Computers and Mathematics with Applications 65 (2013) 1438–1456

Contents lists available at SciVerse ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

Maxallent: Maximizers of all entropies and uncertainty of uncertainty

A.N. Gorban *

Department of Mathematics, University of Leicester, Leicester, LE1 7RH, UK

ARTICLE INFO

Keywords:
Uncertainty
Markov process
Lyapunov function
Entropy
Maxent
Inference

ABSTRACT

The entropy maximum approach (Maxent) was developed as a minimization of the subjective uncertainty measured by the Boltzmann–Gibbs–Shannon entropy. Many new entropies have been invented in the second half of the 20th century. Now there exists a rich choice of entropies for fitting needs. This diversity of entropies gave rise to a Maxent "anarchism". The Maxent approach is now the conditional maximization of an appropriate entropy for the evaluation of the probability distribution when our information is partial and incomplete. The rich choice of non-classical entropies causes a new problem: which entropy is better for a given class of applications? We understand entropy as a *measure of uncertainty which increases in Markov processes.* In this work, we describe the most general ordering of the distribution space, with respect to which all continuous-time Markov processes are monotonic (the Markov order). For inference, this approach results in a *set* of conditionally "most random" distributions. Each distribution from this set is a maximizer of its own entropy. This "uncertainty of uncertainty" is unavoidable in the analysis of non-equilibrium systems. Surprisingly, the constructive description of this set of maximizers is possible. Two decomposition theorems for Markov processes provide a tool for this description.

1. Introduction

Entropy was born in the 19th century as a daughter of energy: $dS = \delta Q/T$. Clausius [1], Boltzmann [2] and Gibbs [3] (and others) had developed the physical notion of entropy. At the same time, the famous Boltzmann's formula $S = k \log W$ had opened the informational interpretation of entropy. In the 20th century, Hartley [4] and Shannon [5] introduced a logarithmic measure of information in electronic communication in order "to eliminate the psychological factors involved and to establish a measure of information in terms of purely physical quantities" [4, p. 536]. Information theory is focused on entropy as a measure of uncertainty of subjective choice. This understanding of entropy was returned from information theory to statistical mechanics by Jaynes [6] as a basis of "subjective" statistical mechanics: "Information theory provides a constructive criterion for setting up probability distributions on the basis of partial knowledge, and leads to a type of statistical inference which is called the maximum entropy estimate. It is least biased estimate possible on the given information; i.e., it is maximally noncommittal with regard to missing information. That is to say, when characterizing some unknown events with a statistical model, we should always choose the one that has Maximum Entropy". This is the brief manifesto of the Maxent (maximum of entropy) methodology.

Entropy is used for measurement of uncertainty in a probability distribution. The Maxent method finds the maximally uncertain distribution under given values of some moments. After Jaynes, this approach became very popular in physics [7,8], statistics [9,10], econometrics [11,12] and other disciplines.

* Tel.: +44 1162231433.
E-mail address: ag153@le.ac.uk.

0898-1221/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2013.01.004

The non-classical entropies were invented by Rényi [13] in the middle of the 20th century, simultaneously with the expansion of the Maxent approach. This invention introduced additional uncertainty in the uncertainty evaluation. Maximization of different entropies produces different probability distributions under the same conditions. Now, one has to select the proper entropy functional to use in the Maxent approach. This choice may be non-obvious. The beautiful and transparent understanding of the Maxent distribution as a unique "least biased estimate possible on the given information" is now destroyed by non-classical entropies. If we consider the non-classical entropies seriously then we have to select the proper entropy for each problem.

If we do not find solid reasons for the entropy selection then we have to accept this "Uncertainty of Uncertainty" (UoU) as the nature of things. In this case, the *set of all the Maxent distributions for different entropies* will evaluate the unknown "maximally uncertain" distribution under given conditions. We call this method of handling the UoU the "maximization of all entropies" or *Maxallent*. If there are some reasons for selection of a class of entropy function then we have to select the conditional maximizer of the entropies from this class.

The widest class of entropies we use in this paper are the *Csiszár–Morimoto conditional entropies (f-divergencies)*. They were introduced by Rényi in his famous work [13] where he proposed also the "Rényi entropy". The f-divergencies were studied further by Csiszar [14] and T. Morimoto [15]. For a discrete probability distribution $P = (p_i)$ and the positive "equilibrium distribution" $P^* = (p_i^*)$, $p_i^* > 0$ the general form of the f-divergence is

$$\boxed{H_h(P \parallel P^*) = \sum_i p_i^* h\left(\frac{p_i}{p_i^*}\right),} \tag{1}$$

where $h(x)$ is a convex function defined on the open ($x > 0$) or closed ($x \geq 0$) semi-axis. We use here the notation $H_h(P \parallel P^*)$ to stress the dependence of H_h both on p_i and p_i^*.

In some practical problems, it is convenient to use a convex function $h(x)$ with singularity at $x = 0$, for example, $h(x) = -\ln x$ (the Burg relative entropy [16]). Therefore, we assume that the function $H_h(P \parallel P^*)$ is defined for positive P and P^*. Convexity of $h(x)$ implies convexity of $H_h(P \parallel P^*)$ as a function of P. It achieves its minimal value on the equilibrium probability, $P = P^*$ $\left(\text{under conditions } \sum_i p_i = 1, \text{ and } p_i > 0\right)$. If $h(x)$ is strictly convex then $H_h(P \parallel P^*)$ is also strictly convex and this minimizer (the equilibrium) is unique.

1.1. Maxallent, approach #1: parametrization by monotonic function of one variable

The standard settings for the Maxent approach are: an event space Ω, a divergency $H_h(P \parallel P^*)$ and a set of moments $M_r(P)$ $(r = 1, \ldots, k)$ are given. Here, P is a probability distribution, P^* is the "maximally disordered" probability distribution ("equilibrium") and $H_h(P \parallel P^*)$ measures the deviation of P from P^*. Of course, for general probability spaces we have to assume that P is absolutely continuous with respect to P^* and that it is possible to compute the divergence $H_h(P \parallel P^*)$. The Maxent problem is: for given values of the moments $M_r(P)$ $(r = 1, \ldots, k)$ find the minimizers of $H_h(P \parallel P^*)$. That is, on the set of probability distributions with given values of $M_r(P)$ $(r = 1, \ldots, k)$ find the distributions that are the closest to the equilibrium P^* if we measure the deviation by $H_h(P \parallel P^*)$. The terminological mess (*Max*ent and *min*imizers) appears due to historical reasons. Divergences measure the differences between distributions and we always look for minimizers of them.

To avoid the irrelevant technicalities we consider discrete distributions. Let $\Omega = \{A_1, A_2, \ldots, A_n\}$ be a finite event space with probability distributions $P = (p_i)$. The set of probability distribution is the standard simplex Δ^{n-1} in $\mathbb{R}^n$. The set of positive distribution $(p_i > 0)$ is Δ_+^{n-1}, the relative interior of the standard simplex.

The Maxent problem for $H_h(P \parallel P^*)$ and given values of moments $\sum_j m_{rj} p_j = M_r$ $(r = 1, \ldots, k)$ reads: find $P \in \Delta^{n-1}$ such that

$$H_h(P \parallel P^*) \to \min, \quad \text{subject to } \sum_j m_{rj} p_j = M_r.$$

The total probability condition gives $\sum_j m_{0j} p_j = 1$ $(m_{0j} = 1, M_0 = 1)$. Assume that $k + 1 < n$ and

$$\operatorname{rank}(m_{rj}) = k + 1 \quad (r = 0, 1, \ldots, k;\ j = 1, 2, \ldots, n).$$

If $\operatorname{rank}(m_{rj}) < k + 1$ then just exclude some moments.

The method of Lagrange multipliers gives for $P \in \Delta_+^{n-1}$

$$h'\left(\frac{p_j}{p_j^*}\right) = \sum_{r=0}^{k} \lambda_r m_{rj} \quad (j = 1, \ldots, n). \tag{2}$$

The derivative h' is a monotonic function. Let h be strictly convex. Then the inverse function $g(y)$ exists, $g(h'(x)) = x$ (for positive x). We can apply the function g to both sides of (2) and write the expression of P and the equations for the Lagrange multipliers λ_r that are just the moment conditions $\sum_j m_{rj}p_j = M_r$:

$$\boxed{\begin{aligned} & p_j = p_j^* g\left(\sum_{r=0}^{k} \lambda_r m_{rj}\right) \quad (j = 1, \dots, n); \\ & \sum_j m_{\rho j} p_j^* g\left(\sum_{r=0}^{k} \lambda_r m_{rj}\right) = M_\rho \quad (\rho = 0, \dots, k). \end{aligned}} \tag{3}$$

Therefore, for the class of the strictly convex functions h all the positive *solutions of the Maxent problem for all f-divergencies are parametrized* (3) *by the monotonic function g*.

The function g should be defined on a real interval $(a, b) = h'((0, \infty))$ (it might be that $a = -\infty$ or $b = \infty$). The image of g should be the real semi-axis $(0, \infty)$ because p/p^* may be any positive number. Therefore, $\lim_{y\to a} g(y) = 0$ and for finite a the function g is defined on $[a, b)$. For each monotonically increasing function g on a real interval (a, b) with $\mathrm{im}\, g = (0, \infty)$, the corresponding solution of the Maxent problem is given by the distribution (3), where λ_i are the solutions of the corresponding equation. This solution of the Maxent problem is the conditional minimizer of $H_h(P \parallel P^*)$ with $h(x) = \int h'(x)\mathrm{d}x$, where $h'(x)$ is the inverse function of $g(y)$, i.e. $h'(x) = y$, where y is the solution to the equation $g(y) = x$. The additive constant in $\int h'(x)\mathrm{d}x$ does not affect the solution of any Maxent problems and may be chosen arbitrarily. Thus, we present the parametric description of the minimizers of all strictly convex divergences $H_h(P \parallel P^*)$. A monotonic function g with the values range $(0, \infty)$ serves as a parameter in this description.

For the existence of a positive distribution P which satisfies (3) the moment conditions $\sum_j m_{rj}p_j = M_r$ $(\rho = 0, \dots, k)$ should be compatible with the positivity of p_i. Of course, for arbitrary g this may be not sufficient for the existence of such a positive distribution. To guarantee the existence of a positive Maxent distribution it is sufficient to add to the function $h(x)$ a term $\varepsilon x \ln x$ with arbitrarily small positive ε. This term creates a logarithmic singularity of $h'(x)$ at zero. It is easy to check that this singularity guarantees the existence of a positive solution of (3) if the moment conditions are compatible with the positivity of p_i. For some applied purposes an additional term $-\varepsilon \ln x$ may be even more convenient [17] because it guarantees the logarithmic singularity of entropy and $h'(x)$ has the singularity $\sim -1/x$ at zero.

In this paper, the question about existence of the positive Maxent distribution is not important. We need only the conditions (3) which are necessary and sufficient for a positive distribution $P = (p_i)$ to provide a minimizer of the given f-divergency under moment conditions.

1.2. Maxallent, approach #2: the Markov order

Any Markov process with equilibrium P^* increases disorder. The classical Boltzmann–Gibbs–Shannon entropy grows in Markov processes. This theorem (the "data processing lemma") was proved in the first paper of Shannon [5] but of course the entropy growth in kinetics was known before (Boltzmann's H-theorem [2] and its generalization for the systems without detailed balance [18]).

A. Rényi proved in the first paper about the non-classical entropies [13] that all f-divergencies (1) decrease in Markov processes with equilibrium P^*. Later on, it was demonstrated that this property characterizes f-divergencies among all functions which can be presented in the form of the sum over states (the "trace form") [19–21].

The generalized data processing lemma was proven [22,23]: For every two positive probability distributions P, Q the divergence $H_h(P \parallel Q)$ decreases under action of a stochastic matrix $A = (a_{ij})$

$$H_h(AP \parallel AQ) \leq \overline{\alpha}(A) H_h(P \parallel Q),$$

where

$$\overline{\alpha}(A) = \frac{1}{2} \max_{i,k} \left\{ \sum_j |a_{ij} - a_{kj}| \right\}$$

is the ergodicity contraction coefficient, $0 \leq \overline{\alpha}(A) \leq 1$.

A second method of handling the UoU is based on a simple remark: "uncertainty of a probability distribution should increase in Markov processes". More precisely, let the most uncertain distribution P^* be given (the equilibrium). If a distribution P' can be obtained from a distribution P in a Markov process with equilibrium P^* then we can assume:

$$\text{uncertainty of } P \leq \text{uncertainty of } P'.$$

Thus, we do not care about the *values* of the uncertainty measure, we just *compare* the uncertainty of distributions: P' is more uncertain than P under given equilibrium P^* (in this sense, the values vanish but the (pre)order appears [21]).

In the Maxent approach, the entropy is used as a (pre)order in the distribution space, not as a function, and the values are not important because any monotonically increasing transformation of the entropy does not change the solution of the Maxent problem. Of course, in some other applications the values of entropy are important: in coding theory (bits per symbol) and in thermodynamics ($\mathrm{d}U = T\mathrm{d}S$) the values of the entropy have a specific important sense. Nevertheless, when

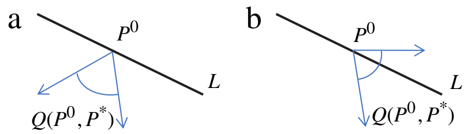

Fig. 1. The local condition (4): P^0 may be an extremely disordered distribution on the condition linear manifold L if the set $P^0 + \mathbf{Q}(P^0, P^*)$ intersects the linear manifold of conditions at the only point P^0; (a) $(P^0 + \mathbf{Q}(P, P^*)) \cap L = \{P^0\}$ and P^0 may be an extremely disordered distribution on L; (b) $(P^0 + \mathbf{Q}(P, P^*)) \cap L \supsetneqq \{P^0\}$ and P^0 has no chance to be an extremely disordered distribution on L. In case (b), there are more disordered distributions on L achievable by the Markov processes from the initial distribution P^0.

we discuss the entropy as a measure of uncertainty and work with the huge population of non-classical entropies, these entropies are, in their essence, (pre)orders on the space of distributions.

We consider the *continuous time Markov processes with a given equilibrium distribution* P^*. By definition, the equilibrium is the unconditionally maximally uncertain distribution. To add the moment conditions we define a linear manifold in the space of distributions. For every non-equilibrium distribution P each Markov process with the equilibrium distribution P^* determines the direction of P evolution, $\mathrm{d}P/\mathrm{d}t$. In this direction, the distribution becomes more uncertain. Let us take this property as a definition of the uncertainty. Instead of an entropy functional we use the transitive closure of this relation, define an order on the space of distributions and call it the "Markov order" [21].

Let $\mathbf{Q}(P, P^*)$ be a cone of possible time derivatives $\mathrm{d}P/\mathrm{d}t$ for a given probability distribution P, the equilibrium P^*, and all Markov processes with equilibrium P^*.

For fixed values of moments, M_r, the conditionally linear manifold L in the space of the probability distributions is given by equations $\sum_j m_{rj}p_j = M_r$ $(r = 0, \ldots, k)$. We can consider $P^0 \in L$ as a possibly extremely disordered distribution on L, if for any Markov process with equilibrium P^* the solution $P(t)$ of the Kolmogorov equation with initial condition $P(0) = P^0$ has no points on the conditionally linear manifold L for $t > 0$ (we assume that P^0 is not a steady state for this process). Instead of this global condition, we consider the local condition (Fig. 1).

Definition 1. The distribution $P^0 \in L \cap \Delta_+^{n-1}$ is a local minimum of the Markov order on $L \cap \Delta_+^{n-1}$ if

$$\boxed{(P^0 + \mathbf{Q}(P^0, P^*)) \cap L = \{P^0\}.} \tag{4}$$

Further, for short, we can omit Δ_+^{n-1} and call P^0 "a local minimum of the Markov order on L". In this definition, we substitute the trajectories $P(t)$ by their tangent directions at point P^0, $\mathrm{d}P(t)/\mathrm{d}t \in \mathbf{Q}(P^0, P^*)$. In Section 2 we justify this substitution and prove that the local condition (4) holds if and only if for every Markov process with equilibrium P^* the solution $P(t)$ of the Kolmogorov equation with initial condition $P(0) = P^0$ has no points on the condition linear manifold L for $t > 0$ (if P^0 is not a steady state for the process).

For applications, we need the local minima condition formalized by Definition 1 and the local order generated by the cone $\mathbf{Q}(P^0, P^*)$ only. The general notion of (global) Markov order appears later, in Section 3, where we prove equivalence of the Maxima of all entropies and the Markov order approaches. Surprisingly, the set of the conditional minimizers of all f-divergencies and the set of the conditionally minimal elements of the Markov order coincide for the same conditions (Section 3). These sets include all reasonable hypotheses about conditionally most uncertain distributions. Let us call the problem of description of all the conditional minima of the Markov order the *Maxallent problem*.

1.3. Main tool: decomposition theorems

The main tools for constructive work with the Markov orders are the decomposition theorems for Markov chains. *The first decomposition theorem* states that every Markov chain with a positive equilibrium distribution is a convex combination of the simple directed cyclic Markov chains with the same equilibrium. The coefficients in this decomposition do not depend on the current probability distribution: the vector field $\mathrm{d}P/\mathrm{d}t$ for a general Markov chain is a convex combination of these vector fields for simple cyclic Markov chains with the same positive equilibrium.

The second decomposition theorem states that for every Markov chain with a positive equilibrium distribution and for any non-equilibrium distribution P the velocity vector $\mathrm{d}P/\mathrm{d}t$ is a convex combination of the velocity vectors for the simple cyclic Markov chains *of the length two* with the same equilibrium (i.e. of the reversible transitions between two states, $A_i \rightleftharpoons A_j$). The coefficients in this decomposition typically depend on the current probability distribution.

The idea of the first decomposition theorem was used by Boltzmann in 1882 [18] in his proof of the H-theorem for systems without detailed balance. (This was his answer to the Lorentz objections [24].) He did not formulate this theorem separately but efficiently used the cycle decomposition for generalization of detailed balance. Later on, his extension of the detailed balance conditions were analyzed by many authors under different names as "cyclic balance", "semi-detailed balance" or "complex balance" (see, for example, the review [25]). Now, the theory of the cycle decomposition is a well developed area of the theory and applications of the random processes [26].

The second decomposition theorem is less known. We found this theorem in the analysis of the Markov order [21]. This decomposition means that for the general first-order kinetics and an arbitrary non-equilibrium probability distribution P there exists a system with detailed balance and the same equilibrium that has the same velocity $\mathrm{d}P/\mathrm{d}t$ at point P [27]: the classes of the general Markov processes and the Markov processes with detailed balance are pointwise equivalent.

The decomposition theorems are discussed in Appendix B in more detail.

2. Local minima of Markov order

Let us consider continuous time Markov chains with n states $A_1, \ldots, A_n$. The *Kolmogorov equation (or master equation)* for the probability distribution $P = (p_i)$ is

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = \sum_{j,\, j\neq i} (q_{ij}p_j - q_{ji}p_i) \quad (i = 1, \ldots, n), \tag{5}$$

where q_{ij} $(i, j = 1, \ldots, n, i \neq j)$ are non-negative.

In this notation, q_{ij} is the *rate constant* for the transition $A_j \to A_i$. Any non-negative values of the coefficients q_{ij} $(i \neq j)$ correspond to a master equation. Therefore, the set of all the Kolmogorov equations (5) may be considered as the positive orthant $\mathbb{R}_+^{n(n-1)}$ in $\mathbb{R}^{n(n-1)}$ with coordinates q_{ij} $(i \neq j)$.

Now, let us restrict our consideration to the set of the Markov chains with the given positive equilibrium distribution P^* $(p_i^* > 0)$.

$$\sum_{j,\, j\neq i} q_{ij}p_j^* = \left(\sum_{j,\, j\neq i} q_{ji}\right) p_i^* \quad \text{for all } i = 1, \ldots, n. \tag{6}$$

This system of uniform linear equations define a cone of the q_{ij} $(i, j = 1, \ldots, n, i \neq j)$ in $\mathbb{R}_+^{n(n-1)}$.

Under the *balance condition* (6), the Kolmogorov equations (5) may be rewritten in a convenient equivalent form:

$$\boxed{\frac{\mathrm{d}p_i}{\mathrm{d}t} = \sum_{j,\, j\neq i} q_{ij}p_j^* \left(\frac{p_j}{p_j^*} - \frac{p_i}{p_i^*}\right) \quad (i = 1, \ldots, n).} \tag{7}$$

We use below one of the f-divergencies (1) with $h(x) = (x - 1)^2$. It is a quadratic divergence, the weighted l_2 distance between P and P^*:

$$H_2(P \parallel P^*) = \sum_i \frac{(p_i - p_i^*)^2}{p_i^*}.$$

With the master equation in the form (7), it is straightforward to calculate the time derivative of $H_2(P \parallel P^*)$

$$\frac{\mathrm{d}H_2(P \parallel P^*)}{\mathrm{d}t} = -\sum_{i,j,\, j\neq i} q_{ij}p_j^* \left(\frac{p_i}{p_i^*} - \frac{p_j}{p_j^*}\right)^2 \leq 0. \tag{8}$$

Each term in the sum is non-negative. The time derivative (8) is *strictly negative* if for a transition $A_j \to A_i$ the rate constant is positive, $q_{ij} > 0$, and $\frac{p_i}{p_i^*} \neq \frac{p_j}{p_j^*}$. Hence, if the state P is not an equilibrium (i.e., the right hand side in (7) is not zero) then $\frac{\mathrm{d}H_2(P\|P^*)}{\mathrm{d}t} < 0$.

An important class of the Markov chains is formed by reversible chains with detailed balance. The *detailed balance* condition reads:

$$q_{ij}p_j^* = q_{ji}p_i^* \quad \text{for all } i, j = 1, \ldots, n. \tag{9}$$

Under this condition, there are only $\frac{n(n-1)}{2}$ independent coefficients among $n(n-1)$ numbers q_{ij}. For example, we can arbitrarily select $q_{ij} \geq 0$ for $i > j$ and then take $q_{ij} = q_{ji}\frac{p_i^*}{p_j^*}$ for $i < j$. So, for given P^*, the cone of the detailed balance systems (9) is a positive orthant in $\mathbb{R}^{\frac{n(n-1)}{2}}$ embedded in $\mathbb{R}_+^{n(n-1)}$. The equilibrium fluxes

$$w_{ij}^* = q_{ij}p_j^* = q_{ji}p_i^* \quad (i > j)$$

are the convenient coordinates in $\mathbb{R}^{\frac{n(n-1)}{2}}$ for a description of systems with detailed balance.

Let $\mathbf{Q}(P, P^*)$ be the set of all possible velocities $\mathrm{d}P/\mathrm{d}t$ at a non-equilibrium distribution P for all Markov chains which obey a given positive equilibrium P^*. According to the second decomposition theorem, the set of all possible velocities $\mathrm{d}P/\mathrm{d}t$ for the chains with detailed balance and the same equilibrium is the same cone $\mathbf{Q}(P, P^*)$. Therefore, $\mathbf{Q}(P, P^*)$ is a convex polyhedral cone and its extreme rays consist of the velocity vectors for two-state Markov chains $A_i \rightleftharpoons A_j$ with rate constants $q_{ji} = \kappa/p_j^*$, $q_{ji} = \kappa/p_i^*$ ($\kappa > 0$).

The construction of the cones of possible velocities was proposed in 1979 [28] for systems with detailed balance in the general setting, for nonlinear chemical kinetics. These systems are represented by stoichiometric equations of the elementary reactions coupled with the reverse reactions:

$$\alpha_{\rho 1}A_1 + \cdots + \alpha_{\rho n}A_n \rightleftharpoons \beta_{\rho 1}A_1 + \cdots + \beta_{\rho n}A_n, \tag{10}$$

where $\alpha_{\rho i}$, $\beta_{\rho i} \geq 0$ are the stoichiometric coefficient, ρ is the reaction number ($\rho = 1, \ldots, m$). The *stoichiometric vector* of the ρth reaction is an n dimensional vector γ_ρ with coordinates $\gamma_{\rho i} = \beta_{\rho i} - \alpha_{\rho i}$. The reaction rate is $w_\rho = w_\rho^+ - w_\rho^-$, where w_ρ^+ is the rate of the direct elementary reaction and w_ρ^- is the rate of the reverse reaction.

The equilibria of the ρth pair of reactions (10) form a hypersurface in the space of concentrations. The intersection of these surfaces for all ρ is the equilibrium (with detailed balance). Each surface of the equilibria of a pair of elementary reactions (10) divides the non-negative orthant of concentrations into three sets: (i) $w_\rho > 0$, (ii) $w_\rho = 0$ (the surface of the equilibria) and (iii) $w_\rho < 0$. All the surfaces of equilibria ($w_\rho = 0$) divide the non-negative orthant of concentrations into compartments. In each compartment, the dominant direction of each reaction (10) is fixed and, hence, the cone of possible velocities is also constant. It is a piecewise constant function of concentrations:

$$\mathbf{Q} = \mathrm{cone}\{\gamma_\rho \mathrm{sign}(w_\rho) | \rho = 1, \ldots, m\},$$

where "cone" stands for the conic hull, that is the set of all linear combinations with non-negative coefficients. Here and below we use the three-valued sign function (with values ± 1 and 0).

Let us apply this construction to Markov chains with detailed balance. Let us join the transitions $A_i \rightleftharpoons A_j$ in pairs (say, $i > j$) and introduce the *stoichiometric vectors* γ^{ji} with coordinates:

$$\gamma_k^{ji} = \begin{cases} -1 & \text{if } k = j, \\ 1 & \text{if } k = i, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Let us rewrite the Kolmogorov equation for the Markov process with detailed balance (9) in the quasichemical form:

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \sum_{i>j} w_{ij}^* \left(\frac{p_j}{p_j^*} - \frac{p_i}{p_i^*}\right) \gamma^{ji}. \tag{12}$$

Here, $w_{ij}^* = q_{ij}p_j^* = q_{ji}p_i^*$ is the *equilibrium flux* from A_i to A_j and back.

The cone of possible velocities for (12) is

$$\boxed{\mathbf{Q}(P, P^*) = \mathrm{cone}\left\{\gamma^{ji} \mathrm{sign}\left(\frac{p_j}{p_j^*} - \frac{p_i}{p_i^*}\right) \Big| i > j\right\}.} \tag{13}$$

The standard simplex of distributions P is divided by linear manifolds $\frac{p_i}{p_i^*} = \frac{p_j}{p_j^*}$ into compartments. They are the polyhedra where the cone of the local Markov order $\mathbf{Q}(P, P^*)$ is constant. The compartments for the Markov chains with the positive equilibrium P^* correspond to various partial orders on the finite set $\{p_i/p_i^*\}$ ($i = 1, \ldots, n$).

Let us describe the compartments and cones in more detail following [21]. For every natural number $k \leq n - 1$ the k-dimensional compartments are enumerated by surjective functions $\sigma : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, k + 1\}$. Such a function defines the partial ordering of quantities $\frac{p_j}{p_j^*}$ inside the compartment:

$$\frac{p_i}{p_i^*} > \frac{p_j}{p_j^*} \quad \text{if } \sigma(i) < \sigma(j); \qquad \frac{p_i}{p_i^*} = \frac{p_j}{p_j^*} \quad \text{if } \sigma(i) = \sigma(j). \tag{14}$$

Let $\mathcal{C}_\sigma$ be the corresponding compartment and Q_σ be the corresponding local Markov order cone ($\mathbf{Q}(P, P^*) = Q_\sigma$ if $P \in \mathcal{C}_\sigma$).

For a given surjection σ the compartment $\mathcal{C}_\sigma$ and the cone Q_σ have the following description:

$$\boxed{\begin{gathered} \mathcal{C}_\sigma = \left\{P \Big| \frac{p_i}{p_i^*} = \frac{p_j}{p_j^*} \quad \text{for } \sigma(i) = \sigma(j) \text{ and } \frac{p_i}{p_i^*} > \frac{p_j}{p_j^*} \text{ for } \sigma(j) = \sigma(i) + 1\right\}; \\ Q_\sigma = \mathrm{cone}\{\gamma^{ij} | \sigma(j) = \sigma(i) + 1\}. \end{gathered}} \tag{15}$$

In Fig. 2, the partition of the standard distribution simplex into compartments, and the cones (angles) of possible velocities are presented for the Markov chains with three states. In the construction of this cone, reversible chains with detailed balance are used. Due to the second decomposition theorem, this construction of the cone of possible velocities is valid for the class of general Markov chains (and not only for reversible chains) with the same equilibrium. It seems quite surprising that the

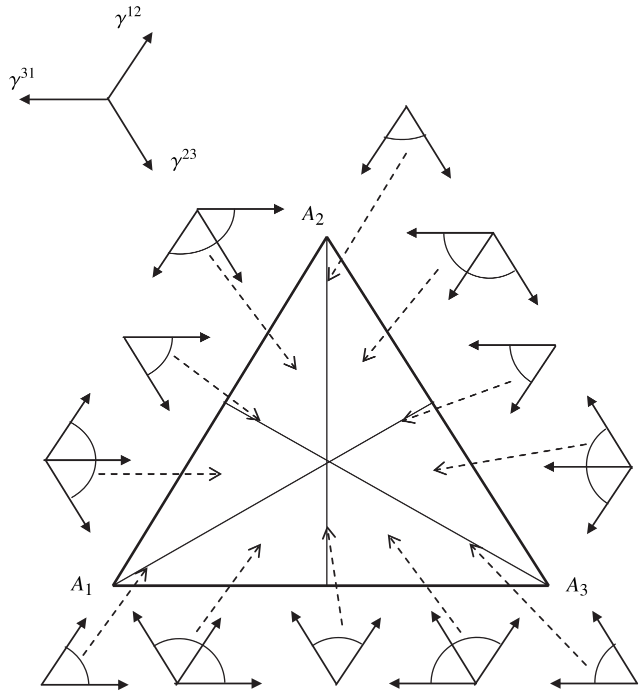

Fig. 2. The cones of possible velocities **Q** for all Markov chains with three states and equilibrium equidistribution ($p_i^* = 1/3$). The triangle of the probability distributions (p_1, p_2, p_3) ($p_i \geq 0, p_1 + p_2 + p_3 = 1$) has the vertices A_i, where $p_i = 1$ and other probabilities are zeros. Equilibrium is the center of the triangle. This triangle is divided by three lines of partial equilibria ($A_i \rightleftharpoons A_i$) into 12 compartments and the equilibrium point. Six compartments are triangles and six other compartments are segments. For all compartments the cones (here the angles) of possible velocities are shown. Each cone is connected with the corresponding compartment by a dashed line. In each cone, the vectors $\gamma^{ji}\text{sign}\left(\frac{p_j}{p_j^*} - \frac{p_i}{p_i^*}\right)$ are presented. For the 2D (triangle) compartments all three vectors are non-zero. For the 1D compartments (segments) only two these vectors are non-zero. The vectors γ^{ji} are presented separately, in the top left corner.

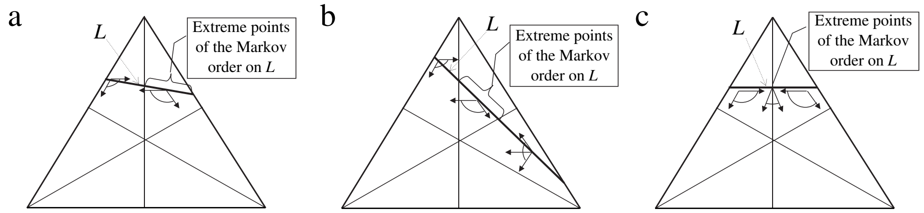

Fig. 3. (a) and (b) Extreme points of the Markov order for the Markov chain with three states and different positions of the condition line L. (c) Extreme points of the Markov order coincide with the partial equilibria, when the moments are just some of p_i.

Markov order for general Markov chains is generated by the reversible Markov chains which satisfy the detailed balance principle.

Let L be a linear manifold in the probability distribution space. Due to Definition 1, $P^0 \in L \cap \Delta_+^{n-1}$ is a local minimum of the Markov order on $L \cap \Delta_+^{n-1}$ if the condition (4) holds.

In Fig. 3 the sets of conditional minimizers are presented for the Markov order on the straight line L for the Markov chain with three states and symmetric equilibrium ($p_i^* = 1/3$). Two general positions of L in the probability triangle are used (Fig. 3(a) and (b)). If L is parallel to one side of the triangle (Fig. 3(c)) then the moments are just some of the p_i and the extreme points of the Markov order on $L \cap \Delta_+^{n-1}$ coincide with the partial equilibria.

Let J be a set of pairs of indexes (i, j) $(i > j)$ and $\mathcal{K}_J$ be the class of kinetic equations (12) with $w_{ij}^* = 0$ for $(i, j) \notin J$ and $w_{ij}^* \geq 0$ for $(i, j) \in J$ $(i \neq j)$. We define $\Phi_J(P^0)$ for an initial distribution P^0 as a set of all values $P(t)$ $(t > 0)$ for solutions $P(t)$ of all equations from the class $\mathcal{K}_J$ with initial value $P(0) = P^0$.

Consider a cone of possible velocities for the set of transitions $A_i \rightleftharpoons A_j$, $(i, j) \in J$:

$$\mathbf{Q}_J(P, P^*) = \text{cone}\left\{\gamma^{ji}\text{sign}\left(\frac{p_j}{p_j^*} - \frac{p_i}{p_i^*}\right) \Big| (i, j) \in J\right\}.$$

The following proposition states that in a vicinity of the distribution P^0 the sets $\Phi_J(P^0)$ and $P^0 + \mathbf{Q}_J(P^0, P^*)$ coincide. This gives a justification of the use of the cone of the tangent directions $\mathbf{Q}_J(P^0, P^*)$ in the definition of the local minima of the Markov order (4).

Proposition 1. *Let* $\frac{p_j}{p_j^*} - \frac{p_i}{p_i^*} \neq 0$ $((i,j) \in J)$ *for a distribution* $P = P^0$. *There exists a vicinity* U *of* P^0 *where* $P^0 + \mathbf{Q}_J(P^0, P^*)$ *coincides with* $\Phi_J(P^0)$*:*

$$(P^0 + \mathbf{Q}_J(P^0, P^*)) \cap U = \Phi_J(P^0) \cap U.$$

Proof. There exists a Euclidean ball B_r around P^0 where $\frac{p_j}{p_j^*} - \frac{p_i}{p_i^*} \neq 0$ $((i,j) \in J)$. Due to (8), inside B_r, the divergence $H_2(P \parallel P^*)$ strictly decreases with λ increasing along any ray $P^0 + \lambda e$, $e \in \mathbf{Q}_J(P, P^*)$ $(\lambda > 0)$. For each ray, we can find the minimum of $H_2(P \parallel P^*)$ in B_r. Let the maximum of these minima be $h_r(P^0)$:

$$h_r(P^0) = \max_{e \in \mathbf{Q}_J(P,P^*)} \left\{ \min_{\lambda > 0} \{H_2(P^0 + \lambda e \parallel P^*) | P^0 + \lambda e \in B_r\} \right\}.$$

By construction, $H_2(P^0 \parallel P^*) > h_r(P^0)$. The set

$$U = \{P \in B_r | H_2(P \parallel P^*) > h_r(P^0)\}$$

is a vicinity of P^0. The intersection $(P^0 + \mathbf{Q}_J(P^0, P^*)) \cap U$ is

$$(P^0 + \mathbf{Q}_J(P^0, P^*)) \cap U = \{P \in (P^0 + \mathbf{Q}_J(P^0, P^*)) | H_2(P \parallel P^*) > h_r(P^0)\}.$$

For any system from $\mathcal{K}_J$ on B_r and for any distribution $P \in (P^0 + \mathbf{Q}_J(P^0, P^*))$ the velocity vector $\mathrm{d}P/\mathrm{d}t$ belongs to $\mathbf{Q}_J(P^0, P^*)$. Obviously, $(P + \mathbf{Q}_J(P^0, P^*)) \subset (P^0 + \mathbf{Q}_J(P^0, P^*))$. Therefore, the solution of this system with the initial condition $P(0) = P^0$ may leave the intersection $(P^0 + \mathbf{Q}_J(P^0, P^*)) \cap U$ through the level surface $H_2(P^0 \parallel P^*) = h_r(P^0)$ only. After that, the solution cannot return to U because in U the values of $H_2(P^0 \parallel P^*)$ are bigger, $H_2(P^0 \parallel P^*) > h_r(P^0)$, and $H_2(P(t) \parallel P^*)$ should decrease in time along every solution of any system from $\mathcal{K}_J$. Thus, one inclusion is proven,

$$(P^0 + \mathbf{Q}_J(P^0, P^*)) \cap U \supseteq \Phi_J(P^0) \cap U.$$

To prove the second inclusion, $(P^0 + \mathbf{Q}_J(P^0, P^*)) \cap U \subseteq \Phi_J(P^0) \cap U$, we have to demonstrate that the solutions $P(t)$ $(P(0) = P^0, t \geq 0)$ of the equations from $\mathcal{K}_J$ cover $(P^0 + \mathbf{Q}_J(P^0, P^*))$ in some vicinity of P^0.

The polyhedral cone $\mathbf{Q}_J(P^0, P^*)$ is covered by the simplicial cones spanned by the sets of linearly independent vectors $\gamma^{ji} \mathrm{sign}\left(\frac{p_j^0}{p_j^*} - \frac{p_i^0}{p_i^*}\right)$. Therefore, it is sufficient to prove the second inclusion for the simplicial cones $\mathbf{Q}_J(P^0, P^*)$.

Let the vectors $\{\gamma^{ji} | (i,j) \in J\}$ be linearly independent. For the simplicity of notation, let us enumerate the states in the order of the values of p_j^0/p_j^*:

$$\frac{p_1^0}{p_1^*} \geq \frac{p_2^0}{p_2^*} \geq \cdots \geq \frac{p_n^0}{p_n^*}.$$

In these notations, $\mathrm{sign}\left(\frac{p_j^0}{p_j^*} - \frac{p_i^0}{p_i^*}\right) = 1$ for all $(i,j) \in J$ because $i > j$ and $\frac{p_j^0}{p_j^*} - \frac{p_i^0}{p_i^*} \neq 0$ for $(i,j) \in J$.

Consider a subset of the cone $\mathbf{Q}_J(P^0, P^*)$ (a "pyramid")

$$\mathcal{Q}_J(P^0, P^*) = \left\{ \sum_{(i,j) \in J} \theta_{ij} \gamma^{ji} \middle| \theta_{ij} \geq 0, \sum_{(i,j) \in J} \theta_{ij} < 1 \right\}. \tag{16}$$

The "base" of this pyramid is a simplex

$$\mathcal{B}_J(P^0, P^*) = \left\{ \sum_{(i,j) \in J} \theta_{ij} \gamma^{ji} \middle| \theta_{ij} \geq 0, \sum_{(i,j) \in J} \theta_{ij} = 1 \right\}.$$

Let $\alpha > 0$ be sufficiently small and, therefore, $\frac{p_j}{p_j^*} - \frac{p_i}{p_i^*} \neq 0$ $((i,j) \in J)$ in $P^0 + \alpha \mathcal{Q}_J(P^0, P^*)$. For this α, a solution $P(t)$ $(t \geq 0)$ of an equation from the class $\mathcal{K}_J$ with initial data $P(0) = P^0$ may leave $P^0 + \alpha \mathcal{Q}_J(P^0, P^*)$ only through its base, $P^0 + \alpha \mathcal{B}_J(P, P^*)$.

Let us prove that if α is sufficiently small then for each point $y \in \mathcal{B}_J(P, P^*)$ there exists a system in $\mathcal{K}_J$ whose solution $P(t)$ $(t > 0, P(0) = P^0)$ leaves $P^0 + a\mathcal{Q}_J(P^0, P^*)$ through the point $P^0 + \alpha y$. This means that $P(t_1) = P^0 + \alpha x$ for some $t_1 > 0$ and $P(t) \in \mathcal{Q}_J(P, P^*)$ for $0 < t < t_1$.

Each vector $x \in \mathcal{B}_J(P, P^*)$ can be expanded into a linear combination of γ^{ji} ($(i, j) \in J$):

$$x = \sum_{(i,j)\in J} \theta_{ij}\gamma^{ji}, \quad \theta_{ij} \geq 0 \quad \text{and} \quad \sum_{(i,j)\in J} \theta_{ij} = 1. \tag{17}$$

With this expansion we define the system $K_x \in \mathcal{K}_J$ by the condition $\frac{dP}{dt}\Big|_{P=P^0} = x$:

$$\frac{dP}{dt} = \sum_{(i,j)\in J} \theta_{ij} \left(\frac{p_j^0}{p_j^*} - \frac{p_i^0}{p_i^*}\right)^{-1} \left(\frac{p_j}{p_j^*} - \frac{p_i}{p_i^*}\right) \gamma^{ji}. \tag{18}$$

(Just take $w_{ij}^* = \theta_{ij}\left(\frac{p_j^0}{p_j^*} - \frac{p_i^0}{p_i^*}\right)$ for $(i, j) \in J$ in (12).) A solution $P(t)$ ($P(0) = P^0$) of this Eq. (18) can be also expanded into a linear combination of γ^{ji} $(i, j) \in J$ ($x \in \mathcal{B}_J(P, P^*)$):

$$P(t) = P^0 + tx + \frac{t^2}{2} f(t, x) = P^0 + \sum_{\theta_{ij}>0} \left(t\theta_{ij} + \frac{t^2}{2}\nu_{ij}(t, \{\theta_{lm}\})\right) \gamma^{ji}, \tag{19}$$

where $\nu_{ij}(t, \{\theta_{lm}\})$ are analytic functions. If x belongs to a face F of the cone $\mathbf{Q}_J(P^0, P^*)$ then $P(t) \in P^0 + F$ for sufficiently small t.

The moment $t = t(\alpha, x)$ when the solution $P(t)$ (19) leaves $P^0 + \alpha\mathcal{Q}_J(P^0, P^*)$ is a root of equation

$$t + \frac{t^2}{2} \sum_{(i,j)\in J} \nu_{ij}(t, \{\theta_{lm}\}) = \alpha.$$

Due to the standard inverse function theorems this root exists and the function $t(\alpha, x)$ is smooth for sufficiently small α for all $x \in \mathcal{B}_J(P, P^*)$, and $t(\alpha, x) = \alpha + o(\alpha)$. The solution $P(t)$ (19) of the system K_x (18) leaves $P^0 + \alpha\mathcal{Q}_J(P^0, P^*)$ at the point $P(t(\alpha, x)) = P^0 + \alpha y(x)$, where $y(x) \in \mathcal{B}_J(P, P^*)$.

To prove that $y(\bullet) : \mathcal{B}_J(P, P^*) \to \mathcal{B}_J(P, P^*)$ is a homeomorphism of the simplex $\mathcal{B}_J(P, P^*)$ onto itself, let us notice that the map $x \mapsto y(x)$ leaves the faces of the simplex $\mathcal{B}_J(P, P^*)$ invariant: vertices transform into themselves, the same for edges, etc.

We use the following topological lemma, the *multidimensional intermediate value theorem*. Consider a continuous map $\Psi : \Delta_n \to \Delta_n$ of the n-dimensional standard simplex into itself. Let each face $F \subset \Delta_n$ be Ψ-invariant, i.e. $\Psi(F) \subset F$. Then Ψ is surjective. The proof is possible by induction in n: for $n = 0$ it is obvious, for $n = 1$ this is just a 1D intermediate value theorem. In all dimensions, it can be proved on the basis of the "no-retraction theorem" [29] and simple inductive topological reasoning, which reduces the general case to the situation when all the faces $F \subset \Delta_n$ consist of fixed points of the map Ψ.

Therefore, for sufficiently small α the solutions $P(t)$ ($P(0) = P^0, t \geq 0$) of the equations from $\mathcal{K}_J$ cover ($P^0 + a\mathcal{Q}_J$ (P^0, P^*)) in some vicinity of P^0. The second inclusion is proven. Let us combine the inclusions and reduce the vicinities, if necessary. □

If $\frac{p_i^0}{p_i^*} \neq \frac{p_j^0}{p_j^*}$ for all i, j ($i \neq j$) then

$$\mathbf{Q}(P^0, P^*) = \mathbf{Q}(P, P^*)$$

for P in some vicinity of P^0. If for some pairs i, j ($i \neq j$) $\frac{p_i^0}{p_i^*} = \frac{p_j^0}{p_j^*}$ (see Fig. 3(c)) then for some $P \in P^0 + \mathbf{Q}(P^0, P^*)$ the cone $\mathbf{Q}(P, P^*)$ may be bigger than $\mathbf{Q}(P^0, P^*)$ even in a small vicinity of P^0. Nevertheless, the set of trajectories $P(t)$ ($t > 0, P(0) = P^0$) remains in $P^0 + \mathbf{Q}(P^0, P^*)$ for sufficiently small t. Let us prove this statement.

Let $\mathcal{K}$ be the class of all master equations with detailed balance with the positive equilibrium P^* (12) with $w_{ij}^* \geq 0$ for all (i, j) ($i > j$). We define $\Phi(P^0)$ for an initial distribution P^0 as a set of all values $P(t)$ ($t > 0$) for solutions $P(t)$ of all equations from the class $\mathcal{K}$ with initial value $P(0) = P^0$.

Proposition 2. *For every probability distribution P^0 there exists a vicinity U of P^0 where $P^0 + \mathbf{Q}(P^0, P^*)$ coincides with $\Phi(P^0)$:*

$$(P^0 + \mathbf{Q}(P^0, P^*)) \cap U = \Phi(P^0) \cap U.$$

Proof. The inclusion $(P^0 + \mathbf{Q}(P^0, P^*)) \cap U \subset \Phi(P^0) \cap U$ is proven in the second part of the proof of Proposition 1 because $\mathcal{K}_J \subset \mathcal{K}$. We have to prove the inclusion $(P^0 + \mathbf{Q}(P^0, P^*)) \cap U \supset \Phi(P^0) \cap U$.

Let us use the combinatorial description of compartments and cones (15). We assume that $P^0 \in \mathcal{C}_\sigma$ for a surjection $\sigma : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, k + 1\}$. Let us recall that $k = \dim \mathcal{C}_\sigma$. If $k = n - 1$ then $\mathcal{C}_\sigma$ is an open subset of the distribution space and the preimage of every $l = 1, 2, \dots, n$ consists of one point. For every $P \in \mathcal{C}_\sigma$ the cone $\mathbf{Q}(P, P^*)$ coincides with $\mathbf{Q}(P^0, P^*)$ and due to Proposition 1 there exists a vicinity U of P^0 where $P^0 + \mathbf{Q}(P^0, P^*)$ coincides with $\Phi(P^0)$.

Let $k < n - 1$. Then for some $i = 1, \dots, k + 1$ the preimage of i includes more than 1 point, $|\sigma^{-1}(i)| > 1$. Let I be the set of such i and $S_i = \sigma^{-1}(i)$ is the preimage of i. Due to (15),

$$\mathbf{Q}(P^0, P^*) = \mathrm{cone}\{\gamma^{ij} | \sigma(j) = \sigma(i) + 1\}.$$

For a sufficiently small ball U_r with the center P^0 and $P \in (P^0 + \mathbf{Q}(P^0, P^*)) \cap U$ the cone $\mathbf{Q}(P, P^*)$ may include also some γ^{ij} with $\sigma(i) = \sigma(j)$ but

$$\mathbf{Q}(P, P^*) \subset \mathrm{cone}\{\gamma^{ij} | \sigma(j) = \sigma(i) + 1 \text{ or } \sigma(j) = \sigma(i)\}. \tag{20}$$

Let us prove that for any Markov chain with equilibrium P^* for sufficiently small time $\tau > 0$ and a ball $U_{r/2}$ with the center P^0 the solutions of the Kolmogorov equations $P(t)$ do not leave $P^0 + \mathbf{Q}(P^0, P^*)$ during the time interval $[0, \tau]$ if $P(0) \in (P^0 + \mathbf{Q}(P^0, P^*)) \cap U_{r/2}$.

A set V is positively invariant with respect to a dynamical system if every motion that starts in V at $t = 0$ remains there for $t > 0$. Let a convex set V be positively invariant with respect to several dynamical system given by Lipschitz vector fields $\mathbf{w}_1, \dots, \mathbf{w}_r$. Then V is positively invariant with respect to any combination $\mathbf{w} = \sum_j f_j \mathbf{w}_j$, where f_j are non-negative functions and $\mathbf{w}$ is a Lipschitz vector field. Therefore, the problem of positive invariance of a convex set with respect to such combinations of vector fields can be "split" into problems of the positive invariance of V with respect to summands $\mathbf{w}_j$. Due to the second decomposition theorem, we can always assume that the vector field of the Kolmogorov equation for the Markov kinetics is a linear combination of the vector fields of the pairs of elementary transitions $A_i \rightleftharpoons A_j$ with the same equilibrium. The coefficients in these combinations are non-negative functions.

The motion $P(t)$ with $P(0) \in (P^0 + \mathbf{Q}(P^0, P^*))$ does not leave $(P^0 + \mathbf{Q}(P^0, P^*))$ in time $t \in [0, \tau]$ if $\mathrm{d}P(t)/\mathrm{d}t \in \mathbf{Q}(P^0, P^*)$ on $[0, \tau]$.

The cone $\mathbf{Q}(P^0, P^*)$ is generated by vectors γ^{ij} with $\sigma(j) = \sigma(i) + 1$. To generate a cone $\mathbf{Q}(P, P^*)$ for a point $P \in U_r$ we have to add to the set of γ^{ij} $(\sigma(j) = \sigma(i) + 1)$ some of γ^{ij} with $\sigma(j) = \sigma(i)$. Let us consider the pyramid (compare to (16))

$$\mathcal{Q}(P^0) = \left\{ \sum_{\sigma(j)=\sigma(i)+1} \theta_{ij} \gamma^{ji} \middle| \theta_{ij} \geq 0, \sum_{(i,j) \in J} \theta_{ij} < 1 \right\}.$$

We will prove that the set $P^0 + a\mathcal{Q}(P^0)$ is positively invariant with respect to any first order kinetics with transitions $A_i \rightleftharpoons A_j$ $(i, j \in S_l)$ and equilibrium P^* for any $l = 1, \dots, k + 1$.

It is sufficient to consider dynamics in projections on the coordinate subspace $\mathbf{R}_{S_l}$ with coordinates p_i, $i \in S_l$ for every $l \in I$ separately. In this space, vectors γ^{ij} $(i, j \in S_l)$ correspond to the standard first order kinetics like (12) with the reduced vector $P \in \mathbf{R}_{S_l}$ but without compulsory unit balance $\left(\sum_{i \in S_l} p_i = const \text{ with any const} > 0\right)$. A projection of P^0 on $\mathbf{R}_{S_l}$, $P^0_{S_l}$ is an equilibrium for this first order kinetics with the balance $\sum_{i \in S_l} p_i = \sum_{i \in S_l} p_i^0$ because $\frac{p_i^0}{p_i^*} = \frac{p_j^0}{p_j^*}$ for $i, j \in S_l$.

The vectors γ^{ij} that generate $\mathbf{Q}(P^0, P^*)$ $(\sigma(j) = \sigma(i) + 1)$ (20) have non-zero projections on $\mathbf{R}_{S_l}$ if and only if either $l = \sigma(j) = \sigma(i) + 1$ or $\sigma(j) = \sigma(i) + 1 = l + 1$. In the first case, $l = \sigma(j) = \sigma(i) + 1$, vector γ^{ij} is the standard basis vector e_j in $\mathbf{R}_{S_l}$. In the second case, $\sigma(j) = \sigma(i) + 1 = l + 1$, we have $\gamma^{ij} = -e_i$. If $l = 1$ then only the second case is possible, and if $l = k + 1$ then only the first case can take place.

Let $V_l = \{P \in \mathbf{R}_{S_l} | p_i \geq 0, \sum_{i \in S_l} p_i < 1\}$. The projection of the pyramid $\mathcal{Q}(P^0)$ onto $\mathbf{R}_{S_l}$ is $\mathrm{conv}(V_l - V_l)$ if $1 < l < k + 1$; it is V_l if $l = k + 1$ and $-V_l$ if $l = 1$. (For sets X, Y, the sum $X + Y$ is the set of all sums $x + y$ $(x \in X, y \in Y)$, the difference $X - Y$ is the set of all differences $x - y$, therefore $V - V$ is not $\{0\}$ if V includes more than one element.)

The set V_l is positively invariant with respect to the first order kinetics in $\mathbf{R}_{S_l}$. Therefore, the following sets are also positively invariant with respect to the first order kinetics in $\mathbf{R}_{S_l}$ with equilibrium $P^0_{S_l}$ for every $a > 0$:

$$P^0_{S_l} + aV_l, \ (P^0_{S_l} + aV_l) - aV_l, \ \mathrm{conv}((P^0_{S_l} + aV_l) - aV_l).$$

Thus, the set $P^0 + a\mathcal{Q}(P^0)$ is positively invariant with respect to any first order kinetics with transitions $A_i \rightleftharpoons A_j$ $(i, j \in S_l)$ and equilibrium P^* for any $l = 1, \dots, k + 1$. A combination of these statements for all $l = 1, \dots, k + 1$ finalizes the proof. □

This proposition finalizes the justification of the use of the cone of the tangent directions $\mathbf{Q}(P^0, P^*)$ in the definition of the local minimum of the Markov order (4).

3. Equivalence of the maxima of all entropies and the Markov order approaches

The cone $\mathbf{Q}(P^0, P^*)$ is a piecewise constant function of P^0: it is the same for all P^0 from one compartment $\mathcal{C}_\sigma$ and, hence, depends on σ only. Therefore, if the condition of the local minimum (4) holds for one $P^0 \in L \cap \mathcal{C}_\sigma$ then it holds also for all elements of $L \cap \mathcal{C}_\sigma$. There is a finite number of compartments $\mathcal{C}_\sigma$.

Let the linear manifold of conditions L be given by the values of moments $\sum_i m_{ri}p_i = M_r$, $L^0 = \ker m$ and $L \cap \Delta_+^{n-1} \neq \emptyset$. The set of all conditional local minima of the Markov order on the linear manifold of conditions L is

$$\boxed{\bigcup \left\{L \cap \mathcal{C}_\sigma \middle| L \cap \mathcal{C}_\sigma \neq \emptyset \text{ and } L^0 \cap Q_\sigma = \{0\}\right\},} \tag{21}$$

where $\mathcal{C}_\sigma$ and Q_σ are defined by (15). It is sufficient to find all σ such that $L \cap \mathcal{C}_\sigma \neq \emptyset$ and $L^0 \cap Q_\sigma = \{0\}$ and then describe the union of the compartments $\mathcal{C}_\sigma$ for these σ.

The approach based on the minimization of all f-divergencies seems to be very different. For all monotonically increasing functions g we have to solve the equations for the Lagrange multipliers and represent the probability distribution in the form (3). Nevertheless, these approaches are equivalent and describe the same set of the "conditionally maximally disordered distributions".

Theorem 1. *A positive distribution $P^0 \in L$ satisfies the local conditional minimum conditions of the Markov order* (4) *if and only if there exists a strictly monotonic function g on $\mathbb{R}$ with $\operatorname{im} g = (0, \infty)$ such that the conditions* (3) *hold for some Lagrange multipliers and for $p_i^0 = p_i$.*

This means that every conditionally minimal distribution of the Markov order on the linear manifold $L \cap \Delta_+^{n-1}$ is a conditional minimum on $L \cap \Delta_+^{n-1}$ of a strictly convex f-divergence (1).

Proof. Due to the classical theorems about separation of convex sets and linear spaces by linear functionals [30], a distribution P^0 satisfies the condition of the local minimum (4) if and only if there exists a linear functional $\psi(P) = \sum_i \psi_i p_i$ such that $\psi|_L = \psi(P^0) = const$ and $\psi(P) > \psi(P^0)$ for every $P \in P^0 + \mathbf{Q}(P^0, P^*)$ if $P \neq P^0$. In other words, $\psi|_{L^0} \equiv 0$ and

$$(\psi_i - \psi_j)\left(\frac{p_i^0}{p_i^*} - \frac{p_j^0}{p_j^*}\right) > 0 \quad \text{if } \frac{p_i^0}{p_i^*} \neq \frac{p_j^0}{p_j^*} \tag{22}$$

according to the definition of $\mathbf{Q}(P, P^*)$ (13). Condition $\psi|_{L^0} \equiv 0$ is equivalent to the existence of the coefficients λ_r such that for all i

$$\psi_i = \sum_r \lambda_r m_{ri}.$$

Condition (22) is equivalent to the existence of a strictly monotonic function $\eta(x)$ defined for $x \geq 0$ such that

$$\psi_i = \eta\left(\frac{p_i^0}{p_i^*}\right).$$

To find such a function $\eta(x)$ we can take the known values ψ_i for $x = p_i^0/p_i^*$ and then use, for example, linear interpolation $\eta(x)$ between p_i^0/p_i^*. To extrapolate $\eta(x)$ from $\max\{p_i^0/p_i^*\}$ to $+\infty$ we can use an increasing linear function. To extrapolate $\eta(x)$ on the interval $(0, \min\{p_i^0/p_i^*\})$ we can use $\varepsilon \log x + const$.

Finally, we can take $h'(x) = \eta(x)$, $h(x) = \int \eta(\xi)\,\mathrm{d}\xi$; and $g(y)$ is the inverse function: $g(\eta(x)) = x$ for $x \geq 0$. The distribution P^0 is the local minimum of $H_h(P \parallel P^*)$ on L.

Conversely, if P^0 is a minimum of a strictly convex Lyapunov function H on L and $\mathrm{d}H/\mathrm{d}t|_{P^0} < 0$ for every Markov chain with equilibrium P^* for which P^0 is a non-equilibrium distribution then we can take

$$\psi_i = -\left.\frac{\partial H}{\partial p_i}\right|_{P^0}.$$

This choice of ψ_i provides (22) (because H is strictly decreasing in time Lyapunov function) and $\psi|_{L^0} \equiv 0$ because grad$H$ is orthogonal to L (the condition of local minimum). □

This equivalence of two definitions of the maximally uncertain distribution under given conditions has several important consequences.

Let us introduce the notion of the (global) *Markov order* [21].

- If for distributions P^0 and P^1 there exists such a Markov process with equilibrium P^* that for the solution of the Kolmogorov equation with $P(0) = P^0$ we have $P(1) = P^1$ then we say that P^0 and P^1 are connected by the Markov preorder [21] with equilibrium P^* and use notation $P^0 \succ_{P^*}^0 P^1$.
- The (global) Markov order is the closed transitive closure of the Markov preorder. For the Markov order with equilibrium P^* we use notation $P^0 \succ_{P^*} P^1$.

The *local Markov order* at point P^0 is just a vector order generated by the tangent cone $\mathbf{Q}(P^0, P^*)$ [21]. We use for this local order the notation $>_{P^0,P*}$:

$$P >_{P^0,P^*} P' \quad \text{if } P' - P \in \mathbf{Q}(P^0, P^*).$$

The proofs of Propositions 1 and 2 give us the possibility to use the relation $P^0 >_{P^0,P^*} P^1$ instead of the Markov preorder for the definition of the Markov order minimizers on linear manifolds. The relation $P^0 >_{P^0,P^*} P^1$ is defined by the local Markov order in a vicinity of P^0:

$$P^1 - P^0 \in \mathbf{Q}(P^0, P^*).$$

The cone $\mathbf{Q}(P^0, P^*)$ depends on P^0, therefore, the relation $P^0 >_{P^0,P^*} P^1$ is antisymmetric locally, in a vicinity of P^0.

Remark 1. It is possible to generate the Markov order by the relation $P^0 >_{P^0,P^*} P^1$. Let us specify the vicinity of P^0 where this relation is defined and introduce a new relation: $P^0 >^0_{P^*} P^1$ if $P^0 >_{P^0,P^*} P^1$ for all $i, j = 1, \dots, n$ and

$$\left(\frac{p_i^0}{p_i^*} - \frac{p_j^0}{p_j^*}\right)\left(\frac{p_i^1}{p_i^*} - \frac{p_j^1}{p_j^*}\right) \geq 0.$$

This condition means that the pairs of numbers $\left(\frac{p_i^0}{p_i^*}, \frac{p_j^0}{p_j^*}\right)$ and $\left(\frac{p_i^1}{p_i^*}, \frac{p_j^1}{p_j^*}\right)$ cannot have an opposite order on the real line. The closed transitive closure of the relation $P^0 >^0_{P^*} P^1$ is the Markov order $P^0 \succ_{P^*} P^1$.

Let L be a linear manifold in the space of distributions. By definition, $P^0 \in L$ is a minimal point on $L \cap \Delta_+^{n-1}$ with respect to the order $\succ_{P^*}$ if and only if there is no point $P^1 \in L \cap \Delta_+^{n-1}$, $P^1 \neq P^0$ such that $P^0 \succ_{P^*} P^1$.

Corollary 1. *$P^0 \in L \cap \Delta_+^{n-1}$ is a minimal point on $L \cap \Delta_+^{n-1}$ with respect to the (global) Markov order if and only if it satisfies the local minimum condition (4).*

Proof. If $P^0 \in L \cap \Delta_+^{n-1}$ is a minimal point on $L \cap \Delta_+^{n-1}$ with respect to the (global) Markov order then it satisfies the condition (4) due to the definition of the Markov order through the transitive closure of the relation $P^0 \succ^0_{P^*} P^1$ and Propositions 1 and 2.

Let P^0 satisfy the local minimum condition (4). Then there exists a divergence $H_h(P \parallel P^*)$ with strictly convex $h(x)$ $(x \geq 0)$ such that P^0 is a local minimum of $H_h(P \parallel P^*)$ on L. Because of strong convexity, this local minimum is a global one. $H_h(P \parallel P^*)$ is a Lyapunov function for all Markov chains with equilibrium P^*. Therefore, a broken line, which is combined from solutions of the Kolmogorov equations for such Markov chains and starts at P^0, leaves a small vicinity of P^0 (Propositions 1 and 2) and never returns in a sufficiently small vicinity of L. Thus, for the closed transitive closure of the relation $P \succ^0_{P^*} P'$, point P^0 is a minimal point on L. □

Of course, there may be infinitely many minimal points of the Markov order on L and each of them corresponds to a different Lyapunov functions $H_h(P \parallel P^*)$.

Another remarkable order on the space of distributions is $P^0 >_{H,P^*} P^1$ if for all strictly convex functions $h(x)$ $(x \geq 0)$

$$H_h(P^0 \parallel P^*) > H_h(P^1 \parallel P^*),$$

that is, P^1 is closer to equilibrium than P^0 with respect to *all* divergencies $H_h(P \parallel P^*)$.

Corollary 2. *For any linear manifold L in the distribution space the minimal elements of the Markov order $\succ_{P^*}$ on $L \cap \Delta_+^{n-1}$ coincide with the minimal elements of the order $>_{H,P^*}$ on $L \cap \Delta_+^{n-1}$.*

Proof. We just have to combine Theorem 1 with Corollary 1. □

Thus, the minimal elements of the orders $\succ_{P^*}$ and $>_{H,P^*}$ on the linear manifolds coincide. Nevertheless, it is necessary to mention the difference between these orders. Let P^0 be a distribution. For $>_{H,P^*}$ the set of distributions $\{P | P^0 >_{H,P^*} P\}$ is convex as an intersection of convex sets $\{P | H_h(P^0 \parallel P^*) > H_h(P \parallel P^*)\}$ for various strictly convex h. This is not the case for the Markov order. The set of distributions $\{P | P^0 \succ_{P^*} P\}$ may be non-convex. The examples may be extracted from the papers [28,31] (see Fig. 4).

Corollary 3. *Let $P^0 \succ_{P^*} P^1$. Then $P^1 \in P^0 + \mathbf{Q}(P^0, P^*)$.*

Proof. Let us apply Corollary 1 to all support hyperplanes L of the convex set $(P^0 + \mathbf{Q}(P^0, P^*))$ for which $(P^0 + \mathbf{Q}(P^0, P^*)) \cap L = \{P^0\}$. □

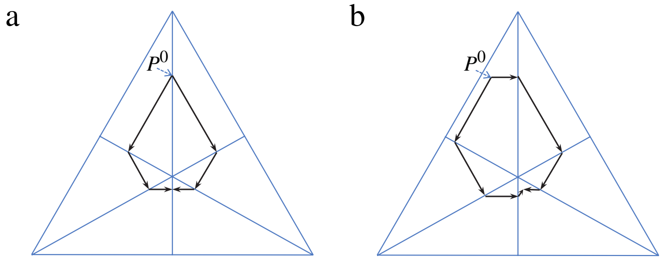

Fig. 4. The set $\{P|P^0 >_{H,P^*}\}$ for different P^0 and for the Markov chains with three states and equilibrium ($p_i^* = 1/3$): (a) $\{P|P^0 >_{H,P^*}\}$ is convex, (b) it is not convex. The border of the set $\{P|P^0 >_{H,P^*}\}$ is highlighted by bold lines. The arrows on these lines correspond to the directions of the extreme rays of the cones $\mathbf{Q}(P, P^*)$ (i.e. the angles represented in Fig. 2).

4. Example: generalization of the normal distribution

In this section, we discuss distributions $p(x)$ on a continuous space of states, the non-negative real semi-axis, $\mathbb{R}_+ = \{x|x \geq 0\}$. We have in mind two classical examples of distributions of the quantity bounded from below: energy (physics) and wealth (economics and microeconomics).

Let two moments be fixed, the total probability $M_0 = \int_0^\infty p(x)\,\mathrm{d}x$ and the average quantity $M_1 = \int_0^\infty xp(x)\,\mathrm{d}x$. The conditional maximization of the classical Boltzmann–Gibbs–Shannon entropy gives:

$$\int p(x)(\ln p(x) - 1)\,\mathrm{d}x \to \min \quad \text{for given } \int_0^\infty p(x) = M_0,\ \int_0^\infty xp(x)\mathrm{d}x = M_1;$$

$$\ln p(x) = \lambda_0 + \lambda_1 x, \qquad p(x) = \exp(\lambda_0 + \lambda_1 x), \quad \exp\lambda_0 = -\lambda_1 \ \exp\lambda_0 = M_1\lambda_1^2;$$

$$\boxed{p^*(x) = \frac{1}{M_1}\exp\left(-\frac{x}{M_1}\right).} \tag{23}$$

This Boltzmann distribution appears always as a first candidate for the equilibrium distribution of an additive conserved quantity bounded from below. Khinchin (1943) clearly explained this law as a version of the limit theorem [32].

Technically, it is not difficult to involve the higher moments and obtain the distribution of the form

$$p(x) = \exp(\lambda_0 + \lambda_1 x + \lambda_2 x^2 + \cdots + \lambda_r x^r). \tag{24}$$

One can expect that this extension of the set of moments may improve the description. This is a traditional belief in Extended Irreversible Thermodynamics (EIT) [7].

There may be many different approaches to evaluation of the quality of the approximation (24) but at least one important property of these functions is wrong: the asymptotic behavior at large x is $p(x) \asymp \exp(-const \times x^r)$. These "super-light" tails of the distribution $p(x)$ change qualitatively with the change of the order r in (24).

If we use, for example, the "regularizing" forth moment in the moment chain for the Boltzmann equation [33] then we corrupt the $\mathrm{e}^{-const\times v^2}$ tails of the Maxwell distribution. Therefore, other approaches which do not modify the tails of the distribution qualitatively (like [8]) may be more appreciated.

The asymptotic behavior of the distribution's tails was thoroughly studied in many cases. Very often, the tails of the distributions are, without a doubt, heavier than normal $\mathrm{e}^{-const\times x^2}$ and definitely are not cut as $\mathrm{e}^{-const\times x^4}$. For example, it is demonstrated that the distribution of money between people has the exponential tail with a possible transformation into a heavier power tail for very rich people [34].

The general solution (3) with the Boltzmann equilibrium (23) gives the following expression instead of (24)

$$p(x) = g(\lambda_0 + \lambda_1 x + \lambda_2 x^2 + \cdots + \lambda_r x^r)\frac{1}{M_1}\exp\left(-\frac{x}{M_1}\right),$$

where g is a monotonically increasing function. In particular, for the moments M_0, M_1 and M_2 we obtain

$$p(x) = g(\lambda_0 + \lambda_1 x + \lambda_2 x^2)\frac{1}{M_1}\exp\left(-\frac{x}{M_1}\right). \tag{25}$$

There are four qualitatively different cases of (25). Let $\lambda_2 \neq 0$ and $\mu = -\frac{\lambda_1}{2\lambda_2}$. Then

$$\boxed{p(x) = \frac{f(x)}{M_1}\exp\left(-\frac{x}{M_1}\right),} \tag{26}$$

and

1. if $\mu \le 0$ and $\lambda_2 > 0$ then $f(x)$ is a monotonically increasing function on $[0, \infty)$;
2. if $\mu \le 0$ and $\lambda_2 < 0$ then $f(x)$ is a monotonically decreasing function on $[0, \infty)$;
3. if $\mu > 0$ and $\lambda_2 > 0$ then $f(x)$ is a monotonically increasing function on $[\mu, \infty)$ and $f(x) = f(2\mu - x)$ for $x \in [0, \mu]$;
4. if $\mu > 0$ and $\lambda_2 < 0$ then $f(x)$ is a monotonically decreasing function on $[\mu, \infty)$ and $f(x) = f(2\mu - x)$ for $x \in [0, \mu]$.

Each of these "generalized normal distributions" (26) is a minimizer of the corresponding f-divergence. For the construction of such a divergence in general case, it is convenient to define the convex functions h in (1) with values on an extended real line with additional possible value $+\infty$. This is a natural general definition of convex functions [30]. In case 1 ($\mu \le 0$, $\lambda_2 > 0$, and f increases), we can take in (25), (26) without loss of generality $\mu = 0, f(x) = g(x^2)$, and $g(y) = f\left(\sqrt{y}\right)$. The monotonically increasing function $g(y)$ is, therefore, defined on $[0, \infty)$ with the set of values $[\underline{g}, \overline{g})$, where $\underline{g} = f(0) \ge 0$, $\overline{g} = \lim_{x\to\infty} f(x) > 0$ and the upper limit may be finite or infinite. The inverse function $\xi(z)$ is defined for $z \in [\underline{g}, \overline{g})$ with the interval of values $[0, \infty)$. Let us take

$$h'(z) = \begin{cases} 0 & \text{if } z < \underline{g}; \\ \xi(z) & \text{if } z \in [\underline{g}, \overline{g}); \\ \infty & \text{if } z \ge \overline{g}; \end{cases} \qquad h(z) = \begin{cases} 0 & \text{if } z < \underline{g}; \\ \displaystyle\int_{\underline{g}}^{z} \xi(\varsigma)\mathrm{d}\varsigma & \text{if } z \in [\underline{g}, \overline{g}]; \\ \infty & \text{if } z > \overline{g}. \end{cases} \tag{27}$$

The improper integral $\int_{\underline{g}}^{\overline{g}} \xi(\varsigma)\mathrm{d}\varsigma$ may take finite or infinite values.

Similarly, in case 2 ($\mu \le 0$ and $\lambda_2 < 0, f$ decreases) we define $g(y) = f\left(\sqrt{-y}\right)$ for $y \in (-\infty, 0]$. The function $g(y)$ monotonically increases and takes values on $(\underline{g}, \overline{g}]$, where $\underline{g} = \lim_{x\to\infty} f(x)$ and $\overline{g} = f(0)$. The inverse function $\xi(z)$ is defined for $z \in (\underline{g}, \overline{g}]$ with the interval of values $(-\infty, 0]$. In this case, we can take

$$h'(z) = \begin{cases} -\infty & \text{if } z < \underline{g}; \\ \xi(z) & \text{if } z \in (\underline{g}, \overline{g}]; \\ 0 & \text{if } z > \overline{g}; \end{cases} \qquad h(z) = \begin{cases} \infty & \text{if } z \le \underline{g}; \\ -\displaystyle\int_{z}^{\overline{g}} \xi(\varsigma)\mathrm{d}\varsigma & \text{if } z \in [\underline{g}, \overline{g}]; \\ 0 & \text{if } z > \overline{g}. \end{cases} \tag{28}$$

In case 3, the construction is almost the same as for the case 1 but $f(x) = g((x-\mu)^2)$ and $g(y) = f\left(\sqrt{y} + \mu\right)$. In this case, $g(y)$ is a monotonically increasing function defined on the interval $[0, \infty)$ with the set of values $[\underline{g}, \overline{g})$, where $\underline{g} = f(\mu)$ and $\overline{g} = \lim_{x\to\infty} f(x)$. Similarly, for case 4, the construction of $h(z)$ is almost the same as in case 2.

Thus, for every distribution in the form (26) we can find a f-divergence $H_h(P \parallel P^*)$, which conditional minimization produces this distribution. For example, if in (26) $f(x) = ax^\beta$ then we can take h in the form $h(z) = \frac{\beta}{\beta+2}(z/a)^{1+2/\beta}$.

5. Conclusion

The Maxallent approach aims to bring some order to the modern anarchy of the measures of disorder. If there is no clear idea which entropy is better then we have to use all of them together.

The Markov order approach was also proposed as an alternative to the entropic anarchism. It is based on the idea that the disorder has to increase in random processes with given equilibrium distribution, which is considered as the maximally disordered state. Here, we have proved that these two approaches produce the same conditional minimizers on the planes of given values of moments (Theorem 1).

In this paper, we have considered several relations between positive distributions:

1. $P^0 \succ^0_{P^*} P^1$ if there exists a Markov chain with equilibrium P^* such that for the solution of the Kolmogorov equation $P(t)$ with $P(0) = P^0$ we have $P(1) = P^1$;
2. $P^0 \succ_{P^*} P^1$ if there exist integrable bounded functions $q_{ij}(t)$ $(i, j = 1, \dots, n, i \neq j, t \ge 0)$ such that $q_{ij}(t)$ satisfy the balance condition (6) for given P^* ($p_i^* > 0$) (for all $t \ge 0$), and $P(1) = P^1$ for solution $P(t)$ of the equations
$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = \sum_{j,\, j\neq i} (q_{ij}(t)p_j - q_{ji}(t)p_i) \quad (i = 1, \dots, n)$$
with $P(0) = P^0$ (that is, $\succ_{P^*}$ is the transitive closure of $\succ^0_{P^*}$);
3. $P^0 >_{H,P^*} P^1$ if $H_h(P^0 \parallel P^*) > H_h(P^1 \parallel P^*)$ for all strictly convex functions $h(x)$ on a semi-axis $x \ge 0$.
4. $P^0 >_{P^0,P^*} P^1$ if $P^1 - P^0 \in \mathbf{Q}(P^0, P^*)$, where $\mathbf{Q}(P^0, P^*)$ is the cone of possible velocities $\mathrm{d}P/\mathrm{d}t$ (13) at point P^0 for all Markov chains with equilibrium P^*.

All these relations are different. Three of them are antisymmetric, and one, $P^0 >_{P^0,P^*} P^1$, is locally antisymmetric, in a vicinity of P^0. Their interrelations are described by the follows implications:

$$(P^0 \succ^0_{P^*} P^1) \Rightarrow (P^0 \succ_{P^*} P^1) \Rightarrow (P^0 >_{H,P^*} P^1) \Rightarrow (P^0 >_{P^0,P^*} P^1).$$

The local Markov order $P^0 >_{P*} P^1$ is the weakest and the connection by a solution of the Kolmogorov equation $P^0 \succ^0_{P*} P^1$ is the strongest of these relations. Nevertheless, locally, in a small vicinity of a positive non-equilibrium distribution P^0, these relations coincide and they define the same set of locally minimal distributions on a linear manifold of conditions L (Propositions 1, 2, Theorem 1, Corollaries 1–3).

Of course, there is the other, the classical way to reduce the variability of the measures of disorder. The divergences $H(P \parallel P^*)$ can be defined by their main properties. This is an axiomatic approach: we postulate some "natural properties" of the divergence, then find the divergences with these properties, evaluate the result and decide whether we have to change the system of axiom or not. The axiomatic approach to definition of entropy was used by Shannon [5] and elaborated in detail by Khinchin [35].

Two distinguished additivity properties are important for the Maxent reasoning:

- Additivity on the algebra of states: $H(P \parallel P^*)$ is a sum in states

$$H(P \parallel P^*) = \sum_i \eta(p_i, p_i^*).$$

- Additivity with respect to the joining of independent subsystems. This means that if P and P^* are products of distributions then $H(P \parallel P^*)$ is the sum of the corresponding entropies: if $P = (p_{jl}) = (q_j r_l)$ and $P^* = (p^*_{jl}) = (q^*_j r^*_l)$ then $H(P \parallel P^*) = H(Q \parallel Q^*) + H(R \parallel R^*)$.

The first additivity property implies that the restriction of the Maxent distribution on a subset of the event space Ω is also a Maxent distribution if the condition functionals are also additive on the algebra of states. The second additivity property implies that the Maxent distribution is a product of the Maxent distributions of subsystems if the condition functionals are additive with respect to the joining of subsystems and the equilibrium is a product of distributions. For more details we refer, for example, to the review in [21].

If we join the first additivity property with the requirement that the divergence should be a Lyapunov function for all Markov chains with equilibrium P^* then we get $H_h(P \parallel P^*)$ of the form (1) [19–21]. If we add the second additivity property and require continuity of $H_h(P \parallel P^*)$ for all values of P (including vectors with some $p_i = 0$) then the classical Boltzmann–Gibbs–Shannon relative entropy will be the only possibility (that is, $H_h(P \parallel P^*)$ with $h(x) = x \ln x$ up to unimportant constant factors and summand). If we relax the requirement of the continuity to the set of strictly positive distributions then we will get the one-parametric family $H_h(P \parallel P^*)$ with $h(x) = \beta x \ln x - (1-\beta) \ln x$ [19,21].

Let us accept the point of view that the divergency is an order. Then the values are not important and all the divergencies connected by a monotonic transformation of a scale, $H = f(H')$ (with a monotonically increasing f), are equivalent. If the first additivity property is valid in one scale, and the second may be valid in another one, then one more one-parametric family appear, the Cressie–Read divergences (see Appendix A) [19,21]. The Tsallis entropy is a particular case of them. The Boltzmann–Gibbs–Shannon relative entropy (or the Kullback–Leibler entropy, which is the same), the convex combination of $H_h(P \parallel P^*)$ and $H_h(P^* \parallel P)$ for $h(x) = x \ln x$, and the Cressie–Read divergences (including the Tsallis relative entropy) form the "entropic aristocracy" distinguished mostly by the additivity properties.

If we accept the additivity on the algebra of states (i.e., the trace form) and the additivity with respect to joining of independent subsystems, both, then we have to use some of these functions. If additivity with respect to joining of independent subsystems seems to be too restrictive then we have to take the wider class of divergencies, for example, $H_h(P \parallel P^*)$ of the form (1). If we reject the requirement of the trace form then the variety of the admissible divergences becomes even richer. This uncertainty in the choice of divergence forces us to use the Maxallent approach.

The Maxallent approach produces a set of conditionally maximally disordered distributions instead of a single distribution that maximizes a selected distinguished entropy in the usual Maxent method. These Maxallent sets of distributions may be considered as probabilistic analogues of the type-2 fuzzy sets introduced by L. Zadeh [36] to capture the uncertainty of the fuzzy systems. The Maxallent approach is invented to manage the uncertainty of the measures of uncertainty. If there is no uncertainty of uncertainty then the set of distributions reduces to a single distribution.

The decomposition theorems for Markov chains provide us with tools for the efficient calculation of the Markov order. Following [27], we compare the general Markov chains and the reversible chains with detailed balance. For any general chain there is a reversible chain with the same velocity vector at a given point. The classes of general and reversible chains locally coincide because they have the same cone of possible velocities at every non-equilibrium distribution (the second decomposition theorem, Appendix B). This theorem gives us the possibility to describe the set of the conditionally maximally uncertain distributions combinatorially, in the finite form (21).

For the classical Boltzmann–Gibbs–Shannon entropy the distribution on $\mathbb{R}_+$ with two given moments has the Gaussian form $a \exp(-b(x-c)^2)$. The class of the Maxallent distributions on $\mathbb{R}_+$ with two given moments is also simple (26) but much richer. It can be produced by multiplication of the Boltzmann distribution (23) by a monotonic function or unimodal function (with one local maximum) or by a function with one local minimum.

There exists an attractive possibility: if a distribution can be obtained in the Maxallent approach then it is a conditional minimum of a divergence. If we find or guess a distribution of the Maxallent type for an empirical system then we can restore the divergence and then use it in the standard Maxent reasoning.

The Maxallent approach is, surprisingly, efficient enough to analyze some practical problems. It gives an answer that does not depend on the subjective choice and, therefore, returns us to the "mission" of information theory: *"to eliminate* *the psychological factors involved . . .*" [4]. At the same time, it has a solid basis in the theory of Lyapunov functions for the Kolmogorov equations.

Now, essential mathematical work on the basic notion of entropy is needed. Gromov suggests that the natural mathematical language for this work will involve nonstandard analysis and category theory [37]. These abstract languages seem to be closer to the basic intuition than the set theory of Cantor and the $\varepsilon - \delta$ reasoning of the classical analysis. Nevertheless, the basic idea of Maxallent is so simple and natural, that it should persist in the future advanced theory of entropy: *order is something that decreases in Markov processes*.

Appendix A. The most popular examples of $H_h(P \parallel P^*)$

The most popular examples of $H_h(P \parallel P^*)$ are [21]:

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

$$H_h(P \parallel P^*) = -\sum_{i,\ p_i > 0} 1. \tag{29}$$

The quantity $-H_h$ is the number of non-zero probabilities p_i and does not depend on P^*. Sometimes it is called the Hartley entropy.

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

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

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

3. $h = x \ln x$,

$$H_h(P \parallel P^*) = \sum_i p_i \ln\left(\frac{p_i}{p_i^*}\right) = D_{\mathrm{KL}}(P \parallel P^*); \tag{30}$$

this is the usual Kullback–Leibler divergence or the relative Boltzmann–Gibbs–Shannon (BGS) entropy;

4. $h = -\ln x$,

$$H_h(P \parallel P^*) = -\sum_i p_i^* \ln\left(\frac{p_i}{p_i^*}\right) = D_{\mathrm{KL}}(P^* \parallel P); \tag{31}$$

this is the relative Burg entropy. It is obvious that 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: $h = \beta x \ln x - (1 - \beta) \ln x$ ($\beta \in [0, 1]$),

$$H_h(P \parallel P^*) = \beta D_{\mathrm{KL}}(P \parallel P^*) + (1 - \beta) D_{\mathrm{KL}}(P^* \parallel P); \tag{32}$$

this convex combination of divergences was used by Gorban in the early 1980s [38] and studied further by Gorban and Karlin [39]. It becomes a symmetric functional of (P, P^*) for $\beta = 1/2$. There exists a special name for this case, "Jeffreys' entropy".

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

$$H_h(P \parallel P^*) = \frac{1}{2} \sum_i \frac{(p_i - p_i^*)^2}{p_i^*} = H_2(P \parallel P^*); \tag{33}$$

this is the quadratic term in the Taylor expansion of the relative Boltzmann–Gibbs–Shannon entropy, $D_{\mathrm{KL}}(P \parallel P^*)$, near equilibrium. We have used its time derivative in (8).

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

$$H_h(P \parallel P^*) = \frac{1}{\lambda(\lambda + 1)} \sum_i p_i \left[\left(\frac{p_i}{p_i^*}\right)^\lambda - 1\right]; \tag{34}$$

this is the Cressie–Read (CR) family of power divergences [40] (the modern exposition of the history, properties and applications of these entropies is presented in [41]). For this family we use the notation $H_{CR\,\lambda}$. If $\lambda \to 0$ then $H_{CR\,\lambda} \to D_{\mathrm{KL}}(P \parallel P^*)$, this is the classical BGS relative entropy; if $\lambda \to -1$ then $H_{CR\,\lambda} \to D_{\mathrm{KL}}(P^* \parallel P)$, this is the relative Burg entropy.

8. For the CR family in the limits $\lambda \to \pm\infty$ only the maximal terms "survive". Exactly as we get the limit l^∞ of l^p norms for $p \to \infty$, we can use $(\lambda(\lambda+1)H_{CR\,\lambda})^{1/|\lambda|}$ for $\lambda \to \pm\infty$ and write in these limits:

$$H_{\mathrm{CR}\,\infty}(P \parallel P^*) = \max_i \left\{ \frac{p_i}{p_i^*} \right\} - 1; \tag{35}$$

$$H_{\mathrm{CR}\,-\infty}(P \parallel P^*) = \max_i \left\{ \frac{p_i^*}{p_i} \right\} - 1. \tag{36}$$

The existence of two limiting divergences $H_{\mathrm{CR}\,\pm\infty}$ seems very natural: there may be two types of extremely non-equilibrium states: with a high excess of current probability p_i above p_i^* and, inversely, with an extremely small current probability p_i with respect to p_i^*.

9. The Tsallis relative entropy [42]: $h = \frac{(x^\alpha - x)}{\alpha - 1}$, $\alpha > 0$,

$$H_h(P \parallel P^*) = \frac{1}{\alpha - 1} \sum_i p_i \left[\left(\frac{p_i}{p_i^*} \right)^{\alpha-1} - 1 \right]. \tag{37}$$

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

Appendix B. The decomposition theorems

The first decomposition theorem. *Every Markov chain with a positive equilibrium is a conic combination of simple cycles with the same equilibrium.*

Proof. If a non-zero Markov chain has a positive equilibrium then it cannot be acyclic: there exists at least one oriented cycle of transitions with non-zero rate constants. The length of this cycle can vary from 2 to n. The set of all Markov chains with a positive equilibrium P^* is an intersection of a linear subspace given by the balance equations (6) with the positive orthant $\mathbb{R}_+^{n(n-1)}$. This is a polyhedral cone which does not include a whole straight line. It is well known in convex geometry that every such polyhedral cone is a convex hull of a finite number of its *extreme rays* [30]. A ray l with direction vector $x \neq 0$ is a set $l = \{\kappa x\}$ ($\kappa \geq 0$). By definition, it is an extreme ray of a cone $\mathbf{Q}$ if for any $u \in l$ and any $x, y \in \mathbf{Q}$, whenever $u = (x+y)/2$, we must have $x, y \in l$.

Any extreme ray of the cone of Markov chains with equilibrium P^* is a simple cycle $A_{i_1} \to \cdots \to A_{i_k} \to A_{i_1}$ with rate constants $q_{i_{j+1}i_j} = \kappa/p_j^*$. Indeed, let a non-zero Markov chain Q with coefficients q_{ij} belong to an extreme ray of this cone. This chain includes a simple cycle with non-zero coefficients, $A_{i_1} \to \cdots \to A_{i_k} \to A_{i_1}$ ($k \leq n$, all the numbers $i_1, \ldots, i_k$ are different, $q_{i_{j+1}\,i_j} > 0$ for $j = 1, \ldots, k$, and $i_{k+1} = i_1$). For sufficiently small κ ($0 < \kappa < \kappa_0$), $q_{i_{j+1}\,i_j} - \frac{\kappa}{p^*_{i_j}} > 0$ ($j = 1, \ldots, k$). Let Q_κ be the same simple cycle with the rate constants $q_{i_{j+1}i_j} = \kappa/p_j^*$. Then for $0 < \kappa < \kappa_0$ vectors $Q \pm Q_\kappa$ also represent Markov chains with the equilibrium P^*. Obviously, $Q = \frac{(Q+Q_\kappa)+(Q-Q_\kappa)}{2}$, hence, Q should be proportional to Q_κ, by the definition of extreme rays.

So, any Markov chain with a positive equilibrium P^* is a linear combination with positive coefficients of the cycles with the same equilibrium. This decomposition is global, it does not depend on the current distribution P. □

The second decomposition theorem. *For every Markov chain with a positive equilibrium P^* and any probability distribution P^0 the vector $\mathrm{d}P/\mathrm{d}t|_{P^0}$ is a conic combination of the vectors $\mathrm{d}P/\mathrm{d}t|_{P^0}$ for the simple cycles of length two $A_i \rightleftharpoons A_j$ with the same equilibrium.*

Proof. Let us start from a simple cycle $A_1 \to A_2 \to \cdots \to A_n \to A_1$ with the constants $q_{i+1i} = 1/p_i^*$, where $p_i^* > 0$ is the equilibrium. At a non-equilibrium distribution P the right hand side of Eq. (5) is the vector $\mathrm{d}P/\mathrm{d}t = \mathbf{v}_n$ with coordinates

$$\frac{\mathrm{d}p_j}{\mathrm{d}t} = (\mathbf{v}_n)_j = \frac{p_{j-1}}{p_{j-1}^*} - \frac{p_j}{p_j^*}. \tag{38}$$

The flux $A_j \to A_{j+1}$ is p_j/p_j^*. Let us find A_j with the minimum value of this flux and, for convenience, let us put this A_j in the first position by a cyclic permutation. We will represent the right hand side vector $\mathbf{v}_n$ in the form

$$\mathbf{v}_n = \mathbf{v}_{n-1} + \kappa \mathbf{v}_2,$$

where $\mathbf{v}_{n-1}$ corresponds to the cycle of the length $n-1$, $A_2 \to \cdots A_n \to A_2$, with the rate constants $q_{i+1i} = 1/p_i^*$ (and the cyclic convention $n+1 = 2$), $\mathbf{v}_2$ corresponds to the cycle of the length 2, $A_1 \rightleftharpoons A_2$, with the rate constants $q_{21} = 1/p_1^*$, $q_{12} = 1/p_2^*$, and $\kappa \geq 0$. Both velocities $\mathbf{v}_{n-1}$ and $\mathbf{v}_2$ should be calculated for the same distribution P.

We find the constant κ from the conditions: $\mathbf{v}_n = \mathbf{v}_{n-1} + \kappa \mathbf{v}_2$ at the point P, hence, the two following reaction schemes, (a) and (b), should have the same velocities, $\mathrm{d}P/\mathrm{d}t$:

$$\text{(a)}\ A_n \xrightarrow{1/p_n^*} A_1 \xrightarrow{1/p_1^*} A_2 \quad \text{and} \quad \text{(b)}\ A_n \xrightarrow{1/p_n^*} A_2;\ A_1 \underset{\kappa/p_2^*}{\overset{\kappa/p_1^*}{\rightleftharpoons}} A_2.$$

From this condition,

$$\kappa = \left(\frac{p_n}{p_n^*} - \frac{p_1}{p_1^*}\right)\left(\frac{p_2}{p_2^*} - \frac{p_1}{p_1^*}\right)^{-1}.$$

The inequality $\kappa \geq 0$ holds because p_1/p_1^* is the minimal value of the flux p_j/p_j^*.

We just delete the vertex with the smallest outgoing flux from the initial cycle of length n and add a cycle of the length 2 with the same equilibrium. Let us repeat this operation for the remaining cycle of length $n-1$, and so on. At the end, the left hand side vector $\mathbf{v}_n$ will be represented as the combination with positive coefficients the vectors $\mathrm{d}P/\mathrm{d}t$ for the cycles of length 2, $A_i \rightleftharpoons A_j$ with the same equilibrium. This is the system with detailed balance. We have to stress here that the set of these transitions and the coefficients κ depend on the current distribution P.

For every distribution P, the velocity $\mathrm{d}P/\mathrm{d}t$ of every cycle with equilibrium P^* is a combination with positive coefficients of the velocities for some cycles of the length two $A_i \rightleftharpoons A_j$ with the same equilibrium. Therefore, the right hand side of the Kolmogorov equation for any Markov chain with equilibrium P^* also allows such a decomposition.

It is necessary to stress that the decomposition of the right hand side of the Kolmogorov equation (5) into a conic combination of cycles of length 2 depends on the ordering of the ratios p_i/p_i^* and cannot be performed for all values of P simultaneously. □

For more details and further references see [27].

References

[1] R. Clausius, Über vershiedene für die Anwendungen bequeme Formen der Hauptgleichungen der Wärmetheorie, Poggendorffs Annalen der Physic und Chemie 125 (1865) 353–400.
[2] L. Boltzmann, Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen, Sitzungsberichte der keiserlichen Akademie der Wissenschaften in Wien 66 (1872) 275–370;
Translation, S.G. Brush, N.S. Hall (Eds.), Further studies on the thermal equilibrium of gas molecules, in: Kinetic Theory of Gases: An Anthology of Classic Papers With Historical Commentary, Imperial Colledge Press, London, UK, 2003, pp. 362–368.
[3] J.W. Gibbs, On the equilibrium of heterogeneous substance, Transactions of the Connecticut Academy (1875–1876) 108–248. (1877–1878) 343–524.
[4] R.V.L. Hartley, Transmission of information, Bell System Technical Journal 7 (1928) 535–563.
[5] C.E. Shannon, A mathematical theory of communication, Bell System Technical Journal 27 (1948) 379–423. 623–656.
[6] E.T. Jaynes, Information theory and statistical mechanics, I; II, Physical Review 106 (1957) 620–630. 108, 171–190.
[7] D. Jou, J. Casas-Vázquez, G. Lebon, Extended Irreversible Thermodynamics, Springer, Berlin–Heidelberg–New York, 2001.
[8] A.N. Gorban, I.V. Karlin, Quasi-equilibrium closure hierarchies for the Boltzmann equation, Physica A 360 (2006) 325–364. arXiv:cond-mat/0305599 [cond-mat.stat-mech].
[9] I. Csiszár, Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems, The Annals of Statistics 19 (4) (1991) 2032–2066.
[10] S.F. Gull, Bayesian inductive inference and maximum entropy, in: G.J. Erickson, C.R. Smith (Eds.), Maximum Entropy and Bayesian Methods in Science, Vol. 1, Foundations, Kluwer, Dordrecht, 1988, pp. 53–74.
[11] A. Golan, G.G. Judge, D. Miller, Maximum Entropy Econometrics: Robust Estimation With Limited Data, Wiley, N.Y, 1996.
[12] G.G. Judge, R.C. Mittelhammer, An Information Theoretic Approach to Econometrics, Cambridge University Press, 2011.
[13] A. Rényi, On measures of entropy and information, in: Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability 1960, Vol. 1, University of California Press, Berkeley, CA, 1961, pp. 547–561.
[14] I. Csiszár, Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten, Publications of the Mathematical Institute of Hungarian Academy of Sciences 8 (1963) 85–108.
[15] T. Morimoto, Markov processes and the H-theorem, Journal of the Physical Society of Japan 12 (1963) 328–331.
[16] J.P. Burg, Maximum entropy spectral analysis, in: D.G. Childers (Ed.), Proceedings of the 37th Meeting of the Society of Exploration Geophysicists; Oklahoma City, OK, USA, 1967, IEEE Press, New York, 1978, pp. 34–39. Reprinted in Modern Spectrum Analysis.
[17] A.N. Gorban, D. Packwood, Allowed and forbidden regimes of entropy balance in lattice Boltzmann collisions, Physical Review E 86 (2012) 025701(R). arXiv:1111.5994 [physics.comp-ph].
[18] L. Boltzmann, Neuer Beweis zweier Sätze über das Wärmegleichgewicht unter mehratomigen Gasmolekülen, Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften in Wien 95 (2) (1887) 153–164.
[19] P. Gorban, Monotonically equivalent entropies and solution of additivity equation, Physica A 328 (2003) 380–390. arXiv:cond-mat/0304131 [cond-mat.stat-mech].
[20] S.-I. Amari, Divergence, Optimization, Geometry, in: C.S. Leung, M. Lee, J.H. Chan (Eds.), Proceedings of the 16th International Conference on Neural Information Processing, in: LNCS, vol. 5863, Springer, Berlin, Germany, 2009, pp. 185–193.
[21] A.N. Gorban, P.A. Gorban, G. Judge, Entropy: the Markov ordering approach, Entropy 12 (5) (2010) 1145–1193. arXiv:1003.1377 [physics.data-an].
[22] J.E. Cohen, Y. Derriennic, G.H. Zbaganu, Majorization, monotonicity of relative entropy and stochastic matrices, Contemporary Mathematics 149 (1993) 251–259.
[23] J.E. Cohen, Y. Iwasa, G. Rautu, M.B. Ruskai, E. Seneta, G. Zbaganu, Relative entropy under mappings by stochastic matrices, Linear Algebra and its Applications 179 (1993) 211–235.
[24] H.-A. Lorentz, Über das Gleichgewicht der lebendigen Kraft unter Gasmolekülen, Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften in Wien 95 (2) (1887) 115–152.
[25] A.N. Gorban, M. Shahzad, The Michaelis–Menten–Stueckelberg theorem, Entropy 13 (5) (2011) 966–1019. arXiv:1008.3296 [physics.chem-ph].
[26] S.L. Kalpazidou, Cycle Representations of Markov Processes, in: Series: Applications of Mathematics, vol. 28, Springer, New York, 2006.
[27] A.N. Gorban, Local equivalence of reversible and general Markov kinetics, Physica A, available online: 23-NOV-2012. http://dx.doi.org/10.1016/j.physa.2012.11.028. arXiv:1205.2052 [physics.chem-ph].
[28] A.N. Gorban, Invariant sets for kinetic equations, Reaction Kinetics and Catalysis Letters 10 (1979) 187–190.
[29] K. Kuratowski, Topology, Vol. 1, Academic Press, New York, 1966.
[30] R.T. Rockafellar, Convex Analysis, Princeton Univ. Press, Princeton, NJ, 1997.
[31] Ch. Zylka, A note on the attainability of states by equalizing processes, Theoretica Chimica Acta 68 (1985) 363–377.
[32] A.I. Khinchin, Mathematical Foundations of Statistical Mechanics, Dover Publications, 1960.
[33] C.D. Levermore, Moment closure hierarchies for kinetic theories, Journal of Statistical Physics 83 (1996) 1021–1065.
[34] V.M. Yakovenko, J.B. Rosser Jr., Colloquium: statistical mechanics of money, wealth, and income, Reviews of Modern Physics 81 (2009) 1703–1725.
[35] A.I. Khinchin, Mathematical Foundations of Information Theory, Dover, New York, 1957.

[36] L.A. Zadeh, The Concept of a linguistic variable and its application to approximate reasoning 1, Information Sciences 8 (1975) 199–249.
[37] M. Gromov, In a search for a structure, in: Proceedings of 6th European Congress of Mathematics, Kraków 2012, the European Mathematical Society Publishing House, 2013 (in press). IHES e-print, In a Search for a Structure, Part 1: On Entropy, 2012. http://www.ihes.fr/~gromov/PDF/structre-serch-entropy-july5-2012.pdf.
[38] A.N. Gorban, Equilibrium Encircling. Equations of Chemical Kinetics and Their Thermodynamic Analysis, Nauka, Novosibirsk, 1984.
[39] A.N. Gorban, I.V. Karlin, Family of additive entropy functions out of thermodynamic limit, Physical Review E 67 (2003) 016104. arXiv:cond-mat/0205511 [cond-mat.stat-mech].
[40] N. Cressie, T. Read, Multinomial goodness of fit tests, Journal of the Royal Statistical Society. Series B 46 (1984) 440–464.
[41] A. Cichocki, S.-I. Amari, Families of alpha- beta- and gamma- divergences: flexible and robust measures of similarities, Entropy 12 (6) (2010) 1532–1568.
[42] C. Tsallis, Possible generalization of Boltzmann–Gibbs statistics, Journal of Statistical Physics 52 (1) (1988) 479–487.