Entropies:
Uncertainty of Uncertainty

Alexander Gorban
University of Leicester, UK

Joint work with P. Gorban (Siberian Federal University, Krasnoyarsk, Russia)

and G. Judge (University of California, Berkeley, CA, USA)

Three Epochs of Entropy

1. The daughter of energy, $\mathrm{d}S = \Delta Q/T$, from Clausius to Boltzmann;

2. The Boltzmann–Gibbs-Shannon entropy $H = -S = \int f(x,v) \ln f(x,v) \mathrm{d}x \mathrm{d}v$;

3. Non-classical entropies: Rényi, Morimoto–Csiszár, Cressie–Read, Tsallis, Bregman,...

Three+ Points of View on Entropies

1. The Boltzmann–Gibbs-Shannon (BGS) entropy is fundamental, all other entropies are just tools for fitting (*"monarchy"*);

2. Each non-classical entropy has its own area of application where it works better than others (*"democracy"*);

3. Some nonclassical entropies are more fundamental than others because of their special properties like various generalizations of the additivity axiom (*"aristocracy"*).

We propose: Entropy may be considered as a measure of uncertainty. Non-uniqueness of the entropy makes the uncertainty uncertain. *In our talk we will try to utilize this uncertainty of uncertainty and to understand how to use all the non-classical entropies together.*

Plan of the Talk

1.Csiszár–Morimoto divergencies and popular non-classical entropies;

2.Rényi entropy;

3. Entropy production by a Markov process;

4. Generalized additivity property and the selected entropies;

5. Entropy: a function or an order?

6. Markov orders;

7. Maxima of all entropies and sets of the most random distributions.

Csiszár–Morimoto divergencies

Most of the relative entropies have the form proposed independently in 1963 by Csiszar and Morimoto

$$H_h(p) = H_h(P\|P^*) = \sum_i p_i^* h\left(\frac{p_i}{p_i^*}\right)$$

where $h(x)$ is a convex function defined on the open ($x > 0$) or closed $x \geq 0$ semi-axis. $H_h(P\|P^*)$ depends on two probability distributions p_i and p_i^*.

The Csiszár–Morimoto divergencies are the Lyapunov functions for all Markov chains with equilibrium $P^* = (p_i^*)$: for the Kolmogorov equations

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = \sum_{j,\, j\neq i} (q_{ij}p_j - q_{ji}p_i)$$

with a positive equilibrium distribution $P^* = (p_j^*)$

$$\frac{\mathrm{d}H_h(P\|P^*)}{\mathrm{d}t} \leq 0\,.$$

The Morimoto H-theorem

$$\frac{\mathrm{d}H_h(P\|P^*)}{\mathrm{d}t} =$$

$$= \sum_{i,j,\, j\neq i} q_{ij} p_j^* \left[h\left(\frac{p_i}{p_i^*}\right) - h\left(\frac{p_j}{p_j^*}\right) + h'\left(\frac{p_i}{p_i^*}\right)\left(\frac{p_j}{p_j^*} - \frac{p_i}{p_i^*}\right)\right] \leq 0$$

The last inequality holds because of the convexity of $h(x)$: $h'(x)(y-x) \leq h(y) - h(x)$ (Jensen's inequality).

The Most Popular Divergences $H_h(P\|P^*)$

1. Let $h(x)$ be the step function, $h(x) = 0$ if $x = 0$ and $h(x) = -1$ if $x > 0$. In this case,

$$H_h(P\|P^*) = -\sum_{i,\ p_i>0} 1$$

$-H_h$ is the number of positive $p_i > 0$ and does not depend on P^* (the Hartley entropy, 1928).

2. $h = |x - 1|$,

$$H_h(P\|P^*) = \sum_i |p_i - p_i^*|$$

this is the l_1-distance between P and P^*.

3. $h = x \ln x$,

$$H_h(P\|P^*) = \sum_i p_i \ln\left(\frac{p_i}{p_i^*}\right) = D_{\mathsf{KL}}(P\|P^*)$$

the Kullback–Leibler divergence (the relative BGS entropy);

4. $h = -\ln x$,

$$H_h(P\|P^*) = -\sum_i p_i^* \ln\left(\frac{p_i}{p_i^*}\right) = D_{\mathsf{KL}}(P^*\|P)$$

The relative Burg entropy. This is again the Kullback–Leibler divergence, but for another order of arguments.

5. Convex combinations of $h = x \ln x$ and $h = -\ln x$ also produces a remarkable family of divergences: $h = \beta x \ln x - (1-\beta) \ln x$ ($\beta \in [0, 1]$),

$$H_h(P\|P^*) = \beta D_{\mathsf{KL}}(P\|P^*) + (1-\beta) D_{\mathsf{KL}}(P^*\|P)$$

The convex combination of divergences becomes a symmetric functional of (P, P^*) for $\beta = 1/2$. There exists a special name for this case, "Jeffreys' entropy".

6. $h = \frac{(x-1)^2}{2}$,

$$H_h(P\|P^*) = \frac{1}{2}\sum_i \frac{(p_i - p_i^*)^2}{p_i^*}$$

This is the quadratic term in the Taylor expansion of the relative Botzmann–Gibbs-Shannon entropy, $D_{\mathsf{KL}}(P\|P^*)$, near equilibrium. Sometimes, this quadratic form is called the Fisher entropy.

7. $h = \frac{x(x^\lambda - 1)}{\lambda(\lambda+1)}$,

$$H_h(P\|P^*) = \frac{1}{\lambda(\lambda+1)} \sum_i p_i \left[\left(\frac{p_i}{p_i^*} \right)^\lambda - 1 \right]$$

This is the Cressie–Read family of power divergences, $H_{\mathsf{CR}\ \lambda}$. If $\lambda \to 0$ then $H_{\mathsf{CR}\ \lambda} \to D_{\mathsf{KL}}(P\|P^*)$, this is the classical BGS relative entropy; if $\lambda \to -1$ then $H_{\mathsf{CR}\ \lambda} \to D_{\mathsf{KL}}(P^*\|P)$, this is the relative Burg entropy.

8. For $H_{\mathrm{CR}\ \lambda}$ only the maximal terms "survive" in the limits $\lambda \to \pm\infty$. For $\lambda \to \pm\infty$ we use $(\lambda(\lambda+1)H_{\mathrm{CR}\ \lambda})^{1/|\lambda|}$:

$$H_{\mathrm{CR}\ \infty}(P\|P^*) = \max_i \left\{\frac{p_i}{p_i^*}\right\} - 1$$

$$H_{\mathrm{CR}\ -\infty}(P\|P^*) = \max_i \left\{\frac{p_i^*}{p_i}\right\} - 1$$

There may be two types of highly non-equilibrium states: with a high excess of current probability p_i above p_i^* and, inversely, with an small current probability p_i with respect to p_i^*.

9. The Tsallis relative entropy corresponds to the choice $h = \frac{(x^\alpha - x)}{\alpha - 1}$, $\alpha > 0$.

$$H_h(P\|P^*) = \frac{1}{\alpha - 1} \sum_i p_i \left[\left(\frac{p_i}{p_i^*} \right)^{\alpha - 1} - 1 \right]$$

For this family we use notation $H_{\mathsf{Ts}\ \alpha}$.

The Rényi entropy (1961) of order $\alpha > 0$ is

$$H_{\mathrm{R}\ \alpha}(P) = \frac{1}{1-\alpha} \log \left(\sum_{i=1}^{n} p_i^{\alpha} \right)$$

The relative Rényi entropy

$$H_{\mathrm{R}\ \alpha}(P \| P^*) = \frac{1}{\alpha - 1} \log \left(\sum_{i=1}^{n} p_i \left(\frac{p_i}{p_i^*} \right)^{\alpha - 1} \right)$$

$$H_{\mathsf{R}\ \alpha}(P\|P^*) = \frac{1}{\alpha-1}\ln c;$$

$$H_{\mathsf{CR}\ \alpha-1}(P\|P^*) = \frac{1}{\alpha(\alpha-1)}(c-1);$$

$$H_{\mathsf{Ts}\ \alpha}(P\|P^*) = \frac{1}{\alpha-1}(c-1)$$

where $c = \sum_i p_i(p_i/p_i^*)^{\alpha-1}$.

Lyapunov Functionals for Markov Chains

The Lyapunov functionals H_h do not depend on the kinetic coefficients q_{ij} directly. They depend on P^*. This independence of the kinetic coefficients is the *universality* property.

Theorem 1. *If a Lyapunov function $H(p)$ for the Markov chain is of the trace–form ($H(p) = \sum_i f(p_i, p_i^*)$) and is universal, then $f(p, p^*) = p^* h(\frac{p}{p^*}) + \text{const}(p^*)$, where $h(x)$ is a convex function of one variable.* P.Gorban, 2003, S. Amari, 2009

Let P and P^* be products of marginal distributions: $p_{jk} = q_j r_k$ and $p^*_{jk} = q^*_j r^*_k$. Then some entropies reveal the additivity property with respect to joining of independent systems.

*The BGS relative entropy

$$D_{\mathsf{KL}}(P\|P^*) = D_{\mathsf{KL}}(Q\|Q^*) + D_{\mathsf{KL}}(R\|R^*).$$

*The Burg entropy

$$D_{\mathsf{KL}}(P^*\|P) = D_{\mathsf{KL}}(Q^*\|Q) + D_{\mathsf{KL}}(R^*\|R) \ .$$

*The Rényi entropy

$$H_{\mathsf{R}\ \alpha}(P\|P^*) = H_{\mathsf{R}\ \alpha}(Q\|Q^*) + H_{\mathsf{R}\ \alpha}(R\|R^*).$$

Let us consider three important properties of the Lyapunov functions $H(P\|P^*)$: (1) *universality*, (2) H is a *trace–form function*, $H(P\|P^*) = \sum_i f(p_i, p_i^*)$ and H is *additive* for composition of independent subsystems.

Theorem 2. *If a function $H(P\|P^*)$ has all the properties 1)-3) simultaneously, then $f(p, p^*) = p_i^* h\left(\frac{p}{p^*}\right)$, $H(P\|P^*) = \sum_i p_i^* h\left(\frac{p_i}{p_i^*}\right)$ where $h(x) = -C_1 \ln x + C_2 x \ln x$, $C_{1,2} \geq 0$.*

Let us allow the entropy be additive in another scale: there exists such a function of one variable $\psi(x)$ that the function $\psi(H(P\|P^*))$ is additive for the union of independent subsystems: if $P = (p_{ij})$, $p_{ij} = q_j r_j$, $p^*_{ij} = q^*_j r^*_j$, then $\psi(H(P\|P^*)) = \psi(H(Q\|Q^*)) + \psi(H(R\|R^*))$.

Theorem 3. *If a C^1-smooth divergence $H(P\|P^*)$ is (1) universal Lyapunov function , (2) a trace–form function and (3) additive in some scale then, up to monotonic transformation, it is either the CR divergence or a convex combination of the Botlzmann–Gibbs–Shannon and the Burg entropies.*

Entropic aristocracy

$$H_{\mathsf{KL-B}\ \beta} = \sum_i [\beta p_i - (1-\beta) p_i^*] \ln\left(\frac{p_i}{p_i^*}\right) \ , \quad \beta \in [0,1]$$

$$H_{\mathsf{CR}\ \lambda} = \frac{1}{\lambda(\lambda+1)} \sum_i p_i \left[\left(\frac{p_i}{p_i^*}\right)^{\lambda} - 1\right] \ , \quad \lambda \in (-\infty, \infty)$$

Entropy: a Function or an Order

MaxEnt

1. An "equilibrium distribution" P^* is given; P^* may be considered as the "most disordered" distribution with respect to some a priori information.

2. We do not know the current distribution P, but we do know some linear functionals, the moments $u(P)$.

3. We do not want to introduce any subjective arbitrariness in the estimation of P and define it as the "most disordered" distribution for given value $u(P) = U$ and equilibrium P^*:

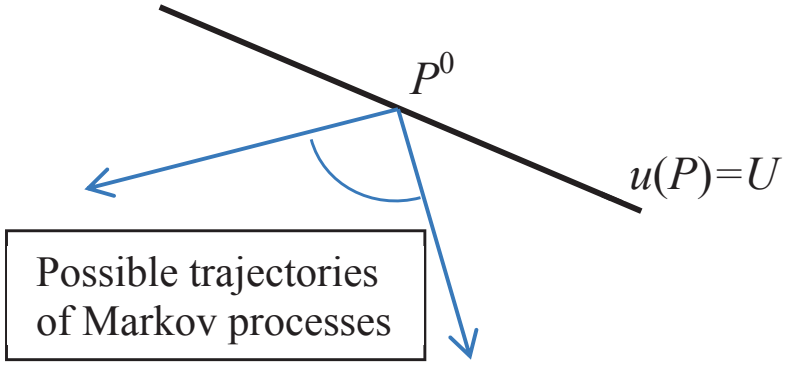
$$H_{...}(P\|P^*) \to \min \quad \text{subject to} \quad u(P) = U$$

Idea of Markov Order

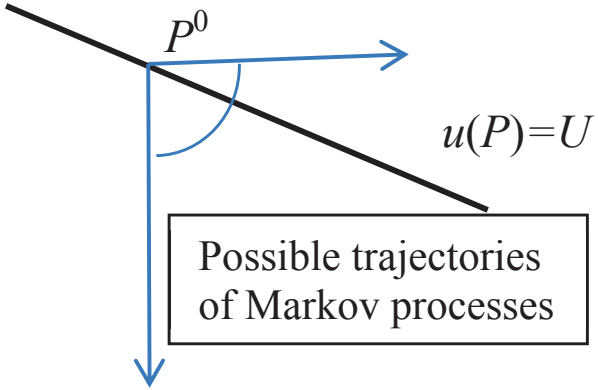
Any Markov process with equilibrium P^* increases disorder.

We can consider P^0 as a possible extremely disordered distribution on the condition plane, if for any Markov process with equilibrium P^* the solution of the Kolmogorov equation $P(t)$ with initial condition $P(0) = P^0$ has no points on the plane $u(P) = U$ for $t > 0$.

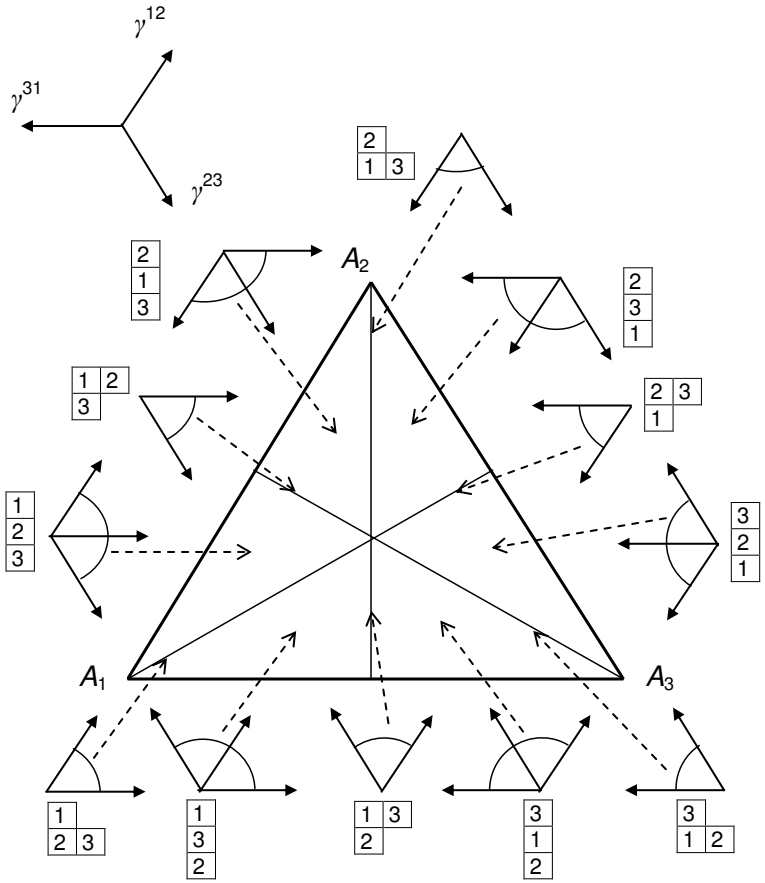
a) P^0 is a possible extremely disordered distribution



b) P^0 is not a possible extremely disordered distribution



Cones of possible velocities $\mathbf{Q}$ for the Markov chain with three states and equilibrium ($p_i^* = 1/3$).



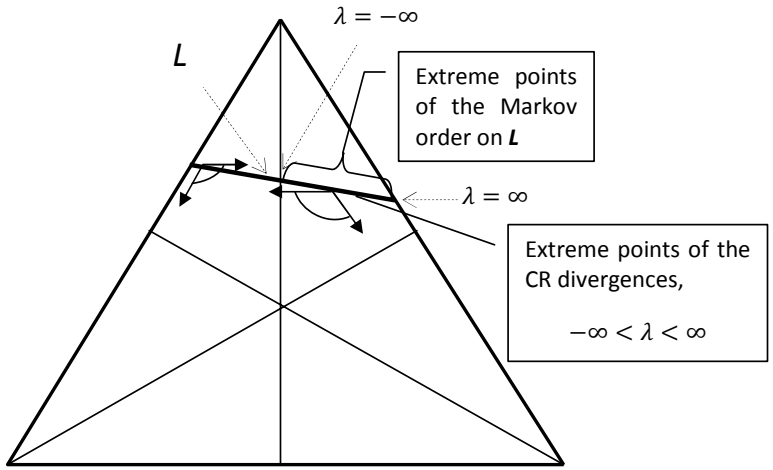
The cone of possible velocities depends in the current distribution P and on equilibrium P^*: $\mathbf{Q}_{(P,P^*)}$

Theorem 4. *Any extreme ray of* $\mathbf{Q}_{(P,P^*)}$ *corresponds to a Markov process with two states and the same equilibrium:* $A_i \rightleftarrows A_j$.

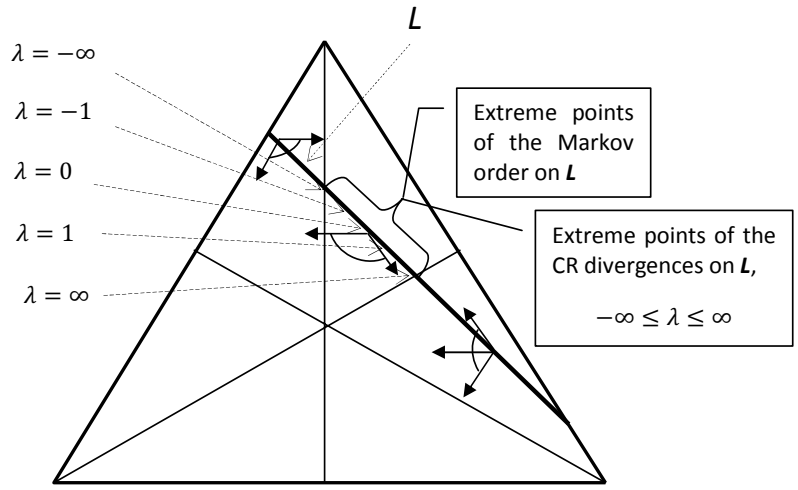
It is quite surprising that the Markov order is generated by the reversible Markov chains which satisfy the detailed balance principle.

The set of conditionally extreme points of the Markov order on the moment plane in two general positions. For several of the most important divergences these minimizers are pointed out. (For the Markov chain with three states and symmetric equilibrium ($p_i^* = 1/3$)).

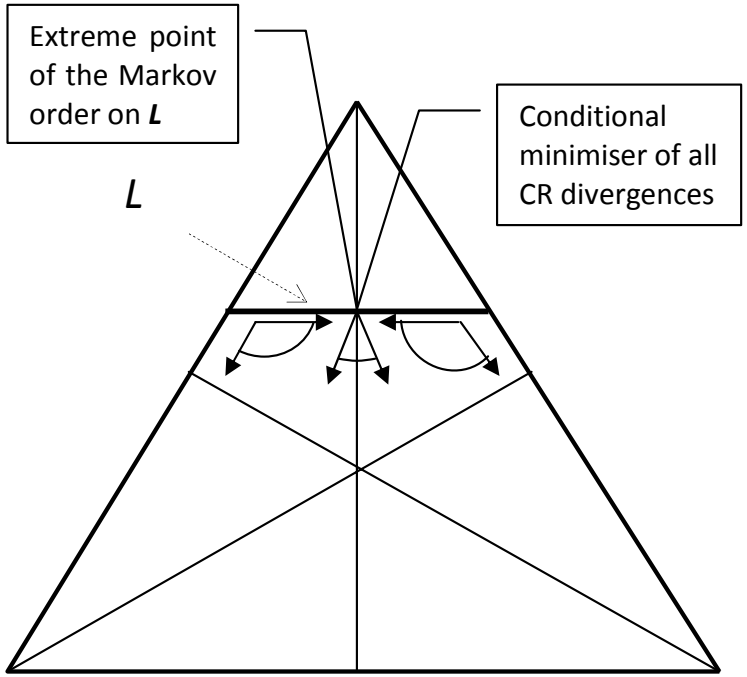
a)



b)



If the moments are just some of p_i then coincide with the unique conditionally extreme point of the Markov order.



Conclusion

1. It would be nice to have "the best entropy" for any class of problems. But from a certain point of view, ambiguity of the entropy choice is unavoidable, and the choice of all conditional optimizers instead of a particular one is a possible way to avoid an arbitrary choice.

2. The set of these minimizers evaluates the possible position of a "maximally random" probability distribution. For many MaxEnt problems the natural solution is not a fixed distribution, but a well defined set of distributions.
3. Minimization of all functions $H_h(p)$, which depend on a functional parameter h, seems to be too complicated. The Markov order allows us to find this set of possible "maximally random" probability distribution.

4. It is quite surprising that the Markov order is generated by the reversible Markov chains which satisfy the detailed balance principle.

5. Locally, the Markov order is described by the piecewise constant polyhedral cones and the description of the set of the maximally random distribution turns into a routine linear programming problem.

6. For the nonequilibrium systems, the most uncertain distribution is not unique: there exists a phenomenon of "uncertainty of uncertainty".

A couple of references

Kolmogorov, A.N. Sulla teoria di Volterra della lotta per l'esistenza. *Giornale Instituto Ital. Attuari* **1936**, *7*, 74–80.

Gorban, A.N. Invariant sets for kinetic equations. *React. Kinet. Catal. Lett.* **1979**, *10*, 187–190.

Glasser, D.; Hildebrandt, D.; Crowe, C. A geometric approach to steady flow reactors: the attainable region and optimisation in concentration space. *Am. Chem. Soc.* **1987**, 1803–1810.

Gorban, A.N.; Gorban, P.A.; Judge, G. Entropy: The Markov Ordering Approach.*Entropy* **2010**, 12, 1145-1193. Available online: http://arxiv.org/abs/1003.1377