $F = \log Z$ is strictly convex:

$$\begin{aligned} \forall \theta_0 \neq \theta_1, \quad \log Z(\alpha\theta_0 + (1-\alpha)\theta_1) &< \alpha \log Z(\theta_0) + (1-\alpha) \log Z(\theta_1), \\ \Leftrightarrow \quad F(\alpha\theta_0 + (1-\alpha)\theta_1) &< \alpha F(\theta_0) + (1-\alpha) F(\theta_1). \end{aligned}$$



Since $f(x) = \exp(x)$ is strictly convex (because $f''(x) = \exp(x) > 0$), we have for all $\alpha \in (0, 1)$:

$$f(\alpha F(\theta_0) + (1-\alpha) F(\theta_1)) < \alpha f(F(\theta_0)) + (1-\alpha) f(F(\theta_1)).$$

Letting $F(\theta) = \log Z(\theta)$ in the above inequality, we get:

$$\exp(\alpha \log Z(\theta_0) + (1-\alpha) \log Z(\theta_1) < \alpha \exp(\log Z(\theta_0)) + (1-\alpha) \exp(\log Z(\theta_1)), \tag{4}$$

$$Z(\theta_0)^\alpha \, Z(\theta_1)^{1-\alpha} < \alpha Z(\theta_0) + (1-\alpha) Z(\theta_1), \tag{5}$$

and therefore we get from Eq. 3 and Eq. 5:

$$\forall \theta_0 \neq \theta_1, Z(\alpha\theta_0 + (1-\alpha\theta_1)) < Z(\theta_0)^\alpha \, Z(\theta_1)^{1-\alpha} < \alpha Z(\theta_0) + (1-\alpha) Z(\theta_1). \tag{6}$$

That is, Z is strictly convex.

Bregman divergences $B_{F=\log Z}$ and $B_{Z=\exp F}$

$$
\begin{aligned}
B_Z(\theta_1 : \theta_2) &= Z(\theta_1) - Z(\theta_2) - \langle \theta_1 - \theta_2, \nabla Z(\theta_2) \rangle \geq 0, \\
B_{\log Z}(\theta_1 : \theta_2) &= \log \left(\frac{Z(\theta_1)}{Z(\theta_2)} \right) - \left\langle \theta_1 - \theta_2, \frac{\nabla Z(\theta_2)}{Z(\theta_2)} \right\rangle \geq 0,
\end{aligned}
$$

And furthermore, we can define **skewed Jensen divergences** from the convex generators:

$$
\begin{aligned}
J_{Z,\alpha}(\theta_1 : \theta_2) &= \alpha Z(\theta_1) + (1-\alpha) Z(\theta_2) - Z(\alpha\theta_1 + (1-\alpha)\theta_2) \geq 0, \\
J_{\log Z,\alpha}(\theta_1 : \theta_2) &= \log \frac{Z(\theta_1)^\alpha Z(\theta_2)^{1-\alpha}}{Z(\alpha\theta_1 + (1-\alpha)\theta_2)} \geq 0.
\end{aligned}
$$

Including the symmetric Jensen divergence when α=1/2:

$$
J_F(\theta_1, \theta_2) = J_{F,\frac{1}{2}}(\theta_1 : \theta_2) = \frac{F(\theta_1) + F(\theta_2)}{2} - F\left(\frac{\theta_1 + \theta_2}{2} \right)
$$

Bhattacharyya distances and Rényi divergences

- Question: If KLD between EF densities = B_F*, to what statistical divergences correspond J_F and $J_{\alpha,F}$?
- Define **scaled skewed Bhattacharyya distances**:

$$D^s_{B,\alpha}(p:q) = -\frac{1}{\alpha(1-\alpha)} \log \int p^\alpha q^{1-\alpha} \mathrm{d}\mu, \quad \alpha \in \mathbb{R}\backslash\{0,1\}$$

which are scaled **Rényi divergences**:

$$D^s_{B,\alpha}(p:q) = \frac{1}{\alpha} D_{R,\alpha}(p:q)$$

$$D_{R,\alpha}(p:q) = \frac{1}{\alpha-1} \log \int p^\alpha q^{1-\alpha} \mathrm{d}\mu$$

Scaling by $(1/\alpha(1\text{-}\alpha))$ allows to unify KLD with Bhattacharyya distances:

$$D^s_{B,\alpha}(p:q) = \begin{cases} -\frac{1}{\alpha(1-\alpha)} \log \int p^\alpha q^{1-\alpha} \mathrm{d}\mu, & \alpha \in \mathbb{R}\backslash\{0,1\}, \\ D_{\mathrm{KL}}(p:q), & \alpha = 1, \\ 4\, D_B(p,q) & \alpha = \frac{1}{2}, \\ D_{\mathrm{KL}}{}^*(p:q) = D_{\mathrm{KL}}(q:p) & \alpha = 0. \end{cases}$$

Bhattacharyya distances and Rényi divergences between densities of an exponential family

Proposition 4 ([32]). *The scaled α-skewed Bhattacharyya distances between two probability densities p_{θ_1} and p_{θ_2} of an exponential family amounts to the scaled α-skewed Jensen divergence between their natural parameters:*

$$D^s_{B,\alpha}(p_{\theta_1} : p_{\theta_2}) = J^s_{F,\alpha}(\theta_1, \theta_2). \quad (3)$$

Cumulant function F

Proof: consider the α-skewed Bhattacharyya similarity coefficient:

$$\begin{aligned}
\rho_\alpha(p_{\theta_1} : p_{\theta_2}) &= \int \exp\left(\langle\theta_1, x\rangle - F(\theta_1)\right)^\alpha \exp\left(\langle\theta_2, x\rangle - F(\theta_2)\right)^{1-\alpha} \mathrm{d}\mu, \\
&= \int \exp(\langle\alpha\theta_1 + (1-\alpha)\theta_2), x\rangle) \exp\left(-(\alpha F(\theta_1) + (1-\alpha)F(\theta_2))\right) \mathrm{d}\mu.
\end{aligned}$$

$$\rho_\alpha(p_{\theta_1} : p_{\theta_2}) = \exp(-(\alpha F(\theta_1) + (1-\alpha)F(\theta_2)) \exp(F(\bar{\theta})) \int \exp(\langle\bar{\theta}, x\rangle - F(\bar{\theta}))\mathrm{d}\mu.$$

$$\rho_\alpha(p_{\theta_1} : p_{\theta_2}) = \exp(-J_{F,\alpha}(\theta_1 : \theta_2))$$

Overview of classical statistical/Jensen divergences

Normalized densities $p_\theta = \exp(x \cdot \theta - F(\theta)) = \frac{\exp(x \cdot \theta)}{Z(\theta)}$

↓

Scaled Rényi α-divergence or α-skewed Bhattacharyya distance

$$\boxed{D^s_{B,\alpha}(p_{\theta_1} : p_{\theta_2}) = \frac{1}{\alpha} D_{R,\alpha}(p_{\theta_1} : p_{\theta_2}) = J^s_{F\alpha}(\theta_1 : \theta_2)}$$

$\alpha \to 0$	$\alpha = \frac{1}{2}$	$\alpha \to 1$
Reverse KLD	4 Bhattacharyya distance	KLD
$D_{\mathrm{KL}}{}^*(p_{\theta_1} : p_{\theta_2}) = B_F(\theta_1 : \theta_2)$		$D_{\mathrm{KL}}(p_{\theta_1} : p_{\theta_2}) = B_F{}^*(\theta_1 : \theta_2)$
Bregman	$4\, D_B(p_{\theta_1}, p_{\theta_2}) = 4\, J_F(\theta_1, \theta_2)$ 4 Jensen	Reverse Bregman

Extended Kullback-Leibler divergences between unnormalized densities: Bregman divergence B_Z

Extend KLD to **unnormalized densities**: $D_{\mathrm{KL}}(\tilde{p}:\tilde{q}) = \int \left(\tilde{p}\log\frac{\tilde{p}}{\tilde{q}} + \tilde{q} - \tilde{p}\right)\mathrm{d}\mu$

$$D_{\mathrm{KL}}(\tilde{p}:\tilde{q}) = H^{\times}(\tilde{p}:\tilde{q}) - H(\tilde{p})$$

$$H^{\times}(\tilde{p}:\tilde{q}) = \int \left(\tilde{p}(x)\log\frac{1}{\tilde{q}(x)} + \tilde{q}(x)\right)\mathrm{d}\mu(x) - 1$$

Reverse extended KLD: $D^{*}_{\mathrm{KL}}(\tilde{p}:\tilde{q}) = D_{\mathrm{KL}}(\tilde{q}:\tilde{p})$

$$\begin{aligned}
D_{\mathrm{KL}}(\tilde{p}_{\theta_1}:\tilde{p}_{\theta_2}) &= \int \left(\tilde{p}_{\theta_1}(x)\log\frac{\tilde{p}_{\theta_1}(x)}{\tilde{p}_{\theta_2}(x)} + \tilde{p}_{\theta_2}(x) - \tilde{p}_{\theta_1}(x)\right)\mathrm{d}\mu(x), \\
&= \int \left(e^{\langle t(x),\theta_1\rangle}\langle\theta_1-\theta_2, t(x)\rangle + e^{\langle t(x),\theta_2\rangle} - e^{\langle t(x),\theta_1\rangle}\right)\mathrm{d}\mu(x), \\
&= \left\langle \int t(x)e^{\langle t(x),\theta_1\rangle}\mathrm{d}\mu(x), \theta_1-\theta_2\right\rangle + Z(\theta_2) - Z(\theta_1), \\
&= \langle\theta_1-\theta_2, \nabla Z(\theta_1)\rangle + Z(\theta_2) - Z(\theta_1) = B_Z(\theta_2:\theta_1),
\end{aligned}$$

$$\nabla Z(\theta) = \int t(x)\tilde{p}_{\theta}(x)\mathrm{d}\mu(x)$$

KLD between arbitrary positive densities

$$\begin{aligned} D_{\mathrm{KL}}(\tilde{p} : \tilde{q}) &= H^{\times}(\tilde{p} : \tilde{q}) - H(\tilde{p}), \\ &= \int \left(\tilde{p} \log \frac{\tilde{p}}{\tilde{q}} + \tilde{q} - \tilde{p} \right) \mathrm{d}\mu, \end{aligned}$$

Consider arbitrary densities (not necessarily exp fams): $p(x) = \frac{\tilde{p}(x)}{Z_p}$ $q(x) = \frac{\tilde{q}(x)}{Z_q}$

$$D_{\mathrm{KL}}(\tilde{p} : \tilde{q}) = Z_p \left(D_{\mathrm{KL}}(p : q) + \log \frac{Z_p}{Z_q} \right) + Z_q - Z_p.$$

$$\begin{aligned} H^{\times}(\tilde{p} : \tilde{q}) &= Z_p H^{\times}(p : q) - Z_p \log Z_q + Z_q - 1, \\ H(\tilde{p}) &= Z_p H(p) - Z_p \log Z_p + Z_p - 1, \end{aligned}$$

Above formula when specialized to densities p and q of exponential family:

$$D_{\mathrm{KL}}(\tilde{p}_{\theta_1} : \tilde{p}_{\theta_2}) = \langle \theta_1 - \theta_2, \nabla Z(\theta_1) \rangle + Z(\theta_2) - Z(\theta_1) = B_Z(\theta_2 : \theta_1)$$

α-divergences between unnormalized densities

- Statistical **α-divergences** between positive measures:

$$D_\alpha(\tilde{p}:\tilde{q}) = \begin{cases} \frac{1}{\alpha(1-\alpha)} \int \left(\alpha\tilde{p} + (1-\alpha)\tilde{q} - \tilde{p}^\alpha \tilde{q}^{1-\alpha}\right) \mathrm{d}\mu, & \alpha \notin \{0,1\} \\ D^*_{\mathrm{KL}}(\tilde{p}:\tilde{q}) = D_{\mathrm{KL}}(\tilde{q}:\tilde{p}) & \alpha = 0, \\ 4\, D^2_H(\tilde{p},\tilde{q}) & \alpha = \frac{1}{2}, \\ D_{\mathrm{KL}}(\tilde{p}:\tilde{q}) & \alpha = 1. \end{cases}$$

- When considering unnormalized exponential family densities:

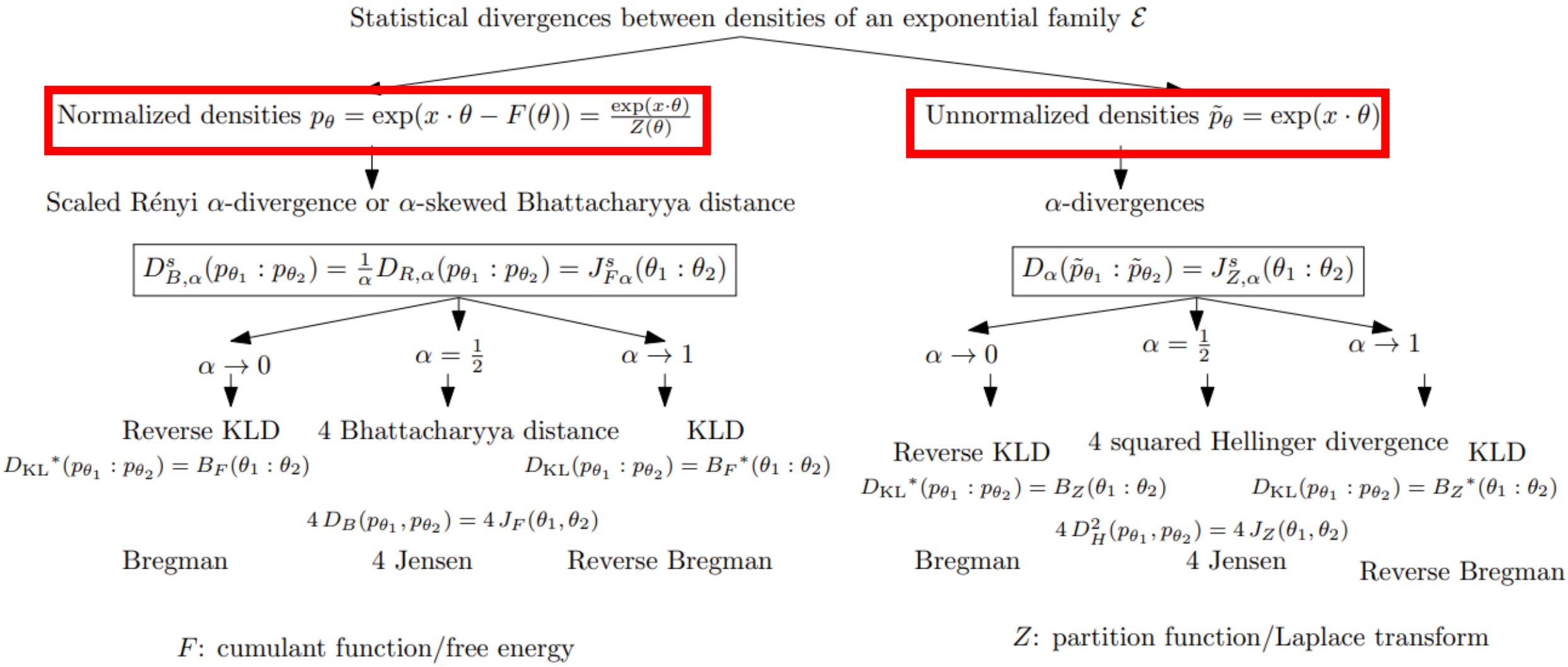
$$D_\alpha(\tilde{p}_{\theta_1}:\tilde{p}_{\theta_2}) = \begin{cases} \frac{1}{\alpha(1-\alpha)} J_{Z,\alpha}(\theta_1:\theta_2), & \alpha \notin \{0,1\} \\ B_Z(\theta_1:\theta_2) & \alpha = 0, \\ 4\, J_Z(\theta_1,\theta_2) & \alpha = \frac{1}{2}, \\ B^*_Z(\theta_1:\theta_2) = B_Z(\theta_2:\theta_1) & \alpha = 1 \end{cases}$$

Partition function Z

Proposition 5. *The α-divergences between unnormalized densities of an exponential family amounts to scaled α-Jensen divergences between their natural parameters for the partition function:*

$$D_\alpha(\tilde{p}_{\theta_1}:\tilde{p}_{\theta_2}) = J^s_{Z,\alpha}(\theta_1:\theta_2).$$

Overview of divergences between (un)normalized EF densities



KLD between normalized and unnormalized densities

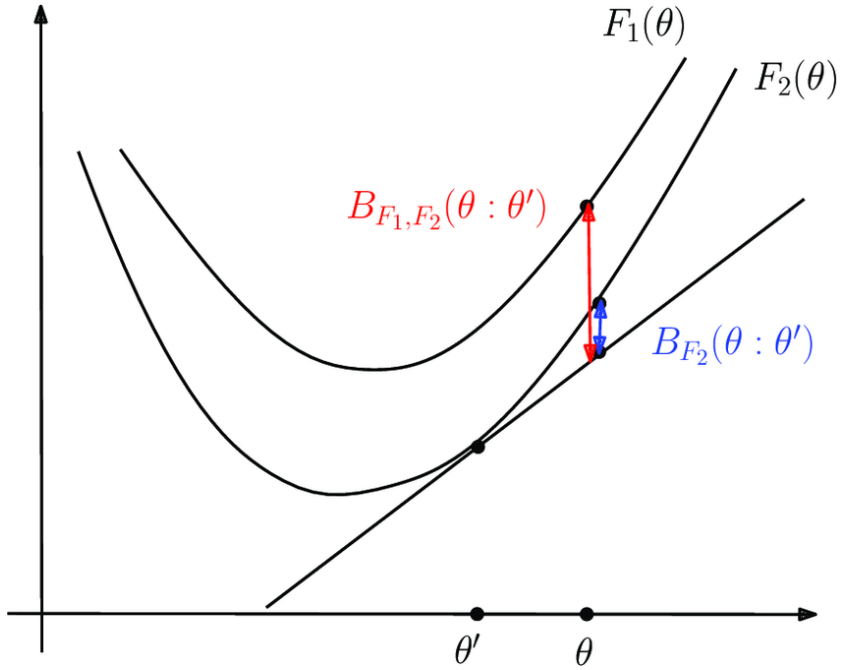
$$\begin{aligned} D_{\mathrm{KL}}(p_{\theta_1} : \tilde{p}_{\theta_2}) &= B_F(\theta_2 : \theta_1) - \log Z(\theta_2) + Z(\theta_2) - 1, \\ &= Z(\theta_2) - 1 - F(\theta_1) - \langle \theta_2 - \theta_1, \nabla F(\theta_2) \rangle, \\ &= B_{Z-1,F}(\theta_2 - \theta_1). \end{aligned}$$

with $Z(\theta) - 1 \geq F(\theta)$

where we generalized Bregman divergences to **duo Bregman pseudo-divergences**:

$$B_{F_1,F_2}(\theta_1 : \theta_2) = F_1(\theta_1) - F_2(\theta_2) - \langle \theta_1 - \theta_2, \nabla F_2(\theta_2) \rangle$$

with $F_1(\theta) = Z(\theta) - 1$ *and* $F_2(\theta) = F(\theta)$



Comparative convexity: (M,N)-convexity

- *Definition*: A function Z is **(M,N)-convex** iff for in α in [0,1]:

$$Z(M(x,y;\alpha,1-\alpha)) \leq N(Z(x),Z(y);\alpha,1-\alpha)$$

- Ordinary convexity: (A,A)-convexity wrt to arithmetic weighted mean

$$A(x,y;\alpha,1-\alpha) = \alpha x + (1-\alpha)y$$

- **Log-convexity**: **(A,G)-convexity** wrt to A/geometric weighted means:

$$G(x,y;\alpha,1-\alpha) = x^{\alpha}y^{1-\alpha}$$

Comparative convexity wrt quasi-arithmetic means

- Kolmogorov-Nagumo-De Finitti **quasi-arithmetic mean** for a strictly monotone generator h(u):

$$M_h(x,y;\alpha,1-\alpha) = h^{-1}(\alpha h(x) + (1-\alpha)h(x)).$$

- Includes **power means** which are *homogeneous means*:

$$M_p(x,y;\alpha,1-\alpha) = (\alpha x^p + (1-\alpha)y^p)^{\frac{1}{p}} = M_{h_p}(x,y;\alpha,1-\alpha), \quad p \neq 0$$

$$h_p(u) = \frac{u^p - 1}{p} \qquad h_p^{-1}(u) = (1+up)^{\frac{1}{p}}$$

Include the **geometric mean** when p→0

Proposition 6 ([1, 34]). *A function $Z(\theta)$ is strictly (M_ρ, M_τ)-convex with respect to two strictly increasing smooth functions ρ and τ if and only if the function $F = \tau \circ Z \circ \rho^{-1}$ is strictly convex.*

Deforming convex functions with comparative convexity

Since log-convexity is $(A = M_{\mathrm{id}}, G = M_{\log})$-convexity, a function Z is strictly log-convex iff $\log \circ Z \circ \mathrm{id}^{-1} = \log \circ Z$ is strictly convex. We have

$$Z = \tau^{-1} \circ F \circ \rho \Leftrightarrow F = \tau \circ Z \circ \rho^{-1}.$$

We consider deformations with **two strictly monotone functions** which preserve convexity and thus induces family of Bregman and Jensen divergences, and families of dually flat spaces:

$$\underbrace{F = \tau \circ Z \circ \rho^{-1}}_{(M_{\rho^{-1}}, M_{\tau^{-1}})\text{-convex when } Z \text{ is convex}} \quad \underset{(\rho^{-1}, \tau^{-1})\text{-deformation}}{\overset{(\rho, \tau)\text{-deformation}}{\rightleftharpoons}} \quad \underbrace{Z = \tau^{-1} \circ F \circ \rho}_{(M_\rho, M_\tau)\text{-convex when } F \text{ is convex}}$$

Deform both (1) the function F (by $\boldsymbol{\tau}^{-1}$) and (2) the argument θ (by ρ) by considering functions $Z = \boldsymbol{\tau}^{-1}(F(\theta))$

Generalizing Bregman divergences with (M,N)-convexity

- Skew Jensen comparative convexity divergence:

Definition:

$$J_{F,\alpha}^{M,N}(p:q) = N_\alpha(F(p),F(q)) - F(M_\alpha(p,q)).$$

Non-negative for **(M,N)-convex generators** F provided regular means M and N (e.g. power means)

Definition 5 (Bregman Comparative Convexity Divergence, BCCD) *The Bregman Comparative Convexity Divergence (BCCD) is defined for a strictly (M,N)-convex function $F : I \to \mathbb{R}$ by*

$$B_F^{M,N}(p:q) = \lim_{\alpha\to 1^-} \frac{1}{\alpha(1-\alpha)} J_{F,\alpha}^{M,N}(p:q) = \lim_{\alpha\to 1^-} \frac{1}{\alpha(1-\alpha)} (N_\alpha(F(p),F(q))) - F(M_\alpha(p,q))) \quad (31)$$

Analogy to limit of skewed Jensen divergences amount to forward/reverse Bregman divergences.

Generalizing Bregman divergences with quasi-arithmetic mean convexity

Theorem 1 (Quasi-arithmetic Bregman divergences, QABD) *Let $F : I \subset \mathbb{R} \to \mathbb{R}$ be a real-valued (M_ρ, M_τ)-convex function defined on an interval I for two strictly monotone and differentiable functions ρ and τ. The quasi-arithmetic Bregman divergence (QABD) induced by the comparative convexity is:*

$$B_F^{\rho,\tau}(p:q) = \frac{\tau(F(p)) - \tau(F(q))}{\tau'(F(q))} - \frac{\rho(p) - \rho(q)}{\rho'(q)} F'(q). \tag{45}$$

Amounts to a **conformal Bregman divergence**:

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

Conformal factor

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

Remark: Conformal Bregman divergences may yield **robustness** in applications

References (partial list)

- John Aczel. **A generalization of the notion of convex functions**. Det Kongelige Norske Videnskabers Selskabs Forhandlinger, Trondheim, 19(24):87–90, 1947
- Katy S Azoury and Manfred K Warmuth. **Relative loss bounds for on-line density estimation with the exponential family of distributions**. Machine learning, 43:211–246, 2001.
- Ole Barndorff-Nielsen. **Information and exponential families**. John Wiley & Sons, 2014.
- Kazuhiro Ishige, Paolo Salani, and Asuka Takatsu. **Hierarchy of deformations in concavity**. Information Geometry, pages 1–19, 2022.
- Thomas Kailath. **The divergence and Bhattacharyya distance measures in signal selection**. IEEE transactions on communication technology, 15(1):52–60, 1967.
- Andrei Kolmogorov. **Sur la notion de la moyenne**. G. Bardi, tip. della R. Accad. dei Lincei, 1930.
- Constantin Niculescu and Lars-Erik Persson. **Convex functions and their applications**, volume 23. Springer, 2018. second edition
- Tim Van Erven and Peter Harremos. **Renyi divergence and Kullback-Leibler divergence**. IEEE Transactions on Information Theory, 60(7):3797–3820, 2014.
- Jun Zhang and Ting-Kam Leonard Wong. **λ-deformed probability families with subtractive and divisive normalizations**. In Handbook of Statistics, volume 45, pages 187–215. Elsevier, 2021.

References and related works

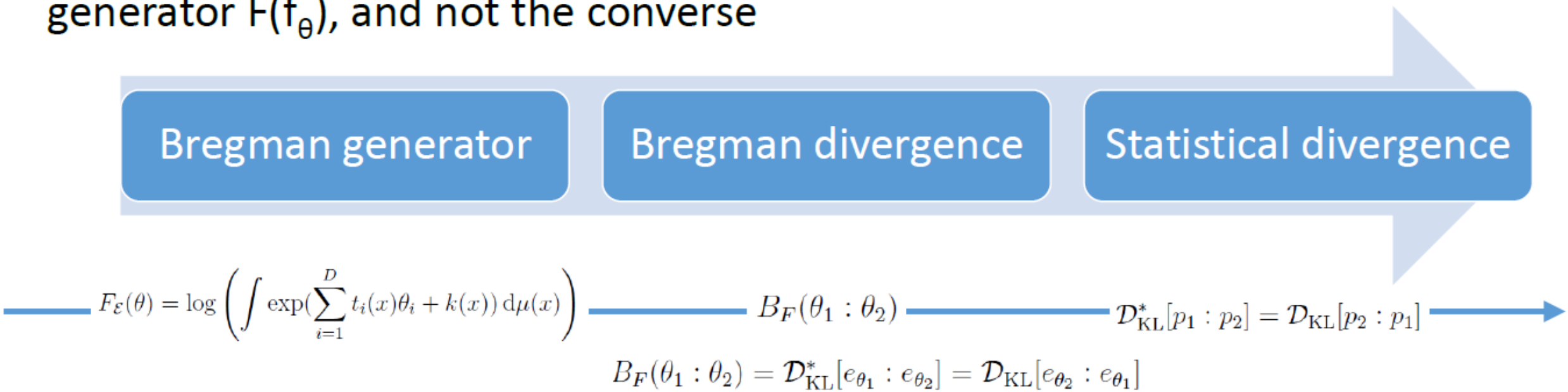
- NF and Richard Nock. "*Generalizing skew Jensen divergences and Bregman divergences with comparative convexity*." *IEEE Signal Processing Letters* 24.8 (2017): 1123-1127.
- "*Divergences induced by dual subtractive and divisive normalizations of exponential families and their convex deformations*." *arXiv preprint arXiv:2312.12849* (2023).
- "*Statistical divergences between densities of truncated exponential families with nested supports: Duo Bregman and duo Jensen divergences*." *Entropy* 24.3 (2022): 421.
- Nock, Richard, FN, and Shun-ichi Amari. "*On conformal divergences and their population minimizers.*" *IEEE Transactions on Information Theory* 62.1 (2015): 527-538.
- Rob Brekelmans and FN. “Variational representations of annealing paths: Bregman information under monotonic embedding”, Information geometry, 2024. Doi: 10.1007/s41884-023-00129-6
- "Quasi-arithmetic Centers, Quasi-arithmetic Mixtures, and the Jensen-Shannon-Divergences." *International Conference on Geometric Science of Information*. Cham: Springer Nature Switzerland, 2023.

- Kullback-Leibler divergence: relative entropy
- Exponential families: Discrete, continuous, measures
- KLD between densities of an EF
- Information geometry of convex function: Dually flat space
- Information geometry of divergence
- Bhattacharyya distance and Rényi divergence
- Jensen divergence
- Overview of classical divergences
- Partition function is log-convex and hence convex
- Bregman divergence wrt Z: KLD between unnormalized EF densities
- Jensen divergence wrt Z: alpha divergences
- Overview of divergences between (un)normalized EF densities
- Comparative convexity
- Comparative convexity wrt quasi-arithmetic means
- Deforming convex functions wrt quasi-arithmetic generators
- (M,N)-Jensen divergence
- (M,N)-Bregman divergence
- Equivalence with a conformal Bregman divergence
- Power Bregman divergences
- Conclusion

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



An elementary introduction to information geometry." Entropy 22.10 (2020)

Bhattacharyya arc: Likelihood Ratio Exponential Family

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

$$\mathcal{E}(p,q) := \left\{ p_\lambda(x) := \frac{p^{1-\lambda}(x)q^\lambda(x)}{Z_\lambda^G(p,q)}, \quad \lambda \in (0,1) \right\} \qquad Z_\lambda^G(p,q) := \int_{\mathcal{X}} p^{1-\lambda}(x)q^\lambda(x)\mathrm{d}\mu(x)$$

- Log-normalizer F(λ) (aka cumulant generating function, log partition function):
- Bhattacharyya arc (geometric mixtures) = **1D exponential family:**

$$\begin{aligned} p_\lambda(x) &= \frac{p_0^{1-\lambda}(x)p_1^\lambda(x)}{Z_\lambda^G(p,q)} \\ &= p_0(x)\exp\left(\lambda\log\left(\frac{p_1(x)}{p_0(x)}\right) - \log Z_\lambda^G(p,q)\right) \\ &= \exp\left(\lambda t(x) - F(\lambda) + k(x)\right) \\ F(\lambda) &:= \log(Z_\lambda^G(p,q)) = \log\left(\int_{\mathcal{X}} p^{1-\lambda}(x)q^\lambda(x)\mathrm{d}\mu(x)\right) \\ &=: -D_\lambda^{\mathrm{Bhat}}[p:q] \end{aligned}$$

Log-likelihood sufficient statistics:

$$t(x) := \log\left(\frac{p_1(x)}{p_0(x)}\right)$$

Base measure is p_0 $\quad k(x) := \log p_0(x)$

$$D_\alpha^{\mathrm{Bhat}}[p:q] := -\log\left(\int_{\mathcal{X}} p^{1-\alpha}(x)q^\alpha(x)\mathrm{d}\mu(x)\right)$$

Generalizing the Geometric Annealing Path using Power Means, UAI 2021

Likelihood Ratio Exponential Families, NeurIPS Workshop on Deep Learning through Information Geometry 2020

Weighted quasi-arithmetic means when α tends to zero:

$M_{\tau}(p,q;1-\alpha;\alpha) = p + \frac{\alpha(\tau(q)-\tau(p))}{\tau'(p)}$