Center of Mechanics and Information Technology,
Institute of Fundamental Technological Research,
Polish Academy of Sciences, ul. Świętokrzyska 21,
00-049 Warsaw, e-mail:ksobcz@ippt.gov.pl

Kazimierz Sobczyk

INFORMATION DYNAMICS
Premises,Challenges and Results

2/2000

WARSZAWA 2000

ISSN 0208-5658

Praca wpłynęła do Redakcji dnia 19 maja 2000 r.

Na prawach rękopisu

Instytut Podstawowych Problemów Techniki PAN
Nakład 100 egz. Ark. 1 ,75 Ark. Druk.2 ,00
Oddano do drukarni w maju 2000 r

ATOS - Poligrafia-Reklama,W-wa,ul.Jana Kazimierza 35/37

Abstract

In various fields of contemporary research *information* and *dynamics* are becoming the key terms. Theoretic information reasoning is well known in physics, especially in thermodynamics where the relationship between the statistical (or, informational) entropy of the system and its thermodynamical entropy has been studied since a long time. Information theory is especially relevant to data processing and statistical inference. Generally speaking, the apparatus of information theory is applicable to any probabilistic system of observations since whenever we make statistical observations (or design and conduct statistical experiments) we seek information.

When the language of information theory (the concepts of entropy, mutual information between random variables and processes, information rate, maximum entropy formalism, information flow etc.) is used in connection with system dynamics we come to the notion of *information dynamics*.

The objective of this report is to show a potential of the basic information theoretic methodology for the analysis of various problems of system dynamics. In particular, we wish to indicate some challenges and expound our recent results on the maximum information entropy approach to the analysis of stochastic dynamical systems.

This report constitutes a little extended and modified version of the paper (under the same title) invited for publication in the journal "Mechanical Systems & Signal Processing", Academic Press (a special issue on: Information, Uncertainty and Decision for Mechanical System Analysis).

Contents

1. INTRODUCTION

Uncertainty of various real phenomena has been for a long time a problem of scientific endeavours. However, the notion of uncertainty has a broad meaning and its mathematical "verbalization" can be performed in various ways. In general, we can say that uncertainty in some situations (under consideration) occurs whenever *information* pertaining to this situation is deficient. It may be incomplete, imprecise, fragmentary, vague etc. If uncertainty is interpreted as randomness then a natural way of quantifying it is probability theory; the information pertaining to random phenomena is then defined in terms of probability. Such an approach to the concept of information will be here of our concern. Other approaches (e.g. algorithmic information, information measures based on possibility theory) have also been proposed (cf. [1]).

A probabilistic approach to description and analysis of a randomness has been the most popular and spectacularly successful. Especially in physics (statistical mechanics, quantum theory) the probabilistic reasoning has essentially contributed to understanding many phenomena and to advance of science. For example, the model of an erratic motion of particle suspended in fluid (known today as Brownian motion or, Wiener process) is an example of enormous power of probability theory. Also in engineering the successes of probabilistic treatment of uncertainty are tremendous (e.g. analysis of noises in radioelectronics, reliability theory, stochastic modelling and analysis of dynamic hazardous phenomena, etc.). Stochastic dynamics is now a greatly advanced field comprising the methods of investigation of various physical/engineering systems subjected to parametric and external random excitation (cf. [2]). It seems, however, that contemporary analyses in stochastic dynamics do not take advantage of the potential which is inherent in probabilistic information theory.

A very likely reason that the theoretic information approach has not been for a long time sufficiently recognized in general dynamics of physical systems might be the fact that information theory in the strictest sense (as originated by Shannon) is commonly joined with the communication systems, the coding theorems and information transmission (cf. [3]). A correct meaning of information theory is, however, much wider. As it has been underlined by Kullback [4], information theory is a branch of mathematical probability theory and mathematical statistics(*). As such, its concepts and methods are applicable to analysis of various physical and engineering systems. Theoretic - information reasoning is well known in physics, especially in thermodynamics where the relationship between the amount of information of the system and its thermodynamical entropy is well established.

Information theory is especially relevant to data processing and statistical inference. As a matter of fact, an information in a technically defined sense was first introduced in statistics by

---

(*) There are possible formulations of information theory without use of probability theory (cf. Ingarden R.S, Urbanik K. [5], also algorithmic information, cf. Kolmogorov [6], Chaitin G.J. [7])

R.A. Fisher (1925) in his work on theory of estimation. His concept of a measure of the amount of information supplied by data about unknown parameter is well known to statisticians. Generally speaking, the apparatus of information theory is applicable to any probabilistic system of observations since whenever we make statistical observations (or design and conduct statistical experiments) we seek information. Hence, the relevant question which arises is, for example: how much information can we infer from a particular set of observations about sampled phenomenon? A particular problem concerns estimation of an unobserved quantity X through observations on another quantity Y; these quantities can be random variables, stochastic processes or random fields. Another question is concerned with optimal design of experiment: how should an experiment be designed to obtain maximum information about a sampled random signal?

Stochastic modelling of various physical phenomena, including those in stochastic dynamics of engineering systems, stimulates a number of additional problems in the analysis of which the information-theoretic approach seems to be very natural. For example, the question of whether a given model satisfactorily represents a real phenomenon (and available empirical information) has always been an important and intriguing one. However, the validity of a model can only be assessed in relative terms. We can only say that a model is satisfactory if it meets the appropriate quality criteria. We have to compare the model predictions with the corresponding characteristics of empirical data. Also, the comparisons between various models are often of interest. In stochastic dynamics and related fields (e.g. in stochastic modelling of fatigue process generated by time-varying stress conditions) various random signals are used and the characterization of their divergence can constitute an important issue. In such situations there is a need for appropriate measures of divergence or diversity between a model and observations, between two models of random signals, etc. The measures which we have in mind (known also as distance measures) can be defined in different ways, but the information-theoretic measures are of great importance (cf. [8], [9]).

When the language of information theory (the concepts of entropy, mutual information between random variables and processes, information rate, maximum entropy analysis etc.) is used in such problems as those mentioned above we come to the notion of *information dynamics*. In fact, in the contemporary theory of dynamical systems information and dynamics become key terms. This is especially visible in the field of nonlinear dynamics which, after the early work of Poincaré, received its main impetus in the 70-ties when deterministic chaos and instabilities were recognized as extremely fruitful and intriguing phenomena. Chaotic dynamical systems have the interesting property that in spite of their deterministic character there appears a stochastic aspect due to extreme sensitivity to the initial conditions. This stochastic property of chaotic nonlinear systems can be quantitatively characterized by Lapunov exponents (which measure the exponential divergence in time of two neighbouring trajectories in a strange attractor), but it can also (and more fruitfully) be specified in the framework of information

theory. For example, the Kolmogorov-Sinai entropy as well as the information-theoretic Shannon entropy and mutual information have been used to characterize the degree of randomness in one-step predictions (of a chaotic motion) based on its whole past and to quantify the rate of information flow in chaotic dynamical systems. Ruelle and Takes (cf. [10]) have underlined the similarity in the behaviour of turbulent flow and strange attractors, suggesting that turbulence results from a strange attractor regime in the Navier-Stokes equations of hydrodynamics. This, in turn allows one to view the turbulence as being governed by information generated by the flow itself.

A need for the information dynamics methodology occurs also in the analysis of stochastic dynamical systems (where randomness is introduced explicitly, e.g. in the form of a random noise). For example, having a system modelled by differential equations with random excitation an interesting characteristics is the global randomness (entropy) of the response per unit time as well as the mutual information rate between two components of a given vectorial system. In spite of this type of problems the modelling and analysis of dynamic stochastic systems requires empirical measurements and deals with statistical problems of time series. But, in the majority of cases the available information on the process in question is deficient. To examine the effects of various type of empirical and numerical uncertainty of dynamical variables on the estimation of systems behaviour the appropriate information measures can be useful. The information dynamics methodology is relevant to a large variety of systems of various physical nature, including mechanical, environmental and structural engineering systems involving complex nonlinear and stochastic interactions. These complexities and uncertainties make that the conventional characteristics of the system in question (e.g. displacement, stresses) can not be quantified in a traditional way. The tools of dynamics and information processing are needed to perform an adequate estimation and prediction analysis.

The intention of this paper is to show a potential of the basic information theoretic methodology for the analysis of various problems of system dynamics. In particular, we wish to indicate some challenges and expound our recent results on the maximum entropy analysis of stochastic dynamical systems.

2. ENTROPY AND INFORMATION

An inherent feature of any random phenomenon is that a result of its observation can not be predicted a priori (i.e. before observation). Let (Γ, F, P) be a basic probability space and $X(\gamma)$, $\gamma \in \Gamma$ be a discrete random variable assuming values x_i with probabilities p_i, $i = 1,2,...,n$. The quantity $H(X)$ defined as

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \qquad (2.1)$$

is called the *Shannon entropy* of a discrete random variable $X(\gamma)$, or the entropy of the probability distribution $(p_1, p_2, \ldots, p_n)$. The negative sign in equation (2.1) makes the entropy of discrete random variable *non-negative*. The logarithm in (2.1) can be taken with arbitrary base greater than one. When a base two is used the unit of entropy is called a *bit* (binary digit). When the natural logarithm (i.e. base e) is taken the unit is called a *nat*. Unless otherwise specified one usually takes all logarithms to base 2, and hence all the entropies are measured in bits. For example, the entropy of a fair coin toss is 1 bit (since $p_1 = p_2 = 0.5$). Of course, entropy can always be changed from one base to another; since $\log_b p = \log_b a \log_a p$ we have $H_b(X) = const. H_a(X)$. We use the convention that $0 \log 0 = 0$, which is easily justified by continuity since $x \log x \to 0$ as $x \to 0$. It should be noticed that the entropy of a discrete random variable depends only on the number of values and their probabilities and does not depend on the actual values taken by the random variable $X(\gamma)$.

A value of entropy of a discrete random variable can be interpreted as an average number of "bits" necessary to differentiate its possible values. When all probabilities except, say, $p_n = 1$, vanish, we are certain that the event x_n will occur. In this case there is no randomness or, one can say there is no information missing and $H = 0$. On the other hand, when all possible events (or, outcomes) are *equiprobable* uncertainty increases with the number n of the events. In such a case $p_i = p$, $i = 1,2,\ldots,n$, $(p = 1/n)$ and

$$H(X) = \log n \tag{2.2}$$

In general, an a priori uncertainty of equiprobable n events can be defined as: $H = f(n)$, where $f(.)$ is some non-negative, increasing function. If we introduce a natural requirement, that f should be additive, we come to the conclusion that $f = \log$. A logarithmic measure of uncertainty was, for the first time, introduced by Hartley (1928).

If $X(\gamma)$ takes infinitely many values $x_1, x_2, \ldots$ with probabilities $p_1, p_2, \ldots$ then the entropy of $X(\gamma)$ is defined by

$$H(X) = -\sum_{i=1}^{\infty} p_i \log p_i \tag{2.3}$$

However, in this case $H(X)$ is not necessarily finite. An important property of entropy of discrete random variable $X(\gamma)$ is that it is invariant under one-to-one transformations: $Y = \varphi(X)$.

Let us consider now continuous random variable $X(\gamma)$ with the probability density $f(x)$. The *entropy* of a continuous random variable (or, the differential entropy of $X(\gamma)$) is defined as

$$H(X) = -\int_S f(x) \log f(x) \mathrm{d}x \tag{2.4}$$

where S is the support set of X, i.e. a set of x-values for which $f(x) \geq 0$. Since the entropy (2.4) depends only on the probability density, it is also denoted as $H(f)$.

It can easily be calculated that entropy of uniform distribution on the interval $[a,b]$ with $b-a=L$, is

$$H_U(X) = \log(b-a) = \log L \tag{2.5}$$

whereas the entropy of Gaussian distribution $N(0,\sigma_X^2)$ is as follows:

$$\begin{aligned} H_G(X) &= \frac{1}{2}\ln\left(2\pi e\sigma_X^2\right) \text{ nats} \\ &= \frac{1}{2}\log_2\left(2\pi e\sigma_X^2\right) \text{ bits} \end{aligned} \tag{2.6}$$

The differential entropy is similar in many ways to the entropy of a discrete random variables, but it has also important differences. Above all, the differential entropy can assume both positive and negative values; e.g. the entropy (2.5) of an uniform distribution is negative for $L<1$. Another important feature of the differential entropy defined by (2.4) is that it *is not a limiting value* of the entropy of a discrete random variable obtained by a discretization of the continuous range of X - values into small sub-intervals of length Δ, and then letting the letter tend to zero. Indeed, in order to make use of the definition (2.1), we replace a continuous probability density $f(x)$ by a discrete distribution denoting by p_i the probabilities $f(x_i)\Delta x_i$. The entropy of such a discretized version of X is

$$\begin{aligned} H(X_{disc}) &= -\sum_i p_i \log p_i = -\sum_i f(x_i)\Delta x_i \log[f(x_i)\Delta x_i] \\ &= -\sum_i f(x_i)\log[f(x_i)]\Delta x_i - \sum_i p_i \log \Delta x_i \end{aligned} \tag{2.7}$$

As the graining becomes finer and finer (i.e. $\Delta x_i \to 0$ for all i), assuming that $f(x)$ is continuous function, we obtain (additionally, assume that all Δx_i are equal, i.e. $\Delta x_i = \Delta$)

$$H(X_{discr.}) \to H(X_{cont.}) - \log \Delta \tag{2.8}$$

However, the last term (i.e. $-\log\Delta$) tends to infinity as $\Delta \to 0$. This means that any continuous probability distribution contains much more randomness than its discretized version, as the difference between $H(X_{discr.})$ and $H(X_{cont.})$ is infinite. The continuous entropy $H(X)$ itself defined by (2.4) can not work as a measure of "global" randomness of the distribution, however the difference, say, $H(X)-H(Y)$ of the entropies characterizes the difference in randomness of X and Y. The above simple reasoning illustrates a deep difference

between entropies of discrete and continuous random variables. In the discrete case, the entropy quantifies randomness in an *absolute* way, whereas in the continuous case the characterization of randomness has only a *relative* meaning. However, very often we just want to know which random variable has a greater or the greatest entropy. In such situations the continuous entropy (2.4) plays an important role.

Definition of the entropy of n-dimensional random variable $\mathbf{X} = [X_1, X_2, \ldots, X_n]$ with the joint probability density $f(x_1, x_2, \ldots, x_n)$ is the extension of (2.4), i.e.

$$H(\mathbf{X}) = -\int_S f(\mathbf{x}) \log f(\mathbf{x}) \, \mathrm{d}\mathbf{x} \tag{2.9}$$

where the integral is a n-fold integral extended over n-dimesional region S supporting $f(\mathbf{x})$.

An interesting property of the differential entropy, generally - in vectorial case - is its transformation rule. Let $\mathbf{X}$ and $\mathbf{Y}$ be n-dimensional random variables such that $\mathbf{Y} = \Phi(\mathbf{X})$ where $\Phi = [\varphi_1, \varphi_2, \ldots, \varphi_n]$ is one-to-one transformation defined on R_n. As is known, the probability density of $\mathbf{Y}$ is

$$f_{\mathrm{Y}}(\mathbf{y}) = f_{\mathrm{X}}[\Psi(\mathbf{y})] \, |\mathrm{J}(\mathbf{y})| \tag{2.10}$$

where $\Psi(\mathbf{y}) = [\psi_1, \psi_2, \ldots, \psi_n]$ is the inverse of Φ, i.e. $\mathbf{x} = \Psi(\mathbf{y})$, and the J is the Jacobian of the inverse transformation, i.e.

$$\mathrm{J}(\mathbf{y}) = \det\left(\frac{\partial \psi_i}{\partial y_j}(\mathbf{y}) \right)_{i,j=1,\ldots,n} \tag{2.11}$$

Therefore,

$$\begin{aligned} H(\mathbf{Y}) &= -\int f_{\mathrm{Y}}(\mathbf{y}) \log f_{\mathrm{Y}}(\mathbf{y}) \, \mathrm{d}\mathbf{y} \\ &= -\int f_{\mathrm{X}}[\Psi(\mathbf{y})] |\mathrm{J}(\mathbf{y})| \log\{ f_{\mathrm{X}}[\Psi(\mathbf{y})] |\mathrm{J}(\mathbf{y})| \} \, \mathrm{d}\mathbf{y} \\ &= -\int f_{\mathrm{X}}(\mathbf{x}) \log\{ f_{\mathrm{X}}(\mathbf{x}) |\mathrm{J}[\Phi(\mathbf{x})]| \} \, \mathrm{d}\mathbf{x} \end{aligned} \tag{2.12}$$

Finally, we have

$$H(\mathbf{Y}) = H(\mathbf{X}) - \int f_{\mathrm{X}}(\mathbf{x}) \log\{ |\mathrm{J}[\Phi(\mathbf{x})]| \} \, \mathrm{d}\mathbf{x} \tag{2.13}$$

The above formula indicates that the entropy of a continuous random variable is *not* invariant under transformation of variables (contrary to the case of discrete random variables). Property (2.13) has been of some concern in the literature since it says that a deterministic

transformation of $\mathbf{X}$ modifies the initial amount of uncertainty involved in $\mathbf{X}$. The second term in (2.13) can be interpreted as the amount of uncertainty involved in the deterministic mapping (cf. [11]). This, however, seems to be an open problem.

If Φ is a linear transformation, i.e. $\mathbf{Y} = \mathbf{AX}$, then $J(\mathbf{y}) = \det \mathbf{A}^{-1} = 1/\det \mathbf{A}$ and we have

$$H(\mathbf{Y}) = H(\mathbf{X}) + \log|\det \mathbf{A}| \tag{2.14}$$

In the analysis of random phenomena with dependent random variables the notion of conditional entropy is important. The *conditional entropy* of a random variable $X(\gamma)$ given $Y = y$ is defined (for continuous random variables) as

$$H(X|y) = -\int f(x|y) \log f(x|y)\, \mathrm{d}x \tag{2.15}$$

The preceding quantity depends on the values of the random variable $Y(\gamma)$, so it is itself a random variable. The *average* conditional entropy of $X(\gamma)$ with respect to $Y(\gamma)$ is defined as

$$\begin{aligned} \overline{H}_Y(X) &= \langle H(X|Y) \rangle = \int H(X|y) f(y)\, \mathrm{d}y \\ &= -\iint f(x,y) \log f(x|y)\, \mathrm{d}x\, \mathrm{d}y \end{aligned} \tag{2.16}$$

An analogous definition of conditional entropy holds for discrete random variables. For both continuous and discrete random variables

$$\begin{aligned} H(X,Y) &= H(X) + \overline{H}_X(Y) = H(Y) + \overline{H}_Y(X) \\ &\leq H(X) + H(Y) \end{aligned} \tag{2.17}$$

Let us define now the concept of *information*. It is clear, that an initial randomness of X, characterized by entropy $H(X)$, decreases when observation of measurement of X is made. In other words, more information is provided on X, the less uncertain its values will be. Hence, it seems to be natural to measure the amount of information about X by the difference in its entropy before and after experiment or measurement. Suppose we are informed that $Y = y_i$ occurred and random variable Y is correlated with X. Then the uncertainty of X is reduced to the conditional entropy $H(X|y_i)$. One can say that the amount of information about X contained in the information "$Y = y_i$" is

$$H(X) - H(X|Y = y_i) \tag{2.18}$$

The average value of (2.18)

$$\sum_i [H(X) - H(X|Y = y_i)]\, P\{Y = y_i\} = H(X) - \overline{H}_Y(X)$$

is the amount of information about X one may expect to obtain by observing Y. So, we come to the following definition. The Shannon *mutual information* (or *transformation*) between random variables X and Y is defined as (cf. [12])

$$I(X,Y) = H(X) - \overline{H}_Y(X) \tag{2.19}$$

From (2.17) one then obtains

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{2.20}$$

For continuous random variables

$$I(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{\infty} f(x,y) \log \frac{f(x,y)}{f_1(x) f_2(y)}\, dx\, dy \tag{2.21}$$

The basic properties of the mutual information are:

a) $I(X,Y) \geq 0$, with equality if only if X and Y are independent

b) $I(X,Y) = I(Y,X)$

c) If $Z = \varphi(Y)$, φ is a continuous mapping, then $I(X,Y) \geq I(X,Z)$ with equality when φ is one-to-one mapping.

The above properties are the same in the discrete and continuous cases. In particular, the mutual information between two continuous random variables is the limit of the mutual information between their discretized versions (cf. (2.8)). Indeed,

$$\begin{aligned} I(X_{discr.}, Y_{discr.}) &= H(X_{discr.}) - H(X_{discr.}|Y_{discr.}) \\ &\approx H(X) - \log\Delta - [\overline{H}_Y(X) - \log\Delta] \\ &= I(X,Y) \end{aligned} \tag{2.22}$$

Remark: It is worth noting that in the process of recognition of the values of a discrete random variable X (via observation or measurements of X itself) the information gained about X is equal to the amount of its entropy decrease. Therefore, a complete information about X, say $I(X)$ or $I(X,X)$, is achived when the entropy of X is reduced to zero. This means that, quantitatively, $I(X) = H(X)$, which agrees with the Boltzmann statement (in the context of statistical mechanics) that the "entropy is a measure of missing information".

3. STATISTICAL INFERENCE; MAXIMUM ENTROPY AND MINIMUM RELATIVE ENTROPY

3.1. Maximum entropy principle

One the main tasks in the analysis of empirical data and the associated statistical inference is to estimate the probability density function when only a partial information about the true distribution is given. Various methods have been elaborated in mathematical statistics to estimate unknown probability distribution (of the population) using the values in a random sample. Often, the available information about unknown probability distribution is given (or, can be expressed) in the form of statistical moments. This means that, in the case of continuous random variable X whose (unknown) probability density is $f(x)$ we have the partial information in the form

$$\int_a^b f(x)\mathrm{d}x = 1 \tag{3.1}$$

$$\langle g_k(X)\rangle = \int_a^b g_k(x) f(x)\mathrm{d}x = m_k \quad , \quad k = 1,2,\ldots,n \tag{3.2}$$

where the functions $g_k(x)$ and numbers m_k, $k = 1,2,\ldots,n$ are given and the integration interval can be finite or infinite, depending on the range of possible values of X.
An important question which arises is: how can the partial information about $f(x)$ contained in (3.1) and (3.2) be used for the best characterization of $f(x)$?

Among the possible approaches to this problem, the *principle of maximum entropy* (introduced first in statistical physics, cf. Jaynes [13], Ingarden [14]) seems to be especially attractive and effective. In general, the principle of maximum entropy states that, of all the probability densities that satisfy the constraints (3.1), (3.2) one should choose the one with the largest Shannon information entropy, i.e. the density $f^*(x)$ which solves the following maximization problem:

$$\max_f H(f) = \max_f \left\{ -\int_a^b f(x)\log f(x)\mathrm{d}x \right\}$$

$$\langle g_k(X)\rangle = \int_a^b g_k(x) f(x)\mathrm{d}x = m_k \quad , \quad k = 1,2,\ldots,n \tag{3.3}$$

$$\int_a^b f(x)\mathrm{d}x = 1$$

Since the entropy characterizes the uncertainty associated with possible realizations of a random variable, maximizing entropy leads to the distribution with the largest informational content. In addition, of all distributions that satisfy given constraints (3.1), (3.2) the maximum entropy distribution $f^*(x)$ is the most unbiased one (we want to be maximally uncertain about what we do not know) and it can be regarded as the most rational, or "most honest" probability distribution, since it does not include any information which is not at our disposal. For this reason, in statistics the maximum entropy distribution has been proposed to serve as a prior distribution in Bayesian inference (cf. [15]). However, the question whether that maximum entropy distribution should be regarded as the prior distribution or the posterior one can not be answered generally and uniquely; the answer may depend on the specific situations. The principle of maximum entropy can be thought of as an extension of the famous Laplace "principle of insufficient reason" which postulates a uniform distribution as being the most "uncertain" in situations in which nothing is known about the variable in question.

In order to solve the maximization problem (3.3) we form the extended functional (the Lagrangian)

$$L = -\int f(x)\ln f(x)\mathrm{d}x - \lambda_0[\int f(x)\mathrm{d}x - 1] - \sum_{k=1}^{n}\lambda_k[\int g_k(x)f(x)\mathrm{d}x - m_k] \tag{3.4}$$

which has the form

$$L = \int F[x, f(x)]\mathrm{d}x \tag{3.5}$$

The function $f(x)$ which maximizes L has to satisfy the equation $\partial F / \partial f(x) = 0$; therefore we get

$$-\ln f(x) - 1 - \lambda_0 - \sum_k \lambda_k g_k(x) = 0 \tag{3.6}$$

Since $H(f)$ is a concave functional we obtain the maximum entropy density $f^*(x)$

$$f^*(x) = C\exp\left\{-\sum_{k=1}^{n}\lambda_k g_k(x)\right\} \tag{3.7}$$

where $C^{-1} = \exp(\lambda_0 + 1)$ and $\lambda_0, \lambda_1, \ldots, \lambda_n$ are unknown Lagrange multipliers and

$$C^{-1} = \int \exp\left\{-\sum_{k=1}^{n}\lambda_k g_k(x)\right\}\mathrm{d}x \tag{3.8}$$

Since $f^*(x)$ has to satisfy the moment constraints (3.2) the Lagrange multipliers: $\lambda_0, \lambda_1, \ldots, \lambda_n$ are determined from the system of equations

$$\int_a^b g_k(x)\exp\left\{-\sum_{k=1}^{n}\lambda_k g_k(x)\right\}\mathrm{d}x = m_k \quad , \quad k = 1,2,\ldots,n \tag{3.9}$$

It can easily be shown (cf. [12]) that $f^*(x)$ uniquely maximizes $H(f)$ over all probability densities $f(x)$ satisfying constraints (3.1), (3.2).

The same maximum entropy method as shown above holds for discrete probabilities $p_1, p_2, \ldots, p_n$ and for multidimensional distributions $f(\mathbf{x})$, $\mathbf{x} \in R_n$.

In the most popular case when m_k in constraints (3.2) are simple moments of random variable X, i.e. $g_k(X) = X^k$, the maximum entropy density is

$$f^*(x) = C\exp\left\{-\sum_{k=1}^{n}\lambda_k x^k\right\} \tag{3.10}$$

The following statements are straightforward conclusions from the maximum entropy principle.

1) When the range of possible values of X is finite, say $[a,b]$, and there are no constraints (except for the natural normalization condition) then the maximum entropy distribution is the uniform distribution over $[a,b]$.
2) When the range of X is $[0,+\infty)$ then:
   a) if no moment constraint is prescribed there is no maximum entropy distribution;
   b) if the mean m is given, then the maximum entropy distribution is the exponential one, i.e.

$$f^*(x) = \mu\exp(-\mu x) \quad , \quad \mu = 1/m \tag{3.11}$$

   and $H_{\max} = 1 + \ln m$.
3) When the range of X is $(-\infty,+\infty)$ then:
   a) if no moment is prescribed, there is no maximum entropy distribution;
   b) if only the mean $\langle X\rangle = m$ is given, there is no maximum entropy distribution;
   c) if two first moments $\langle X\rangle = m_1$, $\langle X^2\rangle = m_2$ are given then the maximum entropy distribution is the normal distribution $N(m_1, \sigma_X^2)$, $\sigma_X^2 = m_2 - m_1^2$.

Although the maximum entropy method has its roots in physics, it has been successfully used in variety of other fields including statistical inference, reliability estimation, pattern recognition and signal processing. For instance, it has been shown that the exponential distribution of air density as a function of height in the earth's atmosphere is the maximum entropy distribution when the mean potential energy is given as a priori information. In mathematical statistics all of the best known distributions turn out to be maximum entropy distributions given simple moment constraints. Even the Cauchy distribution of random variable X (which does not possesses the moments) is a maximum entropy distribution over all

distributions satisfying $\langle g(X)\rangle = \langle \ln(1+X^2)\rangle = m$, where m is given. In the last section of this paper we will show the extension of this idea to the problems of stochastic dynamics.

3.2. Relative entropy

To be able to compare the predictions of two statistical models, as well as to conduct comparison of the specific model with empirical data, one asks how different are, or what is the distance between, corresponding probability distributions?

A distance or divergence between two probability distributions can be measured in different way. Of course, all acceptable distance measures have to possess some basic, natural properties. Let us assume that there are two probability distributions P and Q, which possess densities $p(x)$ and $q(x)$, respectively. It has been recognized that a quite general class of distance measures can be defined as

$$d(p,q) = \left\langle h\left(\frac{p(X)}{q(X)}\right)\right\rangle_P = \int h\left(\frac{p(x)}{q(x)}\right)p(x)\mathrm{d}x \tag{3.12}$$

where $h(\ )$ is some continuous appropriately defined function on $R_+ = [0,\infty)$ and the quantity $\varphi = p(x)/q(x)$ is often called the likelihood ratio.

The best known in information theory and statistics is the *relative entropy* or *Kullback-Leibler divergence* measure ($h(x) = \log x$)

$$d(p,q) = J(p,q) = \int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}x \tag{3.13}$$

In the above definition, we use the convention (based on continuity arguments) that $0\log 0 = 0$ and $p\log(p/0) = \infty$.

The relative entropy $J(p,q)$ is always *non-negative* and is zero if and only if $p = q$. However, it is not true distance since it is not symmetric and does not satisfy the triangle inequality. However, a lack of these two properties in statistical applications is not a serious deficiency. Usually, we are interested in the *directed* divergence from a given (a priori) distribution $q(x)$. Also, we are interested in measuring the difference between two distributions only. The distribution $q(x)$ can be regarded as the *reference distribution*.

It is worth adding that in physics and in theory of dynamical systems (cf. [16], [17]) a "negative" form of the Kullback-Leibler measure is used and called the *conditional entropy* of the density p with respect to the density q. It is denoted $H_c(p|q)$ and defined ($\sup p \subset \sup q$) as

$$H_c(p|q) = -J(p,q) \tag{3.14}$$

If q is the constant density on the state space χ, i.e. $q = 1/\mu_L(\chi)$, then $H_c(p|q) = H(p) - \log \mu_L(\chi)$, where $\mu_L(\chi)$ is the Lebesgue measure of χ. If the state space is normalized, then $q = 1$ and $H_c(p|q) = H_c(p|1) = H(p)$. In this sense the conditional entropy $H_c(p|q)$ is a generalization of the Shannon entropy $H(p)$.

It is interesting to notice the following relationship between K-L divergence J and the Shannon mutual information (2.21). To see this relation, let us assume that we wish to identify an unobsevable variable $X(\gamma)$ on the basis of the observation of the another variable $Y(\gamma)$ that is statistically related to $X(\gamma)$. The two-dimensional analog of (3.13) is

$$\begin{aligned} J(p,q) &= J(f_{XY}(x,y), f_X(x) f_Y(y)) = I(X,Y) \\ &= \int\int f(x,y) \log \frac{f(x,y)}{f_X(x) f_Y(y)} \mathrm{d}x \, \mathrm{d}y \end{aligned} \tag{3.15}$$

This means that the Shannon amount of information about X provided by observation of Y can be regarded to be equal to the Kullback-Leibler divergence between $f_{XY}(x,y)$ and $f_X(x) f_Y(y)$.

The Kullback-Leibler (K-L) relative entropy (3.13) plays an important role in statistical inference (cf. [4], [18]). This is mainly due to the *principle of minimum of relative entropy* (or, minimum discrimination information). Let us consider a general case of n-dimensional random variable $\mathbf{X}(\gamma)$ and assume that its probability density $f(\mathbf{x})$ exists but is unknown. Suppose that our a priori partial information about $f(\mathbf{x})$ is given in the form of the moment constraints (as in maximum entropy method)

$$\langle g_k(\mathbf{X}) \rangle = \int g_k(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x} = m_k \quad , \quad k = 0,1,2,\ldots,m \tag{3.16}$$

where the integration is over the range of random variable $\mathbf{X}$ in R_n, functions $g_k(\mathbf{x})$, and m_k, $k = 0,1,\ldots,m$ are given; we assume here (to include the normalization condition) that $g_0(\mathbf{x}) = 1$ and $m_0 = 1$. Let us also assume that in addition to (3.16) a priori possible estimation of $f(\mathbf{x})$ is $q(\mathbf{x})$. The statistical problem is to select such an estimate $\bar{f}(\mathbf{x})$ of unknown $f(\mathbf{x})$ which satisfies constraints (3.16) and is "as close as possible" to the given density $q(\mathbf{x})$. The solution is to minimize the relative entropy $J(f,q)$, i.e.,

$$\min_f J(f,q) = \min_f \left\{ \int f(\mathbf{x}) \log \frac{f(\mathbf{x})}{q(\mathbf{x})} \mathrm{d}\mathbf{x} \right\} \tag{3.17}$$

subject to the constraints (3.16). This is just the essence of the minimum relative entropy method.

Making use of the Lagrange multipliers method (as in the case of maximum entropy method) one obtains that the minimum relative entropy density has the form

$$\tilde{f}(\mathbf{x}) = q(\mathbf{x})\exp\left\{-\sum_{k=0}^{m}\lambda_k g_k(\mathbf{x})\right\} \tag{3.18}$$

If there is no a priori choice of the distribution $q(\mathbf{x})$ which could be used in (3.17), it is natural to express our ignorance of all $\mathbf{x}$ values to be equaly likely. In this case, the minimum relative entropy method is equivalent to maximization of the entropy. Indeed, let $f(\mathbf{x})$ vanishes outside a domain $D \subset R_n$ with finite volume $|D|$ and q_0 is the density of uniform distribution on D, i.e.,

$$q_0(\mathbf{x}) = \begin{cases} \dfrac{1}{|D|} & , \quad \mathbf{x} \in D \\ 0 & , \quad \mathbf{x} \notin D \end{cases} \tag{3.19}$$

The relative K-L entropy $J(f,q_0)$ is

$$J(f,q_0) = -H(f) + \log|D| \tag{3.20}$$

Therefore, maximization of entropy $H(f)$, assumed here to be finite, is equivalent to minimizing $J(f,q_0)$. Thus, the maximum entropy method can be regarded as a special case of the relative entropy minimization.

The principle of minimum relative-entropy provides a general method of inference about an unknown probability density $f(\mathbf{x})$ when there exists a *prior* estimate $q(\mathbf{x})$ of $f(\mathbf{x})$ and some information (about $f(\mathbf{x})$) in the form of moments constraints. The principle states that, of all the densities that satisfy the constraints, one should choose the *posterior* $f(\mathbf{x})$ with the minimum relative entropy (or, minimum divergence) $J(p,q)$, where $q(\mathbf{x})$ is a *prior* estimate of $f(\mathbf{x})$. This method introduced first by Kullback [4] has been studied extensively by Shore and Johnson [19], [20], who derived it from axioms of consistent inductive inference. In the paper [20] the authors investigate also the conditions for the existence and uniqueness of the solution of the relative entropy minimization problem.

It should be underlined that although the methods of minimum of relative entropy and maximum entropy were originally derived within the context of probability theory they are also applicable to reconstruction of non-probabilistic functions. Such a situation occurs, for example, in pattern recognition and image reconstruction (cf. [21]), where the moment descriptions of various forms have been extensively employed as pattern features.

4. ENTROPY AND INFORMATION IN DYNAMICAL SYSTEMS

4.1 Entropy rate per unit time

Let $X(t)$, $t \in T \subset R_1$ be a stochastic process and let $[X_{t_1}, X_{t_2}, \ldots, X_{t_N}]$ be a "sample" of $X(t)$ for $t = t_1, \ldots, t_N$. The entropy (if it exists)

$$H[X_{t_1}, X_{t_2}, \ldots, X_{t_N}] = \left\langle -\log f(x_{t_1}, \ldots, x_{t_N}) \right\rangle \tag{4.1}$$

in the *N-th order entropy* of the process $X(t)$. However, even if each X_{t_i}, $i = 1,2,\ldots,N$ is a discrete random variable the entropy (4.1) diverges to infinity as $N \to \infty$. A quantity which plays an important role is, therefore, not the limit of $H[X_{t_1}, X_{t_2}, \ldots, X_{t_N}]$, but the rate of the growth of the entropy.

The *entropy rate* (or, the entropy per unit time) of the process $X(t)$ is defined as

$$\overline{H}[X_t] = \lim_{N \to \infty} \frac{1}{N} H[X_{t_1}, X_{t_2}, \ldots, X_{t_N}] \tag{4.2}$$

when the limit exists.

It is clear, that if $X_{t_1}, X_{t_2}, \ldots, X_{t_N}$ are independent and identically distributed random variables then

$$\overline{H}(X_t) = \lim_{N \to \infty} \frac{1}{N} N H(X_{t_1}) = H(X_{t_1}) \tag{4.3}$$

The limit (4.3) exists for any stationary process. If $[X_{t_1}, X_{t_2}, \ldots, X_{t_N}]$ has finite entropy then (for stationary process)

$$\overline{H}(X_t) = H(X_{t_1} | X_0^-) \tag{4.4}$$

where $X_0^- = [X_0, X_{-1}, X_{-2}, \ldots]$ and (denoting $X_{t_1} = X_1$)

$$H(X_1 | X_0^-) = \lim_{N \to \infty} H(X_0, X_{-1}, \ldots, X_{-N}) \tag{4.5}$$

is the *conditional* entropy of X_1 given its past X_0^-. This means that the entropy rate $\overline{H}(X_t)$ is equal to conditional entropy of "one step future" $X_{t_1} = X_1$ when the "past" X_0^- is known.

If we assume that process $X(t)$ is stationary and Gaussian, with spectral density $g_X(\omega)$ then it can be shown that [12]

$$\overline{H}(X_t) = \frac{1}{4\pi}\int_{-\infty}^{+\infty} \ln g_X(\omega)\mathrm{d}\,\omega + \frac{1}{2}\ln 2\pi e \tag{4.6}$$

If the process in question is a stationary time series $\{X_n\}$ then the integration in (4.6) is taken over the interval $[-\pi, \pi]$. If, for example, process $X(t)$ has the spectral density

$$g_X(\omega) = g_0 \frac{\omega^2 + \alpha^2}{\omega^2 + \beta^2}$$

then

$$\overline{H}(X_t) = \frac{g_0}{2\pi}\int_0^{\infty} \ln \frac{\omega^2 + \alpha^2}{\omega^2 + \beta^2}\mathrm{d}\,\omega + \frac{1}{2}\ln 2\pi e = \frac{g_0}{2}(\alpha - \beta) + \ln\sqrt{2\pi e}$$

An important class of discrete-time models of stationary and Gaussian processes is represented in the form of an ARMA process, i.e. *autoregressive-moving-average process* of order (l,m)

$$X_n = \sum_{j=1}^{l} a_j X_{n-j} + \sum_{k=0}^{m} b_k \xi_{n-k} \tag{4.7}$$

where $\{a_j\},\{b_k\}$ are constant coefficients and $\xi_n, \xi_{n-1}, \ldots, \xi_{n-m}$ is a sequence of independent Gaussian random variables of unit variance. As it is seen from (4.7), ARMA model of real random process corresponds to passing a white noise through a discrete filter with the appropriate selected parameters.

It is known that the spectral density of the ARMA process (4.7) is [22]

$$g_X(\omega) = \frac{1}{2\pi}\left|\frac{B(e^{i\omega})}{A(e^{i\omega})}\right|^2 \tag{4.8}$$

where $A(z)$ and $B(z)$ are the polynomials

$$\begin{aligned} A(z) &= \sum_{j=0}^{l} a_{l-j} z^j \;,\quad a_0 = -1 \\ B(z) &= \sum_{k=0}^{m} b_{m-k} z^k \end{aligned} \tag{4.9}$$

The entropy rate (per unit time) of an ARMA (l,m) process as defined above can be calculated as follows.

Denote by $\alpha_1,\ldots,\alpha_l$ the roots of $A(z)$ and by $\beta_1,\ldots,\beta_l$ roots of $B(z)$. It can be assumed that $A(z)$ and $B(z)$ have the roots inside the unit disc, and that these roots do not coincide (if $A(z)$ and $B(z)$ have common roots, this would mean that the process is ARMA process of lower order). Therefore, $|\alpha_j| < 1$ and $|\beta_j| < 1$. Now, $g_X(\omega)$ can be represented as

$$g_X(\omega) = \frac{b_0^2}{2\pi}\frac{\prod_{k=1}^{m}|e^{i\omega}-\beta_k|^2}{\prod_{j=1}^{l}|e^{i\omega}-\alpha_k|^2} \tag{4.10}$$

and

$$\int_{-\pi}^{\pi}\ln g_X(\omega)\mathrm{d}\,\omega = \int_{-\pi}^{\pi}\ln\frac{b_0^2}{2\pi}\mathrm{d}\,\omega + \sum_{k=1}^{m}\int_{-\pi}^{\pi}\ln|e^{i\omega}-\beta_k|\mathrm{d}\,\omega - \sum_{j=1}^{l}\int_{-\pi}^{\pi}\ln|e^{i\omega}-\alpha_j|\mathrm{d}\,\omega$$

By virtue of the Poisson's formula we have if $\alpha \le 1$

$$\int_{-\pi}^{\pi}\ln|e^{i\omega}-\alpha|\mathrm{d}\,\omega = 0$$

Therefore, finally the entropy rate of the ARMA time series (4.7) is

$$\overline{H}(X_n) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\ln\frac{b_0^2}{2\pi}\mathrm{d}\,\omega + \frac{1}{2}\ln 2\pi e = \frac{\ln b_0^2 e}{2} \tag{4.11}$$

Let $X(t)$ and $Y(t)$ be two stationary and Gaussian random processes (of a continuous argument t) with spectral densities $g_X(\omega)$ and $g_Y(\omega)$, respectively. Assume that $X(t)$ and $Y(t)$ are also jointly stationary with the joint spectral density $g_{XY}(\omega)$.

The *rate of the mutual information* between $X(t)$ and $Y(t)$ is given by the formula (cf. [23])

$$I_r(X_t,Y_t) = -\frac{1}{4\pi}\int_{-\infty}^{+\infty}\ln\left[1-\frac{|g_{XY}(\omega)|^2}{g_X(\omega)g_Y(\omega)}\right]\mathrm{d}\,\omega \tag{4.12}$$

In the case, when

$$Y(t) = X(t) + \xi(t) \tag{4.13}$$

when $\xi(t)$ is a stationary process, uncorellated with $X(t)$, formula (4.12) takes the form

$$I_r(X_t, Y_t) = \frac{1}{2\pi} \int_0^{+\infty} \ln\left[1 + \frac{g_X(\omega)}{g_\xi(\omega)}\right] d\omega \tag{4.14}$$

A case of non-Gaussian processes for which the entropy rates can be efficiently evaluated are the Markov chains. A Markov chain is a discrete-parameter stochastic process $X_1, X_2, \ldots$ if, for $n = 1,2,\ldots$

$$P\{X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1\} = P\{X_{n+1} = x_{n+1} | X_n = x_n\}$$

for all possible sets of real numbers $x_1, x_2, \ldots, x_n, x_{n+1}$ belonging to the state space χ. A time invariant Markov chain is characterized by its initial state and a *probability transition* matrix $P = \{P_{ij}\}$, $i,j = 1,2,\ldots,m$, where $P_{ij} = P\{X_{n+1} = j | X_n = i\}$. Let a Markov chain $X_1, X_2, \ldots$ be a stationary one (i.e. P_{ij} do not depend on n). For a stationary Markov chain the entropy rate is

$$\overline{H}(\chi) = \overline{H}(X_1, X_2, \ldots) = H(X_2 | X_1) \tag{4.15}$$

where the conditional entropy is calculated using the given stationary distribution μ. More explicitly,

$$\overline{H}(\chi) = -\sum_{ij} \mu_i P_{ij} \log P_{ij} \tag{4.16}$$

Probabilities P_{ij} characterize a transition of the process from state i to state j in one step. It is of interest to characterize a *trajectory* T_{ij} of a Markov chain from state i to state j. It is a path with initial state i, final state j, and no intervening state equal to j. The probability $P(T_{ij})$ of a trajectory $T_{ij} = ix_2x_3..x_kj$ is given by

$$P(T_{ij}) = P_{ix_2} P_{x_2x_3} \ldots P_{x_kj} \tag{4.17}$$

Let us denote by Θ_{ij} the set of all trajectories from i to j. The entropy H_{ij} of the trajectory T_{ij} from i to j is defined by (cf.[24])

$$H_{ij} = H(T_{ij}) = -\sum_{T_{ij} \in \Theta_{ij}} P(t_{ij}) \log(T_{ij}) \tag{4.18}$$

It has been proven (cf. [24]) that for an irreducible finite state Markov chain with entropy rate (4.16) the entropy H_{ii} of the random trajectory from state i to state i is given by

$$H_{ii} = \frac{\overline{H}(\chi)}{\mu_i} \tag{4.19}$$

This means that the entropy of the random trajectory T_{ii} is the product of the expected number of steps $1 / \mu_i$ to return to state i and the entropy rate $\overline{H}(\chi)$ per step. Paper [24] contains also a general closed form solution for the entropies $H_{ij} = H(T_{ij})$.

4.2. Information flow in dynamical systems

An important class of problems is concerned with the entropy change in dynamical systems, e.g. in the systems described by differential or, differential-integral equations. In fact, such a study was started already by Boltzmann in connection with his kinetic theory of gases. The Boltzmann H-theorem provides an important properties of the kinetic Boltzmann equation (cf. [25]). The essence of this profound result is as follows.

Let $f(\mathbf{r},\mathbf{v},t)$, a function of position, velocity and time, be a density of the probability of finding a particle within the 6-dimensional spatial infinitesimal element $\mathrm{d}\,\mathbf{r}\,\mathrm{d}\,\mathbf{v}$ at time t. This function is governed by the following Boltzmann equation

$$\frac{\partial f}{\partial t} + \mathbf{v}\frac{\partial f}{\partial \mathbf{r}} + \frac{\mathbf{F}}{m}\frac{\partial f}{\partial \mathbf{v}} = Q(f,f) \tag{4.20}$$

where m is the mass of the molecule, and $\mathbf{F}$ characterizes the macroscopic forces acting on the molecules; $Q(f,f)$ is a symbolic denotation of the so called *collision term* representing the rate of change of f due to collisions of particles. Let H_t be a function of t defined as (cf. [25])

$$H_t = \int\int f(\mathbf{r},\mathbf{v},t)\log f(\mathbf{r},\mathbf{v},t)\mathrm{d}\,\mathbf{v}\,\mathrm{d}\,\mathbf{r} \tag{4.21}$$

The Boltzmann H-theorem states that in the isolated system (there is no energy exchange through the boundary of the region filled by gas) H_t defined by (4.21) always decreases with time and can be constant only in the equilibrium state (when $t \to \infty$ and $f(\mathbf{r},\mathbf{v},t)$ is Maxwellian), that is

$$\frac{\mathrm{d}\,H_t}{\mathrm{d}\,t} \le 0 \tag{4.22}$$

(for a proof, cf. [25]).

It is worth noticing that (microscopically) the quantity $-H_t$ being the Shannon entropy, characterizes the uncertainty of a microscopic state. Therefore, (3.14) implies that the evolution of the system is toward the more probable states. At the macroscopic level, adopting

the following relationship between thermodynamical entropy S and H_t for non-equlibrium states

$$S = -\kappa H_t \tag{4.23}$$

we have the known law of increase of entropy S or, the second principle of thermodynamics (for a Boltzmann gas in which the binary interactions between particles are accounted for). Since the Boltzmann H-therorem expresses the thermodynamical law of the entropy increase (in isolated systems) regarded as the mathematical manifestation of the irreversibility, it is interpreted as a statement on irreversibility of the kinetic Boltzmann equation.

There is no need to restrict the Boltzmann approach to classical statistical mechanics of gases and to the associated thermodynamical problems. The change of the entropy due to dynamical systems and related information flow is of interest in many other contexts. The natural questions are, for example, when (under what conditions) does a dynamical change of a system from a state $\mathbf{x}$ to $\mathbf{x}'$ cause the entropy increase $H_{\mathbf{x}} \le H_{\mathbf{x}'}$? For which dynamical systems does the entropy $H(\mathbf{x}_t)$ converge, as $t \to \infty$, to some specific value? Do there exist a state $\mathbf{x}_*$ and the unique limit $\lim_{t\to\infty} H(\mathbf{x}_t)$ such that $\lim_{t\to\infty} H(\mathbf{x}_t) = H(\mathbf{x}_*)$? Although such and related issues are relevant to general dynamical systems (e.g. chaotic, stochastic) they may also be of interest for thermodynamics, especially in investigating of the problems of irreversible and nonequilibruim process (cf. [27], [28]).

Let us consider a general dynamical system evolving in a phase space χ. A large class of systems dynamics can be characterized by the evolution of densities in χ-space. Roughly speaking, a density is an integrable function $f(x)$ defined in χ with values in R, such that $f(x) \ge 0$ and $\|f\|_{L_1} = 1$. For example, if X is a random variable with values in χ characterized by its probability distribution $\Phi(x)$ then (if $\Phi(.)$ is absolutely continuous with respect to the Lebesgue measure μ on χ) $f(x)$ is the Radon-Nikodem derivative of $\Phi(x)$ with respect to μ.

Let us consider a dynamical system governed by a system of ordinary differential equations

$$\frac{\mathrm{d}\, X_i(t)}{\mathrm{d}\, t} = F_i(\mathbf{X}) \quad , \quad i = 1,2,\ldots,n \tag{4.24}$$

where $\mathbf{X}(t) = [X_1, X_2, \ldots, X_n]$ is a state vector in the phase space. As is known (cf. [2]), starting from an initial density $f_0(\mathbf{x})$, the evolution of the time-dependent density $f(\mathbf{x},t)$ is described by the Liouville equation

$$\frac{\partial f}{\partial t} = -\sum_i \frac{\partial (f F_i)}{\partial x_i} \tag{4.25}$$

The density $f(\mathbf{x},t)$ can be represented by a Frobenius-Perron operator P_t (cf. [16]); it is any linear operator P_t from a space L_1 to L_1 such that: (i) $P_t f \geq 0$, (ii) $\|P_t f\| = \|f\|$ for all t and $f \geq 0$, $f \in L_1$. Namely, we can write that $f(\mathbf{x},t) = P_t f(\mathbf{x})$. If some density f_* satisfies $P_t f_* = f_*$ for all t, then f_* is called a stationary density of the operator P_t. For the Liouville equation (4.25) the stationary density f_* is given by the solution of equation

$$-\sum_i \frac{\partial (f, F_i)}{\partial x_i} = 0 \tag{4.26}$$

It has been shown in [17] that the conditional entropy $H_c(f|f_*)$ defined by (3.14) satisfies the equation

$$\frac{\mathrm{d}H_c}{\mathrm{d}t} = 0 \tag{4.27}$$

which implies that the conditional entropy (measuring a "distance" between $f(\mathbf{x},t)$ and $f_*(\mathbf{x})$) does not change during the motion of system (4.24).

The situation is, however, different if a system in question is subjected to random noise, i.e. instead of (4.24) we have a system

$$\frac{\mathrm{d}X_i(t)}{\mathrm{d}t} = F_i(\mathbf{X}) + \sigma_{ij}(\mathbf{X})\xi_i(t) \ , \quad i = 1,2,\ldots,n \tag{4.28}$$

where $\xi_i(t)$ are independent white noises with state-dependent intensity $\sigma^2(\mathbf{X})$. As it is known (cf. [2]) the time-dependent density satisfies the Fokker-Planck-Kolmogorov equation

$$\frac{\partial f}{\partial t} = -\sum_i \frac{\partial (f F_i)}{\partial x_i} + \frac{1}{2}\sum_{ij} \frac{\partial^2 (\sigma_{ij}^2 f)}{\partial x_i \partial x_j} \tag{4.29}$$

The stationary probability density $f_*(\mathbf{x})$ is governed by the equation

$$-\sum_i \frac{\partial (f F_i)}{\partial x_i} + \frac{1}{2}\sum_{ij} \frac{\partial^2 (\sigma_{ij}^2 f)}{\partial x_i \partial x_j} = 0 \tag{4.30}$$

The following assertion holds (cf. [16], [17]). If there exists a unique solution $f(\mathbf{x},t)$ of the Fokker-Planck-Kolmogorov equation (4.29) and there is a unique stationary density

$$\lim_{t\to\infty} f(\mathbf{x},t) = \lim_{t\to\infty} P_t f_0(\mathbf{x}) = f_*(\mathbf{x}) \tag{4.31}$$

then

$$\frac{\mathrm{d}\, H_c(f|f_*)}{\mathrm{d}t} \geq 0 \tag{4.32}$$

and

$$\lim_{t\to\infty} H_c(P_t f|f_*) = 0 \tag{4.33}$$

and (if $\sigma_{ij}(\mathbf{x}) = \sigma g(\mathbf{x})$) the rate of convergence of $H_c(f|f_*)$ to zero is proportional to the noise intensity σ^2. The growth of entropy in time for noisy system represented by inequality (4.32) is associated with irreversibility in time of system (4.28) (cf. [17]) and with some other complex problems of non-equilibrium thermodynamics (cf. [27], [28]).

Another type of problems concerned with the information flow in dynamical systems occurs in the analysis of chaotic dynamical systems. One of the central issues is how the information about the initial conditions is lost and how chaos can be characterized in terms of information theory (cf. [29], [30]). One wishes to characterize the rate of entropy generation in dynamical systems during the chaotic motion as well as the degree of randomness in one or multi-step predictions based on the whole past of a chaotic system dynamics.

Conventionally, dynamical systems are represented as a "flow" in phase space. The complete state of a system at any time is specified by a point in phase space, and the motion of a system in time generates a trajectory or orbit through the space. A "flow" in the phase state is governed by a set of n first-order differential equations of the form (4.24), where n is the dimensionality of the space. Usually in this formalism there is a tacit assumption of reversibility, i.e. that no information is lost as time passes (in such a case the conservation of energy and the Liouville theorem hold). Conservation of the phase volume during the flow means that dimensionality of the solutions remains fixed.

However, in nonconservative or dissipative systems the phenomenon of attraction of the orbits is possible. In two-dimensional phase flow the only (stable) possible attraction sets are fixed points and limit cycles (the Poincare-Bendixon theorem). In higher dimensions it is possible to have such types of flows which expand phase volumes in some dimensions while contracting them in others. In the contemporary physics there is a contention that the effect of these flows is to *create* new information which was not implicitly given in the initial conditions of the flow. For example, it has been argued in [29] that the systems characterized by a class of maps of the unit interval onto itself, "create information, in the sense that any physical realization of them systematically brings the uncertainties". The same interpretation holds for dynamical flows in three dimensions, e.g. for the Lorentz system (governed by a set of three nonlinear differential equations). This type of reasoning leads to understanding that the average rate of information (entropy) creation $\overline{\lambda}$ in a chaotic turbulent system is given by a Lyapunov characteristic exponent. The transition of a system from laminar to turbulent behaviour is characterized by the change of $\overline{\lambda}$ from negative to positive values, corresponding to the change of the system from an information sink to an information source. Shaw [29] argues

that the main quantitative difference between laminar and turbulent flow lies in the direction of information flow between the macroscopic and microscopic length scales. In laminar flow, motion is governed by boundary and initial conditions and no new information is created by the flow, hence the motion is predictable. Turbulent motion is, however, governed by information created continuously by the flow itself; this fact makes both predictability and reversibility impossible. A more quantitative analysis of the intrinsic information flow in chaotic systems is given in [30].

5. MAXIMUM ENTROPY PRINCIPLE FOR STOCHASTIC SYSTEMS

5.1 General idea

The principle of maximum entropy, described in Section 3 and playing a principal role in statistical physics, has recently been developed to stochastic dynamical systems when a prior information is given in the form of moment equations (derivable from given stochastic differential equations by use of the Itô formula or, from the associated Fokker-Planck-Kolmogorov equation) -cf. Sobczyk, Trębicki [31], [32], [33] and Trębicki, Sobczyk [34]. The idea of this extension is as follows.

Let the system of interest be governed by the following stochastic Itô equation for the vector process $\mathbf{Y}(t) = [Y_1(t),...,Y_n(t)]$

$$\mathrm{d}\,\mathbf{Y}(t) = \mathbf{F}\big[\mathbf{Y}(t)\big]\mathrm{d}\,t + \sigma\big[\mathbf{Y}(t)\big]\mathrm{d}\,\mathbf{W}(t,\gamma) \tag{5.1}$$

where $\mathbf{W}(t,\gamma) = [W_1(t,\gamma),...,W_m(t,\gamma)]$ is the m-dimensional Wiener process. Under known conditions (cf. Sobczyk [2]) the solution of (1) is a diffusion Markov process with the following drift vector $\mathbf{A}(\mathbf{y})$ and diffusion matrix $\mathbf{B}(\mathbf{y})$

$$\mathbf{A}(\mathbf{y}) = \mathbf{F}(\mathbf{y}) \quad , \quad \mathbf{B}(\mathbf{y}) = \sigma(\mathbf{y})\sigma^{\mathrm{T}}(\mathbf{y}) \tag{5.2}$$

or, within the Stratonovich interpretation of the equation with white noise

$$\tilde{\mathbf{A}}(\mathbf{y}) = \mathbf{F}(\mathbf{y}) + \frac{1}{2}\sigma(\mathbf{y})\frac{\partial\sigma(\mathbf{y})}{\partial\mathbf{y}} \quad , \quad \tilde{\mathbf{B}}(\mathbf{y}) = \mathbf{B}(\mathbf{y}) \tag{5.3}$$

The equations for moments are derived easily by use of Itô formula to the function $\mathrm{h}_{\mathbf{k}} = Y_1^{k_1}...Y_n^{k_n}$ of the solution and taking the average. The symbol $\mathbf{k}$ denotes here the multi-index i.e. $\mathbf{k}=(k_1,...,k_n)$; we will denote: $|\mathbf{k}|=k_1+...+k_n$ and $|\mathbf{k}|=1,2,..,K$.
The moments of process $\mathbf{Y}(t)$ at time t are defined as usual

$$m_{\mathbf{k}} = \left\langle Y_1^{k_1}Y_2^{k_2}...Y_n^{k_n}\right\rangle_f = \left\langle \mathrm{h}_{\mathbf{k}}(\mathbf{Y}(t))\right\rangle_f \tag{5.4}$$

where $\langle\cdot\rangle_f$ denotes the mean value of the quantity indicated, i.e. $\langle\cdot\rangle_f$ is the integral of $y_1^{k_1}\ldots y_n^{k_n}$ with respect to the true probability density $f(\mathbf{y};t)=f(y_1,\ldots,y_n;t)$ of the solution process. The general form of the moment equations is

$$\frac{\mathrm{d}\,m_{\mathbf{k}}(t)}{\mathrm{d}\,t}=\sum_i\left\langle \mathrm{F}_i\frac{\partial \mathrm{h}_{\mathbf{k}}}{\partial Y_i}\right\rangle_f+\frac{1}{2}\sum_i\sum_{i,j}\left\langle \sigma_{il}\sigma_{jl}\frac{\partial^2 \mathrm{h}_{\mathbf{k}}}{\partial Y_i\partial Y_j}\right\rangle_f \tag{5.5}$$

The initial conditions $m_{\mathbf{k}}(t_0)$ are specified from the given probability density $f(\mathbf{y},t_0)$ of the initial condition $\mathbf{Y}_0(\gamma)$.

If $F_i(\mathbf{Y})$ and $\sigma_{ij}(\mathbf{Y})$ are polynomials with respect to $Y_1,\ldots,Y_n$, equations (5.5) can be represented symbolically as the following infinite hierarchy of equations

$$\frac{\mathrm{d}\,m_{\mathbf{k}}(t)}{\mathrm{d}\,t}=g_{\mathbf{k}}(m_1,\ldots,m_{\mathbf{k}},\ldots)\quad,\quad |\mathbf{k}|=1,2,\ldots \tag{5.6}$$

where $g_{\mathbf{k}}$ are functions of moments specified on the basis of given stochastic system.
The finite set of moment equations usually considered is

$$\frac{\mathrm{d}\,m_{\mathbf{k}}(t)}{\mathrm{d}\,t}=g_{\mathbf{k}}(m_1,\ldots,m_r)\quad,\quad r\geq|\mathbf{k}| \tag{5.7}$$

where $|\mathbf{k}|=1,2,\ldots,K$, and K is a specified number - the highest order of the moment (5.4). Since, in general $r\geq|\mathbf{k}|$ system (5.7) is not closed. The same property has a finite collection of equations from hierarchy (5.5).

According to the spirit of the maximum entropy principle the approximate probability density $p(\mathbf{y};t)$ of the stochastic process $\mathbf{Y}(t)$ governed by general system (5.1) is determined as a result of maximization of the entropy functional

$$\mathrm{H}[p]=-\int p(\mathbf{y};t)\ln p(\mathbf{y};t)\,\mathrm{d}\,\mathbf{y} \tag{5.8}$$

under constrains (5.5) and normalization condition

$$\int p(\mathbf{y};t)\,\mathrm{d}\,\mathbf{y}=1 \tag{5.9}$$

The integration in (5.8) and (5.9) is extended over the range of the possible values $\mathbf{Y}(t)$ for each t.

Let us notice that constraints (5.6) and (5.10) in the maximum entropy scheme can be represented as

$$\frac{\mathrm{d}\,m_{\mathbf{k}}(t)}{\mathrm{d}\,t}=\left\langle \mathrm{G}_{\mathbf{k}}(\mathbf{y})\right\rangle_p \tag{5.10}$$

$$\int p(\mathbf{y};t) - 1 = 0 \tag{5.11}$$

where

$$G_{\mathbf{k}}(\mathbf{y}) = \sum_i F_i \frac{\partial h_{\mathbf{k}}(\mathbf{y})}{\partial y_i} + \frac{1}{2}\sum_i \sum_{i,j} \sigma_{il}\sigma_{jl} \frac{\partial^2 h_{\mathbf{k}}(\mathbf{y})}{\partial y_i \partial y_j} \tag{5.12}$$

5.2 Stationary distributions

Let us consider first the stationary probability distributions of the response process (treated in detail in papers [31], [32]). In this case the moments $m_{\mathbf{k}}(t)$ do not dependent on time and constraints (5.10) in the variational problem (5.8)-(5.12) takes a "non-differential" form

$$\left\langle G_{\mathbf{k}}(\mathbf{Y}) \right\rangle_p = \int G_{\mathbf{k}}(\mathbf{Y}) p(\mathbf{y}) \mathrm{d}\,\mathbf{y} = 0 \tag{5.13}$$

In this case the maximum entropy distribution takes the form (cf. [31], [32])

$$p(\mathbf{y}) = C\exp\{-\sum_{|\mathbf{k}|=1}^{K} \lambda_{\mathbf{k}} G_{\mathbf{k}}(\mathbf{y})\} = C\exp\{-\sum_{k_1+\ldots+k_n=1}^{K} \lambda_{k_1\ldots k_n} G_{k_1\ldots k_n}(y_1,\ldots,y_n)\} \tag{5.14}$$

Unknown Lagrange multipliers $\lambda_{k_1\ldots k_n}$ are determined from constraints (5.13). The normalizing constant C is calculated from equation (5.11). Therefore, the problem is reduced to the solution of system of algebraic (or, transcendental, if nonlinearities are not algebraic) equations for Lagrange multipliers. In general, the standard numerical procedures (e.g. Newton method) constitute, an effective tool for obtaining a solution. Of course, as always, the amount of computational work depends on the order n of the system considered and the number of moment equations taken into account.

In the case of the most common illustrative vibratory system described by the equation

$$\ddot{Y}(t) + \beta\dot{Y}(t) + g(Y) = \xi(t,\gamma), \qquad Y(t_0) = Y_{10}, \quad \dot{Y}(t_0) = Y_{20} \tag{5.15}$$

where $\xi(t,\gamma)$ is a Gaussian white noise with intensity $2D$ (β is a damping coefficient), the corresponding Itô system is ($Y_1 = Y, Y_2 = \dot{Y}$)

$$\begin{aligned} \dot{Y}_1 &= Y_2 \quad , \qquad Y_1(t_0) = Y_{10}, \quad Y_2(t_0) = Y_{20} \\ \dot{Y}_2 &= -\beta Y_2 - g(Y_1) + \sqrt{2D}\,\mathrm{d}\,W(t,\gamma) \end{aligned} \tag{5.16}$$

where $W(t)$ is the Wiener process. The system of moment equations (5.13) is $\left\langle G_{ij}(Y_1,Y_2)\right\rangle$=0 where

$$G_{ij}(Y_1,Y_2)=iY_1^{i-1}Y_2^{j+1}-j[\beta Y_2+g(Y_1)]Y_1^iY_2^{j-1}+Dj(j-1)Y_1^iY_2^{j-2} \tag{5.17}$$
$$i,j=0,1,\ldots,K,\ 0\neq i+j\leq K$$

In the paper [32] the results of calculation have been shown for the case when $0\neq i+j\leq 2=K=2$, (i.e. five moment equations is taken into account) and for Duffing nonlinearity $g(Y_1)=Y_1+\varepsilon Y_1^3$. The results have been compared with the exact solution as well as with statistical linearization and Gram-Charlier expansions. It turned out that the error with respect to the true exact solution decreases with increase of the strength of nonlinearity what distinguishes the maximum entropy method from other procedures. (cf. Figures 1,2).

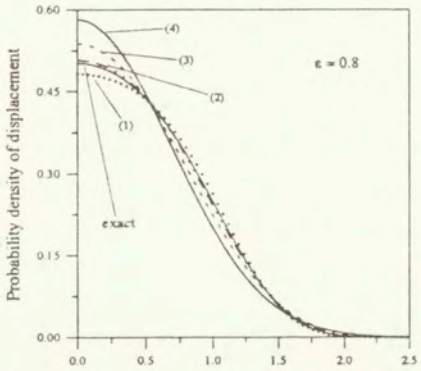

Figure 1. Comparison of the exact displacement distribution of Duffing oscillator with its approximations obtained for small nonlinearity: (1) maximum entropy method; (2) Gram-Charlier exp. 6th order; Gram-Charlier exp. 4th order; (4) statistical linearization

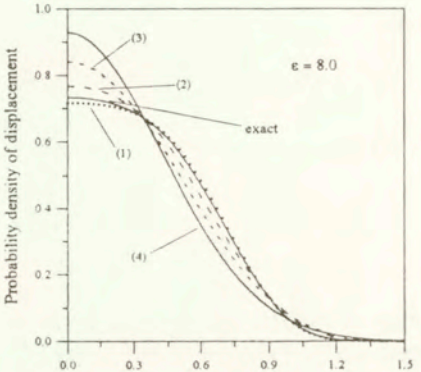

Figure 2.Comparison of the exact displacement distribution of Duffing oscillator with its approximations obtained for strong nonlinearity. Keys as in Fig. 1.

5.3 Non-stationary distributions

It has been tempting to extend the idea of maximum entropy method to nonstationary behaviour of stochastic systems. In such a case, the variational problem for the entropy functional includes time dependent constraints in the form of differential equations for moments. It turns out that if one treats such a time-dependent maximization problem as being a parametric with respect to time (we deal with one-dimensional probability distributions, i.e. with distributions and moments parametrized by t) than the unknown probability density $p(y,t)$ can still be expressed in the exponential form (5.14) in which $\lambda_k=\lambda_k(t)$. So, we have time-dependent Lagrange multipliers which should be determined from the moment restrictions, i.e. from the system of equations

$$\frac{\mathrm{d}\, m_{\mathbf{k}}(t)}{\mathrm{d}\, t} = \int \mathrm{G}_{\mathbf{k}}(\mathbf{y})\, p(\mathbf{y};\lambda_{\mathbf{k}}(t))\,\mathrm{d}\,\mathbf{y} \tag{5.18}$$

$$\frac{\mathrm{d}\, m_{\mathbf{k}}(t)}{\mathrm{d}\, t} = \int \mathrm{h}_{\mathbf{k}}(\mathbf{y})\frac{\partial}{\partial t} p(\mathbf{y};t)\,\mathrm{d}\,\mathbf{y} \tag{5.19}$$

Above equations include differential of $\lambda_{\mathbf{k}}(t)$ and, therefore constitute a system of non-linear differential equations for $\lambda_{\mathbf{k}}(t)$. A detailed presentation of this approach has been given in Trębicki, Sobczyk [34], along with its modification which improves the computational efficiency.

Bearing in mind a computational aspect of the method we wish to mention here an approach which is based on discretized moment equations. The system of moment equations associated with the stochastic system in question has the form

$$\frac{\mathrm{d}\, m_{\mathbf{k}}(t)}{\mathrm{d}\, t} = \frac{\mathrm{d}}{\mathrm{d}\, t}\left\langle \mathbf{Y}^{\mathbf{k}} \right\rangle = \int \mathrm{G}_{\mathbf{k}}(\mathbf{y})\, f(\mathbf{y};t)\,\mathrm{d}\,\mathbf{y} \tag{5.20}$$

where, as above, $\mathbf{k} = (k_1,\ldots,k_n)$ and $\mathbf{Y}^{\mathbf{k}} = Y_1^{k_1} Y_2^{k_2} \ldots Y_n^{k_n}$.
Let us discretize system (5.20) using i.e. the Euler scheme

$$\frac{\mathrm{d}}{\mathrm{d}\, t}\left\langle \mathbf{Y}^{\mathbf{k}} \right\rangle = \frac{1}{\Delta t}\left[\left\langle \mathbf{Y}^{\mathbf{k}}(t+\Delta t)\right\rangle - \left\langle \mathbf{Y}^{\mathbf{k}}(t)\right\rangle\right] \tag{5.21}$$

System of moment equations (5.20) takes the form

$$\begin{aligned} \left\langle \mathbf{Y}^{\mathbf{k}}(t+\Delta t)\right\rangle &= \left\langle \mathbf{Y}^{\mathbf{k}}(t)\right\rangle + \Delta t \int \mathrm{G}_{\mathbf{k}}(\mathbf{y})\, f(\mathbf{y};t)\,\mathrm{d}\,\mathbf{y} \\ \left\langle \mathbf{Y}^{\mathbf{k}}(t_0)\right\rangle &= m_{\mathbf{k}}(t_0) \end{aligned} \tag{5.22}$$

Assuming that the considered time interval is T and denoting by I the number of iterations we get $\Delta t = T/I$, and $\Delta t = t_{i+1} - t_i$, $i=0,1,\ldots,I-1$. System (5.20) can be written as

$$m_{\mathbf{k}}(t_{i+1}) = m_{\mathbf{k}}(t_i) + \Delta t \int \mathrm{G}_{\mathbf{k}}(\mathbf{y}) f(\mathbf{y};t_i)\,\mathrm{d}\,\mathbf{y} \tag{5.23}$$

with $m_{\mathbf{k}}(t_0)$ assumed to be given initial condition. Therefore, the variational - maximum entropy problem can be formulated as: in order to determine the approximation $p(\mathbf{y};t)$ of the true density $f(\mathbf{y};t)$ in the discretization points t_i, $i=1,2,\ldots,I$ we maximize the entropy

$$\mathrm{H}_i[p] = -\int p(\mathbf{y};t_i) \ln p(\mathbf{y};t_i)\,\mathrm{d}\,\mathbf{y} \tag{5.24}$$

under the discretized moment constraints (5.23). In this way the original maximum entropy problem (5.8)-(5.12) for stochastic systems is reduced to the iterative sequence of classical maximum entropy problems (5.23), (5.24). Of course, to start the numerical procedure we determine first $p(\mathbf{y};t_0)$ via maximum entropy algorithm with constraint $m_k(t_0)$.
Application of this idea to specific stochastic mechanical systems along with numerical implementation and graphical illustration of the results can be found in papers [34], [33].

5.4 Relation to statistical thermodynamics: remarks.

The maximum entropy method has been presented in this section as a tool for approximative characterization of the probability distributions of stochastic dynamical systems and is entirely in the spirit of mathematical statistics. Indeed, we have been looking for a "most rational" (or "maximally unbiased by our ignorance") approximative distribution which is in agreement with given information (equations for moments). The physical nature of the system itself may be different depending on the interpretation of the dynamical variables $Y_1, Y_2, \ldots, Y_n$.

If we look at the system from the point of view of statistical physics, then vector $\mathbf{Y} = [Y_1, \ldots, Y_n]$ characterizes the state of the system in the phase space of canonical variables (generalized coordinates and conjugate momenta of particles). In the state of *thermodynamical equilibrium* canonical Gibbs distribution coincides with the maximum entropy distribution when the average energy of system is prescribed (cf. [28]). And more, this maximum entropy can be related to the phenomenological thermodynamical entropy S (cf. formula (4.23)). The existence of a stationary distribution (density) of stochastic system may be associated with the state of thermodynamic equilibrium. Therefore, we can say that the stationary maximum entropy distribution (5.14) of a general Itô stochastic differential system can be regarded as its Gibbs canonical distribution.

The situation is much more complicated when we consider non-stationary systems. In this case a relevant thermodynamics is non-equilibrium statistical thermodynamics. In statistical physics there have been various attempts to construct the non-equilibrium microscopic distributions on the basis of the maximum entropy principle. For example, the so called *local equilibrium* formalism (cf. [28]) uses the assumption that the exact non-equilibrium statistical density is in some sense approximately equal to the local equilibrium one. Other approaches (cf. [35], [36]) use, as constraints the information collected at fixed instant of time, so the distributions obtained can be called "quasi-equilibrium" ones (they do not describe of irreversible processes). Kalashnikov and Zubariev (cf. [28]) showed, however, that the maximum entropy distributions can describe irreversible processes if one looks for maximum entropy given macroscopic quantities not for fixed instant but also for all past time history. Such a priori information in maximum entropy principle takes into account the effects of memory in macroscopic (averaged) quantities and mathematically manifests itself in functional dependence of density p on moments $m_k(t+\tau)$ where $-\infty < \tau \leq 0$.

The microscopic (statistical) formulation of non-equilibrium thermodynamics including the definition of entropy itself, has been since many years a long standing problem. Also, the role of maximum entropy formalism for non-equilibrium processes in physics still remains a subject of continuing debate (cf. [37], [38]).

Acknowledgements.

The author expresses his thanks to Professor J. Telega for a careful reading of the manuscript and numerous corrections as well as to Dr J. Trębicki for many fruitful discussions and for bringing my hand made manuscript to its present form.

REFERENCES

1. Klir G.J., Folger T.A., *Fuzzy Sets, Uncertainty and Information*, Prentice Hall, Englowood, Cliffs, 1988.
2. Sobczyk K., *Stochastic Differential Equations with Application to Physics and Engineering*, Kluwer Acad. Publ., Dordrecht, 1991.
3. Waever W., Shannon C.E., *The Mathematical Theory of Communication*, Univ. of Illinois Press, Urbana, 1949.
4. Kullback S., *Information Theory and Statistics*, Chapman and Hall, N. York, 1959.
5. Ingarden R.S., Urbanik K., Information without probability, *Coll. Math.*, Vol. 9, pp.131-150, 1963.
6. Kolmogorov N., Three approaches to the quantitative definition of information, *Problems of Inform. Transmission*, Vol. 1, pp. 4-7, 1965.
7. Chaitin G.J., *Algorithmic Information Theory*, Cambrige University Press, Cambridge, 1987.
8. Csiszar R.I., Information-type distance measure of probability distributions and indirect observations, *Studia Scient. Mathem. Hungarica*, Vol. 2, pp. 299-318, 1967.
9. Sugimoto S., Wada T., Spectral expressions of information measures of Gaussian time series and their relation to AIC and CAT, *IEEE Trans. on Inform. Theory*, Vol. 34, No. 4, 1988.
10. Ruelle D., Takens F., On the nature of turbulence, *Comm. Math. Phys.*, Vol. 20, pp. 167-172, 1971.
11. Jumarie G., Some approaches to the measures of the amount of information involved by a form, *System Analysis, Modelling and Simulation*, Vol. 3., pp. 479-506, 1986.
12. Cover T.M., Thomas J.A., *Elements of Information Theory*, J.Wiley&Sons, New York, 1991
13. Jaynes E.T., Information theory and statistical mechanics, *Phys. Rev.*, Vol. 106, pp.620-630, 1957.
14. Ingarden R.S., Information theory and variational principles in statistical theories, *Bull. Acad. Polon. Sci., Ser. Math. Astr. Phys.*, **11**, pp. 541-547, 1963
15. Good I.J., Maximum entropy for hypothesis formulation, *Ann. Math. Stat.*, Vol. 34, pp. 911-934, 1963.
16. Lasota A., Mackey M.C., *Chaos, Fractals and Noise; Stochastic Aspects of Dynamics*, Sec. Ed., Springer, New York, 1994

17. Mackey M.C., The dynamic origin of increasing entropy, *Rev. Modern Phys.*, Vol. 61, pp. 763-916, Oct. 1989.
18. Meyer M.E., Gokhale D.V., Kullback-Leibler information measure for studying convergence rates of densities and distributions, *IEEE Trans. Inform. Theory*, Vol. 39, pp.1401-1403, July 1993
19. Share J.E., Johnson R.W., Axiomatic derivation of the principle of maximum cross-entropy, *IEEE Trans. Inform. Theory*, Vol. IT-26, pp.26-37, January 1980.
20. Shore J.E., Johnson R.W., Properties of cross-entropy minimization, *IEEE Trans. Inform. Theory*, Vol. IT-27, pp.472-482, July 1981.
21. Diafari M.A., Demoment G., Maximum entropy and Bayesian approach in tomographic image reconstruction and restoration, in "*Maximum Entropy and Bayesian Methods*", (Ed. Skilling J.), Kluwer Acad. Publ., 1989.
22. Pristley M.B, *Spectral Analysis and Time Series*, Academic Press, 1981.
23. Pinsker M.S., *Information and Information Stability of Random Variables and Processes*, Holden-Day, San Francisco, 1964.
24. Ekroot L., Cover T.M., The entropy of Markov trajectories, *IEEE Trans. Inform.*, Vol. 39, pp.1418-1421, July 1993
25. Cercignani C., *Theory and Application of the Boltzmann Equation*, Academic Press, Edinburgh-London, 1975.
26. Dougherty J.P., Approaches to non-equilibrium statistical mechanics, in in "*Maximum Entropy and Bayesian Methods*", (Ed. Skilling J.), Kluwer Acad. Publ., 1989.
27. Garret A.J.M., Irreversibility, probability and entropy, in: "*From Statistical Physics to Statistical Inference and Back*", (Eds. Grassberger P., Nadal J.P), Kluwer Acad. Publ., 1994.
28. Zubariev D.N., *Non-equlibrium Statistical Thermodynamics*, Nauka, Moscov, 1971, (in Russian).
29. Shaw R., Strange atrractors, chaotic behaviour and information flow, *Z. Naturforsh*, A36, pp. 80-112, 1981.
30. Deco G., Schittenkopf C., Schrümann B., Determining the information flow in dynamical systems from continous probability distributions, Phys. Rev., Letters, Vol. 73, No. 12, March 1997.
31. Sobczyk K., Trębicki J., Maximum entropy principle in stochastic dynamics, *Probabilistic Eng. Mech.*, Vol. 5, No. 3. pp. 1-10, 1990.
32. Sobczyk K., Trębicki J., Maximum entropy principle and non-linear stochastic oscillators, *Physica A*, **193**, pp. 448-468, 1993.
33. Sobczyk K., Trębicki J., Approximate Probability Distributions for Stochastic Systems: Maximum Entropy Method, *Comput. Methods Appl. Mech. Engrg.*, 168, pp.91-111, 1999.
34. Trębicki J., Sobczyk K., Maximum entropy principle and non-stationary distributions of stochastic systems, *Probab. Engng Mechanics*, Vol. 11, pp. 169-178, 1996.
35. Jaynes E.T., *Papers on Probability, Statistics and Statistical Physics*, Reidel, 1983.
36. Lavenda B.H., Scherer C., Statistical inference in equilibrium and non-equilibrium thermodynamics, *Riv. Nuovo Cimento*, Vol. 11, No. 6, 1988.
37. Ingarden R.S, Kossakowski A., Ohya M., *Information Dynamics and Open Systems: Classical and Quantum Approach*, Kluwer Academic Publ., Dordrecht, Boston, 1997.
38. Haken H., *Information and Self-organization: A Macroscopic Approach to Complex Systems*, Springer, Berlin Heidelberg, 1988.