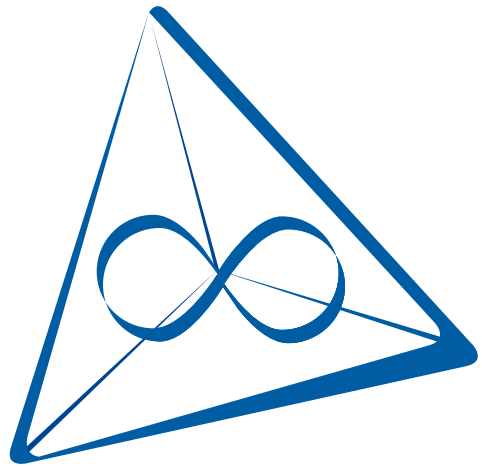
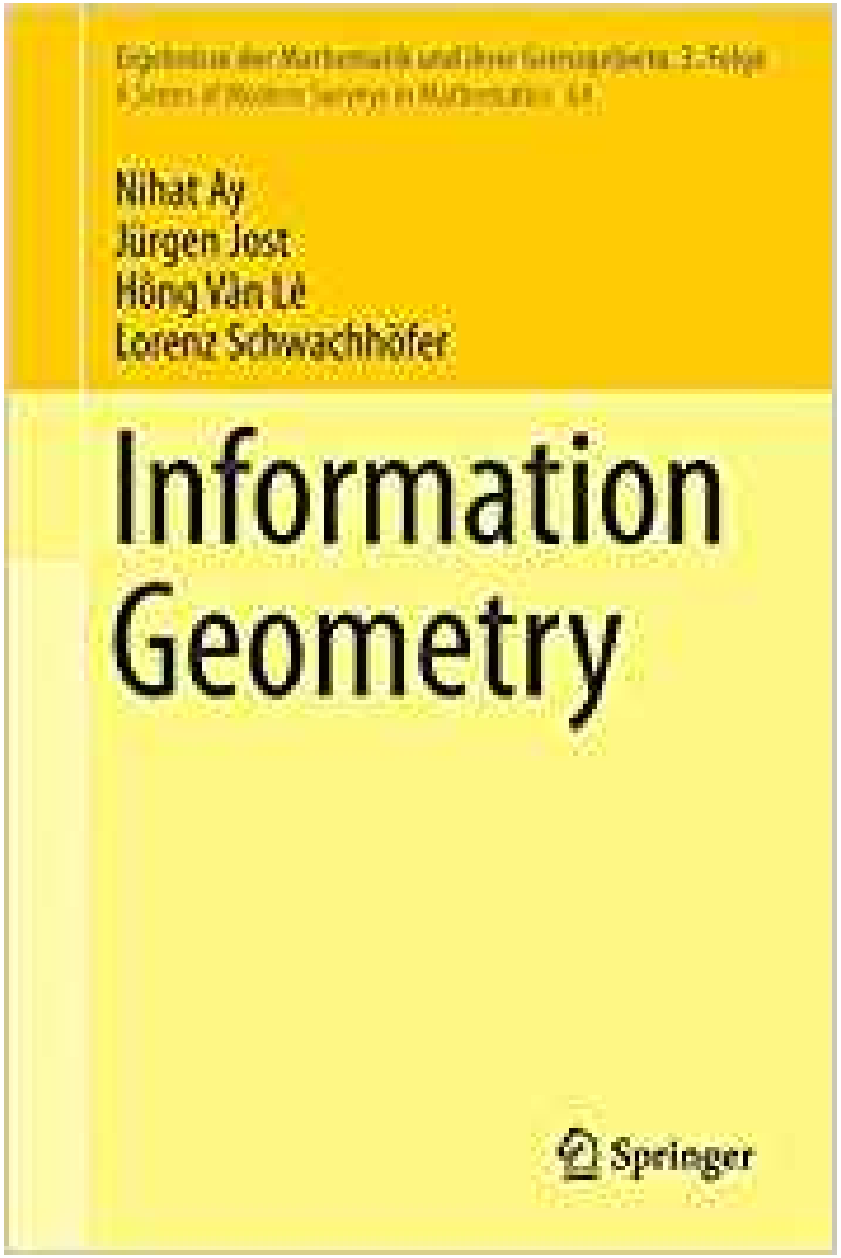
Information Geometry

Jürgen Jost
Max Planck Institute
for
Mathematics in the Sciences
Leipzig




Nihat Ay
Jürgen Jost
Hông Vân Lê
Lorenz Schwachhöfer
Information Geometry
Springer

1. The geometry and functional analysis of families of probability distributions

Introduction

$\mathcal{P}(\Omega)$ is the space of probability distributions on some sample space Ω.

Introduction

$\mathcal{P}(\Omega)$ is the space of probability distributions on some sample space Ω.
A statistical problem consists in finding, approximating, or estimating an unknown probability distribution on Ω, that is, an element of $\mathcal{P}(\Omega)$, according to which samples are drawn from Ω.

Introduction

$\mathcal{P}(\Omega)$ is the space of probability distributions on some sample space Ω.
A statistical problem consists in finding, approximating, or estimating an unknown probability distribution on Ω, that is, an element of $\mathcal{P}(\Omega)$, according to which samples are drawn from Ω.
We usually go about this with some particular assumptions or prior knowledge. In other words, we have some model.

Such a model is a parametrized family of probability distributions on the sample space. The parameter parametrizes the unknown probability measures that govern the distribution of the observable.

Such a model is a parametrized family of probability distributions on the sample space. The parameter parametrizes the unknown probability measures that govern the distribution of the observable. In parametric statistics, the Fisher-Rao metric quantifies how well the parameter can be recovered by observations on the sample space. On the basis of the observations, one then estimates the parameter that best fits the data, for example by using the maximum likelihood estimator.

Such a model is a parametrized family of probability distributions on the sample space. The parameter parametrizes the unknown probability measures that govern the distribution of the observable.
In parametric statistics, the Fisher-Rao metric quantifies how well the parameter can be recovered by observations on the sample space. On the basis of the observations, one then estimates the parameter that best fits the data, for example by using the maximum likelihood estimator.
In Bayesian statistics, one rather starts with a prior distribution over the unknown parameters of the model, which is meant to capture our beliefs about the situation before seeing the data. After observing data, Bayes' formula yields the posterior distribution for these unknowns. This posterior can then be used as a new prior before seeing the next data.

Such a model is a parametrized family of probability distributions on the sample space. The parameter parametrizes the unknown probability measures that govern the distribution of the observable.
In parametric statistics, the Fisher-Rao metric quantifies how well the parameter can be recovered by observations on the sample space. On the basis of the observations, one then estimates the parameter that best fits the data, for example by using the maximum likelihood estimator.
In Bayesian statistics, one rather starts with a prior distribution over the unknown parameters of the model, which is meant to capture our beliefs about the situation before seeing the data. After observing data, Bayes' formula yields the posterior distribution for these unknowns. This posterior can then be used as a new prior before seeing the next data.
Thus, in both parametric and Bayesian statistics, one needs some structure on the space of probability distributions. And for a general and abstract mathematical treatment, this structure should be studied in the most general terms.

Thus, for Bayesian statistics, one should investigate the probability measures on $\mathcal{P}(\Omega)$ (L.H.Duc, H.V.Lê, T.D.Tran, J.J.).
For parametric statistics, one needs a metric on $\mathcal{P}(\Omega)$ that has natural invariance properties. This is the Fisher-Rao metric.

Some notation

Ω a set with a σ-algebra $\mathfrak{B}$.
No further structure on Ω is assumed (like a differentiable or metric structure).

Some notation

Ω a set with a σ-algebra $\mathfrak{B}$.
For a signed measure μ on Ω, we have the total variation of a bounded signed measure

$$\|\mu\|_{TV} := \sup \sum_{i=1}^{n} |\mu(A_i)|, \tag{1}$$

where the supremum is taken over all finite partitions $\Omega = A_1 \dot{\cup} \ldots \dot{\cup} A_n$ with disjoint sets $A_i \in \mathfrak{B}$. If $\|\mu\|_{TV} < \infty$, the signed measure μ is called finite.

Some notation

Ω a set with a σ-algebra $\mathfrak{B}$.
For a signed measure μ on Ω, we have the total variation of a bounded signed measure

$$\|\mu\|_{TV} := \sup \sum_{i=1}^{n} |\mu(A_i)|, \tag{1}$$

where the supremum is taken over all finite partitions $\Omega = A_1 \dot{\cup} \ldots \dot{\cup} A_n$ with disjoint sets $A_i \in \mathfrak{B}$. If $\|\mu\|_{TV} < \infty$, the signed measure μ is called finite.
$\mathcal{S}(\Omega)$=Banach space of signed finite measures on Ω with the total variation norm
$\mathcal{M}(\Omega)$ = finite *non-negative* measures
$\mathcal{P}(\Omega)$ = probability measures on Ω.

$$\mathcal{P}(\Omega) \subset \mathcal{M}(\Omega) \subset \mathcal{S}(\Omega).$$

Some notation

μ_0 a σ-finite non-negative measure,

$$\mathcal{S}(\Omega, \mu_0) \quad := \quad \{\mu = \phi\, \mu_0 \ : \ \phi \in L^1(\Omega, \mu_0)\}$$

space of signed measures dominated by μ_0. Canonical map

$$i_{can} : \mathcal{S}(\Omega, \mu_0) \to L^1(\Omega, \mu_0), \quad \mu \mapsto \frac{d\mu}{d\mu_0},$$

the Radon–Nikodym derivative of μ w.r.t. μ_0.

$$L^1\text{-topology} \quad \|\mu\|_{TV} = \|\phi\|_{L^1(\Omega,\mu_0)} = \left\|\frac{d\mu}{d\mu_0}\right\|_{L^1(\Omega,\mu_0)}.$$

Some notation

μ_0 a σ-finite non-negative measure,

$$\mathcal{S}(\Omega, \mu_0) \quad := \quad \{\mu = \phi\, \mu_0 \ : \ \phi \in L^1(\Omega, \mu_0)\}$$

space of signed measures dominated by μ_0. Canonical map

$$i_{can} : \mathcal{S}(\Omega, \mu_0) \to L^1(\Omega, \mu_0), \quad \mu \mapsto \frac{d\mu}{d\mu_0},$$

the Radon–Nikodym derivative of μ w.r.t. μ_0.

$$L^1\text{-topology} \quad \|\mu\|_{TV} = \|\phi\|_{L^1(\Omega,\mu_0)} = \left\|\frac{d\mu}{d\mu_0}\right\|_{L^1(\Omega,\mu_0)}.$$

Compatible finite non-negative measures μ_1, μ_2 are absolutely continuous with respect to each other, i.e., for some nonnegative ϕ

$$\mu_2 = \phi\mu_1, \quad \text{or, equivalently,} \quad \mu_1 = \phi^{-1}\mu_2. \tag{2}$$

ϕ is the Radon–Nikodym derivative of μ_2 w.r.t. μ_1.

Some notation

μ_0 a σ-finite non-negative measure,

$$\mathcal{S}(\Omega, \mu_0) \quad := \quad \{\mu = \phi\, \mu_0 \ : \ \phi \in L^1(\Omega, \mu_0)\}$$

space of signed measures dominated by μ_0. Canonical map

$$i_{can} : \mathcal{S}(\Omega, \mu_0) \to L^1(\Omega, \mu_0), \quad \mu \mapsto \frac{d\mu}{d\mu_0},$$

the Radon–Nikodym derivative of μ w.r.t. μ_0.

$$L^1\text{-topology} \quad \|\mu\|_{TV} \ = \|\phi\|_{L^1(\Omega,\mu_0)} \ = \ \left\|\frac{d\mu}{d\mu_0}\right\|_{L^1(\Omega,\mu_0)}.$$

Compatible finite non-negative measures μ_1, μ_2 are absolutely continuous with respect to each other, i.e., for some nonnegative ϕ

$$\mu_2 = \phi\mu_1, \quad \text{or, equivalently,} \quad \mu_1 = \phi^{-1}\mu_2. \tag{2}$$

ϕ is the Radon–Nikodym derivative of μ_2 w.r.t. μ_1.

$$\mathcal{M}_+(\Omega, \mu) := \{\phi\mu : \phi \in L^1(\Omega, \mu),\ \phi > 0 \quad \mu\text{-a.e.}\} \tag{3}$$

Compatible finite non-negative measures μ_1, μ_2 are absolutely continuous with respect to each other, i.e., for some nonnegative ϕ

$$\mu_2 = \phi\mu_1, \quad \text{or, equivalently,} \quad \mu_1 = \phi^{-1}\mu_2. \tag{4}$$

Compatibility is an equivalence relation on the space of finite non-negative measures on Ω, and that space is therefore partitioned (stratified) into equivalence classes. The set of such equivalence classes is quite large. For instance, the Dirac measure at any point of Ω generates its own such class, i.e., two different Diracs are singular w.r.t. each other. More generally, in Euclidean space, we can consider Hausdorff measures of subsets of possibly different Hausdorff dimensions.

Compatible finite non-negative measures μ_1, μ_2 are absolutely continuous with respect to each other, i.e., for some nonnegative ϕ

$$\mu_2 = \phi\mu_1, \quad \text{or, equivalently,} \quad \mu_1 = \phi^{-1}\mu_2. \tag{4}$$

Compatibility is an equivalence relation on the space of finite non-negative measures on Ω, and that space is therefore partitioned (stratified) into equivalence classes. The set of such equivalence classes is quite large. For instance, the Dirac measure at any point of Ω generates its own such class, i.e., two different Diracs are singular w.r.t. each other. More generally, in Euclidean space, we can consider Hausdorff measures of subsets of possibly different Hausdorff dimensions.
If (Ω, d) is a metric space, we can consider the Wasserstein distance between measures. Two Dirac measures $\delta(x), \delta(y)$, instead of being singular w.r.t. each other, then have Wasserstein distance $= d(x, y)$.

Parametric models

$$\mathbf{p} : M \to \mathcal{P}(\Omega),$$

a parametric family of probability measures: For each ξ in the parameter space M, we have a probability measure $p(\cdot; \xi)$ on Ω.

Parametric models

$$\mathbf{p} : M \to \mathcal{P}(\Omega),$$

where the parameter $\xi \in M$ should be estimated based on random samples drawn from some unknown probability distribution on Ω, so as to identify a particular $p(\cdot; \xi_0)$ that best fits that sampling distribution.

Parametric models

$$\mathbf{p} : M \to \mathcal{P}(\Omega),$$

a parametric family of probability measures: For each ξ in the parameter space M, we have a probability measure $p(\cdot; \xi)$ on Ω.

Definition

A *statistical model* is a triple $(M, \Omega, \mathbf{p})$ where M is a (finite or infinite-dimensional) Banach manifold and
$\mathbf{p} : M \to \mathcal{P}(\Omega) \subset \mathcal{M}(\Omega) \subset \mathcal{S}(\Omega)$ is a C^1-map.

$$\mathbf{p} : M \to \mathcal{P}(\Omega). \tag{5}$$

How sensitively does $p(x;\xi)$ depend on the parameter ξ? That is, how well can we distinguish between different values of ξ by observing samples x? How well can we estimate ξ?

$$\mathbf{p} : M \to \mathcal{P}(\Omega). \tag{5}$$

How sensitively does $p(x;\xi)$ depend on the parameter ξ?
This sensitivity can be quantified by a Riemannian metric, the Fisher-Rao metric.

$$\mathbf{p} : M \to \mathcal{P}(\Omega). \tag{5}$$

How sensitively does $p(x;\xi)$ depend on the parameter ξ?
This sensitivity can be quantified by a Riemannian metric, the Fisher-Rao metric.
On the parameter space M, it is obtained by pulling back some universal structure from $\mathcal{P}(\Omega)$ via (5).

$$\mathbf{p} : M \to \mathcal{P}(\Omega). \tag{5}$$

How sensitively does $p(x;\xi)$ depend on the parameter ξ?
This sensitivity can be quantified by a Riemannian metric, the Fisher-Rao metric.
On the parameter space M, it is obtained by pulling back some universal structure from $\mathcal{P}(\Omega)$ via (5).
When Ω is infinite, which is not an untypical situation in statistics, however, $\mathcal{P}(\Omega)$ is infinite-dimensional, and therefore functional analytical problems arise.

Invariances

How to determine the metric on $\mathcal{P}(\Omega)$?

Invariances

How to determine the metric on $\mathcal{P}(\Omega)$?

Look at invariances.

Invariances

How to determine the metric on $\mathcal{P}(\Omega)$?

Look at invariances. Consider *statistics*, i.e., mappings

$$\kappa : \Omega \to \Omega' \tag{6}$$

into some other space Ω'.

Invariances

How to determine the metric on $\mathcal{P}(\Omega)$?

Look at invariances. Consider *statistics*, i.e., mappings

$$\kappa : \Omega \to \Omega' \tag{6}$$

into some other space Ω'.

Ω' might be finite; for instance, Ω' could be the index set of some finite partition of Ω, and $\kappa(x)$ records in which member of that partition x is found. In other cases, κ might stand for a specific observable on Ω.

Invariances

How to determine the metric on $\mathcal{P}(\Omega)$?

Look at invariances. Consider *statistics*, i.e., mappings

$$\kappa : \Omega \to \Omega' \tag{6}$$

into some other space Ω'.

What is the possible loss of information about the parameter $\xi \in M$ from the family (5) when we only observe $\kappa(x)$ instead of x itself? The statistic κ is called *sufficient* for the family (5) when no information is lost at all.

Invariances

How to determine the metric on $\mathcal{P}(\Omega)$?

Look at invariances. Consider *statistics*, i.e., mappings

$$\kappa : \Omega \to \Omega' \tag{6}$$

into some other space Ω'.

What is the possible loss of information about the parameter $\xi \in M$ from the family (5) when we only observe $\kappa(x)$ instead of x itself? The statistic κ is called *sufficient* for the family (5) when no information is lost at all.

The information loss quantified by the difference of the Fisher metrics of the originally family $\mathbf{p}$ and the induced family $\kappa_* \mathbf{p}$.

Invariances

How to determine the metric on $\mathcal{P}(\Omega)$?

Look at invariances. Consider *statistics*, i.e., mappings

$$\kappa : \Omega \to \Omega' \tag{6}$$

into some other space Ω'.

What is the possible loss of information about the parameter $\xi \in M$ from the family (5) when we only observe $\kappa(x)$ instead of x itself? The statistic κ is called *sufficient* for the family (5) when no information is lost at all.

The information loss quantified by the difference of the Fisher metrics of the originally family $\mathbf{p}$ and the induced family $\kappa_* \mathbf{p}$.

Theorem

The Fisher metric is uniquely characterized (up to a constant factor) by invariance under sufficient statistics.

Invariances

How to determine the metric on $\mathcal{P}(\Omega)$?
Look at invariances. Consider *statistics*, i.e., mappings

$$\kappa : \Omega \to \Omega' \tag{6}$$

into some other space Ω'.
What is the possible loss of information about the parameter $\xi \in M$ from the family (5) when we only observe $\kappa(x)$ instead of x itself? The statistic κ is called *sufficient* for the family (5) when no information is lost at all.
The information loss quantified by the difference of the Fisher metrics of the originally family $\mathbf{p}$ and the induced family $\kappa_* \mathbf{p}$.

Theorem

The Fisher metric is uniquely characterized (up to a constant factor) by invariance under sufficient statistics.

In the finite case proved already by Chentsov.

Invariances

How to determine the metric on $\mathcal{P}(\Omega)$?
Look at invariances. Consider *statistics*, i.e., mappings

$$\kappa : \Omega \to \Omega' \tag{6}$$

into some other space Ω'.
What is the possible loss of information about the parameter $\xi \in M$ from the family (5) when we only observe $\kappa(x)$ instead of x itself? The statistic κ is called *sufficient* for the family (5) when no information is lost at all.
The information loss quantified by the difference of the Fisher metrics of the originally family $\mathbf{p}$ and the induced family $\kappa_* \mathbf{p}$.

Theorem

The Fisher metric is uniquely characterized (up to a constant factor) by invariance under sufficient statistics.

Remark: When Ω is a differentiable manifold, the Fisher metric is already uniquely determined by invariance under diffeomorphisms of Ω (Bauer-Bruveris-Michor).

When we reparametrize the parameter space M, the Fisher metric transforms appropriately. However, there are particular families $\mathbf{p}$ with particular parametrizations that play an important role. For that, we need to look at the structure of the space $\mathcal{P}(\Omega)$ of probability measures more carefully. Every probability measure is a measure, i.e., there is an embedding

$$\imath : \mathcal{P}(\Omega) \to \mathcal{S}(\Omega) \tag{7}$$

into the space $\mathcal{S}(\Omega)$ of all finite signed measures on Ω, which is a linear space. $p \in \mathcal{P}(\Omega)$ is characterized by $\int_\Omega dp(x) = 1$, and so, $\mathcal{P}(\Omega)$ becomes a convex subset (because of the nonnegativity constraint) of an affine subspace (characterized by the condition $\int_\Omega d\mu(x) = 1$) of the linear space $\mathcal{S}(\Omega)$.

On the other hand, there is also a projection

$$\pi : \mathcal{M}(\Omega) \to \mathcal{P}(\Omega) \tag{8}$$

of the space of nonnegative measures by assigning to each $m \in \mathcal{M}(\Omega)$ the relative measure of subsets. For any measurable subsets $A, B \subset \Omega$ with $m(B) > 0$, $\pi(m)$ looks at the quotients $\frac{m(A)}{m(B)}$, that is, the relative measures of those subsets. That is, a probability measure is now considered as an equivalence class of measures up to a scaling factor.

On the other hand, there is also a projection

$$\pi : \mathcal{M}(\Omega) \to \mathcal{P}(\Omega) \tag{8}$$

of the space of nonnegative measures by assigning to each $m \in \mathcal{M}(\Omega)$ the relative measure of subsets.
$\mathcal{P}(\Omega)$ can be identified with $\pi(\mathcal{M}(\Omega))$, by simply normalizing a measure by $m(\Omega)$.

On the other hand, there is also a projection

$$\pi : \mathcal{M}(\Omega) \to \mathcal{P}(\Omega) \tag{8}$$

of the space of nonnegative measures by assigning to each $m \in \mathcal{M}(\Omega)$ the relative measure of subsets.
$\mathcal{P}(\Omega)$ can be identified with $\pi(\mathcal{M}(\Omega))$, by simply normalizing a measure by $m(\Omega)$.
Thus, $\mathcal{P}(\Omega)$, can be seen as the positive part of a projective space of the linear space $\mathcal{S}(\Omega)$, i.e., as the positive orthant or sector of the unit sphere in $\mathcal{S}(\Omega)$.

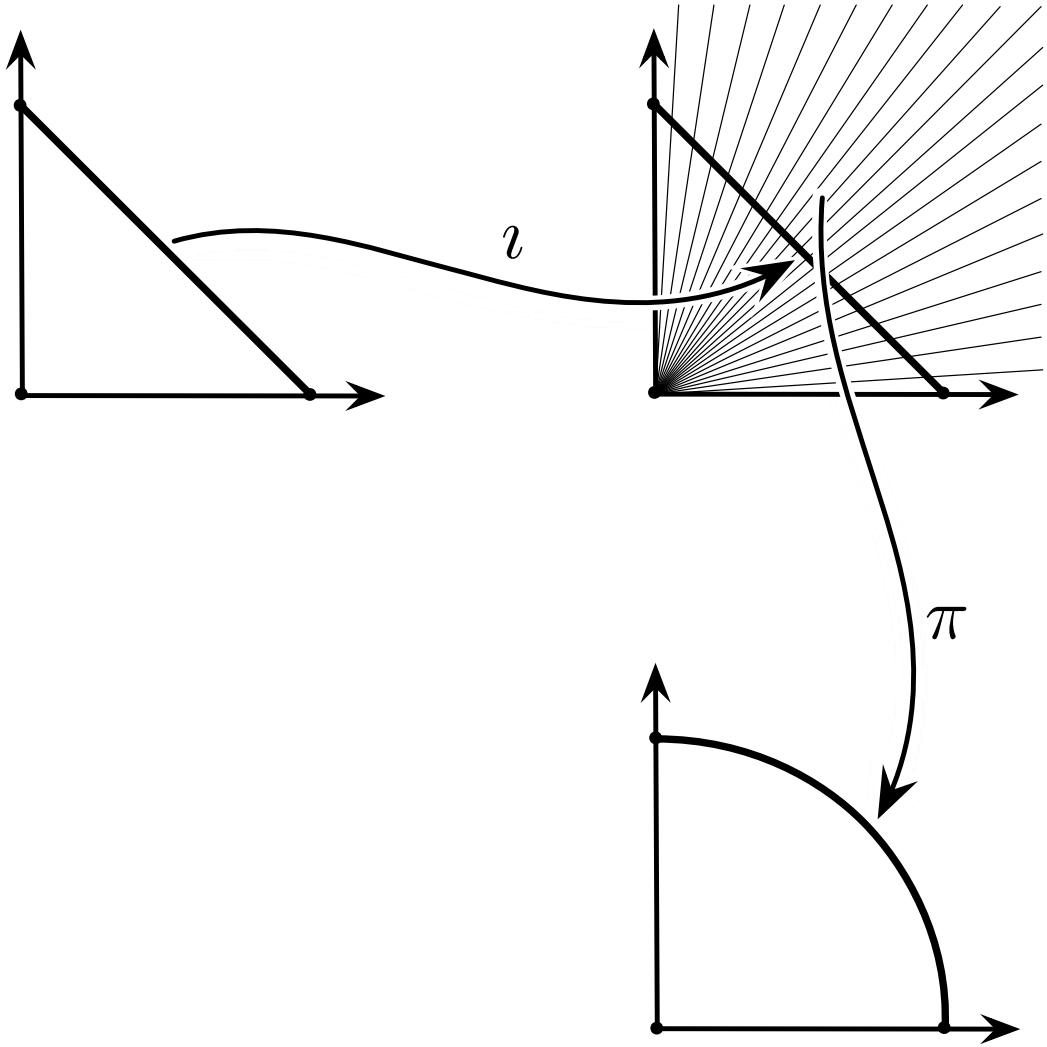


Figure: Natural inclusion and projection.

The Fisher metric

When Ω is finite, the linear space $\mathcal{S}(\Omega)$ is finite-dimensional, and therefore, it can be naturally equipped with a Euclidean metric. This metric then also induces a metric on the unit sphere. Thus, the projection map π from (8) then induces a metric on $\mathcal{P}(\Omega)$. This is the Fisher metric.

The Fisher metric

When Ω is finite, the linear space $\mathcal{S}(\Omega)$ is finite-dimensional, and therefore, it can be naturally equipped with a Euclidean metric. This metric then also induces a metric on the unit sphere. Thus, the projection map π from (8) then induces a metric on $\mathcal{P}(\Omega)$. This is the Fisher metric.

When Ω is infinite, then the space $\mathcal{S}(\Omega)$ is infinite-dimensional, but it does not carry the structure of a Hilbert space. This is a central problem for which the appropriate functional analytic setting has been described above and to which we shall now return.

Families of measures $\mathbf{p}(\xi)$ on Ω parametrized by $\xi \in M$.

Families of measures $\mathbf{p}(\xi)$ on Ω parametrized by $\xi \in M$.
For different ξ, the resulting measures might be quite different and have different null sets. Nevertheless, in many cases, for instance if M is a finite-dimensional manifold, we may write such a family as

$$\mathbf{p}(\xi) = p(\cdot;\xi)\mu_0, \tag{9}$$

for some base measure μ_0 that does not depend on ξ.
$p : \Omega \times M \to \mathbb{R}$ is the density function of $\mathbf{p}$ w.r.t. μ_0, and we then need that $p(\cdot;\xi) \in L^1(\Omega, \mu_0)$ for all ξ.

Families of measures $\mathbf{p}(\xi)$ on Ω parametrized by $\xi \in M$.

$$\mathbf{p}(\xi) = p(\cdot;\xi)\mu_0, \tag{9}$$

for some base measure μ_0 that does not depend on ξ.
$p(\cdot;\xi) \in L^1(\Omega, \mu_0)$ for all ξ.
μ_0 is an auxiliary object, and the construction should not depend on it.

Families of measures $\mathbf{p}(\xi)$ on Ω parametrized by $\xi \in M$.

$$\mathbf{p}(\xi) = p(\cdot;\xi)\mu_0, \tag{9}$$

for some base measure μ_0 that does not depend on ξ.
$p(\cdot;\xi) \in L^1(\Omega, \mu_0)$ for all ξ.
μ_0 is an auxiliary object, and the construction should not depend on it. When we have another probability measure μ_1 with $\mu_1 = \phi\mu_0$ for some positive function ϕ with $\phi \in L^1(\Omega, \mu_0)$ and hence $\phi^{-1} \in L^1(\Omega, \mu_1)$, then $\psi \in L^1(\Omega, \mu_1)$ precisely if $\psi\phi \in L^1(\Omega, \mu_0)$. Thus, the L^1-spaces naturally correspond to each other, and it does not matter which base measure we choose, as long as the different base measures are related by L^1-functions.

The differential of $\mathbf{p}$ in some direction V is

$$d_\xi \mathbf{p}(V) = \partial_V p(\cdot;\xi)\mu_0 \in L^1(\Omega, \mu_0), \tag{10}$$

when this quantity exists.

The differential of $\mathbf{p}$ in some direction V is

$$d_\xi \mathbf{p}(V) = \partial_V p(\cdot;\xi)\mu_0 \in L^1(\Omega, \mu_0), \tag{10}$$

but instead, we should consider the rate of change of $\mathbf{p}(\xi)$ relative to the measure $\mathbf{p}(\xi)$ itself, i.e., the Radon–Nikodym derivative of $d_\xi \mathbf{p}(V)$ w.r.t. $\mathbf{p}(\xi)$, i.e., the *logarithmic derivative*

$$\partial_V \log p(\cdot;\xi) = \frac{d\{d_\xi \mathbf{p}(V)\}}{d\mathbf{p}(\xi)}. \tag{11}$$

The differential of $\mathbf{p}$ in some direction V is

$$d_\xi \mathbf{p}(V) = \partial_V p(\cdot;\xi)\mu_0 \in L^1(\Omega, \mu_0), \tag{10}$$

but instead, we should consider the rate of change of $\mathbf{p}(\xi)$ relative to the measure $\mathbf{p}(\xi)$ itself, i.e., the Radon–Nikodym derivative of $d_\xi \mathbf{p}(V)$ w.r.t. $\mathbf{p}(\xi)$, i.e., the *logarithmic derivative*

$$\partial_V \log p(\cdot;\xi) = \frac{d\{d_\xi \mathbf{p}(V)\}}{d\mathbf{p}(\xi)}. \tag{11}$$

This yields the Fisher metric

$$\mathfrak{g}_\xi(V, W) = \int_\Omega \partial_V \log p(\cdot;\xi)\ \partial_W \log p(\cdot;\xi)\ d\mathbf{p}(\xi). \tag{12}$$

Fisher metric

$$\mathfrak{g}_\xi(V, W) = \int_\Omega \partial_V \log p(\cdot; \xi)\, \partial_W \log p(\cdot; \xi)\, d\mathbf{p}(\xi). \tag{13}$$

What if the density p is not positive almost everywhere?

Fisher metric

$$\mathfrak{g}_\xi(V,W) = \int_\Omega \partial_V \log p(\cdot;\xi)\ \partial_W \log p(\cdot;\xi)\ d\mathbf{p}(\xi). \tag{13}$$

What if the density p is not positive almost everywhere? Introduce the formal square roots

$$\sqrt{\mathbf{p}(\xi)} := \sqrt{p(\cdot;\xi)}\sqrt{\mu_0}, \tag{14}$$

and use the formal computation

$$d_\xi\sqrt{\mathbf{p}}(V) = \frac{1}{2}\partial_V \log p(\cdot;\xi)\sqrt{\mathbf{p}(\xi)} \tag{15}$$

to rewrite (13) as

$$\mathfrak{g}_\xi(V,W) = 4\int_\Omega d\big(d_\xi\sqrt{\mathbf{p}}(V)\cdot d_\xi\sqrt{\mathbf{p}}(W)\big). \tag{16}$$

Fisher metric

$$\mathfrak{g}_\xi(V,W) = \int_\Omega \partial_V \log p(\cdot;\xi)\, \partial_W \log p(\cdot;\xi)\, d\mathbf{p}(\xi). \tag{13}$$

What if the density p is not positive almost everywhere? Introduce the formal square roots

$$\sqrt{\mathbf{p}(\xi)} := \sqrt{p(\cdot;\xi)}\sqrt{\mu_0}, \tag{14}$$

and use the formal computation

$$d_\xi \sqrt{\mathbf{p}}(V) = \frac{1}{2}\partial_V \log p(\cdot;\xi)\sqrt{\mathbf{p}(\xi)} \tag{15}$$

to rewrite (13) as

$$\mathfrak{g}_\xi(V,W) = 4\int_\Omega d(d_\xi\sqrt{\mathbf{p}}(V)\cdot d_\xi\sqrt{\mathbf{p}}(W)). \tag{16}$$

An L^1-condition on $\mathbf{p}(\xi)$ becomes an L^2-condition on $\sqrt{\mathbf{p}(\xi)}$ in (14), and an L^2-condition is precisely what we need in (16) for the derivatives. According to (15), this means that we should now impose an L^2-condition on $\partial_V \log p(\cdot;\xi)$. Again, all this is naturally compatible with a change of base measure.

Sample space Ω with a σ-algebra and space of (positive) measures on Ω. Probability measures can either be considered as measures μ with $\mu(\Omega) = 1$, or as relative measures, that is, considering only quotients $\frac{\mu(A)}{\mu(B)}$ whenever $\mu(B) > 0$. In the first case, we would deal with an infinite dimensional simplex, in the second one with the positive orthant or sector of an infinite-dimensional sphere.

Sample space Ω with a σ-algebra and space of (positive) measures on Ω. Probability measures can either be considered as measures μ with $\mu(\Omega) = 1$, or as relative measures, that is, considering only quotients $\frac{\mu(A)}{\mu(B)}$ whenever $\mu(B) > 0$.
For a base measure μ_0, the space of compatible measures would be $\mathcal{M}_+(\Omega, \mu_0) = \{\phi\mu_0 : \phi \in L^1(\Omega, \mu_0), \phi > 0 \text{ almost everywhere}\}$.
When $\mu_1 = \phi_1\mu_0 \in \mathcal{M}_+(\Omega, \mu_0)$ and $\mu_2 = \phi_2\mu_1 \in \mathcal{M}_+(\Omega, \mu_1)$, then $\mu_2 = \phi_2\phi_1\mu_0 \in \mathcal{M}_+(\Omega, \mu_0)$.

Sample space Ω with a σ-algebra and space of (positive) measures on Ω. Probability measures can either be considered as measures μ with $\mu(\Omega) = 1$, or as relative measures, that is, considering only quotients $\frac{\mu(A)}{\mu(B)}$ whenever $\mu(B) > 0$.
For a base measure μ_0, the space of compatible measures would be $\mathcal{M}_+(\Omega, \mu_0) = \{\phi\mu_0 : \phi \in L^1(\Omega, \mu_0), \phi > 0 \text{ almost everywhere}\}$.
When $\mu_1 = \phi_1\mu_0 \in \mathcal{M}_+(\Omega, \mu_0)$ and $\mu_2 = \phi_2\mu_1 \in \mathcal{M}_+(\Omega, \mu_1)$, then $\mu_2 = \phi_2\phi_1\mu_0 \in \mathcal{M}_+(\Omega, \mu_0)$.
We do not have a multiplicative structure, because if $\phi, \psi \in L^1(\Omega, \mu_0)$, then $\phi\psi$ need not be in $L^1(\Omega, \mu_0)$ itself. The exponential map $f \mapsto e^f$ (defined pointwise, i.e., $e^f(x) = e^{f(x)}$) is not defined for all f. In fact, the natural linear space would be $L^2(\Omega, \mu_0)$, but if $f \in L^2(\Omega, \mu_0)$, then e^f need not be in $L^1(\Omega, \mu_0)$.

There is a natural duality between functions f and measures $\phi\mu$,

$$(f, \phi\mu) = \int_\Omega f\phi d\mu, \tag{17}$$

whenever f and ϕ satisfy appropriate integrability conditions.

There is a natural duality between functions f and measures $\phi\mu$,

$$(f, \phi\mu) = \int_\Omega f\phi d\mu, \tag{17}$$

whenever f and ϕ satisfy appropriate integrability conditions. We can turn (17) into a symmetric pairing by rewriting it as

$$\langle f(\mu)^{1/2}, \phi(\mu)^{1/2}\rangle = \int_\Omega f(d\mu)^{1/2}\phi(d\mu)^{1/2} \tag{18}$$

and require that both factors be in L^2, transforming like $(d\mu)^{1/2}$, i.e., with the square root of the Jacobian of a coordinate transformation (half-densities).

There is a natural duality between functions f and measures $\phi\mu$,

$$(f, \phi\mu) = \int_\Omega f \phi d\mu, \tag{17}$$

whenever f and ϕ satisfy appropriate integrability conditions. We can turn (17) into a symmetric pairing by rewriting it as

$$\langle f(\mu)^{1/2}, \phi(\mu)^{1/2} \rangle = \int_\Omega f (d\mu)^{1/2} \phi (d\mu)^{1/2} \tag{18}$$

and require that both factors be in L^2, transforming like $(d\mu)^{1/2}$, i.e., with the square root of the Jacobian of a coordinate transformation (half-densities). Below, for the Amari–Chentsov structure, we also need $(1/3)$-densities.

There is a natural duality between functions f and measures $\phi\mu$,

$$(f, \phi\mu) = \int_\Omega f \phi d\mu, \tag{17}$$

whenever f and ϕ satisfy appropriate integrability conditions.
We can turn (17) into a symmetric pairing by rewriting it as

$$\langle f(\mu)^{1/2}, \phi(\mu)^{1/2} \rangle = \int_\Omega f (d\mu)^{1/2} \phi (d\mu)^{1/2} \tag{18}$$

and require that both factors be in L^2, transforming like $(d\mu)^{1/2}$, i.e., with the square root of the Jacobian of a coordinate transformation (half-densities). Below, for the Amari–Chentsov structure, we also need $(1/3)$-densities.
Banach spaces $\mathcal{S}^r(\Omega)$ of *formal r-th powers of (signed) measures* for $0 < r \leq 1$.

There is a natural duality between functions f and measures $\phi\mu$,

$$(f, \phi\mu) = \int_\Omega f\phi d\mu, \tag{17}$$

whenever f and ϕ satisfy appropriate integrability conditions.
We can turn (17) into a symmetric pairing by rewriting it as

$$\langle f(\mu)^{1/2}, \phi(\mu)^{1/2}\rangle = \int_\Omega f(d\mu)^{1/2}\phi(d\mu)^{1/2} \tag{18}$$

and require that both factors be in L^2, transforming like $(d\mu)^{1/2}$, i.e., with the square root of the Jacobian of a coordinate transformation (half-densities). Below, for the Amari–Chentsov structure, we also need $(1/3)$-densities.
Banach spaces $\mathcal{S}^r(\Omega)$ of *formal r-th powers of (signed) measures* for $0 < r \leq 1$. $\mathcal{S}^1(\Omega) = \mathcal{S}(\Omega)$ is the Banach space of finite signed measures on Ω with the total variation norm. $\mathcal{S}^{1/2}(\Omega)$ is the space of *signed half densities*, a Hilbert space.

There is a natural duality between functions f and measures $\phi\mu$,

$$(f, \phi\mu) = \int_\Omega f\phi d\mu, \tag{17}$$

whenever f and ϕ satisfy appropriate integrability conditions.
We can turn (17) into a symmetric pairing by rewriting it as

$$\langle f(\mu)^{1/2}, \phi(\mu)^{1/2}\rangle = \int_\Omega f(d\mu)^{1/2}\phi(d\mu)^{1/2} \tag{18}$$

and require that both factors be in L^2, transforming like $(d\mu)^{1/2}$, i.e., with the square root of the Jacobian of a coordinate transformation (half-densities). Below, for the Amari–Chentsov structure, we also need $(1/3)$-densities.
Banach spaces $\mathcal{S}^r(\Omega)$ of *formal r-th powers of (signed) measures* for $0 < r \leq 1$. $\mathcal{S}^1(\Omega) = \mathcal{S}(\Omega)$ is the Banach space of finite signed measures on Ω with the total variation norm. $\mathcal{S}^{1/2}(\Omega)$ is the space of *signed half densities*, a Hilbert space. Inclusions $\mathcal{P}^r(\Omega) \subset \mathcal{M}^r(\Omega) \subset \mathcal{S}^r(\Omega)$ of r-th powers of probability measures.
Rigorous definition of the (formal) tangent bundle $T\mathcal{P}^r(\Omega)$ and $T\mathcal{M}^r(\Omega)$, where $T_\mu\mathcal{M}^r(\Omega) = L^k(\Omega, \mu)$ for $k = 1/r \geq 1$.

Banach spaces $\mathcal{S}^r(\Omega)$ of *formal r-th powers of (signed) measures* for $0 < r \leq 1$.

Banach spaces $\mathcal{S}^r(\Omega)$ of *formal r-th powers of (signed) measures* for $0 < r \leq 1$. $\mathcal{S}^1(\Omega) = \mathcal{S}(\Omega)$ is the Banach space of finite signed measures on Ω with the total variation norm. $\mathcal{S}^{1/2}(\Omega)$ is the space of *signed half densities*, a Hilbert space. Inclusions $\mathcal{P}^r(\Omega) \subset \mathcal{M}^r(\Omega) \subset \mathcal{S}^r(\Omega)$ of r-th powers of probability measures. Rigorous definition of the (formal) tangent bundle $T\mathcal{P}^r(\Omega)$ and $T\mathcal{M}^r(\Omega)$, where $T_\mu\mathcal{M}^r(\Omega) = L^k(\Omega, \mu)$ for $k = 1/r \geq 1$.

Banach spaces $\mathcal{S}^r(\Omega)$ of *formal r-th powers of (signed) measures* for $0 < r \leq 1$. $\mathcal{S}^1(\Omega) = \mathcal{S}(\Omega)$ is the Banach space of finite signed measures on Ω with the total variation norm. $\mathcal{S}^{1/2}(\Omega)$ is the space of *signed half densities*, a Hilbert space. Inclusions $\mathcal{P}^r(\Omega) \subset \mathcal{M}^r(\Omega) \subset \mathcal{S}^r(\Omega)$ of r-th powers of probability measures. Rigorous definition of the (formal) tangent bundle $T\mathcal{P}^r(\Omega)$ and $T\mathcal{M}^r(\Omega)$, where $T_\mu\mathcal{M}^r(\Omega) = L^k(\Omega, \mu)$ for $k = 1/r \geq 1$.

Definition

A *statistical model* $\mathbf{p} : M \to \mathcal{P}(\Omega) \subset \mathcal{M}(\Omega) \subset \mathcal{S}(\Omega)$ is called *k-integrable* if the map

$$\mathbf{p}^{1/k} : M \longrightarrow \mathcal{M}^{1/k}(\Omega) \subset \mathcal{S}^{1/k}(\Omega)$$

is a C^1-map

Affine structures

The structure induced on $\mathcal{P}(\Omega)$ by the projection (8) is dual to the affine structure induced the embedding (7). This dual structure is affine.

Affine structures

The structure induced on $\mathcal{P}(\Omega)$ by the projection (8) is dual to the affine structure induced the embedding (7). This dual structure is affine.

Two possible ways in which a measure can be normalized to become a probability measure.

1. We want to move a probability measure μ to another probability measure ν, straight line

$$\mu + t(\nu - \mu), \text{ with } t \in [0, 1]. \tag{19}$$

When a variation $\mu + t\xi$ should remain a probability measure, we need to subtract $\xi_0 := \xi(\Omega)$,

$$\mu + t(\xi - \xi_0). \tag{20}$$

but this need no longer be nonnegative.

Affine structures

The structure induced on $\mathcal{P}(\Omega)$ by the projection (8) is dual to the affine structure induced the embedding (7). This dual structure is affine.

Two possible ways in which a measure can be normalized to become a probability measure.

1. We want to move a probability measure μ to another probability measure ν, straight line

$$\mu + t(\nu - \mu), \text{ with } t \in [0, 1]. \tag{19}$$

When a variation $\mu + t\xi$ should remain a probability measure, we need to subtract $\xi_0 := \xi(\Omega)$,

$$\mu + t(\xi - \xi_0). \tag{20}$$

but this need no longer be nonnegative.

The geodesic $\mu + t(\xi - \xi_0)$ w.r.t. the affine structure on the simplex may leave the simplex of probability measures. Thus, this affine structure is not complete.

Two possible ways in which a measure can be normalized to become a probability measure.

2. Multiplicative variation

$$\exp(tf)\mu, \text{ with } \exp f \in L^1(\Omega, \mu), \tag{21}$$

which remains nonnegative.

Two possible ways in which a measure can be normalized to become a probability measure.

2 Multiplicative variation

$$\exp(tf)\mu, \text{ with } \exp f \in L^1(\Omega, \mu), \tag{21}$$

which remains nonnegative.

We can consider a linear space of functions f here.

Two possible ways in which a measure can be normalized to become a probability measure.

2. Multiplicative variation

$$\exp(tf)\mu, \text{ with } \exp f \in L^1(\Omega, \mu), \tag{21}$$

which remains nonnegative.
We can consider a linear space of functions f here.
With normalization

$$\frac{\exp(tf)}{Z(t)}\mu \text{ with } Z(t) := \int_\Omega \exp(tf)d\mu. \tag{22}$$

Two possible ways in which a measure can be normalized to become a probability measure.

2 Multiplicative variation

$$\exp(tf)\mu, \text{ with } \exp f \in L^1(\Omega, \mu), \tag{21}$$

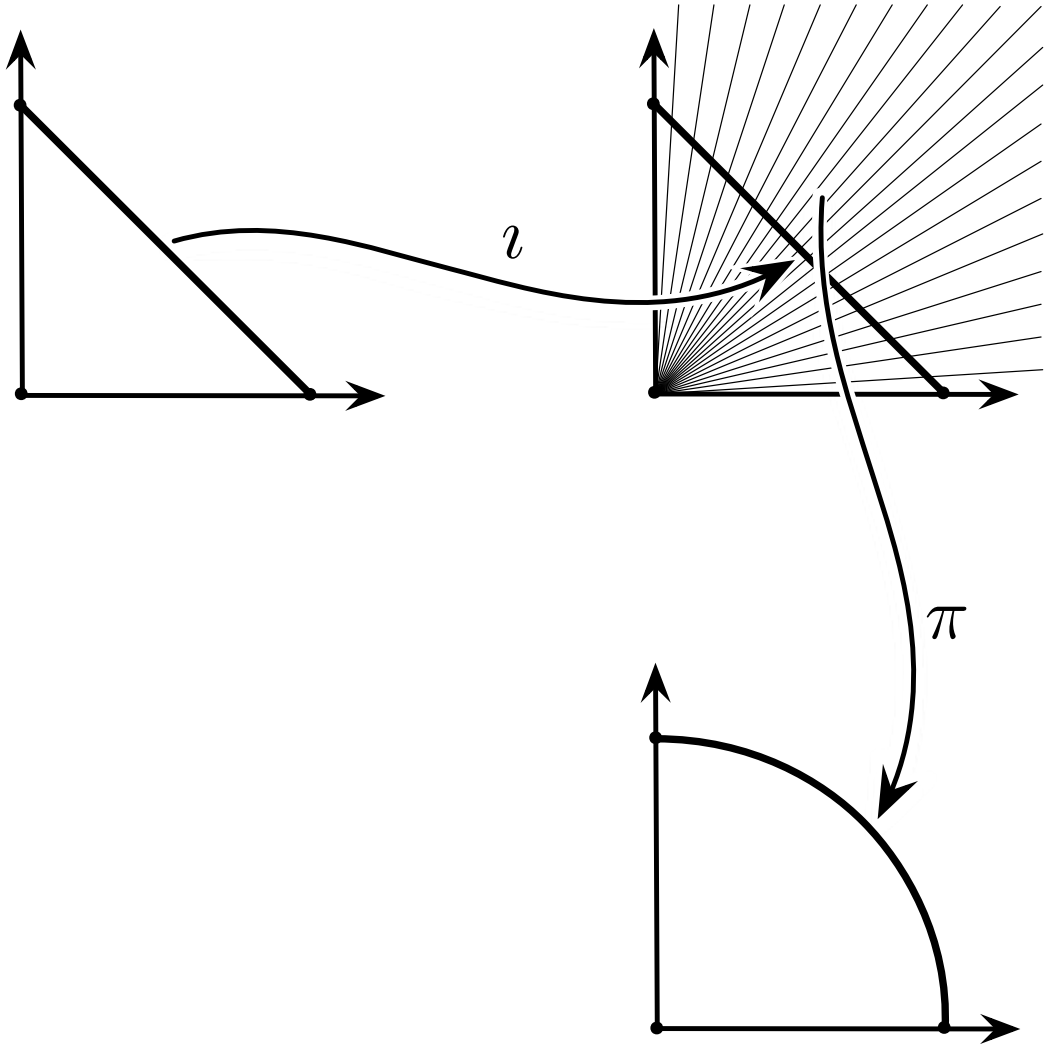
which remains nonnegative.
We can consider a linear space of functions f here.
With normalization

$$\frac{\exp(tf)}{Z(t)}\mu \text{ with } Z(t) := \int_\Omega \exp(tf)d\mu. \tag{22}$$

Family (22) is geodesic for an affine structure. For two probability measures μ, μ_1 with $\mu_1 = \phi\mu$ for some positive ϕ with $\phi \in L^1(\Omega, \mu)$ and hence $\phi^{-1} \in L^1(\Omega, \mu_1)$, then $\exp(tf)\mu$ of μ corresponds to $\frac{\exp(tf)}{\phi}\mu_1$. At the level of the linear spaces, correspondence between f and $f - \log\phi$ which does not depend on the individual f. When $\mu_2 = \psi\mu_1$, then $\mu_2 = \psi\phi\mu$, and the shift is by $\log(\psi\phi) = \log\psi + \log\phi$. But this is precisely what an affine structure amounts to.

Thus, we have identified the second affine structure on the space of probability measures. It possesses a natural exponential map $f \mapsto \exp f$, is naturally adapted to our description of probability measures as equivalence classes of measures, and is complete in contrast to the first affine structure.

The two structures are naturally dual to each other. They are related by a Legendre transform that generalizes the duality between entropy and free energy of statistical mechanics.



This pair of dual affine structures was discovered by Amari and Chentsov, and the tensor describing it is therefore called the Amari–Chentsov tensor. Like the Fisher metric, the Amari–Chentsov tensor is invariant under sufficient statistics, and uniquely characterized by this fact. Spaces with such a pair of dual affine structures turn out to have a richer geometry than simple affine spaces. In particular, such affine structures can be derived from potential functions. In particularly important special cases, these potential functions are the entropy and the free energy of statistical mechanics.

This pair of dual affine structures was discovered by Amari and Chentsov, and the tensor describing it is therefore called the Amari–Chentsov tensor. Like the Fisher metric, the Amari–Chentsov tensor is invariant under sufficient statistics, and uniquely characterized by this fact. Spaces with such a pair of dual affine structures turn out to have a richer geometry than simple affine spaces. In particular, such affine structures can be derived from potential functions. In particularly important special cases, these potential functions are the entropy and the free energy of statistical mechanics.

Thus, there is a natural connection between information geometry and statistical mechanics. Of course, there is also a natural connection between statistical mechanics and information theory, through the analogy between Boltzmann–Gibbs entropy and Shannon information. In many interesting cases within statistical mechanics, the interaction of physical elements can be described in terms of a graph or, more generally, in terms of a hypergraph. This leads to families of Boltzmann–Gibbs distributions that are known as hierarchical or graphical models.

2. The Fisher metric and the Amari-Chentsov tensor

A little differential geometry

Einstein summation convention $a^i b_i := \sum_{i=1}^{d} a^i b_i$

A little differential geometry

Tangent vectors are dual to 1-forms

$$dx^i \left(\frac{\partial}{\partial x^j} \right) = \delta^i_j. \tag{23}$$

A little differential geometry

Tangent vectors are dual to 1-forms

$$dx^i \left(\frac{\partial}{\partial x^j} \right) = \delta^i_j. \tag{23}$$

Coordinate changes $x = x(y)$ yield

$$\frac{\partial}{\partial x^i} = \frac{\partial y^\alpha}{\partial x^i} \frac{\partial}{\partial y^\alpha} \tag{24}$$

$$dx^i = \frac{\partial x^i}{\partial y_\alpha} dy^\alpha. \tag{25}$$

A little differential geometry

Tangent vectors are dual to 1-forms

$$dx^i \left(\frac{\partial}{\partial x^j} \right) = \delta^i_j. \tag{23}$$

Coordinate changes $x = x(y)$ yield

$$\frac{\partial}{\partial x^i} = \frac{\partial y^\alpha}{\partial x^i} \frac{\partial}{\partial y^\alpha} \tag{24}$$

$$dx^i = \frac{\partial x^i}{\partial y_\alpha} dy^\alpha. \tag{25}$$

Riemannian metric

$$\langle V, W \rangle = g_{ij} v^i w^j \tag{26}$$

for $V = v^i \frac{\partial}{\partial x^i}, W = w^i \frac{\partial}{\partial x^i}$.

A *connection* ∇ is a rule for differenting a vector field in the direction of a vector, satisfying for all (smooth) vector fields V, W_1, W_2 and functions f

$$\begin{aligned}
\nabla_{V_1+V_2} W &= \nabla_{V_1} W + \nabla_{V_2} W \\
\nabla_{fV} W &= f \nabla_V W \\
\nabla_V (W_1 + W_2) &= \nabla_V W_1 + \nabla_V W_2 \\
\nabla_V (fW) &= V(f) W + f \nabla_V W
\end{aligned}$$

A *connection* ∇ is a rule for differenting a vector field in the direction of a vector, satisfying for all (smooth) vector fields V, W_1, W_2 and functions f

$$\begin{aligned}
\nabla_{V_1+V_2} W &= \nabla_{V_1} W + \nabla_{V_2} W \\
\nabla_{fV} W &= f \nabla_V W \\
\nabla_V (W_1 + W_2) &= \nabla_V W_1 + \nabla_V W_2 \\
\nabla_V (fW) &= V(f) W + f \nabla_V W
\end{aligned}$$

In local coordinates expressed through *Christoffel symbols*

$$\nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} = \Gamma_{ij}^k \frac{\partial}{\partial x^k}. \tag{27}$$

A connection ∇ is called torsion free if

$$\nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} = \nabla_{\frac{\partial}{\partial x^j}} \frac{\partial}{\partial x^i}, \tag{28}$$

that is, if

$$\Gamma^k_{ij} = \Gamma^k_{ji} \text{ for all } i, j.$$

A connection ∇ is called torsion free if

$$\nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} = \nabla_{\frac{\partial}{\partial x^j}} \frac{\partial}{\partial x^i}, \tag{28}$$

that is, if

$$\Gamma^k_{ij} = \Gamma^k_{ji} \text{ for all } i, j.$$

Given a Riemannian metric, there is a unique connection ∇^0, called the Levi-Civita connection, that is torsion free and satisfies

$$X\langle V, W\rangle = \langle \nabla^0_X V, W\rangle + \langle V, \nabla^0_X W\rangle \text{ for all } X, V, W.$$

Its Christoffels satisfy

$$\Gamma^k_{ij} = \frac{1}{2} g^{kl} \left(\frac{\partial}{\partial x^j} g_{il} + \frac{\partial}{\partial x^i} g_{jl} - \frac{\partial}{\partial x^l} g_{ij} \right). \tag{29}$$

A connection ∇ is called torsion free if

$$\nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} = \nabla_{\frac{\partial}{\partial x^j}} \frac{\partial}{\partial x^i}, \tag{28}$$

that is, if

$$\Gamma_{ij}^k = \Gamma_{ji}^k \text{ for all } i, j.$$

A connection ∇ is called *flat* if we can find local coordinates with

$$\Gamma_{ij}^k \equiv 0.$$

The Fisher metric

$$g_{ij}(\xi) = \mathbb{E}_{\mathbf{p}(\xi)}\left(\frac{\partial}{\partial \xi^i}\log\, p(\cdot;\xi)\frac{\partial}{\partial \xi^j}\log\, p(\cdot;\xi)\right),$$
$$= \int_\Omega \frac{\partial}{\partial \xi^i}\log\, p(x;\xi)\frac{\partial}{\partial \xi^j}\log\, p(x;\xi)\; p(x;\xi)dx, \qquad (29)$$

and so

$$\frac{\partial}{\partial \xi^k} g_{ij}(\xi) = \mathbb{E}_{\mathbf{p}}\left(\frac{\partial}{\partial \xi^k}\frac{\partial}{\partial \xi^i}\log\, p\;\frac{\partial}{\partial \xi^j}\log\, p\right)$$
$$+ \mathbb{E}_{\mathbf{p}}\left(\frac{\partial}{\partial \xi^i}\log\, p\;\frac{\partial}{\partial \xi^k}\frac{\partial}{\partial \xi^j}\log\, p\right)$$
$$+ \mathbb{E}_{\mathbf{p}}\left(\frac{\partial}{\partial \xi^i}\log\, p\;\frac{\partial}{\partial \xi^j}\log\, p\;\frac{\partial}{\partial \xi^k}\log\, p\right). \qquad (30)$$

Therefore,

$$\Gamma^{(0)}_{ijk} = \mathbb{E}_{\mathbf{p}}\left(\frac{\partial^2}{\partial \xi^i \partial \xi^j}\log\, p\;\frac{\partial}{\partial \xi^k}\log\, p + \frac{1}{2}\frac{\partial}{\partial \xi^i}\log\, p\;\frac{\partial}{\partial \xi^j}\log\, p\;\frac{\partial}{\partial \xi^k}\log\, p\right) \qquad (31)$$

yields the Levi-Civita connection $\nabla^{(0)}$ for the Fisher metric.

The Fisher metric and its LC-connection

$$g_{ij}(\xi) = \mathbb{E}_{\mathbf{p}(\xi)}\left(\frac{\partial}{\partial \xi^i}\log\, p(\cdot;\xi)\frac{\partial}{\partial \xi^j}\log\, p(\cdot;\xi)\right)$$

$$\Gamma^{(0)}_{ijk} = \mathbb{E}_{\mathbf{p}}\left(\frac{\partial^2}{\partial \xi^i \partial \xi^j}\log\, p\ \frac{\partial}{\partial \xi^k}\log\, p + \frac{1}{2}\frac{\partial}{\partial \xi^i}\log\, p\ \frac{\partial}{\partial \xi^j}\log\, p\ \frac{\partial}{\partial \xi^k}\log\, p\right).$$

The Fisher metric and its LC-connection

$$g_{ij}(\xi) = \mathbb{E}_{\mathbf{p}(\xi)}\left(\frac{\partial}{\partial \xi^i}\log\, p(\cdot;\xi)\frac{\partial}{\partial \xi^j}\log\, p(\cdot;\xi)\right)$$

$$\Gamma^{(0)}_{ijk} = \mathbb{E}_{\mathbf{p}}\left(\frac{\partial^2}{\partial \xi^i \partial \xi^j}\log\, p\,\frac{\partial}{\partial \xi^k}\log\, p + \frac{1}{2}\frac{\partial}{\partial \xi^i}\log\, p\,\frac{\partial}{\partial \xi^j}\log\, p\,\frac{\partial}{\partial \xi^k}\log\, p\right).$$

However, the LC-connection is not the most interesting connection here!

The Fisher metric and its LC-connection

$$g_{ij}(\xi) = \mathbb{E}_{\mathbf{p}(\xi)}\left(\frac{\partial}{\partial \xi^i}\log\, p(\cdot;\xi)\frac{\partial}{\partial \xi^j}\log\, p(\cdot;\xi)\right)$$

$$\Gamma^{(0)}_{ijk} = \mathbb{E}_{\mathbf{p}}\left(\frac{\partial^2}{\partial \xi^i \partial \xi^j}\log\, p\,\frac{\partial}{\partial \xi^k}\log\, p + \frac{1}{2}\frac{\partial}{\partial \xi^i}\log\, p\,\frac{\partial}{\partial \xi^j}\log\, p\,\frac{\partial}{\partial \xi^k}\log\, p\right).$$

However, the LC-connection is not the most interesting connection here!

A connection is *flat* if we can find coordinates for which its Christoffels vanish identically.

The Fisher metric and its LC-connection

$$g_{ij}(\xi) = \mathbb{E}_{\mathbf{p}(\xi)}\left(\frac{\partial}{\partial \xi^i}\log\, p(\cdot;\xi)\frac{\partial}{\partial \xi^j}\log\, p(\cdot;\xi)\right)$$

$$\Gamma^{(0)}_{ijk} = \mathbb{E}_{\mathbf{p}}\left(\frac{\partial^2}{\partial \xi^i \partial \xi^j}\log\, p\,\frac{\partial}{\partial \xi^k}\log\, p + \frac{1}{2}\frac{\partial}{\partial \xi^i}\log\, p\,\frac{\partial}{\partial \xi^j}\log\, p\,\frac{\partial}{\partial \xi^k}\log\, p\right).$$

However, the LC-connection is not the most interesting connection here!

A connection is *flat* if we can find coordinates for which its Christoffels vanish identically.
The LC-connection is not flat, but we shall now find two natural flat connections.

The Amari–Chentsov tensor

More generally, we can define a family $\nabla^{(\alpha)}$, $-1 \leq \alpha \leq 1$, of connections via

$$\begin{aligned}\Gamma_{ijk}^{(\alpha)} &= \mathbb{E}_{\mathbf{p}}\left(\frac{\partial^2}{\partial\xi^i\partial\xi^j}\log p\,\frac{\partial}{\partial\xi^k}\log p + \frac{1-\alpha}{2}\frac{\partial}{\partial\xi^i}\log p\,\frac{\partial}{\partial\xi^j}\log p\,\frac{\partial}{\partial\xi^k}\log p\right)\\ &= \Gamma_{ijk}^{(0)} - \frac{\alpha}{2}\mathbb{E}_{\mathbf{p}}\left(\frac{\partial}{\partial\xi^i}\log p\,\frac{\partial}{\partial\xi^j}\log p\,\frac{\partial}{\partial\xi^k}\log p\right). \qquad (32)\end{aligned}$$

This structure is more compactly encoded by the *Amari–Chentsov tensor*

$$\begin{aligned}T_{ijk} =& \mathbb{E}_{\mathbf{p}}\left(\frac{\partial}{\partial\xi^i}\log p\,\frac{\partial}{\partial\xi^j}\log p\,\frac{\partial}{\partial\xi^k}\log p\right)\\ =& \int_\Omega \frac{\partial}{\partial\xi^i}\log p(x;\xi)\frac{\partial}{\partial\xi^j}\log p(x;\xi)\frac{\partial}{\partial\xi^k}\log p(x;\xi)\,p(x;\xi)dx.\end{aligned}$$

Expectation value of score vanishes

$$\mathbb{E}_{\mathbf{p}}\left(\frac{\partial}{\partial\xi^i}\log\, p\right)=0$$

(since $\int p=1$), and

$$g_{ij}=\mathbb{E}_{\mathbf{p}}\left(\frac{\partial}{\partial\xi^i}\log\, p\frac{\partial}{\partial\xi^j}\log\, p\right), \tag{33}$$

$$T_{ijk}=\mathbb{E}_{\mathbf{p}}\left(\frac{\partial}{\partial\xi^i}\log\, p\,\frac{\partial}{\partial\xi^j}\log\, p\,\frac{\partial}{\partial\xi^k}\log\, p\right). \tag{34}$$

Lemma

All the connections $\nabla^{(\alpha)}$ are torsion free.

Proof.

A connection is torsion free iff its Christoffel symbols Γ_{ijk} are symmetric in the indices i and j. (32) exhibits that symmetry. □

Lemma

All the connections $\nabla^{(\alpha)}$ are torsion free.

Proof.

A connection is torsion free iff its Christoffel symbols Γ_{ijk} are symmetric in the indices i and j. (32) exhibits that symmetry. □

Lemma

The connections $\nabla^{(-\alpha)}$ and $\nabla^{(\alpha)}$ are dual to each other.

Connections ∇ and ∇^* are dual if

$$Z\langle V, W\rangle = \langle \nabla_Z V, W\rangle + \langle V, \nabla^*_Z W\rangle \tag{35}$$

for all tangent vectors Z and vector fields V, W.
The LC-connection ∇^0 is selfdual:

$$Z\langle V, W\rangle = \langle \nabla^0_Z V, W\rangle + \langle V, \nabla^0_Z W\rangle.$$

Lemma

All the connections $\nabla^{(\alpha)}$ are torsion free.

Proof.

A connection is torsion free iff its Christoffel symbols Γ_{ijk} are symmetric in the indices i and j. (32) exhibits that symmetry. □

Lemma

The connections $\nabla^{(-\alpha)}$ and $\nabla^{(\alpha)}$ are dual to each other.

Proof.

$$\Gamma_{ijk}^{(-\alpha)} + \Gamma_{ijk}^{(\alpha)} = 2\Gamma_{ijk}^{(0)}$$

yields $\frac{1}{2}(\nabla^{(-\alpha)} + \nabla^{(\alpha)}) = \nabla^{(0)}$ which implies that the two connections are dual to each other. □

Exponential families

$$p(x;\vartheta) = \exp(\gamma(x) + f_i(x)\vartheta^i - \psi(\vartheta)) \tag{35}$$

depending on parameters ϑ^i (with suitable integrability conditions).

Exponential families

$$p(x;\vartheta) = \exp(\gamma(x) + f_i(x)\vartheta^i - \psi(\vartheta)) \tag{35}$$

depending on parameters ϑ^i.

$$\frac{\partial}{\partial\vartheta^j}\log p(x;\vartheta) = (f_j(x) - \mathbb{E}_{\mathbf{p}}(f_j))p(x;\vartheta). \tag{36}$$

Since $\mathbb{E}_{\mathbf{p}}\left(\frac{\partial}{\partial\vartheta^k}\log p\right) = 0$, and

$$\Gamma^{(\alpha)}_{ijk} = \mathbb{E}_{\mathbf{p}}\left(\frac{\partial^2}{\partial\vartheta^i\partial\vartheta^j}\log p\,\frac{\partial}{\partial\vartheta^k}\log p + \frac{1-\alpha}{2}\frac{\partial}{\partial\vartheta^i}\log p\,\frac{\partial}{\partial\vartheta^j}\log p\,\frac{\partial}{\partial\vartheta^k}\log p\right.$$

$$\Gamma^{(1)}_{ijk} = \mathbb{E}_{\mathbf{p}}\left(\frac{\partial^2}{\partial\vartheta^i\partial\vartheta^j}\log p\,\frac{\partial}{\partial\vartheta^k}\log p\right)$$

$$= -\frac{\partial^2}{\partial\vartheta^i\partial\vartheta^j}\psi(\vartheta)\,\mathbb{E}_{\mathbf{p}}\left(\frac{\partial}{\partial\vartheta^k}\log p\right) = 0.$$

Exponential families

$$p(x;\vartheta) = \exp(\gamma(x) + f_i(x)\vartheta^i - \psi(\vartheta)) \tag{35}$$

depending on parameters ϑ^i.

$$\frac{\partial}{\partial\vartheta^j}\log p(x;\vartheta) = (f_j(x) - \mathbb{E}_{\mathbf{p}}(f_j))p(x;\vartheta). \tag{36}$$

Since $\mathbb{E}_{\mathbf{p}}\left(\frac{\partial}{\partial\vartheta^k}\log p\right) = 0$,

$$\begin{aligned}\Gamma^{(1)}_{ijk} &= \mathbb{E}_{\mathbf{p}}\left(\frac{\partial^2}{\partial\vartheta^i\partial\vartheta^j}\log\, p\ \frac{\partial}{\partial\vartheta^k}\log\, p\right)\\ &= -\frac{\partial^2}{\partial\vartheta^i\partial\vartheta^j}\psi(\vartheta)\ \mathbb{E}_{\mathbf{p}}\left(\frac{\partial}{\partial\vartheta^k}\log\, p\right) = 0.\end{aligned}$$

Lemma

ϑ yields an affine coordinate system for the so-called exponential connection $\nabla^{(1)}$ which is flat.

Mixture families

$$p(x;\eta) = c(x) + \sum_{i=1}^{d} g^i(x)\eta_i, \tag{37}$$

an affine family of probability measures depending on parameters η_i

Mixture families

$$p(x;\eta) = c(x) + \sum_{i=1}^{d} g^i(x)\eta_i, \tag{37}$$

$$(\int c(x)dx = 1, \quad \int g^i(x)dx = 0 \ \text{ for all } i)$$

$$\frac{\partial}{\partial \eta_i} \log p(x;\eta) = \frac{g^i(x)}{p(x;\eta)},$$

$$\frac{\partial^2}{\partial \eta_i \partial \eta_j} \log p(x;\eta) = -\frac{g^i(x)g^j(x)}{p(x;\eta)^2},$$

$$\frac{\partial^2}{\partial \eta_i \partial \eta_j} \log p + \frac{\partial}{\partial \eta_i} \log p \, \frac{\partial}{\partial \eta_j} \log p = 0,$$

$$\Gamma^{(-1)}_{ijk} = 0.$$

Mixture families

$$p(x;\eta) = c(x) + \sum_{i=1}^{d} g^i(x)\eta_i, \tag{37}$$

$$(\int c(x)dx = 1, \quad \int g^i(x)dx = 0 \ \text{ for all } i)$$

$$\frac{\partial}{\partial \eta_i} \log p(x;\eta) = \frac{g^i(x)}{p(x;\eta)},$$

$$\frac{\partial^2}{\partial \eta_i \partial \eta_j} \log p(x;\eta) = -\frac{g^i(x)g^j(x)}{p(x;\eta)^2},$$

$$\frac{\partial^2}{\partial \eta_i \partial \eta_j} \log p + \frac{\partial}{\partial \eta_i} \log p \, \frac{\partial}{\partial \eta_j} \log p = 0,$$

$$\Gamma^{(-1)}_{ijk} = 0.$$

Lemma

η is an affine coordinate system for the flat mixture connection $\nabla^{(-1)}$.

Amari-Nagaoka: Consider a triple consisting of a Riemannian metric and two torsion-free flat connections ∇ and ∇^* that are dual to each other.

Amari-Nagaoka: Consider a triple consisting of a Riemannian metric and two torsion-free flat connections ∇ and ∇^* that are dual to each other.
We choose affine coordinates $\vartheta^1, ..., \vartheta^d$, for ∇; the vector fields $\partial_i := \frac{\partial}{\partial \vartheta^i}$ are then parallel. We define vector fields ∂^j via

$$\langle \partial_i, \partial^j \rangle = \delta_i^j \quad \left(= \begin{cases} 1 & \text{for } i = j \\ 0 & \text{else} \end{cases} \right). \tag{38}$$

We have for any vector V

$$0 = V \langle \partial_i, \partial^j \rangle = \langle \nabla_V \partial_i, \partial^j \rangle + \langle \partial_i, \nabla^*_V \partial^j \rangle,$$

and since ∂_i is parallel for ∇, ∂^j is parallel for ∇^*. Since ∇^* is torsion-free, also $[\partial^j, \partial^k] = 0$ for all j and k, and we may find ∇^*-affine coordinates η_j with $\partial^j = \frac{\partial}{\partial \eta_j}$. The position of the indices (upper or lower) is important because it indicates the transformation behavior under coordinate changes. For example, if when changing the ϑ-coordinates ∂_i transforms as a vector (contravariantly), then ∂^j transforms as a 1-form (covariantly). For changes of the η-coordinates, the rules are reversed.

$$\partial_i = \frac{\partial}{\partial \vartheta^i}, \partial^j = \frac{\partial}{\partial \eta_j}, \quad \partial^j = (\partial^j \vartheta^i)\partial_i \quad \text{and } \partial_i = (\partial_i \eta_j)\partial^j \tag{39}$$

as the transition rules between the ϑ- and η-coordinates.

$$g_{ij} := \langle \partial_i, \partial_j \rangle, \quad g^{ij} := \langle \partial^i, \partial^j \rangle, \tag{40}$$

we obtain from $\langle \partial_i, \partial^j \rangle = \delta_i^j$

$$\frac{\partial \eta_j}{\partial \vartheta^i} = g_{ij}, \quad \frac{\partial \vartheta^i}{\partial \eta_j} = g^{ij}. \tag{41}$$

$$\partial_i = \frac{\partial}{\partial \vartheta^i}, \partial^j = \frac{\partial}{\partial \eta_j}, \quad \partial^j = (\partial^j \vartheta^i)\partial_i \quad \text{and } \partial_i = (\partial_i \eta_j)\partial^j \tag{39}$$

as the transition rules between the ϑ- and η-coordinates.

$$g_{ij} := \langle \partial_i, \partial_j \rangle, \quad g^{ij} := \langle \partial^i, \partial^j \rangle, \tag{40}$$

we obtain from $\langle \partial_i, \partial^j \rangle = \delta_i^j$

$$\frac{\partial \eta_j}{\partial \vartheta^i} = g_{ij}, \quad \frac{\partial \vartheta^i}{\partial \eta_j} = g^{ij}. \tag{41}$$

Theorem

There exist strictly convex potential functions $\varphi(\eta), \psi(\vartheta)$ with

$$\begin{aligned} \eta_i = \partial_i \psi(\vartheta), \qquad & \vartheta^i = \partial^i \varphi(\eta), \\ g_{ij} &= \partial_i \partial_j \psi, \\ g^{ij} &= \partial^i \partial^j \varphi. \end{aligned}$$

Theorem

There exist strictly convex potential functions $\varphi(\eta), \psi(\vartheta)$ *with*

$$\eta_i = \partial_i \psi(\vartheta), \qquad \vartheta^i = \partial^i \varphi(\eta), \tag{42}$$

$$g_{ij} = \partial_i \partial_j \psi, \tag{43}$$

$$g^{ij} = \partial^i \partial^j \varphi. \tag{44}$$

Proof.

Local solvability of first equation of (42) from symmetry

$$\partial_i \eta_j = g_{ij} = g_{ji} = \partial_j \eta_i. \tag{45}$$

Moreover, we obtain

$$g_{ij} = \partial_i \partial_j \psi. \tag{46}$$

Thus, ψ is strictly convex. Same for φ. □

Duality:

$$\varphi := \vartheta^i \eta_i - \psi \tag{47}$$

from which

$$\partial^i \varphi = \vartheta^i + \frac{\partial \vartheta^j}{\partial \eta_i} \eta_j - \frac{\partial \vartheta^j}{\partial \eta_i} \frac{\partial}{\partial \vartheta^j} \psi = \vartheta^i.$$

Since ψ and φ are strictly convex, the relation

$$\varphi(\eta) + \psi(\vartheta) = \vartheta^i \eta_i \tag{48}$$

means that they are related by Legendre transformations,

$$\varphi(\eta) = \max_{\vartheta} (\vartheta^i \eta_i - \psi(\vartheta)), \tag{49}$$

$$\psi(\vartheta) = \max_{\eta} (\vartheta^i \eta_i - \phi(\eta)). \tag{50}$$

Of course, all these formulae are valid locally, i.e., where ψ and φ are defined. In fact, the construction can be reversed, and all that is needed locally is a convex function $\psi(\vartheta)$ of some local coordinates.

All can be derived from a strictly convex function $\psi(\vartheta)$: metric

$$g_{ij} = \partial_i \partial_j \psi$$

and α-connection

$$\Gamma^{(\alpha)}_{ijk} = \Gamma^{(0)}_{ijk} - \frac{\alpha}{2} \partial_i \partial_j \partial_k \psi$$

where $\Gamma^{(0)}_{ijk}$ is the Levi-Civita connection for g_{ij}. Since

$$\Gamma^{(0)}_{ijk} = \frac{1}{2}(g_{ik,j} + g_{jk,i} - g_{ij,k}) = \frac{1}{2} \partial_i \partial_j \partial_k \psi, \tag{51}$$

we have

$$\Gamma^{(\alpha)}_{ijk} = \frac{1}{2}(1-\alpha) \partial_i \partial_j \partial_k \psi, \tag{52}$$

and since this is symmetric in i and j, $\nabla^{(\alpha)}$ is torsion free. Since $\Gamma^{(\alpha)}_{ijk} + \Gamma^{(-\alpha)}_{ijk} = 2\Gamma^{(0)}_{ijk}$, $\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$ are dual to each other.

$$T_{ijk} = \partial_i \partial_j \partial_k \psi \tag{53}$$

is the 3-symmetric tensor. In particular, $\Gamma^{(1)}_{ijk} = 0$, and so $\nabla^{(1)}$ defines a flat structure, and the coordinates ϑ are affine coordinates for $\nabla^{(1)}$.

In the ϑ -coordinates, the curvature of the LC- connection becomes

$$\begin{aligned} R^k_{lij} &= \frac{1}{2} g^{kn} g^{mr} (\partial_j \partial_n \partial_r \psi \; \partial_i \partial_l \partial_m \psi - \partial_j \partial_l \partial_m \psi \; \partial_i \partial_n \partial_r \psi \\ &\qquad + \frac{1}{2} \partial_j \partial_l \partial_r \psi \; \partial_i \partial_m \partial_n \psi - \frac{1}{2} \partial_j \partial_m \partial_n \psi \; \partial_i \partial_l \partial_r \psi) \\ &= \frac{1}{4} (T^k_{jr} T^r_{\ell i} - T^k_{ir} T^r_{\ell j}) \end{aligned}$$

It can be computed from the second and third derivatives of ψ; no fourth derivatives are involved. The curvature tensor is a quadratic expression of coefficients of the 3-symmetric tensor.

In the ϑ -coordinates, the curvature of the LC- connection becomes

$$\begin{aligned} R^k_{lij} &= \frac{1}{2} g^{kn} g^{mr} (\partial_j \partial_n \partial_r \psi \; \partial_i \partial_l \partial_m \psi - \partial_j \partial_l \partial_m \psi \; \partial_i \partial_n \partial_r \psi \\ &\qquad + \frac{1}{2} \partial_j \partial_l \partial_r \psi \; \partial_i \partial_m \partial_n \psi - \frac{1}{2} \partial_j \partial_m \partial_n \psi \; \partial_i \partial_l \partial_r \psi) \\ &= \frac{1}{4} (T^k_{jr} T^r_{\ell i} - T^k_{ir} T^r_{\ell j}) \end{aligned}$$

It can be computed from the second and third derivatives of ψ; no fourth derivatives are involved. The curvature tensor is a quadratic expression of coefficients of the 3-symmetric tensor. If

$$g^{ij} = \delta^{ij},$$

we get

$$\begin{aligned} R^k_{lij} &= \frac{1}{4} (\partial_j \partial_m \partial_k \psi \; \partial_i \partial_m \partial_l \psi - \partial_j \partial_m \partial_l \psi \; \partial_i \partial_m \partial_k \psi) \quad (54) \\ &= \frac{1}{4} (T_{jkm} T_{i\ell m} - T_{j\ell m} T_{ikm}) \quad \text{(sum over } m\text{)}. \end{aligned}$$

In the ϑ -coordinates, the curvature of the LC- connection becomes

$$\begin{aligned} R^k_{lij} &= \frac{1}{2} g^{kn} g^{mr} (\partial_j \partial_n \partial_r \psi \; \partial_i \partial_l \partial_m \psi - \partial_j \partial_l \partial_m \psi \; \partial_i \partial_n \partial_r \psi \\ &\qquad + \frac{1}{2} \partial_j \partial_l \partial_r \psi \; \partial_i \partial_m \partial_n \psi - \frac{1}{2} \partial_j \partial_m \partial_n \psi \; \partial_i \partial_l \partial_r \psi) \\ &= \frac{1}{4} (T^k_{jr} T^r_{\ell i} - T^k_{ir} T^r_{\ell j}) \end{aligned}$$

It can be computed from the second and third derivatives of ψ; no fourth derivatives are involved. The curvature tensor is a quadratic expression of coefficients of the 3-symmetric tensor. If

$$g^{ij} = \delta^{ij},$$

we get

$$\begin{aligned} R^k_{lij} &= \frac{1}{4} (\partial_j \partial_m \partial_k \psi \; \partial_i \partial_m \partial_l \psi - \partial_j \partial_m \partial_l \psi \; \partial_i \partial_m \partial_k \psi) \quad (54) \\ &= \frac{1}{4} (T_{jkm} T_{i\ell m} - T_{j\ell m} T_{ikm}) \quad \text{(sum over } m\text{)}. \end{aligned}$$

When this is 0, we have a Frobenius manifold (Witten–Dijkgraaf–Verlinde–Verlinde condition).

The dual connection is $\nabla^{(-1)}$, with Christoffels

$$\Gamma^{(-1)}_{ijk} = \partial_i \partial_j \partial_k \psi. \tag{55}$$

The dually affine coordinates η are again

$$\eta_j = \partial_j \psi, \tag{56}$$

$$\text{and so also } g_{ij} = \partial_i \eta_j. \tag{57}$$

The potential is again obtained by a Legendre transform

$$\varphi(\eta) = \max_{\vartheta}(\vartheta^i \eta_i - \psi(\vartheta)), \quad \psi(\vartheta) + \varphi(\eta) - \vartheta \cdot \eta = 0, \tag{58}$$

$$\vartheta^j = \partial^j \varphi(\eta), \quad g^{ij} = \frac{\partial \vartheta^j}{\partial \eta_i} = \partial^i \partial^j \varphi(\eta). \tag{59}$$

Christoffels for LC connection for metric g^{ij} w.r.t. ϑ

$$\tilde{\Gamma}^{ijk} = -\Gamma_{ijk} = -\frac{1}{2} \partial_i \partial_j \partial_k \psi, \tag{60}$$

and so

$$\tilde{\Gamma}^{(\alpha)ijk} = \tilde{\Gamma}^{ijk} - \frac{\alpha}{2} \partial_i \partial_j \partial_k \psi = -\Gamma^{(-\alpha)}_{ijk},$$

W.r.t. dual g^{ij}, α and $-\alpha$ reverse roles, $\tilde{\Gamma}^{(1)} = -\Gamma^{(-1)} = 0$ in η.

Summarizing

Theorem

A dually flat structure, i.e., a Riemannian metric g together with two flat connections ∇ and ∇^* that are dual with respect to g is locally equivalent to the datum of a single convex function ψ, where convexity here refers to local coordinates ϑ and not to any metric.

Definition

For $p, q \in M$, a differentiable parameter manifold, the *canonical divergence* of Amari-Nagaoka is

$$D(p\|q) := \psi(p) + \varphi(q) - \vartheta^i(p)\eta_i(q). \tag{61}$$

Definition

For $p, q \in M$, a differentiable parameter manifold, the *canonical divergence* of Amari-Nagaoka is

$$D(p\|q) := \psi(p) + \varphi(q) - \vartheta^i(p)\eta_i(q). \tag{61}$$

$$D(p\|q) \geq 0, \tag{62}$$

and

$$D(p\|q) = 0 \Longleftrightarrow p = q. \tag{63}$$

Theorem

The divergence is characterized by the relation

$$D(p\|q) + D(q\|r) - D(p\|r) = (\vartheta^i(p) - \vartheta^i(q))(\eta_i(r) - \eta_i(q)) \quad (64)$$

since $\psi(q) + \varphi(q) = \vartheta^i(q)\eta_i(q)$.

Theorem

The divergence is characterized by the relation

$$D(p\|q) + D(q\|r) - D(p\|r) = (\vartheta^i(p) - \vartheta^i(q))(\eta_i(r) - \eta_i(q)) \quad (64)$$

Generalization of the cosine formula in Hilbert spaces,

$$\frac{1}{2}\|p-q\|^2 + \frac{1}{2}\|q-r\|^2 - \frac{1}{2}\|p-r\|^2 = \langle p-q, r-q\rangle.$$

Theorem

The divergence is characterized by the relation

$$D(p\|q) + D(q\|r) - D(p\|r) = (\vartheta^i(p) - \vartheta^i(q))(\eta_i(r) - \eta_i(q)) \quad (64)$$

Corollary

The ∇-geodesic from q to p is $t\vartheta^i(p) + (1-t)\vartheta^i(q)$ (ϑ^i affine for ∇), and the ∇^*-geodesic from q to r is $t\eta_i(r) + (1-t)\eta_i(q)$. If the two geodesics are orthogonal at q, Pythagoras relation

$$D(p\|r) = D(p\|q) + D(q\|r). \quad (65)$$

Theorem

The divergence is characterized by the relation

$$D(p\|q) + D(q\|r) - D(p\|r) = (\vartheta^i(p) - \vartheta^i(q))(\eta_i(r) - \eta_i(q)) \quad (64)$$

Corollary

The ∇-geodesic from q to p is $t\vartheta^i(p) + (1-t)\vartheta^i(q)$ (ϑ^i affine for ∇), and the ∇^*-geodesic from q to r is $t\eta_i(r) + (1-t)\eta_i(q)$. If the two geodesics are orthogonal at q, Pythagoras relation

$$D(p\|r) = D(p\|q) + D(q\|r). \quad (65)$$

Corollary

Let $N \subset M$ be autoparallel for ∇^*, $p \in M$. Then

$$q = \operatorname{argmin}_{r \in N} D(p\|r) \quad (66)$$

iff the ∇-geodesic from p to q is orthogonal to N at q.

Exponential family

$$p(x;\vartheta) = \exp(\gamma(x) + f_i(x)\vartheta^i - \psi(\vartheta)), \tag{67}$$

with

$$\psi(\vartheta) = \log \int \exp(\gamma(x) + f_i(x)\vartheta^i)dx, \tag{68}$$

that is,

$$p(x;\vartheta) = \frac{1}{Z(\vartheta)} \exp(\gamma(x) + f_i(x)\vartheta^i) \tag{69}$$

with the expression

$$Z(\vartheta) := \int \exp(\gamma(x) + f_i(x)\vartheta^i)dx = e^{\psi(\vartheta)}, \tag{70}$$

zustandssumme or *partition function* in statistical mechanics.

Exponential family

$$p(x;\vartheta) = \exp(\gamma(x) + f_i(x)\vartheta^i - \psi(\vartheta)), \tag{67}$$

with

$$\psi(\vartheta) = \log \int \exp(\gamma(x) + f_i(x)\vartheta^i)dx, \tag{68}$$

that is,

$$p(x;\vartheta) = \frac{1}{Z(\vartheta)} \exp(\gamma(x) + f_i(x)\vartheta^i) \tag{69}$$

with the expression

$$Z(\vartheta) := \int \exp(\gamma(x) + f_i(x)\vartheta^i)dx = e^{\psi(\vartheta)}, \tag{70}$$

zustandssumme or *partition function* in statistical mechanics.

$$\frac{\partial^k Z(\vartheta)}{\partial\vartheta^{i_1}\dots\partial\vartheta^{i_k}} = \int f_{i_1}\cdots f_{i_k} \exp(\gamma(x) + f_i(x)\vartheta^i)dx, \tag{71}$$

and hence

$$\mathbb{E}_p(f_{i_1}\cdots f_{i_k}) = \frac{1}{Z(\vartheta)} \frac{\partial^k Z(\vartheta)}{\partial\vartheta^{i_1}\dots\partial\vartheta^{i_k}}. \tag{72}$$

$$\eta_i(\vartheta) := \int f_i(x)p(x;\vartheta)dx = \mathbb{E}_p(f_i), \quad \text{expect. of coeff. of } \vartheta^i \text{ w.r.t.} p(\cdot;\vartheta) \tag{73}$$

$$\eta_i = \partial_i\psi \quad \text{from (72)} \tag{74}$$

$$g_{ij} = \partial_i\partial_j\psi \quad \text{for the Fisher information metric,} \tag{75}$$

as computed above. Dual potential

$$\begin{aligned}\varphi(\eta) &= \vartheta^i\eta_i - \psi(\vartheta)\\ &= \int(\log\, p(x;\vartheta) - \gamma(x))p(x;\vartheta)dx,\end{aligned} \tag{76}$$

with entropy $-\int \log\, p(x;\vartheta)\; p(x;\vartheta)dx$. Divergence

$$\begin{aligned}D(p\|q) &= \psi(\vartheta) - \vartheta^i\int f_i(x)q(x;\eta)dx + \int(\log q(x;\eta) - \gamma(x))q(x;\eta)dx\\ &= \int(\log q(x) - \log p(x))q(x)dx,\end{aligned} \tag{77}$$

the dual of the *Kullback–Leibler divergence*.

$$
\begin{aligned}
\eta_{ij} &= \int f_i(x) f_j(x) \exp(\gamma(x) + f_k(x)\vartheta^k - \psi(\vartheta)) dx \\
&= \exp(-\psi(\vartheta)) \frac{\partial^2}{\partial\vartheta^i \partial\vartheta^j} \int \exp(\gamma(x) + f_k(x)\vartheta^k) dx \\
&= \exp(-\psi(\vartheta)) \frac{\partial}{\partial\vartheta^i} \int f_j(x) \exp(\gamma(x) + f_k(x)\vartheta^k) dx \\
&= \exp(-\psi(\vartheta)) \frac{\partial}{\partial\vartheta^i} (\exp(\psi(\vartheta))\eta_j) \\
&= \exp(-\psi(\vartheta))\eta_j \frac{\partial}{\partial\vartheta^i} \int \exp(\gamma(x) + f_k(x)\vartheta^k) dx + \frac{\partial\eta_j}{\partial\vartheta^i} \\
&= \eta_i\eta_j + g_{ij}. \qquad (78)
\end{aligned}
$$

Theorem

$$
g_{ij} = \eta_{ij} - \eta_i\eta_j. \qquad (79)
$$

(Coordinates ϑ^i as weights for observables $f_i(x)$ on the basis of $\gamma(x)$, our metric $g_{ij}(\vartheta)$ is the covariance matrix of those observables at the given weights or coordinates. □

What to remember

Exponential and mixture families are dual to each other, with potential functions given by the entropy and the free energy, and the canonical divergence being the Kullback-Leibler divergence.

3. Complexity measures

Shannon Information

$$H(X) = H(p_1, \ldots, p_n) = -\sum_i p_i \log_2 p_i \text{ (bits)} \qquad (80)$$

Shannon Information

$$H(X) = H(p_1, \ldots, p_n) = -\sum_i p_i \log_2 p_i \text{ (bits)} \qquad (80)$$

is the expected *reduction of uncertainty*, i.e., the *information gain*, if we learn which concrete value x_i of the random variable X from a known distribution p with probabilities $p_i = p(x_i)$ is realized.

Shannon Information

$$H(X) = H(p_1, \ldots, p_n) = -\sum_i p_i \log_2 p_i \text{ (bits)} \tag{80}$$

is the expected *reduction of uncertainty*, i.e., the *information gain*, if we learn which concrete value x_i of the random variable X from a known distribution p with probabilities $p_i = p(x_i)$ is realized.

Mutual information of X and Y as information gain about X from knowing Y,

$$MI(X:Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = MI(Y:X) \tag{81}$$

is *symmetric* and has a *difference structure*.

Shannon Information

$$H(X) = H(p_1, \ldots, p_n) = -\sum_i p_i \log_2 p_i \text{ (bits)} \tag{80}$$

is the expected *reduction of uncertainty*, i.e., the *information gain*, if we learn which concrete value x_i of the random variable X from a known distribution p with probabilities $p_i = p(x_i)$ is realized.

Mutual information of X and Y as information gain about X from knowing Y,

$$MI(X:Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = MI(Y:X) \tag{81}$$

Summary:

$$H(X) = MI(X:Y) + H(X|Y) \tag{82}$$

$H(X) =$ how much you learn from observing X
$MI(X:Y) =$ how much you learn about X by observing Y
$H(X|Y) =$ how much you learn from observing X when you already know Y

Shannon Information

$$H(X) = H(p_1, \ldots, p_n) = -\sum_i p_i \log_2 p_i \text{ (bits)} \tag{80}$$

is the expected *reduction of uncertainty*, i.e., the *information gain*, if we learn which concrete value x_i of the random variable X from a known distribution p with probabilities $p_i = p(x_i)$ is realized.

Mutual information of X and Y as information gain about X from knowing Y,

$$MI(X:Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = MI(Y:X) \tag{81}$$

Iteration of conditioning process

$$MI(X:Y|Z) = H(X|Z) - H(X|Y,Z) \tag{82}$$

quantifies how much additional mutual information between X and Y can be gained when we already know Z.

Maximum entropy

E.Jaynes' maximum-entropy principle: Take the least biased estimate possible on the given information, that is, don't put any information into your model that is not based on the observed data.

Maximum entropy

E.Jaynes' maximum-entropy principle: Take the least biased estimate possible on the given information, that is, don't put any information into your model that is not based on the observed data. Look for p with maximal entropy $H(p)$ under constraint that expectation values of observables f be reproduced,

$$E_p f_\alpha = \sum_i f_\alpha^i p_i \text{ for } \alpha = 1, \ldots, A. \tag{83}$$

Maximum entropy

E.Jaynes' maximum-entropy principle: Take the least biased estimate possible on the given information, that is, don't put any information into your model that is not based on the observed data. Look for p with maximal entropy $H(p)$ under constraint that expectation values of observables f be reproduced,

$$E_p f_\alpha = \sum_i f_\alpha^i p_i \text{ for } \alpha = 1, \ldots, A. \tag{83}$$

Solution is exponential distribution

$$p_j = \frac{1}{Z} \exp(\sum_\alpha \lambda_\alpha f_\alpha^j) \quad \text{with } Z = \sum_i \exp(\sum_\alpha \lambda_\alpha f_\alpha^i). \tag{84}$$

In particular, when there are no observations,

$$p_j = \frac{1}{n} \text{ for } j = 1, \ldots, n. \tag{85}$$

Kullback-Leibler divergence

Relative entropy for two probability distributions p, q

$$D(p\|q)) = \begin{cases} \sum_i p_i \log_2 \frac{p_i}{q_i} & \text{if } \operatorname{supp} p \subset \operatorname{supp} q \\ \infty & \text{else} \end{cases} \tag{86}$$

Kullback-Leibler divergence

Relative entropy for two probability distributions p, q

$$D(p\|q)) = \begin{cases} \sum_i p_i \log_2 \frac{p_i}{q_i} & \text{if } \operatorname{supp} p \subset \operatorname{supp} q \\ \infty & \text{else} \end{cases} \tag{86}$$

is positive ($D(p\|q)) > 0$ if $p \neq q$), but not symmetric ($D(p\|q)) \neq D(q\|p)$).

Kullback-Leibler divergence

Relative entropy for two probability distributions p, q

$$D(p\|q)) = \begin{cases} \sum_i p_i \log_2 \frac{p_i}{q_i} & \text{if supp } p \subset \text{supp } q \\ \infty & \text{else} \end{cases} \tag{86}$$

is positive ($D(p\|q)) > 0$ if $p \neq q$), but not symmetric ($D(p\|q)) \neq D(q\|p)$).

Example: The mutual information is the KL-divergence between the joint distribution and the product of the marginals,

$$MI(X:Y) = D(p(x,y)||p(x)p(y)). \tag{87}$$

Kullback-Leibler divergence

Relative entropy for two probability distributions p, q

$$D(p\|q)) = \begin{cases} \sum_i p_i \log_2 \frac{p_i}{q_i} & \text{if supp } p \subset \text{supp } q \\ \infty & \text{else} \end{cases} \tag{86}$$

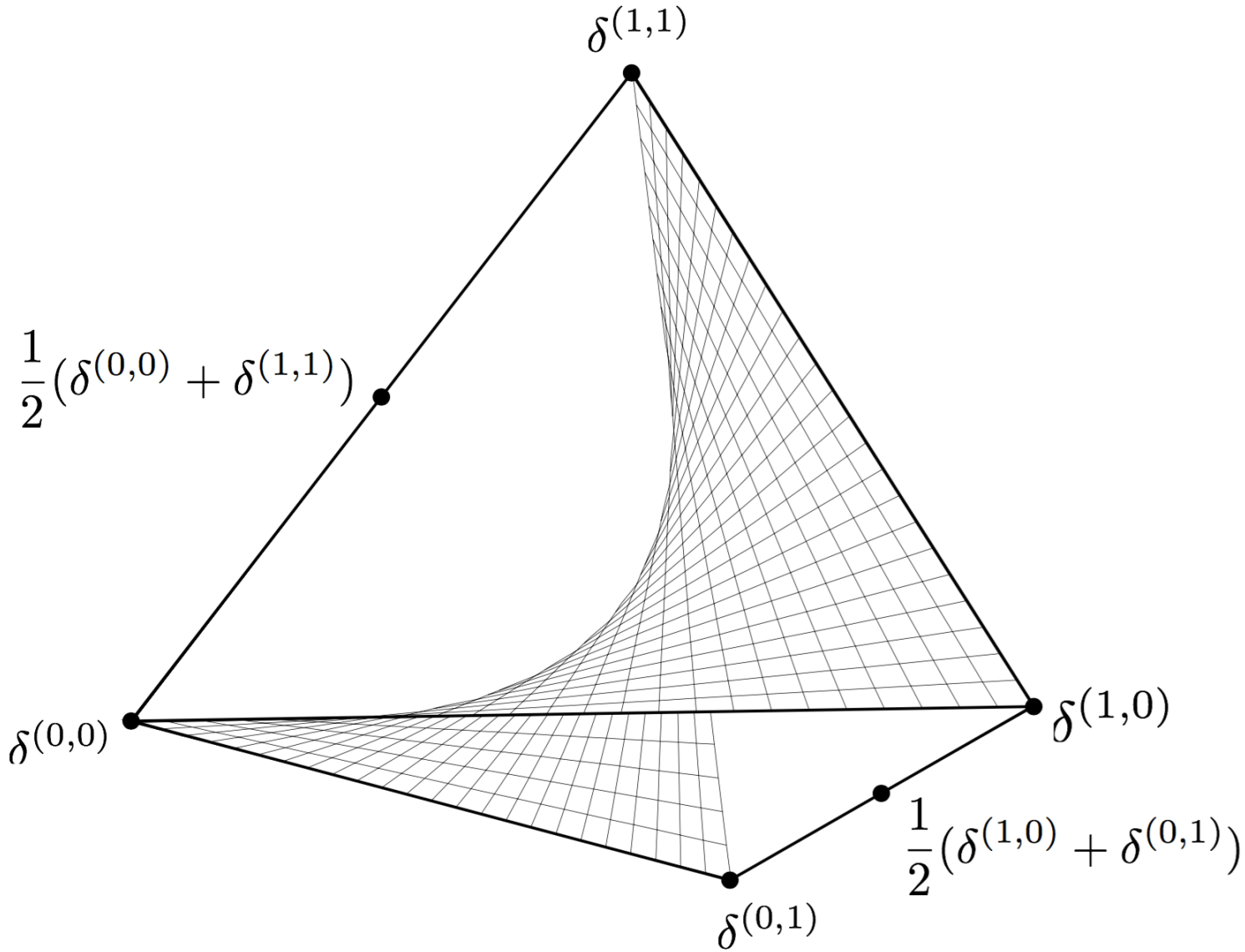
is positive ($D(p\|q)) > 0$ if $p \neq q$), but not symmetric ($D(p\|q)) \neq D(q\|p)$).

Example: The mutual information is the KL-divergence between the joint distribution and the product of the marginals,

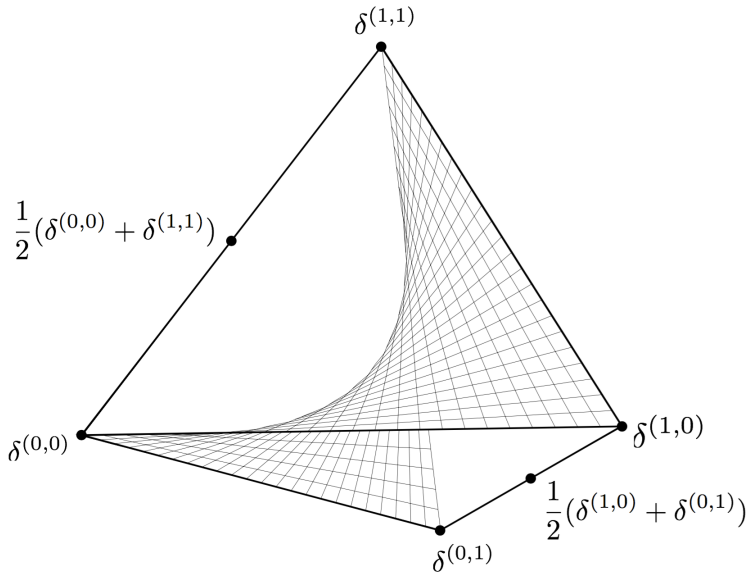
$$MI(X:Y) = D(p(x,y)||p(x)p(y)). \tag{87}$$

Among all distributions $p(x,y)$ with the same marginals $p(x) = \sum_y p(x,y), p(y) = \sum_x p(x,y)$, the product distribution $p(x)p(y)$ has the largest entropy.

Product distribution $p(x)p(y)$ has largest entropy among all distributions $p(x, y)$ with same marginals.

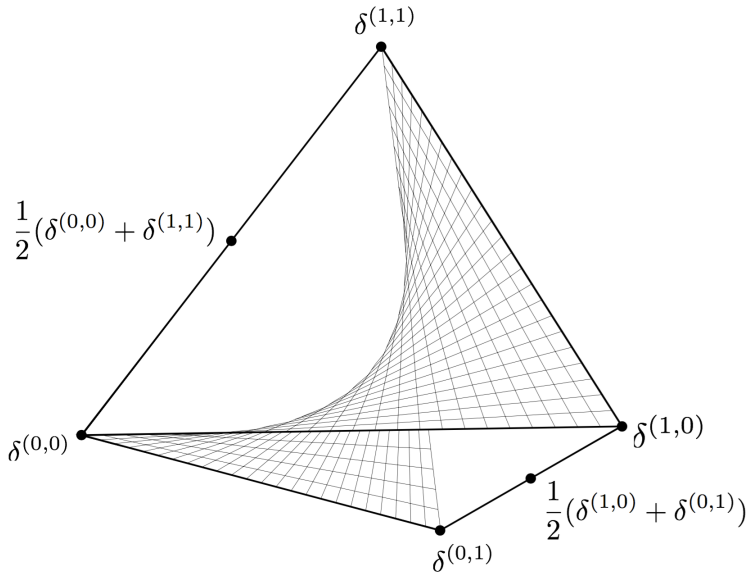


Product distribution $p(x)p(y)$ has largest entropy among all distributions $p(x, y)$ with same marginals.



3-dimensional simplex for the probability distributions on two binary variables and its 2-dimensional subfamily of product distributions. The extreme points of the simplex are the Dirac measures $\delta^{(x,y)}, x, y = 0, 1$. Maximization of distance from family of product distributions leads to distributions with support cardinality two (perfect correlation or anticorrelation).

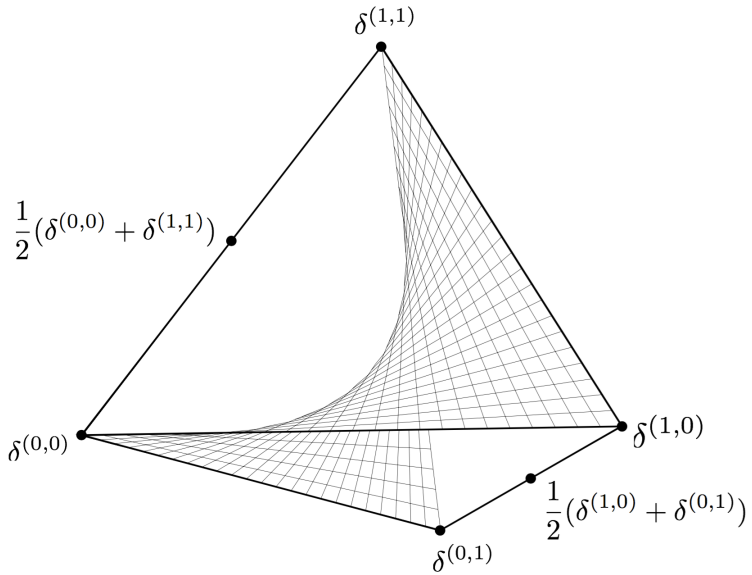
Product distribution $p(x)p(y)$ has largest entropy among all distributions $p(x, y)$ with same marginals.



simplex for the probability distributions on two binary variables and subfamily of product distributions. Project π a given distribution onto the product family $\mathcal{E}$ to maximize entropy while preserving the marginals,

$$\begin{aligned} D(p \,\|\, \mathcal{E}) \quad &:= \quad \inf_{q \in \mathcal{E}} D(p \,\|\, q) \;=\; D(p \,\|\, \pi(p)) \\ &= \quad H_{\pi(p)}(X, Y) - H_p(X, Y). \end{aligned} \tag{88}$$

Product distribution $p(x)p(y)$ has largest entropy among all distributions $p(x, y)$ with same marginals.



A family of all distributions with the same marginals is a mixture family $\mathcal{M}$. Maximizing entropy in such a family is the projection onto an exponential family $\mathcal{E}$. The two families satisfy the Pythagoras relation

$$D(p\|r) = D(p\|q) + D(q\|r) \text{ for } p \in \mathcal{M}, r \in \mathcal{E}$$

where $q = \operatorname{argmin}_{r \in \mathcal{E}} D(p\|r)$.

Hierarchical models

More generally, maximize entropy while preserving marginals among subsets of variables. For instance, for a distribution on 3 variables, we could prescribe all single and pairwise marginals.

Hierarchical models

More generally, maximize entropy while preserving marginals among subsets of variables.
State set V; consider hierarchy

$$\mathfrak{S}_1 \subseteq \mathfrak{S}_2 \subseteq \ldots \subseteq \mathfrak{S}_{N-1} \subseteq \mathfrak{S}_N := 2^V, \tag{89}$$

$\pi_{\mathfrak{S}_k}$ = projection on $\mathcal{E}_{\mathfrak{S}_k}$, $p^{(k)} := \pi_{\mathfrak{S}_k}(p)$. Pythagorean relation

$$D(p^{(l)} \,\|\, p^{(m)}) = \sum_{k=m}^{l-1} D(p^{(k+1)} \,\|\, p^{(k)}), \tag{90}$$

for $l, m = 1, \ldots, N-1$, $m < l$. In particular,

$$D(p \,\|\, p^{(1)}) = \sum_{k=1}^{N-1} D(p^{(k+1)} \,\|\, p^{(k)} b). \tag{91}$$

Take configurations with correlations of order $\leq k$, to get
Complexity measure:[1] with weight vector
$\alpha = (\alpha_1, \ldots, \alpha_{N-1}) \in \mathbb{R}^{N-1}$

$$C_\alpha(p) := \sum_{k=1}^{N-1} \alpha_k \, D(p \,\|\, p^{(k)}) \tag{92}$$

$$= \sum_{k=1}^{N-1} \beta_k \, D(p^{(k+1)} \,\|\, p^{(k)}), \tag{93}$$

with $\beta_k := \sum_{l=1}^{k} \alpha_l$.
$p^{(k)}$ is the distribution of highest entropy among all those with the same correlations of order $\leq k$ as p.
Weighted sum of higher order correlation structure.

[1] Ay, Olbrich, Bertschinger, Jost, *Chaos*, 2011

Take configurations with correlations of order $\leq k$, to get
Complexity measure:[1] with weight vector
$\alpha = (\alpha_1, \ldots, \alpha_{N-1}) \in \mathbb{R}^{N-1}$

$$C_\alpha(p) := \sum_{k=1}^{N-1} \alpha_k \, D(p \,\|\, p^{(k)}) \tag{92}$$

$$= \sum_{k=1}^{N-1} \beta_k \, D(p^{(k+1)} \,\|\, p^{(k)}), \tag{93}$$

with $\beta_k := \sum_{l=1}^{k} \alpha_l$.
$p^{(k)}$ is the distribution of highest entropy among all those with the same correlations of order $\leq k$ as p.
Examples:

- Tononi-Sporns-Edelman complexity: $\alpha_k = \frac{k}{N}$

[1] Ay, Olbrich, Bertschinger, Jost, *Chaos*, 2011

Examples (ctd.):

- Stationary stochastic process X_n: Conditional entropy

$$h_p(X_n) \;:=\; H_p(X_n \,|\, X_1, \ldots, X_{n-1}).$$

Entropy rate or *Kolmogorov–Sinai entropy*

$$h_p(X) \;:=\; \lim_{n\to\infty} h_p(X_n) \;=\; \lim_{n\to\infty} \frac{1}{n} H_p(X_1, \ldots, X_n), \quad (94)$$

Excess entropy (Grassberger)

$$\begin{aligned} E_p(X) \;&:=\; \lim_{n\to\infty} \sum_{k=1}^{n} (h_p(X_k) - h_p(X)) \\ &=\; \lim_{n\to\infty} (H_p(X_1, \ldots, X_n) - n h_p(X)) \quad (95) \\ &=\; \lim_{n\to\infty} \underbrace{\sum_{k=1}^{n-1} \frac{k}{n-k} D(p_n{}^{(k+1)} \,\|\, p_n{}^{(k)})}_{=:E_p(X_n)}. \quad (96) \end{aligned}$$

measures the non-extensive part of the entropy, i.e. amount of entropy of each element that *exceeds* the entropy rate.

Literature


Author,

Titel,

Info.


Author,

Titel,

Info.


Author,

Titel,

Info.


Author,

Titel,

Info.