Statistical mechanics from the point of view of information geometry

Jan Naudts

Universiteit Antwerpen

Rio, October 30, 2013

Part I Statistical Models
Part II Information Geometry

Part I

Statistical Models

Outline of Part I

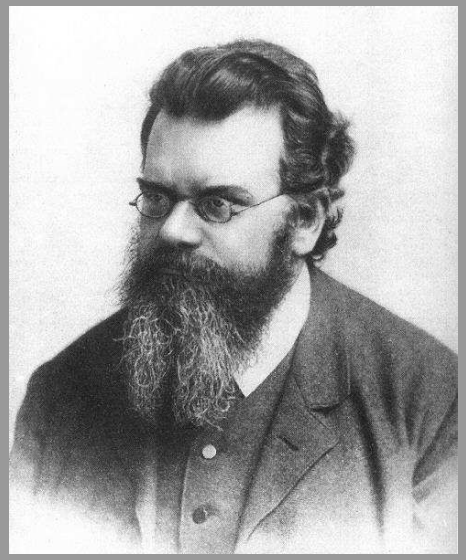
The exponential family of models

Generalized exponential families

Deformed exponential and logarithmic functions

Statistical mechanics is statistics applied to mechanics

The weight of history



Ludwig Boltzmann (1844–1906) lived before Andrei Kolmogorov (1903–1987).
Statistical physics existed before probability theory was axiomatized.
Physicists still partly use terminology found in the book of Gibbs.
Statisticians introduced a different terminology.
$\Rightarrow$ Sometimes we need a dictionary.

The exponential family of models

In physics a model is determined by its Hamiltonian H. For instance (HO),

$$H(q,p) = \frac{1}{2m}p^2 + \frac{1}{2}m\omega^2 q^2.$$

We use the Hamiltonian in the canonical ensemble of statistical physics to calculate the Boltzmann-Gibbs distribution

$$p(q,p) = \frac{1}{Z(\beta)}e^{-\beta H(q,p)}.$$

Mathematicians then say that the model belongs to the exponential family of models.
For any model of the exponential family one has the identity

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\ln Z(\beta) = -\langle H\rangle$$

$(= \mathbb{E}H$ for mathematicians).

Mean-field models do NOT belong to the exponential family!

Many models do not belong the exponential family.

Example The configurational pdf of any real gas in the microcanonical ensemble.

$$H = \frac{1}{2m}\sum_{n=1}^{N} |\mathbf{p}_n|^2 + V(\mathbf{q}_1, \mathbf{q}_2, \cdots \mathbf{q}_N).$$

$$f_E(\mathbf{q}_1, \mathbf{q}_2, \cdots \mathbf{q}_N) \sim \int \mathrm{d}\mathbf{p}_1 \cdots \int \mathrm{d}\mathbf{p}_N \, \delta(E - H).$$

This model with parameter E belongs to the q-exponential family with

$$q = 1 - \frac{2}{3N - 2}.$$

J. Naudts and M. Baeten, *Non-extensivity of the Configurational Density Distribution in the Classical Microcanonical Ensemble,*
Entropy **11**, 285–294 (2009).

Generalized exponential families

The Maximum Entropy Principle applied to the Tsallis entropy S_q yields a probability distribution which becomes Boltzmann-Gibbs for $q = 1$.

It belongs to the q-exponential family.

Or $2 - q$?

The intersection of q-exponential families with different q-values is empty.

Different Hamiltonians lead to different models belonging to the same q-exponential family.

An important model is the q-Gaussian.

Why are generalized exponential families a new (and hot) topic in mathematics?

Historically, mathematicians have considered another kind of generalization.

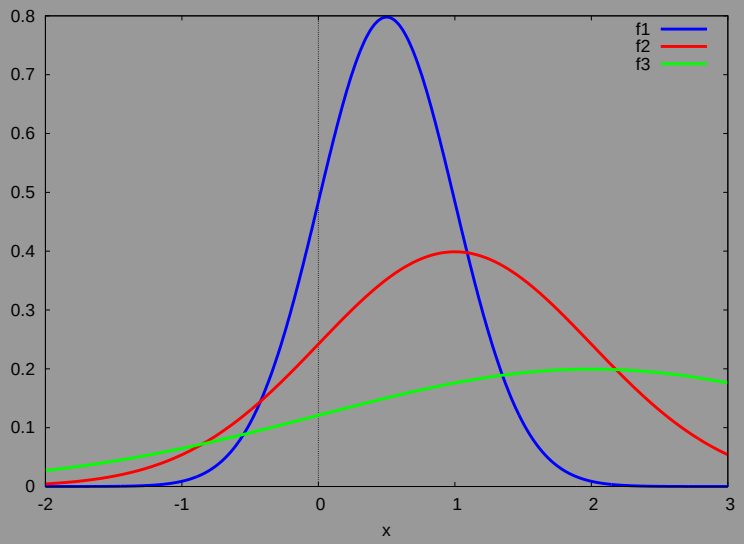
Efron (1975): A model can be curved.
Models of the exponential family are flat, not curved.
Removing parameters of a flat model can make it curved.

Example

Consider the normal distributions $p_{\mu,\sigma}(x)$ with mean μ and standard deviation σ

$$p_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}.$$

With parameters μ and σ it belongs to the exponential family.

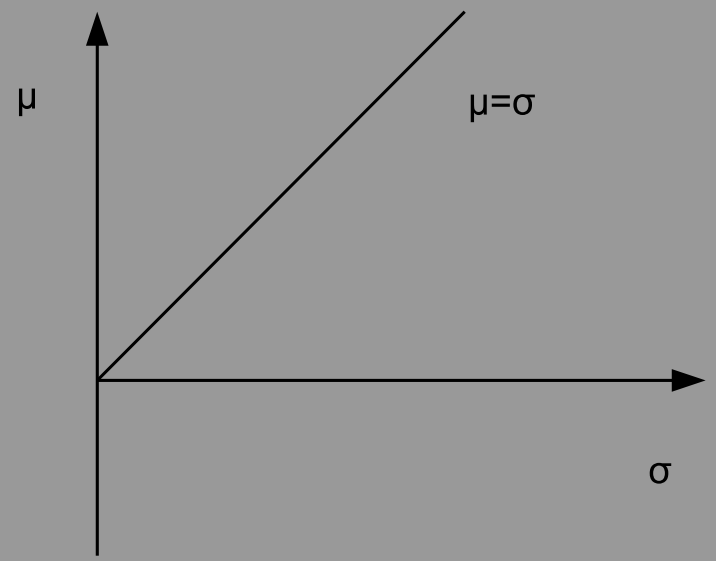


Consider the subset of normal distributions for which $\mu = \sigma$

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\theta^2}} e^{-(x-\theta)^2/2\theta^2}.$$

The one-parameter family p_θ does not belong to the exponential family. It is known to be curved.

Why curved?



Introduce canonical coordinates $\theta_1 = \sigma^{-2}$ and $\theta_2 = -\mu\sigma^{-2}$.

The line $\mu = \sigma$ becomes the curve $\theta_1 + \theta_2^2 = 0$.

Indeed, one can write the pdf as

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

The two 'Hamiltonians' are $H_1(x) = \frac{1}{2}x^2$ and $H_2(x) = x$.

Statistics

Why are generalized exponential families of interest for statisticians?

Vishwanathan and Ding:
Inference with q-exponentials is more robust.
An outlier in the tail of an exponential is very unlikely,
more unlikely than for a q-exponential with algebraic tail.

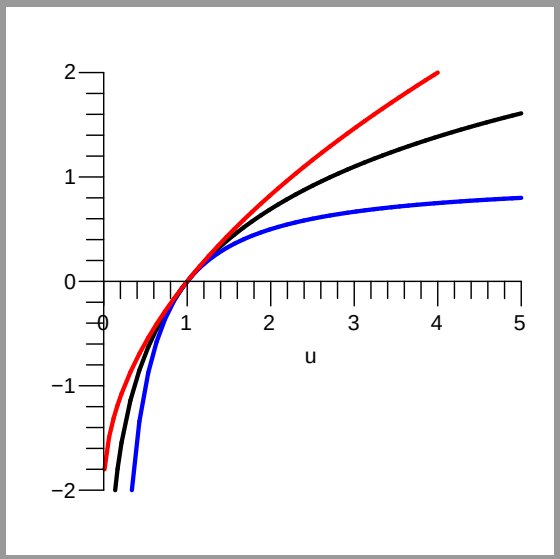
T. D. Sears, *Generalized Maximum Entropy, Convexity, and Machine Learning,* PhD thesis, Australian National University, 2008.
Nan Ding, *Statistical machine learning in the t -exponential family of distributions,* PhD thesis, Purdue, 2013.
Nan Ding, S.V. N. Vishwanathan, *t-Logistic Regression,* Adv. Neural Inf. Proc. Sys. **23**, 514–522 (2010).

Note They change notation: p, q in statistics are probabilities.
t, replacing q as a parameter, comes from the t-student distribution which belongs to the q-exponential family.

Deformed exponential and logarithmic functions

The q-deformed logarithmic function is defined by

$$\ln_q(x) = \frac{1}{1-q}\left(x^{1-q} - 1\right).$$



In the limit $q = 1$ this reduces to the natural logarithm.

The inverse function is the q-deformed exponential function

$$\exp_q(x) = [1 + (1-q)x]_+^{1/(1-q)}.$$

The figure shows $\ln_q(u)$ for $q = 0.5$ (red), 1 (black), and 2 (blue).

Definition of the q-exponential family

$$p_\theta(x) = c(x)\exp_q\left(-\alpha(\theta) - \theta^k H_k(x)\right). \quad (*)$$

But alternative definitions exist!
Information geometry can help to give a covariant definition, one which does not involve canonical coordinates.

(*) depends on the choice of parameters!

WHAT ARE THE NUMBERS THAT EXPERIMENTS PROVIDE?

Constantino Tsallis
Centro Brasileiro de Pesquisas Físicas - Rua Xavier Sigaud, 150 - 22290-180 - Rio de Janeiro - RJ

On the basis of a recently proposed generalization of Boltzmann-Gibbs Statistical Mechanics and Thermodynamics, we argue that the numbers provided by experimental measurements are to be interpreted as q-expectation values $\langle \hat{O} \rangle_q \equiv Tr\, \hat{\rho}^q \hat{O}$, where $\hat{O}$ is the observable, $\hat{\rho}$ is the density operator and the real index q characterizes the corresponding (generically nonextensive) entropy and depends on some general characteristics of the system. The familiar association with the mean value $\langle \hat{O} \rangle_1 \equiv \mathrm{Tr}\, \hat{\rho}\, \hat{O}$ as well as the extensivity of standard additive observables are recovered only for the Boltzmann-Gibbs particular case ($q = 1$), $\forall\, \hat{\rho}$, or for pure states, $\forall q$. This interpretation leaves untouched the standard additivity and conservation of energy for *pure states*, but, unless $q = 1$, modifies the definition and additivity of internal energies for *statistical mixtures*.

Keywords: measure; entropy; ensembles; generalized statistical mechanics and thermodynamics.

...One who brings
A mind not to be changed by place or time.
The mind is its own place, and in itself
Can make a heaven of hell, a hell of heaven.
Paradise Lost (1658-1665), John Milton.

the mean value $Tr\, \hat{\rho}\, \hat{O}$. We shall argue here that this is only a *particular* (though extremely ubiquitous, hence important) case. We propose instead to generically interpret the experimental result as the *q-expectation value*

$$\langle \hat{O} \rangle_q \equiv Tr\, \hat{\rho}^q \hat{O} \qquad (2)$$

C. Tsallis, *What are the numbers that experiments provide?*
Quimica Nova 17, **468**, 115 (1994).

Part II

Information Geometry

Outline of Part II

Massieu function

Fisher information

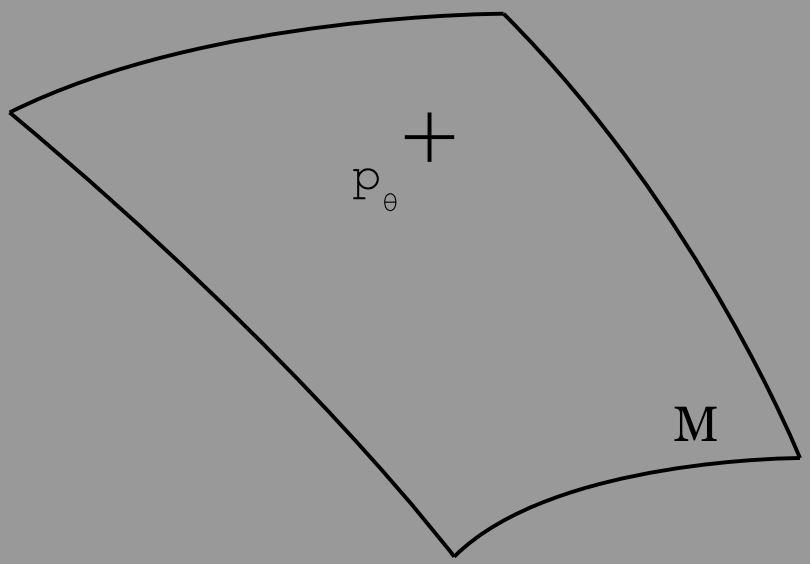
Divergence

Extended Definition of Fisher Information

Example revisited

Models may be curved.
This is the start of geometry applied to statistics

Wikipedia Information geometry is a branch of mathematics that applies the techniques of differential geometry to the field of probability theory.



A model is a *statistical manifold*.
A point of the statistical manifold is determined by the value of the parameters $\theta_1, \theta_2, \cdots, \theta_n$.

The parametrization of a model is not unique.
An equilibrium state can be identified by the inverse temperature β or by the internal energy U.
Parameter transformations are well-known from thermodynamics.

Massieu function

The Massieu function is given by $\Phi(\beta) = \ln Z(\beta)$.
Its Legendre transform is the entropy $S(U)$

$$S(U) = \inf_{\beta}\{\Phi(\beta) + \beta U\}.$$

They satisfy the dual relations

$$\frac{\mathrm{d}\Phi}{\mathrm{d}\beta} = -U \quad \text{and} \quad \frac{\mathrm{d}S}{\mathrm{d}U} = \theta.$$

Mathematicians call this duality a Hessian structure.

Distances on the manifold of model points are determined by a metric tensor $g_{k,l}$. which is given by

$$g_{k,l} = \frac{\partial^2 \Phi}{\partial\theta^k \partial\theta^l}.$$

Fisher information

The standard expression for the Fisher information matrix

$$I_{k,l}(\theta) = \int \mathrm{d}x\, p_\theta(x) \left(\frac{\partial}{\partial\theta_k} \ln p_\theta\right) \left(\frac{\partial}{\partial\theta_l} \ln p_\theta\right)$$

Theorem If the model belongs to the exponential family then $I_{k,l}(\theta) = g_{k,l}(\theta)$.

For the q-deformed Fisher information different expressions are found in the literature

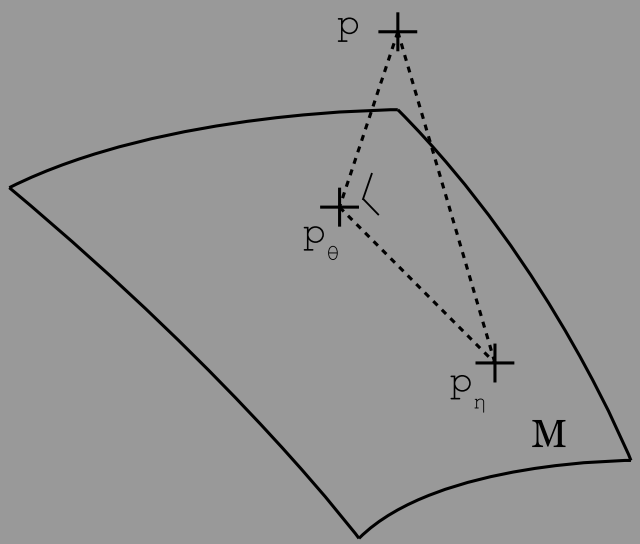
$$I_{k,l}(\theta) \sim \int \mathrm{d}x\, p_\theta^q(x) \left(\frac{\partial}{\partial\theta_k} \ln p_\theta\right) \left(\frac{\partial}{\partial\theta_l} \ln p_\theta\right) \quad \text{(Plastino et al)}$$

$$I_{k,l}(\theta) \sim \int \mathrm{d}x\, p_\theta^{2-q}(x) \left(\frac{\partial}{\partial\theta_k} \ln p_\theta\right) \left(\frac{\partial}{\partial\theta_l} \ln p_\theta\right) \quad \text{(Naudts).}$$

The difference has to do with the $q \leftrightarrow 2-q$-symmetry.

Fitting data — the geometric approach

- Fitted model parameters deviate from the unknown exact values.
- The deviation can be measured using the Fisher information matrix.



The fitting procedure itself can be described as an orthogonal projection onto the manifold of model parameters.

- The relative entropy can be used as a distance between measured data and model points.

Divergence

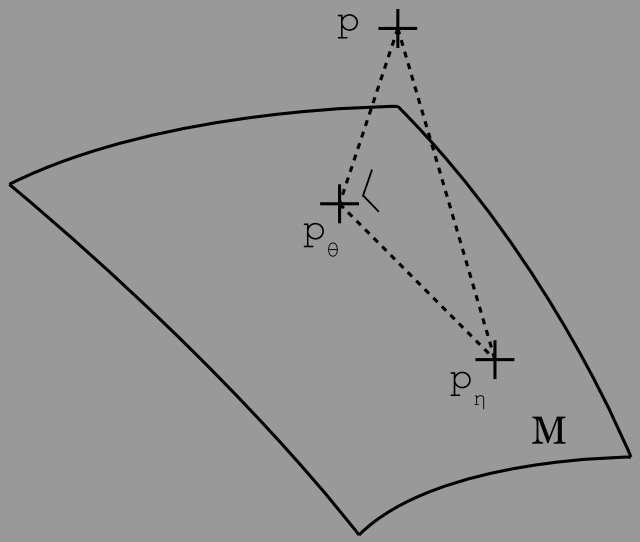
The relative entropy is in fact a relative Massieu function.
It is defined by

$$D(p||p_\beta) = [S(p_\beta) - \beta\langle H\rangle_\beta] - [S(p) - \beta\langle H\rangle_p]\,.$$

Note that $\Phi(\beta) = \sup_p\{S(p) - \beta\langle H\rangle_p\}$.

Hence $D(p||p_\beta) \geq 0$ with equality if and only if $p = p_\beta$.

- Max of Massieu is equivalent with minimal free energy.
- Mathematicians call $D(p||p_\beta)$ a divergence.
- It is a kind of squared distance between de measured data p and the model point p_β.
- The standard definition of the Kullback-Leibler divergence is $D(p||p_\theta) = \int \mathrm{d}p(x)\,\ln\dfrac{p(x)}{p_\theta(x)}$.



Pythagorean Theorem If the model belongs to the exponential family and p_θ minimizes $D(p||p_\theta)$ then

$$D(p||p_\eta) = D(p||p_\theta) + D(p_\theta||p_\eta).$$

- This property does not involve coordinates.
- A generalized exponential family must have the same property.

Extended Definition of Fisher Information

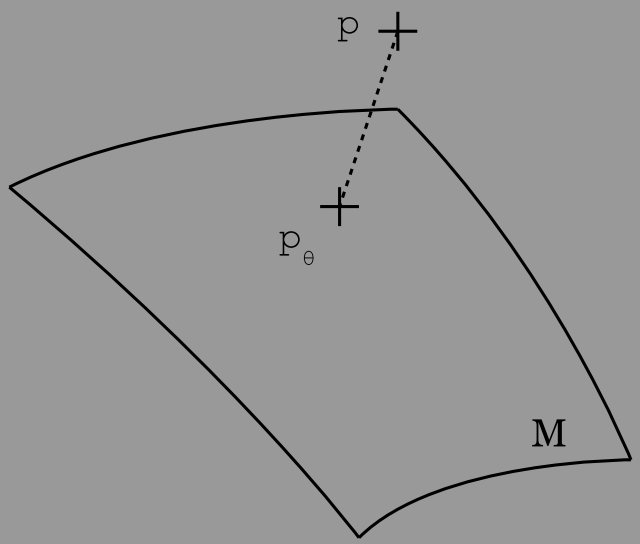
The Fisher information matrix is obtained from the divergence by

$$I_{k,l}(\theta) = \frac{\partial^2}{\partial\theta^k \partial\theta^l} D(p||p_\theta)\Big|_{p=p_\theta}.$$

Definition The *extended* Fisher information of a pdf p is

$$I_{k,l}(p) = \frac{\partial^2}{\partial\theta^k \partial\theta^l} D(p||p_\theta),$$

where p_θ minimizes $D(p||p_\theta)$.



Proposition $I_{k,l}(p)$ is covariant.

Proposition If p_θ belongs to the exponential family then $I_{k,l}(p)$ is constant along the orthogonal projection line.

These properties also hold for generalized exponential families.

J. Naudts and B. Anthonis, *The exponential family in abstract information theory,* GSI 2013 LNCS proceedings, F. Nielsen and F. Barbaresco eds., (Springer, 2013), p. 265–272.

Example revisited

Consider the manifold of normal distributions $p_{\mu,\sigma}(x)$ with mean μ and standard deviation σ

$$p_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}.$$

Consider the submanifold of normal distributions for which $\mu = \sigma$

$$p_{\theta}(x) = \frac{1}{\sqrt{2\pi\theta^2}} e^{-(x-\theta)^2/2\theta^2}.$$

Let us show that $I(p_{\mu,\sigma})$ is not constant along the projection lines. (orthogonal projection of $p_{\mu,\sigma}(x)$ on the normal distributions for which $\mu = \sigma$).

The Kullback-Leibler divergence $D(p_{\mu,\sigma}||p_\theta)$ is minimal when θ is the positive root of the equation

$$\theta^2 + \mu\theta = \mu^2 + \sigma^2.$$

The Fisher information $I(p_{\mu,\sigma})$ equals

$$I(p_{\mu,\sigma}) = \frac{\theta^2 + \mu^2 + \sigma^2}{\theta^4}.$$

It is not constant on the projection lines —
it cannot be written as a function of θ alone.

This implies that the subset satisfying $\mu = \sigma$ is *not* an exponential family.

Kyoto 2009