Information Geometry from Divergences

Armin van de Venn, **Prof. Tomoi Koide**

June 30, 2026

Laboratory of Theoretical Physics
University of Tartu

Background

Idea: Use Differential Geometry to study Statistics/Probability Theory

- Fisher information matrix:

$$g_{ij}(\theta) = \mathbb{E}_{p(x|\theta)}\left[\frac{\partial}{\partial\theta_i}\log p(x|\theta)\,\frac{\partial}{\partial\theta_j}\log p(x|\theta)\right]$$

  "How sensitive is the distribution to small changes of the parameters"
- C. R. Rao noticed (≈ 1945) that this defines a Riemannian metric on a parametric family of probability distributions
- Shun'ichi Amari (pioneer of AI!) developed the modern framework of information geometry, including dual affine connections and divergences in the late 1970s and early 1980s

Statistical Manifolds

A **statistical manifold** is a smooth family of probability distributions

$$\mathcal{M} = \left\{ p(x|\theta) \mid \theta = (\theta^1, \ldots, \theta^n) \in \Theta \subseteq \mathbb{R}^n \right\}$$

- Each element of $\mathcal{M}$ is a probability distribution
- Topology usually inherited from Θ
- Smooth structure defined with the charts $\varphi\colon \mathcal{M} \to \mathbb{R}^n$, $\varphi(p(x|\theta)) = \theta$
- The parameters θ^i act as local coordinates on $\mathcal{M}$, thus $\mathcal{M}$ is n-dimensional

Example:

$$\mathcal{M} = \left\{ \mathcal{N}(x|\mu, \sigma^2) \mid (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_{>0} \subseteq \mathbb{R}^2 \right\}$$

is a two-dimensional statistical manifold.

Divergences

Metric

A metric is a map

$$d\colon X \times X \to \mathbb{R}$$

satisfying

(i) $d(x, y) \geq 0$

(ii) $d(x, y) = 0 \iff x = y$

(iii) $d(x, y) = d(y, x)$

(iv) $d(x, z) \leq d(x, y) + d(y, z)$

Divergence

A divergence is a smooth map

$$D\colon \mathcal{M} \times \mathcal{M} \to \mathbb{R}$$

satisfying

(i) $D(p, q) \geq 0$

(ii) $D(p, q) = 0 \iff p = q$

In general D is not symmetric and does not satisfy the triangle inequality.

Induced Geometry

Observation: A divergence has a minimum along the diagonal

- Near the diagonal $\theta' = \theta$, the linear terms vanish and

$$D(\theta, \theta + d\theta) = \frac{1}{2} H_{ij}(\theta) d\theta^i d\theta^j + O(d\theta^3),$$

where

$$H_{ij}(\theta) \coloneqq \partial'_i \partial'_j D\big|_{\theta=\theta'}, \quad \partial'_i D \coloneqq \frac{\partial}{\partial \theta'^i} D(\theta, \theta').$$

Additional Requirement

We require that H_{ij} is positive definite, i.e.

$$H_{ij} v^i v^j > 0 \qquad \text{for all } v \neq 0.$$

Induced Geometry

For two vector fields $X, Y \in \mathfrak{X}(\mathcal{M})$ we define

$$D[X|Y] \coloneqq X^{(1)} Y^{(2)} D\Big|_{\theta=\theta'},$$

where $X^{(1)}$ and $Y^{(2)}$ are the canonical lifts to the first and second factor of $\mathcal{M} \times \mathcal{M}$: $X^{(1)}_{(p,q)} = (X_p, 0)$, $Y^{(2)}_{(p,q)} = (0, Y_q)$. In other words $X^{(1)} = X^i(\theta)\frac{\partial}{\partial \theta^i}$, $Y^{(2)} = Y^i(\theta')\frac{\partial}{\partial \theta'^i}$.

Metric Tensor

$$g(X, Y) \coloneqq -D[X|Y]$$

$$g_{ij} = -\,\partial_i \partial'_j D\big|_{\theta=\theta'} = \partial'_i \partial'_j D\big|_{\theta=\theta'}$$

Connection

$$g(\nabla_X Y, Z) \coloneqq -D[XY|Z]$$

$$\Gamma^l{}_{ji} g_{kl} = -\,\partial_i \partial_j \partial'_k D\big|_{\theta=\theta'}$$

Dual Divergence

Existence of D implies the existence of the **dual divergence**

$$D^*(p, q) \coloneqq D(q, p).$$

It satisfies the properties of a divergence as well. D^* induces

- $g^*_{ij} = -\partial_i \partial'_j D(\theta', \theta)\big|_{\theta'=\theta} = -\partial_j \partial'_i D(\theta, \theta')|_{\theta=\theta'} = g_{ji}$
- $\Gamma^* \neq \Gamma$

One can show:

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

Two connections ∇, ∇^* that satisfy this relation are said to be **dual** wrt the metric g.

Dual Divergence

Notice that

$$\Gamma^{l}{}_{ji} g_{kl} = -\partial_i \partial_j \dots \qquad \Gamma^{*l}{}_{ji} g_{kl} = -\partial'_i \partial'_j \dots$$

hence ∇ and ∇^* are both **torsion-free**. However:

$$\nabla_i g_{jk} = -T_{jik} \qquad \nabla^*_i g_{jk} = T_{kij}\,,$$

where

$$T^{l}_{ik} \coloneqq \Gamma^{l}{}_{ik} - \Gamma^{*l}{}_{ik}$$

is the so-called **Amari–Chentsov Tensor** (or **cubic tensor**) which we know as the nonmetricity.

Nonmetricity naturally emerges within the setup of information geometry and is not added by hand! In fact, nonmetricity arises due to the asymmetry of the divergence!

Case Study: Exponential Families

Probability distributions of form

$$p(x|\theta) = h(x)\exp\big(\theta^i F_i(x) - \psi(\theta)\big), \qquad \psi(\theta) = \ln\left(\int h(x)\exp\big(\theta^i F_i(x)\big)\mathrm{d}x\right).$$

- **Example:** Gaussian distributions

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

with choice $h(x) = 1$, $F(x) = (x, x^2)$, $(\theta^1, \theta^2) = (\mu/\sigma^2, -1/(2\sigma^2))$, and

$$\psi(\theta) = -\frac{(\theta^1)^2}{4\theta^2} + \frac{1}{2}\ln(\pi) - \frac{1}{2}\ln\big(-\theta^2\big).$$

Case Study: Exponential Families

Standard choice of divergence: **Kullback–Leibler (KL) divergence**

$$D_{\mathrm{KL}}(p, q) = \int p(x) \ln\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x.$$

- Related to maximum likelihood estimation
- Induces Fisher information metric
- Behaves naturally for exponential families

Induced geometry:

$$g_{ij}(\theta) = \partial_i \partial_j \psi(\theta)$$

$$\Gamma^{l}{}_{ij}(\theta) g_{kl}(\theta) = 0$$

$$\Gamma^{*l}{}_{ij}(\theta) g_{kl}(\theta) = \partial_i \partial_j \partial_k \psi(\theta)$$

Dual Coordinates

Since $g_{ij}(\theta) = \partial_i \partial_j \psi(\theta)$ is positive definite, $\psi(\theta)$ is strictly convex. Therefore we may Legendre-transform

$$\phi(\eta) = \theta^i \eta_i - \psi(\theta)$$

with

$$\eta_i = \frac{\partial \psi(\theta)}{\partial \theta^i}.$$

The coordinates η are called **dual coordinates** (dual to θ).

- **Example:** Gaussian distributions

$$\theta^1 = \frac{\eta_1}{\eta_2 - \eta_1^2}, \qquad \theta^2 = \frac{1}{2\left(\eta_1^2 - \eta_2\right)}$$

and

$$\phi(\eta) = -\frac{1}{2} \ln\bigl(2\pi e(\eta_2 - \eta_1^2)\bigr).$$

Application: Brownian Bridge

A **Brownian bridge** (or **pinned Brownian motion**) describes Brownian motion that has a fixed start $(x_a = 0, t_a = 0)$ and a fixed end (x_b, t_b). Its probability distribution at some $t \in (0, t_b)$ is given by a Gaussian distribution:

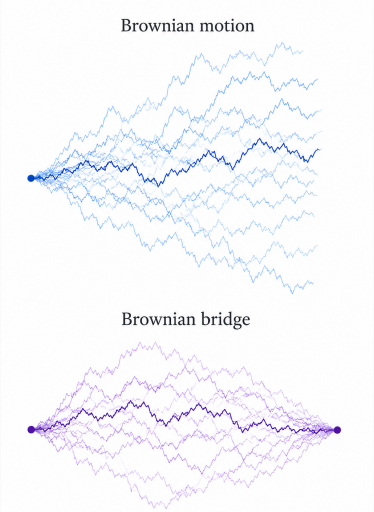
$$\mathcal{N}\big(x \mid \mu_{\mathrm{BB}}(t), \sigma^2_{\mathrm{BB}}(t)\big)\,,$$

where

$$\mu_{\mathrm{BB}}(t) = \frac{t}{t_b} x_b,$$

$$\sigma^2_{\mathrm{BB}}(t) = 2\mathfrak{D}\frac{t(t_b - t)}{t_b}.$$

Here, $\mathfrak{D}$ is the diffusion coefficient.



Application: Brownian Bridge

Idea: Use information geometry in order to describe the Brownian bridge

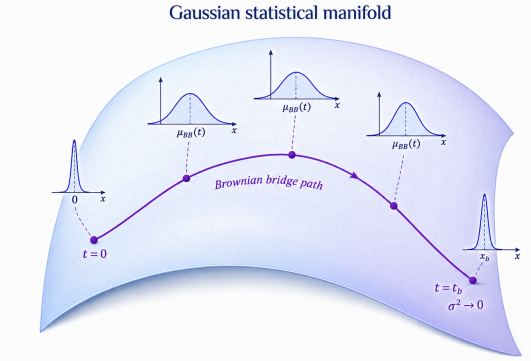
At any $t \in (0, t_b)$ the distribution of the Brownian Bridge is Gaussian. Consider a path within the Gaussian statistical manifold:

$$(\theta^1, \theta^2) = \left(\frac{\mu_{\text{geo}}(t)}{\sigma^2_{\text{geo}}(t)}, \frac{-1}{2\sigma^2_{\text{geo}}(t)} \right)$$

and

$$\psi(\theta) = -\frac{(\theta^1)^2}{4\theta^2} + \frac{1}{2}\ln(\pi) - \frac{1}{2}\ln(-\theta^2).$$

Parameters θ^i are a bit awkward: $\sigma_{\text{BB}} \to 0$ as $t \to t_b$, therefore $\theta^i \to \infty$.
Consider instead the dual coordinates.



Application: Brownian Bridge

The dual coordinates are

$$\eta_1 = \frac{\partial \psi(\theta)}{\partial \theta^1} = -\frac{\theta^1}{2\theta^2} = \mu_{\text{geo}}(t)$$
$$\eta_2 = \frac{\partial \psi(\theta)}{\partial \theta^2} = \frac{(\theta^1)^2}{4(\theta^2)^2} - \frac{1}{2\theta^2} = \mu^2_{\text{geo}}(t) + \sigma^2_{\text{geo}}(t).$$

Dual connection (using KL divergence) in dual coordinates:

$$\Gamma^{*k}{}_{ij}(\eta) = 0.$$

Hence, geodesics in dual coordinates can be described by straight lines. Let us therefore make the ansatz

$$\eta_i(t) = c_i t + d_i.$$

Application: Brownian Bridge

We already know the mean and variance of a Brownian Bridge at the start and endpoint, hence we can find the coefficients of the straight lines. Then:

$$\mu_{\text{geo}}(t) = \frac{t}{t_b}\, x_b,$$

$$\sigma^2_{\text{geo}}(t) = \frac{x_b^2}{t_b^2}\, t(t_b - t).$$

Comparing, we indeed find $\mu_{\text{BB}}(t) = \mu_{\text{geo}}(t)$. Furthermore $\sigma^2_{\text{BB}}(t) = \sigma^2_{\text{geo}}(t)$ only if the diffusion coefficient is given by

$$\mathfrak{D} = \frac{1}{2}\, \frac{x_b^2}{t_b}.$$

Time evolution of a **canonical** Brownian bridge is identical to a geodesic on the associated statistical manifold!

Relation to Gravity & AI

Analogies with Gravity:

- spacetime $\leftrightarrow$ statistical manifold
- spacetime interval $\leftrightarrow$ statistical distinguishability
- nonmetricity $\leftrightarrow$ Amari-Chentsov tensor
- free fall along geodesics $\leftrightarrow$ geodesic motion of stochastic processes?

(Bold) Conjecture

Stochastic processes are described by geodesics on an appropriate statistical manifold

Applications in AI:

- Optimization: Natural gradient descent
- Classification: Pattern recognition and clustering
- Generative AI: diffusion models

. . . see e.g. Amari, "Information Geometry and Its Applications"; Karczewski et al. arXiv:2505.17517;

Thank you!