*Entropy* **2010**, *12*, 1145-1193; doi:10.3390/e12051145

OPEN ACCESS

*entropy*

**ISSN 1099-4300**

www.mdpi.com/journal/entropy

*Article*

# Entropy: The Markov Ordering Approach

**Alexander N. Gorban [1,⋆], Pavel A. Gorban [2] and George Judge [3]**

[1]Department of Mathematics, University of Leicester, Leicester, UK

[2]Institute of Space and Information Technologies, Siberian Federal University, Krasnoyarsk, Russia

[3]Department of Resource Economics, University of California, Berkeley, CA, USA

⋆ Author to whom correspondence should be addressed; E-mail: ag153@le.ac.uk.

*Received: 1 March 2010; in revised form: 30 April 2010 / Accepted: 4 May 2010 / Published: 7 May 2010 / Corrected Postprint 9 November 2013*

---

**Abstract:** The focus of this article is on entropy and Markov processes. We study the properties of functionals which are invariant with respect to monotonic transformations and analyze two invariant "additivity" properties: (i) existence of a monotonic transformation which makes the functional additive with respect to the joining of independent systems and (ii) existence of a monotonic transformation which makes the functional additive with respect to the partitioning of the space of states. All Lyapunov functionals for Markov chains which have properties (i) and (ii) are derived. We describe the most general ordering of the distribution space, with respect to which all continuous-time Markov processes are monotonic (the *Markov order*). The solution differs significantly from the ordering given by the inequality of entropy growth. For inference, this approach results in a convex compact set of conditionally "most random" distributions.

**Keywords:** Markov process; Lyapunov function; entropy functionals; attainable region; MaxEnt; inference

---

## 1. Introduction

### *1.1. A Bit of History: Classical Entropy*

Two functions, energy and entropy, rule the Universe.

In 1865 R. Clausius formulated two main laws [1]:

1. The energy of the Universe is constant.

2. The entropy of the Universe tends to a maximum.

The universe is isolated. For non-isolated systems energy and entropy can enter and leave, the change in energy is equal to its income minus its outcome, and the change in entropy is equal to entropy production inside the system plus its income minus outcome. The entropy production is always positive.

Entropy was born as a daughter of energy. If a body gets heat $\Delta Q$ at the temperature $T$ then for this body $\mathrm{d}S = \Delta Q/T$. The total entropy is the sum of entropies of all bodies. Heat goes from hot to cold bodies, and the total change of entropy is always positive.

Ten years later J.W. Gibbs [2] developed a general theory of equilibrium of complex media based on the entropy maximum: the equilibrium is the point of the conditional entropy maximum under given values of conserved quantities. The entropy maximum principle was applied to many physical and chemical problems. At the same time J.W. Gibbs mentioned that entropy maximizers under a given energy are energy minimizers under a given entropy.

The classical expression $\int p \ln p$ became famous in 1872 when L. Boltzmann proved his $H$-theorem [3]: the function

$$H = \int f(x,v) \ln f(x,v) \mathrm{d}x \mathrm{d}v$$

decreases in time for isolated gas which satisfies the Boltzmann equation (here $f(x,v)$ is the distribution density of particles in phase space, $x$ is the position of a particle, $v$ is velocity). The statistical entropy was born: $S = -kH$. This was the one-particle entropy of a many-particle system (gas).

In 1902, J.W. Gibbs published a book "Elementary principles in statistical dynamics" [4]. He considered ensembles in the many-particle phase space with probability density $\rho(p_1, q_1, \ldots p_n, q_n)$, where $p_i, q_i$ are the momentum and coordinate of the $i$th particle. For this distribution,

$$S = -k \int \rho(p_1, q_1, \ldots p_n, q_n) \ln(\rho(p_1, q_1, \ldots p_n, q_n)) \mathrm{d}q_1 \ldots \mathrm{d}q_n \mathrm{d}p_1 \ldots \mathrm{d}p_n \quad (1)$$

Gibbs introduced the canonical distribution that provides the entropy maximum for a given expectation of energy and gave rise to the entropy maximum principle (MaxEnt).

The Boltzmann period of history was carefully studied [5]. The difference between the Boltzmann entropy which is defined for coarse-grained distribution and increases in time due to gas dynamics, and the Gibbs entropy, which is constant due to dynamics, was analyzed by many authors [6,7]. Recently, the idea of two functions, energy and entropy which rule the Universe was implemented as a basis of two-generator formalism of nonequilibrium thermodynamics [8,9].

In information theory, R.V.L. Hartley (1928) [10] 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". He defined information in a text of length $n$ in alphabet of s symbols as $H = n \log s$.

In 1948, C.E. Shannon [11] generalized the Hartley approach and developed "a mathematical theory of communication", that is information theory. He measured information, choice and uncertainty by the entropy:

$$S = -\sum_{i=1}^{n} p_i \log p_i \quad (2)$$

Here, $p_i$ are the probabilities of a full set of $n$ events ($\sum_{i=1}^{n} p_i = 1$). The quantity $S$ is used to measure of how much "choice" is involved in the selection of the event or of how uncertain we are of the outcome. Shannon mentioned that this quantity form will be recognized as that of entropy, as defined in certain formulations of statistical mechanics. The classical entropy (1), (2) was called the Boltzmann–Gibbs–Shannon entropy (BGS entropy). (In 1948, Shannon used the *concave* function (2), but under the same notation $H$ as for the Boltzmann *convex* function. Here we use $H$ for the convex $H$-function, and $S$ for the concave entropy.)

In 1951, S. Kullback and R.A. Leibler [12] supplemented the BGS entropy by the relative BGS entropy, or the Kullback–Leibler divergence between the current distribution $P$ and some "base" (or "reference") distribution $Q$:

$$D_{\mathrm{KL}}(P\|Q) = \sum_i p_i \log \frac{p_i}{q_i} \tag{3}$$

The Kullback–Leibler divergence is always non-negative $D_{\mathrm{KL}}(P\|Q) \geq 0$ (the Gibbs inequality). It is not widely known that this "distance" has a very clear physical interpretation. This function has been well known in physical thermodynamics since 19th century under different name. If $Q$ is an equilibrium distribution at the same temperature as $P$ has, then

$$D_{\mathrm{KL}}(P\|Q) = \frac{F(P) - F(Q)}{kT} \tag{4}$$

where $F$ is free energy and $T$ is thermodynamic temperature. In physics, $F = U - TS$, where physical entropy $S$ includes an additional multiplier $k$, the Boltzmann constant. The thermodynamic potential $-F/T$ has its own name, Massieu function. Let us demonstrate this interpretation of the Kullback–Leibler divergence. The equilibrium distribution $Q$ provides the conditional entropy (2) maximum under a given expectation of energy $\sum_i u_i q_i = U$ and the normalization condition $\sum_i q_i = 1$. With the Lagrange multipliers $\mu_U$ and $\mu_0$ we get the equilibrium Boltzmann distribution:

$$q_i = \exp(-\mu_0 - \mu_U u_i) = \frac{\exp(-\mu_U u_i)}{\sum_i \exp(-\mu_U u_i)} \tag{5}$$

The Lagrange multiplier $\mu_U$ is in physics (by definition) $1/kT$, so $S(Q) = \mu_0 + \frac{U}{kT}$, hence, $\mu_0 = -\frac{F(Q)}{kT}$. For the Kullback–Leibler divergence this formula gives (4).

After the classical work of Zeldovich (1938, reprinted in 1996 [13]), the expression for free energy in the "Kullback–Leibler form"

$$F = kT \sum_i c_i \left( \ln \left( \frac{c_i}{c_i^*(T)} \right) - 1 \right)$$

where $c_i$ is concentration and $c_i^*(T)$ is the equilibrium concentration of the $i$th component, is recognized as a useful instrument for the analysis of kinetic equations (especially in chemical kinetics [14,15]).

Each given positive distribution $Q$ could be represented as an equilibrium Boltzmann distribution for given $T > 0$ if we take $u_i = -kT \log q_i + u_0$ for an arbitrary constant level $u_0$. If we change the order of arguments in the Kullback–Leibler divergence then we get the relative Burg entropy [16,17]. It has a much more exotic physical interpretation: for a current distribution $P$ we can define the "auxiliary energy" functional $U_P$ for which $P$ is the equilibrium distribution under a given temperature $T$. We can

calculate the auxiliary free energy of any distribution $Q$ and this auxiliary energy functional: $F_P(Q)$. (Up to an additive constant, for $P = P^*$ this $F_P(Q)$ turns into the classical free energy, $F_P^*(Q) = F(Q)$.) In particular, we can calculate the auxiliary free energy of the physical equilibrium, $F_P(P^*)$. The relative Burg entropy is

$$D_{\mathrm{KL}}(P^*\|P) = \frac{F_P(P^*) - F_P(P)}{kT}$$

This functional should also decrease in any Markov process with given equilibrium $P^*$.

Information theory developed by Shannon and his successors focused on entropy as a measure of uncertainty of subjective choice. This understanding of entropy was returned from information theory to statistical mechanics by E.T. Jaynes as a basis of "subjective" statistical mechanics [18,19]. He followed Wigner's idea "entropy is an antropocentric concept". The entropy maximum approach was declared as a minimization of the subjective uncertainty. This approach gave rise to a MaxEnt "anarchism". It is based on a methodological hypothesis that everything unknown could be estimated by the principle of the entropy maximum under the condition of fixed known quantities. At this point the classicism in entropy development changed to a sort of scientific modernism. The art of model fitting based on entropy maximization was developed [20]. The principle of the entropy maximum was applied to plenty of problems: from many physical problems [21], chemical kinetics and process engineering [15] to econometrics [22,23] and psychology [24]. Many new entropies were invented and now one has rich choice of entropies for fitting needs [25]. The most celebrated of them are the Rényi entropy [26], the Burg entropy [16,17], the Tsallis entropy [27,28] and the Cressie–Read family [29,30]. The nonlinear generalized averaging operations and generalized entropy maximization procedures were suggested [31].

Following this impressive stream of works we understand the MaxEnt approach as conditional maximization of entropy for the evaluation of the probability distribution when our information is partial and incomplete. The entropy function may be the classical BGS entropy or any function from the rich family of non-classical entropies. This rich choice causes a new problem: which entropy is better for a given class of applications?

The MaxEnt "anarchism" was criticized many times as a "senseless fitting". Arguments pro and contra the MaxEnt approach with non-classical entropies (mostly the Tsallis entropy [27]) were collected by Cho [32]. This sometimes "messy and confusing situation regarding entropy-related studies has provided opportunities for us: clearly there are still many very interesting studies to pursue" [33].

### *1.2. Key Points*

In this paper we do not pretend to invent new entropies. (There appear new functions as limiting cases of the known entropy families, but this is not our main goal). Entropy is understood in this paper as a measure of uncertainty which increases in Markov processes. In our paper we consider a Markov process as a semigroup on the space of positive probability distributions. The state space is finite. Generalizations to compact state spaces are simple. We analyze existent relative entropies (divergences) using several simple ideas:

1. In Markov processes probability distributions $P(t)$ monotonically approach equilibrium $P^*$: divergence $D(P(t)\|P^*)$ monotonically decrease in time.

2. In most applications, conditional minimizers and maximizers of entropies and divergences are used, but the values are not. This means that the system of level sets is more important than the functions' values. Hence, most of the important properties are invariant with respect to monotonic transformations of entropy scale.

3. The system of level sets should be the same as for additive functions: after some rescaling the divergences of interest should be additive with respect to the joining of statistically independent systems.

4. The system of level sets should after some rescaling the divergences of interest should have the form of a sum (or integral) over states $\sum_i f(p_i, p_i^*)$, where the function $f$ is the same for all states. In information theory, divergences of such form are called *separable*, in physics the term *trace–form functions* is used

The first requirement means that if a distribution becomes more random then it becomes closer to equilibrium (Markov process decreases the information excess over equilibrium). For example, classical entropy increases in all Markov processes with uniform equilibrium distributions. This is why we may say that the distribution with higher entropy is more random, and why we use entropy conditional maximization for the evaluation of the probability distribution when our information is partial and incomplete.

It is worth to mention that some of the popular Bregman divergences, for example, the squared Euclidean distance or the Itakura–Saito divergence, do not satisfy the first requirement (see Section 4.3).

The second idea is just a very general methodological thesis: to evaluate an instrument (a divergence) we have to look how it works (produces conditional minimizers and maximizers). The properties of the instrument which are not related to its work are not important. The number three allows to separate variables if the system consists of independent subsystems, the number four relates to separation of variables for partitions of the space of probability distributions.

Amongst a rich world of relative entropies and divergences, only two families meet these requirements. Both were proposed in 1984. The Cressie–Read (CR) family [29,30]:

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

and the convex combination of the Burg and Shannon relative entropies proposed in [34] and further analyzed in [35,36]:

$$H(P\|P^*) = \sum_i (\beta p_i - (1-\beta) p_i^*) \log \left( \frac{p_i}{p_i^*} \right) \ , \ \ \beta \in [0,1]$$

When $\lambda \to 0$ the CR divergence tends to the KL divergence (the relative Shannon entropy) and when $\lambda \to -1$ it turns into the Burg relative entropy. The Tsallis entropy was introduced four years later [27] and became very popular in thermostatistics (there are thousands of works that use or analyze this entropy [37]). The Tsallis entropy coincides (up to a constant multiplier $\lambda + 1$) with the CR entropy for $\lambda \in ]-1, \infty[$ and there is no need to study it separately (see discussion in Section 2.2).

A new problem arose: which entropy is better for a specific problem? Many authors compare performance of different entropies and metrics for various problems (see, for example, [39,40]). In any case study it may be possible to choose "the best" entropy but in general we have no sufficient reasons for such a choice. We propose a possible way to avoid the choice of the best entropy.

Let us return to the idea: the distribution $Q$ is more random than $P$ if there exists a continuous-time Markov process (with given equilibrium distribution $P^*$) that transforms $P$ into $Q$. We say in this case that $P$ and $Q$ are connected by the *Markov preorder* with equilibrium $P^*$ and use notation $P \succ^0_{P^*} Q$. The *Markov order* $\succ_{P^*}$ is the transitive closure of the Markov preorder.

If a priori information gives us a set of possible distributions $W$ then the conditionally "maximally random distributions" (the "distributions without additional information", the "most indefinite distributions" in $W$) should be the minimal elements in $W$ with respect to Markov order. If a Markov process (with equilibrium $P^*$) starts at such a minimal element $P$ then it cannot produce any other distribution from $W$ because all distributions which are more random that $P$ are situated outside $W$. In this approach, the maximally random distributions under given a priori information may be not unique. Such distributions form a set which plays the same role as the standard MaxEnt distribution. For the moment based a priori information the set $W$ is an intersection of a linear manifold with the simplex of probability distributions, the set of minimal elements in $W$ is also polyhedron and its description is available in explicit form. In low-dimensional case it is much simpler to construct this polyhedron than to find the MaxEnt distributions for most of specific entropies.

### *1.3. Structure of the Paper*

The paper is organized as follows. In Section 2 we describe the known non-classical divergences (relative entropies) which are the Lyapunov functions for the Markov processes. We discuss the general construction and the most popular families of such functions. We pay special attention to the situations, when different divergences define the same order on distributions and provide the same solutions of the MaxEnt problems (Section 2.2). In two short technical Sections 2.3 and 2.4 we present normalization and symmetrization of divergences (similar discussion of these operations was published very recently [38].

The divergence between the current distribution and equilibrium should decrease due to Markov processes. Moreover, divergence between any two distributions should also decrease (the generalized data processing Lemma, Section 3).

Definition of entropy by its properties is discussed in Section 4. Various approaches to this definition were developed for the BGS entropy by Shannon [11], [41] and by other authors for the Rényi entropy [43,44], the Tsallis entropy [42], the CR entropy and the convex combination of the BGS and Burg entropies [46]. Csiszár [45] axiomatically characterized the class of Csiszár–Morimoto divergences (formula (6) below). Another characterization of this class was proved in [46] (see Lemma 1, Section 4.3 below).

From the celebrated properties of entropy [47] we selected the following three:

1. Entropy should be a Lyapunov function for continuous-time Markov processes;

2. Entropy is additive with respect to the joining of independent systems;

3. Entropy is additive with respect to the partitioning of the space of states (*i.e.*, has the *trace–form*).

To solve the MaxEnt problem we have to find the maximizers of entropy (minimizers of the relative entropy) under given conditions. For this purpose, we have to know the sublevel sets of entropy, but not its values. We consider entropies with the same system of sublevel sets as equivalent ones (Section 2.2). From this point of view, all important properties have to be invariant with respect to monotonic transformations of the entropy scale. Two last properties from the list have to be substituted by the following:

2'. There exists a monotonic transformation which makes entropy additive with respect to the joining of independent systems (Section 4.2);

3'. There exists a monotonic transformation which makes entropy additive with respect to the partitioning of the space of states (Section 4.1).

Several "No More Entropies" Theorems are proven in Section 4.3: if an entropy has properties 1, 2' and 3' then it belongs to one of the following one-parametric families: to the Cressie–Read family, or to a convex combination of the classical BGS entropy and the Burg entropy (may be, after a monotonic transformation of scale).

It seems very natural to consider divergences as orders on distribution spaces (Section 5.1), the sublevel sets are the lower cones of this orders. For several functions, $H_1(P), \dots, H_k(P)$ the sets $\{Q \mid H_i(P) > H_i(Q) \text{ for all } i\}$ give the simple generalization of the sublevel sets. In Section 5 we discuss the more general orders in which continuous time Markov processes are monotone, define the Markov order and fully characterize the local Markov order. The Markov chains with detailed balance define the Markov order for general Markov chains (Section 5.2). It is surprising that there is no necessity to consider other Markov chains for the order characterization (Section 5.2) because no reversibility is assumed in this analysis.

In Section 6.1 we demonstrate how is it possible to use the Markov order to reduce the uncertainty in the standard settings when a priori information is given about values of some moments. Approaches to construction of the most random distributions are presented in Section 6.2.

Various approaches for the definition of the reference distribution (or the generalized canonical distribution) are compared in Section 7.

In Conclusion we briefly discuss the main results.

## 2. Non-Classical Entropies

### *2.1. The Most Popular Divergences*

Csiszár–Morimoto Functions $H_h$

During the time of modernism plenty of new entropies were proposed. Esteban and Morales [25] attempted to systemize many of them in an impressive table. Nevertheless, there are relatively few

entropies in use now. Most of the relative entropies have the form proposed independently in 1963 by I. Csiszar [49] and T. Morimoto [48]:

$$H_h(p) = H_h(P\|P^*) = \sum_i p_i^* h\left(\frac{p_i}{p_i^*}\right) \tag{6}$$

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

These relative entropies are the Lyapunov functions for all Markov chains with equilibrium $P^* = (p_i^*)$. Moreover, they have the relative entropy contraction property given by the generalized data processing lemma (Section 3.2 below).

For $h(x) = x\log x$ this function coincides with the Kullback–Leibler divergence from the current distribution $p_i$ to the equilibrium $p_i^*$. Some practically important functions $h$ have singularities at 0. For example, if we take $h(x) = -\log x$, then the correspondent $H_h$ is the relative Burg entropy $H_h = -\sum_i p_i^* \log(p_i/p_i^*) \to \infty$ for $p_i \to 0$.

Required Properties of the Function $h(x)$

Formally, $h(x)$ is an extended real-valued proper convex function on the closed positive real half-line $[0, \infty[$. An *extended real-valued function* can take real values and infinite values $\pm\infty$. A *proper function* has at least one finite value. An extended real valued function on a convex set $U$ is called *convex* if its *epigraph*

$$\mathrm{epi}(h) = \{(x, y) \mid x > 0, y \geq h(x)\}$$

is a convex set [50]. For a proper function this definition is equivalent to the *Jensen inequality*

$$h(ax + (1-a)y) \leq ah(x) + (1-a)h(y) \quad \text{for all} \quad x, y \in U,\ a \in [0,1]$$

It is assumed that the function $h(x)$ takes finite values on the open positive real half-line $]0, \infty[$ but the value at point $x = 0$ may be infinite. For example, functions $h(x) = -\log x$ or $h(x) = x^{-\alpha}$ ($\alpha > 0$) are allowed. A convex function $h(x)$ with finite values on the open positive real half-line is continuous on $]0, \infty[$ but may have a discontinuity at $x = 0$. For example, the step function, $h(x) = 0$ if $x = 0$ and $h(x) = -1$ if $x > 0$, may be used.

A convex function is differentiable almost everywhere. Derivative of $h(x)$ is a monotonic function which has left and right limits at each point $x > 0$. An inequality holds: $h'(x)(y - x) \leq h(y) - h(x)$ (Jensen's inequality in the differential form). It is valid also for left and right limits of $h'$ at any point $x > 0$.

Not everywhere differentiable functions $h(x)$ are often used, for example, $h(x) = |x - 1|$. Nevertheless, it is convenient to consider the twice differentiable on $]0, \infty[$ functions $h(x)$ and to produce a non-smooth $h(x)$ (if necessary) as a limit of smooth convex functions. We use widely this possibility.

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,

$$H_h(P\|P^*) = -\sum_{i,\ p_i>0} 1 \tag{7}$$

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. $h = |x-1|$,
$$H_h(P\|P^*) = \sum_i |p_i - p_i^*|$$
this is the $l_1$-distance between $P$ and $P^*$.

3. $h = x\ln x$,
$$H_h(P\|P^*) = \sum_i p_i \ln\left(\frac{p_i}{p_i^*}\right) = D_{\mathrm{KL}}(P\|P^*) \tag{8}$$
this is the usual Kullback–Leibler divergence or the relative BGS entropy;

4. $h = -\ln x$,
$$H_h(P\|P^*) = -\sum_i p_i^* \ln\left(\frac{p_i}{p_i^*}\right) = D_{\mathrm{KL}}(P^*\|P) \tag{9}$$
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: $h = \beta x\ln x - (1-\beta)\ln x$ ($\beta \in [0,1]$),
$$H_h(P\|P^*) = \beta D_{\mathrm{KL}}(P\|P^*) + (1-\beta)D_{\mathrm{KL}}(P^*\|P) \tag{10}$$
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. $h = \frac{(x-1)^2}{2}$,
$$H_h(P\|P^*) = \frac{1}{2}\sum_i \frac{(p_i - p_i^*)^2}{p_i^*} \tag{11}$$
This is the quadratic term in the Taylor expansion of the relative Botzmann–Gibbs-Shannon entropy, $D_{\mathrm{KL}}(P\|P^*)$, near equilibrium. Sometimes, this quadratic form is called the Fisher entropy.

7. $h = \frac{x(x^\lambda - 1)}{\lambda(\lambda+1)}$,
$$H_h(P\|P^*) = \frac{1}{\lambda(\lambda+1)}\sum_i p_i\left[\left(\frac{p_i}{p_i^*}\right)^\lambda - 1\right] \tag{12}$$
This is the CR family of power divergences [29,30]. For this family we use notation $H_{\mathrm{CR}\,\lambda}$. If $\lambda \to 0$ then $H_{\mathrm{CR}\,\lambda} \to D_{\mathrm{KL}}(P\|P^*)$, this is the classical BGS relative entropy; if $\lambda \to -1$ then $H_{\mathrm{CR}\,\lambda} \to D_{\mathrm{KL}}(P^*\|P)$, 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 $l^\infty$ of $l^p$ norms for $p \to \infty$, we can use the root $(\lambda(\lambda+1)H_{\mathrm{CR}\,\lambda})^{1/|\lambda|}$ for $\lambda \to \pm\infty$ and write in these limits the divergences:
$$H_{\mathrm{CR}\,\infty}(P\|P^*) = \max_i\left\{\frac{p_i}{p_i^*}\right\} - 1 \tag{13}$$

$$H_{\mathrm{CR}\ -\infty}(P\|P^*) = \max_i \left\{\frac{p_i^*}{p_i}\right\} - 1 \tag{14}$$

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 [27] corresponds to the choice $h = \frac{(x^\alpha - x)}{\alpha - 1}$, $\alpha > 0$.

$$H_h(P\|P^*) = \frac{1}{\alpha - 1}\sum_i p_i \left[\left(\frac{p_i}{p_i^*}\right)^{\alpha-1} - 1\right] \tag{15}$$

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

Rényi Entropy

The Rényi entropy of order $\alpha > 0$ is defined [26] as

$$H_{\mathrm{R}\ \alpha}(P) = \frac{1}{1-\alpha}\log\left(\sum_{i=1}^n p_i^\alpha\right) \tag{16}$$

It is a concave function, and

$$H_{\mathrm{R}\ \alpha}(P) \to S(P)$$

when $\alpha \to 1$, where $S(P)$ is the Shannon entropy.

When $\alpha \to \infty$, the Rényi entropy has a limit $H_\infty(X) = -\log \max_{i=1,\dots n} p_i$, which has a special name "Min-entropy".

It is easy to get the expression for a relative Rényi entropy $H_{\mathrm{R}\ \alpha}(P\|P^*)$ from the requirement that it should be a Lyapunov function for any Markov chain with equilibrium $P^*$:

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

For the Min-entropy, the correspondent divergence (the relative Min-entropy) is

$$H_\infty(P\|P^*) = \log \max_{i=1,\dots n}\left(\frac{p_i}{p_i^*}\right)$$

It is obvious from (22) below that $\max_{i=1,\dots n}(p_i/p_i^*)$ is a Lyapunov function for any Markov chain with equilibrium $P^*$, hence, the relative Min-entropy is also the Lyapunov functional.

### *2.2. Entropy Level Sets*

A *level set* of a real-valued function $f$ is a set of the form :

$$\{x \mid f(x) = c\}$$

where $c$ is a constant (the "level"). It is the set where the function takes on a given constant value. A *sublevel set* of $f$ is a set of the form

$$\{x \mid f(x) \le c\}$$

A *superlevel set* of $f$ is given by the inequality with reverse sign:

$$\{x \mid f(x) \geq c\}$$

The intersection of the sublevel and the superlevel sets for a given value $c$ is the level set for this level.

In many applications, we do not need the entropy values, but rather the order of these values on the line. For any two distributions $P, Q$ we have to compare which one is closer to equilibrium $P^*$, *i.e.*, to answer the question: which of the following relations is true: $H(P\|P^*) > H(Q\|P^*)$, $H(P\|P^*) = H(Q\|P^*)$ or $H(P\|P^*) < H(Q\|P^*)$? To solve the MaxEnt problem we have to find the maximizers of entropy (or, in more general settings, the minimizers of the relative entropy) under given conditions. For this purpose, we have to know the sublevel sets, but not the values. All the MaxEnt approach does not need the values of the entropy but the sublevel sets are necessary.

Let us consider two functions, $\phi$ and $\psi$ on a set $U$. For any $V \subset U$ we can study conditional minimization problems $\phi(x) \to \min$ and $\psi(x) \to \min$, $x \in V$. The sets of minimizers for these two problems coincide for any $V \subset U$ if and only if the functions $\phi$ and $\psi$ have the same sets of sublevel sets. It should be stressed that here just the sets of sublevel sets have to coincide without any relation to values of level.

Let us compare the level sets for the Rényi, the Cressie-Read and the Tsallis families of divergences (for $\alpha - 1 = \lambda$ and for all values of $\alpha$). The values of these functions are different, but the level sets are the same (outside the Burg limit, where $\alpha \to 0$): for $\alpha \neq 0, 1$

$$H_{\mathrm{R}\ \alpha}(P\|P^*) = \frac{1}{\alpha - 1}\ln c;\ \ H_{\mathrm{CR}\ \alpha-1}(P\|P^*) = \frac{1}{\alpha(\alpha-1)}(c-1);\ \ H_{\mathrm{Ts}\ \alpha}(P\|P^*) = \frac{1}{\alpha-1}(c-1) \quad (17)$$

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

Beyond points $\alpha = 0, 1$

$$H_{\mathrm{CR}\ \alpha-1}(P\|P^*) = \frac{1}{\alpha(\alpha-1)}\exp((\alpha-1)H_{\mathrm{R}\ \alpha}(P\|P^*)) = \frac{1}{\alpha}H_{\mathrm{Ts}\ \alpha}(P\|P^*)$$

For $\alpha \to 1$ all these divergences turn into the Shannon relative entropy. Hence, if $\alpha \neq 0$ then for any $P$, $P^*$, $Q$, $Q^*$ the following equalities A, B, C are equivalent, A⇔B⇔C:

$$\begin{aligned} &\text{A.}\ \ H_{\mathrm{R}\ \alpha}(P\|P^*) = H_{\mathrm{R}\ \alpha}(Q\|Q^*) \\ &\text{B.}\ \ H_{\mathrm{CR}\ \alpha+1}(P\|P^*) = H_{\mathrm{CR}\ \alpha+1}(Q\|Q^*) \\ &\text{C.}\ \ H_{\mathrm{Ts}\ \alpha}(P\|P^*) = H_{\mathrm{Ts}\ \alpha}(Q\|Q^*) \end{aligned} \quad (18)$$

This equivalence means that we can select any of these three divergences as a basic function and consider the others as functions of this basic one.

For any $\alpha \geq 0$ and $\lambda = \alpha + 1$ the Rényi, the Cressie–Read and the Tsallis divergences have the same family of sublevel sets. Hence, they give the same maximizers to the conditional relative entropy minimization problems and there is no difference which entropy to use.

The CR family has a more convenient normalization factor $1/\lambda(\lambda+1)$ which gives a proper convexity for all powers, both positive and negative, and provides a sensible Burg limit for $\lambda \to -1$ (in contrary, when $\alpha \to 0$ both the Rényi and Tsallis entropies tend to 0).

When $\alpha < 0$ then for the Tsallis entropy function $h = \frac{(x^\alpha - x)}{\alpha - 1}$ loses convexity, whereas for the Cressie-Read family convexity persists for all values of $\lambda$. The Rényi entropy also loses convexity for $\alpha < 0$. Neither the Tsallis, nor the Rényi entropy were invented for use with negative $\alpha$.

There may be a reason: for $\alpha < 0$ the function $x^\alpha$ is defined for $x > 0$ only and has a singularity at $x = 0$. If we assume that the divergence should exist for all non-negative distributions, then the cases $\alpha \leq 0$ should be excluded. Nevertheless, the Burg entropy which is singular at zeros is practically important and has various attractive properties. The Jeffreys' entropy (the symmetrized Kullback–Leibler divergence) is also singular at zero, but has many important properties. We can conclude at this point that it is not obvious that we have to exclude any singularity at zero probability. It may be useful to consider positive probabilities instead ("nature abhors a vacuum").

Finally, for the MaxEnt approach (conditional minimization of the relative entropy), the Rényi and the Tsallis families of divergences ($\alpha > 0$) are particular cases of the Cressie–Read family because they give the same minimizers. For $\alpha \leq 0$ the Rényi and the Tsallis relative entropies lose their convexity, while the Cressie–Read family remains convex for $\lambda \leq -1$ too.

### *2.3. Minima and normalization*

For a given $P^*$, the function $H_h$ achieves its minimum on the hyperplane $\sum_i p_i = \sum_i p_i^* =$const at equilibrium $p_i^*$, because at this point

$$\mathrm{grad} H_h = (h'(1), \ldots h'(1)) = h'(1)\mathrm{grad}\left(\sum_i p_i\right)$$

The transformation $h(x) \to h(x) + ax + b$ just shifts $H_h$ by constant value: $H_h \to H_h + a\sum_i p_i + b = H_h + a + b$. Therefore, we can always assume that the function $h(x)$ achieves its minimal value at point $x = 1$, and this value is zero. For this purpose, one should just transform $h$:

$$h(x) := h(x) - h(1) - h'(1)(x - 1) \tag{19}$$

This normalization transforms $x \ln x$ into $x \ln x - (x - 1)$, $-\ln x$ into $-\ln x + (x - 1)$, and $x^\alpha$ into $x^\alpha - 1 - \alpha(x - 1)$. After normalization $H_h(P\|P^*) \geq 0$. If the normalized $h(x)$ is strictly positive outside point $x = 1$ ($h(x) > 0$ if $x \neq 1$) then $H_h(P\|P^*) = 0$ if and only if $P = P^*$ (*i.e.*, in equilibrium).

The normalized version of any divergence $H_h(P\|P^*)$ could be produced by the normalization transformation $h(x) := h(x) - h(1) - h'(1)(x - 1)$ and does not need separate discussion.

### *2.4. Symmetrization*

Another technical issue is symmetry of a divergence. If $h(x) = x \ln x$ then both $H_h(P\|P^*)$ (the KL divergence) and $H_h(P^*\|P)$ (the relative Burg entropy) are the Lyapunov functions for the Markov chains, and $H_h(P^*\|P) = H_g(P\|P^*)$ with $g(x) = -\ln x$. Analogously, for any $h(x)$ we can write $H_h(P^*\|P) = H_g(P\|P^*)$ with

$$g(x) = xh\left(\frac{1}{x}\right) \tag{20}$$

If $h(x)$ is convex on $\mathbf{R}_+$ then $g(x)$ is convex on $\mathbf{R}_+$ too because

$$g''(x) = \frac{1}{x^3} h''\left(\frac{1}{x}\right)$$

The transformation (20) is an involution:

$$xg\left(\frac{1}{x}\right) = h(x)$$

The fixed points of this involution are such functions $h(x)$ that $H_h(P\|P^*)$ is symmetric with respect to transpositions of $P$ and $P^*$. There are many such $h(x)$. An example of symmetric $H_h(P\|P^*)$ gives the choice $h(x) = -\sqrt{x}$:

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

After normalization, we get

$$h(x) := \frac{1}{2}(\sqrt{x} - 1)^2\,; \quad H_h(P\|P^*) = \frac{1}{2}\sum_i (\sqrt{p_i} - \sqrt{p_i^*})^2$$

Essentially (up to a constant addition and multiplier) this function coincides with a member of the CR family, $H_{\mathrm{CR}\ -\frac{1}{2}}$ (12), and with one of the Tsallis relative entropies $H_{\mathrm{Ts}\ \frac{1}{2}}$ (15). The involution (20) is a linear operator, hence, for any convex $h(x)$ we can produce its symmetrization:

$$h_{\mathrm{sym}}(x) = \frac{1}{2}(h(x) + g(x)) = \frac{1}{2}\left(h(x) + xh\left(\frac{1}{x}\right)\right)$$

For example, if $h(x) = x\log x$ then $h_{\mathrm{sym}}(x) = \frac{1}{2}(x\log x - \log x)$; if $h(x) = x^\alpha$ then $h_{\mathrm{sym}}(x) = \frac{1}{2}(x^\alpha + x^{1-\alpha})$.

## 3. Entropy Production and Relative Entropy Contraction

### *3.1. Lyapunov Functionals for Markov Chains*

Let us consider continuous time Markov chains with positive equilibrium probabilities $p_j^*$. The dynamics of the probability distribution $p_i$ satisfy the Master equation (the Kolmogorov equation):

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = \sum_{j,\, j\neq i} (q_{ij}p_j - q_{ji}p_i) \tag{21}$$

where coefficients $q_{ij}$ $(i \neq j)$ are non-negative. For chains with a positive equilibrium distribution $p_j^*$ another equivalent form is convenient:

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = \sum_{j,\, j\neq i} q_{ij}p_j^*\left(\frac{p_j}{p_j^*} - \frac{p_i}{p_i^*}\right) \tag{22}$$

where $p_i^*$ and $q_{ij}$ are connected by identity

$$\sum_{j,\, j\neq i} q_{ij}p_j^* = \left(\sum_{j,\, j\neq i} q_{ji}\right) p_i^* \tag{23}$$

The time derivative of the Csiszár–Morimoto function $H_h(p)$ (6) due to the Master equation is

$$\frac{\mathrm{d}H_h(P\|P^*)}{\mathrm{d}t} = \sum_{i,j,\,j\neq i} q_{ij}p_j^* \left[ h\left(\frac{p_i}{p_i^*}\right) - h\left(\frac{p_j}{p_j^*}\right) + h'\left(\frac{p_i}{p_i^*}\right)\left(\frac{p_j}{p_j^*} - \frac{p_i}{p_i^*}\right)\right] \leq 0 \tag{24}$$

To prove this formula, it is worth to mention that for any $n$ numbers $h_i$, $\sum_{i,j,\,j\neq i} q_{ij}p_j^*(h_i - h_j) = 0$. The last inequality holds because of the convexity of $h(x)$: $h'(x)(y-x) \leq h(y) - h(x)$ (Jensen's inequality).

Inversely, if

$$h(x) - h(y) + h'(y)(x-y) \leq 0 \tag{25}$$

for all positive $x, y$ then $h(x)$ is convex on $\mathbf{R}_+$. Therefore, if for some function $h(x)$ $H_h(p)$ is the Lyapunov function for all the Markov chains with equilibrium $P^*$ then $h(x)$ is convex on $\mathbf{R}_+$.

The Lyapunov functionals $H_h$ do not depend on the kinetic coefficients $q_{ij}$ directly. They depend on the equilibrium distribution $p^*$ which satisfies the identity (23). This independence of the kinetic coefficients is the *universality* property.

### *3.2. "Lyapunov Divergences" for Discrete Time Markov Chains*

The Csiszár–Morimoto functions (6) are also Lyapunov functions for discrete time Markov chains. Moreover, they can serve as a "Lyapunov distances" [51] between distributions which decreases due to time evolution (and not only the divergence between the current distribution and equilibrium). In more detail, let $A = (a_{ij})$ be a stochastic matrix in columns:

$$a_{ij} \geq 0, \quad \sum_i a_{ij} = 1 \text{ for all } j$$

The *ergodicity contraction coefficient* for $A$ is a number $\overline{\alpha}(A)$ [52,53]:

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

$0 \leq \overline{\alpha}(A) \leq 1$.

Let us consider in this subsection the normalized Csiszár–Morimoto divergences $H_h(P\|Q)$ (19): $h(1) = 0$, $h(x) \geq 0$.

**Theorem about relative entropy contraction.** *(The generalized data processing Lemma.) For each two probability positive distributions $P, Q$ the divergence $H_h(P\|Q)$ decreases under action of stochastic matrix $A$ [54,55]:*

$$H_h(AP\|AQ) \leq \overline{\alpha}(A)H_h(P\|Q) \tag{26}$$

The generalizations of this theorem for general Markov kernels seen as operators on spaces of probability measures was given by [56]. The shift in time for continuous-time Markov chain is a column-stochastic matrix, hence, this contraction theorem is also valid for continuous-time Markov chains.

The question about a converse theorem arises immediately. Let the contraction inequality hold for two pairs of positive distributions $(P, Q)$ and $(U, V)$ and for all $H_h$:

$$H_h(U\|V) \leq H_h(P\|Q) \tag{27}$$

Could we expect that there exists such a stochastic matrix $A$ that $U = AP$ and $V = AP$? The answer is positive:

**The converse generalized data processing lemma.** *Let the contraction inequality (27) hold for two pairs of positive distributions $(P, Q)$ and $(U, V)$ and for all normalized $H_h$. Then there exists such a column-stochastic matrix $A$ that $U = AP$ and $V = AQ$ [54].*

This means that for the system of inequalities (27) (for all normalized convex functions $h$ on $]0, \infty[$) is necessary and sufficient for existence of a (discrete time) Markov process which transform the pair of positive distributions $(P, Q)$ in $(U, V)$. It is easy to show that for continuous-time Markov chains this theorem is not valid: the attainable regions for them are strictly smaller than the set given by inequalities (27) and could be even non-convex (see [62] and Section 8.1 below).

## 4. Definition of Entropy by its Properties

### *4.1. Separation of Variables for Partition of the State Space*

An important property of separation of variables is valid for all divergences which have the form of a sum of convex functions $f(p_i, p_i^*)$. Let the set of states be divided into two subsets, $I_1$ and $I_2$, and let the functionals $u^1, \ldots u^m$ be linear. We represent each probability distribution as a direct sum $P = P^1 \oplus P^2$, where $P^{1,2}$ are restrictions of $P$ on $I_{1,2}$.

Let us consider the problem

$$H(P\|P^*) \to \min$$

subject to conditions $u^i(P) = U_i$ for a set of linear functionals $u^i(P)$.

The solution $P^{\min}$ to this problem has a form $P^{\min} = P_1^{\min} \oplus P_2^{\min}$, where $P^{1,2}$ are solutions to the problems

$$H(P^{1,2}\|P^{*\,1,2}) \to \min$$

subject to conditions $u^i(P_{1,2}) = U_i^{1,2}$ and $\sum_{i\in I_{1,2}} p_i^{1,2} = \pi_{1,2}$ for some redistribution of the linear functionals values, $U_i = U_i^1 + U_i^2$, and of the total probability, $1 = \pi_1 + \pi_2$ ($\pi_{1,2} \geq 0$) .

The solution to the divergence minimization problem is composed from solutions of the partial maximization problems. Let us call this property the *separation of variables for incompatible events* (because $I_1 \cap I_2 = \emptyset$).

This property is trivially valid for the Tsallis family (for $\alpha > 0$, and for $\alpha < 0$ with the change of minimization to maximization) and for the CR family. For the Rényi family it also holds (for $\alpha > 0$, and for $\alpha < 0$ with the change from minimization to maximization), because the Rényi entropy is a function of those trace–form entropies, their level sets coincide.

A simple check shows that this separation of variables property holds also for the convex combination of Shannon's and Burg's entropies, $\beta D_{\mathrm{KL}}(P\|P^*) + (1-\beta)D_{\mathrm{KL}}(P^*\|P)$.

### *4.2. Additivity Property*

The *additivity property* with respect to joining of subsystems is crucial both for the classical thermodynamics and for the information theory.

Let us consider a system which is result of joining of two subsystems. A state of the system is an ordered pair of the states of the subsystems and the space of states of the system is the Cartesian product of the subsystems spaces of state. For systems with finite number of states this means that if the states of subsystems are enumerated by indexes $j$ and $k$ then the states of the system are enumerated by pairs $jk$. The probability distribution for the whole system is $p_{jk}$, and for the subsystems the probability distributions are the marginal distributions $q_j = \sum_k p_{jk}$, $r_k = \sum_j p_{jk}$.

The *additive functions of state* are defined for each state of the subsystems and for a state of the whole system they are sums of these subsystem values:

$$u_{jk} = v_j + w_k$$

where $v_j$ and $w_k$ are functions of the subsystems state.

In classical thermodynamics such functions are called the *extensive quantities*. For expected values of additive quantities the similar additivity condition holds:

$$\sum_{j,k} u_{jk} p_{jk} = \sum_{j,k} (v_j + w_k) p_{ik} = \sum_j v_j q_j + \sum_k w_k r_k \tag{28}$$

Let us consider these expected values as functionals of the probability distributions: $u(\{p_{jk}\})$, $v(\{q_j\})$ and $w(\{r_k\})$. Then the additivity property for the expected values reads:

$$u(\{p_{jk}\}) = v(\{q_j\}) + w(\{r_k\}) \tag{29}$$

where $q_j$ and the $r_k$ are the marginal distributions.

Such a linear additivity property is impossible for non-linear entropy functionals, but under some independence conditions the entropy can behave as an extensive variable.

Let $P$ be a product of marginal distributions. This means that the subsystems are statistically independent: $p_{jk} = q_j r_k$. Assume also that the distribution $P^*$ is also a product of marginal distributions $p^*_{jk} = q^*_j r^*_k$. Then some entropies reveal the additivity property with respect to joining of independent systems.

1. The BGS relative entropy $D_{\mathrm{KL}}(P\|P^*) = D_{\mathrm{KL}}(Q\|Q^*) + D_{\mathrm{KL}}(R\|R^*)$.
2. The Burg entropy $D_{\mathrm{KL}}(P^*\|P) = D_{\mathrm{KL}}(Q^*\|Q) + D_{\mathrm{KL}}(R^*\|R)$ . It is obvious that a convex combination of the Shannon and Burg entropies has the same additivity property.
3. The Rényi entropy $H_{\mathrm{R}\,\alpha}(P\|P^*) = H_{\mathrm{R}\,\alpha}(Q\|Q^*) + H_{\mathrm{R}\,\alpha}(R\|R^*)$. For $\alpha \to \infty$ the Min-entropy also inherits this property.

This property implies the separation of variables for the entropy maximization problems if the system consists of independent subsystems, $p_{jk} = q_j r_k$. Let functionals $u^1(\{p_{jk}\}), \dots u^m(\{p_{jk}\})$ be additive (28) (29) and let the relative entropy $H(P\|P^*)$ be additive with respect to joining of independent

systems. Assume that in equilibrium subsystems are also independent, $p^*_{jk} = q^*_j r^*_k$. Then the solution to the problem

$$H(P\|P^*) \to \min$$

subject to conditions

$$u^i(P) = U_i \ \ (i = 1, \dots m); \ \ p_{jk} = q_j r_k \tag{30}$$

is $p^{\min}_{jk} = q^{\min}_j r^{\min}_k$, where $q^{\min}_j$, $r^{\min}_k$ are solutions of partial problems:

$$H(Q\|Q^*) \to \min$$

subject to the conditions

$$v^i(Q) = V_i \ \ (i = 1, \dots m)$$

and

$$H(R\|R^*) \to \min$$

subject to the conditions

$$w^i(Q) = W_i \ \ (i = 1, \dots m)$$

for some redistribution of the additive functionals values $U_i = V_i + W_i$.

Let us call this property the *separation of variables for independent subsystems*.

Neither the CR, nor the Tsallis divergences families have the additivity property. It is proven [46] that a function $H_h$ has the additivity property if and only if it is a convex combination of the Shannon and Burg entropies. See also Theorem 3 in Appendix.

Nevertheless, both the CR and the Tsallis families have the property of separation of variables for independent subsystems because of the coincidence of the level sets with the additive function, the Rényi entropy (for all $\alpha > 0$).

The Tsallis entropy family has absolutely the same property of separation of variables as the Rényi entropy. To extend this property of the Rényi Tsallis entropies for negative $\alpha$, we have to change there min to max.

For the CR family the result sounds even better: because of better normalization, the separation of variables is valid for $H_{\mathrm{CR}\,\lambda} \to \min$ problem for all values $\lambda \in ]-\infty, \infty[$.

The condition of independence of subsystems $p_{jk} = q_j r_k$ in (30) cannot be relaxed: if we assume $p^*_{jk} = q^*_j r^*_k$ only then the correlations between subsystems may emerge in the solution of the minimization problem. For example, without assumption of independence, for the Burg entropy, the method of Lagrange multipliers gives ($\phi_i$ and $\psi_i$ are the Lagrange multipliers):

$$\frac{p^*_{jk}}{p^{\min}_{jk}} = \sum_i (\phi_i v^i_j + \psi_i w^i_k)$$

and the subsystems are not independent in this state even if they are independent in equilibrium and the conditions are additive. These emergent correlations may be considered as spurious [57] or may be interpreted as sensible ones for some finite systems far from thermodynamic limit for modelling of non-canonic ensembles [35]. In any case, the use of entropies which are additive with respect to joining of independent subsystem does not guarantee independence of subsystems but allows only to separate variables under condition of independence.

The stronger condition was used by Shore and Johnson [58] in the axiomatic derivation of the principle of maximum entropy and the principle of minimum divergence (or 'cross-entropy'). They postulated that the MaxEnt distribution for the whole system is the product of the distributions of the subsystems if the known information (conditions) is the information about subsystems (Axiom III). Independence of subsystems in this axioms is not assumed but should be the consequence of the entropy maximization. This axiom can be called 'separation of variables under independent conditions'. They supplement this assumption by the separation of variables for partition of the state space (Axiom IV), by the condition of uniqueness of the MaxEnt distribution (Axiom I), and by the requirement of the invariance with respect to the coordinate transformations (Axiom II). All these axioms together give the unique classical BGS entropy. For further discussion see [57].

Violation of the Shore and Johnson Axiom III leads to correlation between subsystems and this is an essential difference of the non-classical MaxEnt ensembles from the classical canonical ensembles.

We use the weaker assumption of separation of variables for *independent subsystems and additive conditions*. Its violation leads to much more counterintuitive consequences: Subsystems remain independent (condition) and other conditions are additive (30) but the solution of the MaxEnt problem is the product of distributions which are not solutions of the partial MaxEnt problems. In other words, the probability distribution for a subsystem is modified just by existence of another subsystem without any interactions and correlations.

It seems to be difficult to find a reason for such a behavior and therefore the assumption of separation of variables for independent subsystems and additive conditions is a sensible axiom. It is weaker than the Shore and Johnson Axiom III [58] and, therefore, leads to a wider family of entropies than just a classical BGS entropy. This wider family includes the CR family (12) and the convex combination of the Shannon and the Burg entropies (10).

The question arises: is there any new divergence that has the following three properties: (i) the divergence $H(P\|P^*)$ should decrease in Markov processes with equilibrium $P^*$, (ii) for minimization problems the separation of variables for independent subsystems holds and (iii) the separation of variables for incompatible events holds. A *new* divergence means here that it is not a function of a divergence from the CR family or from the convex combination of the Shannon and the Burg entropies.

The answer is: no, any divergence which has these three properties and is defined and differentiable for positive distributions is a monotone function of $H_h$ for $h(x) = \alpha p^\alpha$ ($\alpha \in ]-\infty, \infty[$, $\alpha \neq 0, 1$), that is, essentially, the CR family (12), or $h(x) = \beta x \ln x - (1-\beta)\ln x$ ($\beta \in [0,1]$). If we relax the differentiability property, then we have to add to the CR family the limits for $\lambda \to \pm\infty$. For $\lambda \to +\infty$ we get the CR analogue of min-entropy

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

The limiting case for the CR family for $\lambda \to -\infty$ is less known but is also a continuous and piecewise differential Lyapunov function for the Master equation:

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

Both properties of separation of variables are based on the specific additivity properties: additivity with respect to the composition of independent systems and additivity with respect to the partitioning

of the space of states. Separation of variables can be considered as a weakened form of additivity: not the minimized function should be additive but there exists such a monotonic transformation of scale after which the function becomes additive (and different transformations may be needed for different additivity properties).

### *4.3. "No More Entropies" Theorems*

The classical Shannon work included the characterization of entropy by its properties. This meant that the classical notion of entropy is natural, and *no more entropies* are expected. In the seminal work of Rényi, again the characterization of entropy by its properties was proved, and for this, extended family the *no more entropies* theorem was proved too. In this section, we prove the next *no more entropies* theorem, where two one-parametric families are selected as sensible: the CR family and the convex combination of Shannon's and Burg's entropies. They are two branches of solutions of the correspondent functional equation and intersect at two points: Shannon's entropy ($\lambda = 1$ in the CR family) and Burg's entropy ($\lambda = 0$). We consider entropies as equivalent if their level sets coincide. In that sense, the Rényi entropy and the Tsallis entropy (with $\alpha > 0$) are equivalent to the CR entropy with $\alpha - 1 = \lambda$, $\lambda > -1$.

Following Rényi, we consider entropies of *incomplete distributions*: $p_i \geq 0$, $\sum_i p_i \leq 1$. The divergence $H(P\|P^*)$ is a $C^1$ smooth function of a pair of positive generalized probability distributions $P = (p_i)$, $p_i > 0$ and $P^* = (p_i^*)$, $p_i^* > 0$, $i = 1, \dots n$.

The following 3 properties are required for characterization of the "natural" entropies.

1. To provide the separation of variables for incompatible events together with the symmetry property we assume that the divergence is separable, possibly, after a scaling transformation: there exists such a function of two variables $f(p, p^*)$ and a monotonic function of one variable $\phi(x)$ that $H(P\|P^*) = \phi(\sum_i f(p_i, p_i^*))$. This formula allows us to define $H(P\|P^*)$ for all $n$.

2. $H(P\|P^*)$ is a Lyapunov function for the Kolmogorov equation (22) for any Markov chain with equilibrium $P^*$. (One can call these functions the *universal* Lyapunov functions because they do not depend on the kinetic coefficients directly, but only on the equilibrium distribution $P^*$.)

3. To provide separation of variables for independent subsystems we assume that $H(P\|P^*)$ is additive (possibly after a scaling transformation): 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 1.** *If a $C^1$-smooth divergence $H(P\|P^*)$ satisfies the conditions 1-3 then, up to monotonic transformation, it is either the CR divergence $H_{\mathrm{CR}\,\lambda}$ or a convex combination of the Botlzmann–Gibbs–Shannon and the Burg entropies, $H_h(P\|P^*) = \beta D_{\mathrm{KL}}(P\|P^*) + (1-\beta) D_{\mathrm{KL}}(P^*\|P)$.*

In a paper [46] this family was identified as the Tsallis relative entropy with some abuse of language, because in the Tsallis entropy the case with $\alpha < 0$ is usually excluded.

First of all, let us prove that any function which satisfies the conditions 1 and 2 is a monotone function of a Csiszár–Morimoto function (6) for some convex smooth function $h(x)$. This was mentioned in 2003 by P. Gorban [46]. Recently, a similar statement was published by S. Amari (Theorem 1 in [59]).

**Lemma 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 $f(p, p^*) = p^* h(\frac{p}{p^*}) + \mathrm{const}(p^*)$, where $h(x)$ is a convex function of one variable.*

*Proof.* Let us consider a Markov chain with two states. For such a chain

$$\frac{\mathrm{d}p_1}{\mathrm{d}t} = q_{12}p_2^*\left(\frac{p_2}{p_2^*} - \frac{p_1}{p_1^*}\right) = -q_{21}p_1^*\left(\frac{p_1}{p_1^*} - \frac{p_2}{p_2^*}\right) = -\frac{\mathrm{d}p_2}{\mathrm{d}t} \tag{31}$$

If $H$ is a Lyapunov function then $\dot{H} \leq 0$ and the following inequality holds:

$$\left(\frac{\partial f(p_2, p_2^*)}{\partial p_2} - \frac{\partial f(p_1, p_1^*)}{\partial p_1}\right)\left(\frac{p_1}{p_1^*} - \frac{p_2}{p_2^*}\right) \leq 0$$

We can consider $p_1, p_2$ as independent variables from an open triangle $D = \{(p_1, p_2) \mid p_{1,2} > 0,\ p_1 + p_2 < 1\}$. For this purpose, we can include the Markov with two states into a chain with three states and $q_{3i} = q_{i3} = 0$.

If for a continuous function of two variables $\psi(x, y)$ in an open domain $D \subset \mathbb{R}^2$ an inequality $(\psi(x_1, y_1) - \psi(x_2, y_2))(y_1 - y_2) \leq 0$ holds then this function does not depend on $x$ in $D$. Indeed, let there exist such values $x_{1,2}$ and $y$ that $\psi(x_1, y) \neq \psi(x_2, y)$, $\psi(x_1, y) - \psi(x_2, y) = \varepsilon > 0$. We can find such $\delta > 0$ that $(x_1, y + \Delta y) \in D$ and $|\psi(x_1, y + \Delta y) - \psi(x_1, y)| < \varepsilon/2$ if $|\Delta y| < \delta$. Hence, $\psi(x_1, y + \Delta y) - \psi(x_2, y) > \varepsilon/2 > 0$ if $|\Delta y| < \delta$. At the same time $(\psi(x_1, y + \Delta y) - \psi(x_2, y))\Delta y \leq 0$, hence, for a positive $0 < \Delta y < \delta$ we have a contradiction. Therefore, the function $\frac{\partial f(p,p^*)}{\partial p}$ is a monotonic function of $\frac{p}{p^*}$, hence, $f(p, p^*) = p^* h(\frac{p}{p^*}) + \mathrm{const}(p^*)$, where $h$ is a convex function of one variable. □

This lemma has important corollaries about many popular divergences $H(P(t)\|P^*)$ which are not Lyapunov functions of Markov chains. This means that there exist such distributions $P_0$ and $P^*$ and a Markov chain with equilibrium distribution $P^*$ that due to the Kolmogorov equations

$$\left.\frac{\mathrm{d}H(P(t)\|P^*)}{\mathrm{d}t}\right|_{t=0} > 0$$

if $P(0) = P_0$. This Markov process increases divergence between the distributions $P, P^*$ (in a vicinity of $P_0$) instead of making them closer. For example,

**Corollary 1.** *The following Bregman divergences [60] are not universal Lyapunov functions for Markov chains:*

- *Squared Euclidean distance $B(P\|P^*) = \sum_i (p_i - p_i^*)^2$;*
- *The Itakura–Saito divergence [61] $B(P\|P^*) = \sum_i \left(\frac{p_i}{p_i^*} - \log\frac{p_i}{p_i^*} - 1\right)$.* □

These divergences violate the requirement: due to the Markov process distributions always monotonically approach equilibrium. (Nevertheless, among the Bregman divergences there exists a universal Lyapunov function for Markov chains, the Kulback–Leibler divergence.)

We place the proof of Theorem 1 in Appendix.

**Remark**. If we relax the requirement of smoothness and consider in conditions of Theorem 1 just continuous functions, then we have to add to the answer the limit divergences,

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

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

## 5. Markov Order

### *5.1. Entropy: a Function or an Order?*

Theorem 1 gives us all of the divergences for which (i) the Markov chains monotonically approach their equilibrium, (ii) the level sets are the same as for a separable (sum over states) divergence and (iii) the level sets are the same as for a divergence which is additive with respect to union of independent subsystems.

We operate with the level sets and their orders, compare where the divergence is larger (for monotonicity of the Markov chains evolution), but the values of entropy are not important by themselves. We are interested in the following order: $P$ precedes $Q$ with respect to the divergence $H_{...}(P\|P^*)$ if there exists such a continuous curve $P(t)$ ($t \in [0,1]$) that $P(0) = P$, $P(1) = Q$ and the function $H(t) = H_{...}(P(t)\|P^*)$ monotonically decreases on the interval $t \in [0,1]$. This property is invariant with respect to a monotonic (increasing) transformation of the divergence. Such a transformation does not change the conditional minimizers or maximizers of the divergence.

There exists one important property that is not invariant with respect to monotonic transformations. The increasing function $F(H)$ of a convex function $H(P)$ is not obligatorily a convex function. Nevertheless, the sublevel sets given by inequalities $H(P) \leq a$ coincide with the sublevel sets $F(H(P)) \leq F(a)$. Hence, sublevel sets for $F(H(P))$ remain convex.

The Jensen inequality

$$H(\theta P + (1-\theta)Q) \leq \theta H(P) + (1-\theta)H(Q)$$

($\theta \in [0,1]$) is not invariant with respect to monotonic transformations. Instead of them, there appears the *max form analogue of the Jensen inequality* (quasiconvexity [64]):

$$H(\theta P + (1-\theta)Q) \leq \max\{H(P), H(Q)\}\,, \quad \theta \in [0,1] \tag{32}$$

This inequality is invariant with respect to monotonically increasing transformations and it is equivalent to convexity of sublevel sets.

**Proposition 1.** *All sublevel sets of a function $H$ on a convex set $V$ are convex if and only if for any two points $P, Q \in V$ and every $\theta \in [0,1]$ the inequality (32) holds.* □

It seems very natural to consider divergences as orders on distribution spaces, and discuss only properties which are invariant with respect to monotonic transformations. From this point of view, the CR family appears absolutely naturally from the additivity (ii) and the "sum over states" (iii) axioms, as well as the convex combination $\beta D_{\mathrm{KL}}(P\|P^*) + (1-\beta)D_{\mathrm{KL}}(P^*\|P)$ ($\alpha \in [0,1]$), and in the above property context there are no other smooth divergences.

### *5.2. Description of Markov Order*

The CR family and the convex combinations of Shannon's and Burg relative entropies are distinguished families of divergences. Apart from them there are many various "divergences", and even

the Csiszár–Morimoto functions (6) do not include all used possibilities. Of course, most users prefer to have an unambiguous choice of entropy: it would be nice to have "the best entropy" for any class of problems. But from some point of view, ambiguity of the entropy choice is unavoidable. In this section we will explain why the choice of entropy is necessarily non unique and demonstrate that for many MaxEnt problems the natural solution is not a fixed distribution, but a well defined set of distributions.

The most standard use of divergence in many application is as follows:

1. On a given space of states an "equilibrium distribution" $P^*$ is given. If we deal with the probability distribution in real kinetic processes then it means that without any additional restriction the current distribution will relax to $P^*$. In that sense, $P^*$ is the most disordered distribution. On the other hand, $P^*$ 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 $P^*$. That is, we define $P$ as solution to the problem:

$$H_{...}(P\|P^*) \to \min \quad \text{subject to } u(P) = U \tag{33}$$

   Without the condition $u(P) = U$ the solution should be simply $P^*$.

Now we have too many entropies and do not know what is the optimal choice of $H_{...}$ and what should be the optimal estimate of $P$. In this case the proper question may be: *which $P$ could not be such an optimal estimate*? We can answer the exclusion question. Let for a given $P^0$ the condition hold, $u(P^0) = U$. If there exists a Markov process with equilibrium $P^*$ such that at point $P^0$ due to the Kolmogorov equation (22)

$$\frac{\mathrm{d}P}{\mathrm{d}t} \neq 0 \text{ and } \frac{\mathrm{d}(u(P))}{\mathrm{d}t} = 0$$

then $P^0$ cannot be the optimal estimate of the distribution $P$ under condition $u(P) = U$.

The motivation of this approach is simple: any Markov process with equilibrium $P^*$ increases disorder and brings the system "nearer" to the equilibrium $P^*$. If at $P^0$ it is possible to move along the condition plane towards the more disordered distribution then $P^0$ cannot be considered as an extremely disordered distribution on this plane. On the other hand, 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 (22) $P(t)$ with initial condition $P(0) = P^0$ has no points on the plane $u(P) = U$ for $t > 0$.

Markov process here is considered as a "randomization". Any set $C$ of distributions can be divided in two parts: the distributions which retain in $C$ after some non-trivial randomization and the distributions which leave $C$ after any non-trivial randomization. The last are the maximally random elements of $C$: they cannot become more random and retain in $C$. Conditional minimizers of relative entropies $H_h(P\|P^*)$ in $C$ are maximally random in that sense.

There are too many functions $H_h(P\|P^*)$ for effective description of all their conditional minimizers. Nevertheless, we can describe the maximally random distributions directly, by analysis of Markov processes.

To analyze these properties more precisely, we need some formal definitions.

**Definition 1.** *(Markov preorder). 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 with equilibrium $P^*$ and use notation $P^0 \succ^0_{P^*} P^1$.*

**Definition 2.** *Markov order is the closed transitive closure of the Markov preorder. For the Markov order with equilibrium $P^*$ we use notation $P^0 \succ_{P^*} P^1$.*

For a given $P^* = (p^*_i)$ and a distribution $P = (p_i)$ the set of all vectors $v$ with coordinates

$$v_i = \sum_{j,\, j\neq i} q_{ij} p^*_j \left( \frac{p_j}{p^*_j} - \frac{p_i}{p^*_i} \right)$$

where $p^*_i$ and $q_{ij} \geq 0$ are connected by identity (23) is a closed convex cone. This is a cone of all possible time derivatives of the probability distribution at point $P$ for Markov processes with equilibrium $P^* = (p^*_i)$. For this cone, we use notation $\mathbf{Q}_{(P,P^*)}$

**Definition 3.** *For each distribution $P$ and a $n$-dimensional vector $\Delta$ we say that $\Delta <_{(P,P^*)} 0$ if $\Delta \in \mathbf{Q}_{(P,P^*)}$. This is the local Markov order.*

**Proposition 2.** $\mathbf{Q}_{(P,P^*)}$ *is a proper cone,* i.e., *it does not include any straight line.*

*Proof.* To prove this proposition its is sufficient to analyze the formula for entropy production (for example, in form (24)) and mention that for strictly convex $h$ (for example, for traditional $x \ln x$ or $(x-1)^2/2$) $\mathrm{d}H_h/\mathrm{d}t = 0$ if and only if $\mathrm{d}P/\mathrm{d}t = 0$. If the cone $\mathbf{Q}_{(P,P^*)}$ includes both vectors $x$ and $-x$ ($x \neq 0$ it means that there exist Markov chains with equilibrium $P^*$ and with opposite time derivatives at point $P$. Due to the positivity of entropy production (24) this is impossible. □

The connection between the local Markov order and the Markov order gives the following proposition, which immediately follows from definitions.

**Proposition 3.** $P^0 \succ_{P^*} P^1$ *if and only if there exists such a continuous almost everywhere differentiable curve $P(t)$ in the simplex of probability distribution that $P(0) = P^0$, $P(1) = P^1$ and for all $t \in [0,1]$, where $P(t)$ is differentiable,*

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} \in \mathbf{Q}_{(P(t),P^*)} \qquad \square \tag{34}$$

For our purposes, the following estimate of the Markov order through the local Markov order is important.

**Proposition 4.** *If $P^0 \succ_{P^*} P^1$ then $P^0 >_{(P^0,P^*)} P^1$,* i.e., *$P^1 - P^0 \in \mathbf{Q}_{(P,P^*)}$.*

This proposition follows from the characterization of the local order and detailed description of the cone $\mathbf{Q}_{(P(t),P^*)}$ (Theorem 2 below).

Let us recall that a convex pointed cone is a convex envelope of its extreme rays. A ray with directing vector $x$ is a set of points $\lambda x$ ($\lambda \geq 0$). We say that $l$ is an extreme ray of $\mathbf{Q}$ if for any $u \in l$ and any $x, y \in \mathbf{Q}$, whenever $u = (x+y)/2$, we must have $x, y \in l$. To characterize the extreme rays of the cones of the local Markov order $\mathbf{Q}_{(P,P^*)}$ we need a graph representation of the Markov chains. We use the notation $A_i$ for states (vertices), and designate transition from state $A_i$ to state $A_j$ by an arrow (edge) $A_i \to A_j$. This transition has its transition intensity $q_{ji}$ (the coefficient in the Kolmogorov equation (21)).

**Lemma 2.** *Any extreme ray of the cone $\mathbf{Q}_{(P,P^*)}$ corresponds to a Markov process which transition graph is a simple cycle*

$$A_{i_1} \to A_{i_2} \to \dots A_{i_k} \to A_{i_1}$$

*where $k \leq n$, all the indices $i_1, \dots i_k$ are different, and transition intensities for a directing vector of such an extreme ray $q_{i_{j+1}\, i_j}$ may be selected as $1/p^*_{i_j}$:*

$$q_{i_{j+1}\, i_j} = \frac{1}{p^*_{i_j}} \tag{35}$$

*(here we use the standard convention that for a cycle $q_{i_{k+1}\, i_k} = q_{i_1\, i_k}$).*

*Proof.* First of all, let us mention that if for three vectors $x, y, u \in \mathbf{Q}_{(P,P^*)}$ we have $u = (x+y)/2$ then the set of transitions with non-zero intensities for corresponding Markov processes for $x$ and $y$ are included in this set for $u$ (because negative intensities are impossible). Secondly, just by calculation of the free variables in the equations (23) (with additional condition) we find that the the amount of non-zero intensities for a transition scheme which represents an extreme ray should be equal to the amount of states included in the transition scheme. Finally, there is only one scheme with $k$ vertices, $k$ edges and a positive equilibrium, a simple oriented cycle. □

**Theorem 2.** *Any extreme ray of the cone $\mathbf{Q}_{(P,P^*)}$ corresponds to a Markov process whose transition graph is a simple cycle of the length 2: $A_i \rightleftarrows A_j$. A transition intensities $q_{ij},\ q_{ji}$ for a directing vector of such an extreme ray may be selected as*

$$q_{ij} = \frac{1}{p^*_j}\ ,\quad q_{ji} = \frac{1}{p^*_i} \tag{36}$$

*Proof.* Due to Lemma 2, it is sufficient to prove that for any distribution $P$ the right hand side of the Kolmogorov equation (22) for a simple cycle with transition intensities (35) is a conic combination (the combination with non-negative real coefficients) of the right hand sides of this equation for simple cycles of the length 2 at the same point $P$. Let us prove this by induction. For the cycle length 2 it is trivially true. Let this hold for the cycle lengths $2, \dots n-1$. For a cycle of length $n$, $A_{i_1} \to A_{i_2} \to \dots A_{i_k} \to A_{i_1}$, with transition intensities given by (35) the right hand side of the Kolmogorov equation is the vector $v$ with coordinates

$$v_{i_j} = \frac{p_{i_{j-1}}}{p^*_{i_{j-1}}} - \frac{p_{i_j}}{p^*_{i_j}}$$

(under the standard convention regarding cyclic order). Other coordinates of $v$ are zeros. Let us find the minimal value of $p_{i_j}/p^*_{i_j}$ and rearrange the indices by a cyclic permutation to put this minimum in the first place:

$$\min_j \left\{ \frac{p_{i_j}}{p^*_{i_j}} \right\} = \frac{p_{i_1}}{p^*_{i_1}}$$

The vector $v$ is a sum of two vectors: a directing vector for the cycle $A_{i_2} \to \dots A_{i_k} \to A_{i_2}$ of the length $n-1$ with transition intensities given by formula (35) (under the standard convention about the cyclic order for this cycle) and a vector

$$\frac{\frac{p_{i_n}}{p^*_{i_n}} - \frac{p_{i_1}}{p^*_{i_1}}}{\frac{p_{i_2}}{p^*_{i_2}} - \frac{p_{i_1}}{p^*_{i_1}}} v^2$$

where $v^2$ is the directing vector for a cycle of length 2, $A_{i_1} \rightleftarrows A_{i_2}$ which can have only two non-zero coordinates:

$$v^2_{i_1} = \frac{p_{i_2}}{p^*_{i_2}} - \frac{p_{i_1}}{p^*_{i_1}} = -v^2_{i_2}$$

The coefficient in front of $v^2$ is positive because $p_{i_1}/p^*_{i_1}$ is the minimal value of $p_{i_j}p^*_{i_j}$. A case when $p_{i_1}/p^*_{i_1} = p_{i_2}/p^*_{i_2}$ does not need special attention because it is equivalent to the shorter cycle $A_{i_1} \to A_{i_3} \to \dots A_{i_k} \to A_{i_1}$ ($A_{i_2}$ could be omitted). A conic combination of conic combinations is a conic combination again. □

It is quite surprising that the local Markov order and, hence, the Markov order also are generated by the reversible Markov chains which satisfy the detailed balance principle. We did not include any reversibility assumptions, and studied the general Markov chains. Nevertheless, for the study of orders, the system of cycles of length 2 all of which have the same equilibrium is sufficient.

### *5.3. Combinatorics of Local Markov Order*

Let us describe the local Markov order in more detail. First of all, we represent kinetics of the reversible Markov chains. For each pair $A_i, A_j$ ($i \neq j$) we select an arbitrary order in the pair and write the correspondent cycle of the length 2 in the form $A_i \leftrightarrows A_j$. For this cycle we introduce the directing vector $\gamma^{ij}$ with coordinates

$$\gamma^{ij}_k = -\delta_{ik} + \delta_{jk} \tag{37}$$

where $\delta_{ik}$ is the Kronecker delta. This vector has the $i$th coordinate $-1$, the $j$th coordinate $1$ and other coordinates are zero. Vectors $\gamma^{ij}$ are parallel to the edges of the standard simplex in $R^n$. They are antisymmetric in their indexes: $\gamma^{ij} = -\gamma^{ji}$.

We can rewrite the Kolmogorov equation in the form

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \sum_{\text{pairs } ij} \gamma^{ij} w_{ji} \tag{38}$$

where $i \neq j$, each pair is included in the sum only once (in the preselected order of $i, j$) and

$$w_{ji} = r_{ji} \left( \frac{p_i}{p^*_i} - \frac{p_j}{p^*_j} \right)$$

The coefficient $r_{ji} \geq 0$ satisfies the detailed balance principle:

$$r_{ji} = q_{ji}p_i^* = q_{ij}p_j^* = r_{ij}$$

We use the three-value sign function:

$$\mathrm{sign}x = \begin{cases} -1, & \text{if } x < 0; \\ 0, & \text{if } x = 0; \\ 1, & \text{if } x > 0 \end{cases} \tag{39}$$

With this function we can rewrite Equation (38) again as follows:

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \sum_{\text{pairs } ij,\ r_{ji}\neq 0} r_{ji}\gamma^{ij}\mathrm{sign}\left(\frac{p_i}{p_i^*} - \frac{p_j}{p_j^*}\right)\left|\frac{p_i}{p_i^*} - \frac{p_j}{p_j^*}\right| \tag{40}$$

The non-zero coefficients $r_{ji}$ may be arbitrary positive numbers. Therefore, using Theorem 2, we immediately find that the cone of the local Markov order at point $P$ is

$$\mathbf{Q}_{(P,P^*)} = \mathrm{cone}\left\{\gamma^{ij}\mathrm{sign}\left(\frac{p_i}{p_i^*} - \frac{p_j}{p_j^*}\right) \;\middle|\; r_{ji} > 0\right\} \tag{41}$$

where cone$\{\}$ stands for the conic hull.

The number $\mathrm{sign}\left(\frac{p_i}{p_i^*} - \frac{p_j}{p_j^*}\right)$ is 1, when $\frac{p_i}{p_i^*} > \frac{p_j}{p_j^*}$, $-1$, when $\frac{p_i}{p_i^*} < \frac{p_j}{p_j^*}$ and 0, when $\frac{p_i}{p_i^*} = \frac{p_j}{p_j^*}$. For a given $P^*$, the standard simplex of distributions $P$ is divided by planes $\frac{p_i}{p_i^*} = \frac{p_j}{p_j^*}$ into convex polyhedra where functions $\mathrm{sign}\left(\frac{p_i}{p_i^*} - \frac{p_j}{p_j^*}\right)$ are constant. In these polyhedra the cone of the local Markov order (41) $\mathbf{Q}_{(P,P^*)}$ is also constant. Let us call these polyhedra *compartments*.

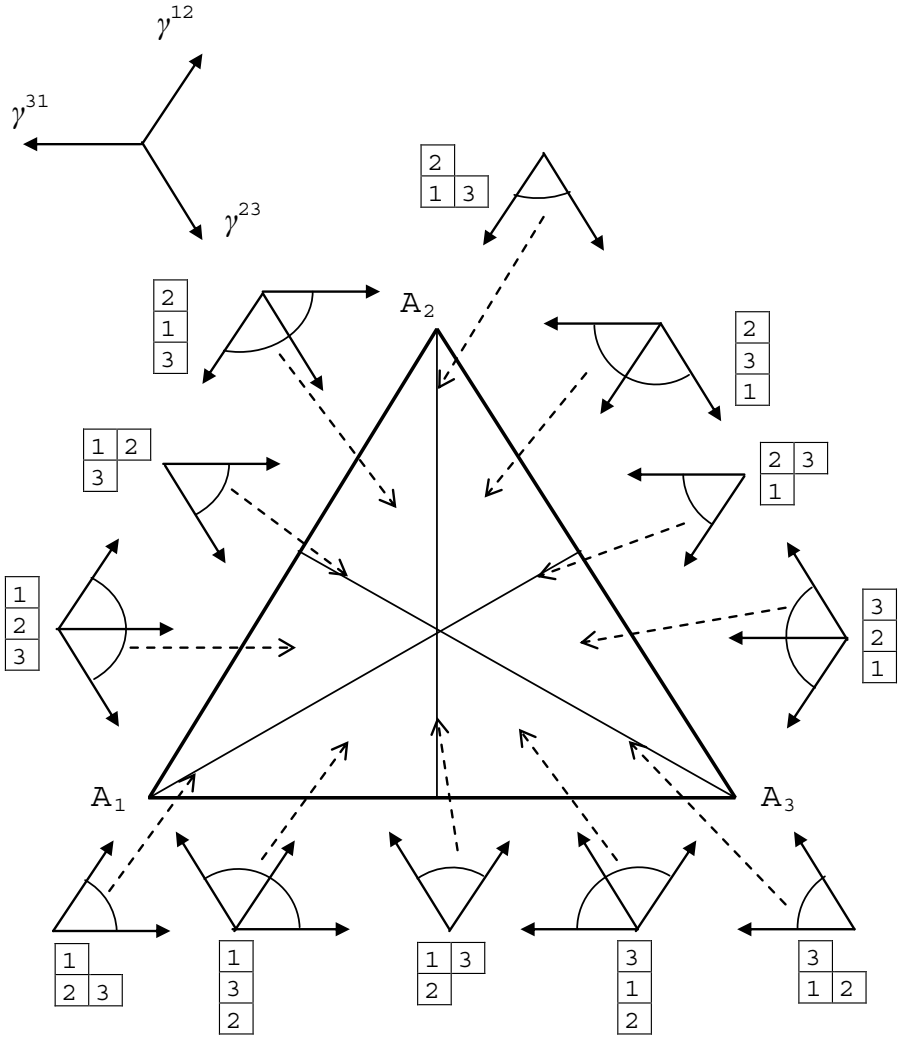
In Figure 1 we represent compartments and cones of the local Markov order for the Markov chains with three states, $A_{1,2,3}$. The reversible Markov chain consists of three reversible transitions $A_1 \leftrightarrows A_2 \leftrightarrows A_3 \leftrightarrows A_1$ with corresponding directing vectors $\gamma^{12} = (-1, 1, 0)^\top$; $\gamma^{23} = (0, -1, 1)^\top$; $\gamma^{31} = (1, 0, -1)^\top$. The topology of the partitioning of the standard simplex into compartments and the possible values of the cone $\mathbf{Q}_{(P,P^*)}$ do not depend on the position of the equilibrium distribution $P^*$.

Let us describe all possible compartments and the correspondent local Markov order cones. For every natural number $k \leq n-1$ the $k$-dimensional compartments are numerated 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:

$$\frac{p_i}{p_i^*} > \frac{p_j}{p_j^*} \text{ if } \sigma(i) < \sigma(j); \quad \frac{p_i}{p_i^*} = \frac{p_j}{p_j^*} \text{ if } \sigma(i) = \sigma(j) \tag{42}$$

Let us use for the correspondent compartment notation $\mathcal{C}_\sigma$ and for the Local Markov order cone $Q_\sigma$. Let $k_i$ be a number of elements in preimage of $i$ ($i = 1, \ldots, k$): $k_i = |\{j \mid \sigma(j) = i\}|$. It is convenient to represent surjection $\sigma$ as a tableau with $k$ rows and $k_i$ cells in the $i$th row filled by numbers from $\{1, 2, \ldots, n\}$. First of all, let us draw diagram, that is a finite collection of cells arranged in left-justified rows. The $i$th row has $k_i$ cells. A tableau is obtained by filling cells with numbers $\{1, 2, \ldots, n\}$. Preimages of $i$ are located in the $i$th row. The entries in each row are increasing. (This is convenient to avoid ambiguity of the representation of the surjection $\sigma$ by the diagram.) Let us use for tableaus the same notation as for the corresponding surjections.

**Figure 1**. Compartments $\mathcal{C}_\sigma$, corresponding cones $\mathbf{Q}_\sigma$ (the angles) and all tableaus $\sigma$ for the Markov chain with three states (the choice of equilibrium ($p_i^* = 1/3$), does not affect combinatorics and topology of tableaus, compartments and cones).




Let a tableau $A$ have $k$ rows. We say that a tableau $B$ follows $A$ (and use notation $A \to B$) if $B$ has $k-1$ rows and $B$ can be produced from $A$ by joining of two neighboring rows in $A$ (with ordering the numbers in the joined row). For the transitive closure of the relation $\to$ we use notation $\Rightarrow$.

**Proposition 5.** $r\partial Q_\sigma = \bigcup_{\sigma \Rightarrow \varsigma} Q_\varsigma$ $\square$

Here $r\partial U$ stands for the "relative boundary" of a set $U$ in the minimal linear manifold which includes $U$.

The following Proposition characterizes the local order cone through the surjection $\sigma$. It is sufficient to use in definition of $Q_\sigma$ (41) vectors $\gamma^{ij}$ (37) with $i$ and $j$ from the neighbor rows of the diagram (see Figure 1).

**Proposition 6.** *For a given surjection $\sigma$ compartment $\mathcal{C}_\sigma$ and cone $Q_\sigma$ have the following description:*

$$\mathcal{C}_\sigma = \left\{ P \mid \frac{p_i}{p_i^*} = \frac{p_j}{p_j^*} \text{ for } \sigma(i) = \sigma(j) \text{ and } \frac{p_i}{p_i^*} > \frac{p_j}{p_j^*} \text{ for } \sigma(j) = \sigma(i) + 1 \right\} \tag{43}$$

$$Q_\sigma = \mathrm{cone}\{\gamma^{ij} \mid \sigma(j) = \sigma(i) + 1\} \qquad \square \tag{44}$$

Compartment $\mathcal{C}_\sigma$ is defined by equalities $\frac{p_i}{p_i^*} = \frac{p_j}{p_j^*}$ where $i, j$ belong to one row of the tableau $\sigma$ and inequalities $\frac{p_i}{p_i^*} > \frac{p_j}{p_j^*}$ where $j$ is situated in a row one step down from $i$ in the tableau ($\sigma(j) = \sigma(i) + 1$). Cone $Q_\sigma$ is a conic hull of $\sum_{i=1}^{k-1} k_i k_{i+1}$ vectors $\gamma^{ij}$. For these vectors, $j$ is situated in a row one step down from $i$ in the tableau. Extreme rays of $Q_\sigma$ are products of the positive real half-line on vectors $\gamma^{ij}$ (44).

Each compartment has the *lateral faces* and the *base*. We call the face a lateral face, if its closure includes the equilibrium $P^*$. The base of the compartment belongs to a border of the standard simplex of probability distributions.

To enumerate all the lateral faces of a $k$-dimensional compartment $\mathcal{C}_\sigma$ of codimension $s$ (in $\mathcal{C}_\sigma$) we have to take all subsets with $s$ elements in $\{1, 2, \ldots, k\}$. For any such a subset $J$ the correspondent $k - s$-dimensional lateral face is given by additional equalities $\frac{p_i}{p_i^*} = \frac{p_j}{p_j^*}$ for $\sigma(j) = \sigma(i) + 1$, $i \in J$.

**Proposition 7.** *All $k - s$-dimensional lateral faces of a $k$-dimensional compartment $\mathcal{C}_\sigma$ are in bijective correspondence with the $s$-element subsets $J \subset \{1, 2, \ldots, k\}$. For each $J$ the correspondent lateral face is given in $\mathcal{C}_\sigma$ by equations*

$$\frac{p_i}{p_i^*} = \frac{p_j}{p_j^*} \ \text{for all} \ \ i \in J \ \ \text{and} \ \ \sigma(j) = \sigma(i) + 1 \qquad \square \tag{45}$$

The 1-dimensional lateral faces (extreme rays) of compartment $\mathcal{C}_\sigma$ are given by selection of one number from $\{1, 2, \ldots, k\}$ (this number is the complement of $J$). For this number $r$, the correspondent 1-dimensional face is a set parameterized by a positive number $a \in ]1, a_r]$, $a_r = 1/\sum_{\sigma(i) \le r} p_i^*$:

$$\begin{gathered} \frac{p_i}{p_i^*} = a, \ \ \text{for} \ \ \sigma(i) \le r\,; \ \ \frac{p_i}{p_i^*} = b, \ \ \text{for} \ \ \sigma(i) > r\,; \\ a > 1 > b \ge 0, \ a \sum_{i,\, \sigma(i) \le r} p_i^* + b \sum_{i,\, \sigma(i) > r} p_i^* = 1 \end{gathered} \tag{46}$$

The compartment $\mathcal{C}_\sigma$ is the interior of the $k$-dimensional simplex with vertices $P^*$ and $v_r$ ($r = 1, 2, \ldots k$). The vertex $v_r$ is the intersection of the correspondent extreme ray (46) with the border of the standard simplex of probability distributions: $P = v_r$ if

$$p_i = p_i^* a_r, \ \ \text{for} \ \ \sigma(i) \le r; \ \ p_i = 0 \ \ \text{for} \ \ \sigma(i) > r \tag{47}$$

The base of the compartment $\mathcal{C}_\sigma$ is a $k - 1$-dimensional simplex with vertices $v_r$ ($r = 1, 2, \ldots k$).

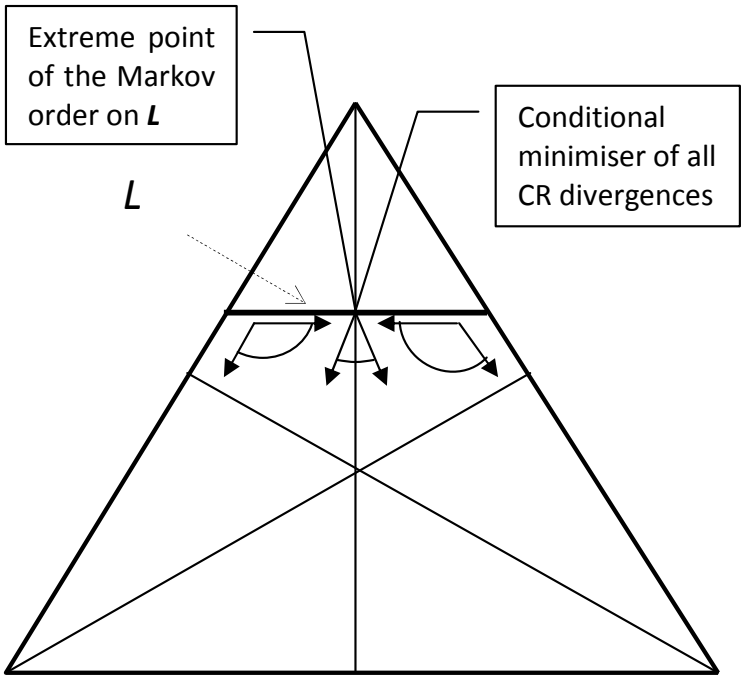
It is necessary to stress that we use the reversible Markov chains for construction of the general Markov order due to Theorem 2.

## 6. The "Most Random" and Conditionally Extreme Distributions

### *6.1. Conditionally Extreme Distributions in Markov Order*

The Markov order can be used to reduce the uncertainty in the standard settings. Let the plane $L$ of the known values of some moments be given: $u^i(P) = U_i$ on $L$. Assume also that the "maximally

**Figure 2**. If the moments are just some of $p_i$ then all points of conditionally minimal divergence are the same for all the main divergences and coincide with the unique conditionally extreme point of the Markov order (example for the Markov chain with three states, symmetric equilibrium ($p_i^* = 1/3$)) and the moment plane $p_2$ =const.




disordered" distribution (equilibrium) $P^*$ is known and we assume that the probability distribution is $P^*$ if there is no restrictions. Then, the standard way to evaluate $P$ for given moment conditions $u^i(P) = U_i$ is known: just to minimize $H_{...}(P\|P^*)$ under these conditions. For the Markov order we also can define the *conditionally extreme points* on $L$.

**Definition 4.** *Let $L$ be an affine subspace of $\mathbf{R}^n$, $\Sigma_n$ be a standard simplex in $\mathbf{R}^n$. A probability distribution $P \in L \cap \Sigma_n$ is a conditionally extreme point of the Markov order on $L$ if*

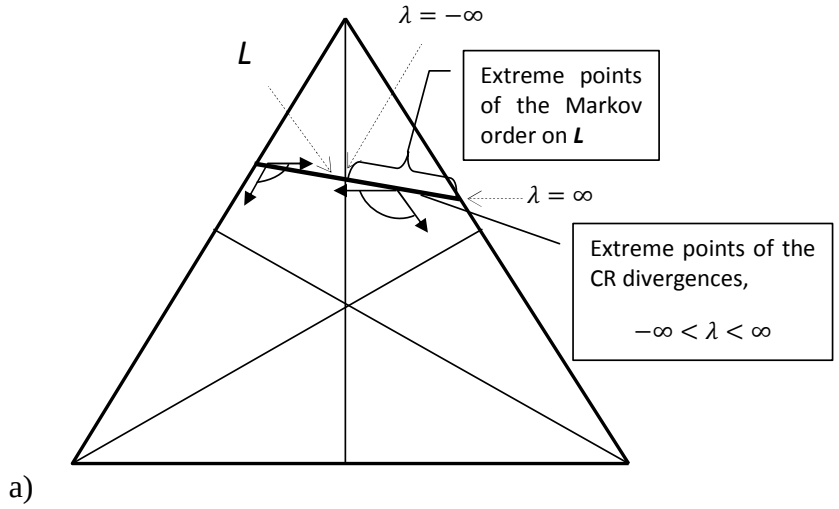
$$(P + \mathbf{Q}_{(P,P^*)}) \cap L = \{P\}$$

It is useful to compare this definition to the condition of the extremum of a differentiable function $H$ on $L$: $\mathrm{grad}H \perp L$.

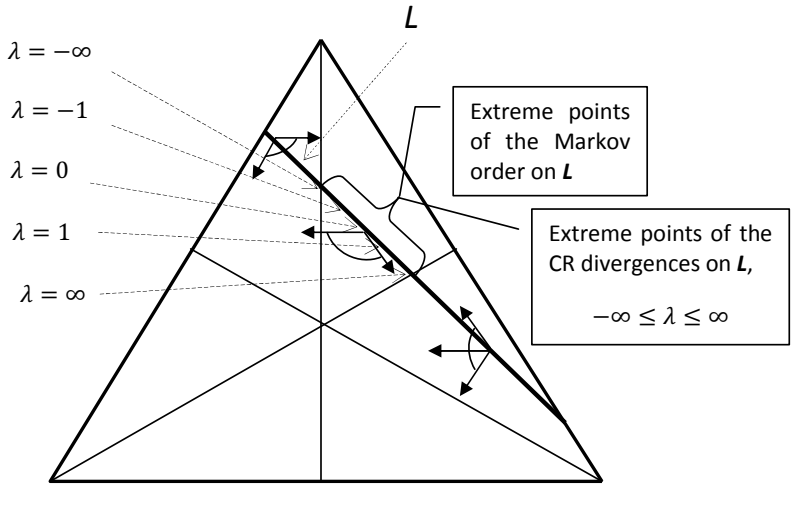
First of all, it is obvious that in the case when all the moments $u^i(P)$ are just some of the values $p_i$, then there exists only one extreme point of the Markov order on $L$, and this point is, at the same time, the conditional minimum on $L$ of all Csiszár–Morimoto functions $H_h(P)$ (6) (see, for example, Figure 2). This situation is unstable, and for a small perturbation of $L$ the set of extreme points of the Markov order on $L$ includes the intersection of $L$ with one of compartments (Figure 3a). For the Markov chains with three states, each point of this intersection is a conditional minimizer of one of the CR divergences (see Fig. 3a). Such a situation persists for all $L$ in general positions (Figure 3b). The extreme points of the family $\beta D_{\mathrm{KL}}(P\|P^*) + (1-\beta)D_{\mathrm{KL}}(P^*\|P)$ form an interval which is strictly inside the interval of the extreme points of the Markov order on $L$. For higher dimensions of $L \cap \Sigma_n$ the Markov order on $L$ also

includes the intersection of $L$ with some compartments, however the conditional minimizers of the CR divergences form a curve there, and extreme points of the family $\beta D_{\mathrm{KL}}(P\|P^*) + (1-\beta)D_{\mathrm{KL}}(P^*\|P)$ on $L$ form another curve. These two curves intersect at two points ($\lambda = 0, -1$), which correspond to the BGS and Burg relative entropies.

**Figure 3**. The set of conditionally extreme points of the Markov order on the moment plane in two general positions. For the main divergences the points of conditionally minimal divergence are distributed in this set. For several of the most important divergences these minimizers are pointed out. In this simple example each extreme point of the Markov order is at the same time a minimizer of one of the $H_{\mathrm{CR}\,\lambda}$ ($\lambda \in ]-\infty, +\infty[$) (examples for the Markov chain with three states, symmetric equilibrium ($p_i^* = 1/3$)).




a)




b)

### *6.2. How to Find the Most Random Distributions?*

Let the plane $L$ of the known values of some moments be given: $u^i(P) = \sum_j u^i_j p_j = U_i$ $(i = 1, \dots m)$ on $L$. For a given divergence $H(P\|P^*)$ we are looking for a conditional minimizer $P$:

$$H(P\|P^*) \to \min \quad \text{subject to} \quad u^i(P) = U_i (i = 1, \dots m) \tag{48}$$

We can assume that $H(P\|P^*)$ is convex. Moreover, usually it is one of the Csiszár–Morimoto functions (6). This is very convenient for numerical minimization because the matrix of second derivatives is diagonal. Let us introduce the Lagrange multipliers $\mu_i$ $(i = 1, \dots m)$ and write the system of equations ($\mu_0$ is the Lagrange multiplier for the total probability identity $\sum_j p_j = 1$ :

$$\begin{aligned} &\frac{\partial H}{\partial p_j} = \mu_0 + \sum_{i=1}^{m} \mu_i u^i_j \; ; \\ &\sum_{j=1}^{n} u^i_j p_j = U_i \; ; \\ &\sum_{j=1}^{n} p_j = 1 \end{aligned} \tag{49}$$

Here we have $n + m + 1$ equations for $n + m + 1$ unknown variables ($p_j$, $\mu_i$, $\mu_0$).

Usually $H$ is a convex function with a diagonal matrix of second variables and the method of choice for solution of this equation (49) is the Newton method. On the $l + 1$st iteration to find $P^{l+1} = P^l + \Delta P$ we have to solve the following system of linear equations

$$\begin{aligned} &\sum_{s=1}^{n} \left.\frac{\partial^2 H}{\partial p_j \partial p_s}\right|_{P=P^l} \Delta p_s = \mu_0 + \sum_{i=1}^{m} \mu_i u^i_j - \left.\frac{\partial H}{\partial p_j}\right|_{P=P^l} \; ; \\ &\sum_{j=1}^{n} u^i_j \Delta p_j = 0 \; ; \\ &\sum_{j=1}^{n} \Delta p_j = 0 \end{aligned} \tag{50}$$

For a diagonal matrix of the second derivatives the first $n$ equations can be explicitly resolved. If for the solution of this system (50) the positivity condition $p^l_j + \Delta p_j > 0$ does not hold (for some of $j$) then we should decrease the step, for example by multiplication $\Delta P := \theta \Delta P$, where

$$0 < \theta < \min_{p^l_i + \Delta p_i < 0} \frac{p^l_i}{|\Delta p_i|}$$

For initial approximation we can take any positive normalized distribution which satisfies the conditions $u^i(P) = U_i$ $(i = 1, \dots m)$.

For the Markov orders the set of conditionally extreme distributions consists of intersections of $L$ with compartments.

Here we find this set for one moment condition of the form $u(P) = \sum_j u_j p_j = U$. First of all, assume that $U \neq U^*$, where $U^* = u(P^*) = \sum_j u_j p^*_j$ (if $U = U^*$ then equilibrium is the single conditionally

extreme distribution). In this case, the set of conditionally extreme distributions is the intersection of the condition hyperplane with the closure of one compartment and can be described by the following system of equations and inequalities (under standard requirements $p_i \geq 0$, $\sum_i p_i = 1$ ):

$$\begin{aligned} &\sum_j u_j p_j = U; \\ &\frac{p_i}{p_i^*} \geq \frac{p_j}{p_j^*} \text{ if } u_i(U-U^*) \geq u_j(U-U^*) \end{aligned} \tag{51}$$

(hence, $\frac{p_i}{p_i^*} = \frac{p_j}{p_j^*}$ if $u_i = u_j$).

To find this solution it is sufficient to study dynamics of $u(P)$ due to equations (38) and to compare it with dynamics of $u(P)$ due to a model system $\dot{P} = P^* - P$. This model system is also a Markov chain and, therefore, $P^* - P \in \mathbf{Q}_{(P,P^*)}$. Equations and inequalities (51) mean that the set of conditionally extreme distributions is the intersection of the condition hyperplane with the closure of compartment $\mathcal{C}$. In $\mathcal{C}$, numbers $\frac{p_i}{p_i^*}$ have the same order on the real line as numbers $u_i(U-U^*)$ have, these two tuples of numbers correspond to the same tableau $\sigma$ and $\mathcal{C} = \mathcal{C}_\sigma$.

For several linearly independent conditions there exists a condition plane $L$:

$$u^i(P) = \sum_j u_j^i p_j = U_i \ \ (i = 1, \ldots m) \tag{52}$$

Let us introduce the $m$-dimensional space $T$ with coordinates $u^i$. Operator $u(P) = (u^i(P))$ maps the distribution space into $T$ and the affine manifold $L$ (52) maps into a point with coordinates $u^i = U_i$.

If $P^* \in L$ then the problem is trivial and the only extreme distribution of the Markov order on $L$ is $P^*$. Let us assume that $P^* \notin L$.

For each distribution $P \in L$ we can study the possible direction of motions of projection distributions onto $T$ due to the Markov processes.

First of all, let us mention that if $u(\gamma^{ij}) = 0$ then the transitions $A_i \leftrightarrows A_j$ move the distribution along $L$. Hence, for any conditionally extreme distribution $P \in L$ this transition $A_i \leftrightarrows A_j$ should be in equilibrium and the partial equilibrium condition holds: $\frac{p_i}{p_i^*} = \frac{p_j}{p_j^*}$.

Let us consider processes with $u(\gamma^{ij}) \neq 0$. If there exists a convex combination (40) of vectors $u(\gamma^{ij})\text{sign}\left(\frac{p_i}{p_i^*} - \frac{p_j}{p_j^*}\right)$ $(u(\gamma^{ij}) \neq 0)$ that is equal to zero then $P$ cannot be an extreme distribution of the Markov order on $L$.

These two conditions for vectors $\gamma^{ij}$ with $u(\gamma^{ij}) = 0$ and for the set of vectors with non-zero projection on the condition space define the extreme distributions of the Markov order on the condition plane $L$ for several conditions.

## 7. Generalized Canonical Distribution

### *7.1. Reference Distributions for Main Divergences*

A system with equilibrium $P^*$ is given and expected values of some variables $u_j(P) = U_j$ are known. We need to find a distribution $P$ with these values $u_j(P) = U_j$ and is "the closest" to the equilibrium distribution under this condition.

This distribution parameterized through expectation values is often called the *reference distribution* or *generalized canonical distribution*. After Gibbs and Jaynes, the standard statement of this problem is an optimization problem:

$$H(P\|P^*) \to \min, \quad u_j(P) = U_j$$

for appropriate divergence $H(P\|P^*)$. If the number of conditions is $m$ then this optimization problem can be often transformed into $m+1$ equations with $m+1$ unknown Lagrange multipliers.

In this section, we study the problem of the generalized canonical distributions for single condition $u(P) = \sum_{i=1}^n u_i p_i = U, U \neq U^*$.

For the Csiszár–Morimoto functions $H_h(P\|P^*)$

$$\frac{\partial H_h}{\partial p_i} = h'\left(\frac{p_i}{p_i^*}\right) \tag{53}$$

We assume that the function $h'(x)$ has an inverse function $g$: $g(h'(x)) = x$ for any $x \in ]0,\infty[$. The method of Lagrange multipliers gives for the generalized canonical distribution:

$$\frac{\partial H_h}{\partial p_i} = \mu_0 \frac{\partial(\sum_{j=1}^n p_j)}{\partial p_i} + \mu \frac{\partial U}{\partial p_i}, \; h'\left(\frac{p_i}{p_i^*}\right) = \mu_0 + \mu u_i, \; \sum_{i=1}^n p_i = 1, \; \sum_{i=1}^n p_i u_i = U \tag{54}$$

As a result, we get the final expression for the distribution

$$p_i = p_i^* g(\mu_0 + u_i \mu)$$

and equations for Lagrange multipliers $\mu_0$ and $\mu$:

$$\sum_{i=1}^n p_i^* g(\mu_0 + u_i\mu) = 1, \; \sum_{i=1}^n p_i^* g(\mu_0 + u_i\mu) u_i = U \tag{55}$$

If the image of $h'(x)$ is the whole real line ($h'(]0,\infty[) = R$) then for any real number $y$ the value $g(y) \geq 0$ is defined and there exist no problems about positivity of $p_i$ due to (55).

For the BGS relative entropy $h'(x) = \ln x$ (we use the normalized $h(x) = x\ln x - (x-1)$ (19)). Therefore, $g(x) = \exp x$ and for the generalized canonical distribution we get

$$p_i = p_i^* \mathrm{e}^{\mu_0} \mathrm{e}^{u_i\mu}, \; \mathrm{e}^{-\mu_0} = \sum_{i=1}^n p_i^* \mathrm{e}^{u_i\mu}, \; \text{`} \sum_{i=1}^n p_i^* u_i \mathrm{e}^{u_i\mu} = U \sum_{i=1}^n p_i^* \mathrm{e}^{u_i\mu} \tag{56}$$

As a result, we get one equation for $\mu$ and an explicit expression for $\mu_0$ through $\mu$.

These $\mu_0$ and $\mu$ have the opposite sign comparing to (5) just because the formal difference between the entropy maximization and the relative entropy minimization. Equation (56) is essentially the same as (5).

For the Burg entropy $h'(x) = -\frac{1}{x}$, $g(x) = -\frac{1}{x}$ too and

$$p_i = -\frac{p_i^*}{\mu_0 + u_i\mu} \tag{57}$$

For the Lagrange multipliers $\mu_0, \mu$ we have a system of two algebraic equations

$$\sum_{i=1}^n \frac{p_i^*}{\mu_0 + u_i\mu} = -1, \; \sum_{i=1}^n \frac{p_i^* u_i}{\mu_0 + u_i\mu} = -U \tag{58}$$

For the convex combination of the BGS and Burg entropies $h'(x) = \beta \ln x - \frac{1-\beta}{x}$ $(0 < \beta < 1)$, and the function $x = g(y)$ is a solution of a transcendent equation

$$\beta \ln x - \frac{1-\beta}{x} = y \tag{59}$$

Such a solution exists for all real $y$ because this $h'(x)$ is a (monotonic) bijection of $]0, \infty[$ on the real line.

Solution to Equation (59) can be represented through a special function, the Lambert function [65]. This function is a solution to the transcendent equation

$$w\mathrm{e}^w = z$$

and is also known as $W$ function, $\Omega$ function or modified logarithm $\mathrm{lm}z$ [36]. Below we use the main branch $w = \mathrm{lm}z$ for which $\mathrm{lm}z > 0$ if $z > 0$ and $\mathrm{lm}0 = 0$. Let us write (59) in the form

$$\ln x - \frac{\delta}{x} = -\Lambda \tag{60}$$

where $\delta = (1-\beta)/\beta$, $\Lambda = -y/\beta$. Then

$$x = \mathrm{e}^{-\Lambda}\mathrm{e}^{\mathrm{lm}(\delta \mathrm{e}^{\Lambda})}$$

Another equivalent representation of the solution gives

$$x = \frac{\delta}{\mathrm{lm}(\delta \mathrm{e}^{\Lambda})}$$

Indeed, let us take $z = \delta/x$ and calculate exponent of both sides of (60). After simple transformations, we obtain $z\mathrm{e}^z = \delta \mathrm{e}^{\Lambda}$.

The identity $\mathrm{lm}a = \ln a - \ln \mathrm{lm}a$ is convenient for algebraic operations with this function. Many other important properties are collected in [65].

The generalized canonical distribution for the convex combination of the BGS and Burg divergence is [36]

$$p_i = p_i^* \mathrm{e}^{-\Lambda_i}\mathrm{e}^{\mathrm{lm}(\delta \mathrm{e}^{\Lambda_i})} = \frac{\delta p_i^*}{\mathrm{lm}(\delta \mathrm{e}^{\Lambda_i})} \tag{61}$$

where $\Lambda_i = -\frac{1}{\beta}(\mu_0 + u_i\mu)$, $\delta = (1-\beta)/\beta$ and equations (55) hold for the Lagrange multipliers.

For small $1-\beta$ (small addition of the Burg entropy to the BGS entropy) we have

$$p_i = p_i^* \left( \mathrm{e}^{-\Lambda_i} + \frac{1-\beta}{\beta} - \frac{(1-\beta)^2}{2\beta^2}\mathrm{e}^{\Lambda_i} \right) + o((1-\beta)^2)$$

For the CR family $h(x) = \frac{x(x^\lambda - 1)}{\lambda(\lambda+1)}$, $h'(x) = \frac{(\lambda+1)x^\lambda - 1}{\lambda(\lambda+1)}$, $g(x) = (\frac{\lambda(\lambda+1)x+1}{(\lambda+1)})^{\frac{1}{\lambda}}$ and

$$p_i = p_i^* \left( \frac{\lambda(\lambda+1)(\mu_0 + u_i\mu) + 1}{(\lambda+1)} \right)^{\frac{1}{\lambda}} \tag{62}$$

For $\lambda = 1$ (a quadratic divergence) we easily get linear equations and explicit solutions for $\mu_0$ and $\mu$. If $\lambda = \frac{1}{2}$ then equations for the Lagrange multipliers (55) become quadratic and also allow explicit

solution. The same is true for $\lambda = \frac{1}{3}$ and $\frac{1}{4}$ but explicit solutions to the correspondent cubic or quartic equations are too cumbersome.

We studied the generalized canonical distributions for one condition $u(P) = U$ and main families of entropies. For the BGS entropy, the method of Lagrange multipliers gives one transcendent equation for the multiplier $\mu_1$ and explicit expression for $\mu_0$ as a function of $\mu_1$ (56). In general, for functions $H_h$, the method gives a system of two equations (55). For the Burg entropy this is a system of algebraic equation (58). For a convex combination of the BGS and the Burg entropies the expression for generalized canonical distribution function includes the special Lambert function (61). For the CR family the generalized canonical distribution is presented by formula (62). for several values of $\lambda$ it can be represented in explicit form. The Tsallis entropy family is a subset of the CR family (up to constant multipliers).

### *7.2. Polyhedron of Generalized Canonical Distributions for the Markov Order*

The set of the most random distributions with respect to the Markov order under given condition consists of those distributions which may be achieved by randomization which has the given equilibrium distribution and does not violate the condition.

In the previous section, this set was characterized for a single condition $\sum_i p_i u_i = U$, $U \neq U^*$ by a system of inequalities and equations (51). It is a polyhedron that is an intersection of the closure of one compartment with the hyperplane of condition. Here we construct the dual description of this polyhedron as a convex envelope of the set of extreme points (vertices).

The Krein–Milman theorem gives general backgrounds of such a representation of convex compact sets in locally convex topological vector spaces [66]: a compact convex set is the closed convex hull of its extreme points. (An extreme point of a convex set $K$ is a point $x \in K$ which cannot be represented as an average $x = \frac{1}{2}(y + z)$ for $y, z \in K$, $y, z \neq x$.)

Let us assume that there are $k+1 \leq n$ different numbers in the set of numbers $u_i(U-U^*)$. There exists the unique surjection $\sigma : \{1, 2, \ldots n\} \to \{1, 2, \ldots k+1\}$ with the following properties: $\sigma(i) < \sigma(j)$ if and only if $u_i(U-U^*) > u_j(U-U^*)$ (hence, $\sigma(i) = \sigma(j)$ if and only if $u_i(U-U^*) = u_j(U-U^*)$). The polyhedron of generalized canonical distributions is the intersection of the condition plane $\sum_i p_i u_i = U$ with the closure of $\mathcal{C}_\sigma$.

This closure is a simplex with vertices $P^*$ and $v_r$ ($r = 1, 2, \ldots k$) (47). The vertices of the intersection of this simplex with the condition hyperplane belong to edges of the simplex, hence we can easily find all of them: the edge $[x, y]$ has nonempty intersection with the condition hyperplane if either $u(x) \geq U \& u(y) \leq U$ or $u(x) \leq U \& u(y) \geq U$. This intersection is a single point $P$ if $u(x) \neq u(y)$:

$$P = \lambda x + (1 - \lambda)y, \quad \lambda = \frac{u(y) - U}{u(y) - u(x)} \tag{63}$$

If $u(x) = u(y)$ then the intersection is the whole edge, and the vertices are $x$ and $y$.

For example, if $U$ is sufficiently close to $U^*$ then the intersection is a simplex with $k$ vertices $w_r$ ($r = 1, 2, \ldots k$). Each $w_r$ is the intersection of the edge $[P^*, v_r]$ with the condition hyperplane.

Let us find these vertices explicitly. We have a system of two equations

$$\begin{aligned} a\sum_{i,\,\sigma(i)\le r} p_i^* + b\sum_{i,\,\sigma(i)>r} p_i^* = 1\,; \\ a\sum_{i,\,\sigma(i)\le r} u_i p_i^* + b\sum_{i,\,\sigma(i)>r} u_i p_i^* = U \end{aligned} \tag{64}$$

Position of the vertex $w_r$ on the edge $[P^*, v_r]$ is given by the following expressions

$$\begin{aligned} &\frac{p_i}{p_i^*} = a, \quad \text{for} \quad \sigma(i) \le r\,; \quad \frac{p_i}{p_i^*} = b, \quad \text{for} \quad \sigma(i) > r \\ &a = 1 + \frac{(U-U^*)\sum_{i,\,\sigma(i)>r} p_i^*}{\sum_{i,\,\sigma(i)>r} p_i^* \sum_{i,\,\sigma(i)\le r} u_i p_i^* - \sum_{i,\,\sigma(i)\le r} p_i^* \sum_{i,\,\sigma(i)>r} u_i p_i^*} \\ &b = 1 - \frac{(U-U^*)\sum_{i,\,\sigma(i)\le r} p_i^*}{\sum_{i,\,\sigma(i)>r} p_i^* \sum_{i,\,\sigma(i)\le r} u_i p_i^* - \sum_{i,\,\sigma(i)\le r} p_i^* \sum_{i,\,\sigma(i)>r} u_i p_i^*} \end{aligned} \tag{65}$$

If $b \ge 0$ for all $r$ then the polyhedron of generalized canonical distributions is a simplex with vertices $w_r$. If the solution becomes negative for some $r$ then the set of vertices changes qualitatively and some of them belong to the base of $\mathcal{C}_\sigma$. For example, in Figure 3a the interval of the generalized canonical distribution (1D polyhedron) has vertices of two types: one belongs to the lateral face, another is situated on the basement of the compartment. In Figure 3b both vertices belong to the lateral faces.

Vertices $w_r$ on the edges $[P^*, v_r]$ have very special structure: the ratio $p_i/p_i^*$ can take for them only two values, it is either $a$ or $b$.

Another form for representation of vertices $w_r$ (65) can be found as follows. $w_r$ belongs to the edge $[P^*, v_r]$, hence, $w_r = \lambda P^* + (1-\lambda)v_r$ for some $\lambda \in [0,1]$. Equation for the value of $\lambda$ follows from the condition $u(w_r) = U$: $\lambda U^* + (1-\lambda)u(v_r) = U$. Hence, we can use (63) with $x = P^*$, $y = v_r$.

For sufficiently large value of $U - U^*$ for some of these vertices $b$ loses positivity, and instead of them the vertices on edges $[v_r, v_q]$ (47) appear.

There exists a vertex on the edge $[v_r, v_q]$ if either $u(v_r) \ge U \& u(v_q) \le U$ or $u(v_r) \le U \& u(v_q) \ge U$. If $u(v_r) \ne u(v_q)$ then his vertex has the form $P = \lambda v_r + (1-\lambda)v_q$ and for $\lambda$ the condition $u(P) = U$ gives (63) with $x = v_r$, $y = v_q$. If $u(v_r) = u(v_q)$ then the edge $[u(v_r), u(v_q)]$ belongs to the condition plane and the extreme distributions are $u(v_r)$ $u(v_q)$.

For each of $v_r$ the ratio $p_i/p_i^*$ can take only two values: $a_r$ or 0. Without loss of generality we can assume that $q > r$. For a convex combination $\lambda v_r + (1-\lambda)v_q$ $(1 > \lambda > 0)$ the ratio $p_i/p_i^*$ can take three values: $\lambda a_r + (1-\lambda)a_q$ (for $\sigma(i) \le r$), $(1-\lambda)a_q$ (for $r < \sigma(i) \le q$) and 0 (for $\sigma(i) > q$).

The case when a vertex is one of the $v_r$ is also possible. In this case, there are two possible values of $p_i/p_i^*$, it is either $a_r$ or 0.

All the generalized canonical distributions from the polyhedron are convex combinations of its extreme points (vertices). If the set of vertices is $\{w_r\}$, then for any generalized canonical distributions $P = \sum \lambda_i w_i$ ($\lambda_i \ge 0$, $\sum_i \lambda_i = 1$). The vertices can be found explicitly. Explicit formulas for the extreme generalized canonical distributions are given in this section: (65) and various applications of (63). These formulas are based on the description of compartment $\mathcal{C}_\sigma$ given in Proposition 7 and Equation (47).

## 8. History of the Markov Order

### *8.1. Continuous Time Kinetics*

We have to discuss the history of the Markov order in the wider context of orders, with respect to which the solutions of kinetic equations change monotonically in time. The Markov order is a nice and constructive example of such an order and at the same time the prototype of all of them (similarly the Master Equation is a simple example of kinetic equations and, at the same time, the prototype of all kinetic equations).

The idea of orders and attainable domains (the lower cones of these orders) in phase space was developed in many applications: from biological kinetics to chemical kinetics and engineering. A kinetic model includes information of various levels of detail and of variable reliability. Several types of building block are used to construct a kinetic model. The system of these building blocks can be described, for example, as follows:

1. The list of components (in chemical kinetics) or populations (in mathematical ecology) or states (for general Markov chains);

2. The list of elementary processes (the reaction mechanism, the graph of trophic interactions or the transition graph), which is often supplemented by the lines or surfaces of partial equilibria of elementary processes;

3. The reaction rates and kinetic constants.

We believe that the lower level information is more accurate and reliable: we know the list of component better than the mechanism of transitions, and our knowledge of equilibrium surfaces is better than the information about exact values of kinetic constants.

It is attractive to use the more reliable lower level information for qualitative and quantitative study of kinetics. Perhaps, the first example of such a analysis was performed in biological kinetics.

In 1936, A.N. Kolmogorov [67] studied the dynamics of a pair of interacting populations of prey ($x$) and predator ($y$) in general form:

$$\dot{x} = xS(x,y), \quad \dot{y} = yW(x,y)$$

under monotonicity conditions: $\partial S(x,y)/\partial y < 0$, $\partial W(x,y)/\partial y < 0$. The zero isoclines, the lines at which the rate of change for one population is zero (given by equations $S(x,y) = 0$ or $W(x,y) = 0$), are graphs of two functions $y(x)$. These isoclines divide the phase space into compartments (generically with curvilinear borders). In every compartment the angle of possible directions of motion is given (compare to Figure 1).

Analysis of motion in these angles gives information about dynamics without an exact knowledge of the kinetic constants. The geometry of the zero isoclines intersection together with some monotonicity conditions give important information about the system dynamics [67] without exact knowledge of the right hand sides of the kinetic equations.

This approach to population dynamics was further developed by many authors and applied to various problems [68,69]. The impact of this work on population dynamics was analyzed by K. Sigmund in review [70].

It seems very attractive to use an attainable region instead of the single trajectory in situations with incomplete information or with information with different levels of reliability. Such situations are typical in many areas of engineering. In 1964, F. Horn proposed to analyze the attainable regions for chemical reactors [71]. This approach was applied both to linear and nonlinear kinetic equations and became popular in chemical engineering. It was applied to the optimization of steady flow reactors [72], to batch reactor optimization by use of tendency models without knowledge of detailed kinetics [73] and for optimization of the reactor structure [74]. Analysis of attainable regions is recognized as a special geometric approach to reactor optimization [75] and as a crucially important part of the new paradigm of chemical engineering [76]. Plenty of particular applications was developed: from polymerization [77] to particle breakage in a ball mill [78]. Mathematical methods for study of attainable regions vary from the Pontryagin's maximum principle [79] to linear programming [80], the Shrink-Wrap algorithm [81] and convex analysis.

The connection between attainable regions, thermodynamics and stoichiometric reaction mechanisms was studied by A.N. Gorban in the 1970s. In 1979, he demonstrated how to utilize the knowledge about partial equilibria of elementary processes to construct the attainable regions [62].

He noticed that the set (a cone) of possible direction for kinetics is defined by thermodynamics and the reaction mechanism (the system of the stoichiometric equation of elementary reactions).

Thermodynamic data are more robust than the reaction mechanism and the reaction rates are known with lower accuracy than the stoichiometry of elementary reactions. Hence, there are two types of attainable regions. The first is the thermodynamic one, which use the linear restrictions and the thermodynamic functions [82]. The second is generated by thermodynamics and stoichiometric equations of elementary steps (but without reaction rates) [62,83].

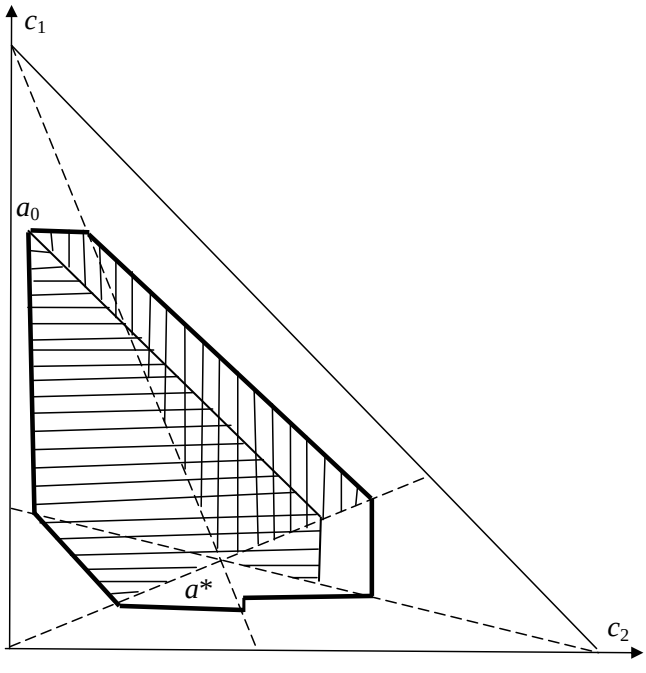
It was demonstrated that the attainable regions significantly depend on the transition mechanism (Figure 4) and it is possible to use them for the mechanisms discrimination [84].

Already simple examples demonstrate that the sets of distributions which are accessible from a given initial distribution by Markov processes with equilibrium are, in general, *non-convex* polytopes [62,85] (see, for example, the outlined region in Figure 4, or, for particular graphs of transitions, any of the shaded regions there). This non-convexity makes the analysis of attainability for continuous time Markov processes more difficult (and also more intriguing).

This approach was developed for all thermodynamic potentials and for open systems as well [34]. Partially, the results are summarized in [14,63].

This approach was rediscovered by F.J. Krambeck [86] for linear systems, that is, for Markov chains, and by R. Shinnar and other authors [87] for more general nonlinear kinetics. There was even an open discussion about priority [89]. Now this geometric approach is applied to various chemical and industrial processes.

**Figure 4**. Attainable regions from an initial distribution $a_0$ for a linear system with three components $A_1, A_2, A_3$ in coordinates $c_1, c_2$ (concentrations of $A_1, A_2$) ($c_3 = \text{const} - c_1 - c_2$) [62]: for a full mechanism $A_1 \rightleftarrows A_2 \rightleftarrows A_3 \rightleftarrows A_1$ (outlined region), for a two-step mechanism $A_1 \rightleftarrows A_2$, $A_1 \rightleftarrows A_3$ (horizontally shaded region) and for a two-step mechanism $A_1 \rightleftarrows A_2$, $A_2 \rightleftarrows A_3$ (vertically shaded region). Equilibrium is $a^*$. The dashed lines are partial equilibria.




### *8.2. Discrete Time Kinetics*

In our paper we deal mostly with continuous time Markov chains. For the discrete time Markov chains, the attainable regions have two important properties: they are convex and symmetric with respect to permutations of states. Because of this symmetry and convexity, the discrete time Markov order is characterized in detail. As far as we can go in history, this work was begun in early 1970s by A. Uhlmann and P.M. Alberti. The results of the first 10 years of this work were summarized in monograph [90]. A more recent bibliography (more than 100 references) is collected in review [91].

This series of work was concentrated mostly on processes with uniform equilibrium (doubly stochastic maps). The relative majorization, which we also use in Section 5, and the Markov order with respect to a non-uniform equilibrium was introduced by P. Harremoës in 2004 [92]. He used formalism based on the Lorenz diagrams.

## 9. Conclusion

Is playing with non-classical entropies and divergences just an extension to the fitting possibilities (no sense—just fitting)? We are sure now that this is not the case: two one-parametric families of non-classical divergences are distinguished by the very natural properties:

1. They are Lyapunov functions for all Markov chains;

2. They become additive with respect to the joining of independent systems after a monotone transformation of scale;

3. They become additive with respect to a partitioning of the state space after a monotone transformation of scale.

Two families of smooth divergences (for positive distributions) satisfy these requirements: the Cressie–Read family [29,30]

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

and the convex combination of the Burg and Shannon relative entropies [34,35]:

$$H(P\|P^*) = \sum_i (\beta p_i - (1-\beta)p_i^*)\log\left(\frac{p_i}{p_i^*}\right)\ ,\ \ \beta\in[0,1]$$

If we relax the differentiability property, then we have to add to the the CR family two limiting cases:

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

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

Beyond these two distinguished one-parametric families there is the whole world of the Csiszár–Morimoto Lyapunov functionals for the Master equation (6). These functions monotonically decrease along any solution of the Master equation. The set of all these functions can be used to reduce the uncertainty by conditional minimization: for each $h$ we could find a conditional minimizer of $H_h(p)$.

Most users prefer to have an unambiguous choice of entropy: 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. 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.

The task to minimize functions $H_h(p)$ which depend on a functional parameter $h$ seems too complicated. The Markov order gives us another way for the evaluation of the set of possible "maximally random" probability distribution, and this evaluation is, in some sense, the best one. We defined the Markov order, studied its properties and demonstrated how it can be used to reduce uncertainty.

It is quite surprising that the Markov order is generated by the reversible Markov chains which satisfy the detailed balance principle. We did not include any reversibility assumptions and studied the general Markov chains. There remain some questions about the structure and full description of the global Markov order. Nevertheless, to find the set of conditionally extreme ("most random") probability distributions, we need the local Markov order only. This local order is fully described in Section 5.2 and has a very clear geometric structure. For a given equilibrium distribution $P^*$, the simplex of probability distributions is divided by $n(n-1)/2$ hyperplanes of "partial equilibria" (this terminology

comes from chemical kinetics [62,63]): $\frac{p_i}{p_i^*} = \frac{p_j}{p_j^*}$ (there is one hyperplane for each pair of states $(i, j)$). In each compartment a cone of all possible time derivatives of the probability distribution is defined as a conic envelope of $n(n-1)/2$ vectors (41). The extreme rays of this cone are explicitly described in Proposition 6 (44). This cone defines the local Markov order. When we look for conditionally extreme distributions, this cone plays the same role as a hyperplane given by entropy growth condition ( $\mathrm{d}S/\mathrm{d}t > 0$) in the standard approach.

For the problem of the generalized canonical (or reference) distribution the Markov order gives a polyhedron of the extremely disordered distributions. The vertices of that polyhedron can be computed explicitly.

The construction of efficient algorithms for numerical calculation of conditionally extreme compacts in high dimensions is a challenging task for our future work as well as the application of this methodology to real life problems.

**Acknowledgements**

Suggestions from Mike George, Marian Grendar, Ivan Tyukin and anonymous referees are gratefully acknowledged.

**References**

1. Clausius, R. Über vershiedene für die Anwendungen bequeme Formen der Hauptgleichungen der Wärmetheorie. *Poggendorffs Annalen der Physic und Chemie* **1865**, *125*, 353–400.
2. Gibbs, J.W. On the equilibrium of heterogeneous substance. *Trans. Connect. Acad.* **1875–1876**, 108–248; **1877–1878**, 343–524.
3. Boltzmann, L. Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen. *Sitzungsberichte der keiserlichen Akademie der Wissenschaften* **1872**, *66*, 275–370. Translation: Further studies on the thermal equilibrium of gas molecules, In *Kinetic Theory of Gases: An Anthology of Classic Papers With Historical Commentary*; Brush, S.G.; Hall, N.S., Eds.; Imperial Colledge Press: London, UK, 2003; pp. 362–368.
4. Gibbs, J.W. *Elementary principles in statistical mechanics*; Ox Bow Press: New York, NY, USA, 1981.
5. Villani, C. $H$-theorem and beyond: Boltzmann's entropy in today's mathematics. In *Boltzmann's Legacy*; Gallavotti, G.; Reiter, W.L.; Yngvason J., Eds.; EMS Publishing House: Zürich, Switzerland, 2008; pp. 129–145.
6. Jaynes, E.T. Gibbs versus Boltzmann entropy. *Am. J. Phys.* **1965**, *33*, 391–398.
7. Goldstein, S.; Lebowitz, J.L. On the (Boltzmann) entropy of non-equilibrium systems. *Physica D* **2004**, *193*, 53–66.
8. Grmela, M.; Öttinger, H.C. Dynamics and thermodynamics of complex fluids. I. Development of a general formalism. *Phys. Rev. E* **1997**, *56*, 6620–6632.
9. Öttinger, H.C. *Beyond Equilibrium Thermodynamics*; Wiley-Interscience: Hoboken, NJ, USA, 2005.
10. Hartley, R.V.L. Transmission of information. *Bell System Technical Journal* **1928**, *July*, 535–563.

11. Shannon, C.E. A mathematical theory of communication. *The Bell System Technical Journal* **1948**, *27*, 379–423, 623–656.
12. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Statist.* **1951**, *22*: 79–86.
13. Zeldovich, Y.B. proof of the uniqueness of the solution of the equations of the law of mass action. In *Selected Works of Yakov Borisovich Zeldovich*; Volume 1, Ostriker, J.P., Ed.; Princeton University Press: Princeton, NJ, USA, 1996; pp. 144–148.
14. Yablonskii, G.S.; Bykov, V.I.; Gorban, A.N.; Elokhin, V.I. *Kinetic Models of Catalytic Reactions* (Series "Comprehensive Chemical Kinetics," Volume 32); Elsevier: Amsterdam, The Netherlands, 1991.
15. Hangos, K.M. Engineering model reduction and entropy-based lyapunov functions in chemical reaction kinetics. *Entropy* **2010**, *12*, 772–797.
16. Burg, J.P. Maximum entropy spectral analysis. In *Proceedings of the 37th meeting of the Society of Exploration Geophysicists*; Oklahoma City, OK, USA, 1967. Reprinted in *Modern Spectrum Analysis*; Childers, D.G., Ed.; IEEE Press: New York, NY, USA, 1978; pp. 34–39.
17. Burg, J.P. The relationship between maximum entropy spectra and maximum likelihood spectra. *Geophysics* **1972**, *37*, 375–376.
18. Jaynes, E.T. Information Theory and Statistical Mechanics, *Phys. Rev.* **1957**, *106*, 620–630.
19. Jaynes, E.T. Information theory and statistical mechanics. II. *Phys. Rev.* **1957**, *108*, 171–190.
20. Harremoës P.; Topsøe F. Maximum entropy fundamentals. *Entropy* **2001**, *3*, 191–226.
21. Beck, C. Generalized information and entropy measures in physics. *Contemp. Phys.* **2009**, *50*, 495–510.
22. Mittelhammer, R.; Judge, G.; Miller, D. *Econometric Foundations*; Cambridge University Press: New York, NY, USA, 2000.
23. Van Akkeren, M.; Judge, G.; Mittelhammer, R. Generalized moment based estimation and inference. *J. Econom.* **2002**, *107*, 127–148. 127-148.
24. Myers, T.S.; Daniel, N.; Osherson, D.N. On the psychology of ampliative inference. *Psychol. Sci.* **1992**, *3*, 131–135.
25. Esteban, M.D.; Morales, D. A summary of entropy statistics, *Kybernetica* **1995**, *31*, 337–346.
26. Rényi, A. On measures of entropy and information.In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability 1960*; University of California Press: Berkeley, CA, USA, 1961; Volume 1; pp. 547–561.
27. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.* **1988**, *52*, 479–487.
28. Abe, S., Okamoto, Y. Eds. *Nonextensive Statistical Mechanics and its Applications*; Springer: Heidelberg, Germany, 2001.
29. Cressie, N.; Read, T. Multinomial Goodness of Fit Tests. *J. R. Stat. Soc. Ser. B* **1984**, *46*, 440–464.
30. Read, T.R., Cressie, N.A. *Goodness of Fit Statistics for Discrete Multivariate Data*; Springer: New York, NY, USA, 1988.
31. Bagci, G.B.; Tirnakli, U. On the way towards a generalized entropy maximization procedure. *Phys. Lett. A* **2009**, *373*, 3230–3234.
32. Cho, A. A fresh take on disorder, or disorderly science? *Science* **2002**, *297*, 1268-1269.
33. Lin, S-K. Diversity and entropy. *Entropy* **1999**, *1*, 1–3.

34. Gorban, A.N. *Equilibrium Encircling. Equations of Chemical kinetics and their Thermodynamic Analysis*; Nauka: Novosibirsk, Russia, 1984.
35. Gorban, A.N.; Karlin, I.V. Family of additive entropy functions out of thermodynamic limit. *Phys. Rev. E* **2003** *67*, 016104. Available online: http://arxiv.org/abs/cond-mat/0205511 (accessed on 4 May 2010)
36. Gorban, A.N.; Karlin, I.V.; Öttinger, H.C. The additive generalization of the Boltzmann entropy. *Phys. Rev. E* **2003** *67*, 067104. Available online: http://arxiv.org/abs/cond-mat/0209319 (accessed on 4 May 2010).
37. *Nonextensive statistical mechanics and thermodynamics: bibliography*. Available online: http://tsallis.cat.cbpf.br/TEMUCO.pdf (accessed on 4 May 2010).
38. Petz, D. From $f$-divergence to quantum quasi-entropies and their use. *Entropy* **2010**, *12*, 304–325.
39. Cachin, C. Smooth entropy and Rényi entropy, In *Proceedings of EUROSCRIPT'97*; W. Fumy at al, eds., LNCS, Vol. 1233, Springer: New York, NY, USA, 1997; pp. 193–208.
40. Davis, J.V.; Kulis, B.; Jain, P.; Sra, S.; Dhillon, I.S. Information-theoretic metric learning. *ICML* **2007**, *227*, 209–216.
41. Rényi, A. *Probability theory*; North-Holland: Amsterdam, The Netherlands, 1970.
42. Abe, S. Axioms and uniqness theorem for Tsallis entropy. *Phys. Lett. A* **2000**, *271*, 74–79.
43. Aczel, J. Measuring information beyond communication theory. *Information Processing and Management* **1984**, *20*, 383–395.
44. Aczel, J.; Daroczy, Z. *On Measures of Information and Their Characterizations*; Academic Press: New York, NY, USA, 1975.
45. Csiszár, I. Information measures: a critical survey. In *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions and the Eighth European Meeting of Statisticians, Prague, Czech Republic, 18 August–23 August 1974*; Academia: Prague, Czech Republic, 1978; Volume B, pp. 73–86.
46. Gorban, P. Monotonically equivalent entropies and solution of additivity equation. *Physica A* **2003**, *328*, 380–390. Available online: http://arxiv.org/abs/cond-mat/0304131 (accessed on 4 May 2010).
47. Wehrl, A. General properties of entropy. *Rev. Mod. Phys.* **1978**, *50*, 221–260.
48. Morimoto, T. Markov processes and the $H$-theorem. *J. Phys. Soc. Jap.* **1963**, *12*, 328–331.
49. Csiszár, I. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Magyar. Tud. Akad. Mat. Kutató Int. Közl.* **1963**, *8*, 85–108.
50. Rockafellar, R.T. *Convex Analysis*; Princeton University Press: Princeton, NJ, USA, 1970. Reprint: 1997.
51. Liese, F.; Vajda, I. *Convex Statistical Distances*; Teubner: Leipzig, Germany, 1987.
52. Dobrushin, R.L. Central limit theorem for non-stationary Markov chains I, II. *Theory Probab. Appl.* **1956**, *1*, 163–80, 329–383.
53. Seneta, E. *Nonnegative Matrices and Markov Chains*; Springer: New York, USA, 1981.
54. Cohen, J.E.; Derriennic, Y.; Zbaganu GH.; Majorization, Monotonicity of Relative Entropy and stochastic matrices. *Contemp. Math.* **1993**, *149*, 251–259.

55. Cohen, J.E.; Iwasa, Y.; Rautu, G.; Ruskai, M.B.; Seneta, E.; Zbaganu, G. Relative entropy under mappings by stochastic matrices. *Linear. Alg. Appl.* **1993**, *179*, 211–235.
56. Del Moral, P.; Ledoux, M.; Miclo, L. On contraction properties of Markov kernels. *Probab. Theory Relat. Field* **2003**, *126*, 395–420.
57. Pressé, S.; Ghosh, K.; Lee, J.; Dill, K.A. Nonadditive Entropies Yield Probability Distributions with Biases not Warranted by the Data. *Phys. Rev. Lett.* **2013**, *111*, 180604.
58. Shore, J.E.; Johnson, R.W. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory* **1980**, *IT-26*, 26–37.
59. Amari, S. *Divergence, Optimization, Geometry*. In Proceedings of the 16th International Conference on Neural Information Processing; Leung, C.S., Lee, M., Chan, J.H., Eds.; LNCS 5863; Springer: Berlin, Germany, 2009; pp. 185–193.
60. Bregman, L.M. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics* **1967**, *7*, 200–217.
61. Itakura, F.; Saito, S. Analysis synthesis telephony based on the maximum likelihood method. In *Proceedings of the 6th International Congress on Acoustics*; Tokyo, Japan, 1968; pp. C17–C20. Reprinted in *Speech Synthesis*, Flanagan, J.L., Rabiner, L.R., Eds.; Dowden, Hutchinson & Ross: Stroudsburg, PA, USA, 1973; pp. 289–292.
62. Gorban, A.N. Invariant sets for kinetic equations. *React. Kinet. Catal. Lett.* **1979**, *10*, 187–190.
63. Gorban, A.N.; Kaganovich, B.M.; Filippov, S.P.; Keiko, A.V.; Shamansky, V.A.; Shirkalin, I.A. *Thermodynamic Equilibria and Extrema: Analysis of Attainability Regions and Partial Equilibria*; Springer: New York, NY, USA, 2006.
64. Sion, M. On general minimax theorems. *Pacific J. Math.* **1958**, *8*, 171–176.
65. Corless, R.M.; Gonnet, G.H.; Hare, D.E.G.; Jeffrey, D.J.; Knuth, D.E. On the Lambert $W$ function. *Adv. Comput. Math.* **1996**, *5*, 329–359.
66. Edwards, R.E. *Functional Analysis: Theory and Applications*; Dover Publications: New York, NY, USA, 1995.
67. Kolmogorov, A.N. Sulla teoria di Volterra della lotta per l'esistenza. *Giornale Instituto Ital. Attuari* **1936**, *7*, 74–80.
68. May, R.M.; Leonard, W.J. Nonlinear Aspects of Competition Between Three Species. *SIAM Journal on Applied Mathematics* **1975**, *29*, 243–253.
69. Bazykin, A.D. *Nonlinear dynamics of interacting populations*; World Scientific Publishing: Singapore, 1998.
70. Sigmung, K. Kolmogorov and population dynamics. In *Kolmogorovs Heritage in Mathematics*, Charpentier, É.; Lesne, A.; Nikolski, N.K., Eds.; Springer: Berlin, Germany, 2007; pp. 177–186.
71. Horn, F. Attainable regions in chemical reaction technique. In *The Third European Symposium on Chemical Reaction Engineering*; Pergamon Press: London, UK, 1964, pp. 1–10.
72. 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.

73. Filippi-Bossy, C.; Bordet, J.; Villermaux, J.; Marchal-Brassely, S.; Georgakis C. Batch reactor optimization by use of tendency models. *Comput. Chem. Eng.* **1989**, *13*, 35–47
74. Hildebrandt, D.; Glasser, D. The attainable region and optimal reactor structures. *Chem. Eng. Sci.* **1990**, *45*, 2161–2168.
75. Feinberg, M.; Hildebrandt, D. Optimal reactor design from a geometric viewpoint–I. Universal properties of the attainable region. *Chem. Eng. Sci.* **1997**, *52*, 1637–1665.
76. Hill, M. Chemical product engineering – the third paradigm. *Comput. Chem. Eng.* **2009**, *33*, 947–953.
77. Smith, R.L.; Malone, M.F. Attainable regions for polymerization reaction systems. *Ind. Eng. Chem. Res.* **1997**, *36*, 1076–1084.
78. Metzger, M.J.; Glasser, D., Hausberger, B.; Hildebrandt, D.; Glasser B.J. Use of the attainable region analysis to optimize particle breakage in a ball mill. *Chem. Eng. Sci.* **2009** *64*, 3766–3777.
79. McGregor, C; Glasser, D; Hildebrandt, D. The attainable region and Pontryagin's maximum principle. *Ind. Eng. Chem. Res.* **1999**, *38*, 652–659.
80. Kauchali, S.; Rooney, W. C.; Biegler, L. T.; Glasser, D.; Hildebrandt, D. Linear programming formulations for attainable region analysis. *Chem. Eng. Sci.* **2002**, *57*, 2015–2028.
81. Manousiouthakis, VI; Justanieah, AM; Taylor, LA. The Shrink-Wrap algorithm for the construction of the attainable region: an application of the IDEAS framework. *Comput. Chem. Eng.* **2004**, *28*, 1563–1575.
82. Gorban, A.N. Methods for qualitative analysis of chemical kinetics equations. In *Numerical Methods of Continuum Mechanics*; Institute of Theoretical and Applied Mechanics: Novosibirsk, USSR, 1979; Volume 10, pp. 42–59.
83. Gorban, A.N.; Yablonskii, G.S.; Bykov, V.I. Path to equilibrium. In *Mathematical Problems of Chemical Thermodynamics*; Nauka: Novosibirsk, 1980; pp. 37–47 (in Russian). English translation: *Int. Chem. Eng.* **1982**, *22*, 368–375.
84. Gorban, A.N.; Yablonskii, G.S. On one unused possibility in planning of kinetic experiment. *Dokl. Akad. Nauk SSSR* **1980**, *250*, 1171–1174.
85. Zylka, Ch. A note on the attainability of states by equalizing processes. *Theor. Chim. Acta* **1985**, *68*, 363–377.
86. Krambeck, F.J. Accessible composition domains for monomolecular systems. *Chem. Eng. Sci.* **1984**, *39*, 1181–1184.
87. Shinnar, R.; Feng, C.A. Structure of complex catalytic reactions: thermodynamic constraints in kinetic modeling and catalyst evaluation. *Ind. Eng. Chem. Fundam.* **1985**, *24*, 153–170.
88. Shinnar, R. Thermodynamic analysis in chemical process and reactor design. *Chem. Eng. Sci.* **1988**, *43*, 2303–2318.
89. Bykov, V.I. Comments on "Structure of complex catalytic reactions: thermodynamic constraints in kinetic modeling and catalyst evaluation". *Ind. Eng. Chem. Res.* **1987**, *26*, 1943–1944.
90. Alberti, P.M.; Uhlmann, A. *Stochasticity and Partial Order – Doubly Stochastic Maps and Unitary Mixing*; Mathematics and its Applications 9; D. Reidel Publ. Company: Dordrecht-Boston-London, 1982.

91. Alberti, P.M.; Crell, B.; Uhlmann, A.; Zylka, C. Order structure (majorization) and irreversible processes. In *Vernetzte Wissenschaften–Crosslinks in Natural and Social Sciences*; Plath, P.J.; Hass, E.-Chr., Eds.; Logos Verlag: Berlin, Germany, 2008; pp. 281–290.
92. Harremoës, P. A new look on majorization. In *Proceedings ISITA 2004, Parma, Italy*; IEEE/SITA, 2004; pp. 1422–1425. ISBN 4-902087-08-1.
93. Harremoës P.; Tishby, N. The information bottleneck revisited or how to choose a good distortion measure. In *Proceedings ISIT 2007, Nice, France*; IEEE Information Theory Society, 2007; pp. 566–571. DOI: 10.1109/ISIT.2007.4557285.
94. Aczél, J. *Lectures on Functional Equations and Their Applications*; Academic Press: New York, NY, USA, 1966.


## Appendix

*Proof of Theorem 1*. The problem is to find all such universal and trace–form Lyapunov functions $H$ for Markov chains, that there exists a monotonous function $F$, such that $F(H(P)) = F(H(Q)) + F(H(R))$ if $P = p_{ij} = q_i r_j$.

With Lemma 1 we get that

$$H(P) = \sum_{i,j} q_i^* r_j^* h\left(\frac{q_i r_j}{q_i^* r_j^*}\right), \quad H(Q) = \sum_i q_i^* h\left(\frac{q_i}{q_i^*}\right), \quad H(R) = \sum_j r_j^* h\left(\frac{r_j}{r_j^*}\right)$$

Let $F(x)$ and $h(x)$ be differentiable as many times as needed. Differentiating the equality $F(H(P)) = F(H(Q)) + F(H(R))$ on $r_1$ and $q_1$ taking into account that $q_n = 1 - \sum_{i=1}^{n-1} q_i$ and $r_m = 1 - \sum_{j=1}^{m-1} r_j$ we get that $F'(H(P))H''_{q_1 r_1}(P) = -F''(H(P))H'_{q_1}(P)H'_{r_1}(P)$, or, if $-\frac{F'(H(P))}{F''(H(P))} = G(H(P))$ then

$$G(H(P)) = \frac{H'_{q_1}(P)H'_{r_1}(P)}{H''_{q_1 r_1}(P)} \tag{66}$$

It is possible if and only if every linear differential operator of the first order, which annulates $H(P)$ and $\sum p_i$, annulates also

$$\frac{H'_{q_1}(P)H'_{r_1}(P)}{H''_{q_1 r_1}(P)} \tag{67}$$

and it means that every differential operator which has the form

$$D = \left(\frac{\partial H(P)}{\partial q_\gamma} - \frac{\partial H(P)}{\partial q_\alpha}\right)\frac{\partial}{\partial q_\beta} + \left(\frac{\partial H(P)}{\partial q_\beta} - \frac{\partial H(P)}{\partial q_\gamma}\right)\frac{\partial}{\partial q_\alpha} + \left(\frac{\partial H(P)}{\partial q_\alpha} - \frac{\partial H(P)}{\partial q_\beta}\right)\frac{\partial}{\partial q_\gamma} \tag{68}$$

annulates (67). For $\beta = 2, \alpha = 3, \gamma = 4$ we get the following equation

$$\begin{aligned}
&F_1(Q,R)\left[h'\left(\tfrac{q_2 r_1}{q_2^* r_1^*}\right) - h'\left(\tfrac{q_2 r_m}{q_2^* r_m^*}\right) + \tfrac{q_2 r_1}{q_2^* r_1^*}h''\left(\tfrac{q_2 r_1}{q_2^* r_1^*}\right) - \tfrac{q_2 r_m}{q_2^* r_m^*}h''\left(\tfrac{q_2 r_m}{q_2^* r_m^*}\right)\right] + \\
&F_2(Q,R)\left[h'\left(\tfrac{q_3 r_1}{q_3^* r_1^*}\right) - h'\left(\tfrac{q_3 r_m}{q_3^* r_m^*}\right) + \tfrac{q_3 r_1}{q_3^* r_1^*}h''\left(\tfrac{q_3 r_1}{q_3^* r_1^*}\right) - \tfrac{q_3 r_m}{q_3^* r_m^*}h''\left(\tfrac{q_3 r_m}{q_3^* r_m^*}\right)\right] + \\
&F_3(Q,R)\left[h'\left(\tfrac{q_4 r_1}{q_4^* r_1^*}\right) - h'\left(\tfrac{q_4 r_m}{q_4^* r_m^*}\right) + \tfrac{q_4 r_1}{q_4^* r_1^*}h''\left(\tfrac{q_4 r_1}{q_4^* r_1^*}\right) - \tfrac{q_4 r_m}{q_4^* r_m^*}h''\left(\tfrac{q_4 r_m}{q_4^* r_m^*}\right)\right] = 0
\end{aligned} \tag{69}$$

where

$$F_1(Q,R)=\sum_j r_j\left[h'\left(\frac{q_4 r_j}{q_4^* r_j^*}\right)-h'\left(\frac{q_3 r_j}{q_3^* r_j^*}\right)\right];$$
$$F_2(Q,R)=\sum_j r_j\left[h'\left(\frac{q_2 r_j}{q_2^* r_j^*}\right)-h'\left(\frac{q_4 r_j}{q_4^* r_j^*}\right)\right];$$
$$F_3(Q,R)=\sum_j r_j\left[h'\left(\frac{q_3 r_j}{q_3^* r_j^*}\right)-h'\left(\frac{q_2 r_j}{q_2^* r_j^*}\right)\right]$$

If we apply the differential operator $\frac{\partial}{\partial r_2}-\frac{\partial}{\partial r_3}$, which annulates the conservation law $\sum_j r_j=1$, to the left part of (69), and denote $f(x)=xh''(x)+h'(x)$, $x_1=\frac{q_2}{q_2^*}$, $x_2=\frac{q_3}{q_3^*}$, $x_3=\frac{q_4}{q_4^*}$, $y_1=\frac{r_1}{r_1^*}$, $y_2=\frac{r_m}{r_m^*}$, $y_3=\frac{r_2}{r_2^*}$, $y_4=\frac{r_3}{r_3^*}$, we get the equation

$$\begin{aligned}&(f(x_3y_3)-f(x_2y_3)-f(x_3y_4)+f(x_2y_4))(f(x_1y_1)-f(x_1y_2))+\\&(f(x_1y_3)-f(x_3y_3)-f(x_1y_4)+f(x_3y_4))(f(x_2y_1)-f(x_2y_2))+\\&(f(x_2y_3)-f(x_1y_3)-f(x_2y_4)+f(x_1y_4))(f(x_3y_1)-f(x_3y_2))=0\end{aligned}\tag{70}$$

or, after differentiation on $y_1$ and $y_3$ and denotation $g(x)=f'(x)$

$$\begin{aligned}&x_1g(x_1y_1)(x_3g(x_3y_3)-x_2g(x_2y_3))+x_2g(x_2y_1)(x_1g(x_1y_3)-\\&-x_3g(x_3y_3))+x_3g(x_3y_1)(x_2g(x_2y_3)-x_1g(x_1y_3))=0\end{aligned}\tag{71}$$

If $y_3=1$, $y_1\neq 0$, $\varphi(x)=xg(x)$, we get after multiplication (71) on $y_1$

$$\varphi(x_1y_1)(\varphi(x_3)-\varphi(x_2))+\varphi(x_2y_1)(\varphi(x_1)-\varphi(x_3))+\varphi(x_3y_1)(\varphi(x_2)-\varphi(x_1))=0\tag{72}$$

It implies that for every three positive numbers $\alpha$, $\beta$, $\gamma$ the functions $\varphi(\alpha x)$, $\varphi(\beta x)$, $\varphi(\gamma x)$ are linearly dependent, and for $\varphi(x)$ the differential equation

$$ax^2\varphi''(x)+bx\varphi'(x)+c\varphi(x)=0\tag{73}$$

holds. This differential equation has solutions of two kinds:

1. $\varphi(x)=C_1x^{k_1}+C_2x^{k_2}$, $k_1\neq k_2$, $k_1$ and $k_2$ are real or complex-conjugate numbers.
2. $\varphi(x)=C_1x^k+C_2x^k\ln x$.

Let us check, which of these solutions satisfy the functional equation (72).

1. $\varphi(x)=C_1x^{k_1}+C_2x^{k_2}$. After substitution of this into (72) and calculations we get

$$C_1C_2(y_1^{k_1}-y_1^{k_2})(x_1^{k_1}x_3^{k_2}-x_1^{k_1}x_2^{k_2}+x_1^{k_2}x_2^{k_1}-x_2^{k_1}x_3^{k_2}+x_2^{k_2}x_3^{k_1}-x_1^{k_2}x_3^{k_1})=0$$

This means that $C_1=0$, or $C_2=0$, or $k_1=0$, or $k_2=0$ and the solution of this kind can have only the form $\varphi(x)=C_1x^k+C_2$.

2. $\varphi(x) = C_1 x^k + C_2 x^k \ln x$. After substitution of this into (72) and some calculations if $y_1 \neq 0$ we get
$$C_2^2((x_1^k - x_2^k)x_3^k \ln x_3 + (x_3^k - x_1^k)x_2^k \ln x_2 + (x_2^k - x_3^k)x_1^k \ln x_1) = 0$$
This means that either $C_2 = 0$ and the solution is $\varphi(x) = C_1 x^k$ or $k = 0$ and the solution is $\varphi(x) = C_1 + C_2 \ln x$.

So, the equation (72) has two kinds of solutions:

1. $\varphi(x) = C_1 x^k + C_2$,
2. $\varphi(x) = C_1 + C_2 \ln x$

Let us solve the equation $f(x) = xh''(x) + h'(x)$ for each of these two cases.

1. $\varphi(x) = C_1 x^k + C_2$, $g(x) = C_1 x^{k-1} + \frac{C_2}{x}$, there are two possibilities:
   1.1) $k = 0$. Then $g(x) = \frac{C}{x}$, $f(x) = C \ln x + C_1$, $h(x) = C_1 x \ln x + C_2 \ln x + C_3 x + C_4$;
   1.2) $k \neq 0$. Then $f(x) = Cx^k + C_1 \ln x + C_2$, and here are also two possibilities:
      1.2.1) $k = -1$. Then $h(x) = C_1 \ln^2 x + C_2 x \ln x + C_3 \ln x + C_4 x + C_5$;
      1.2.2) $k \neq -1$. Then $h(x) = C_1 x^{k+1} + C_2 x \ln x + C_3 \ln x + C_4 x + C_5$;
2. $\varphi(x) = C_1 + C_2 \ln x$; $g(x) = C_1 \frac{\ln x}{x} + \frac{C_2}{x}$; $f(x) = C_1 \ln^2 x + C_2 \ln x + C_3$; $h(x) = C_1 x \ln^2 x + C_2 x \ln x + C_3 \ln x + C_4 x + C_5$.

(We have renamed constants during the calculations).

For the next step let us check, which of these solutions remains a solution to equation (69). The result is that there are just two families of functions $h(x)$ such, that equation (69) holds:

1. $h(x) = Cx^k + C_1 x + C_2$, $k \neq 0$, $k \neq 1$,
2. $h(x) = C_1 x \ln x + C_2 \ln x + C_3 x + C_4$.

The function $h(x)$ should be convex. This condition determines the signs of coefficients $C_i$.

The corresponding divergence $H(P\|P^*)$ is either one of the CR entropies or a convex combination of Shannon's and Burg's entropies up to a monotonic transformation. □

*Characterization of Additive Trace–form Lyapunov Functions for Markov Chains.* We will consider three important properties of Lyapunov functions $H(P\|P^*)$:

1. *Universality*: $H$ is a Lyapunov function for Markov chains (22) with a given equilibrium $P^*$ for every possible values of kinetic coefficients $k_{ij} \geq 0$.
2. $H$ is a *trace–form function*.
$$H(P\|P^*) = \sum_i f(p_i, p_i^*) \tag{74}$$
where $f$ is a differentiable function of two variables.

3. $H$ is *additive* for composition of independent subsystems. It means that if $P = p_{ij} = q_i r_j$ and $P^* = p^*_{ij} = q^*_i r^*_j$ then $H(P\|P^*) = H(Q\|Q^*) + H(R\|R^*)$.

Here and further we suppose $0 < p_i, p^*_i, q_i, q^*_i, r_i, r^*_i < 1$.

We consider the additivity condition as a functional equation and solve it. The following theorem describes all Lyapunov functions for Markov chains, which have all three properties 1) - 3) simultaneously.

Let $f(p, p^*)$ be a twice differentiable function of two variables.

**Theorem 3.** *If a function $H(P\|P^*)$ has all the properties 1)-3) simultaneously, then*

$$f(p, p^*) = p^*_i h\left(\frac{p}{p^*}\right), \quad H(P\|P^*) = \sum_i p^*_i h\left(\frac{p_i}{p^*_i}\right) \tag{75}$$

*where*

$$h(x) = C_1 \ln x + C_2 x \ln x, \; C_1 \le 0, \; C_2 \ge 0 \tag{76}$$

*Proof.* We follow here the P. Gorban proof [46]. Another proof of this theorem was proposed in [93]. Due to Lemma 1 let us take $H(P\|P^*)$ in the form (75). Let $h$ be twice differentiable in the interval $]0, +\infty[$. The additivity equation

$$H(P\|P^*) - H(Q\|Q^*) - H(R\|R^*) = 0 \tag{77}$$

holds. Here (in (77))

$$q_n = 1 - \sum_{i=1}^{n-1} q_i, \; r_m = 1 - \sum_{j=1}^{m-1} r_j, \; P = p_{ij} = q_i r_j$$

$$H(P\|P^*) = \sum_{i,j} q^*_i r^*_j h\left(\frac{q_i r_j}{q^*_i r^*_j}\right), \; H(Q\|Q^*) = \sum_i q^*_i h\left(\frac{q_i}{q^*_i}\right), \; H(R\|R^*) = \sum_j r^*_j h\left(\frac{r_j}{r^*_j}\right)$$

Let us take the derivatives of this equation first on $q_1$ and then on $r_1$. Then we get the equation ($g(x) = h'(x)$)

$$g(\tfrac{q_1 r_1}{q^*_1 r^*_1}) - g(\tfrac{q_n r_1}{q^*_n r^*_1}) - g(\tfrac{q_1 r_m}{q^*_1 r^*_m}) + g(\tfrac{q_n r_m}{q^*_n r^*_m}) +$$
$$+\tfrac{q_1 r_1}{q^*_1 r^*_1} g'(\tfrac{q_1 r_1}{q^*_1 r^*_1}) - \tfrac{q_n r_1}{q^*_n r^*_1} g'(\tfrac{q_n r_1}{q^*_n r^*_1}) - \tfrac{q_1 r_m}{q^*_1 r^*_m} g'(\tfrac{q_1 r_m}{q^*_1 r^*_m}) + \tfrac{q_n r_m}{q^*_n r^*_m} g'(\tfrac{q_n r_m}{q^*_n r^*_m}) = 0$$

Let us denote $x = \frac{q_1 r_1}{q^*_1 r^*_1}$, $y = \frac{q_n r_1}{q^*_n r^*_1}$, $z = \frac{q_1 r_m}{q^*_1 r^*_m}$, and $\psi(x) = g(x) + x g'(x)$. It is obvious that if $n$ and $m$ are more than 2, then $x$, $y$ and $z$ are independent and can take any positive values. So, we get the functional equation:

$$\psi\left(\frac{yz}{x}\right) = \psi(y) + \psi(z) - \psi(x) \tag{78}$$

Let's denote $C_2 = -\psi(1)$ and $\psi_1(\alpha) = \psi(\alpha) - \psi(1)$ and take $x = 1$. We get then

$$\psi_1(yz) = \psi_1(y) + \psi_1(z) \tag{79}$$

the Cauchy functional equation [94]. The solution of this equation in the class of measurable functions is $\psi_1(\alpha) = C_1 \ln \alpha$, where $C_1$ is constant. So we get $\psi(x) = C_1 \ln x + C_2$ and $g(x) + x g'(x) = C_1 \ln x + C_2$.

The solution is $g(x) = \frac{C_3}{x} + C_1 \ln x + C_2 - C_1$; $h(x) = \int (\frac{C_3}{x} + C_1 \ln x + C_2 - C_1)dx = C_3 \ln x + C_1 x \ln x + (C_2 - 2C_1)x + C_4$, or, renaming constants, $h(x) = C_1 \ln x + C_2 x \ln x + C_3 x + C_4$. In the expression for $h(x)$ there are two parasite constants $C_3$ and $C_4$ which occurs because the initial equation was differentiated twice. So, $C_3 = 0$, $C_4 = 0$ and $h(x) = C_1 \ln x + C_2 x \ln x$. Because $h$ is convex, we have $C_1 \leq 0$ and $C_2 \geq 0$. □

So, any universal additive trace–form Lyapunov function for Markov chains is a convex combination of the BGS entropy and the Burg entropy.