

Čencov's Theorem and the Fisher Metric

1 The Probability Simplex

The open probability simplex is

$$\Delta_n = \left\{ p = (p_1, \dots, p_n) \in \mathbb{R}^n \mid p_i > 0, \quad \sum_{i=1}^n p_i = 1 \right\}.$$

A point $p \in \Delta_n$ represents a probability distribution on the finite set

$$\{1, \dots, n\},$$

with

$$\mathbb{P}(i) = p_i.$$

Since the coordinates satisfy the constraint

$$\sum_i p_i = 1,$$

the tangent space at p is

$$T_p \Delta_n = \left\{ v = (v_1, \dots, v_n) \in \mathbb{R}^n \mid \sum_i v_i = 0 \right\}.$$

Thus a tangent vector v represents an infinitesimal variation of a probability distribution that preserves total mass.

2 The Fisher Metric

The Fisher metric on Δ_n is

$$g_p(v, w) = \sum_{i=1}^n \frac{v_i w_i}{p_i}, \quad v, w \in T_p \Delta_n.$$

In particular,

$$g_p(v, v) = \sum_i \frac{v_i^2}{p_i}.$$

If the probability distribution is parametrized by coordinates

$$p_i = p_i(\theta^1, \dots, \theta^m),$$

then

$$dp_i = \sum_a \frac{\partial p_i}{\partial \theta^a} d\theta^a,$$

and the metric tensor is

$$g_{ab} = \sum_i \frac{1}{p_i} \frac{\partial p_i}{\partial \theta^a} \frac{\partial p_i}{\partial \theta^b}.$$

Equivalently,

$$g_{ab} = \sum_i p_i \frac{\partial \log p_i}{\partial \theta^a} \frac{\partial \log p_i}{\partial \theta^b}.$$

3 Square-Root Coordinates

Let

$$\psi_i = \sqrt{p_i}.$$

Then

$$\sum_i \psi_i^2 = 1, \quad \psi_i > 0.$$

Thus the simplex is mapped to the positive orthant of the unit sphere

$$S^{n-1} = \left\{ \psi \in \mathbb{R}^n : \sum_i \psi_i^2 = 1 \right\}.$$

Since

$$dp_i = 2\psi_i d\psi_i,$$

we obtain

$$\frac{dp_i^2}{p_i} = 4 d\psi_i^2.$$

Hence

$$g_F = \sum_i \frac{dp_i^2}{p_i} = 4 \sum_i d\psi_i^2.$$

Therefore the Fisher metric is four times the ordinary spherical metric on the positive orthant of the unit sphere.

Equivalently, if we use

$$x_i = 2\sqrt{p_i},$$

then

$$\sum_i x_i^2 = 4$$

and the Fisher metric becomes exactly the Euclidean metric induced on the sphere of radius 2.

★ Spherical Simplex Geometry. Let $\Delta', \Delta'' \subset \Delta_n$ be two faces of the n simplex. If the two do not intersect (this implies that $\dim(\Delta') + \dim(\Delta'') \leq n-1$), then the Fisher (spherical) distance between the points $x' \in \Delta'$ and $x'' \in \Delta''$ is $\text{dist}_{2S^n}(x', x'') = \pi$ for **all** pairs of points in these faces.

4 Stochastic Maps

Let

$$K = (K_{ij})$$

be an $n \times m$ matrix satisfying

$$K_{ij} \geq 0, \quad \sum_{j=1}^m K_{ij} = 1$$

for every i .

Such a matrix is called a *stochastic matrix* or *Markov kernel*.

It defines a linear map

$$T_K : \Delta_n \longrightarrow \Delta_m$$

by

$$q_j = (T_K p)_j = \sum_{i=1}^n p_i K_{ij}.$$

In matrix notation,

$$q = pK.$$

Probabilistically,

$$K_{ij} = \mathbb{P}(Y = j \mid X = i).$$

A deterministic map

$$f : \{1, \dots, n\} \longrightarrow \{1, \dots, m\}$$

is a special stochastic map, with

$$K_{ij} = \begin{cases} 1, & j = f(i), \\ 0, & j \neq f(i). \end{cases}$$

Then

$$q_j = \sum_{i: f(i)=j} p_i.$$

5 Statistics

Given a map $S : X \rightarrow Y$, a statistician call X "*sample space*", Y "*value space*" and if X comes with a probability measure or a class of these, then T is called "*statistics*".

For a finite sample space

$$X = \{1, \dots, n\},$$

a statistic is a map

$$S : X \longrightarrow Y.$$

If

$$Y = \{1, \dots, m\},$$

then S linearly extends to what a statistician calls "*stochastic map*"

$$S_* : \Delta_n \longrightarrow \Delta_m$$

given by

$$(S_*p)_j = \sum_{i:S(i)=j} p_i,$$

and says that passing from X to the statistic $S(X)$ means "forgetting some of the information contained in X ."

6 Monotonicity Under Stochastic Maps

Let

$$T : \Delta_n \longrightarrow \Delta_m$$

be a stochastic (i.e. linear) map.

For the Fisher metric,

$$g_{T(p)}(dT_p(v), dT_p(v)) \leq g_p(v, v).$$

Equivalently,

$$|dT_p(v)|_F \leq |v|_F.$$

Thus stochastic processing decreases (possibly non-strictly) Fisher length.

7 Stochastic Embeddings

A stochastic map

$$E : \Delta_n \longrightarrow \Delta_m$$

is called a *stochastic embedding* if there exists a stochastic map

$$S : \Delta_m \longrightarrow \Delta_n$$

such that

$$S \circ E = \text{id}_{\Delta_n}.$$

In the statistics language (as it is done by GPT) one introduces additional random refinement without losing information about the original state, which is done as follows

Replace each outcome i by

$$(i, 1), \dots, (i, r_i).$$

Choose positive numbers

$$\alpha_{ia} > 0, \quad \sum_{a=1}^{r_i} \alpha_{ia} = 1.$$

Define

$$E(p)_{ia} = \alpha_{ia} p_i.$$

Thus

$$p_i \longmapsto (\alpha_{i1} p_i, \dots, \alpha_{ir_i} p_i).$$

The original distribution can be recovered by summing:

$$p_i = \sum_a E(p)_{ia}.$$

This implies that there is a stochastic map

$$S : \Delta_m \longrightarrow \Delta_n$$

such that

$$S \circ E = \text{id}_{\Delta_n}.$$

For the Fisher metric,

$$E^* g_F^{(m)} = g_F^{(n)}.$$

Indeed,

$$\sum_{i,a} \frac{d(\alpha_{ia} p_i)^2}{\alpha_{ia} p_i} = \sum_{i,a} \alpha_{ia} \frac{dp_i^2}{p_i} = \sum_i \frac{dp_i^2}{p_i}.$$

Thus stochastic embeddings are isometric for the Fisher metric.

8 Sufficient Statistics

Suppose

$$\mathcal{P} = \{p_\theta : \theta \in \Theta\} \subset \Delta_n$$

is a *statistical family*.

A statistic

$$S : X = \{1, \dots, n\} \longrightarrow Y = \{1, \dots, m\}$$

is called *sufficient* for $\mathcal{P}$ if the value of $S(X)$ contains all the information about θ that was present in X .

One standard definition is that the conditional distribution

$$\mathbb{P}_\theta(X = i \mid S(X) = y)$$

is independent of θ .

Equivalently, there exists a stochastic recovery map

$$R : \Delta_m \longrightarrow \Delta_n$$

such that

$$R(S_* p_\theta) = p_\theta$$

for every θ .

Thus

$$R \circ S_* = \text{id}$$

on the statistical family.

This, as presented by chat GPT, is most confusing, since one starts with deterministic maps $X \rightarrow Y$ and then switches to stochastic ones, i.e. linear, $\Delta_m \rightarrow \Delta_n$.

For deterministic $X \rightarrow Y$ sufficient statistics are those injective on $\mathcal{P}$. Nothing interesting.

The right setting is to make the above definition for "stochastic maps" $X \rightarrow Y$ understood as linear maps between the corresponding simplices, where "sufficiently" (stochastic invertibility) is more meaningful.

9 Why Sufficiency Gives Equality

For every stochastic map T , monotonicity gives

$$g_{T(p)}(dT(v), dT(v)) \leq g_p(v, v).$$

Suppose T is sufficient and admits a stochastic recovery map R such that

$$R \circ T = \text{id}$$

on the statistical model.

Then

$$g_{T(p)}(dT(v), dT(v)) \leq g_p(v, v).$$

Applying monotonicity to R ,

$$g_p(v, v) = g_{R(T(p))}(dR(dT(v)), dR(dT(v))) \leq g_{T(p)}(dT(v), dT(v)).$$

Therefore

$$g_{T(p)}(dT(v), dT(v)) = g_p(v, v).$$

So a sufficient statistic is an isometry on the statistical model.

10 Čencov's Theorem

Consider, for every n , a Riemannian metric

$$g^{(n)}$$

on the probability simplex Δ_n .

Assume these metrics are compatible with stochastic changes of sample space in the sense that stochastic embeddings are isometries and sufficient statistics preserve the metric.

Then there exists a constant

$$c > 0$$

independent of n such that

$$g^{(n)} = c g_F^{(n)}.$$

That is,

$$\boxed{g_p(v, w) = c \sum_i \frac{v_i w_i}{p_i}}.$$

Thus, up to multiplication by one universal positive constant, the Fisher metric is the unique Riemannian metric on finite probability spaces compatible with statistically information-preserving transformations.

Exercise. Recall the above ★ and prove Čencov theorem.

11 A Concrete Interpretation

The theorem can be understood through two elementary operations.

Splitting

One may replace an outcome of probability p_i by several sub-outcomes:

$$p_i \longmapsto \alpha_1 p_i, \dots, \alpha_r p_i, \quad \sum_a \alpha_a = 1.$$

This is a stochastic embedding.

Grouping

One may forget which sub-outcome occurred and replace the refined probabilities by their sum:

$$(\alpha_1 p_i, \dots, \alpha_r p_i) \longmapsto p_i.$$

This is a statistic. When it exactly reverses the splitting, it is sufficient.

Thus Čencov's theorem says that invariance under all such information-preserving refinements and reductions forces

$$g_F = \sum_i \frac{dp_i^2}{p_i}$$

up to an overall positive constant.

12 Kullback–Leibler Divergence

For

$$p, q \in \Delta_n,$$

the Kullback–Leibler divergence is

$$D_{\text{KL}}(p||q) = \sum_i p_i \log \frac{p_i}{q_i}.$$

It is generally not symmetric:

$$D_{\text{KL}}(p||q) \neq D_{\text{KL}}(q||p).$$

Let

$$q = p + dp, \quad \sum_i dp_i = 0.$$

Then

$$D_{\text{KL}}(p||p + dp) = \frac{1}{2} \sum_i \frac{dp_i^2}{p_i} + O(|dp|^3).$$

Hence

$$D_{\text{KL}}(p||p + dp) = \frac{1}{2} g_F(dp, dp) + O(|dp|^3).$$

Thus the second-order part of KL divergence is precisely the Fisher metric.

13 Half-Densities

Let X be a manifold and let

$$\rho$$

be a positive probability density on X , normalized by

$$\int_X \rho = 1.$$

Define the positive half-density

$$\psi = \sqrt{\rho}.$$

Then

$$\int_X \psi^2 = 1.$$

Since

$$\rho = \psi^2, \quad d\rho = 2\psi d\psi,$$

we obtain

$$\int_X \frac{(d\rho)^2}{\rho} = 4 \int_X (d\psi)^2.$$

Hence the Fisher metric is, up to the factor 4, the Hilbert-space metric on the unit sphere of positive half-densities.

14 Summary

The basic objects are

probability distribution $p \in \Delta_n$,

stochastic map $p \mapsto pK$,

statistic deterministic or randomized processing of observations,

stochastic embedding information-preserving refinement,

sufficient statistic information-preserving reduction.

The Fisher metric

$$g_F = \sum_i \frac{dp_i^2}{p_i}$$

is monotone under arbitrary stochastic maps and invariant under information-preserving ones.

Čencov's theorem states that these invariance properties determine the Fisher metric uniquely, up to multiplication by a positive constant.

Under the square-root transformation

$$p_i \mapsto \sqrt{p_i},$$

the Fisher metric becomes the ordinary spherical metric up to scale.

References

1. N. N. Čencov, *Statistical Decision Rules and Optimal Inference*.
2. S. Amari and H. Nagaoka, *Methods of Information Geometry*.
3. S. Amari, *Information Geometry and Its Applications*.