Entropy, **2001,** *3*, 191–226

entropy
ISSN 1099-4300
www.mdpi.org/entropy/

Maximum Entropy Fundamentals

P. Harremoës[1] **and F. Topsøe**[2,*]

1. Rønne Allé, Søborg, Denmark
email: moes@post7.tele.dk
2. Department of Mathematics; University of Copenhagen; Denmark
email: topsoe@math.ku.dk

Received: 12 September 2001/ Accepted: 18 September 2001 / Published: 30 September 2001

Abstract: In its modern formulation, the *Maximum Entropy Principle* was promoted by E.T. Jaynes, starting in the mid-fifties. The principle dictates that one should look for a distribution, consistent with available information, which maximizes the entropy. However, this principle focuses only on distributions and it appears advantageous to bring information theoretical thinking more prominently into play by also focusing on the "observer" and on coding. This view was brought forward by the second named author in the late seventies and is the view we will follow-up on here. It leads to the consideration of a certain game, the *Code Length Game* and, via standard game theoretical thinking, to a principle of *Game Theoretical Equilibrium.* This principle is more basic than the Maximum Entropy Principle in the sense that the search for one type of optimal strategies in the Code Length Game translates directly into the search for distributions with maximum entropy.

In the present paper we offer a self-contained and comprehensive treatment of fundamentals of both principles mentioned, based on a study of the Code Length Game. Though new concepts and results are presented, the reading should be instructional and accessible to a rather wide audience, at least if certain mathematical details are left aside at a first reading.

The most frequently studied instance of entropy maximization pertains to the *Mean Energy Model* which involves a moment constraint related to a given function, here taken to represent “energy”. This type of application is very well known from the literature with hundreds of applications pertaining to several different fields and will also here serve as important illustration of the theory. But our approach reaches further, especially regarding the study of continuity properties of the entropy function, and this leads to new results which allow a discussion of models with so-called *entropy loss*. These results have tempted us to speculate over the development of natural languages. In fact, we are able to relate our theoretical findings to the empirically found *Zipf's law* which involves statistical aspects of words in a language. The apparent irregularity inherent in models with entropy loss turns out to imply desirable stability properties of languages.

Keywords: Maximum Entropy, Minimum Risk, Game Theoretical Equilibrium, Information Topology, Nash Equilibrium Code, Entropy Loss, Partition Function, Exponential Family, continuity of entropy, hyperbolic distributions, Zipf's law.

---

*Research of both authors supported by the Danish Natural Science Research Council, by INTAS (project 00-738) and by IMPAN–BC, the Banach Center, Warsaw.

1 The Maximum Entropy Principle – overview and a generic example

The *Maximum Entropy Principle* as conceived in its modern form by Jaynes, cf. [11], [12] and [13], is easy to formulate: "Given a model of probability distributions, choose the distribution with highest entropy." With this choice you single out the most significant distribution, the least biased one, the one which best represents the "true" distribution. The sensibility of this principle in a number of situations is well understood and discussed at length by Jaynes, in particular.

The principle is by now well established and has numerous applications in physics, biology, demography, economy etc. For practically all applications, the key example which is taken as point of departure – and often the only example discussed – is that of models prescribed by moment conditions. We refer to Kapur, [14] for a large collection of examples as well as a long list of references.

In this section we present models defined by just one moment condition. These special models will later be used to illustrate theoretical points of more technical sections to follow.

Our approach will be based on the introduction of a two-person zero-sum game. The principle which this leads to, called the principle of *Game Theoretical Equilibrium* is taken to be even more basic than the Maximum Entropy Principle. In fact, from this principle you are led directly to the Maximum Entropy Principle and, besides, new interesting features emerge naturally by focusing on the interplay between a system and the observer of the system. As such the new principle is in conformity with views of quantum physics, e.g. we can view the principle of Game Theoretical Equilibrium as one way of expressing certain sides of the notion of complementarity as advocated by Niels Bohr in a precise mathematical way.

To be more specific, let us choose the language of physics and assume that on the set of natural numbers $\mathbb{N}$ we have given a function E, the *energy function*. This function is assumed to be bounded below. Typically, E will be non-negative. Further, we specify a certain finite *energy level*, λ, and take as our model all probability distributions with mean energy λ. We assume that the energy E_i in "state" $i \in \mathbb{N}$ goes fast enough to infinity that the entropies of distributions in the model remain bounded. In particular, this condition is fulfilled if $E_i = \infty$ for all i sufficiently large – the corresponding states are then "forbidden states" – and in this case the study reduces to a study of models with finite support.

Once you have accepted the Maximum Entropy Principle, this leads to a search for a maximum entropy distribution in the model. It is then tempting to introduce Lagrange multipliers and to solve the constrained optimization problem you are faced with in the standard manner. In fact, this is what practically all authors do and we shall briefly indicate this approach.

We want to maximize entropy $H = -\sum_1^\infty p_i \log p_i$ subject to the moment condition $\sum_1^\infty p_i E_i =$

λ and subject to the usual constraints $p_i \geq 0$; $i \in \mathbb{N}$ and $\sum_1^\infty p_i = 1$. Introducing Lagrange multipliers $-\beta$ and μ, we are led to search for a solution for which all partial derivatives of the function $H - \beta \sum_1^\infty p_i E_i + \mu \sum_1^\infty p_i$ vanish. This leads to the suggestion that the solution is of the form

$$p_i = \frac{\exp(-\beta E_i)}{Z(\beta)}\,;\ i \geq 1 \tag{1.1}$$

for some value of β for which the *partition function* Z defined by

$$Z(\beta) = \sum_{i=1}^{\infty} \exp(-\beta E_i) \tag{1.2}$$

is finite.

The approach is indeed very expedient. But there are difficulties connected with it. Theoretically, we have to elaborate on the method to be absolutely certain that it leads to the solution, even in the finite case when $E_i = \infty$ for i sufficiently large. Worse than this, in the infinite case there may not be any solution at all. This is connected with the fact that there may be no distribution of the form (1.1) which satisfies the required moment condition. In such cases it is not clear what to do.

Another concern is connected with the observation that the method of Lagrange multipliers is a completely general tool, and this very fact indicates that in any particular situation there may possibly be other ways forward which better reflect the special structure of the problem at hand. Thus one could hope to discover new basic features by appealing to more intrinsic methods.

Finally we note that the method of Lagrange multipliers cannot handle all models of interest. If the model is refined by just adding more moment constraints, this is no great obstacle. Then the distributions and partition functions that will occur instead of (1.1) and (1.2) will work with inner products of the form $\beta' E_i' + \beta'' E_i'' + \cdots$ in place of the simple product βE_i. In fact, we shall look into this later. Also other cases can be handled based on the above analysis, e.g. if we specify the geometric mean, this really corresponds to a linear constraint by taking logarithms, and the maximum entropy problem can be solved as above (and leads to interesting distributions in this case, socalled *power laws*). But for some problems it may be difficult, or even impossible to use natural extensions of the standard method or to use suitable transformations which will reduce the study to the standard set-up. In such cases, new techniques are required in the search for a maximum entropy distribution. As examples of this difficulty we point to models involving binomial or empirical distributions, cf. [8] and [22].

After presentation of preliminary material, we introduce in Section 3 the basic concepts related to the game we shall study. Then follows a section which quickly leads to familiar key results. The method depends on the information- and game-theoretical point of view. This does not lead to complete clarification. For the proper understanding of certain phenomena, a more thorough theoretical discussion is required and this is taken up in the remaining sections.

New results are related to so-called *entropy loss* – situations where a maximum entropy distribution does not exist. In the last section, these type of models are related to *Zipf's law* regarding statistical aspects of the semantics of natural languages.

Mathematical justification of all results is provided. Some technical results which we shall need involve special analytical tools regarding Dirichlet series and are delegated to an appendix.

2 Information theoretical preliminaries

Let $\mathbb{A}$, the *alphabet*, be a discrete set, either finite or countably infinite and denote by $\tilde{}M_+^1(\mathbb{A})$, respectively $M_+^1(\mathbb{A})$ the set of non-negative measures P on $\mathbb{A}$ (with the discrete Borel structure) such that $P(\mathbb{A}) \le 1$, respectively $P(\mathbb{A}) = 1$. The elements in $\mathbb{A}$ can be thought of in many ways, e.g. as *letters* (for purely information theoretical or computer science oriented studies), as *pure states* (for applications to quantum physics) or as *outcomes* (for models of probability theory and statistics).

For convenience, $\mathbb{A}$ will always be taken to be the set $\mathbb{N}$ of natural numbers or a finite section thereof, and elements in $\mathbb{A}$ are typically referred to by indices like $i, j, \cdots$.

Measures in $M_+^1(\mathbb{A})$ are *probability distributions*, or just *distributions*, measures in $\tilde{}M_+^1(\mathbb{A})$ are *general distributions* and measures in $\tilde{}M_+^1(\mathbb{A}) \setminus M_+^1(\mathbb{A})$ are *incomplete distributions*. For $P, Q, \cdots \in \tilde{}M_+^1(\mathbb{A})$, the point masses are, typically, denoted by $p_i, q_i, \cdots$.

By $\tilde{}K(\mathbb{A})$, we denote the set of all mappings $\kappa : \mathbb{A} \to [0;\infty]$ which satisfy *Kraft's inequality*

$$\sum_{i\in\mathbb{A}} \exp(-\kappa_i) \le 1\,. \tag{2.3}$$

Elements in $\tilde{}K(\mathbb{A})$ are *general codes*. The values of a general code κ are denoted κ_i. The terminology is motivated by the fact that if $\kappa \in \tilde{}K(\mathbb{A})$ and if the base for the exponential in (2.3) is 2, then there exists a binary prefix-free code such that the i'th code word consists of approximately κ_i binary digits.

By $K(\mathbb{A})$ we denote the set of mappings $\kappa : \mathbb{A} \to [0;\infty]$ which satisfy *Kraft's equality*

$$\sum_{i\in\mathbb{A}} \exp(-\kappa_i) = 1\,. \tag{2.4}$$

This case corresponds to codes without superfluous digits. For further motivation, the reader may wish to consult [23] or standard textbooks such as [3] and [6].

Elements in $K(\mathbb{A})$ are *compact codes*, for short just *codes*.

For mathematical convenience, we shall work with exponentials and logarithms to the base e.

For $\kappa \in {}^{\sim}K(\mathbb{A})$ and $i \in \mathbb{A}$, κ_i is the *code length* associated with i or, closer to the intended interpretation, we may think of κ_i as the code length of the code word which we imagine κ associates with i. There is a natural bijective correspondance between ${}^{\sim}M_+^1(\mathbb{A})$ and ${}^{\sim}K(\mathbb{A})$, expressed notationally by writing $P \leftrightarrow \kappa$ or $\kappa \leftrightarrow P$, and defined by the formulas

$$\kappa_i = -\log p_i\,, \quad p_i = \exp(-\kappa_i)\,.$$

Here the values $\kappa_i = \infty$ and $p_i = 0$ correspond to eachother. When the above formulas hold, we call (κ, P) a *matching pair* and we say that κ is *adapted* to P or that P is the general distribution which *matches* κ. If $\mathcal{P} \subseteq M_+^1(\mathbb{A})$ and $\kappa \in K(\mathbb{A})$, we say that κ is $\mathcal{P}$-*adapted* if κ is adapted to one of the distributions in $\mathcal{P}$. Note that the correspondance $\kappa \leftrightarrow P$ also defines a bijection between $M_+^1(\mathbb{A})$ and $K(\mathbb{A})$.

The *support* of κ is the set of $i \in \mathbb{A}$ with $\kappa_i < \infty$. Thus, with obvious notation, $\mathrm{supp}\,(\kappa) = \mathrm{supp}\,(P)$ where P is the distribution matching κ and $\mathrm{supp}\,(P)$ is the usual support of P.

For expectations – always w.r.t. genuine probability distributions – we use the bracket notation. Thus, for $P \in M_+^1(\mathbb{A})$ and $f : \mathbb{A} \to [-\infty; \infty]$, we put

$$\langle f, P \rangle = \sum_{i \in \mathbb{A}} f(i) p_i$$

whenever this is a well-defined extended real number. Mostly, our functions will be non-negative and then $\langle f, P \rangle$ will of course be a well defined number in $[0; \infty]$. In particular this is the case for *average code length* defined for $\kappa \in {}^{\sim}K(\mathbb{A})$ and $P \in M_+^1(\mathbb{A})$ by

$$\langle \kappa, P \rangle = \sum_{i \in \mathbb{A}} \kappa_i p_i.$$

Entropy and divergence are defined as usual, i.e., for $P \in M_+^1(\mathbb{A})$, the *entropy* of P is given by

$$H(P) = -\sum_{i \in \mathbb{A}} p_i \log p_i \tag{2.5}$$

or, equivalently, by $H(P) = \langle \kappa, P \rangle$ where κ is the code adapted to P. And for $P \in M_+^1(\mathbb{A})$ and $Q \in {}^{\sim}M_+^1(\mathbb{A})$ we define the *divergence* (or *relative entropy*) between P and Q by

$$D(P\|Q) = \sum_{i \in \mathbb{A}} p_i \log \frac{p_i}{q_i}\,. \tag{2.6}$$

Divergence is well defined with $0 \leq D(P\|Q) \leq \infty$ and $D(P\|Q) = 0$ if and only if $P = Q$.

The topological properties which we shall find useful for codes and for distributions do not quite go in parallel. On the coding side we consider the space ${}^{\sim}K(\mathbb{A})$ of all general codes and remark that this space is a metrizable, compact and convex Hausdorff space. This may be seen by embedding $\sim K(\mathbb{A})$ in the space $[0;\infty]^{\mathbb{A}}$ of all functions on $\mathbb{A}$ taking values in the compact space $[0;\infty]$. The topology on $\sim K(\mathbb{A})$ then is the topology of pointwise convergence. This is the only topology we shall need on $\sim K(\mathbb{A})$.

On the distribution side we shall primarily consider probability distributions but on the corresponding space, $M_+^1(\mathbb{A})$, we find it useful to consider two topologies, the *usual*, pointwise topology and then a certain stronger non-metrizable topology, the *information topology*.

As to the usual topology on $M_+^1(\mathbb{A})$ we remind the reader that this is a metrizable topology, indeed it is metrized by total variation defined by

$$V(P,Q) = \sum_i |p_i - q_i|.$$

We write $P_n \xrightarrow{V} P$ for convergence and $\overline{\mathcal{P}}^V$, $\overline{\text{co}}^V\mathcal{P}$ etc. for closure in this topology (the examples show the closure of $\mathcal{P}$ and of the convex hull of $\mathcal{P}$, respectively).

As to the information topology – the second topology which we need on the space $M_+^1(\mathbb{A})$ – this can be described as the strongest topology such that, for $(P_n)_{n\geq 1} \subseteq M_+^1(\mathbb{A})$ and $P \in M_+^1(\mathbb{A})$, $\lim_{n\to\infty} D(P_n\|P) = 0$ implies that the sequence $(P_n)_{n\geq 1}$ converges to P. Convergence in this topology is denoted $P_n \xrightarrow{D} P$. We only need convergence in this topology for sequences, not for generalized sequences or nets. Likewise, we only need sequential closure and $\overline{\mathcal{P}}^{D_\sigma}, \overline{\text{co}}^{D_\sigma}\mathcal{P}$, or what the case may be denotes sequential closure. Thus $\overline{\mathcal{P}}^{D_\sigma}$ denotes the set of distributions P for which there exists a sequence $(P_n)_{n\geq 1}$ of distributions in $\mathcal{P}$ with $P_n \xrightarrow{D} P$. The necessary and sufficient condition that $P_n \xrightarrow{D} P$ holds is that $D(P_n\|P) \to 0$ as $n \to \infty$. We warn the reader that the corresponding statement for nets (generalized sequences) is wrong – only the sufficiency part holds generally. For the purposes of this paper, the reader needs only worry about sequences but it is comforting to know that the sequential notion $P_n \xrightarrow{D} P$ is indeed a topological notion of convergence. Further details will be in [9].

An important connection between total variation and divergence is expressed by *Pinsker's inequality*:

$$D(P\|Q) \geq \frac{1}{2}V(P,Q)^2\,, \tag{2.7}$$

which shows that convergence in the information topology is stronger than convergence in total variation.

The functions of relevance to us, entropy and divergence, have important continuity properties: $P \curvearrowright H(P)$ is lower semi-continuous on $M_+^1(\mathbb{A})$ and $(P,Q) \curvearrowright D(P\|Q)$ is jointly lower semi-continuous on $M_+^1(\mathbb{A}) \times {\sim} M_+^1(\mathbb{A})$. These continuity properties even hold w.r.t. the usual, pointwise topology. Details may be found in [23].

3 The Code Length Game, introduction

In this section $\mathcal{P}$ is a non-empty subset of $M_+^1(\mathbb{A})$, neutrally referred to as the *model*. In specific applications it may be more appropriate with other terminology, e.g. the *preparation space* or the *statistical model*. Distributions in $\mathcal{P}$ are called *consistent* distributions.

With $\mathcal{P}$ we associate a two-person zero-sum game, called the *Code Length Game* over $\mathcal{P}$. In this game, Player I chooses a consistent distribution, and Player II chooses a general code. The *cost-function*, seen from the point of view of Player II, is the map $\mathcal{P} \times {}^{\sim}K(\mathbb{A}) \to [0;\infty]$ given by the average code length:

$$(P,\kappa) \curvearrowright \langle \kappa, P \rangle .$$

This game was introduced in [20], see also [15], [21], [10], [8] and [22]. Player I may be taken to represent "the system", "Nature", "God" or $\cdots$, whereas Player II represents "the observer", "the statistician" or $\cdots$.

We can motivate the game introduced in various ways. The personification of the two participants in the game is natural as far as Player II is concerned since, in many situations, we can identify ourselves with that person. Also, the objective of Player II appears well motivated. To comment on this in more detail, we first remind the reader that we imagine that there is associated a real code consisting of binary sequences to $\kappa \in {}^{\sim}K(\mathbb{A})$ and that κ merely tells us what the code lengths of the various code words are.

We can think of a specific code in at least three different ways: as a *representation* of the letters in $\mathbb{A}$, as a means for *identification* of these letters and – the view we find most fruitful – as a strategy for making *observations* from a source generating letters from $\mathbb{A}$. The two last views are interrelated. In fact, for the strategy of observation which we have in mind, we use the code to identify the actual outcome by posing a succession of questions, starting with the question "is the first binary digit in the code word corresponding to the outcome a 1?" , then we ask for the second binary digit and so on until it is clear to us which letter is the actual outcome from the source. The number of questions asked is the number of binary digits in the corresponding code word.

The cost function can be interpretated as mean representation time, mean identification time or mean observation time and it is natural for Player II to attempt to minimize this quantity. The sense in assuming that Player I has the opposite aim, namely to maximize the cost function is more dubious. The arguments one can suggest to justify this, thereby motivating the zero-sum character of the Code Length Game, are partly natural to game theory in general, partly can be borrowed from Jaynes' reasoning behind his Maximum Entropy Principle. Without going into lengthy discussions we give some indications: Though we do not seriously imagine that Player I is a "real" person with rational behaviour, such thoughts regarding the fictive Player I reflect back on our own conceptions. With our fictitious assumptions we express our own modelling. If all we know is the model $\mathcal{P}$ and if, as is natural, all we strive for is minimization of the cost function, we cannot do better than *imagining* that Player I is a real person behaving rationally in a way which is least favourable to us. Any other assumption would, typically, lead to non-sensical results which would reveal that we actually knew more than first expressed and therefore, as a consequence, we should change the model in order better to reflect our level of knowledge.

To sum up, we have argued that the observer should be allowed freely to choose the means of observation, that codes offer an appropriate technical tool for this purpose and that the choice of a specific code should be dictated by the wish to minimize mean observation time, modelled adequately by the chosen cost function. Further, the more fictitious views regarding Player I and the behaviour of that player, really reflect on the adequacy and completeness of our modelling. If our modelling is precise, the assumptions regarding Player I are sensible and general theory of two-person zero-sum games can be expected to lead to relevant and useful results.

The overall principle we shall apply, we call the principle of *Game Theoretical Equilibrium.* It is obtained from general game theoretical considerations applied to the Code Length Game. No very rigid formulation of this principle is necessary. It simply dictates that in the study of a model, we shall investigate standard game theoretical notions such as equilibrium and optimal strategies.

According to our basic principle, Player I should consider, for each possible strategy $P \in \mathcal{P}$, the infimum of $\langle \kappa, P \rangle$ over $\kappa \in {}^{\sim}K(\mathbb{A})$. This corresponds to the optimal response of Player II to the chosen strategy. The infimum in question can easily be identified by appealing to an important identity which we shall use frequently in the following. The identity connects average code length, entropy and divergence and states that

$$\langle \kappa, P \rangle = H(P) + D(P \| Q), \tag{3.8}$$

valid for any $\kappa \in {}^{\sim}K(\mathbb{A})$ and $P \in M_+^1(\mathbb{A})$ with Q the (possibly incomplete) distribution matching κ. The identity is called the *linking identity.* As $D(P\|Q) \geq 0$ with equality if and only if $P = Q$, an immediate consequence of the linking identity is that entropy can be conceived as minimal average code length:

$$H(P) = \min_{\kappa \in {}^{\sim}K(\mathbb{A})} \langle \kappa, P \rangle \,. \tag{3.9}$$

The minimum is attained for the code adapted to P and, provided $H(P) < \infty$, for no other code.

Seen from the point of view of Player I, the optimal performance is therefore achieved by maximizing entropy. The maximum value to strive for is called the *maximum entropy value* ($H_{\max}$-value) and is given by

$$H_{\max}(\mathcal{P}) = \sup_{P \in \mathcal{P}} H(P) \,. \tag{3.10}$$

On the side of Player II – the "coding side" – we consider, analogously, for each $\kappa \in {}^{\sim}K(\mathbb{A})$ the *associated risk* given by

$$R(\kappa|\mathcal{P}) = \sup_{P\in\mathcal{P}} \langle \kappa, P\rangle \tag{3.11}$$

and then the *minimum risk value* ($R_{\min}$-value)

$$R_{\min}(\mathcal{P}) = \inf_{\kappa\in{}^{\sim}K(\mathbb{A})} R(\kappa|\mathcal{P})\,. \tag{3.12}$$

This is the value to strive for for Player II.

We have now looked at each side of the game separately. Combining the two sides, we are led to the usual concepts, well known from the theory of two-person zero-sum games. Thus, the model $\mathcal{P}$ is in *equilibrium* if $H_{\max}(\mathcal{P}) = R_{\min}(\mathcal{P}) < \infty$, and in this case, $H_{\max}(\mathcal{P}) = R_{\min}(\mathcal{P})$ is the *value* of the game. Note that as a "supinf" is bounded by the corresponding "infsup", the inequality

$$H_{\max}(\mathcal{P}) \le R_{\min}(\mathcal{P}) \tag{3.13}$$

always holds.

The concept of *optimal strategies* also follows from general considerations. For Player I, this is a consistent distribution with maximal entropy, i.e. a distribution $P \in \mathcal{P}$ with $H(P) = H_{\max}(\mathcal{P})$. And for Player II, an optimal strategy is a code $\kappa^* \in {}^{\sim}K(\mathbb{A})$ such that $R(\kappa^*|\mathcal{P}) = R_{\min}(\mathcal{P})$. Such a code is also called a *minimum risk code* ($R_{\min}$-code).

4 Cost-stable codes, partition functions and exponential families

The purpose of this section is to establish a certain sufficient condition for equilibrium and to identify the optimal strategies for each of the players in the Code Length Game. This cannot always be done but the simple result presented here already covers most applications. Furthermore, the approach leads to familiar concepts and results. This will enable the reader to judge the merits of the game theoretical method as compared to a more standard approach via the introduction of Lagrange multipliers.

As in the previous section, we consider a model $\mathcal{P} \subseteq M_+^1(\mathbb{A})$. Let $\kappa^* \in K(\mathbb{A})$ together with its matching distribution P^* be given and assume that P^* is consistent. Then we call κ^* a *Nash equilibrium code* for the model $\mathcal{P}$ if

$$\langle \kappa^*, P\rangle \le \langle \kappa^*, P^*\rangle\,;\; P \in \mathcal{P} \tag{4.14}$$

and if $H(P^*) < \infty$. The terminology is adapted from mathematical economy, cf. e.g. Aubin [2]. The requirement can be written $R(\kappa^*|\mathcal{P}) \le H(P^*) < \infty$. Note that here we insist that a Nash equilibrium code be $\mathcal{P}$-adapted. This condition will later be relaxed.

Theorem 4.1. *Let $\mathcal{P}$ be a model and assume that there exists a $\mathcal{P}$-adapted Nash equilibrium code κ^*, say, with matching distribution P^*. Then $\mathcal{P}$ is in equilibrium and both players have optimal strategies. Indeed, P^* is the unique optimal strategy for Player I and κ^* the unique optimal strategy for Player II.*

Proof. Since $R(\kappa^*|\mathcal{P}) \le H(P^*)$, $R_{\min}(\mathcal{P}) \le H_{\max}(\mathcal{P})$. As the opposite inequality always holds by (3.13), $\mathcal{P}$ is in equilibrium, the value of the Code Length Game associated with $\mathcal{P}$ is $H(P^*)$ and κ^* and P^* are optimal strategies.

To establish the uniqueness of κ^*, let κ be any code distinct from κ. Let P be the distribution matching κ. Then, by the linking identity,

$$R(\kappa|\mathcal{P}) \ge \langle \kappa, P^* \rangle = H(P^*) + D(P^* \| P) > H(P^*)\,,$$

hence κ is not optimal.

For the uniqueness proof of P^*, let P be a consistent distribution distinct from P^*. Then, again by the linking identity,

$$H(P) < H(P) + D(P\|P^*) = \langle \kappa^*, P \rangle \le H(P^*)\,,$$

and P cannot be optimal. □

As we shall see later, the existence of a Nash equilibrium code is, essentially, also necessary for the conclusion of the theorem. [1] This does not remove the difficulty of actually finding the Nash equilibrium code in concrete cases of interest. In many cases it turns out to be helpful to search for codes with stronger properties. A code κ^* is a *cost-stable* code for $\mathcal{P}$ if there exists $h < \infty$ such that $\langle \kappa^*, P \rangle = h$ for all $P \in \mathcal{P}$. Clearly, a cost-stable code with a consistent matching distribution is a Nash equilibrium code. Therefore, we obtain the following corollary from Theorem 4.1:

Corollary 4.2. *If κ^* is a cost-stable code for $\mathcal{P}$ and if the matching distribution P^* is consistent, then $\mathcal{P}$ is in equilibrium and κ^* and P^* are the unique optimal strategies pertaining to the Code Length Game.*

[1] The reader may want to note that it is in fact easy to prove directly that if $\mathcal{P}$ is convex, $H_{\max}(\mathcal{P})$ finite and P^* a consistent distribution with maximum entropy, then the adapted code κ^* must be a Nash equilibrium code. To see this, let P_0 and P_1 be distributions with finite entropy and put $P_\alpha = (1-\alpha)P_0 + \alpha P_1$. Then $h(\alpha) = H(P_\alpha)$; $0 \le \alpha \le 1$ is strictly concave and $h'(\alpha) = \langle \kappa_\alpha, P_1 \rangle - \langle \kappa_\alpha, P_0 \rangle$ with κ_α the code adapted to P_α. From this it is easy to derive the stated result. A more complete result is given in Theorem 7.3.

In order to illustrate the usefulness of this result, consider the case of a model $\mathcal{P}$ given by finitely many linear constraints, say

$$\mathcal{P} = \{P \in M_+^1(\mathbb{A}) \mid \langle E_1, P\rangle = \lambda_1, \dots, \langle E_n, P\rangle = \lambda_n\} \tag{4.15}$$

with $E_1, \dots, E_n$ real-valued functions bounded from below and $\lambda_1, \dots, \lambda_n$ real-valued constants. Let us search for cost-stable codes κ for $\mathcal{P}$. Clearly, any code of the form

$$\kappa = \alpha + \beta_1 E_1 + \cdots + \beta_n E_n = \alpha + \overline{\beta} \cdot \overline{E} \tag{4.16}$$

is cost-stable. Here, α and the β's denote constants, $\overline{\beta}$ and $\overline{E}$ vectors and a dot signifies scalar products of vectors. For κ defined by (4.16) to define a code we must require that $\kappa \geq 0$ and, more importantly, that Kraft's equality (2.4) holds. We are thus forced to assume that the *partition function* evaluated at $\overline{\beta} = (\beta_1, \dots, \beta_n)$ is finite, i.e. that

$$Z(\overline{\beta}) = \sum_{i \in \mathbb{A}} \exp(-\overline{\beta} \cdot \overline{E_i}) \tag{4.17}$$

is finite, and that $\alpha = \log Z(\overline{\beta})$. When these conditions are fulfilled, $\kappa = \kappa^{\overline{\beta}}$ defined by

$$\kappa^{\overline{\beta}} = \log Z(\overline{\beta}) + \overline{\beta} \cdot \overline{E} \tag{4.18}$$

defines a cost-stable code with individual code lengths given by

$$\kappa_i^{\overline{\beta}} = \log Z(\overline{\beta}) + \overline{\beta} \cdot \overline{E}_i\,. \tag{4.19}$$

The matching distribution $P^{\overline{\beta}}$ is given by the point probabilities

$$P_i^{\overline{\beta}} = \frac{\exp(-\overline{\beta} \cdot \overline{E}_i)}{Z(\overline{\beta})}\,. \tag{4.20}$$

In most cases where linear models occur in the applications, one will be able to adjust the parameters in $\overline{\beta}$ such that $P^{\overline{\beta}}$ is consistent. By Corollary 4.2, the entropy maximization problem will then be solved. However, not all cases can be settled in this way as there may not exist a consistent maximum entropy distribution.

We have seen that the search for cost-stable codes led us to consider the well-known partition function and also the well-known *exponential family* consisting of distributions $(P^{\overline{\beta}})$ with $\overline{\beta}$ ranging over all vectors $\overline{\beta} \in \mathbb{R}^n$ for which $Z(\overline{\beta}) < \infty$.

From our game theoretical point of view, the family of codes $(\kappa^{\overline{\beta}})$ with $Z(\overline{\beta}) < \infty$ has at least as striking features as the corresponding family of distributions. We shall therefore focus on both types of objects and shall call the family of matching pairs $(\kappa^{\overline{\beta}}, P^{\overline{\beta}})$ with $\overline{\beta}$ ranging over vectors with $Z(\overline{\beta}) < \infty$ for the *exponential family* associated with the set $\overline{E} = (E_1, \ldots, E_n)$ of functions on $\mathbb{A}$ or associated with the family of models one can define from $\overline{E}$ by choosing $\overline{\lambda} = (\lambda_1, \ldots, \lambda_n)$ and considering $\mathcal{P}$ given by (4.15).

We stress that the huge literature on exponential families displays other families of discrete distributions than those that can be derived from the above definition. In spite of this we maintain the view that an information theoretical definition in terms of codes (or related objects) is more natural than the usual structural definitions. We shall not pursue this point vigorously here as it will require the consideration of further games than the simple Code Length Game.

In order to further stress the significance of the class of cost stable codes we mention a simple continuity result:

Theorem 4.3. *If a model $\mathcal{P}$ has a cost-stable code, the entropy function H is continuous when restricted to $\mathcal{P}$.*

Proof. Assume that $\langle \kappa^*, P \rangle = h < \infty$ for all $P \in \mathcal{P}$. Then $H(P) + D(P\|P^*) = h$ for $P \in \mathcal{P}$ with P^* the distribution matching κ^*. As the sum of the two lower semi-contimuous functions in this identity is a constant function, each of the functions, in particular the entropy function, must be continuous. □

As we have already seen, the notion of cost-stable codes is especially well suited to handle models defined by linear constraints. In section 6 we shall take this up in more detail.

The following sections will be more technical and mathematically abstract. This appears necessary in order to give a comprehensive treatment of all basic aspects related to the Cost Length Game and to the Maximum Entropy Principle.

5 The Code Length Game, further preparations

In section 3 we introduced a minimum of concepts that enabled us to derive the useful results of section 4. With that behind us as motivation and background material, we are ready to embark on a more thorough investigation which will lead to a clarification of certain obscure points, especially related to the possibility that a consistent distribution with maximal entropy may not exist. In this section we point out certain results and concepts which will later be useful.

In view of our focus on codes it is natural to look upon divergence in a different way, as *redundancy*. Given is a code $\kappa \in {}^{\sim}K(\mathbb{A})$ and a distribution $P \in M_+^1(\mathbb{A})$. We imagine that we use κ to code letters from $\mathbb{A}$ generated by a "source" and that P is the "true" distribution of the letters. The optimal performance is, according to (3.9), represented by the entropy $H(P)$ whereas the actual performance is represented by the number $\langle \kappa, P \rangle$. The difference $\langle \kappa, P \rangle - H(P)$ is then taken as the *redundancy*. This is well defined if $H(P) < \infty$ and then coincides with $D(P\|Q)$ where Q denotes the distribution matching κ. As $D(P\|Q)$ is always well defined, we use this quantity for our technical definition: The *redundancy of* $\kappa \in {}^{\sim}K(\mathbb{A})$ *against* $P \in M_+^1(\mathbb{A})$ is denoted $D(P\|\kappa)$ and defined by

$$D(P\|\kappa) = D(P\|Q) \text{ with } \kappa \leftrightarrow Q\,.$$

Thus $D(P\|\kappa)$ and $D(P\|Q)$ can be used synonymously and reflect different ways of thinking. Using redundancy rather than divergence, the linking identity takes the following form:

$$\langle \kappa, P\rangle = H(P) + D(P\|\kappa)\,. \tag{5.21}$$

We shall often appeal to basic concavity and convexity properties. Clearly, the entropy function is concave as a minimum of affine functions, cf. (3.9). However, we need a more detailed result which also implies strict concavity. The desired result is the following identity

$$H(\sum_\nu \alpha_\nu P_\nu) = \sum_\nu \alpha_\nu H(P_\nu) + \sum_\nu \alpha_\nu D(P_\nu\|\overline{P})\,, \tag{5.22}$$

where $\overline{P} = \sum_\nu \alpha_\nu P_\nu$ is any finite or countably infinite convex combination of probability distributions. This follows by the linking identity.

A closely related identity involves divergence and states that, with notation as above and with Q denoting an arbitrary general distribution,

$$\sum_\nu \alpha_\nu D(P_\nu\|Q) = D\left(\sum_\nu \alpha_\nu P_\nu \|Q\right) + \sum_\nu \alpha_\nu D(P_\nu|\overline{P})\,. \tag{5.23}$$

The identity shows that divergence $D(\cdot\|Q)$ is strictly convex. A proof can be found in [23].

For the remainder of the section we consider a model $\mathcal{P} \subseteq M_+^1(\mathbb{A})$ and the associated Code Length Game.

By $\mathrm{supp}\,(\mathcal{P})$ we denote the *support* of $\mathcal{P}$, i.e. the set of $i \in \mathbb{A}$ for which there exists $P \in \mathcal{P}$ with $p_i > 0$. Thus, $\mathrm{supp}\,(\mathcal{P}) = \cup_{P\in\mathcal{P}}\mathrm{supp}\,(P)$, the union of the usual supports of all consistent distributions. Often one may restrict attention to models with *full support*, i.e. to models with $\mathrm{supp}\,(\mathcal{P}) = \mathbb{A}$. However, we shall not make this assumption unless pointed out specifically.

Recall that distributions in $\mathcal{P}$ are said to be *consistent*. Often, it is more appropriate to consider distributions in $\overline{\mathcal{P}}^{\sigma}$. These distributions are called *essentially consistent distributions*. Using these distributions we relax the requirements to a distribution with maximum entropy, previously only considered for consistent distributions. Accordingly, a distribution P^* is called a *maximum entropy* *distribution* ($H_{\max}$-distribution) if P^* is essentially consistent and $H(P^*) = H_{\max}(\mathcal{P})$. We warn the reader that the usual definition in the literature insists on the requirement of consistency. Nevertheless, we find the relaxed requirement of essential consistency more adequate. For one thing, lower semi-continuity of the entropy function implies that

$$H_{\max}(\mathcal{P}) = H_{\max}(\overline{\mathcal{P}}^{\sigma}) = H_{\max}(\overline{\mathcal{P}}^{V}) \tag{5.24}$$

and this points to the fact that the models $\overline{\mathcal{P}}^{\sigma}$ and $\overline{\mathcal{P}}^{V}$ behave in the same way as $\mathcal{P}$. This view is further supported by the observation that for any $\kappa \in {}^{\sim}K(\mathbb{A})$,

$$R(\kappa|\mathcal{P}) = R(\kappa|\overline{\mathrm{co}}^{\,V}\mathcal{P})\,. \tag{5.25}$$

This follows as the map $P \curvearrowright \langle \kappa, P\rangle$ is lower semi-continuous and affine. As a consequence,

$$R_{\min}(\mathcal{P}) = R_{\min}(\overline{\mathrm{co}}^{\,V})\,. \tag{5.26}$$

It follows that all models with $\mathcal{P} \subseteq \mathcal{P}' \subseteq \overline{\mathcal{P}}^{V}$ behave similarly as far as the Code Length Game is concerned. The reason why we do not relax further the requirement of a $H_{\max}$-distribution from $P^* \in \overline{\mathcal{P}}^{\sigma}$ to $P^* \in \overline{\mathcal{P}}^{V}$ is firstly, that we hold the information topology for more relevant for our investigations than the usual topology. Secondly, we shall see that the property $P^* \in \overline{\mathcal{P}}^{\sigma}$ which is stronger than $P^* \in \overline{\mathcal{P}}^{V}$ can in fact be verified in the situations we have in mind (see Theorem 6.2).

The fact that a consistent $H_{\max}$-distribution may not exist leads to further important notions. Firstly, a sequence $(P_n)_{n\geq 1}$ of distributions is said to be *asymptotically optimal* if all the P_n are consistent and if $H(P_n) \to H_{\max}(\mathcal{P})$ for $n \to \infty$. And, secondly, a distribution P^* is the *maximum entropy attractor* ($H_{\max}$-attractor) if P^* is essentially consistent and if $P_n \xrightarrow{D} P^*$ for every asymptotically optimal sequence $(P_n)_{n\geq 1}$.

As an example, consider the (uninteresting!) model of all deterministic distributions. For this model, the $H_{\max}$-attractor does not exist and there is no unique $H_{\max}$-distribution. For more sensible models, the $H_{\max}$-attractor P^* will exist, but it may not be the $H_{\max}$-distribution as lower semi-continuity only quarantees the inequality $H(P^*) \leq H_{\max}(\mathcal{P})$, not the corresponding equality.

Having by now refined the concepts related to the distribution side of the Code Length Game, we turn to the coding side.

It turns out that we need a localized variant of the risk associated with certain codes. The codes we shall consider are, intuitively, all codes which the observer (Player II) off-hand finds it worth while to consider. If $P \in \mathcal{P}$ is the "true" distribution, and the observer knows this, he will choose the code adapted to P in order to minimize the average code length. As nature (Player

I) could from time to time change the choice of $P \in \mathcal{P}$, essentially any strategy in the closure of $\mathcal{P}$ could be approached. With these remarks in mind we find it natural for the observer only to consider $\overline{\mathcal{P}}^{D_\sigma}$-adapted codes in the search for reasonable strategies.

Assume now that the observer decides to choose a $\overline{\mathcal{P}}^{D_\sigma}$-adapted code κ. Let $P \in \overline{\mathcal{P}}^{D_\sigma}$ be the distribution which matches κ. Imagine that the choice of κ is dictated by a strong belief that the true distribution is P or some distribution very close to P (in the information topology!). Then the observer can evaluate the associated risk by calculating the *localized risk* associated with κ which is defined by the equation:

$$R_{\mathrm{loc}}(\kappa|\mathcal{P}) = \sup_{(P_n)\subseteq\mathcal{P}, P_n \overset{D}{\to} P} \ \limsup_{n\to\infty} \langle \kappa, P_n \rangle \,, \tag{5.27}$$

where the supremum is over the class of all sequences of consistent distributions which converge in the information topology to P. Note that we insist on a definition which operates with sequences.

Clearly, the normal "global" risk must be at least as large as localized risk, therefore, for any $\overline{\mathcal{P}}^{D_\sigma}$-adapted code,

$$R_{\mathrm{loc}}(\kappa|\mathcal{P}) \le R(\kappa|\mathcal{P}) \,. \tag{5.28}$$

A further and quite important inequality is the following:

$$R_{\mathrm{loc}}(\kappa|\mathcal{P}) \le H_{\max}(\mathcal{P}) \,. \tag{5.29}$$

This inequality is easily derived from the defining relation (5.27) by writing $\langle \kappa, P_n \rangle$ in the form $H(P_n) + D(P_n \| P)$ with $\kappa \leftrightarrow P$, noting also that $P_n \overset{D}{\to} P$ implies that $D(P_n\|P) \to 0$. We note that had we allowed nets in the defining relation (5.27), a different and sometimes strictly larger quantity would result and (5.29) would not necessarily hold.

As the last preparatory result, we establish pretty obvious properties of an eventual optimal strategy for the observer, i.e. of an eventual $R_{\min}$-code.

Lemma 5.1. *Let $\mathcal{P} \subseteq M_+^1(\mathbb{A})$ with $R_{\min}(\mathcal{P}) < \infty$ be given. Then the $R_{\min}$-code is unique and if it exists, say $R(\kappa^*|\mathcal{P}) = R_{\min}(\mathcal{P})$, then κ^* is compact with* $\mathrm{supp}\,(\kappa^*) = \mathrm{supp}\,(\mathcal{P})$.

Proof. Assume that $\kappa^* \in {}^{\sim}K(\mathbb{A})$ is a $R_{\min}$-code. As $R(\kappa^*|\mathcal{P}) < \infty$, $\mathrm{supp}\,(\mathcal{P}) \subseteq \mathrm{supp}\,(\kappa^*)$. Then consider an $a_0 \in \mathrm{supp}\,(\kappa^*)$ and assume, for the purpose of an indirect proof, that $a_0 \in \mathbb{A}\backslash\mathrm{supp}\,(\mathcal{P})$. Then the code κ obtained from κ^* by putting $\kappa(a_0) = \infty$ and keeping all other values fixed, is a general non-compact code which is not identically $+\infty$. Therefore, there exists $\varepsilon > 0$ such that $\kappa - \varepsilon$ is a compact code. For any $P \in \mathcal{P}$, we use the fact that $a_0 \notin \mathrm{supp}\,(P))$ to conclude that $\langle \kappa - \varepsilon, P \rangle = \langle \kappa^* - \varepsilon, P \rangle$, hence $R(\kappa - \varepsilon|\mathcal{P}) = R(\kappa^*|\mathcal{P}) - \varepsilon$, contradicting the minimality property of κ^*. Thus we conclude that $\operatorname{supp}(\kappa^*) = \operatorname{supp}(\mathcal{P})$. Similarly, it is clear that κ^* must be compact – since otherwise, $\kappa^* - \varepsilon$ would be more efficient than κ^* for some $\varepsilon > 0$.

In order to prove uniqueness, assume that both κ_1 and κ_2 are $R_{\min}$-codes for $\mathcal{P}$. If we assume that $\kappa_1 \neq \kappa_2$, then $\kappa_1(a) \neq \kappa_2(a)$ holds for some a in the common support of the codes κ_1 and κ_2 and then, by the geometric/arithmetic inequality, we see that $\frac{1}{2}(\kappa_1 + \kappa_2)$ is a general non-compact code. For some $\varepsilon > 0$, $\frac{1}{2}(\kappa_1 + \kappa_2) - \varepsilon$ will then also be a code and as this code is seen to be more efficient than κ_1 and κ_2, we have arrived at a contradiction. Thus $\kappa_1 = \kappa_2$, proving the uniqueness assertion. □

6 Models in equilibrium

Let $\mathcal{P} \subseteq M_+^1(\mathbb{A})$. By definition, the requirement of equilibrium is one which involves the relationship between both sides of the Code Length Game. The main result of this section shows that the requirement can be expressed in terms involving only one of the sides of the game, either distributions or codes.

Theorem 6.1 (conditions for equilibrium). *Let $\mathcal{P} \subseteq M_+^1(\mathbb{A})$ be a model and assume that $H_{\max}(\mathcal{P}) < \infty$. Then the following conditions are equivalent:*

(i) $\mathcal{P}$ is in equilibrium,

(ii) $H_{\max}(\operatorname{co}\mathcal{P}) = H_{\max}(\mathcal{P})$,

(iii) there exists a $\overline{\mathcal{P}}^D$-adapted code κ^* such that

$$R(\kappa^*|\mathcal{P}) = R_{\mathrm{loc}}(\kappa^*|\mathcal{P})\,.$$

Proof. (i) ⇒ (iii): Here we assume that $H_{\max}(\mathcal{P}) = R_{\min}(\mathcal{P})$. In particular, $R_{\min}(\mathcal{P}) < \infty$. As the map $\kappa \curvearrowright R(\kappa|\mathcal{P})$ is lower semi-continuous on ${}^\sim K(\mathbb{A})$ (as the supremum of the maps $\kappa \curvearrowright \langle \kappa, P\rangle$; $P \in \mathcal{P}$), and as ${}^\sim K(\mathbb{A})$ is compact, the minimum of $\kappa \curvearrowright R(\kappa|\mathcal{P})$ is attained. Thus, there exists $\kappa^* \in {}^\sim K(\mathbb{A})$ such that $R(\kappa^*|\mathcal{P}) = R_{\min}(\mathcal{P})$. As observed in Lemma 5.1, κ^* is a compact code and κ^* is the unique $R_{\min}$-code.

For $P \in \mathcal{P}$,

$$H(P) + D(P\|\kappa^*) = \langle \kappa^*, P\rangle \leq R_{\min}(\mathcal{P}) = H_{\max}(\mathcal{P})\,.$$

It follows that $D(P_n\|\kappa^*) \to 0$ for any asymptotically optimal sequence $(P_n)_{n\geq 1}$. In other words, the distribution $P^* \in M_+^1(\mathbb{A})$ which matches κ^* is the $H_{\max}$- attractor of the model.

We can now consider any asymptotically optimal sequence $(P_n)_{n\geq 1}$ in order to conclude that

$$R_{\mathrm{loc}}(\kappa^*|\mathcal{P}) \geq \limsup_{n\to\infty}\langle\kappa^*, P_n\rangle = \limsup_{n\to\infty}(H(P_n) + D(P_n\|\kappa^*))$$
$$= H_{\max}(\mathcal{P}) \geq R_{\min}(\mathcal{P}) = R(\kappa^*|\mathcal{P})\,.$$

By (5.28), the assertion of (iii) follows.

(iii) $\Rightarrow$ (ii): Assuming that (iii) holds, we find from (5.25), (3.13) and (5.29) that

$$H_{\max}(\mathrm{co}\,\mathcal{P}) \leq R_{\min}(\mathrm{co}\,\mathcal{P}) = R_{\min}(\mathcal{P}) \leq R(\kappa^*|\mathcal{P})$$
$$\leq R_{\mathrm{loc}}(\kappa^*|\mathcal{P}) \leq H_{\max}(\mathcal{P})$$

and the equality of (ii) must hold.

(ii) $\Rightarrow$ (i): For this part of the proof we fix a specific asymptotically optimal sequence $(P_n)_{n\geq 1} \subseteq \mathcal{P}$. We assume that (ii) holds. For each n and m we observe that by (5.22) and (2.7), with $M = \frac{1}{2}(P_n + P_m)$,

$$H_{\max}(\mathcal{P}) = H_{\max}(co\,\mathcal{P}) \geq H(\frac{1}{2}(P_n + P_m))$$
$$= \frac{1}{2}H(P_n) + \frac{1}{2}H(P_m) + \frac{1}{2}D(P_n\|M) + \frac{1}{2}D(P_m\|M)$$
$$\geq \frac{1}{2}H(P_n) + \frac{1}{2}H(P_m) + \frac{1}{8}V(P_n, P_m)^2\,.$$

It follows that $(P_n)_{n\geq 1}$ is a Cauchy sequence with respect to total variation, hence there exists $P^* \in M_+^1(\mathbb{A})$ such that $P_n \overset{V}{\to} P^*$.

Let κ^* be the code adapted to P^*. In order to evaluate $R(\kappa^*|\mathcal{P})$ we consider any $P \in \mathcal{P}$. For a suitable sequence $(\varepsilon_n)_{n\geq 1}$ of positive numbers converging to zero, we consider the sequence $(Q_n)_{n\geq 1} \subseteq \mathcal{P}$ given by

$$Q_n = (1-\varepsilon_n)P_n + \varepsilon_n P\,;\ n \geq 1\,.$$

By (5.22) we find that

$$H_{\max}(\mathcal{P}) = H_{\max}(co\,\mathcal{P}) \geq H(Q_n) \geq (1-\varepsilon_n)H(P_n) + \varepsilon_n H(P) + \varepsilon_n D(P\|Q_n)\,,$$

hence

$$H(P) + D(P\|Q_n) \leq H(P_n) + \frac{1}{\varepsilon_n}\left(H_{\max}(\mathcal{P}) - H(P_n)\right)\,.$$

As $Q \curvearrowright D(P\|Q)$ is lower semi-continuous, we conclude from this that

$$H(P) + D(P\|P^*) \leq H_{\max}(\mathcal{P}) + \liminf_{n\to\infty}\frac{1}{\varepsilon_n}\left(H_{\max}(\mathcal{P}) - H(P_n)\right)\,.$$

Choosing the ε_n's appropriately, e.g. $\varepsilon_n = (H_{\max}(\mathcal{P}) - H(P_n))^{\frac{1}{2}}$, it follows that $H(P) + D(P\|P^*) \le H_{\max}(\mathcal{P})$, i.e. that $\langle \kappa^*, P\rangle \le H_{\max}(\mathcal{P})$. As this holds for all $P \in \mathcal{P}$, $R(\kappa^*|\mathcal{P}) \le H_{\max}(\mathcal{P})$ follows. Thus $R_{\min}(\mathcal{P}) \le H_{\max}(\mathcal{P})$, hence equality must hold here, and we have proved that $\mathcal{P}$ is in equilibrium, as desired. □

It is now easy to derive the basic properties which hold for a system in equilibrium.

Theorem 6.2 (models in equilibrium). *Assume that $\mathcal{P} \subseteq M_+^1(\mathbb{A})$ is a model in equilibrium. Then the following properties hold:*

(i) There exists a unique $H_{\max}$-attractor and for this distribution, say P^*, the inequality

$$R_{\min}(\mathcal{P}) + D(P^*\|\kappa) \le R(\kappa|\mathcal{P}) \tag{6.30}$$

holds for all $\kappa \in {}^{\sim}K(\mathbb{A})$.

(ii) There exists a unique $R_{\min}$-code and for this code, say κ^*, the inequality

$$H(P) + D(P\|\kappa^*) \le H_{\max}(\mathcal{P}) \tag{6.31}$$

holds for every $P \in \mathcal{P}$, even for every $P \in \overline{\mathrm{co}}^V\mathcal{P}$. The $R_{\min}$-code is compact.

Proof. The existence of the $R_{\min}$-code was established by the compactness argument in the beginning of the proof of Theorem 6.1. The inequality (6.31) for $P \in \mathcal{P}$ is nothing but an equivalent form of the inequality $R(\kappa^*|\mathcal{P}) \le H_{\max}(\mathcal{P})$ and this inequality immediately implies that the distribution P^* matching κ^* is the $H_{\max}$-attractor. The extension of the validity of (6.31) to $P \in \overline{\mathrm{co}}^V\mathcal{P}$ follows from (5.26).

To prove (6.30), let $(P_n)_{n\ge 1} \subseteq \mathcal{P}$ be asymptotically optimal. Then

$$\begin{aligned} R(\kappa|\mathcal{P}) &\ge \limsup_{n\to\infty} \langle \kappa, P_n\rangle \\ &= \limsup_{n\to\infty} (H(P_n) + D(P_n\|\kappa)) \\ &\ge H_{\max}(\mathcal{P}) + \liminf_{n\to\infty} D(P_n\|\kappa) \\ &\ge H_{\max}(\mathcal{P}) + D(P^*\|\kappa)\,, \end{aligned}$$

which is the desired conclusion as $H_{\max}(\mathcal{P}) = R_{\min}(\mathcal{P})$. □

If $\mathcal{P}$ is a model in equilibrium, we refer to the pair (κ^*, P^*) from Theorem 6.2 as the *optimal matching pair pair.* Thus κ^* denotes the $R_{\min}$-code and P^* the $H_{\max}$-attractor.

Combining Theorem 6.2 and Theorem 6.1 we realize that, unless $R(\kappa|\mathcal{P}) = \infty$ for every $\kappa \in {}^{\sim}K(\mathbb{A})$, there exists a unique $R_{\min}$-code. The matching distribution is the $H_{\max}$-attractor for the model co $(\mathcal{P})$. We also note that simple examples (even with $\mathbb{A}$ a two-element set) show that $\mathcal{P}$ may have a $H_{\max}$-attractor without $\mathcal{P}$ being in equilibrium and this attractor may be far away from the $H_{\max}$-attractor for co $(\mathcal{P})$.

Corollary 6.3. *For a model $\mathcal{P} \subseteq M_+^1(\mathbb{A})$ in equilibrium, the $R_{\min}$-code and the $H_{\max}$-attractor form a matching pair: $\kappa^* \leftrightarrow P^*$, and for any matching pair (κ, P) with $P \in \overline{\text{co}}^V \mathcal{P}$,*

$$V(P, P^*) \leq (R(\kappa|\mathcal{P}) - H(P))^{\frac{1}{2}}.$$

Proof. Combining (6.30) with (6.31) it folows that for $\kappa \leftrightarrow P$ with $P \in \overline{\text{co}}^V \mathcal{P}$,

$$D(P\|P^*) + D(P^*\|P) \leq R(\kappa|\mathcal{P}) - H(P)$$

and the result follows from Pinskers inequality, (2.7). □

Corollary 6.3 may help us to judge the approximate position of the $H_{\max}$-attractor P^* even without knowing the value of $H_{\max}(\mathcal{P})$. Note also that the proof gave the more precise bound $J(P, P^*) \leq R(\kappa|\mathcal{P}) - H(P)$ with assumptions as in the theorem and with $J(\cdot\,,\,\cdot)$ denoting *Jeffrey's measure of discrimination*, cf. [3] or [16].

Corollary 6.4. *Assume that the model $\mathcal{P}$ has a cost-stable code κ^* and let P^* be the matching distribution. Then $\mathcal{P}$ is in equilibrium and has (κ^*, P^*) as optimal matching pair if and only if P^* is essentially consistent.*

Proof. By definition, an $H_{\max}$-attractor is essentially consistent. Therefore, the necessity of the condition $P^* \in \overline{\mathcal{P}}^{D_\sigma}$ is trivial. For the proof of sufficiency, assume that $\langle \kappa^*, P \rangle = h$ for all $P \in \mathcal{P}$ with h a finite constant. Clearly then, $H_{\max}(\mathcal{P}) \leq h$. Now, let $(P_n)_{n\geq 1}$ be a sequence of consistent distributions with $P_n \xrightarrow{D} P$. By the linking identity, $H(P_n) + D(P_n\|P^*) = h$ for all n, and we see that $H_{\max}(\mathcal{P}) \geq h$. Thus $H_{\max}(\mathcal{P}) = h$ and (P_n) is asymptotically optimal. By Theorem 6.2, the sequence converges in the information topology to the $H_{\max}$-attractor which must then be P^*. The result follows. □

Note that this result is a natural further development of Corollary 4.2.

Corollary 6.5. *Assume that $\mathcal{P} \subseteq M_+^1(\mathbb{A})$ is a model in equilibrium. Then all models $\mathcal{P}'$ with $\mathcal{P} \subseteq \mathcal{P}' \subseteq \overline{\text{co}}^V \mathcal{P}$ are in equilibrium too and they all have the same optimal matching pair.*

Proof. If $\mathcal{P} \subseteq \mathcal{P}' \subseteq \overline{\text{co}}^V \mathcal{P}$ then

$$H_{\max}(\text{co }\mathcal{P}') \leq H_{\max}(\overline{\text{co}}^V \mathcal{P}) = H_{\max}(\text{co }\mathcal{P}) = H_{\max}(\mathcal{P}) \leq H_{\max}(\mathcal{P}')$$

and we see that $\mathcal{P}'$ is in equilibrium. As an asymptotically optimal sequence for $\mathcal{P}$ is also asymptotically optimal for $\mathcal{P}'$, it follows that $\mathcal{P}'$ has the same $H_{\max}$-attractor, hence also the same optimal matching pair, as $\mathcal{P}$. □

Another corollary is the following result which can be used as a basis for proving certain limit theorems, cf. [22].

Corollary 6.6. *Let $(\mathbb{A}, \mathcal{P}_n)_{n\geq 1}$ be a sequence of models and assume that they are all in equilibrium with $\sup_{n\geq 1} H_{\max}(\mathcal{P}_n) < \infty$ and that they are nested in the sense that $co(\mathcal{P}_1) \subseteq co(\mathcal{P}_2) \subseteq \cdots$. Let there further be given a model $\mathcal{P}$ such that*

$$\bigcup_{n\geq 1} \mathcal{P}_n \subseteq \mathcal{P} \subseteq \overline{\text{co}}^V(\bigcup_{n\geq 1} \mathcal{P}_n).$$

Then $\mathcal{P}$ is in equilibrium too, and the sequence of $H_{\max}$-attractors of the $\mathcal{P}_n$'s converges in divergence to the $H_{\max}$-attractor of $\mathcal{P}$.

Clearly, the corollaries are related and we leave it to the reader to extend the argument in the proof of Corollary 6.5 so that it also covers the case of Corollary 6.6.

We end this section by developing some results on models given by linear conditions, thereby continuing the preliminary results from sections 1 and 4. We start with a general result which uses the following notion: A distribution P^* is *algebraically inner* in the model $\mathcal{P}$ if, for every $P \in \mathcal{P}$ there exists $Q \in \mathcal{P}$ such that P^* is a convex combination of P and Q.

Lemma 6.7. *If the model $\mathcal{P}$ is in equilibrium and has a $H_{\max}$-distribution P^* which is algebraically inner in $\mathcal{P}$, then P^* is cost-stable.*

Proof. Let κ^* be the code adapted to P^*. To any $P \in \mathcal{P}$ we determine $Q \in \mathcal{P}$ such that P^* is a convex combination of these two distributions. Then, as $\langle \kappa^*, P\rangle \leq H_{\max}(\mathcal{P})$ and $\langle \kappa^*, Q\rangle \leq H_{\max}(\mathcal{P})$ and as a convex combination gives $\langle \kappa^*, P^*\rangle \leq H_{\max}(\mathcal{P})$ we must conclude that $\langle \kappa^*, P\rangle = \langle \kappa^*, Q\rangle$ since $\langle \kappa^*, P^*\rangle$ is in fact equal to $H_{\max}(\mathcal{P})$. Therefore, κ^* is cost-stable. ☐

Theorem 6.8. *If the alphabet $\mathbb{A}$ is finite and the model $\mathcal{P}$ affine, then the model is in equilibrium and the $R_{\min}$-code is cost-stable.*

Proof. We may assume that $\mathcal{P}$ is closed. By Theorem 6.1, the model is in equilibrium and by continuity of the entropy function, the $H_{\max}$-attractor is a $H_{\max}$-distribution. For the $R_{\min}$-code κ^*, $\text{supp}\,(P^*) = \text{supp}\,(\mathcal{P})$ by Lemma 5.1. As $\mathbb{A}$ is finite we can then conclude that P^* is algebraically inner and Lemma 6.7 applies. ☐

We can now prove the following result:

Theorem 6.9. *Let $\mathcal{P}$ be a non-empty model given by finitely many linear constraints as in* (4.15)*:*

$$\mathcal{P} = \{P \in M_+^1(\mathbb{A}) \mid \langle E_1, P\rangle = \lambda_1, \ldots, \langle E_n, P\rangle = \lambda_n\}.$$

Assume that the functions $E_1, \cdots, E_n, 1$ are linearly independent and that $H_{\max}(\mathcal{P}) < \infty$. Then the model is in equilibrium and the optimal matching pair (κ^*, P^*) belongs to the exponential family defined by (4.18) *and* (4.20). *In particular, κ^* is cost-stable.*

Proof. The model is in equilibrium by Theorem 6.1. Let (κ^*, P^*) be the corresponding optimal matching pair. If $\mathbb{A}$ is finite the result follows by Theorem 6.8 and some standard linear algebra.

Assume now that $\mathbb{A}$ is infinite. Choose an asymptotically optimal sequence $(P_n)_{n\geq 1}$. Let $\mathbb{A}_0$ be a finite subset of $\mathbb{A}$, chosen sufficiently large (see below), and denote by $\mathcal{P}_n$ the convex model of all $P \in \mathcal{P}$ for which $p_i = P_{n,i}$ for all $i \in \mathbb{A}\backslash\mathbb{A}_0$. Let P_n^* be the $H_{\max}$-attractor for $\mathcal{P}_n$ and κ_n^* the adapted code. Then this code is cost-stable for $\mathcal{P}_n$ and of the form

$$\kappa_{n,i}^* = \alpha_n + \sum_{\nu=1}^{n} \beta_{n,\nu} \cdot E_\nu(i)\,;\; i \in \mathbb{A}_0\,.$$

If the set $\mathbb{A}_0$ is sufficiently large, the constants appearing here are uniquely determined. We find that $(P_n^*)_{n\geq 1}$ is asymptotically optimal for $\mathcal{P}$, and therefore, $P_n^* \xrightarrow{D} P^*$. It follows that the constants $\beta_{n,\nu}$ and α_n converge to some constants β_ν and α and that

$$\kappa_i^* = \alpha + \sum_{\nu=1}^{n} \beta_\nu \cdot E_\nu(i)\,;\; i \in \mathbb{A}_0$$

As $\mathbb{A}_0$ can be chosen arbitrarily large, the constants α and β_ν must be independent of $\mathbb{A}_0$ with $i \in \mathbb{A}_0$ and the above equation must hold for all $i \in \mathbb{A}$. □

Remark. Extensions of the result just proved may well be possible, but care has to be taken. For instance, if we consider models obtained by infinitely many linear constraints, the result does not hold. As a simple instance of this, the reader may consider the case where the model is a "line", viz. the affine hull generated by the two distributions P, Q on $\mathbb{A} = \mathbb{N}$ given by $p_i = 2^{-i}$; $i \geq 1$ and $q_i = (\zeta(3) \cdot i^3)^{-1}$; $i \geq 1$. This model is in equilibrium with P as $H_{\max}$-distribution, but the adapted code is not cost-stable. These facts can be established quite easily via the results quoted in the footnote following the proof of Theorem 4.1.

7 Entropy-continuous models

In the sequel we shall only discuss models in equilibrium. Such models can be quite different regarding the behaviour of the entropy function near the maximum. We start with a simple observation.

Lemma 7.1. *If $\mathcal{P}$ is in equilibrium and the $H_{\max}$-value $H_{\max}(\mathcal{P})$ is attained on $\overline{\mathcal{P}}^V$, it is only attained for the $H_{\max}$-attractor.*

Proof. Assume that $P \in \overline{\mathcal{P}}^V$ and that $H(P) = H_{\max}(\mathcal{P})$. Choose a sequence $(P_n)_{n\geq 1} \subseteq \mathcal{P}$ which converges to P in total variation. By lower semi-continuity and as $H(P) = H_{\max}(\mathcal{P})$ we see that (P_n) is asymptotically optimal. Therefore, for the $H_{\max}$-attractor P^*, $P_n \xrightarrow{D} P^*$, hence also $P_n \xrightarrow{V} P^*$. It follows that $P = P^*$. □

Lemma 7.2. *For a model $\mathcal{P}$ in equilibrium and with $H_{\max}$-attractor P^* the following conditions are equivalent:*

(i) $H : \overline{\mathcal{P}}^V \to \mathbb{R}_+$ *is continuous at P^* in the topology of total variation,*

(ii) $H : \overline{\mathcal{P}}^\sigma \to \mathbb{R}_+$ *is sequentially continuous at P^* in the information topology,*

(iii) $H(P^*) = H_{\max}(\mathcal{P})$.

Proof. Clearly, (i) implies (ii).

Assume that (ii) holds and let $(P_n)_{n\geq 1} \subseteq \mathcal{P}$ be asymptotically optimal. Then $P_n \xrightarrow{D} P^*$. By assumption, $H(P_n) \to H(P^*)$ and (iii) follows since $H(P_n) \to H_{\max}(\mathcal{P})$ also holds.

Finally, assume that (iii) holds and let $(P_n)_{n\geq 1} \subseteq \overline{\mathcal{P}}^V$ satisfy $P_n \xrightarrow{V} P^*$. By lower semi-continuity,

$$H_{\max}(\mathcal{P}) = H(P^*) \leq \liminf_{n\to\infty} H(P_n) \leq \limsup_{n\to\infty} H(P_n) \leq H_{\max}(\overline{\mathcal{P}}^V) = H_{\max}(\mathcal{P})$$

and $H(P_n) \to H(P^*)$ follows. Thus (i) holds. □

A model $\mathcal{P}$ in equilibrium is *entropy-continuous* if $H(P^*) = H_{\max}(\mathcal{P})$ with P^* the $H_{\max}$-attractor. In the opposite case we say that there is an *entropy loss.*

We now discuss entropy-continuous models. As we shall see, the previously introduced notion of Nash equilibrium code, cf. Section 4, is of central importance in this connection. We need this concept for any $\overline{\mathcal{P}}^{D_\sigma}$-adapted code. Thus, by definition, a code κ^* is a *Nash equilibrium code* if κ^* is $\overline{\mathcal{P}}^{D_\sigma}$-adapted and if

$$R(\kappa^*|\mathcal{P}) \leq H(P^*) < \infty\,. \tag{7.32}$$

We stress that the definition is used for any model $\mathcal{P}$ (whether or not it is known beforehand that the model is in equilibrium). We shall see below that a Nash equilibrium code is unique.

Theorem 7.3 (entropy-continuous models). *Let $\mathcal{P} \subseteq M_+^1(\mathbb{A})$ be a model. The following conditions are equivalent:*

(i) $\mathcal{P}$ *is in equilibrium and entropy-continuous,*

(ii) $\mathcal{P}$ is in equilibrium and has a maximum entropy distribution,

(iii) $\mathcal{P}$ has a Nash equilibrium code.

If these conditions are fulfilled, the $H_{\max}$-distribution is unique and coincides with the $H_{\max}$-attractor. Likewise, the Nash equilibrium code is unique and it coincides with the $R_{\min}$-code.

Proof. (i) $\Rightarrow$ (ii): This is clear since, assuming that (i) holds, the $H_{\max}$-attractor must be a $H_{\max}$-distribution.

(ii) $\Rightarrow$ (iii): Assume that $\mathcal{P}$ is in equilibrium and that $P_0 \in \overline{\mathcal{P}}^{D_\sigma}$ is a $H_{\max}$-distribution. Let (κ^*, P^*) be the optimal matching pair pair. Applying Theorem 6.2, (6.31) with $P = P_0$, we conclude that $D(P_0\|P^*) = 0$, hence $P_0 = P^*$. Then we find that

$$R(\kappa^*|\mathcal{P}) = R_{\min}(\mathcal{P}) = H_{\max}(\mathcal{P}) = H(P_0) = H(P^*)$$

and we see that κ^* is a Nash equilibrium code.

(iii) $\Rightarrow$ (i): If κ^* is a Nash equilibrium code for $\mathcal{P}$, then

$$R_{\min}(\mathcal{P}) \leq R(\kappa^*|\mathcal{P}) \leq H(P^*) \leq H_{\max}(\mathcal{P}) \leq R_{\min}(\mathcal{P})$$

and we conclude that $\mathcal{P}$ is in equilibrium and that κ^* is the minimum risk code.

In establishing the equivalence of (i)–(iii) we also established the uniqueness assertions claimed. □

The theorem generalizes the previous result, Theorem 4.1. We refer to section 4 for results which point to the great applicability of results like Theorem 7.3.

8 Loss of entropy

We shall study a model $\mathcal{P}$ in equilibrium. By previous results we realize that for many purposes we may assume that $\mathcal{P}$ is a closed, convex subset of $M_+^1(\mathbb{A})$ with $H_{\max}(\mathcal{P}) < \infty$. Henceforth, these assumptions are in force.

Denote by (κ^*, P^*) the optimal matching pair associated with $\mathcal{P}$. By the *disection* of $\mathcal{P}$ we understand the decomposition of $\mathcal{P}$ consisting of all non-empty sets of the form

$$\mathcal{P}_x = \{P \in \mathcal{P} \mid \langle \kappa^*, P\rangle = x\}\,. \tag{8.33}$$

Let Δ denote the set of $x \in \mathbb{R}$ with $\mathcal{P}_x \neq \varnothing$. As $R(\kappa^*|\mathcal{P}) = R_{\min}(\mathcal{P}) = H_{\max}(\mathcal{P})$, and as $\mathcal{P}$ is convex with $P^* \in \mathcal{P}$, Δ is a subinterval of $[0; H_{\max}(\mathcal{P})]$ which contains the interval $[H(P^*), H_{\max}(\mathcal{P})[$.

Clearly, κ^* is a cost-stable code for all models $\mathcal{P}_x; x \in \Delta$. Hence, by Theorem 4.3 the entropy function is continuous on each of the sets $\mathcal{P}_x; x \in \Delta$.

Each set $\mathcal{P}_x$; $x \in \Delta$ is a sub-model of $\mathcal{P}$ and as each $\mathcal{P}_x$ is convex with $H_{\max}(\mathcal{P}_x) < \infty$, these sub-models are all in equilibrium. The linking identity shows that for all $P \in \mathcal{P}_x$,

$$H(P) + D(P\|P^*) = x\,. \tag{8.34}$$

This implies that $H_{\max}(\mathcal{P}_x) \le x$, a sharpening of the trivial inequality $H_{\max}(\mathcal{P}_x) \le H_{\max}(\mathcal{P})$. From (8.34) it also follows that maximizing entropy $H(\cdot)$ over $\mathcal{P}_x$ amounts to the same thing as minimizing divergence $D(\cdot\|P^*)$ over $\mathcal{P}_x$. In other words, the $H_{\max}$-attractor of $\mathcal{P}_x$ may, alternatively, be characterized as the *I-projection* of P^* on $\mathcal{P}_x$, i.e. as the unique distribution P_x for which $Q_n \xrightarrow{D} P_x$ for every sequence $(Q_n) \subseteq \mathcal{P}_x$ for which $D(Q_n\|P^*)$ converges to the infimum of $D(Q\|P^*)$ with $Q \in \mathcal{P}_x$.*

Further basic results are collected below:

Theorem 8.1 (disection of models). *Let $\mathcal{P}$ be a convex model in equilibrium with optimal matching pair (κ^*, P^*) and assume that $P^* \in \mathcal{P}$. Then the following properties hold for the disection $(\mathcal{P}_x)_{x\in\Delta}$ defined by (8.33):*

(i) The set Δ is an interval with $\sup\Delta = H_{\max}(\mathcal{P})$. *A necessary and sufficient condition that $H_{\max}(\mathcal{P}) \in \Delta$ is that $\mathcal{P}$ is entropy-continuous. If $\mathcal{P}$ has entropy loss, Δ contains the non-degenerate interval* $[H(P^*), H_{\max}(\mathcal{P})[$.

(ii) The entropy function is continuous on each sub-model $\mathcal{P}_x$; $x \in \Delta$,

(iii) Each sub-model $\mathcal{P}_x$; $x \in \Delta$ is in equilibrium and the $H_{\max}$-attractor for $\mathcal{P}_x$ is the I-projection of P^* on $\mathcal{P}_x$,

(iv) For $x \in \Delta$, $H_{\max}(\mathcal{P}_x) \le x$ and the following bi-implications hold, where P^*_x denotes the $H_{\max}$-attractor of $\mathcal{P}_x$:

$$H_{\max}(\mathcal{P}_x) = x \iff P^*_x = P^* \iff x \ge H(P^*)\,. \tag{8.35}$$

Proof. (i)–(ii) as well as the inequality $H_{\max}(\mathcal{P}_x) \le x$ of (iii) were proved above.

For the proof of (iv) we consider an $x \in \Delta$ and let $(P_n) \subseteq \mathcal{P}_x$ be an asymptotically optimal sequence for $\mathcal{P}_x$. Then the condition $H_{\max}(\mathcal{P}_x) = x$ is equivalent with the condition $H(P_n) \to x$,

*Terminology is close to that adopted by Csiszár, cf. [4], [5], who first developed the concept for closed models. This was later extended, using a different terminology, in Topsøe [20]. In this paper we refrain from a closer study of I-projections and refer the reader to sources just cited.

and the condition $P^*_x = P^*$ is equivalent with the condition $P_n \xrightarrow{D} P^*$. In view of the equality $x = H(P_n) + D(P_n \| P^*)$ we now realize that the first bi-implication of (8.35) holds. For the second bi-implication we first remark that as $x \geq H(P^*_x)$ holds generally, if $P^*_x = P^*$ then $x \geq H(P^*)$ must hold.

For the final part of the proof of (iv), we assume that $x \geq H(P^*)$. The equality $H_{\max}(\mathcal{P}_x) = x$ is evident if $x = H(P^*)$. We may therefore assume that $H(P^*) < x < H_{\max}(\mathcal{P})$. We now let (P_n) denote an asymptotically optimal sequence for the full model $\mathcal{P}$ such that $H(P_n) \geq x$; $n \geq 1$. As $\langle \kappa^*, P^* \rangle \leq x \leq \langle \kappa^*, P_n \rangle$ for all n, we can find a sequence $(Q_n)_{n \geq 1}$ of distributions in $\mathcal{P}_x$ such that each Q_n is a convex combination of the form $Q_n = \alpha_n P^* + \beta_n P_n$. By (5.23), $D(Q_n \| P^*) \leq \beta_n D(P_n \| P^*) \to 0$. Thus P^* is essentially consistent for $\mathcal{P}_x$ and as the code adapted to P^* is cost-stable for this model, Corollary 6.4 implies that the model has P^* as its $H_{\max}$-attractor. □

A distribution P^* is said to have *potential entropy loss* if the distribution is the $H_{\max}$-attractor of a model in equilibrium with entropy loss. As we shall see, this amounts to a very special behaviour of the point probabilities. The definition we need at this point we first formulate quite generally for an arbitrary distribution P. With P we consider the *density function* Ω associated with the adapted code, cf. the appendix. In terms of P this function is given by:

$$\Omega(t) = \#\{i \in \mathbb{A} \mid p_i \geq \exp(-t)\} \tag{8.36}$$

($\#$ = "number of elements in"). We can now define a *hyperbolic distribution* as a distribution P such that

$$\limsup_{t \to \infty} \frac{\log \Omega(t)}{t} = 1\,. \tag{8.37}$$

Clearly, $\Omega(t) \leq \exp(t)$ for each t so that the equality in the defining relation may just as well be replaced by the inequality "$\geq$" .

We note that zero point probabilities do not really enter into the definition, therefore we may assume without any essential loss of generality that all point probabilities are positive. And then, we may as well assume that the point probabilities are ordered: $p^*_1 \geq p^*_2 \geq \cdots$. In this case, it is easy to see that (8.37) is equivalent with the requirement

$$\liminf_{i \to \infty} \frac{\log p^*_i}{\log \frac{1}{i}} = 1\,. \tag{8.38}$$

In the sequel we shall typically work with distributions which are ordered in the above sense. The terminology regarding hyperbolic distributions is inspired by [19] but goes back further, cf. [24]. In these references the reader will find remarks and results pertaining to this and related types of distributions and their discovery from empirical studies which we will also comment on in the next section.

We note that in (8.38) the inequality "$\leq$" is trivial as $p_i \leq \frac{1}{i}$ for every $i \in \mathbb{A}$. Therefore, in more detail, a distribution with ordered point probabilities is hyperbolic if and only if, for every $a > 1$,

$$p_i^* \geq \frac{1}{i^a} \tag{8.39}$$

for infinitely many indices.

Theorem 8.2. *Every distribution with infinite entropy is hyperbolic.*

Proof. Assume that P is not hyperbolic and that the point probabilities are ordered. Then there exists $a > 1$ such that $p_i \geq i^{-a}$ for all sufficiently large i. As the distribution with point probabilities equal to i^{-a}, properly normalized, has finite entropy, the result follows. □

With every model $\mathcal{P}$ in equilibrium we associate a *partition function* and an *exponential family*, simply by considering the corresponding objects associated with the $R_{\min}$-code for the model in question. This then follows the definition given in Section 4, but for the simple case where there is only one "energy function" with the $R_{\min}$-code playing the role of the energy function.

Theorem 8.3 (maximal models). *Let $\mathcal{P} \subseteq M_+^1(\mathbb{A})$ be given and assume that there exists a model $\mathcal{P}'$ such that $\mathcal{P}' \supseteq \mathcal{P}$, $\mathcal{P}'$ is in equilibrium and $H_{\max}(\mathcal{P}') = H_{\max}(\mathcal{P})$. Then $\mathcal{P}$ itself must be in equilibrium. Furthermore, there exists a largest model $\mathcal{P}_{\max}$ with the stated properties, namely the model*

$$\mathcal{P}_{\max} = \{P \in M_+^1(\mathbb{A}) \mid \langle \kappa^*, P \rangle \leq H_{\max}(\mathcal{P})\}\,, \tag{8.40}$$

where κ^* denotes the minimum risk code of $\mathcal{P}$. Finally, any model $\mathcal{P}'$ with $\mathcal{P} \subseteq \mathcal{P}' \subseteq \mathcal{P}_{\max}$ is in equilibrium and has the same optimal matching pair as $\mathcal{P}$.

Proof. Choose $\mathcal{P}'$ with the stated properties. By Theorem 6.1,

$$H_{\max}(\mathrm{co}\,\mathcal{P}) \leq H_{\max}(\mathrm{co}\,\mathcal{P}') = H_{\max}(\mathcal{P}') = H_{\max}(\mathcal{P})\,,$$

hence $\mathcal{P}$ is in equilibrium. Let κ^* be the $R_{\min}$-code of $\mathcal{P}$ in accordance with Theorem 6.2 and consider $\mathcal{P}_{\max}$ defined by (8.40).

Now let $\mathcal{P}' \supseteq \mathcal{P}$ be an equilibrium model with $H_{\max}(\mathcal{P}') = H_{\max}(\mathcal{P})$. As an asymptotically optimal sequence for $\mathcal{P}$ is also asymptotically optimal for $\mathcal{P}'$, we realize that $\mathcal{P}'$ has the same $H_{\max}$-attractor, hence also the same $R_{\min}$-code as $\mathcal{P}$. Thus $R(\kappa^*|\mathcal{P}') = R_{\min}(\mathcal{P}') = H_{\max}(\mathcal{P}') = H_{\max}(\mathcal{P})$ and it follows that $\mathcal{P}' \subseteq \mathcal{P}_{\max}$.

Clearly, $\mathcal{P}_{\max}$ is convex and $H_{\max}(\mathcal{P}_{\max}) = H_{\max}(\mathcal{P}) < \infty$, hence $\mathcal{P}_{\max}$ is in equilibrium by Theorem 6.1.

The final assertion of the theorem follows by one more application of Theorem 6.1. □

The models which can arise as in Theorem 8.3 via (8.40) are called *maximal models.*

Let $\kappa^* \in K(\mathbb{A})$ and $0 \leq h < \infty$. Put

$$\mathcal{P}_{\kappa^*,h} = \{P \in M_+^1(\mathbb{A}) \mid \langle \kappa^*, P \rangle \leq h\}\,. \tag{8.41}$$

We know that any maximal model must be of this form. Naturally, the converse does not hold. An obvious necessary condition is that the entropy of the matching distribution be finite. But we must require more. Clearly, the models in (8.41) are in equilibrium but it is not clear that they have κ^* as $R_{\min}$-code and h as $H_{\max}$-value.

Theorem 8.4. *A distribution $P^* \in M_+^1(\mathbb{N})$ with finite entropy has potential entropy loss if and only if it is hyperbolic.*

Proof. We may assume that the point probabilities of P are ordered.

Assume first that P^* is not hyperbolic and that P^* is the attractor for some model. Consider the corresponding maximal models $\mathcal{P}_{\kappa^*,h}$ and consider a value of h with $H(P^*) \leq h \leq H_{\max}(\mathcal{P})$. Let γ be the abscissa of convergence associated with κ^* and let Φ be defined as in the appendix. As $\gamma < 1$, we can choose $\beta > \gamma$ such that $\Phi(\beta) = h$. Now both P^* and Q_β given by

$$Q_\beta = \frac{\exp(-\beta\kappa_i)}{Z(\beta)}$$

are attractors for $\mathcal{P}_{\kappa^*,h}$ and hence equal. It follows that $h = H(P^*)$. Next we show that a hyperbolic distribution has potential entropy loss.

Consider the maximal models $\mathcal{P}_{\kappa^*,h}$. Each one of these models is given by a single linear constraint. Therefore, the attractor is element in the corresponding exponential family. The abscissa of convergence is 1 and, therefore, the range of the map $\Phi : [1;\infty[\to \mathbb{R}$ is $]\kappa_1^*; H(P^*)]$. For $h \in]\kappa_1^*; H(P^*)]$, there exists a consistent maximum entropy distribution. Assume that $h_0 > H(P^*)$ and that the attractor equals

$$Q_\beta = \frac{\exp(-\beta\kappa_i)}{Z(\beta)}.$$

By Theorem 8.1, Q_β must be attractor for all $\mathcal{P}_{\kappa*,h}$ with $h \in [\Phi(\beta); h_0]$. Especially, this holds for $h = H(P^*)$. This shows that $P^* = Q_\beta$. By Theorem 8.1 the conclusion is now clear. □

9 Zipf's law

Zipf's law is an empirically discovered relationship for the relative frequencies of the words of a natural language. The law states that

$$\log(f_i) \approx a \log(\frac{1}{i}) + b$$

where f_i is the relative frequency of the i'th most common word in the language, and where a and b denote constants. For large values of i we then have

$$a \approx \frac{\log f_i}{\log \frac{1}{i}}$$

The constants a and b depend on the language, but for many languages $a \approx 1$, see [24].

Now consider an ideal language where the frequencies of words is described by a hyperbolic probability distribution P^*. Assume that the entropy of the distribution is finite. We shall discribe in qualitative terms the consequences of these assumptions as they can be derived from the developed theory, especially Theorem 8.4. We shall see that our asumption introduces a kind of stability of the language which is desirable in most situations.

Small children with a limited vocabulary will use the few words they know with relative frequencies very different from the probabilities described by P^*. They will only form simple sentences, and at this stage the number of bits per word will be small in the sense that the entropy of the childs probability distribution is small. Therefore the parents will often be able to understand the child even though the pronounciation is poor. The parents will, typically, talk to their children with a lower bit rate than they normally use, but with a higher bit rate than their children. Thereby new words and grammatical structures will be presented to the child, and, adopting elements of this structure, the child will be able to increase its bit rate. At a certain stage the child will be able to communicate at a reasonably high rate (about $H(P^*)$). Now the child knows all the basic words and structures of the language.

The child is still able to increase its bit rate, but from now on this will make no significant change in the relative frequencies of the words. Bit rates higher than $H(P^*)$ are from now on obtained by the introduction of specialized words, which occur seldom in the language as a whole. The introduction of new specialized words can be continued during the rest of the life. Therefore one is able to express even complicated ideas without changing the basic structure of the language, indeed there is no limit, theoretically, to the bit rate at which one can communicate without change of basic structure.

We realize that in view of our theoretical results, specifically Theorem 8.4, the features of a natural language as just discussed are only possible if the language obeys Zipf's law. Thus we have the striking phenomenon that the apparent "irregular" behaviour of models with entropy loss (or just potential entropy loss) is actually the key to desirable stability, the fact that for such models you can increase the bit rate, the level of communication, and maintain the basic features of the language. One could even speculate that modelling based on entropy loss lies behind the phenomenon that many will realize as a fact, viz. that "we can talk without thinking" . We just start talking using basic structure of the language (and rather common words) and then from time to time stick in more informative words and phrases in order to give our talk more semantic content, but in doing so, we use relatively infrequent words and structures, thus not violating basic principles – hence still speaking recognizably danish, english or what the case may be, so that also the receiver or listener feels at ease and recognizes our talk as unmistakenly danish, english or ...

We see that very informative speeking can be obtained by use of infrequent expressions. Therefore a conversation between, say 2 physicists may use English supplied with specialized words like electron and magnetic flux. We recognize their language as English because the basic words and grammer is the same in all English. The specialists only have to know special words, not a special grammer. In this sense the languages are stable. If the entropy of our distribution is infinite the language will behave in just about the same manner as described above. In fact one would not feel any difference between a language with finite entropy and a language with infinite entropy.

We see that it is convienient that a language follows a Zipf's law, but the information theoretic methods also gives some explanation of how the language may have evolved into a state which obeys Zipf's law. The set of hyperbolic distributions is convex. Therefore if 2 information sources both follows Zipf's law then so do their mixture, and if 2 information sources both approximately follows Zipf's law their mixture will do this even more. The information sources may be from different languages, but it is more interesting to consider a small child learning the language. The child gets input from different sources: the mother, father, other children ect. trying to imitate their language the child will use the words with frequences which are closer to Zipf's law the the sources. As the language develops during the centuries the frequences will converge to a hyperbolic distribution.

Here we have discussed entropy as bit per word and not bit per letter. The letters give an encoding of the words which should primarilly be understood by others, and therefore the encoding cannot just be changed to obtain a better data compression. To stress the difference between bit per word and bit per letter we remark the the words are the basic semantic structure in the language. Therefore we may have an internal representation of the words which has very little to do with their length when spoken, which could explain that it is often much easier to remember a long word in a language you understand than a short word in a language you do not understand. It would be interesting to compare these ideas with empirical measurements of the entropy here considered but, precisely in the regime where Zipf's law holds, such a study is very difficult as convergence of estimators of the entropy is very slow, cf. [1].

A The partition function

In this appendix we collect some basic facts about partition functions associated with one linear constraint.

The point of departure is a code $\kappa \in K(\mathbb{A})$. With κ we associate the *partition function* $Z = Z_\kappa$ which maps $\mathbb{R}$ into $]0, \infty]$, given by

$$Z(\beta) = \sum_{i \in I} e^{-\beta \kappa_i} . \tag{A.42}$$

Here we adopt the convention that $e^{-\beta\kappa_i} = 0$ if $\beta = 0$ and $\kappa_i = \infty$. Clearly, Z is decreasing on $[1; \infty[$ and $Z(\beta) \to 0$ for $\beta \to \infty$ (note that, given K, $e^{-\beta\kappa_i} \le e^{-K} e^{-\kappa_i}$ for all i when β is large enough).

The series defining Z is a Dirichlet-series, cf. Hardy and Riesz [7] or Mandelbrojt [17]. The *abscissa of convergence* we denote by γ. Thus, by definition, $Z(\beta) < \infty$ for $\beta > \gamma$ and $Z(\beta) = \infty$ for $\beta < \gamma$. As $Z(1) = 1$, $\gamma \le 1$. If $\operatorname{supp}(\kappa)$ is infinite, $Z(\beta) = \infty$ for $\beta \le 0$, hence $\gamma \ge 0$. If $\operatorname{supp}(\kappa)$ is finite, $Z(\beta) < \infty$ for all $\beta \in \mathbb{R}$ and we then find that $\gamma = -\infty$. Mostly, we shall have the case when $\operatorname{supp}(\kappa)$ is infinite in mind.

We shall characterize γ analytically. This is only a problem when $\operatorname{supp}(\kappa)$ is infinite. So assume that this is the case and also assume, for the sake of convenience, that κ has full support and that the indexing set I is the set of natural numbers: $I = \mathbb{N}$, and that $\kappa_1 \le \kappa_2 \le \cdots$.

Lemma A.1. *With assumptions as just introduced,*

$$\gamma = \limsup_{i \to \infty} \frac{\log i}{\kappa_i} . \tag{A.43}$$

Proof. First assume that $\beta > \gamma$ with γ defined by (A.43). Then, for some $\alpha > 1$, $\alpha \log i / \kappa_i \le \beta$ for all sufficiently large values of i. For these values of i, $e^{-\beta\kappa_i} \le i^{-\alpha}$ and we conclude that $Z(\beta) < \infty$. Conversely, assume that, for some value of $\beta, Z(\beta) < \infty$. Then $\beta > 0$ and $\square$

Remark. If no special ordering on $\mathbb{A}$ is given, the abscissa of convergence can be expressed analytically via the *density function* $\Omega : \mathbb{R} \to \mathbb{N}_0$ ($\mathbb{N}_0 = \mathbb{N} \cup \{0\}$) which is defined by

$$\Omega(t) = \#\{a \in \mathbb{A} \mid \kappa(a) \le t\} \tag{A.44}$$

($\#$ = "number of elements in"). In fact, as follows easily from (A.43),

$$\gamma = \limsup_{t \to \infty} \frac{\log \Omega(t)}{t} . \tag{A.45}$$

We can now introduce the *exponential family* associated with the model κ. It is the family of distributions (Q_β) with β ranging over all values with $Z(\beta) < \infty$ which is defined by

$$Q_\beta(a_i) = \frac{e^{-\beta\kappa_i}}{Z(\beta)}\,;\; i \in I\,. \tag{A.46}$$

The family of adapted codes, denoted (ρ_β), is also of significance. These codes are given by

$$\rho_\beta(a_i) = \log Z(\beta) + \beta\kappa_i\,;\; i \in I\,. \tag{A.47}$$

We also need certain approximations to Z, Q_β and ρ_β. For convenience we stick to the assumption $I = \mathbb{N}$, $\kappa_1 \le \kappa_2 \le \dots$. We then define Z_n, $Q_{n,\beta}$ and $\rho_{n,\beta}$ by

$$Z_n(\beta) = \sum_{i=1}^{n} e^{-\beta\kappa_i}\,; \qquad \beta \in \mathbb{R}\,, \tag{A.48}$$

$$Q_{n,\beta}(a_i) = \frac{e^{-\beta\kappa_i}}{Z_n(\beta)}\,; \qquad i \le n\,, \tag{A.49}$$

$$\rho_{n,\beta}(a_i) = \log\, Z_n(\beta) + \beta\kappa_i\,; \qquad i \le n\,, \tag{A.50}$$

it being understood that $\operatorname{supp}(Q_{n,\beta}) = \operatorname{supp}(\rho_{n,\beta}) = \{1, 2, \dots, n\}$. Formally, the approximating quantities could be obtained from (non-compact) codes obtained from κ by replacing κ_i by the value ∞ for $i > n$.

We are particularly interested in the mean values $\langle\kappa, Q_\beta\rangle$ and $\langle\kappa, Q_{n,\beta}\rangle$, and define functions Φ and Φ_n; $n \ge 1$ by

$$\Phi(\beta) = \langle\kappa, Q_\beta\rangle\,; \qquad Z(\beta) < \infty\,, \tag{A.51}$$

$$\Phi_n(\beta) = \langle\kappa, Q_{n,\beta}\rangle\,; \qquad \beta \in \mathbb{R}\,. \tag{A.52}$$

Note that $\Phi(1) = H(P)$ and that

$$\Phi(\beta) = -Z'(\beta)/Z(\beta) \;\; = -\frac{d}{d\beta}\log\, Z(\beta)\,, \tag{A.53}$$

$$\Phi_n(\beta) = -Z_n'(\beta)/Z_n(\beta) = -\frac{d}{d\beta}\log\, Z_n(\beta)\,. \tag{A.54}$$

Furthermore, $-Z'$ is a Dirichlet series with the same abscissa of convergence as Z, hence $\Phi(\beta)$ is well defined and finite for all $\beta > \gamma$.

Lemma A.2. *With assumptions and notation as above, the following properties hold:*

(i) $\Phi_1 \le \Phi_2 \le \dots$,

(ii) Φ_n *is strictly decreasing on* $\mathbb{R}$ *(except if* $\kappa_1 = \cdots = \kappa_n$*),*

(iii) $\lim_{\beta\to\infty} \Phi_n(\beta) = \kappa_1$, $\lim_{\beta\to-\infty} \Phi_n(\beta) = \kappa_n$,

(iv) Φ *is strictly decreasing on* $]\gamma, \infty[$,

(v) $\lim_{\beta\to\infty} \Phi(\beta) = \kappa_1$,

(vi) $\Phi(\gamma)$ *is infinite if and only if* $-Z'(\gamma) = \infty$,

(vii) *If* $-Z'(\beta_0) < \infty$, *then* $\Phi_n \to \Phi$, *uniformly on* $[\beta_0, \infty[$,

(viii) $\lim_{n\to\infty} \Phi_n(\gamma) = \Phi(\gamma^+)$, *the limit from the right at* γ,

(ix) *for every* $\beta < \gamma$, $\lim_{n\to\infty} \Phi_n(\beta) = \infty$.

Proof. (i) follows from

$$\Phi_{n+1}(\beta) - \Phi_n(\beta) = \frac{e^{-\beta\kappa_{n+1}}}{Z_{n+1}(\beta)Z_n(\beta)} \sum_{i=1}^{n} (\kappa_{n+1} - \kappa_i) e^{-\beta\kappa_i} ,$$

(ii) from

$$\Phi_n'(\beta) = -\frac{1}{Z_n(\beta)^2} \sum_{n\geq i>j} (\kappa_i - \kappa_j)^2 e^{-\beta(\kappa_i+\kappa_j)}$$

and (iv) from an obvious extension of this formula. Writing Φ_n in the form

$$\Phi_n(\beta) = \sum_{i=1}^{n} \frac{\kappa_i}{\sum_{j=1}^{n} e^{\beta(\kappa_i-\kappa_j)}} ,$$

we derive the limiting behaviour of (iii) and, with a little care, the limit relation of (v) follows in a similar way.

(vii) follows from (A.53) and (A.54) since the convergences $Z_n(x) \to Z(x)$ and $Z_n'(x) \to Z'(x)$ hold uniformly on $[\beta_0, \infty[$ when $-Z'(\beta_0) < \infty$. Actually, the uniform convergence is first derived only for intervals of the form $[\beta_0, K]$. By (i) and (v) it is easy to extend the uniform convergence to $[\beta_0, \infty[$.

It is now clear that for $n \geq 1$ and $x > \gamma$, $\Phi_n(x) \leq \Phi(x) \leq \Phi(\gamma^+)$, hence $\lim_{n\to\infty} \Phi_n(\gamma) \leq \Phi(\gamma^+)$. On the other hand, for $x > \gamma$, $\lim_{n\to\infty} \Phi_n(\gamma) \geq \lim_{n\to\infty} \Phi_n(x) = \Phi(x)$. We conclude that (viii) holds.

If $-Z'(\gamma) < \infty$, then $Z(\gamma) < \infty$ and $\lim_{n\to\infty} \Phi_n(\gamma) = -Z'(\gamma)/Z(\gamma) < \infty$. By (viii), this shows that $\Phi(\gamma^+) < \infty$. Now assume that $-Z'(\gamma) = \infty$. If $Z(\gamma) < \infty$, it is easy to see that $\Phi(\gamma^+) = \infty$. If also $Z(\gamma) = \infty$, we choose to $N \geq 1$, $n_0 > N$ such that

$$Q_{n,\gamma}(\{1, 2, \ldots, N\}) \leq \frac{1}{2} \text{ for } n \geq n_0 .$$

Then, for $n \geq n_0$,

$$\Phi_n(\gamma) = \sum_{i=1}^{n} \kappa_i Q_{n,\gamma}(i) \geq \kappa_N Q_{n,\gamma}(\{N+1, \ldots, n\}) \geq \frac{1}{2}\kappa_N .$$

This shows that $\Phi_n(\gamma) \to \infty$, hence $\Phi(\gamma^+) = \infty$. We have now proved (vi).

In order to prove (ix), let $\beta < \gamma$ and choose, to a given K, n_0 such that $\kappa_i \geq K$ for $i \geq n_0$. Then, for $n \geq n_0$,

$$\Phi_n(\beta) \geq \frac{K \sum\limits_{i=n_0}^{n} e^{-\beta\kappa_i}}{\sum\limits_{i=1}^{n} e^{-\beta\kappa_i}} \geq K \left(1 + \frac{\sum\limits_{i=1}^{n_0-1} e^{-\beta\kappa_i}}{\sum\limits_{i=n_0}^{n} e^{-\beta\kappa_i}} \right)^{-1} .$$

As $\sum_{n_0}^{\infty} e^{-\beta\kappa_i} = \infty$, we see that for n sufficiently large, $\Phi_n(\beta) \geq K/2$ and (ix) follows. □

Remark. The formula for Φ_n' and Φ' can be interpreted more probabilistically. Consider Φ' and remark first that when we consider κ as a random variable defined on the discrete probability space $(\mathbb{A}, Q_\beta) = (\mathbb{N}, Q_\beta)$, then $\Phi(\beta)$ is the expectation of this random variable. A simple calculation shows that $\Phi'(\beta)$ is the variance of this random variable.

References

[1] A. Antos and I. Kontoyiannis: “Convergence Properties of Functional Estimates for Discrete Distributions,” to appear.

[2] J.-P. Aubin, *Optima and equilibria. An introduction to nonlinear analysis.*, Berlin: Springer, 1993.

[3] T.M. Cover and J.A. Thomas, *Information Theory*, New York: Wiley, 1991.

[4] I. Csiszár, “I-divergence geometry of probability distributions and minimization problems,” *Ann. Probab.*, vol. 3, pp. 146–158, 1975.

[5] I. Csiszár, "Sanov property, generalized I-projection and a conditional limit theorem," *Ann. Probab.*, vol. 12, pp. 768–793, 1984.

[6] R.G. Gallager, *Information Theory and reliable Communication*. New York: Wiley, 1968.

[7] G.H. Hardy and M. Riesz, *The general Theory of Dirichlet's series*. Cambridge: Cambridge University Press, 1915.

[8] P. Harremoës, "Binomial and Poisson Distributions as Maximum Entropy Distributions," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2039–2041, 2001.

[9] P. Harremoës, "The Information Topology," In preparation

[10] D. Haussler, "A general Minimax Result for Relative Entropy," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1276–1280, 1997.

[11] E. T. Jaynes, "Information Theory and Statistical Mechanics," *Physical Reviews*, vol. 106, pp. 620–630, vol. 108, pp. 171–190, 1957.

[12] E. T. Jaynes, "Clearing up mysteries – The original goal," in *Maximum Entropy and Bayesian Methods*, J. Skilling (ed.), Kluwer, Dordrecht, 1989.

[13] http://bayes.wustl.edu [ONLINE] – a web page dedicated to Edwin T. Jaynes, maintained by L. Brethorst.

[14] J.N. Kapur, "Maximum Entropy Models in Science and Engineering," New York: Wiley, 1993 (first edition 1989).

[15] D. Kazakos, "Robust Noiceless Source Coding Through a Game Theoretic Approach," *IEEE Trans. Inform. Theory*, vol. 29, pp. 577–583, 1983.

[16] S. Kullback, "Information Theory and Statistics," New York: Wiley, 1959 (Dover edition 1968).

[17] S. Mandelbrot, "Series de Dirichlet," Paris: Gauthier-Villars, 1969.

[18] B. B. Mandelbrot, "On the theory of word frequencies and on related Markovian models of discourse," in R. Jacobsen (ed.): "Structures of Language and its Mathematical Aspects," New York, American Mathematical Society, 1961.

[19] M. Schroeder, "Fractals, Chaos, Power Laws," New York: W. H. Freeman, 1991.

[20] F. Topsøe, "Information theoretical Optimization Techniques," *Kybernetika*, vol. 15, pp. 8–27, 1979.

[21] F. Topsøe, "Game theoretical equilibrium, maximum entropy and minimum information discrimination," in *Maximum Entropy and Bayesian Methods*, A. Mohammad-Djafari and G. Demoments (eds.), pp. 15–23, Kluwer, Dordrecht, 1993.

[22] F. Topsøe, "Maximum Entropy versus Minimum Risk and Applications to some classical discrete Distributions," submitted for publication

[23] F. Topsøe, "Basic Concepts, Identities and Inequalities – the Toolkit of Information Theory," http://www.mdpi.org/entropy/ [ONLINE], *Entropy*, vol. 3, pp. 162–190, 2001.

[24] G. K. Zipf, "Human Behavior and the Principle of Least Effort," Addison-Wesley, Cambridge, 1949.