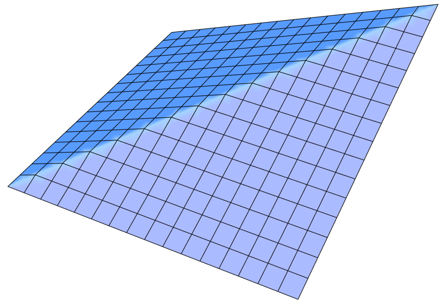

# STA 4531H: Information Geometry

Prof. Leonard Wong • Fall 2022 • University of Toronto
M 12:00-2:00PM and W 1:00-2:00PM in MY 320

---

Transcribed, LaTeXed, and edited by Rob Zimmerman

**Note:** These are personal lecture notes, not official course materials. They may contain errors (which by default you should attribute to me, Rob Zimmerman, rather than the course instructor). The permanent home of these — and all of my other lecture notes — is `https://rob-zimmerman.github.io/coursenotes`.

## Contents

Lecture 1 — Monday, September 12, 2022

# 1. Statistical Models and Divergences

## Statistical Models and Coordinate Systems

Information geometry brings geometry into probability and statistics in two distinct ways. The state space of an observation may itself be non-Euclidean, as in directional statistics on a sphere. Alternatively, a family of probability distributions may be treated as a geometric space whose points are distributions. The second viewpoint is the starting point here. A divergence will first provide a local notion of separation, and its second- and third-order behaviour will later produce a Riemannian metric and a pair of affine connections.

Fix a measure space $(\Omega, \mathcal{A}, \nu)$, where $\nu$ is a dominating measure.

**Definition 1.1** A **dominated statistical model** is a family

$$\mathcal{M} = \{\mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$$

of probability measures on $(\Omega, \mathcal{A})$ such that $\mathbb{P}_{\boldsymbol{\theta}} \ll \nu$ for every $\boldsymbol{\theta} \in \Theta$. Its densities are

$$p_{\boldsymbol{\theta}}(x) \coloneqq \frac{\mathrm{d}\mathbb{P}_{\boldsymbol{\theta}}}{\mathrm{d}\nu}(x).$$

The model is **identifiable** when $\mathbb{P}_{\boldsymbol{\theta}} = \mathbb{P}_{\boldsymbol{\theta}'}$ implies $\boldsymbol{\theta} = \boldsymbol{\theta}'$.

The parameter set will ordinarily be an open subset of $\mathbb{R}^d$. This keeps the discussion finite dimensional; an infinite dimensional family requires additional topological and functional analytic structure.

For a first example, fix $\sigma > 0$. The normal location model is

$$\mathcal{M} = \{\mathcal{N}(\boldsymbol{\theta}, \sigma^2 \mathbf{I}_d) : \boldsymbol{\theta} \in \mathbb{R}^d\},$$

with Lebesgue density

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{\theta}\|^2}{2\sigma^2}\right).$$

The parameter space is Euclidean, and the divergence between two members will recover a scalar multiple of squared Euclidean distance.

The basic discrete model is the open simplex

$$\Delta^d := \left\{ \boldsymbol{p} = (p_0, \ldots, p_d) \in (0,1)^{d+1} : \sum_{i=0}^{d} p_i = 1 \right\}. \tag{1.1}$$

Its points are the strictly positive probability distributions on a set with $d+1$ elements. One global coordinate system is given by the log odds

$$\theta^i = \log\left(\frac{p_i}{p_0}\right), \quad i = 1, \ldots, d.$$

Writing $\theta^0 = 0$, the inverse transformation is

$$p_i(\boldsymbol{\theta}) = \frac{\mathrm{e}^{\theta^i}}{\sum_{j=0}^{d} \mathrm{e}^{\theta^j}} = \exp\bigl(\theta^i - \phi(\boldsymbol{\theta})\bigr) \qquad \text{and} \qquad \phi(\boldsymbol{\theta}) = \log\left(1 + \sum_{j=1}^{d} \mathrm{e}^{\theta^j}\right). \tag{1.2}$$

Thus the boundary of the closed simplex corresponds to points at infinity in the exponential coordinates.

**Definition 1.2** Let $\boldsymbol{F} = (F_1, \ldots, F_d) \colon \Omega \to \mathbb{R}^d$ be a vector of statistics. An **exponential family** has densities

$$p_{\boldsymbol{\theta}}(x) = \exp\bigl(\boldsymbol{\theta}^\top \boldsymbol{F}(x) - \phi(\boldsymbol{\theta})\bigr), \tag{1.3}$$

where the natural parameter space is

$$\Theta = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d : \int_\Omega \mathrm{e}^{\boldsymbol{\theta}^\top \boldsymbol{F}(x)} \, \mathrm{d}\nu(x) < \infty \right\}$$

and the cumulant generating function is

$$\phi(\boldsymbol{\theta}) = \log\left(\int_\Omega \mathrm{e}^{\boldsymbol{\theta}^\top \boldsymbol{F}(x)} \, \mathrm{d}\nu(x)\right).$$

The natural parameter space is convex, although it need not be open, and $\phi$ is convex. The simplex representation in (1.2) is an exponential family with indicator statistics. The normal location family is also an exponential family: using $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$ as the reference measure gives

$$\frac{\mathrm{d}\mathcal{N}(\boldsymbol{\theta}, \sigma^2 \mathbf{I}_d)}{\mathrm{d}\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)}(\boldsymbol{x}) = \exp\left(\frac{\boldsymbol{\theta}^\top \boldsymbol{x}}{\sigma^2} - \frac{\|\boldsymbol{\theta}\|^2}{2\sigma^2}\right).$$

By contrast, a location-scale Student distribution with fixed degrees of freedom is not an exponential family, and neither is a uniform model whose support depends on its parameter.

**Definition 1.3** Given densities $p_0, \ldots, p_d$ with respect to the same reference measure, their **mixture family** consists of

$$p(x; \boldsymbol{w}) = \sum_{i=0}^{d} w_i p_i(x), \quad \boldsymbol{w} \in \overline{\Delta}^d.$$

The open simplex is simultaneously an exponential family and a mixture family. These descriptions lead to different interpolations. For two densities with common support, the mixture interpolation is

$$p_t^{(m)}(x) = (1-t)p_0(x) + tp_1(x), \quad 0 \le t \le 1,$$

whereas the exponential interpolation is

$$p_t^{(e)}(x) = \frac{p_0(x)^{1-t} p_1(x)^t}{Z(t)} \qquad \text{and} \qquad Z(t) = \int_\Omega p_0(x)^{1-t} p_1(x)^t \, \mathrm{d}\nu(x).$$

These paths will become geodesics for two different affine connections. Their distinct geometry on a simplex with three points is suggested in Figure 1.

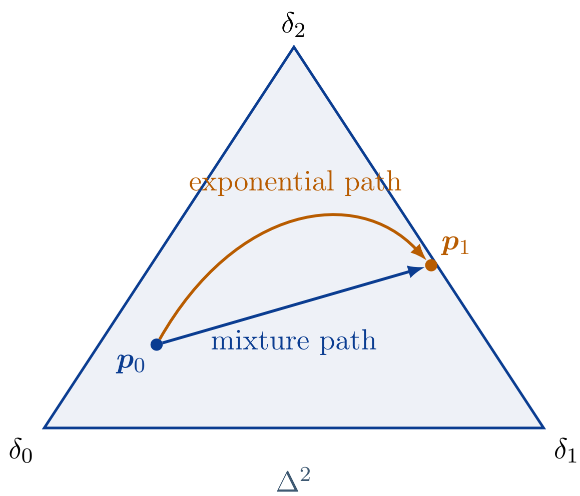


Figure 1

**Exercise 1.4** Let $p_0$ and $p_1$ be the densities of $\mathcal{N}(\mu_0, \sigma_0^2)$ and $\mathcal{N}(\mu_1, \sigma_1^2)$. Show that their exponential interpolation is normal, find its mean and variance, and compare it with the ordinary mixture interpolation.

### Lecture 2 — Wednesday, September 14, 2022

## Smooth Manifolds

A parameterization is a coordinate choice rather than part of the underlying statistical object. For the family $\mathcal{N}(\mu, \sigma^2)$ on the real line, we may use the mean and standard deviation $(\mu, \sigma)$, the natural parameters

$$\theta^1 = \frac{\mu}{\sigma^2} \qquad \text{and} \qquad \theta^2 = -\frac{1}{2\sigma^2},$$

or the moment parameters

$$\eta_1 = \mu \qquad \text{and} \qquad \eta_2 = \mu^2 + \sigma^2.$$

The transformations among these parameterizations are diffeomorphisms. Information geometry seeks constructions whose meaning does not depend on which of them is selected.

**Definition 2.1** Let $M$ be a Hausdorff, second countable topological space. A **chart** is a pair $(U, \varphi)$ in which $U \subseteq M$ is open and $\varphi\colon U \to \varphi(U) \subseteq \mathbb{R}^d$ is a homeomorphism onto an open set. Two charts $(U, \varphi)$ and $(V, \psi)$ are **smoothly compatible** when the transition maps

$$\psi \circ \varphi^{-1} \qquad \text{and} \qquad \varphi \circ \psi^{-1}$$

are smooth wherever they are defined. A **smooth atlas** is a covering collection of pairwise compatible charts, and a **smooth $d$-dimensional manifold** is $M$ equipped with a maximal smooth atlas.

The compatibility condition ensures that differentiability is independent of the chart. The picture with two charts in Figure 2 makes the local Euclidean structure explicit.

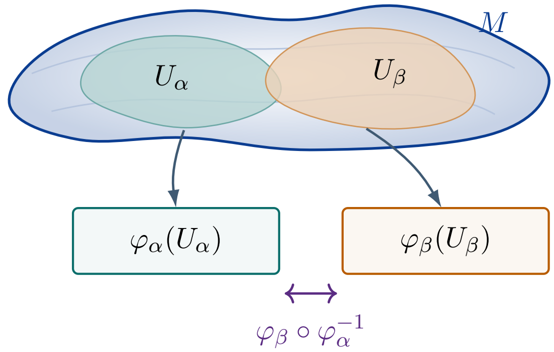


FIGURE 2

**Definition 2.2** Let $M$ and $N$ be smooth manifolds. A map $F\colon M \to N$ is **smooth** when every local coordinate representation

$$\psi \circ F \circ \varphi^{-1}$$

is smooth in the Euclidean sense. A bijection $F$ is a **diffeomorphism** when both $F$ and $F^{-1}$ are smooth.

An open subset of $\mathbb{R}^d$ has its standard smooth structure. Consequently, an identifiable model $\{\mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ inherits a smooth structure from an open parameter domain $\Theta$ whenever the parameterization and its inverse are sufficiently regular. Products of charts similarly give the product manifold $M \times N$.

For the normal family, the cumulant generating function is a single smooth function on the statistical manifold even though its formula changes with the coordinates. Using $\mathrm{d}\nu(x) = \mathrm{d}x/\sqrt{2\pi}$ avoids an inessential additive constant. In the natural coordinates above,

$$\phi(\boldsymbol{\theta}) = -\frac{(\theta^1)^2}{4\theta^2} - \frac{1}{2}\log\bigl(-2\theta^2\bigr),$$

whereas in the coordinates $(\mu, \sigma)$ it is

$$\phi(\mu, \sigma) = \frac{\mu^2}{2\sigma^2} + \log \sigma.$$

The formulas represent the same scalar field. This distinction between an intrinsic object and its coordinate representation will recur for metrics, connections, and curvature.

## Lecture 3 — Monday, September 19, 2022

### Statistical Divergences

A metric is often too restrictive for statistical comparison. Relative entropy, for example, is asymmetric and does not satisfy the triangle inequality, but its orientation records which distribution plays the role of the reference model.

**Definition 3.1** Let $M$ be a set. A **divergence** is a map

$$D\colon M \times M \to [0, \infty]$$

such that $D(p\|q) = 0$ if and only if $p = q$. The reversed divergence is

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

No symmetry or triangle inequality is required. Additional smoothness and nondegeneracy assumptions will be imposed when a divergence is used to induce differential geometry.

**Definition 3.2** For probability measures $\mathbb{P}$ and $\mathbb{Q}$ on the same measurable space, the **Kullback–Leibler divergence**, or **relative entropy**, is

$$\mathrm{KL}(\mathbb{P}\|\mathbb{Q}) \coloneqq \begin{cases} \int_\Omega \log\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right) \mathrm{d}\mathbb{P}, & \mathbb{P} \ll \mathbb{Q}, \\ +\infty, & \mathbb{P} \not\ll \mathbb{Q}. \end{cases} \tag{3.1}$$

If $\mathbb{P}$ and $\mathbb{Q}$ have densities $p$ and $q$ with respect to $\nu$, then

$$\mathrm{KL}(\mathbb{P}\|\mathbb{Q}) = \int_\Omega p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}\nu(x).$$

On a finite state space, this becomes

$$\mathrm{KL}(\boldsymbol{p}\|\boldsymbol{q}) = \sum_{i=0}^{d} p_i \log\left(\frac{p_i}{q_i}\right).$$

**Proposition 3.3** Relative entropy is nonnegative, and $\mathrm{KL}(\mathbb{P}\|\mathbb{Q}) = 0$ if and only if $\mathbb{P} = \mathbb{Q}$.

**Proof.** If $\mathbb{P} \not\ll \mathbb{Q}$, the divergence is infinite by definition. Suppose, therefore, that $\mathbb{P} \ll \mathbb{Q}$, and put

$$R = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}.$$

The inequality $\log r \geq 1 - 1/r$ for $r > 0$ gives

$$\mathrm{KL}(\mathbb{P}\|\mathbb{Q}) = \int_{\{R>0\}} R \log R \,\mathrm{d}\mathbb{Q} \geq \int_{\{R>0\}} (R-1)\,\mathrm{d}\mathbb{Q} = 1 - \mathbb{Q}(R>0) \geq 0.$$

Equality in both inequalities requires $R = 1$ on $\{R > 0\}$ and $\mathbb{Q}(R > 0) = 1$, so $\mathbb{P} = \mathbb{Q}$. The converse is immediate. □

For normal location distributions with common covariance $\sigma^2\mathbf{I}_d$, direct calculation gives

$$\mathrm{KL}\big(\mathcal{N}(\boldsymbol{\theta}, \sigma^2\mathbf{I}_d)\|\mathcal{N}(\boldsymbol{\theta}', \sigma^2\mathbf{I}_d)\big) = \frac{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2}{2\sigma^2}. \tag{3.2}$$

The asymmetry disappears on this restricted family, revealing its Euclidean geometry.

Relative entropy also gives a population interpretation of maximum likelihood. Let $p_{\boldsymbol{\theta}}$ be a dominated model, and let $X_1, \ldots, X_n$ be independent and identically distributed observations from a density $p$, which need not belong to the model. The log likelihood is

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(X_i).$$

Taking the expected normalized log likelihood gives

$$\mathbb{E}_p\left[\frac{1}{n}\ell_n(\boldsymbol{\theta})\right] = \int_\Omega p(x) \log p_{\boldsymbol{\theta}}(x)\,\mathrm{d}\nu(x).$$

On the other hand,

$$\mathrm{KL}(p\|p_{\boldsymbol{\theta}}) = \int_\Omega p(x)\log p(x)\,\mathrm{d}\nu(x) - \int_\Omega p(x)\log p_{\boldsymbol{\theta}}(x)\,\mathrm{d}\nu(x) = -H_\nu(p) - \mathbb{E}_p\left[\frac{1}{n}\ell_n(\boldsymbol{\theta})\right],$$

where

$$H_\nu(p) \coloneqq -\int_\Omega p(x)\log p(x)\,\mathrm{d}\nu(x)$$

is the Shannon entropy relative to $\nu$. Since the first term does not depend on $\boldsymbol{\theta}$, maximizing expected log likelihood is equivalent to projecting $p$ onto the model by minimizing $\mathrm{KL}(p\|p_{\boldsymbol{\theta}})$.

The same orientation appears in large deviations. Suppose $X$ takes values in $\{1,\ldots,k\}$ with distribution $\boldsymbol{p}$, and let $\widehat{\boldsymbol{p}}_n$ be the empirical distribution of an independent and identically distributed sample. If a sequence of empirical types $\boldsymbol{q}^{(n)}$ converges to $\boldsymbol{q}$, Stirling's formula yields

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\widehat{\boldsymbol{p}}_n = \boldsymbol{q}^{(n)}\big) = -\,\mathrm{KL}(\boldsymbol{q}\|\boldsymbol{p}). \tag{3.3}$$

Thus atypical empirical distributions have probabilities of exponential order $\exp(-n\,\mathrm{KL}(\boldsymbol{q}\|\boldsymbol{p}))$. The empirical type and the law generating the data have different roles, so asymmetry is essential rather than a defect.

## Relative Entropy and Related Divergences

Several important divergences arise by relaxing or reorganizing properties of relative entropy.

**Definition 3.4** For $\alpha > 0$ with $\alpha \neq 1$, let $\mu$ dominate both $\mathbb{P}$ and $\mathbb{Q}$, and write $p = \mathrm{d}\mathbb{P}/\,\mathrm{d}\mu$ and $q = \mathrm{d}\mathbb{Q}/\,\mathrm{d}\mu$. The **Rényi divergence** of order $\alpha$ is

$$H_\alpha(\mathbb{P}\|\mathbb{Q}) := \frac{1}{\alpha-1} \log\left(\int_\Omega p(x)^\alpha q(x)^{1-\alpha}\,\mathrm{d}\mu(x)\right),$$

with the usual extended-value conventions. This value is independent of the common dominating measure. In particular, when $\alpha > 1$ it is infinite unless $\mathbb{P} \ll \mathbb{Q}$; for $0 < \alpha < 1$, singular components need not force it to be infinite. For a density $p$ with respect to $\nu$, the associated entropy is

$$H_\alpha(p) = \frac{1}{1-\alpha} \log\left(\int_\Omega p(x)^\alpha\,\mathrm{d}\nu(x)\right).$$

Both quantities converge to their Kullback–Leibler and Shannon counterparts as $\alpha \to 1$. The closely related Tsallis entropy is

$$H^{\mathrm{T}}_\alpha(p) = \frac{1}{\alpha-1}\left(1 - \int_\Omega p(x)^\alpha\,\mathrm{d}\nu(x)\right).$$

It is a monotone transformation of $H_\alpha(p)$ and is useful when power-law behaviour replaces exponential behaviour.

**Definition 3.5** Let $f\colon [0,\infty) \to \mathbb{R} \cup \{+\infty\}$ be convex, finite on $(0,\infty)$, and normalized by $f(1) = 0$. For $\mathbb{P} \ll \mathbb{Q}$, the ***f*-divergence** generated by $f$ is

$$D_f(\mathbb{P}\|\mathbb{Q}) \coloneqq \int_\Omega f\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right) \mathrm{d}\mathbb{Q}.$$

Jensen's inequality again gives nonnegativity. If $f$ is strictly convex at 1, equality holds only when $\mathbb{P} = \mathbb{Q}$; without such a condition, the functional need not separate distinct measures. Important choices of the generator include the following.

1. The generator $f(u) = u \log u$ gives $D_f(\mathbb{P}\|\mathbb{Q}) = \mathrm{KL}(\mathbb{P}\|\mathbb{Q})$.
2. The generator $f(u) = -\log u$ gives $D_f(\mathbb{P}\|\mathbb{Q}) = \mathrm{KL}(\mathbb{Q}\|\mathbb{P})$ when the measures are equivalent.
3. The generator $f(u) = |u-1|/2$ gives the total variation distance

$$D_f(\mathbb{P}\|\mathbb{Q}) = \sup_{A \in \mathcal{A}} |\mathbb{P}(A) - \mathbb{Q}(A)|.$$

4. The generator $f(u) = 4(1-\sqrt{u})$ gives a multiple of squared Hellinger distance:

$$D_f(\mathbb{P}\|\mathbb{Q}) = 2\int_\Omega \left(\sqrt{p} - \sqrt{q}\right)^2 \mathrm{d}\nu.$$

5. For $\alpha \neq \pm 1$, the generator

$$f_\alpha(u) = \frac{4}{1-\alpha^2}\left(1 - u^{(1+\alpha)/2}\right)$$

defines the $\alpha$-divergence and, when the measures are equivalent, satisfies

$$D_{f_\alpha}(\mathbb{P}\|\mathbb{Q}) = D_{f_{-\alpha}}(\mathbb{Q}\|\mathbb{P}).$$

Another construction begins with convex approximation rather than a change of generator.

**Definition 3.6** Let $\Omega_0 \subseteq \mathbb{R}^d$ be convex, and let $\phi\colon \Omega_0 \to \mathbb{R} \cup \{+\infty\}$ be convex and differentiable on the relative interior of $\operatorname{dom}\phi$. For $\boldsymbol{x} \in \operatorname{dom}\phi$ and $\boldsymbol{x}' \in \operatorname{ri}(\operatorname{dom}\phi)$, the **Bregman divergence** generated by $\phi$ is

$$B_\phi[\boldsymbol{x} : \boldsymbol{x}'] \coloneqq \phi(\boldsymbol{x}) - \phi(\boldsymbol{x}') - \nabla\phi(\boldsymbol{x}')^\top(\boldsymbol{x} - \boldsymbol{x}'). \tag{3.4}$$

This is the vertical error in the first-order approximation to $\phi$, as illustrated in Figure 3. Convexity makes the error nonnegative, and strict convexity makes it vanish only on the diagonal.

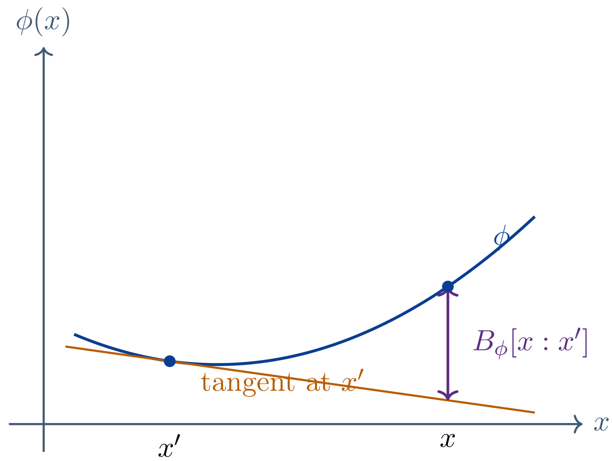


FIGURE 3

For $\phi(\boldsymbol{x}) = \|\boldsymbol{x}\|^2/2$, one obtains

$$B_\phi[\boldsymbol{x} : \boldsymbol{x}'] = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2.$$

For the closed simplex, take negative Shannon entropy,

$$\phi(\boldsymbol{p}) = \sum_{i=0}^{d} p_i \log p_i.$$

Since $(\nabla\phi(\boldsymbol{p}'))_i = 1 + \log p_i'$, substitution into (3.4) gives

$$\begin{aligned} B_\phi[\boldsymbol{p} : \boldsymbol{p}'] &= \sum_{i=0}^{d} \left[p_i \log p_i - p_i' \log p_i' - (1 + \log p_i')(p_i - p_i')\right] \\ &= \sum_{i=0}^{d} p_i \log\left(\frac{p_i}{p_i'}\right) = \mathrm{KL}(\boldsymbol{p}\|\boldsymbol{p}'). \end{aligned}$$

Thus relative entropy on the simplex is both an $f$-divergence and a Bregman divergence. The latter description depends on an affine coordinate system because convexity is not preserved by an arbitrary reparameterization.

There are useful non-Bregman divergences as well. Suppose $d$ assets have positive initial and final prices $x_i'$ and $x_i$, and a portfolio has weights $\pi_i > 0$ with total weight one. Jensen's inequality gives

$$\log\left(\sum_{i=1}^{d} \pi_i \frac{x_i}{x_i'}\right) \geq \sum_{i=1}^{d} \pi_i \log\left(\frac{x_i}{x_i'}\right).$$

The difference is the portfolio's excess growth rate. For a suitable exponentially concave function $\varphi$, it takes the logarithmic divergence form

$$D[\boldsymbol{x} : \boldsymbol{x}'] = \log\Big(1 + \nabla\varphi(\boldsymbol{x}')^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}')\Big) - \big(\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{x}')\big). \tag{3.5}$$

This replaces a linear supporting approximation by a logarithmic one and anticipates the geometry of constant curvature developed later.

Finally, optimal transport compares measures by moving mass through the underlying state space. For probability measures $\mu$ and $\nu$ on $\mathbb{R}^d$ with finite $p$th moments, let $\Pi(\mu, \nu)$ denote their couplings. For $p \geq 1$, the $p$-Wasserstein distance is

$$W_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\boldsymbol{x} - \boldsymbol{y}\|^p \, \mathrm{d}\pi(\boldsymbol{x}, \boldsymbol{y})\right)^{1/p}. \tag{3.6}$$

Unlike relative entropy, it may remain finite for measures with disjoint supports. For instance, if

$$\mu = \mathcal{N}(0, \sigma^2) \qquad \text{and} \qquad \nu = \mathcal{N}(\varepsilon, \sigma^2),$$

then translation is optimal and $W_p(\mu, \nu) = |\varepsilon|$, whereas (3.2) gives

$$\mathrm{KL}(\mu \| \nu) = \frac{\varepsilon^2}{2\sigma^2}.$$

The two comparisons therefore encode genuinely different geometries.

---

**Lecture 4 — Wednesday, September 21, 2022**

# 2. Riemannian Geometry and Fisher Information

## Tangent Spaces and Vector Fields

Differential geometry replaces a nonlinear space by a linear approximation near each point. An embedded surface has a familiar tangent plane, but the definition must not depend on an embedding. The intrinsic construction uses velocities of curves, as suggested in Figure 4.

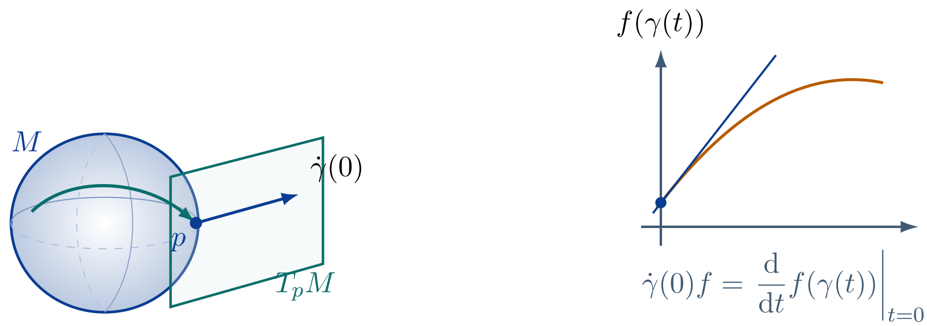


FIGURE 4

Let $\mathcal{F}(M)$ be the collection of smooth functions from $M$ to $\mathbb{R}$. If $\gamma\colon (-\varepsilon, \varepsilon) \to M$ is smooth and $\gamma(0) = p$, its velocity at zero acts on $f \in \mathcal{F}(M)$ by

$$\dot\gamma(0)f := \frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t))\Big|_{t=0}. \tag{4.1}$$

> **Definition 4.1** The **tangent space** $T_pM$ is the set of velocity operators (4.1) of smooth curves through $p$.

Let $(U, \varphi)$ be a chart near $p$, with coordinate functions $x^1, \dots, x^d$. If

$$\varphi(\gamma(t)) = \big(\gamma^1(t), \dots, \gamma^d(t)\big),$$

the chain rule gives

$$\dot\gamma(0)f = \sum_{i=1}^{d} \dot\gamma^i(0) \frac{\partial f}{\partial x^i}(p). \tag{4.2}$$

This motivates the coordinate tangent vector

$$\frac{\partial}{\partial x^i}\Big|_p f := \frac{\partial f}{\partial x^i}(p).$$

**Theorem 4.2** For every $p \in M$, the tangent space $T_pM$ is a $d$-dimensional real vector space. In any local coordinates $x^1, \dots, x^d$, the vectors

$$\left.\frac{\partial}{\partial x^1}\right|_p, \dots, \left.\frac{\partial}{\partial x^d}\right|_p$$

form a basis.

**Proof.** Equation (4.2) shows that every velocity has the form

$$v = \sum_{i=1}^{d} v^i \left.\frac{\partial}{\partial x^i}\right|_p.$$

Conversely, a curve whose coordinate representation has initial velocity $(v^1, \dots, v^d)$ realizes this operator, so $T_pM$ is closed under linear combinations and the displayed vectors span it. If

$$\sum_{i=1}^{d} a^i \left.\frac{\partial}{\partial x^i}\right|_p = 0,$$

apply the operator to each coordinate function $x^j$. This gives $a^j = 0$ for every $j$, proving linear independence. $\square$

Equivalently, a tangent vector is a derivation: it is a map $v \colon \mathcal{F}(M) \to \mathbb{R}$ that is linear over $\mathbb{R}$ and satisfies the Leibniz rule

$$v(fg) = f(p)v(g) + g(p)v(f).$$

This characterization makes the construction entirely intrinsic.

If $y^1, \dots, y^d$ is another coordinate system, the chain rule gives the basis transformation

$$\left.\frac{\partial}{\partial x^i}\right|_p = \sum_{j=1}^{d} \frac{\partial y^j}{\partial x^i}(p) \left.\frac{\partial}{\partial y^j}\right|_p. \tag{4.3}$$

For a smooth map $F \colon M \to N$, a curve $\gamma$ through $p$ is carried to $F \circ \gamma$ through $F(p)$. This defines the **differential**

$$\mathrm{d}F_p \colon T_pM \to T_{F(p)}N \qquad \text{and} \qquad (\mathrm{d}F_p v)h = v(h \circ F).$$

If $x^1, \dots, x^m$ and $y^1, \dots, y^n$ are coordinates on $M$ and $N$, then

$$\mathrm{d}F_p \left( v^j \left. \frac{\partial}{\partial x^j} \right|_p \right) = \sum_{i=1}^{n} \sum_{j=1}^{m} v^j \frac{\partial (y^i \circ F)}{\partial x^j}(p) \left. \frac{\partial}{\partial y^i} \right|_{F(p)}. \tag{4.4}$$

Thus the differential is represented by the usual Jacobian matrix.

The disjoint union

$$TM := \bigcup_{p \in M} T_pM$$

is the **tangent bundle**. Local coordinates on $M$ induce coordinates

$$(x^1, \dots, x^d, v^1, \dots, v^d)$$

on $TM$.

**Definition 4.3** A **vector field** is a smooth map $X\colon M \to TM$ such that $X_p \in T_pM$ for every $p$. In local coordinates it has the form

$$X = \sum_{i=1}^{d} X^i \frac{\partial}{\partial x^i},$$

where the coefficient functions $X^i$ are smooth.

For vector fields $X$ and $Y$, the commutator

$$[X, Y]f := X(Yf) - Y(Xf)$$

is again a vector field and is called their **Lie bracket**.

## Riemannian Metrics

An inner product on each tangent space permits measurements of length and angle. Smooth variation from point to point turns these local measurements into geometry.

**Definition 4.4** A **Riemannian metric** on $M$ is a family of inner products

$$g_p\colon T_pM \times T_pM \to \mathbb{R}, \quad p \in M,$$

such that $p \mapsto g_p(X_p, Y_p)$ is smooth for all smooth vector fields $X$ and $Y$. The pair $(M, g)$ is a **Riemannian manifold**.

For $u, v \in T_pM$, write $\langle u, v\rangle_p = g_p(u, v)$. The norm and angle are determined by

$$\|v\|_p = \sqrt{\langle v, v\rangle_p} \qquad \text{and} \qquad \cos\vartheta = \frac{\langle u, v\rangle_p}{\|u\|_p\|v\|_p}.$$

The vectors are orthogonal when their inner product vanishes. The length of a smooth curve $\gamma\colon [a, b] \to M$ is

$$L_g(\gamma) = \int_a^b \|\dot\gamma(t)\|_{\gamma(t)}\,\mathrm{d}t. \tag{4.5}$$

This length is unchanged by any regular monotone reparameterization. A length minimizing curve between its endpoints is a Riemannian geodesic.

**Definition 4.5** For $f \in \mathcal{F}(M)$, the **Riemannian gradient** $\operatorname{grad} f(p) \in T_pM$ is the unique tangent vector satisfying

$$vf = \langle \operatorname{grad} f(p), v\rangle_p \quad \text{for every } v \in T_pM.$$

Existence and uniqueness follow from the finite dimensional Riesz representation theorem. In coordinates, define

$$g_{ij}(p) := g_p\left(\left.\frac{\partial}{\partial x^i}\right|_p, \left.\frac{\partial}{\partial x^j}\right|_p\right).$$

If $u = u^i\partial_i$ and $v = v^j\partial_j$, then

$$g_p(u, v) = \sum_{i,j=1}^d g_{ij}(p)u^iv^j. \tag{4.6}$$

Writing $\mathbf{G}(p) = (g_{ij}(p))$ and $\mathbf{G}(p)^{-1} = (g^{ij}(p))$, the coordinate vector of the Riemannian gradient is

$$\operatorname{grad} f = \sum_{i,j=1}^d g^{ij}\frac{\partial f}{\partial x^j}\frac{\partial}{\partial x^i}. \tag{4.7}$$

Under a coordinate change $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{y})$, the metric matrix transforms as

$$\widetilde{\mathbf{G}}(\boldsymbol{y}) = \mathbf{J}(\boldsymbol{y})^\mathsf{T}\mathbf{G}(\boldsymbol{x}(\boldsymbol{y}))\mathbf{J}(\boldsymbol{y}) \qquad \text{and} \qquad \mathbf{J}(\boldsymbol{y}) = \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}^\mathsf{T}}. \tag{4.8}$$

This is the coordinate law that will make the Fisher information matrix geometric.

**Example 4.6** Let $\Omega_0 \subseteq \mathbb{R}^d$ be open and convex, and suppose $\phi\colon \Omega_0 \to \mathbb{R}$ is smooth with positive definite Hessian. Then

$$g_{\boldsymbol{x}}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \nabla^2 \phi(\boldsymbol{x}) \boldsymbol{v}$$

defines a Riemannian metric. Such a metric is called a **Hessian metric**.

If $N \subseteq M$ is a smooth submanifold, its tangent space is naturally a subspace of $T_pM$. The metric induced on $N$ is simply the restriction

$$g_p^N = g_p\big|_{T_pN \times T_pN}.$$

This pullback principle will appear repeatedly when a parameterized model is embedded into a larger space of distributions.

## Metrics Induced by Divergences

The simplest local model is the squared Mahalanobis divergence

$$D(\boldsymbol{x}\|\boldsymbol{x}') = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}')^\top \mathbf{A}(\boldsymbol{x} - \boldsymbol{x}'),$$

where $\mathbf{A}$ is positive definite. Along the curve $\boldsymbol{x} + t\boldsymbol{v}$,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} D(\boldsymbol{x} + t\boldsymbol{v}\|\boldsymbol{x})\bigg|_{t=0} = \boldsymbol{v}^\top \mathbf{A}\boldsymbol{v}.$$

Thus the quadratic term of the divergence is a squared tangent norm. The same idea works intrinsically.

For a smooth function $D(p\|q)$ on $M \times M$ and a vector field $X$, write $X_pD$ when $X$ differentiates the first argument and $X_qD$ when it differentiates the second.

**Definition 4.7** A smooth divergence $D\colon M \times M \to [0, \infty)$ is an **information geometric divergence** when it is finite on a neighbourhood of the diagonal and

$$-X_pX_qD(p\|q)\big|_{q=p} > 0$$

for every vector field $X$ with $X_p \neq 0$.

**Theorem 4.8** Let $D$ be an information geometric divergence. Then

$$g_p(X_p, Y_p) := -X_p Y_q D(p\|q)\big|_{q=p} \tag{4.9}$$

defines a Riemannian metric. Moreover,

$$\begin{aligned} -X_p Y_q D(p\|q)\big|_{q=p} &= -X_q Y_p D(p\|q)\big|_{q=p} \\ &= X_p Y_p D(p\|q)\big|_{q=p} = X_q Y_q D(p\|q)\big|_{q=p}. \end{aligned} \tag{4.10}$$

**Proof.** Choose coordinates $\boldsymbol{x}$ for $p$ and an independent copy $\boldsymbol{x}'$ for $q$. If $X = a^i \partial_i$ and $Y = b^j \partial_j$, then

$$-X_p Y_q D(p\|q)\big|_{q=p} = -a^i(p) b^j(p) \left.\frac{\partial^2 D}{\partial x^i \partial x'^j}\right|_{\boldsymbol{x}'=\boldsymbol{x}},$$

so the value depends only on $X_p$ and $Y_p$ and is bilinear. Since $D(p\|p) = 0$ and the diagonal is a minimum,

$$\left.\frac{\partial D}{\partial x^i}\right|_{\boldsymbol{x}'=\boldsymbol{x}} = 0 \qquad \text{and} \qquad \left.\frac{\partial D}{\partial x'^j}\right|_{\boldsymbol{x}'=\boldsymbol{x}} = 0.$$

Differentiate the second identity along the diagonal. The chain rule gives

$$0 = \left.\frac{\partial^2 D}{\partial x^i \partial x'^j}\right|_{\boldsymbol{x}'=\boldsymbol{x}} + \left.\frac{\partial^2 D}{\partial x'^i \partial x'^j}\right|_{\boldsymbol{x}'=\boldsymbol{x}}.$$

Repeating the argument with the first identity gives (4.10); equality of mixed partial derivatives gives symmetry. Positive definiteness is precisely the nondegeneracy condition in Definition 4.7, and smoothness follows from smoothness of $D$. $\square$

In coordinates, Theorem 4.8 reads

$$g_{ij}(\boldsymbol{x}) = -\left.\frac{\partial^2 D(\boldsymbol{x}\|\boldsymbol{x}')}{\partial x^i \partial x'^j}\right|_{\boldsymbol{x}'=\boldsymbol{x}}. \tag{4.11}$$

The metric is unchanged when $D$ is replaced by $D^*$; asymmetry first appears at third order.

For the Bregman divergence in (3.4), differentiation gives

$$g_{ij}(\boldsymbol{x}) = \frac{\partial^2 \phi}{\partial x^i \partial x^j}(\boldsymbol{x}),$$

so its induced metric is the Hessian metric from Example 4.6.

On the simplex, tangent vectors satisfy

$$T_{\boldsymbol{p}}\Delta^d = \left\{ \boldsymbol{v} \in \mathbb{R}^{d+1} : \sum_{i=0}^{d} v_i = 0 \right\}.$$

Applying the construction based on second derivatives to relative entropy along $\boldsymbol{p} + t\boldsymbol{v}$ yields

$$g_{\boldsymbol{p}}(\boldsymbol{u}, \boldsymbol{v}) = \sum_{i=0}^{d} \frac{u_i v_i}{p_i}. \tag{4.12}$$

The square-root map

$$\boldsymbol{p} \longmapsto 2\sqrt{\boldsymbol{p}} := 2(\sqrt{p_0}, \ldots, \sqrt{p_d})$$

sends the simplex into the positive orthant of the sphere of radius two and turns (4.12) into the Euclidean metric restricted to that sphere. Consequently, Fisher–Rao geodesics on the simplex become arcs of great circles, as shown in Figure 5.

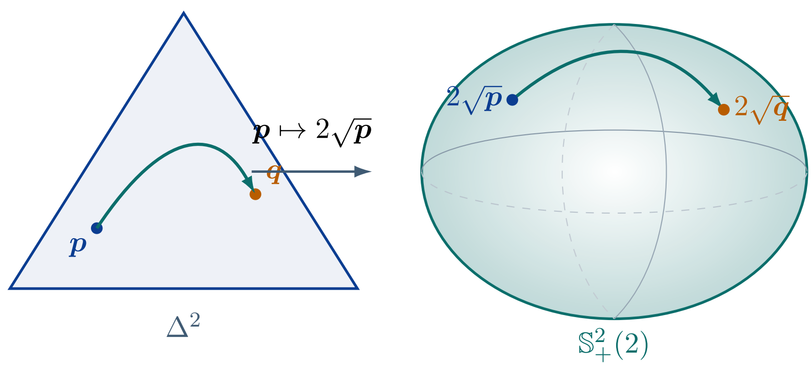


FIGURE 5

---

### Lecture 5 — Monday, September 26, 2022

## Fisher–Rao Geometry

Return to a dominated model $\mathcal{M} = \{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. Assume that the support is independent of $\boldsymbol{\theta}$, that $p_{\boldsymbol{\theta}}$ is smooth in the parameter, and that differentiation may be interchanged with integration. Differentiating the normalization identity gives

$$0 = \frac{\partial}{\partial \theta^i} \int_\Omega p_{\boldsymbol{\theta}}(x)\, \mathrm{d}\nu(x) = \int_\Omega \frac{\partial}{\partial \theta^i} \log p_{\boldsymbol{\theta}}(x) p_{\boldsymbol{\theta}}(x)\, \mathrm{d}\nu(x). \tag{5.1}$$

Thus the score

$$\boldsymbol{s}_{\boldsymbol{\theta}}(x) \coloneqq \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(x)$$

has mean zero under $\mathbb{P}_{\boldsymbol{\theta}}$.

**Definition 5.1** The **Fisher information matrix** is the covariance matrix of the score:

$$I_{ij}(\boldsymbol{\theta}) \coloneqq \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial \theta^i} \log p_{\boldsymbol{\theta}}(X) \frac{\partial}{\partial \theta^j} \log p_{\boldsymbol{\theta}}(X)\right]. \tag{5.2}$$

Under the same regularity conditions, a second differentiation of the normalization identity gives the information identity

$$I_{ij}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial \theta^i \partial \theta^j} \log p_{\boldsymbol{\theta}}(X)\right]. \tag{5.3}$$

**Theorem 5.2** If $\mathbf{I}(\boldsymbol{\theta})$ is positive definite, its entries transform according to (4.8). Hence

$$g_{ij}(\boldsymbol{\theta}) = I_{ij}(\boldsymbol{\theta})$$

defines a Riemannian metric independent of the coordinates, called the **Fisher–Rao metric**.

**Proof.** Under a reparameterization $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\eta})$, the chain rule gives

$$\nabla_{\boldsymbol{\eta}} \log p = \mathbf{J}(\boldsymbol{\eta})^{\mathsf{T}} \nabla_{\boldsymbol{\theta}} \log p \qquad \text{and} \qquad \mathbf{J}(\boldsymbol{\eta}) = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}^{\mathsf{T}}}.$$

Taking the covariance of both sides yields precisely the metric transformation law (4.8). □

For the family $\mathcal{N}(\mu, \sigma^2)$, direct calculation gives

$$\mathbf{I}(\mu, \sigma) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}. \tag{5.4}$$

Set $u = \mu/\sqrt{2}$ and $v = \sigma$. Then

$$\mathrm{d}s^2 = 2\, \frac{\mathrm{d}u^2 + \mathrm{d}v^2}{v^2}. \tag{5.5}$$

Thus the normal family is a rescaled Poincaré upper half-plane with constant curvature $-1/2$. Its geodesics are vertical lines and semicircles orthogonal to the boundary $v = 0$, as shown in Figure 6. In the original $(\mu, \sigma)$-coordinates, the latter become half ellipses.

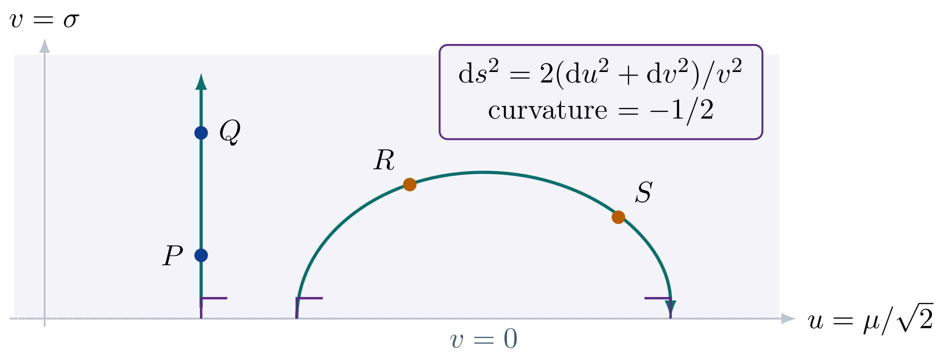


FIGURE 6

For the simplex in exponential coordinates (1.2), the score at state $k$ is

$$\frac{\partial}{\partial \theta^i} \log p_k(\boldsymbol{\theta}) = \delta_{ik} - p_i(\boldsymbol{\theta}),$$

and therefore

$$g_{ij}(\boldsymbol{\theta}) = \delta_{ij} p_i(\boldsymbol{\theta}) - p_i(\boldsymbol{\theta}) p_j(\boldsymbol{\theta}). \tag{5.6}$$

Transforming this metric back to the probability coordinates gives (4.12).

**Proposition 5.3** Subject to the regularity assumptions above, the metric induced by $\mathrm{KL}(p_{\boldsymbol{\theta}} \| p_{\boldsymbol{\theta}'})$ is the Fisher–Rao metric.

**Proof.** Using (4.11) and differentiating under the integral sign,

$$\begin{aligned} g_{ij}(\boldsymbol{\theta}) &= -\frac{\partial^2}{\partial \theta^i \partial \theta'^j} \int_\Omega p_{\boldsymbol{\theta}}(x) \log\left(\frac{p_{\boldsymbol{\theta}}(x)}{p_{\boldsymbol{\theta}'}(x)}\right) \mathrm{d}\nu(x) \bigg|_{\boldsymbol{\theta}'=\boldsymbol{\theta}} = \int_\Omega \frac{1}{p_{\boldsymbol{\theta}}(x)} \frac{\partial p_{\boldsymbol{\theta}}}{\partial \theta^i}(x) \frac{\partial p_{\boldsymbol{\theta}}}{\partial \theta^j}(x)\, \mathrm{d}\nu(x) \\ &= I_{ij}(\boldsymbol{\theta}). \end{aligned}$$

□

In particular, nearby distributions satisfy the local expansion

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}} \| \mathbb{P}_{\boldsymbol{\theta}}) = \frac{1}{2} \Delta\boldsymbol{\theta}^\top \mathbf{I}(\boldsymbol{\theta}) \Delta\boldsymbol{\theta} + O\big((\|\Delta\boldsymbol{\theta}\|^3\big).$$

The Fisher metric therefore measures local statistical distinguishability.

## Sufficiency and the Characterization Theorem

**Definition 5.4** For a model $\{\mathbb{P}_{\boldsymbol{\theta}}\}$ and an observation $X$, a statistic $T(X)$ is **sufficient** when the conditional distribution of $X$ given $T(X)$ does not depend on $\boldsymbol{\theta}$.

On a finite state space $\mathcal{X}$, the statistic determines a partition

$$\mathcal{X} = \bigcup_{y \in \mathcal{Y}} \mathcal{X}_y \qquad \text{and} \qquad \mathcal{X}_y = \{x : T(x) = y\}.$$

If $p_{\boldsymbol{\theta}}$ is the distribution of $X$, the induced distribution of $Y = T(X)$ is

$$q_{\boldsymbol{\theta}}(y) = \sum_{x \in \mathcal{X}_y} p_{\boldsymbol{\theta}}(x).$$

Write the conditional law as

$$r_{\boldsymbol{\theta}}(x \mid y) = \frac{p_{\boldsymbol{\theta}}(x)}{q_{\boldsymbol{\theta}}(y)}, \quad x \in \mathcal{X}_y.$$

Sufficiency means that $r_{\boldsymbol{\theta}}(x \mid y)$ is constant in $\boldsymbol{\theta}$. In that case, aggregation and reconstruction form the two maps in Figure 7.

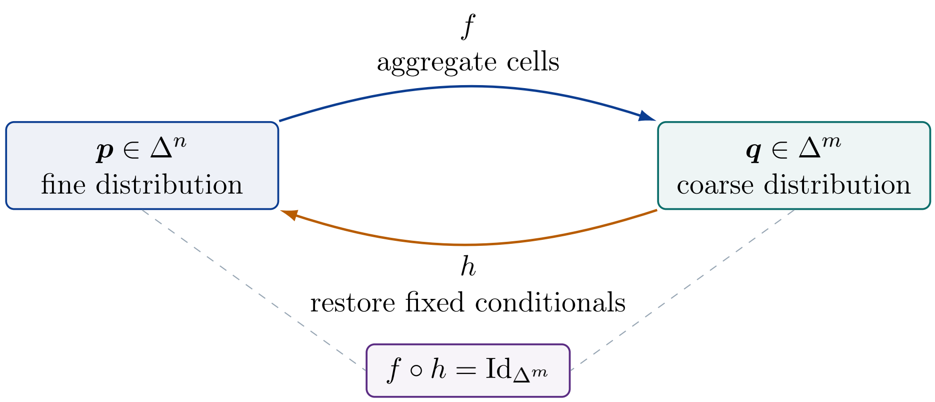


FIGURE 7

**Theorem 5.5** Let $g^{\mathcal{X}}$ and $g^{\mathcal{Y}}$ be the Fisher–Rao metrics before and after applying a statistic $T$. Then

$$g^{\mathcal{Y}} \preceq g^{\mathcal{X}}.$$

Equality holds when $T$ is sufficient.

**Proof.** The factorization

$$p_{\boldsymbol{\theta}}(x) = q_{\boldsymbol{\theta}}(y) r_{\boldsymbol{\theta}}(x \mid y), \quad x \in \mathcal{X}_y,$$

gives the score decomposition

$$\partial_i \log p_{\boldsymbol{\theta}}(x) = \partial_i \log q_{\boldsymbol{\theta}}(y) + \partial_i \log r_{\boldsymbol{\theta}}(x \mid y).$$

The conditional score has conditional mean zero, so the cross terms vanish after taking expectations. Consequently,

$$g_{ij}^{\mathcal{X}}(\boldsymbol{\theta}) = g_{ij}^{\mathcal{Y}}(\boldsymbol{\theta}) + \sum_{y \in \mathcal{Y}} q_{\boldsymbol{\theta}}(y) \sum_{x \in \mathcal{X}_y} r_{\boldsymbol{\theta}}(x \mid y) \partial_i \log r_{\boldsymbol{\theta}}(x \mid y) \partial_j \log r_{\boldsymbol{\theta}}(x \mid y).$$

The second matrix is positive semidefinite, which proves the inequality. If $T$ is sufficient, all of its entries vanish because the conditional law is independent of $\boldsymbol{\theta}$. □

The converse invariance principle uniquely identifies the Fisher metric.

**Theorem 5.6 (Chentsov's theorem, finite case)** Suppose every open finite simplex $\Delta^n$ is equipped with a Riemannian metric $g^{(n)}$. Assume every sufficient statistic embedding $h\colon \Delta^m \to \Delta^n$ of the form shown in Figure 7 is an isometry onto its image. Then there is a constant $c > 0$, independent of $n$, such that

$$g^{(n)} = c\, g_F^{(n)},$$

where

$$g_{F,\boldsymbol{p}}^{(n)}(\boldsymbol{u}, \boldsymbol{v}) = \sum_{i=0}^{n} \frac{u_i v_i}{p_i}.$$

**Proof.** At the uniform distribution

$$\overline{\boldsymbol{p}} = \left( \frac{1}{n+1}, \ldots, \frac{1}{n+1} \right),$$

every coordinate permutation is an isometry. Permutation invariance forces the ambient coefficient matrix to have the form

$$g_{ij}^{(n)}(\overline{\boldsymbol{p}}) = C_n \delta_{ij} + D_n.$$

Tangent vectors have coordinate sum zero, so the term involving $D_n$ contributes nothing and

$$g^{(n)}_{\overline{\boldsymbol{p}}}(\boldsymbol{u},\boldsymbol{v}) = C_n \sum_{i=0}^{n} u_i v_i.$$

Fix a rational point $\boldsymbol{q} \in \Delta^m$ with

$$q_j = \frac{k_j}{n+1}, \quad k_j \in \mathbb{N} \qquad \text{and} \qquad \sum_{j=0}^{m} k_j = n+1.$$

Partition the fine state space into cells of sizes $k_j$ and reconstruct with uniform conditionals within each cell. Then $h(\boldsymbol{q}) = \overline{\boldsymbol{p}}$, and

$$(\mathrm{d}h_{\boldsymbol{q}}\boldsymbol{u})_i = \frac{u_j}{k_j} \quad \text{when } i \text{ belongs to cell } j.$$

The isometry assumption yields

$$g^{(m)}_{\boldsymbol{q}}(\boldsymbol{u},\boldsymbol{v}) = C_n \sum_{j=0}^{m} k_j \frac{u_j}{k_j}\frac{v_j}{k_j} = \frac{C_n}{n+1} \sum_{j=0}^{m} \frac{u_j v_j}{q_j}.$$

The left side does not depend on the chosen common denominator, so $C_n/(n+1) = c$ for a universal constant $c > 0$. Rational points are dense, and continuity extends the formula to every point of every simplex. $\square$

Two standard constructions follow. The Fisher volume element gives the Jeffreys prior, which is invariant under reparameterization:

$$\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \propto \sqrt{\det \mathbf{I}(\boldsymbol{\theta})}\,\mathrm{d}\boldsymbol{\theta},$$

while replacing the Euclidean gradient by (4.7) produces the natural gradient

$$\operatorname{grad} f(\boldsymbol{\theta}) = \mathbf{I}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}).$$

### Lecture 6 — Wednesday, September 28, 2022

# 3. Bregman Divergence and Dually Flat Geometry

## Legendre Duality

Let $\phi\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper function. Its convex conjugate is

$$\psi(\boldsymbol{y}) = \phi^*(\boldsymbol{y}) \coloneqq \sup_{\boldsymbol{x}\in\mathbb{R}^d} \{\boldsymbol{x}^\mathsf{T}\boldsymbol{y} - \phi(\boldsymbol{x})\}. \tag{6.1}$$

When $\phi$ is lower semicontinuous and convex, $\phi^{**} = \phi$. The definition immediately gives the Fenchel–Young inequality

$$\phi(\boldsymbol{x}) + \psi(\boldsymbol{y}) - \boldsymbol{x}^\mathsf{T}\boldsymbol{y} \geq 0. \tag{6.2}$$

**Definition 6.1** For a convex function $\phi$, the **subdifferential** at $\boldsymbol{x}$ is

$$\partial\phi(\boldsymbol{x}) \coloneqq \left\{\boldsymbol{y} \in \mathbb{R}^d : \phi(\boldsymbol{x}') \geq \phi(\boldsymbol{x}) + \boldsymbol{y}^\mathsf{T}(\boldsymbol{x}' - \boldsymbol{x}) \text{ for every } \boldsymbol{x}'\right\}.$$

Equality holds in (6.2) precisely when $\boldsymbol{y} \in \partial\phi(\boldsymbol{x})$. At a differentiability point this means $\boldsymbol{y} = \nabla\phi(\boldsymbol{x})$.

**Lemma 6.2** Let $\Omega_0 \subseteq \mathbb{R}^d$ be open and convex, and suppose $\phi$ is smooth on $\Omega_0$.

1. If $\phi$ is strictly convex, then $\nabla\phi$ is injective.
2. If $\nabla^2\phi(\boldsymbol{x})$ is positive definite for every $\boldsymbol{x} \in \Omega_0$, then $\nabla\phi$ is a diffeomorphism from $\Omega_0$ onto the open set
$$\Omega_0^* \coloneqq \nabla\phi(\Omega_0).$$
On this range, $\psi$ is smooth and $\nabla\psi = (\nabla\phi)^{-1}$.

**Proof.** For (1), suppose $\nabla\phi(\boldsymbol{x}) = \nabla\phi(\boldsymbol{x}')$. Adding the two Bregman divergences gives

$$B_\phi[\boldsymbol{x} : \boldsymbol{x}'] + B_\phi[\boldsymbol{x}' : \boldsymbol{x}] = \big(\nabla\phi(\boldsymbol{x}) - \nabla\phi(\boldsymbol{x}')\big)^\mathsf{T}(\boldsymbol{x} - \boldsymbol{x}') = 0.$$

Both terms are nonnegative, and strict convexity forces $\boldsymbol{x} = \boldsymbol{x}'$.

For (2), the inverse function theorem applies because the Jacobian $\nabla^2\phi$ is invertible everywhere. Injectivity makes the local inverses agree globally, and the image is open. Equality in (6.2) gives

$$\phi(\boldsymbol{x}) + \psi(\nabla\phi(\boldsymbol{x})) - \boldsymbol{x}^\top \nabla\phi(\boldsymbol{x}) = 0.$$

Differentiation shows that $\nabla\psi(\boldsymbol{y}) = (\nabla\phi)^{-1}(\boldsymbol{y})$. □

The image $\Omega_0^*$ need not be convex without an additional boundary condition.

**Definition 6.3** A differentiable strictly convex function $\phi\colon \Omega_0 \to \mathbb{R}$ is of **Legendre type** when

$$\|\nabla\phi(\boldsymbol{x}_n)\| \longrightarrow \infty$$

whenever a sequence $(\boldsymbol{x}_n)$ in $\Omega_0$ converges to a boundary point of $\Omega_0$.

Under the standard additional hypotheses that the proper closed convex extension is essentially smooth and essentially strictly convex, Legendre type is preserved by conjugacy. In that case the interiors of the effective domains of $\phi$ and $\phi^*$ are exchanged by $\nabla\phi$ and $\nabla\phi^*$.

## Primal and Dual Geometry

Assume from now on that $\nabla^2\phi$ is positive definite on $\Omega_0$. Regard $M = \Omega_0$ as a manifold with **primal coordinate** $\boldsymbol{x}$ and **dual coordinate**

$$\boldsymbol{y} = \nabla\phi(\boldsymbol{x}).$$

**Theorem 6.4** Let $\psi = \phi^*$. If

$$\boldsymbol{y} = \nabla\phi(\boldsymbol{x}) \qquad \text{and} \qquad \boldsymbol{y}' = \nabla\phi(\boldsymbol{x}'),$$

then

$$B_\phi[\boldsymbol{x} : \boldsymbol{x}'] = B_\psi[\boldsymbol{y}' : \boldsymbol{y}] = \phi(\boldsymbol{x}) + \psi(\boldsymbol{y}') - \boldsymbol{x}^\top \boldsymbol{y}'. \tag{6.3}$$

**Proof.** The equality condition in (6.2) gives

$$\phi(\boldsymbol{x}') + \psi(\boldsymbol{y}') = \boldsymbol{x}'^\top \boldsymbol{y}'.$$

Substituting this identity into (3.4) gives the second line of (6.3). Interchanging $\phi$ with $\psi$ and using $\nabla\psi = (\nabla\phi)^{-1}$ gives the first line. □

For $p, q \in M$, write $\boldsymbol{x}_p, \boldsymbol{y}_p$ for their two coordinate representations and define

$$D(p\|q) = \phi(\boldsymbol{x}_p) + \psi(\boldsymbol{y}_q) - \boldsymbol{x}_p^\top \boldsymbol{y}_q. \tag{6.4}$$

The induced metric is Hessian in both coordinate systems:

$$\mathbf{G}_{\boldsymbol{x}}(p) = \nabla^2\phi(\boldsymbol{x}_p) \qquad \text{and} \qquad \mathbf{G}_{\boldsymbol{y}}(p) = \nabla^2\psi(\boldsymbol{y}_p). \tag{6.5}$$

**Lemma 6.5** The two Hessians satisfy

$$\nabla^2\psi(\boldsymbol{y}_p) = \nabla^2\phi(\boldsymbol{x}_p)^{-1}.$$

**Proof.** Differentiate the identity $\nabla\psi \circ \nabla\phi = \mathrm{Id}$. □

Let

$$\partial_i := \frac{\partial}{\partial x^i} \qquad \text{and} \qquad \partial^j := \frac{\partial}{\partial y_j}.$$

The relation between the inverse Hessians gives the biorthogonality identity

$$\langle \partial_i, \partial^j \rangle_p = \delta_i^j. \tag{6.6}$$

Consequently, if $u = a^i\partial_i$ and $v = b_j\partial^j$, then

$$\langle u, v \rangle_p = \sum_{i=1}^d a^i b_i.$$

The two coordinate systems therefore supply mutually dual frames.

Each coordinate system also supplies a flat notion of parallel transport. The **primal transport** keeps the coefficients in the $\partial_i$ frame constant:

$$\Pi_{pq}\left(\sum_{i=1}^d a^i \partial_i\big|_p\right) = \sum_{i=1}^d a^i \partial_i\big|_q.$$

The **dual transport** keeps the coefficients in the $\partial^j$ frame constant:

$$\Pi^*_{pq}\left(\sum_{j=1}^d b_j \partial^j\big|_p\right) = \sum_{j=1}^d b_j \partial^j\big|_q.$$

Both transports are path independent.

**Definition 6.6** A curve is a **primal geodesic** when its velocity is constant under primal transport, and it is a **dual geodesic** when its velocity is constant under dual transport.

Equivalently, primal geodesics are straight lines in $\boldsymbol{x}$ and dual geodesics are straight lines in $\boldsymbol{y}$. Unless the divergence is symmetric, neither family generally agrees with the Riemannian geodesics.

**Proposition 6.7** Let $U$ and $V$ be vector fields along a curve. If $U$ is primal parallel and $V$ is dual parallel, then $\langle U, V\rangle$ is constant along the curve.

**Proof.** Write $U = a^i\partial_i$ and $V = b_j\partial^j$. Parallelism makes the coefficients $a^i$ and $b_j$ constant, while (6.6) gives

$$\langle U, V\rangle = \sum_{i=1}^d a^i b_i,$$

which is independent of the curve parameter. □

The metric together with these two flat transports is the **dually flat geometry** induced by Bregman divergence.

## Pythagorean and Projection Theorems

The dual coordinate frames turn the Euclidean orthogonality calculation into a divergence identity. The three simultaneous views are shown in Figure 8.

**Theorem 6.8 (Generalized Pythagorean theorem)** Let $p, q, r \in M$. Suppose the relevant primal and dual line segments remain in their coordinate domains. Then

$$D(q\|p) + D(r\|q) = D(r\|p) \tag{6.7}$$

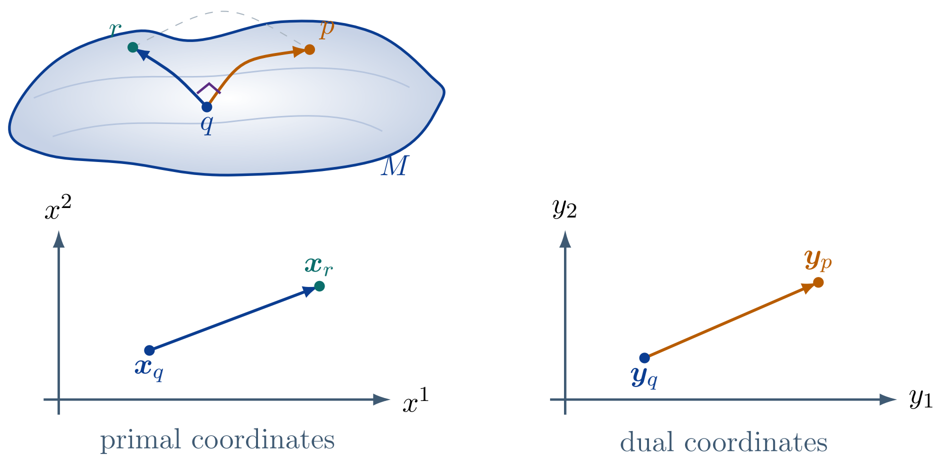


FIGURE 8

if and only if the primal geodesic from $q$ to $r$ is orthogonal at $q$ to the dual geodesic from $q$ to $p$.

**Proof.** Using the self-dual representation (6.4) and the identity

$$\phi(\boldsymbol{x}_q) + \psi(\boldsymbol{y}_q) = \boldsymbol{x}_q^{\mathsf{T}} \boldsymbol{y}_q,$$

one obtains

$$D(q\|p) + D(r\|q) - D(r\|p) = \big(\boldsymbol{x}_r - \boldsymbol{x}_q\big)^{\mathsf{T}} \big(\boldsymbol{y}_p - \boldsymbol{y}_q\big).$$

The first vector is the primal coordinate velocity from $q$ to $r$, and the second is the dual coordinate velocity from $q$ to $p$. By biorthogonality (6.6), their Euclidean pairing is exactly their Riemannian inner product at $q$. It vanishes precisely when the two geodesics are orthogonal.

*Forward implication.* If (6.7) holds, the left-hand side of the displayed difference is zero, so the geodesic velocities are orthogonal.

*Reverse implication.* If the geodesic velocities are orthogonal, their pairing is zero, and the displayed difference gives (6.7). □

**Definition 6.9** A submanifold $S \subseteq M$ is **primal autoparallel** when its image in primal coordinates is relatively open and convex. It is **dual autoparallel** when the analogous condition holds in dual coordinates.

**Theorem 6.10** Let $S$ be primal autoparallel and fix $p \in M$. If

$$\min_{q \in S} D(q \| p)$$

has a solution $q^*$, then the solution is unique. Moreover, the dual geodesic from $p$ to $q^*$ is orthogonal to $S$ at $q^*$. Conversely, this orthogonality characterizes $q^*$.

**Proof.** In primal coordinates the objective is

$$\boldsymbol{x} \longmapsto \phi(\boldsymbol{x}) + \psi(\boldsymbol{y}_p) - \boldsymbol{x}^\top \boldsymbol{y}_p.$$

It is strictly convex, so a minimizer on the convex set $\boldsymbol{x}(S)$ is unique. At an interior minimizer $\boldsymbol{x}_{q^*}$, every tangent direction $\boldsymbol{a}$ to $S$ satisfies

$$0 = \boldsymbol{a}^\top \big(\nabla\phi(\boldsymbol{x}_{q^*}) - \boldsymbol{y}_p\big) = \boldsymbol{a}^\top (\boldsymbol{y}_{q^*} - \boldsymbol{y}_p).$$

*Forward implication.* The vector $\boldsymbol{a}$ is a primal tangent direction to $S$, while $\boldsymbol{y}_{q^*} - \boldsymbol{y}_p$ is the dual velocity from $p$ to $q^*$. Biorthogonality proves the stated orthogonality condition.

*Reverse implication.* If the dual velocity is orthogonal to $S$, the directional derivative of the objective vanishes along every tangent direction at $q^*$. Convexity of the objective and of $\boldsymbol{x}(S)$ then makes $q^*$ the global minimizer. □

An analogous statement exchanges primal and dual roles for minimizing $D(p\|q)$ over a dual autoparallel submanifold.

---

### Lecture 7 — Monday, October 03, 2022

## Exponential Families

Consider the exponential family

$$p_{\boldsymbol{\theta}}(x) = \exp\big(\boldsymbol{\theta}^\top \boldsymbol{F}(x) - \phi(\boldsymbol{\theta})\big), \tag{7.1}$$

where the natural parameter domain $\Theta$ is open, $\phi$ is smooth with positive definite Hessian, and differentiation under the integral sign is valid. Take $\boldsymbol{\theta}$ as the primal coordinate and

$$\boldsymbol{\eta} = \nabla\phi(\boldsymbol{\theta})$$

as the dual coordinate.

**Theorem 7.1** The cumulant generating function and its conjugate have the following statistical interpretations.

1. The dual coordinate is the expectation parameter:

$$\boldsymbol{\eta} = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{F}(X)].$$

2. Bregman divergence is relative entropy with reversed parameter order:

$$B_\phi[\boldsymbol{\theta} : \boldsymbol{\theta}'] = \mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}'} \| \mathbb{P}_{\boldsymbol{\theta}}).$$

   In particular, the Hessian metric of $\phi$ is the Fisher–Rao metric.

3. The dual potential is negative Shannon entropy:

$$\psi(\boldsymbol{\eta}) = \int_\Omega p_{\boldsymbol{\theta}}(x)\log p_{\boldsymbol{\theta}}(x)\,\mathrm{d}\nu(x) = -H_\nu(\mathbb{P}_{\boldsymbol{\theta}}).$$

**Proof.** For (1), differentiating the log normalizing constant gives

$$\nabla\phi(\boldsymbol{\theta}) = \frac{\int_\Omega \boldsymbol{F}(x)\mathrm{e}^{\boldsymbol{\theta}^\mathsf{T}\boldsymbol{F}(x)}\,\mathrm{d}\nu(x)}{\int_\Omega \mathrm{e}^{\boldsymbol{\theta}^\mathsf{T}\boldsymbol{F}(x)}\,\mathrm{d}\nu(x)} = \int_\Omega \boldsymbol{F}(x)p_{\boldsymbol{\theta}}(x)\,\mathrm{d}\nu(x).$$

Using this identity in the Bregman formula gives

$$B_\phi[\boldsymbol{\theta} : \boldsymbol{\theta}'] = \phi(\boldsymbol{\theta}) - \phi(\boldsymbol{\theta}') - (\boldsymbol{\theta} - \boldsymbol{\theta}')^\mathsf{T}\mathbb{E}_{\boldsymbol{\theta}'}[\boldsymbol{F}(X)] = \int_\Omega p_{\boldsymbol{\theta}'}(x)\log\left(\frac{p_{\boldsymbol{\theta}'}(x)}{p_{\boldsymbol{\theta}}(x)}\right)\mathrm{d}\nu(x),$$

which proves (2). Finally, equality in Fenchel–Young duality yields

$$\begin{aligned}\psi(\boldsymbol{\eta}) &= \boldsymbol{\theta}^\mathsf{T}\boldsymbol{\eta} - \phi(\boldsymbol{\theta}) \\ &= \int_\Omega p_{\boldsymbol{\theta}}(x)\big(\boldsymbol{\theta}^\mathsf{T}\boldsymbol{F}(x) - \phi(\boldsymbol{\theta})\big)\,\mathrm{d}\nu(x)\end{aligned}$$

$$= \int_\Omega p_{\boldsymbol{\theta}}(x) \log p_{\boldsymbol{\theta}}(x) \, \mathrm{d}\nu(x),$$

proving (3). □

The expectation parameter is exactly the quantity matched by maximum likelihood.

**Theorem 7.2** Given observations $x_1, \dots, x_n$, any interior maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ satisfies

$$\widehat{\boldsymbol{\eta}} = \nabla\phi(\widehat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{F}(x_i). \tag{7.2}$$

**Proof.** The log likelihood is

$$\ell_n(\boldsymbol{\theta}) = \boldsymbol{\theta}^\mathsf{T} \sum_{i=1}^{n} \boldsymbol{F}(x_i) - n\phi(\boldsymbol{\theta}).$$

Setting its gradient to zero gives (7.2). □

Independent replication preserves the dimension of the sufficient statistic because the joint model has statistic

$$\sum_{i=1}^{n} \boldsymbol{F}(X_i)$$

and cumulant generating function $n\phi$.

The self-dual representation also rewrites the likelihood itself. If $\boldsymbol{F}(x)$ lies in the effective domain of $\psi$, then

$$p_{\boldsymbol{\theta}}(x) = \exp\bigl(-B_\psi[\boldsymbol{F}(x) : \boldsymbol{\eta}] + \psi(\boldsymbol{F}(x))\bigr). \tag{7.3}$$

For the normal location family, $\phi(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|^2/2$ and (7.3) reduces to a Gaussian kernel built from squared Euclidean distance. In general, $B_\psi$ replaces squared error.

One application is principal component analysis for exponential families. Given data $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ in the dual domain and a dimension $k$, one minimizes

$$\min_A \min_{\boldsymbol{\theta}_i \in A} \frac{1}{n} \sum_{i=1}^{n} B_\psi[\boldsymbol{y}_i : \nabla\phi(\boldsymbol{\theta}_i)], \tag{7.4}$$

where $A$ ranges over $k$-dimensional affine subsets of the primal domain. The dual image $\nabla\phi(A)$ is primal autoparallel, and each fitted point is a dual projection. This geometry is shown in Figure 9.

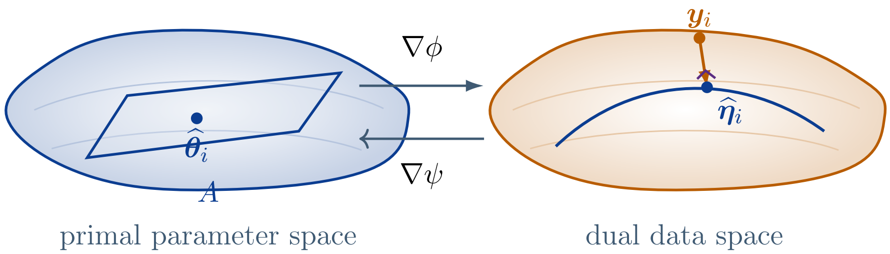


FIGURE 9

## Barycentres and Maximum Entropy

Equation (7.2) is a variational statement about Bregman centres.

**Definition 7.3** Let $\mathbb{P}$ be a probability measure on the domain of $\psi$. A **right Bregman barycentre** is a minimizer

$$\widehat{\boldsymbol{\eta}} \in \arg\min_{\boldsymbol{\eta}} \int B_\psi[\boldsymbol{y} : \boldsymbol{\eta}]\, \mathrm{d}\mathbb{P}(\boldsymbol{y}).$$

A left Bregman barycentre reverses the two arguments.

**Proposition 7.4** Suppose

$$\overline{\boldsymbol{y}} = \int \boldsymbol{y}\, \mathrm{d}\mathbb{P}(\boldsymbol{y})$$

exists and belongs to the dual domain. Then the unique right Bregman barycentre is $\overline{\boldsymbol{y}}$.

**Proof.** For any admissible $\boldsymbol{\eta}$,

$$\int (B_\psi[\boldsymbol{y} : \boldsymbol{\eta}] - B_\psi[\boldsymbol{y} : \overline{\boldsymbol{y}}])\, \mathrm{d}\mathbb{P}(\boldsymbol{y}) = \psi(\overline{\boldsymbol{y}}) - \psi(\boldsymbol{\eta}) - \nabla\psi(\boldsymbol{\eta})^\mathsf{T}(\overline{\boldsymbol{y}} - \boldsymbol{\eta}) = B_\psi[\overline{\boldsymbol{y}} : \boldsymbol{\eta}] \geq 0.$$

Strict convexity gives uniqueness. □

When $\nabla\psi(\boldsymbol{Y})$ is integrable, the left barycentre $\widehat{\boldsymbol{\eta}}_{\mathrm{L}}$ is characterized instead by

$$\nabla\psi(\widehat{\boldsymbol{\eta}}_{\mathrm{L}}) = \int \nabla\psi(\boldsymbol{y})\, \mathrm{d}\mathbb{P}(\boldsymbol{y}), \tag{7.5}$$

Indeed, differentiating the left barycentre objective with respect to its first argument gives

$$\nabla\psi(\boldsymbol{\eta}) - \int \nabla\psi(\boldsymbol{y})\,\mathrm{d}\mathbb{P}(\boldsymbol{y}),$$

and strict convexity gives uniqueness. Thus the left barycentre is an arithmetic mean in primal coordinates. By contrast, Proposition 7.4 says that the right barycentre is an arithmetic mean in dual coordinates. The two constructions are compared in Figure 10.


FIGURE 10

Applying Proposition 7.4 to the empirical distribution of $\boldsymbol{F}(x_1), \ldots, \boldsymbol{F}(x_n)$ recovers the maximum likelihood equation (7.2). It also yields a variational characterization of dual geodesics: if

$$\boldsymbol{y}_t = (1-t)\boldsymbol{y}_0 + t\boldsymbol{y}_1, \quad 0 \le t \le 1,$$

then

$$\boldsymbol{y}_t = \arg\min_{\boldsymbol{y}} \left\{(1-t)B_\psi[\boldsymbol{y}_0 : \boldsymbol{y}] + tB_\psi[\boldsymbol{y}_1 : \boldsymbol{y}]\right\}.$$

Exponential families also solve entropy maximization problems. Let $\boldsymbol{c} \in \mathbb{R}^d$ be fixed and consider all probability measures $\mathbb{P}$ equivalent to $\nu$ such that

$$\mathbb{E}_{\mathbb{P}}[\boldsymbol{F}(X)] = \boldsymbol{c}.$$

**Theorem 7.5** Suppose there is a parameter $\boldsymbol{\theta}^*$ such that

$$\nabla\phi(\boldsymbol{\theta}^*) = \boldsymbol{c}.$$

Then $\mathbb{P}_{\boldsymbol{\theta}^*}$ uniquely maximizes Shannon entropy over all feasible $\mathbb{P}$ for which the displayed quantities are well defined. More precisely,

$$H_\nu(\mathbb{P}_{\boldsymbol{\theta}^*}) - H_\nu(\mathbb{P}) = \mathrm{KL}(\mathbb{P}\|\mathbb{P}_{\boldsymbol{\theta}^*}). \tag{7.6}$$

**Proof.** Let $p = \mathrm{d}\mathbb{P}/\mathrm{d}\nu$. Feasibility and the form of the exponential family give

$$\mathrm{KL}(\mathbb{P}\|\mathbb{P}_{\boldsymbol{\theta}^*}) = \int_\Omega p(x)\log\left(\frac{p(x)}{p_{\boldsymbol{\theta}^*}(x)}\right)\mathrm{d}\nu(x) = -H_\nu(\mathbb{P}) - \boldsymbol{\theta}^{*\mathsf{T}}\boldsymbol{c} + \phi(\boldsymbol{\theta}^*).$$

By (3) and Fenchel equality,

$$-\boldsymbol{\theta}^{*\mathsf{T}}\boldsymbol{c} + \phi(\boldsymbol{\theta}^*) = H_\nu(\mathbb{P}_{\boldsymbol{\theta}^*}).$$

This proves (7.6). Nonnegativity and the equality condition in Proposition 3.3 prove uniqueness. □

## Mirror Descent

Let $\Omega_0 \subseteq \mathbb{R}^d$ be open and convex, and consider minimizing a differentiable function $f\colon \Omega_0 \to \mathbb{R}$. Euclidean gradient descent can be written as the proximal update

$$\boldsymbol{x}_{k+1} = \operatorname*{arg\,min}_{\boldsymbol{x}\in\Omega_0}\left\{f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^{\mathsf{T}}(\boldsymbol{x}-\boldsymbol{x}_k) + \frac{1}{2\delta_k}\|\boldsymbol{x}-\boldsymbol{x}_k\|^2\right\}.$$

The first-order condition gives the familiar update

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \delta_k\nabla f(\boldsymbol{x}_k).$$

Mirror descent replaces the Euclidean penalty by a Bregman divergence:

$$\boldsymbol{x}_{k+1} = \operatorname*{arg\,min}_{\boldsymbol{x}\in\Omega_0}\left\{f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^{\mathsf{T}}(\boldsymbol{x}-\boldsymbol{x}_k) + \frac{1}{\delta_k}B_\phi[\boldsymbol{x}:\boldsymbol{x}_k]\right\}. \tag{7.7}$$

The first-order condition becomes

$$\nabla\phi(\boldsymbol{x}_{k+1}) = \nabla\phi(\boldsymbol{x}_k) - \delta_k\nabla f(\boldsymbol{x}_k).$$

Thus, with $\boldsymbol{y}_k = \nabla\phi(\boldsymbol{x}_k)$,

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k - \delta_k\nabla f(\boldsymbol{x}_k) \qquad \text{and} \qquad \boldsymbol{x}_{k+1} = \nabla\psi(\boldsymbol{y}_{k+1}). \tag{7.8}$$

The coordinate flow is summarized in Figure 11.

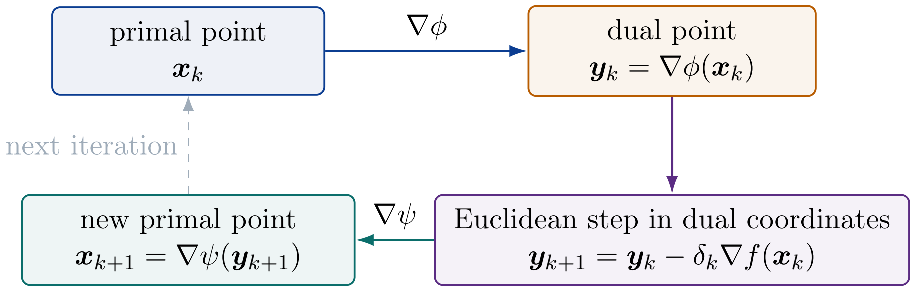


FIGURE 11

For the simplex, choose negative Shannon entropy. The Bregman penalty is relative entropy, and a Lagrange multiplier for the constraint that the components sum to one gives the entropic update

$$p^i_{k+1} = \frac{p^i_k \exp\bigl(-\delta_k \partial_i f(\boldsymbol{p}_k)\bigr)}{\sum_{j=0}^d p^j_k \exp\bigl(-\delta_k \partial_j f(\boldsymbol{p}_k)\bigr)}. \tag{7.9}$$

The positivity and normalization constraints are built into the update.

**Theorem 7.6** Assume $f$ is convex and there is $L > 0$ such that

$$B_f[\boldsymbol{x} : \boldsymbol{x}'] \le L B_\phi[\boldsymbol{x} : \boldsymbol{x}']$$

for all $\boldsymbol{x}, \boldsymbol{x}' \in \Omega_0$. Suppose the updates in (7.7) exist and remain in $\Omega_0$. With a constant step size $0 < \delta \le 1/L$, mirror descent satisfies

$$f(\boldsymbol{x}_k) \le \inf_{\boldsymbol{x} \in \Omega_0} \left\{ f(\boldsymbol{x}) + \frac{B_\phi[\boldsymbol{x} : \boldsymbol{x}_0]}{\delta k} \right\}. \tag{7.10}$$

**Proof.** Define the convex auxiliary function

$$\xi_k(\boldsymbol{x}) = f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}_k) + \frac{1}{\delta} B_\phi[\boldsymbol{x} : \boldsymbol{x}_k].$$

The update makes $\nabla \xi_k(\boldsymbol{x}_{k+1}) = 0$, so the Bregman three-point identity gives

$$\xi_k(\boldsymbol{x}_{k+1}) = \xi_k(\boldsymbol{x}) - B_{\xi_k}[\boldsymbol{x} : \boldsymbol{x}_{k+1}].$$

The relative smoothness assumption and $\delta \le 1/L$ imply

$$\xi_k(\boldsymbol{x}_{k+1}) \ge f(\boldsymbol{x}_{k+1}).$$

After expanding $B_{\xi_k}$ and using convexity of $f$, one obtains, for every $\boldsymbol{x} \in \Omega_0$,

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}) \leq \frac{1}{\delta}\left(B_\phi[\boldsymbol{x} : \boldsymbol{x}_k] - B_\phi[\boldsymbol{x} : \boldsymbol{x}_{k+1}]\right).$$

The objective values are nonincreasing. Summing from 0 to $k-1$ telescopes the Bregman terms and gives (7.10). $\square$

The limit in continuous time is the Riemannian gradient flow

$$\dot{\gamma}(t) = -\operatorname{grad} f(\gamma(t)). \tag{7.11}$$

Writing $\boldsymbol{x}(t)$ and $\boldsymbol{y}(t) = \nabla\phi(\boldsymbol{x}(t))$, equations (4.7) and (6.5) give

$$\dot{\boldsymbol{x}}(t) = -\nabla^2\phi(\boldsymbol{x}(t))^{-1}\nabla f(\boldsymbol{x}(t)) \qquad \text{and} \qquad \dot{\boldsymbol{y}}(t) = -\nabla f(\boldsymbol{x}(t)).$$

Thus (7.8) is the forward Euler discretization of the gradient flow in dual coordinates.

For fixed $q \in M$, the gradient flow of $p \mapsto D(q\|p)$ traces a reparameterized primal geodesic toward $q$, while the gradient flow of $p \mapsto D(p\|q)$ traces a reparameterized dual geodesic. For the second statement, for example,

$$\nabla_{\boldsymbol{x}} D(p\|q) = \boldsymbol{y}_p - \boldsymbol{y}_q,$$

so the dual coordinate satisfies

$$\dot{\boldsymbol{y}}(t) = -(\boldsymbol{y}(t) - \boldsymbol{y}_q),$$

which remains on the straight segment to $\boldsymbol{y}_q$.

**Lecture 8 — Wednesday, October 05, 2022**

# 4. Dualistic Geometry and Curvature

## Affine Connections and Parallel Transport

On a vector space, vectors based at different points may be compared without further choices. A manifold has no such canonical identification between its tangent spaces. An affine connection supplies the missing rule: it differentiates one vector field in the direction of another and thereby defines parallel transport.

Write $\mathfrak{X}(M)$ for the space of smooth vector fields on $M$ and $C^\infty(M)$ for the smooth real-valued functions on $M$.

**Definition 8.1** An **affine connection** on $M$ is an $\mathbb{R}$-bilinear map

$$\nabla\colon \mathfrak{X}(M) \times \mathfrak{X}(M) \longrightarrow \mathfrak{X}(M) \qquad \text{and} \qquad (X, Y) \longmapsto \nabla_X Y,$$

such that, for $f \in C^\infty(M)$,

$$\nabla_{fX} Y = f \nabla_X Y \qquad \text{and} \qquad \nabla_X (fY) = (Xf)Y + f\nabla_X Y. \tag{8.1}$$

The vector field $\nabla_X Y$ is the **covariant derivative** of $Y$ in the direction $X$.

The connection is linear over smooth functions in its direction argument, but the product rule in its second argument is essential. Consequently, a connection is not itself a tensor. Its **torsion** is the tensor

$$\mathrm{Tor}(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y]. \tag{8.2}$$

The connection is **torsion-free** when (8.2) vanishes for all vector fields $X$ and $Y$. Only torsion-free connections will be needed below.

On an open set $\Omega_0 \subseteq \mathbb{R}^d$, the Euclidean rule

$$\nabla_X Y = \sum_{k=1}^{d} X(Y^k) \frac{\partial}{\partial x^k} \qquad \text{and} \qquad Y = \sum_{k=1}^{d} Y^k \frac{\partial}{\partial x^k}, \tag{8.3}$$

is a torsion-free connection. It is the flat connection associated with the coordinates $\boldsymbol{x}$. The primal and dual transports in the Bregman construction are precisely the flat connections associated with the $\boldsymbol{x}$- and $\boldsymbol{y}$-coordinate systems, respectively.

Let $\partial_i = \partial/\partial x^i$ be the coordinate vector fields on a chart. The smooth functions $\Gamma^k_{ij}$ determined by

$$\nabla_{\partial_i}\partial_j = \sum_{k=1}^{d} \Gamma^k_{ij}\partial_k \tag{8.4}$$

are the **Christoffel symbols of the second kind**. If

$$X = \sum_{i=1}^{d} X^i \partial_i \qquad \text{and} \qquad Y = \sum_{j=1}^{d} Y^j \partial_j,$$

then the connection has the coordinate form

$$\nabla_X Y = \sum_{k=1}^{d} \left( \sum_{i=1}^{d} X^i \partial_i Y^k + \sum_{i,j=1}^{d} X^i Y^j \Gamma^k_{ij} \right) \partial_k. \tag{8.5}$$

Because coordinate vector fields commute, torsion-freeness is equivalent to

$$\Gamma^k_{ij} = \Gamma^k_{ji} \tag{8.6}$$

in every coordinate system.

If a Riemannian metric $g$ is also present, the **Christoffel symbols of the first kind** are

$$\Gamma_{ijk} = g(\nabla_{\partial_i}\partial_j, \partial_k) = \sum_{m=1}^{d} \Gamma^m_{ij} g_{mk}. \tag{8.7}$$

Thus

$$\Gamma^k_{ij} = \sum_m \Gamma_{ijm} g^{mk}.$$

Neither kind of Christoffel symbol is a tensor: the symbols may vanish in one coordinate system and fail to vanish after a nonlinear change of coordinates.

Let $\gamma \colon [a,b] \to M$ be a smooth curve and let

$$V(t) = \sum_{k=1}^{d} V^k(t)\, \partial_k|_{\gamma(t)}$$

be a vector field along $\gamma$. Extending $V$ and $\dot\gamma$ locally and then taking their covariant derivative gives a well-defined vector field along the curve. In coordinates,

$$\nabla_{\dot\gamma} V = \sum_{k=1}^{d} \left( \dot V^k + \sum_{i,j=1}^{d} \dot x^i V^j \Gamma_{ij}^k(\gamma) \right) \partial_k|_\gamma . \tag{8.8}$$

**Definition 8.2** A vector field $V$ along $\gamma$ is **$\nabla$-parallel** if $\nabla_{\dot\gamma} V = 0$. A curve $\gamma$ is a **$\nabla$-geodesic** if its velocity is parallel along itself:

$$\nabla_{\dot\gamma}\dot\gamma = 0.$$

Equation (8.8) turns parallel transport into a linear system of ordinary differential equations. Therefore, for each $v \in T_{\gamma(a)}M$, there is a unique parallel field $V$ with $V(a) = v$. Its terminal value defines a linear isomorphism

$$\Pi_\gamma^\nabla : T_{\gamma(a)}M \longrightarrow T_{\gamma(b)}M \qquad \text{and} \qquad v \longmapsto V(b). \tag{8.9}$$

The inverse is transport along the reversed curve.

Applying (8.8) to $V = \dot\gamma$ yields the geodesic equations

$$\ddot x^k(t) + \sum_{i,j=1}^{d} \Gamma_{ij}^k(\gamma(t)) \dot x^i(t) \dot x^j(t) = 0, \quad k = 1, \ldots, d. \tag{8.10}$$

For the Euclidean flat connection, the Christoffel symbols vanish and (8.10) reduces to $\ddot{\boldsymbol{x}} = \mathbf{0}$. Hence its geodesics are straight lines traced at constant speed. More generally, any coordinate system in which all Christoffel symbols vanish makes the local affine geometry look Euclidean.

## Levi–Civita and Dual Connections

A Riemannian metric selects one distinguished affine connection, but information geometry naturally uses a pair of connections around it.

**Definition 8.3** An affine connection $\nabla$ on a Riemannian manifold $(M, g)$ is **metric-compatible** if

$$Zg(X, Y) = g(\nabla_Z X, Y) + g(X, \nabla_Z Y) \tag{8.11}$$

for all vector fields $X, Y, Z$.

Metric compatibility says that parallel transport preserves inner products. Indeed, if $U$ and $V$ are parallel along $\gamma$, then (8.11) gives

$$\frac{\mathrm{d}}{\mathrm{d}t} g(U, V) = g(\nabla_{\dot\gamma} U, V) + g(U, \nabla_{\dot\gamma} V) = 0.$$

**Theorem 8.4 (Fundamental theorem of Riemannian geometry)** Every Riemannian manifold $(M, g)$ has a unique torsion-free, metric-compatible affine connection. It is called the **Levi–Civita connection** of $g$.

**Proof.** If such a connection exists, torsion-freeness and metric compatibility force the **Koszul formula**

$$\begin{aligned} 2g(\nabla_X Y, Z) &= Xg(Y, Z) + Yg(Z, X) - Zg(X, Y) \\ &\quad - g(X, [Y, Z]) + g(Y, [Z, X]) + g(Z, [X, Y]). \end{aligned} \tag{8.12}$$

Since $g$ is nondegenerate, (8.12) uniquely determines $\nabla_X Y$. Conversely, define $\nabla_X Y$ by requiring (8.12) for every $Z$. Direct substitution shows that the resulting operator satisfies (8.1), is torsion-free, and obeys (8.11). □

Taking $X = \partial_i$, $Y = \partial_j$, and $Z = \partial_k$ in (8.12) gives

$$\Gamma^{\mathrm{LC}}_{ijk} = \frac{1}{2}\left(\partial_i g_{jk} + \partial_j g_{ki} - \partial_k g_{ij}\right). \tag{8.13}$$

Raising the last index with the inverse metric gives the Christoffel symbols of the second kind.

The descriptions based on minimizing length and having zero acceleration now meet. A sufficiently short Levi–Civita geodesic minimizes Riemannian length between its endpoints. The qualification is local: a geodesic that continues far enough need not remain minimizing.

**Definition 8.5** Let $(M, g)$ be Riemannian, and let $\nabla$ and $\nabla^*$ be torsion-free affine connections. They are **dual** with respect to $g$ if

$$Zg(X, Y) = g(\nabla_Z X, Y) + g(X, \nabla^*_Z Y) \tag{8.14}$$

for all vector fields $X, Y, Z$. The quadruple $(M, g, \nabla, \nabla^*)$ is a **statistical manifold**, and $(g, \nabla, \nabla^*)$ is its **dualistic structure**.

If $\Gamma_{kij}$ and $\Gamma^*_{kij}$ denote the Christoffel symbols of the first kind with the direction index written first, then (8.14) becomes

$$\partial_k g_{ij} = \Gamma_{kij} + \Gamma^*_{kji}. \tag{8.15}$$

For fixed $g$ and $\nabla$, this relation determines $\nabla^*$ uniquely.

The midpoint of the dual pair is the Levi–Civita connection:

$$\nabla^{\mathrm{LC}} = \frac{1}{2}(\nabla + \nabla^*). \tag{8.16}$$

Indeed, averaging (8.14) with the identity obtained by interchanging $\nabla$ and $\nabla^*$ proves metric compatibility, while torsion-freeness is preserved by affine combinations.

More generally, the ***α*-connections** associated with a dual pair are

$$\nabla^{(\alpha)} = \frac{1+\alpha}{2}\nabla + \frac{1-\alpha}{2}\nabla^*, \quad \alpha \in \mathbb{R}. \tag{8.17}$$

Then $\nabla^{(0)} = \nabla^{\mathrm{LC}}$, and $\nabla^{(\alpha)}$ is dual to $\nabla^{(-\alpha)}$.

For a regular statistical model $\{p_{\boldsymbol{\theta}}\}$, let

$$\ell_{\boldsymbol{\theta}}(x) = \log p_{\boldsymbol{\theta}}(x) \qquad \text{and} \qquad \partial_i = \frac{\partial}{\partial \theta^i}.$$

Amari's $\alpha$-connection is described by

$$\Gamma^{(\alpha)}_{ijk}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\partial_i\partial_j\ell_{\boldsymbol{\theta}}\,\partial_k\ell_{\boldsymbol{\theta}} + \frac{1-\alpha}{2}\partial_i\ell_{\boldsymbol{\theta}}\partial_j\ell_{\boldsymbol{\theta}}\partial_k\ell_{\boldsymbol{\theta}}\right]. \tag{8.18}$$

The connections with parameters $\alpha$ and $-\alpha$ are dual under the Fisher–Rao metric. In particular, the 0-connection is its Levi–Civita connection.

Duality also has a direct transport interpretation. If $U$ is $\nabla$-parallel and $V$ is $\nabla^*$-parallel along the same curve, then

$$\frac{\mathrm{d}}{\mathrm{d}t} g(U(t), V(t)) = 0. \tag{8.19}$$

Thus the two transports may distort lengths separately while preserving their mutual pairing. Figure 12 depicts this complementary behaviour.

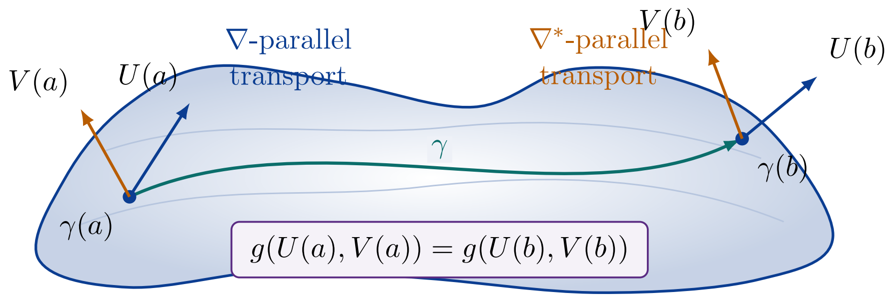


FIGURE 12

## Dualistic Structures from Divergences

The metric in Theorem 4.8 records the quadratic part of a divergence along the diagonal. Its third derivatives record the two affine connections. The resulting hierarchy is summarized in Figure 13: the second-order jet gives the local quadratic geometry, while the third-order jet separates the primal and dual transports. Curvature records how those transports vary from point to point.

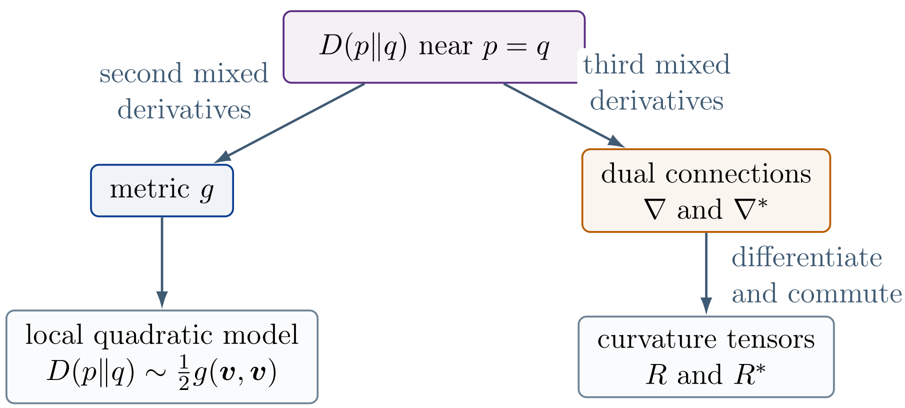


FIGURE 13

**Theorem 8.6** Let $D$ be an information geometric divergence on $M$, and let $g$ be its induced metric. Define $\nabla$ and $\nabla^*$ by

$$g(\nabla_X Y, Z) = -X_p Y_p Z_q D(p\|q)\big|_{q=p}, \tag{8.20}$$
$$g(\nabla^*_X Y, Z) = -X_q Y_q Z_p D(p\|q)\big|_{q=p}. \tag{8.21}$$

Then $\nabla$ and $\nabla^*$ are torsion-free affine connections that are dual with respect to $g$.

**Proof.** Mixed partial derivatives commute, so (8.20) and (8.21) are symmetric in $X$ and $Y$ after the correction by the Lie bracket. This gives torsion-freeness. Linearity in $X$ and the product rule in $Y$ follow by differentiating and using that the first derivatives of $D$ vanish on the diagonal.

It remains to verify duality. In coordinates $\boldsymbol{x}$ and $\boldsymbol{x}'$, define

$$\Gamma_{ijk} = -\left.\frac{\partial^3 D}{\partial x^i \partial x^j \partial x'^k}\right|_{\boldsymbol{x}'=\boldsymbol{x}}, \tag{8.22}$$

$$\Gamma^*_{ijk} = -\left.\frac{\partial^3 D}{\partial x'^i \partial x'^j \partial x^k}\right|_{\boldsymbol{x}'=\boldsymbol{x}}. \tag{8.23}$$

Differentiating (4.11) along the diagonal requires differentiating both arguments. Hence

$$\partial_k g_{ij} = \Gamma_{kij} + \Gamma^*_{kji},$$

which is exactly (8.15). □

The adjoint divergence $D^*(p\|q) = D(q\|p)$ interchanges the two connections. If $D$ is symmetric, the connections coincide; by (8.16), both must then equal the Levi–Civita connection. The asymmetry of a divergence is therefore encoded infinitesimally by the separation between $\nabla$ and $\nabla^*$.

For relative entropy on a regular statistical model, the induced pair is the pair of Amari connections with parameters 1 and $-1$. On an exponential family, one is flat in natural coordinates and the other is flat in expectation coordinates, recovering the Bregman geometry of Theorem 7.1.

The converse reconstruction is not unique. A dualistic structure fixes only the relevant second- and third-order derivatives on the diagonal. For example, if $h\colon [0,\infty) \to [0,\infty)$ is smooth near zero with

$$h(0) = 0 \qquad \text{and} \qquad h'(0) = 1,$$

then $D$ and $h \circ D$ induce the same metric and connections. Terms that vanish to fourth order along the diagonal are likewise invisible to the dualistic structure. Canonical divergences become available only when additional geometry, such as dual flatness or constant curvature, removes part of this ambiguity.

## Tensors and Curvature

Curvature measures the failure of parallel transport to be path independent. Its definition is most naturally expressed through tensors, so a short review is useful.

The **cotangent space** at $p$ is the dual vector space

$$T_p^*M \coloneqq (T_pM)^*.$$

Its elements are covectors. If $x^1, \dots, x^d$ are local coordinates, then $\mathrm{d}x^1|_p, \dots, \mathrm{d}x^d|_p$ form the basis dual to $\partial_1|_p, \dots, \partial_d|_p$:

$$\mathrm{d}x^j|_p(\partial_i|_p) = \delta_i^j. \tag{8.24}$$

**Definition 8.7** For nonnegative integers $r$ and $s$, a **tensor of type $(r, s)$** at $p$ is a multilinear map

$$A\colon \underbrace{T_p^*M \times \cdots \times T_p^*M}_{r \text{ factors}} \times \underbrace{T_pM \times \cdots \times T_pM}_{s \text{ factors}} \longrightarrow \mathbb{R}.$$

A **tensor field** is a smoothly varying choice $p \mapsto A_p$.

The tensor product combines an $(r, s)$-tensor $A$ and an $(r', s')$-tensor $B$ into an $(r+r', s+s')$-tensor $A \otimes B$ by multiplying their evaluations on the corresponding arguments. Consequently, any tensor has a coordinate expansion in products of the basis vectors and covectors. For example,

$$g = \sum_{i,j=1}^{d} g_{ij}\, \mathrm{d}x^i \otimes \mathrm{d}x^j. \tag{8.25}$$

A smooth function is a $(0, 0)$-tensor field, a differential one-form is a $(0, 1)$-tensor field under the convention that counts vector arguments second, and a Riemannian metric is a $(0, 2)$-tensor field.

The connection $\nabla$ is not tensorial because of the product rule in (8.1). In contrast, the difference of two affine connections is tensorial. On a statistical manifold, define the **skewness tensor**

$$C(X, Y, Z) \coloneqq g\big((\nabla^* - \nabla)_X Y, Z\big). \tag{8.26}$$

The torsion-free and duality identities imply that $C$ is symmetric in all three arguments. Equations (8.16) and (8.26) show that $(g, C)$ carries the same information as the dualistic structure:

$$g(\nabla_X Y, Z) = g(\nabla_X^{\mathrm{LC}} Y, Z) - \frac{1}{2} C(X, Y, Z) \qquad \text{and} \qquad g(\nabla_X^* Y, Z) = g(\nabla_X^{\mathrm{LC}} Y, Z) + \frac{1}{2} C(X, Y, Z). \tag{8.27}$$

**Definition 8.8** The **curvature tensor** of an affine connection $\nabla$ is

$$R(X, Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z. \tag{8.28}$$

The connection is **flat** if $R$ vanishes identically.

Although its definition involves derivatives of vector fields, the cancellations in (8.28) make $R$ linear over $C^\infty(M)$ in each argument. It is therefore a tensor of type $(1,3)$. Writing

$$R(\partial_i, \partial_j)\partial_k = \sum_{\ell=1}^{d} R_{ijk}{}^{\ell} \partial_\ell,$$

one obtains

$$R_{ijk}{}^{\ell} = \partial_i \Gamma^{\ell}_{jk} - \partial_j \Gamma^{\ell}_{ik} + \sum_{m=1}^{d} \left( \Gamma^{m}_{jk} \Gamma^{\ell}_{im} - \Gamma^{m}_{ik} \Gamma^{\ell}_{jm} \right). \tag{8.29}$$

The order-$\varepsilon^2$ mismatch in parallel transport around a small coordinate parallelogram is curvature. More precisely, let the sides at $p$ be $\varepsilon X$ and $\varepsilon Y$, transport $Z \in T_pM$ once around the loop, and call the result $\widetilde{Z}_\varepsilon$. With an orientation consistent with (8.28),

$$R(X,Y)Z = \lim_{\varepsilon \downarrow 0} \frac{\widetilde{Z}_\varepsilon - Z}{\varepsilon^2}. \tag{8.30}$$

This infinitesimal holonomy is shown in Figure 14.

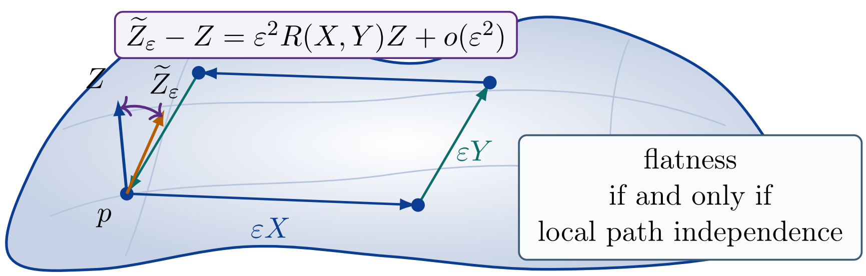


FIGURE 14

For a torsion-free connection, flatness is equivalent to the local existence of coordinates in which every Christoffel symbol vanishes. In those coordinates, parallel transport is locally independent of the chosen path, and affine geodesics are straight lines.

## Lecture 9 — Monday, October 10, 2022

## Dually Flat Manifolds

The two curvatures of a statistical manifold are coupled. Denote the curvature tensors of $\nabla$ and $\nabla^*$ by $R$ and $R^*$.

**Proposition 9.1** For all vector fields $X, Y, Z, W$,

$$g(R(X,Y)Z, W) = -g(Z, R^*(X,Y)W). \tag{9.1}$$

Consequently, $\nabla$ is flat if and only if $\nabla^*$ is flat.

**Proof.** Apply (8.14) successively to the three terms in (8.28). The second derivatives of $g(Z,W)$ cancel because

$$XY - YX = [X,Y],$$

while the remaining terms combine into

$$-g\big(Z, \nabla^*_X \nabla^*_Y W - \nabla^*_Y \nabla^*_X W - \nabla^*_{[X,Y]} W\big).$$

This proves (9.1).

*Forward implication.* If $R = 0$, then (9.1) and nondegeneracy of $g$ imply $R^* = 0$.

*Reverse implication.* If $R^* = 0$, the same identity and nondegeneracy imply $R = 0$. □

**Definition 9.2** A statistical manifold $(M, g, \nabla, \nabla^*)$ is **dually flat** when the equivalent flatness conditions in Proposition 9.1 hold.

Dual flatness is exactly the local geometry generated by Bregman divergence.

**Theorem 9.3** Let $(M, g, \nabla, \nabla^*)$ be dually flat. Near every point there are primal coordinates $\boldsymbol{x}$, dual coordinates $\boldsymbol{y}$, and strictly convex Legendre dual potentials $\phi$ and $\psi$ such that

$$\boldsymbol{y} = \nabla\phi(\boldsymbol{x}) \qquad \text{and} \qquad \boldsymbol{x} = \nabla\psi(\boldsymbol{y}). \tag{9.2}$$

Moreover,

$$g_{ij} = \frac{\partial^2 \phi}{\partial x^i \partial x^j} \qquad \text{and} \qquad g^{ij} = \frac{\partial^2 \psi}{\partial y_i \partial y_j}, \tag{9.3}$$

and the local Bregman divergence

$$D(p\|q) = \phi(\boldsymbol{x}_p) - \phi(\boldsymbol{x}_q) - \nabla\phi(\boldsymbol{x}_q)^{\mathsf{T}}(\boldsymbol{x}_p - \boldsymbol{x}_q) \tag{9.4}$$

induces the given dualistic structure.

**Proof.** Flatness and torsion-freeness of $\nabla$ provide local affine coordinates $\boldsymbol{x}$ in which $\Gamma^k_{ij} = 0$. The coordinate fields $\partial_i$ are therefore $\nabla$-parallel. Define vector fields $\partial^j$ by

$$g(\partial_i, \partial^j) = \delta_i^j. \tag{9.5}$$

For every vector field $Z$, duality gives

$$0 = Zg(\partial_i, \partial^j) = g(\nabla_Z \partial_i, \partial^j) + g(\partial_i, \nabla^*_Z \partial^j) = g(\partial_i, \nabla^*_Z \partial^j).$$

Hence each $\partial^j$ is $\nabla^*$-parallel. Torsion-freeness then gives $[\partial^i, \partial^j] = 0$, so these fields are coordinate fields for local dual coordinates $\boldsymbol{y}$.

Biorthogonality implies

$$g_{ij} = \frac{\partial y_i}{\partial x^j} \qquad \text{and} \qquad g^{ij} = \frac{\partial x^i}{\partial y_j}. \tag{9.6}$$

The symmetry of $g_{ij}$ is the integrability condition for a local potential $\phi$ with $\boldsymbol{y} = \nabla\phi(\boldsymbol{x})$. Positive definiteness of $g$ makes $\phi$ strictly convex. Its Legendre transform $\psi$ satisfies $\boldsymbol{x} = \nabla\psi(\boldsymbol{y})$, and (9.3) follows. The calculations in Theorem 8.6 applied to (9.4) recover $g$, $\nabla$, and $\nabla^*$. $\square$

The word "canonical" is local here. Different affine charts alter $\phi$ by the corresponding affine transformation but leave (9.4) unchanged wherever both descriptions are defined. Globally, topology may prevent a single affine coordinate system or potential from covering the entire manifold.

**Lecture 10 — Wednesday, October 12, 2022**

# 5. Optimal Transport and Curved Information Geometry

## Optimal Transport and $c$-Duality

Optimal transport compares probability measures through the least expensive way to rearrange one into the other. Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces, let $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, and let

$$c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

be a measurable cost.

**Definition 10.1** The **Monge problem** is

$$\inf_{T_{\#}\mu=\nu} \int_{\mathcal{X}} c(x, T(x))\,\mathrm{d}\mu(x), \tag{10.1}$$

where the infimum is taken over measurable maps $T\colon \mathcal{X} \to \mathcal{Y}$ and $T_{\#}\mu = \mu \circ T^{-1}$ is the pushforward measure.

The deterministic constraint may be infeasible. For instance, no map can push two equally weighted atoms onto three equally weighted atoms. Kantorovich's relaxation allows a unit of mass to split.

**Definition 10.2** A **coupling** of $\mu$ and $\nu$ is a probability measure $\pi$ on $\mathcal{X} \times \mathcal{Y}$ with first marginal $\mu$ and second marginal $\nu$. Write $\Pi(\mu, \nu)$ for the set of all couplings. The **Kantorovich transport cost** is

$$\mathcal{T}_c(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\,\mathrm{d}\pi(x, y). \tag{10.2}$$

The feasible set is never empty because it contains $\mu \otimes \nu$. It is convex, and the objective in (10.2) is linear in $\pi$. A transport map is recovered as the special coupling

$$\pi = (\mathrm{Id}, T)_{\#}\mu,$$

whose conditional distribution at $x$ is the point mass at $T(x)$.

Adding terms that depend on only one marginal does not change the optimizer. If

$$\widetilde{c}(x,y) = c(x,y) + a(x) + b(y),$$

then every coupling satisfies

$$\int \widetilde{c}\,\mathrm{d}\pi = \int c\,\mathrm{d}\pi + \int a\,\mathrm{d}\mu + \int b\,\mathrm{d}\nu.$$

The last two terms are fixed over $\Pi(\mu,\nu)$.

If $(\mathcal{X}, d)$ is a metric space and $r \geq 1$, the cost $c(x,y) = d(x,y)^r$ gives the ***r*-Wasserstein distance**

$$W_r(\mu,\nu) = \left( \inf_{\pi\in\Pi(\mu,\nu)} \int d(x,y)^r \,\mathrm{d}\pi(x,y) \right)^{1/r}. \tag{10.3}$$

It is a metric on the measures with finite $r$th moment. Since

$$W_r(\delta_x, \delta_y) = d(x,y),$$

the map $x \mapsto \delta_x$ embeds the base space isometrically into its Wasserstein space. This geometry differs sharply from relative entropy: two mutually singular measures may lie a finite Wasserstein distance apart even though their relative entropy is infinite.

Optimality is governed by a convex duality adapted to the cost.

**Definition 10.3** For $\varphi\colon \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$, its ***c*-transform** is

$$\varphi^c(y) = \inf_{x\in\mathcal{X}} \{c(x,y) - \varphi(x)\}. \tag{10.4}$$

The corresponding transform of a function on $\mathcal{Y}$ is defined analogously. The function $\varphi$ is ***c*-concave** if $\varphi \not\equiv -\infty$ and $\varphi^{cc} = \varphi$.

For a $c$-concave $\varphi$, its ***c*-superdifferential** is

$$\partial^c\varphi = \{(x,y) : \varphi(x) + \varphi^c(y) = c(x,y)\}. \tag{10.5}$$

When precisely one such $y$ corresponds to $x$, write $y = \nabla^c\varphi(x)$ and call it the ***c*-gradient**.

The defining inequality

$$\varphi(x) + \varphi^c(y) \leq c(x,y) \tag{10.6}$$

is tight on $\partial^c\varphi$. This immediately gives cyclic optimality.

**Proposition 10.4** If $(x_i, y_i) \in \partial^c\varphi$ for $i = 1, \ldots, N$, then, for every permutation $\sigma$ of $\{1, \ldots, N\}$,

$$\sum_{i=1}^{N} c(x_i, y_i) \leq \sum_{i=1}^{N} c(x_i, y_{\sigma(i)}). \tag{10.7}$$

**Proof.** Equality in (10.6) gives

$$\sum_{i=1}^{N} c(x_i, y_i) = \sum_{i=1}^{N} \varphi(x_i) + \sum_{i=1}^{N} \varphi^c(y_i).$$

Permuting the $y_i$ leaves the last sum unchanged. Applying (10.6) to each pair $(x_i, y_{\sigma(i)})$ proves (10.7). $\square$

For the quadratic cost on $\mathbb{R}^d$,

$$c(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2,$$

$c$-concavity is ordinary convexity in disguise. Specifically, $\varphi$ is $c$-concave precisely when

$$\Phi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2 - \varphi(\boldsymbol{x}) \tag{10.8}$$

is lower semicontinuous and convex; under the natural identification, $\partial^c\varphi$ is the ordinary subdifferential $\partial\Phi$. This follows by expanding the squared distance in (10.4) and recognizing a Legendre–Fenchel transform.

**Theorem 10.5 (Fundamental theorem of optimal transport)** Assume $c$ is continuous, bounded below, and bounded above by an integrable sum $a(x) + b(y)$. For $\pi \in \Pi(\mu, \nu)$, the following are equivalent:

(i) The coupling $\pi$ solves (10.2).

(ii) The support of $\pi$ is $c$-cyclically monotone.

(iii) The support of $\pi$ is contained in $\partial^c\varphi$ for an integrable $c$-concave potential $\varphi$.

The theorem is the infinite-dimensional analogue of complementary slackness. Indeed, the dual linear program is

$$\mathcal{T}_c(\mu,\nu) = \sup_{\varphi(x)+\psi(y)\le c(x,y)} \left\{ \int \varphi \, \mathrm{d}\mu + \int \psi \, \mathrm{d}\nu \right\}, \tag{10.9}$$

and an optimal pair may be chosen as $(\varphi, \varphi^c)$. Weak duality follows at once by integrating the pointwise constraint against a coupling. Under the hypotheses of Theorem 10.5, compactness and lower semicontinuity supply strong duality and dual attainment.

**Theorem 10.6 (Brenier's theorem)** Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, and suppose that $\mu$ is absolutely continuous with respect to Lebesgue measure. For the quadratic cost, there is a convex function $\Phi$ such that

$$T = \nabla\Phi$$

is the unique optimal transport map up to $\mu$-null sets.

Thus a convex gradient is simultaneously a mirror map and an optimal transport map. Figure 15 emphasizes the difference between the deterministic graph of $T$ and a diffuse coupling.

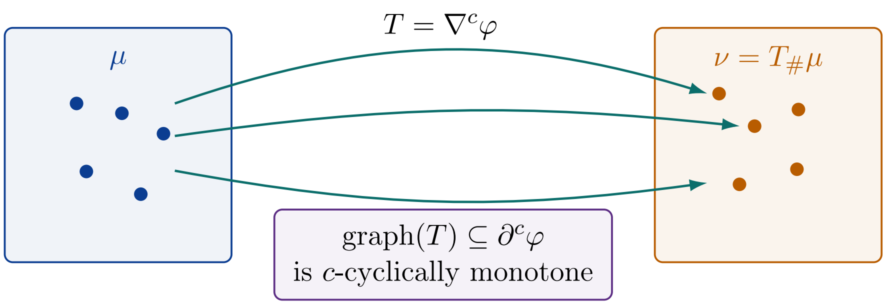


FIGURE 15

## The $c$-Divergence

The slack in the potential inequality (10.6) becomes a divergence when it is restricted to an optimal transport graph. Let $M$ be a $d$-dimensional manifold, let $c\colon M \times M \to \mathbb{R}$ be smooth, and suppose $\varphi$ and $\psi = \varphi^c$ are smooth. Assume that the $c$-gradient

$$T = \nabla^c\varphi$$

is a diffeomorphism onto its image. Its graph

$$G = \{(p, T(p)) : p \in M\} \subseteq M \times M \tag{10.10}$$

is a $d$-dimensional submanifold. The two projections provide primal and dual coordinate descriptions of the same point of $G$.

**Definition 10.7** For $z = (p, q)$ and $z' = (p', q')$ in $G$, the ***c*-divergence** is

$$D_c(z\|z') = c(p, q') - \varphi(p) - \psi(q'). \tag{10.11}$$

The potential inequality makes $D_c$ nonnegative. If equality in (10.6) identifies only the graph of $T$, then $D_c(z\|z') = 0$ precisely when $z = z'$. In local coordinates $\boldsymbol{x}$ for the first factor and $\boldsymbol{y}$ for the second, suppose the graph is $c$-monotone and the nondegeneracy condition

$$\det\left[\frac{\partial^2 c}{\partial x^i \partial y^j}(p, q)\right]_{i,j=1}^{d} \neq 0 \tag{10.12}$$

holds. The induced quadratic form is positive semidefinite by $c$-monotonicity and nonsingular by this condition together with the local diffeomorphism property of $T$; it is therefore positive definite. Thus $D_c$ is an information geometric divergence on $G$.

The geometry of (10.11) is visible in Figure 16. The mixed pair $(p, q')$ lies off the optimal graph; its vertical and horizontal projections identify $z$ and $z'$. The divergence is the potential slack at that mixed pair.

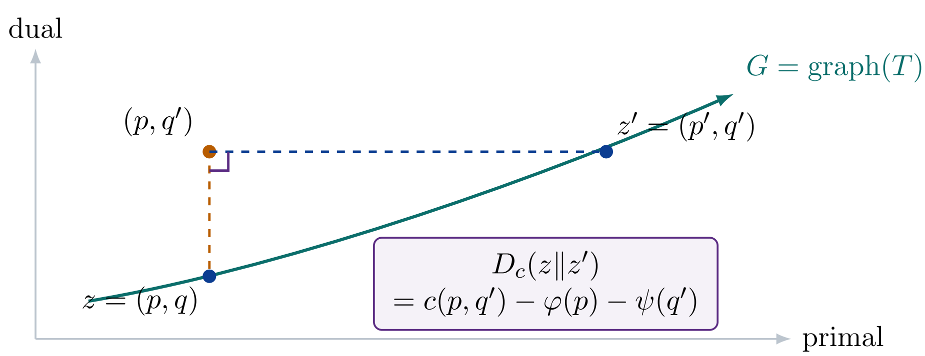


FIGURE 16

**Proposition 10.8** Let $\nu = T_{\#}\mu$, suppose the hypotheses of Kantorovich duality hold, and let $\pi \in \Pi(\mu, \nu)$. For $(p, q')$ distributed according to $\pi$, set

$$z = (p, T(p)) \qquad \text{and} \qquad z' = (T^{-1}(q'), q').$$

Then

$$\int D_c(z\|z')\,\mathrm{d}\pi(p,q') = \int c(p,q')\,\mathrm{d}\pi(p,q') - \mathcal{T}_c(\mu,\nu). \tag{10.13}$$

**Proof.** Since $(\varphi,\psi)$ is an optimal pair of Kantorovich potentials,

$$\mathcal{T}_c(\mu,\nu) = \int \varphi\,\mathrm{d}\mu + \int \psi\,\mathrm{d}\nu.$$

Integrating (10.11) against $\pi$ and using its two marginal constraints gives (10.13). □

For the quadratic cost, put

$$\Phi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2 - \varphi(\boldsymbol{x}) \qquad \text{and} \qquad \psi(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y}\|^2 - \Phi^*(\boldsymbol{y}).$$

Then $T = \nabla\Phi$, and

$$D_c(z\|z') = \Phi(\boldsymbol{x}) + \Phi^*(\boldsymbol{y}') - \boldsymbol{x}^\top\boldsymbol{y}' = B_\Phi[\boldsymbol{x} : \boldsymbol{x}'] = B_{\Phi^*}[\boldsymbol{y}' : \boldsymbol{y}]. \tag{10.14}$$

Thus Bregman divergence is the optimality gap attached to quadratic transport.

At the opposite extreme, every information geometric divergence is formally a $c$-divergence. Set $c(p,q) = D(p\|q)$, choose zero potentials, and take the identity transport. The graph is the diagonal in $M \times M$, and (10.11) returns $D$. The value of the framework is not this tautological representation, but the nontrivial cases in which the transport cost already carries useful geometry.

---

### Lecture 11 — Monday, October 17, 2022

## The Kim–McCann Pseudo-Riemannian Framework

The $c$-divergence lives on the optimal graph $G$, but the cost is defined on the full product $M \times M$. Kim and McCann's construction places an indefinite metric on this ambient product and recovers the information geometry of $G$ by restriction and projection.

**Definition 11.1** For $z = (p, q)$ and $z' = (p', q')$ in $M \times M$, the **cross-difference** is

$$\delta(z, z') = c(p, q') + c(p', q) - c(p, q) - c(p', q'). \tag{11.1}$$

This is the change in cost when the two assignments

$$p \longmapsto q \qquad \text{and} \qquad p' \longmapsto q'$$

are replaced by the crossed assignments, as illustrated in Figure 17.

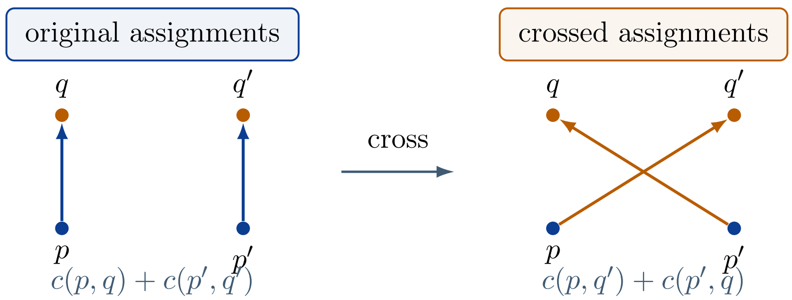


FIGURE 17

Therefore, $c$-cyclical monotonicity implies $\delta(z, z') \geq 0$ when $z, z' \in G$. On the optimal graph, the cross-difference is exactly the symmetrized $c$-divergence:

$$\delta(z, z') = D_c(z \| z') + D_c(z' \| z). \tag{11.2}$$

Indeed, the four potential terms in the two divergences collapse to $c(p, q) + c(p', q')$ because the potential inequality is tight on $G$.

**Definition 11.2** A **pseudo-Riemannian metric** $h$ is a smooth, symmetric, nondegenerate bilinear form on each tangent space; positive definiteness is not required. If the associated matrix has $r$ positive and $s$ negative eigenvalues, then $(r, s)$ is the **signature** of $h$.

Nondegeneracy is enough to define a Levi–Civita connection by the Koszul formula and hence a curvature tensor. Vectors may now have positive, negative, or zero squared length. Nonzero vectors with $h(v, v) = 0$ are called **null**.

Take local coordinates $\boldsymbol{x}$ and $\boldsymbol{y}$ on the two copies of $M$, and put

$$\mathbf{C}(\boldsymbol{x}, \boldsymbol{y}) = \left[ \frac{\partial^2 c}{\partial x^i \partial y^j}(\boldsymbol{x}, \boldsymbol{y}) \right]_{i,j=1}^{d}. \tag{11.3}$$

Assume that $\mathbf{C}$ is invertible.

**Proposition 11.3** The block matrix

$$\mathbf{h} = \frac{1}{2}\begin{bmatrix} \mathbf{0} & -\mathbf{C} \\ -\mathbf{C}^{\top} & \mathbf{0} \end{bmatrix} \tag{11.4}$$

defines a pseudo-Riemannian metric of signature $(d, d)$ on $M \times M$. For $v = (\boldsymbol{a}, \boldsymbol{b}) \in T_pM \oplus T_qM$,

$$h(v, v) = -\boldsymbol{a}^{\top}\mathbf{C}\boldsymbol{b}. \tag{11.5}$$

Equivalently, $h$ is the mixed second derivative of the cross-difference along the diagonal of $(M \times M)^2$.

**Proof.** Symmetry is explicit in (11.4). If $h((\boldsymbol{a}, \boldsymbol{b}), \cdot) = 0$, testing separately against vectors in the two factors gives

$$\mathbf{C}^{\top}\boldsymbol{a} = \mathbf{0} \qquad \text{and} \qquad \mathbf{C}\boldsymbol{b} = \mathbf{0}.$$

Invertibility forces $\boldsymbol{a} = \boldsymbol{b} = \mathbf{0}$, so $h$ is nondegenerate. A singular value decomposition of $\mathbf{C}$ transforms the metric into $d$ independent blocks with one positive and one negative eigenvalue each. Its signature is therefore $(d, d)$. $\square$

For the quadratic cost, $\mathbf{C} = -\mathbf{I}$, and

$$\mathbf{h} = \frac{1}{2}\begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}. \tag{11.6}$$

In dimension one, the two factor directions are null, while their diagonal and anti-diagonal combinations have opposite signs. Figure 18 shows how a $c$-monotone graph sits in the positive region of this split geometry.

The restriction statement is fundamental.

**Theorem 11.4** Let $D_c$ be the $c$-divergence on the optimal graph $G$, and let $g$ be the metric it induces. Then

$$h|_{T_zG \times T_zG} = g_z \tag{11.7}$$

for every $z \in G$.

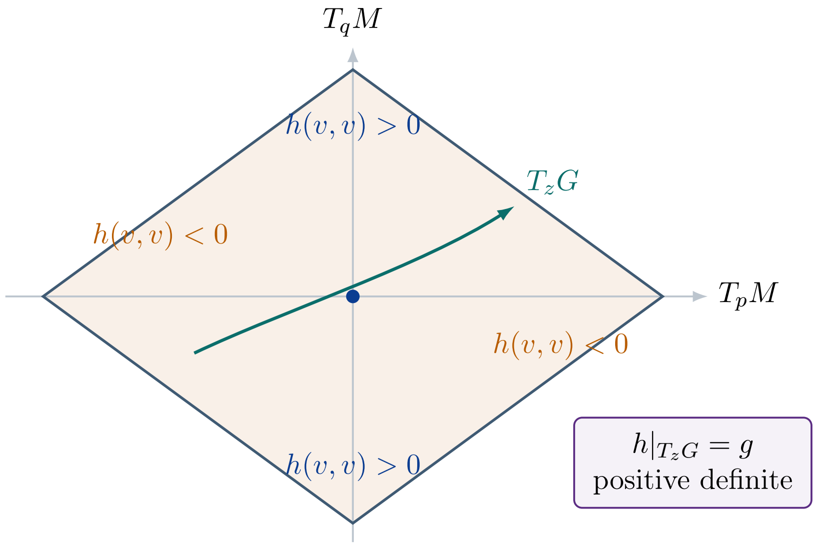


FIGURE 18

Let $\overline{\overline{\nabla}}$ be the Levi–Civita connection of $h$, and, on $M \times T(M)$, define the two projections

$$\pi_0(p,q) = (p, T(p)), \tag{11.8}$$
$$\pi_1(p,q) = (T^{-1}(q), q). \tag{11.9}$$

If $\overline{\overline{\nabla}}$ is projected to $TG$ using $\mathrm{d}\pi_0$ and $\mathrm{d}\pi_1$, the resulting connections are the primal and dual connections induced by $D_c$, respectively.

**Proof.** In primal coordinates, a tangent vector to $G$ has the form

$$v = (\boldsymbol{a}, \mathrm{d}T_p\boldsymbol{a}).$$

Substitution into (11.5) gives

$$h(v,v) = -\boldsymbol{a}^\top \mathbf{C}(p, T(p))\, \mathrm{d}T_p\boldsymbol{a},$$

which is the second mixed derivative of $D_c$ and hence equals $g(v,v)$.

For the connection statement, the Koszul formula applied to the block metric shows that its only nonzero Christoffel symbols differentiate entirely within the first factor or entirely within the second. The first family is

$$\overline{\Gamma}_{ij}^{k} = \sum_{m=1}^{d} \frac{\partial^3 c}{\partial x^i \partial x^j \partial y^m} (\mathbf{C}^{-1})^{mk}, \tag{11.10}$$

and the second family is its analogue with the factors swapped. Projecting the first family by $\mathrm{d}\pi_0$ reproduces (8.22); projecting the second by $\mathrm{d}\pi_1$ reproduces (8.23). $\square$

The ambient metric is indefinite, yet every $c$-monotone optimal graph is spacelike: its induced metric is positive definite. In this way, one cost produces two linked geometries. The ambient pseudo-Riemannian curvature controls transport regularity, while the induced metric and projected connections describe statistical separation along the optimal graph.

## Entropic Optimal Transport

Entropic regularization replaces the linear transport problem by a strictly convex one. For $\varepsilon > 0$, define

$$\mathcal{T}_{c,\varepsilon}(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \left\{ \int c \,\mathrm{d}\pi + \varepsilon \,\mathrm{KL}(\pi \| \mu \otimes \nu) \right\}. \tag{11.11}$$

Under mild hypotheses, strict convexity of relative entropy gives a unique optimizer $\pi_\varepsilon$. In the discrete setting, alternating rescaling of its two marginals yields the Sinkhorn algorithm.

The optimizer has the Gibbs form

$$\frac{\mathrm{d}\pi_\varepsilon}{\mathrm{d}(\mu \otimes \nu)}(p, q) = \exp\left( \frac{\varphi_\varepsilon(p) + \psi_\varepsilon(q) - c(p, q)}{\varepsilon} \right), \tag{11.12}$$

where $\varphi_\varepsilon$ and $\psi_\varepsilon$ are **Schrödinger potentials**. They are chosen so that the density in (11.12) has marginals $\mu$ and $\nu$.

As $\varepsilon \downarrow 0$, entropic transport concentrates around unregularized optimal plans under suitable compactness and integrability conditions. Suppose the Schrödinger potentials converge to Kantorovich potentials $\varphi$ and $\psi$, with errors negligible on the scale under consideration, and suppose the limiting plan is concentrated on the graph $G$. For a pair $(p, q)$ off the graph, let

$$z = (p, T(p)) \qquad \text{and} \qquad z' = (T^{-1}(q), q).$$

Then the exponent in (11.12) has the leading behaviour

$$\frac{\mathrm{d}\pi_\varepsilon}{\mathrm{d}(\mu \otimes \nu)}(p, q) \approx \exp\left( -\frac{D_c(z \| z')}{\varepsilon} \right). \tag{11.13}$$

The concentration mechanism in (11.13) is depicted in Figure 19. Points with appreciable mass are progressively confined to a narrow neighbourhood of the optimal graph as $\varepsilon$ decreases.

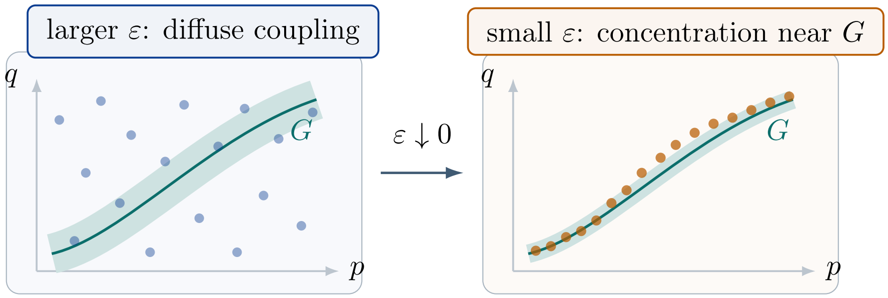


FIGURE 19

Thus, under these convergence assumptions, the $c$-divergence is the leading exponential penalty for leaving the optimal graph. It also enters higher-order small-$\varepsilon$ expansions of the regularized transport cost, where the metric, dual connections, and ambient curvature describe increasingly fine corrections.

---

### Lecture 12 — Wednesday, October 19, 2022

## Logarithmic Divergence and $\lambda$-Duality

Bregman geometry is flat because it is built from the bilinear cost $-\boldsymbol{x}^\top \boldsymbol{y}$. A logarithmic deformation of that cost produces a comparably explicit geometry with nonzero constant curvature.

Fix $\lambda \in \mathbb{R} \setminus \{0\}$. On the region where $1 + \lambda \boldsymbol{x}^\top \boldsymbol{y} > 0$, define the **logarithmic cost**

$$c_\lambda(\boldsymbol{x}, \boldsymbol{y}) = -\frac{1}{\lambda} \log\bigl(1 + \lambda \boldsymbol{x}^\top \boldsymbol{y}\bigr). \tag{12.1}$$

The classical pairing is recovered continuously:

$$\lim_{\lambda \to 0} c_\lambda(\boldsymbol{x}, \boldsymbol{y}) = -\boldsymbol{x}^\top \boldsymbol{y}. \tag{12.2}$$

**Definition 12.1** Let $\Omega_0 \subseteq \mathbb{R}^d$ be open and convex. A smooth function $\phi\colon \Omega_0 \to \mathbb{R}$ is **regular $c_\lambda$-convex** if

$$\Phi_\lambda(\boldsymbol{x}) := \frac{\exp(\lambda \phi(\boldsymbol{x})) - 1}{\lambda} \tag{12.3}$$

has positive definite Hessian and

$$1 - \lambda \nabla \phi(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{x} > 0 \tag{12.4}$$

for every $\boldsymbol{x} \in \Omega_0$.

The exponential transform in (12.3) is an ordinary convex generator. Rewriting its tangent inequality in logarithmic form leads to the curved analogue of Bregman divergence.

**Theorem 12.2** Let $\phi$ be regular $c_\lambda$-convex. Then the following statements hold:

(i) The **$\boldsymbol{\lambda}$-logarithmic divergence**

$$L_{\lambda,\phi}[\boldsymbol{x} : \boldsymbol{x}'] = \phi(\boldsymbol{x}) - \phi(\boldsymbol{x}') - \frac{1}{\lambda} \log\left(1 + \lambda \nabla\phi(\boldsymbol{x}')^\mathsf{T}(\boldsymbol{x} - \boldsymbol{x}')\right) \tag{12.5}$$

is nonnegative and vanishes precisely when $\boldsymbol{x} = \boldsymbol{x}'$.

(ii) The **$\boldsymbol{\lambda}$-gradient**

$$\nabla^{c_\lambda}\phi(\boldsymbol{x}) = \frac{\nabla\phi(\boldsymbol{x})}{1 - \lambda\nabla\phi(\boldsymbol{x})^\mathsf{T}\boldsymbol{x}} \tag{12.6}$$

is a diffeomorphism from $\Omega_0$ onto an open dual domain $\Omega_0'$. Every $\boldsymbol{x} \in \Omega_0$ and $\boldsymbol{y} \in \Omega_0'$ satisfy

$$1 + \lambda\boldsymbol{x}^\mathsf{T}\boldsymbol{y} > 0.$$

(iii) If $\boldsymbol{y} = \nabla^{c_\lambda}\phi(\boldsymbol{x})$, define

$$\psi(\boldsymbol{y}) = -c_\lambda(\boldsymbol{x}, \boldsymbol{y}) - \phi(\boldsymbol{x}). \tag{12.7}$$

Then $\phi$ and $\psi$ are $c_\lambda$-conjugate:

$$\phi(\boldsymbol{x}) = \sup_{\boldsymbol{y}' \in \Omega_0'} \{-c_\lambda(\boldsymbol{x}, \boldsymbol{y}') - \psi(\boldsymbol{y}')\}, \tag{12.8}$$

$$\psi(\boldsymbol{y}) = \sup_{\boldsymbol{x}' \in \Omega_0} \{-c_\lambda(\boldsymbol{x}', \boldsymbol{y}) - \phi(\boldsymbol{x}')\}. \tag{12.9}$$

Moreover,

$$(\nabla^{c_\lambda}\phi)^{-1}(\boldsymbol{y}) = \nabla^{c_\lambda}\psi(\boldsymbol{y}) = \frac{\nabla\psi(\boldsymbol{y})}{1 - \lambda\nabla\psi(\boldsymbol{y})^\mathsf{T}\boldsymbol{y}}. \tag{12.10}$$

**Proof.** Convexity of $\Phi_\lambda$ gives

$$\Phi_\lambda(\boldsymbol{x}) \geq \Phi_\lambda(\boldsymbol{x}') + \nabla\Phi_\lambda(\boldsymbol{x}')^\mathsf{T}(\boldsymbol{x} - \boldsymbol{x}'). \tag{12.11}$$

Since

$$\nabla\Phi_\lambda = \exp(\lambda\phi)\nabla\phi,$$

equation (12.11), after multiplication by the appropriate sign of $\lambda$ and taking logarithms, is equivalent to

$$L_{\lambda,\phi}[\boldsymbol{x} : \boldsymbol{x}'] \geq 0.$$

Strict convexity gives equality only on the diagonal. The stationarity condition in the $c_\lambda$-transform is

$$\nabla\phi(\boldsymbol{x}) = \frac{\boldsymbol{y}}{1+\lambda\boldsymbol{x}^\top\boldsymbol{y}},$$

which rearranges to (12.6). The inverse function theorem and the positive definite Hessian assumption give local invertibility; regularity gives the stated global diffeomorphism onto the image. Substitution of the stationary pair into the transform yields (12.7). Repeating the same calculation with the two variables interchanged gives (12.10) and the conjugacy identities. □

On the manifold parameterized by primal coordinates $\boldsymbol{x}$ and dual coordinates $\boldsymbol{y} = \nabla^{c_\lambda}\phi(\boldsymbol{x})$, the divergence has the self-dual form

$$D(p\|q) = \phi(\boldsymbol{x}_p) + \psi(\boldsymbol{y}_q) - \frac{1}{\lambda}\log\bigl(1+\lambda\boldsymbol{x}_p^\top\boldsymbol{y}_q\bigr). \tag{12.12}$$

As $\lambda \to 0$, (12.5) converges to $B_\phi[\boldsymbol{x} : \boldsymbol{x}']$.

Differentiating (12.5) gives the induced metric

$$\mathbf{G}_\lambda(\boldsymbol{x}) = \nabla^2\phi(\boldsymbol{x}) + \lambda\nabla\phi(\boldsymbol{x})\nabla\phi(\boldsymbol{x})^\top = \exp(-\lambda\phi(\boldsymbol{x}))\nabla^2\Phi_\lambda(\boldsymbol{x}). \tag{12.13}$$

It is therefore a conformal Hessian metric. There is also an exact connection to Bregman divergence:

$$L_{\lambda,\phi}[\boldsymbol{x} : \boldsymbol{x}'] = -\frac{1}{\lambda}\log\left(1-\lambda\exp(-\lambda\phi(\boldsymbol{x}))B_{\Phi_\lambda}[\boldsymbol{x} : \boldsymbol{x}']\right). \tag{12.14}$$

The logarithmic tangent model and its gap are illustrated in Figure 20.

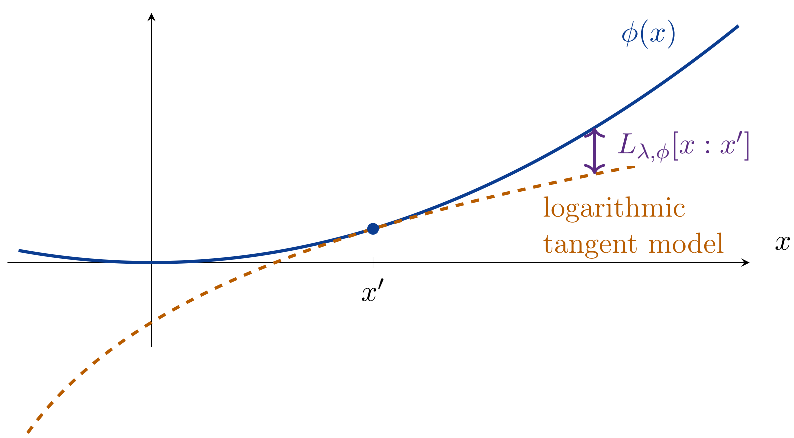


FIGURE 20

The curvature enters through projective rather than affine flatness.

**Definition 12.3** Two torsion-free connections $\nabla$ and $\widetilde{\nabla}$ are **projectively equivalent** if there is a one-form $\tau$ such that

$$\widetilde{\nabla}_X Y = \nabla_X Y + \tau(X)Y + \tau(Y)X. \tag{12.15}$$

A connection is **projectively flat** if it is projectively equivalent to a flat connection.

Projectively equivalent connections have the same unparameterized geodesics. Thus a projectively flat connection has geodesic traces that become straight lines in suitable coordinates, although their affine parameterizations need not be linear.

**Definition 12.4** An affine connection $\nabla$ has **constant sectional curvature** $\kappa$ with respect to $g$ if

$$R(X,Y)Z = \kappa\big(g(Y,Z)X - g(X,Z)Y\big) \tag{12.16}$$

for all vector fields $X, Y, Z$.

**Theorem 12.5** The dualistic structure induced by $L_{\lambda,\phi}$ has the following properties.

(i) The primal connection is projectively equivalent to the flat connection in $\boldsymbol{x}$-coordinates, and the dual connection is projectively equivalent to the flat connection in $\boldsymbol{y}$-coordinates.

(ii) Both connections have constant sectional curvature $\lambda$.

(iii) If the primal geodesic from $q$ to $r$ and the dual geodesic from $q$ to $p$ meet orthogonally at $q$, then

$$D(q\|p) + D(r\|q) = D(r\|p). \tag{12.17}$$

The converse holds whenever the required primal and dual geodesic segments remain inside their coordinate domains.

The theorem makes $\lambda$ a genuine curvature parameter. When $\lambda = 0$, projective flatness becomes ordinary flatness and the Bregman Pythagorean theorem is recovered. Moreover, every dualistic structure that is dually projectively flat with nonzero constant curvature $\lambda$ locally admits potentials $\phi, \psi$ for which (12.12) is a canonical divergence. Logarithmic divergence is therefore the counterpart of Bregman divergence in constant curvature.

## The $\lambda$-Exponential Family

The cost $c_\lambda$ produces a power-law deformation of the exponential function:

$$\exp\bigl(-c_\lambda(\boldsymbol{x}, \boldsymbol{y})\bigr) = \Bigl[1 + \lambda \boldsymbol{x}^\top \boldsymbol{y}\Bigr]_+^{1/\lambda}, \tag{12.18}$$

where $[t]_+ = \max\{t, 0\}$ and the entire power-law expression is defined to be zero when its bracket is nonpositive. This convention is needed when the exponent is negative. The formula is used as an ordinary real power wherever the bracket is positive. This suggests a family compatible with $\lambda$-duality in the same way that ordinary exponential families are compatible with Legendre duality.

**Definition 12.6** Let $(\mathcal{X}, \mathcal{A}, \nu)$ be a measure space and let $\boldsymbol{F} \colon \mathcal{X} \to \mathbb{R}^d$ be measurable. Define

$$\phi(\boldsymbol{\theta}) = \log \int_{\mathcal{X}} \Bigl[1 + \lambda \boldsymbol{\theta}^\top \boldsymbol{F}(x)\Bigr]_+^{1/\lambda} \mathrm{d}\nu(x). \tag{12.19}$$

On the natural parameter domain where (12.19) is finite, the densities

$$p_{\boldsymbol{\theta}}(x) = \Big[1 + \lambda \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{F}(x)\Big]_{+}^{1/\lambda} \exp\big(-\phi(\boldsymbol{\theta})\big) \tag{12.20}$$

form a **$\boldsymbol{\lambda}$-exponential family**. The function $\phi$ is its **divisive potential**.

When $\lambda \to 0$, the power term in (12.20) converges to $\exp\big(\boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{F}(x)\big)$, so the ordinary exponential family is recovered.

Put

$$q = 1 - \lambda. \tag{12.21}$$

The $q$-exponential

$$\exp_q(t) = [1 + (1-q)t]_{+}^{1/(1-q)} \tag{12.22}$$

has the same power-law form. A conventional $q$-exponential family often places the normalizing potential inside $\exp_q$; this is subtractive normalization. Formula (12.20) normalizes by an external exponential factor. The divisive normalization is the one that fits the $c_\lambda$-conjugacy exactly.

**Example 12.7** For $q < 3$, the centred $q$-Gaussian scale family on $\mathbb{R}$ has densities proportional to

$$p_\theta(x) = \sqrt{\theta}\, \exp_q(-\theta x^2), \quad \theta > 0. \tag{12.23}$$

It is a $\lambda$-exponential family with $\lambda = 1 - q$ and statistic $F(x) = -x^2$. Its divisive potential has the form

$$\phi(\theta) = -\frac{1}{2} \log \theta + C_\lambda. \tag{12.24}$$

For $q < 1$, its positive-density region is the parameter-dependent interval

$$|x| < \frac{1}{\sqrt{(1-q)\theta}}.$$

For $q = 1$, it is Gaussian; for $1 < q < 3$, it has polynomial tails and is a rescaled Student distribution. It cannot be normalized when $q \geq 3$.

Figure 21 compares the profiles after normalizing their heights. Negative $\lambda$ produces heavier tails, while positive $\lambda$ produces compact support.

The central identities of the family parallel the corresponding identities for exponential families, with ordinary expectations replaced by escort expectations.

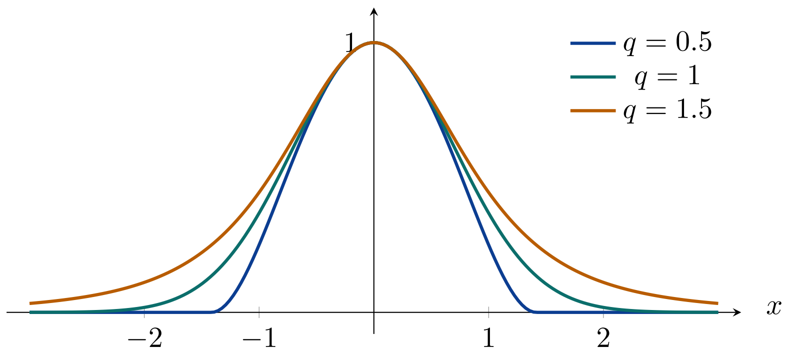


FIGURE 21

**Theorem 12.8** Consider the family in (12.20). Assume that $q = 1 - \lambda > 0$, the natural parameter domain is open, the support is independent of $\boldsymbol{\theta}$, and differentiation under the integral sign is valid to the required orders. Then the following statements hold.

(i) The divisive potential satisfies the regularity conditions in Definition 12.1, provided its transformed Hessian is nondegenerate.

(ii) For $\boldsymbol{\theta}, \boldsymbol{\theta}'$ in the parameter domain,

$$L_{\lambda,\phi}[\boldsymbol{\theta} : \boldsymbol{\theta}'] = H_q(\mathbb{P}_{\boldsymbol{\theta}'} \| \mathbb{P}_{\boldsymbol{\theta}}). \tag{12.25}$$

(iii) The dual parameter $\boldsymbol{\eta} = \nabla^{c_\lambda} \phi(\boldsymbol{\theta})$ is the **escort expectation**

$$\boldsymbol{\eta} = \frac{\int_{\mathcal{X}} \boldsymbol{F}(x) p_{\boldsymbol{\theta}}(x)^q \, \mathrm{d}\nu(x)}{\int_{\mathcal{X}} p_{\boldsymbol{\theta}}(x)^q \, \mathrm{d}\nu(x)}. \tag{12.26}$$

(iv) The dual potential is negative Rényi entropy:

$$\psi(\boldsymbol{\eta}) = -H_q(p_{\boldsymbol{\theta}}) \qquad \text{and} \qquad \boldsymbol{\eta} = \nabla^{c_\lambda} \phi(\boldsymbol{\theta}). \tag{12.27}$$

(v) The metric induced by $L_{\lambda,\phi}$ is

$$g = q g^{\mathrm{F}}, \tag{12.28}$$

where $g^{\mathrm{F}}$ is the Fisher–Rao metric of the family.

(vi) Among densities $p$ equivalent to $\nu$ that have finite Rényi quantities and the same escort expectation (12.26), $p_{\boldsymbol{\theta}}$ maximizes Rényi entropy of order $q$. Equality is attained only by $p = p_{\boldsymbol{\theta}}$ almost everywhere.

**Proof.** Write

$$A_{\boldsymbol{\theta}}(x) = 1 + \lambda \boldsymbol{\theta}^\top \boldsymbol{F}(x)$$

on the common support. Differentiating the normalizing identity gives

$$\nabla\phi(\boldsymbol{\theta}) = \exp(-\phi(\boldsymbol{\theta})) \int_{\mathcal{X}} A_{\boldsymbol{\theta}}(x)^{1/\lambda-1} \boldsymbol{F}(x)\, \mathrm{d}\nu(x). \tag{12.29}$$

Since $q/\lambda = 1/\lambda - 1$, a short calculation yields

$$1 - \lambda \boldsymbol{\theta}^\top \nabla\phi(\boldsymbol{\theta}) = \exp(-\phi(\boldsymbol{\theta})) \int_{\mathcal{X}} A_{\boldsymbol{\theta}}(x)^{1/\lambda-1}\, \mathrm{d}\nu(x) > 0. \tag{12.30}$$

Dividing (12.29) by (12.30) proves (12.26) and the second regularity condition.

For the Rényi identity, direct substitution gives

$$\int p_{\boldsymbol{\theta}'}^{q} p_{\boldsymbol{\theta}}^{1-q}\, \mathrm{d}\nu = \exp\big(\lambda(\phi(\boldsymbol{\theta}') - \phi(\boldsymbol{\theta}))\big) \left(1 + \lambda \nabla\phi(\boldsymbol{\theta}')^\top (\boldsymbol{\theta} - \boldsymbol{\theta}')\right).$$

Applying the definition of Rényi divergence and using $q - 1 = -\lambda$ proves (12.25). The local second-order expansion of Rényi divergence is $q$ times the Fisher–Rao metric, proving (12.28).

Equation (12.30) also gives

$$H_q(p_{\boldsymbol{\theta}}) = \phi(\boldsymbol{\theta}) + \frac{1}{\lambda} \log\big(1 - \lambda \boldsymbol{\theta}^\top \nabla\phi(\boldsymbol{\theta})\big).$$

Comparing this expression with (12.7) proves (12.27). Finally, let $p$ have the same escort expectation $\boldsymbol{\eta}$ as $p_{\boldsymbol{\theta}}$. Since $1 - q = \lambda$,

$$\begin{aligned}
H_q(\mathbb{P} \| \mathbb{P}_{\boldsymbol{\theta}}) &= -\frac{1}{\lambda} \log\left(\int p(x)^q p_{\boldsymbol{\theta}}(x)^\lambda\, \mathrm{d}\nu(x)\right) \\
&= -\frac{1}{\lambda} \log\left(\mathrm{e}^{-\lambda\phi(\boldsymbol{\theta})} \int p(x)^q \big(1 + \lambda \boldsymbol{\theta}^\top \boldsymbol{F}(x)\big)\, \mathrm{d}\nu(x)\right) \\
&= -H_q(p) - \frac{1}{\lambda} \log\big(1 + \lambda \boldsymbol{\theta}^\top \boldsymbol{\eta}\big) + \phi(\boldsymbol{\theta}).
\end{aligned}$$

The same identity with $p = p_{\boldsymbol{\theta}}$ makes the left-hand side zero, so

$$H_q(\mathbb{P} \| \mathbb{P}_{\boldsymbol{\theta}}) = H_q(p_{\boldsymbol{\theta}}) - H_q(p) \geq 0.$$

Strictness of the divergence gives the equality condition. □

Unlike an ordinary exponential family, a fixed-dimensional $\lambda$-exponential family is generally not closed under independent sampling. For a one-parameter model with $A_\theta(x) = 1 + \lambda\theta F(x)$, two observations have joint power term

$$A_\theta(x_1)A_\theta(x_2) = 1 + \lambda\theta\big(F(x_1) + F(x_2)\big) + \lambda^2\theta^2 F(x_1)F(x_2). \tag{12.31}$$

The model becomes a curved one-dimensional subfamily of an ambient family with at least two statistics. Higher sample sizes create higher-order interaction statistics. This loss of a fixed-dimensional sufficient statistic is one of the main computational differences from ordinary exponential families.

## Directions Beyond the Regular Finite-Dimensional Theory

The regular finite-dimensional setting exposes a coherent geometry, but several important regimes lie beyond it.

1. Statistical inference for generalized exponential families requires scalable optimization, approximation, and data-driven selection of deformation parameters such as $\lambda$ or $q$.
2. Infinite-dimensional statistical manifolds are naturally formulated using functional analysis. Practical use calls for finite-dimensional approximations that preserve the relevant metric, divergence, and dual transports.
3. Singular models, parameter-dependent supports, and nonidentifiable parameterizations require geometry without the usual regularity assumptions. Mixture models and neural networks are prominent examples.
4. Classical asymptotic information geometry keeps the dimension fixed while sample size grows. High-dimensional regimes, in which dimension and sample size grow together, may exhibit different geometric limits.
5. A metric and dual connections define primal and dual Laplace-type operators. Understanding the corresponding diffusion processes could connect stochastic dynamics with dualistic geometry.
6. The relation between optimal transport and information geometry extends beyond the restric-

tion theorem in Theorem 11.4. Entropic asymptotics, transport regularity, and nonquadratic costs suggest further links among divergence, curvature, and statistical computation.

Across these directions, the recurring principle is that a global contrast or transport cost determines local geometry through its derivatives. The metric captures second-order distinguishability, dual connections capture third-order asymmetry, and curvature records the obstruction to simultaneously straightening both statistical coordinate systems.

# Appendix: Lecture and Tutorial Index