Maximum Entropy versus Minimum Risk and Applications to some classical discrete Distributions

Flemming Topsøe*
Department of Mathematics
University of Copenhagen, Denmark

(final version of 05.03.01)

Abstract

The game which can be taken to lie behind the maximum entropy principle is studied. Refining previous techniques, we present a comprehensive and satisfactory theoretical discussion of the fundamentals of this game in its simplest setting. The results are illustrated by concrete examples pertaining to well known classical models.

Keywords. Maximum entropy, minimum risk, maximum entropy attractor, central code, convergence in divergence, geometric distribution, binomial distribution, Poisson distribution, multinomial distribution, empirical distribution.

1 Introduction, background information

Let \( \mathbb{A} \), the alphabet, be a countable set, \( M^1_+ (\mathbb{A}) \) the set of probability distributions over \( \mathbb{A} \) and \( K(\mathbb{A}) \) the set of (idealized) codes, i.e. the set of mappings \( \kappa : \mathbb{A} \to [0, \infty] \) with \( \sum_{a \in \mathbb{A}} \exp(-\kappa(a)) = 1 \). We say that \((\kappa, P)\) is a matching pair if \( \kappa(a) = -\ln P(a) \) for each \( a \in \mathbb{A} \) (“ln” is used for the natural logarithm). We may also express this by saying, e.g. that \( \kappa \) is adapted to \( P \), or that \( P \) is the distribution matching \( \kappa \).

*Research supported in part by the Danish Natural Science Research Council.
We use \( \langle \cdot, P \rangle \) to denote mean value w.r.t. \( P \), \( H = H(\cdot) \) to denote entropy and \( D = D(\cdot \| \cdot) \) to denote information divergence. For any \( P \in M_1^+(\mathcal{A}) \) and any \( \kappa \in K(\mathcal{A}) \),

\[
\langle \kappa, P \rangle = H(P) + D(P\|Q),
\]

where \( Q \) is the distribution matching \( \kappa \). This is the \textit{linking identity}.

A slight variation of concepts is often natural. If \( P \) is a distribution and \( \kappa \) a code, then the \textit{redundancy of} \( P \) \textit{given} \( \kappa \), thought of as the redundancy when using \( \kappa \) to code events governed by the “true” distribution \( P \), is denoted by \( D(P\|\kappa) \) and defined to be equal to \( D(P\|Q) \) where \( Q \) is the distribution matching \( \kappa \). Thus we may rewrite the linking identity in the form

\[
\langle \kappa, P \rangle = H(P) + D(P\|\kappa).
\]

A sequence \((P_n)_{n \geq 1} \subseteq M_1^+(\mathcal{A})\) \textit{converges in divergence} to \( P \in M_1^+(\mathcal{A}) \) if \( D(P_n\|P) \to 0 \). We also express this by writing \( P_n \overset{D}{\to} P \). Convergence in divergence is stronger than convergence in the usual topology of \( M_1^+(\mathcal{A}) \) (note that total variation metrizes \( M_1^+(\mathcal{A}) \) and apply Pinskers inequality). At times we find it convenient to say that \( P_n \) \textit{converges in entropy} to \( P \) if \( H(P_n) \to H(P) \). In general, this will of course not say all that much but for the specific situations we have in mind, this kind of convergence is even stronger than convergence in divergence and often requires a special argument.

We shall exploit a game closely related to the \textit{maximum entropy principle}. This game was introduced by the author in [24], cf. also [25].

It may be reasonable to use the generic term “information space” for any mathematical object which reflects the knowledge available in a given situation. We shall only consider the simplest case when this makes sense. Thus, to us, an \textit{information space} is a pair \((\mathcal{A}, \mathcal{P})\), where \( \mathcal{A} \) – the alphabet as above – is a discrete set and \( \mathcal{P} \) is a subset of \( M_1^+(\mathcal{A}) \). We shall mostly use the relatively neutral terminology \textit{model} for the set \( \mathcal{P} \). If you have applications to quantum physics in mind, it would be better to call \( \mathcal{P} \) the \textit{preparation space} – and distributions in \( \mathcal{P} \) \textit{individual preparations} – whereas, if you think in terms of statistical concepts, it would be natural to refer to \( \mathcal{P} \) as a \textit{statistical model} and, perhaps, to parametrize the distributions in \( \mathcal{P} \). The concept has of course been studied extensively in one form or the other. The view which we favour is forcefully put forward by Jaynes, cf. e.g. [14], where he stresses the distinction between distributions as the “truth” about “reality” and as a means of expressing our \textit{knowledge} about reality.

The distributions in \( \mathcal{P} \) are referred to as \textit{consistent distributions}. A distribution \( P \in M_1^+(\mathcal{A}) \) is \textit{essentially consistent} if there exists a sequence of consistent distributions which converges to \( P \) in divergence.
We consider the two-person-zero-sum game defined by the code length map which maps \((\kappa, P) \in K(\mathbb{A}) \times \mathcal{P}\) into \(\langle \kappa, P \rangle\). The set \(\mathcal{P}\) is the strategy set for the system ("Player I", if you wish) and \(K(\mathbb{A})\) is the strategy set for the observer ("Player II", if you wish). It is the objective of the observer to minimize the average code length, whereas the system attempts to maximize this quantity. For \(\kappa \in K(\mathbb{A})\),

\[ R(\kappa) = \sup_{P \in \mathcal{P}} \langle \kappa, P \rangle \]

is the risk associated with \(\kappa\) and

\[ R_{\min} = \inf_{\kappa \in K(\mathbb{A})} R(\kappa) \]

is the minimum risk of the model, written as \(R_{\min}(\mathcal{P})\) when required. The corresponding notions for the system are the infima over \(\kappa \in K(\mathbb{A})\) of \(\langle \kappa, P \rangle\) which, by (1), we recognize as the entropy \(H(P)\), and the supremum over \(P \in \mathcal{P}\) of these quantities which then is the maximum entropy value \(H_{\max} = H_{\max}(\mathcal{P})\). The game is referred to as the code length game or, more technically, the \(H_{\max}/R_{\min}\)-game.\(^1\)

Clearly, \(H_{\max} \leq R_{\min}\). If \(H_{\max} = R_{\min}\), this is the value of the game and if, furthermore, \(R_{\min} < \infty\), we say that \((\mathbb{A}, \mathcal{P})\), or just \(\mathcal{P}\), is in equilibrium.

A minimum risk code (\(R_{\min}\)-code) is an optimal strategy for the observer, i.e. a code \(\kappa\) with \(R(\kappa) = R_{\min}\). A maximum entropy distribution (\(H_{\max}\)-distribution) is an essentially consistent distribution \(P\) with \(H(P) = H_{\max}\). Thus, an optimal strategy for the system is the same as a consistent \(H_{\max}\)-distribution. A sequence \((P_n)_{n \geq 1}\) of consistent distributions is asymptotically optimal if \(H(P_n) \to H_{\max}\).

Another concept is important. We say that \(P^* \in M_1^+(\mathbb{A})\) is the maximum entropy attractor (the \(H_{\max}\)-attractor) if \(P_n \overset{D}{\to} P^*\) for every asymptotically optimal sequence \((P_n)_{n \geq 1}\). Clearly, the \(H_{\max}\)-attractor need not exist (consider, for example, the model of all deterministic distributions), but if it does, it is unique. If \(P^*\) is the \(H_{\max}\)-attractor, then \(P^*\) is essentially consistent, and \(H(P^*) \leq H_{\max}\). Therefore, it must be the unique maximum entropy distribution if \(H(P^*) = H_{\max}\).

Basic information about the \(H_{\max}/R_{\min}\)-game is contained in the result below which follows directly from theorems 1, 2 and 3 of [24], cf. also Theorem 1 of [25] (note the use of "co" for "convex hull"):

\(^1\)In [24] and [25] this game is called the absolute game in contrast to certain relative games which are of significance also for continuous distributions.
Theorem 1. The information space \((A, \mathcal{P})\) is in equilibrium if and only if \(H_{\text{max}}(\text{co}(\mathcal{P})) = H_{\text{max}}(\mathcal{P}) < \infty\). If this condition is fulfilled, there exists a unique minimum risk code, \(\kappa^*\), as well as a, likewise unique, maximum entropy attractor, \(P^*\), and \((\kappa^*, P^*)\) is a matching pair.

For a model in equilibrium, we refer to the matching pair, the existence of which is ensured by this result, as the ideal matching pair associated with the model.

We warn the reader that in Theorem 1, strict inequality in \(H(\mathcal{P}^*) \leq H_{\text{max}}\) may hold, thus the maximum entropy distribution need not exist. In the more typical case when \(H(\mathcal{P}^*) = H_{\text{max}}\) does hold, we say that entropy is preserved.

When we consider the definition of the \(H_{\text{max}}\)-attractor, we realize that Theorem 1 is, partly, a limit theorem. It is easy to deduce a further result where this feature is emphasized (note the use of “co” for “closed convex hull”):

Theorem 2. Let \((A, \mathcal{P}_n)_{n \geq 1}\) be a sequence of information spaces and assume that they are all in equilibrium with \(\sup_{n \geq 1} H_{\text{max}}(\mathcal{P}_n) < \infty\) and that they are nested in the sense that \(\text{co}(\mathcal{P}_1) \subseteq \text{co}(\mathcal{P}_2) \subseteq \cdots\). Let there further be given a model \(\mathcal{P}\) such that

\[
\bigcup_{n \geq 1} \mathcal{P}_n \subseteq \mathcal{P} \subseteq \text{co}\left(\bigcup_{n \geq 1} \mathcal{P}_n\right).
\]

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

Proof. Clearly,

\[
\sup_{n \geq 1} H_{\text{max}}(\mathcal{P}_n) \leq H_{\text{max}}(\mathcal{P}) \leq H_{\text{max}}(\text{co}(\mathcal{P})) \leq H_{\text{max}}\left(\text{co}\left(\bigcup_{n \geq 1} \mathcal{P}_n\right)\right) = H_{\text{max}}\left(\bigcup_{n \geq 1} \text{co}(\mathcal{P}_n)\right) = \sup_{n \geq 1} H_{\text{max}}(\text{co}(\mathcal{P}_n)) = \sup_{n \geq 1} H_{\text{max}}(\mathcal{P}_n)
\]

hence, by Theorem 1, \(\mathcal{P}\) is in equilibrium. The convergence assertion follows as the sequence of \(H_{\text{max}}\)-attractors for the \(\mathcal{P}_n\)'s is asymptotically optimal for \(\mathcal{P}\). \(\square\)
2 Criteria for optimality

The proof of Theorem 1 (not repeated here) is an existence proof and does not give much of a clue as to how one finds the ideal matching pair in any given situation. Therefore, there is a need to develop criteria which will facilitate the search for optimal strategies. In this respect the following concept turns out to be particularly useful. The code $\kappa^*$ is the central code for $(A, P)$ if the distribution $P^*$ which matches $\kappa^*$ is essentially consistent and $R(\kappa^*) = H(P^*) < \infty$. We shall soon see that the central code is unique. Furthermore, in typical situations, the central code does exist, cf. Theorem 4 below.

For a number of naturally occurring models, the central code is also stable (in the sense discussed in [25]), i.e. $\langle \kappa^*, P \rangle$ is finite and independent of $P$ for every consistent distribution $P$. There may be many stable codes. If a stable code has a consistent matching distribution, it must be the central code. Often, the central code can be found in this way, i.e. by first searching for stable codes. But we stress that the central code may have an inconsistent matching distribution. Note that, in principle, it is possible to check if a code is central without knowing $H_{\text{max}}$ or $R_{\text{min}}$, whereas a direct check if a given distribution is the $H_{\text{max}}$-attractor or a $H_{\text{max}}$-distribution requires that $H_{\text{max}}$ is known.

Generalizing Theorem 2 of [25] we can now establish the following criteria:

**Theorem 3.** Let $(A, P)$ be an information space and assume that the central code, $\kappa^*$, exists. Let $P^*$ be the distribution matching $\kappa^*$. Then $(A, P)$ is in equilibrium and $(\kappa^*, P^*)$ is the ideal matching pair. For $P \in P$ and for $\kappa \in K(A)$, the following sharper versions of the trivial inequalities $H(P) \leq H_{\text{max}}$ and $R_{\text{min}} \leq R(\kappa)$ hold:

\[
H(P) + D(P \parallel P^*) \leq H_{\text{max}}(P), \quad (2)
\]

\[
R_{\text{min}}(P) + D(P^* \parallel \kappa) \leq R(\kappa). \quad (3)
\]

**Proof.** As $P^*$ is essentially consistent, we may choose $(P_n)_{n \geq 1} \subseteq P$ such that $D(P_n \parallel P^*) \to 0$. Then, by the linking identity and by lower semi-continuity of the entropy function,

\[
R(\kappa^*) \geq \limsup_{n \to \infty} \langle \kappa^*, P_n \rangle = \limsup_{n \to \infty} (H(P_n) + D(P_n \parallel P^*)) = \limsup_{n \to \infty} H(P_n) \geq \liminf_{n \to \infty} H(P_n) \geq H(P^*) = R(\kappa^*).
\]
It follows that
\[
\lim_{n \to \infty} H(P_n) = H_{\text{max}} = R_{\text{min}} = R(\kappa^*) = H(P^*).
\]
In particular, \( P \) is in equilibrium, \( (P_n) \) is asymptotically optimal, \( \kappa^* \) is a minimum risk code and \( P^* \) a maximum entropy distribution.

If \( \kappa \) is any code, then
\[
R(\kappa) \geq \limsup_{n \to \infty} \langle \kappa, P_n \rangle = \limsup_{n \to \infty} (H(P_n) + D(P_n \| \kappa))
\]
\[
= H_{\text{max}} + \limsup_{n \to \infty} D(P_n \| \kappa) = R_{\text{min}} + \limsup_{n \to \infty} D(P_n \| \kappa)
\]
\[
\geq R_{\text{min}} + D(P^* \| \kappa),
\]
where, at last, we used the lower semi-continuity of \( D(\cdot \| \kappa) \) and the fact that \( P_n \overset{D}{\to} P \), hence \( P_n \to P^* \). Thus (3) holds and \( \kappa^* \) is the unique minimum risk code.

For \( Q \in \mathcal{P} \),
\[
H(Q) + D(Q \| P^*) = \langle \kappa^*, Q \rangle \leq R(\kappa^*) = H_{\text{max}},
\]
thus (2) holds. Therefore, \( P^* \) is the \( H_{\text{max}} \)-attractor as well as the unique maximum entropy distribution.

The proof shows that, with assumptions and notation as above, for any sequence \( (P_n)_{n \geq 1} \) of consistent distributions, the conditions that \( (P_n)_{n \geq 1} \) converge in divergence to \( P^* \) and that \( (P_n)_{n \geq 1} \) be asymptotically optimal are equivalent.

If \( (\mathbb{A}, \mathcal{P}) \) is in equilibrium and preserves entropy, any asymptotically optimal sequence of distributions does of course converge in entropy to the \( H_{\text{max}} \)-attractor. This points to the information spaces which are in equilibrium and which preserve entropy as the most important ones. Let us collect some facts for this class of spaces:

**Theorem 4.** Assume that the information space \( (\mathbb{A}, \mathcal{P}) \) is in equilibrium and denote by \( (\kappa^*, P^*) \) the ideal matching pair. Then the following conditions are equivalent:

(i) \( (\mathbb{A}, \mathcal{P}) \) preserves entropy,

(ii) \( (\mathbb{A}, \mathcal{P}) \) has a central code (necessarily \( \kappa^* \)),

(iii) \( (\mathbb{A}, \mathcal{P}) \) has a maximum entropy distribution (necessarily \( P^* \)).
(iv) every asymptotically optimal sequence of distributions converges in entropy to $P^*$, 

(v) there exists an asymptotically optimal sequence $(P_n)_{n \geq 1}$, of distributions such that 

$$\lim_{n \to \infty} \langle \kappa^*, P_n \rangle = \langle \kappa^*, P^* \rangle \; (= H(P^*))$$.

We leave the simple proof, based on the linking identity and the preceding theory, to the interested reader.

3 Some classical models and associated distributions

We shall study some of the classical distributions. Our aim is to demonstrate the usefulness of a purely information theoretical approach. This point of view is not new. Without being comprehensive we mention earlier research by Linnik [19], Čencov [3], Csiszár [8] and Barron [1].

Our findings can be considered as a companion to the recent correspondence [9] by Harremoës where focus was on convexity properties and detailed approximations regarding the binomial and Poisson distributions. We shall derive basic properties by as simple considerations as possible based on the $H_{\text{max}}/R_{\text{min}}$-game. In order to stress the point of view taken, we shall, slightly provocatively, redefine the classical distributions involved.

As an illustrative example, consider first a finite alphabet $A$ and the uniform distribution over $A$ which we define as the maximum entropy distribution for $P = M_1^1(A)$. Of course, this makes good sense and leads to the usual uniform distribution (directly or via Theorem 3, say). The point is that the information theoretical approach stresses the importance of this distribution as the zero-knowledge-distribution.

The concrete information spaces which we shall study are connected with the alphabets $A_n = \{0, 1, 2, \ldots, n\}; \; n \geq 1,$ and $A^* = \{0, 1, 2, \ldots\}$. We now use $E(P)$ for the meanvalue of a random variable with distribution $P$.

For $0 \leq \lambda \leq n$, $B_n(\lambda) \subseteq M_1^1(A_n)$ is the set of distributions of sums of $n$ independent Bernoulli variables for which the sum has mean value $\lambda$. And $G_n(\lambda)$ is the set of all $P \in M_1^1(A_n)$ with mean value $\lambda$: $E(P) = \lambda$. Using the natural embedding of the sets $M_1^1(A_n)$ in $M_1^1(A^*)$, we put $B^*(\lambda) = \bigcup B_n(\lambda)$ and $G^*(\lambda) = \bigcup G_n(\lambda)$, the unions being over all $n \geq \lambda$. Clearly, for
0 ≤ p ≤ 1, \( B_1(p) = G_1(p) = \{ \text{BIN}(1, p) \} \), \( \text{BIN}(1, p) \) denoting the Bernoulli distribution with parameter (success probability) \( p \).

By \( B_\infty(\lambda) \) we denote the set of distributions in \( M_+^1(\mathbb{R}^*) \) of infinite sums of independent \( \text{BIN}(1, p_n) \)-distributed random variables with \( \sum_{n=1}^{\infty} p_n = \lambda \) (by the Borel-Cantelli Lemma this makes good sense). By \( G_\infty(\lambda) \) we denote the set of all \( P \in M_+^1(\mathbb{R}^*) \) with mean value \( \lambda \). We shall use the notation \( X(I) \), where \( X \) could stand for \( B_n, B^*, B_\infty, G_n, G^* \) or \( G_\infty \) and where \( I \) is some subset of \([0, \infty[\), for the union of \( X(\lambda) \) over \( \lambda \in I \). For instance, \( G_\infty([0, \lambda]) \) is the set of \( P \in M_+^1(\mathbb{R}^*) \) with mean value at most \( \lambda \).

For the appropriate parameter values, we now define the binomial distribution \( \text{BIN}(n, p) \), the geometric distribution \( \text{GEO}(n, \lambda) \), the geometric distribution \( \text{GEO}(\lambda) \) and the Poisson distribution \( \text{POI}(\lambda) \) as the \( H_{\text{max}} \)-distribution of \( B_n(np) \), of \( G_n(\lambda) \), of \( G^*(\lambda) \) and of \( B^*(\lambda) \), respectively.

It is not clear beforehand that these definitions make sense. We shall consider this problem in the next sections.

### 4 The geometric distributions

The simplest cases to handle are the geometric distributions since, for these, the relevant models are convex.

Consider first the family \( P_x; 0 \leq x < 1 \), of distributions on \( \mathbb{R}^* \) determined by the equation

\[
P_x(k) = P_x(0) \cdot x^k; \quad k \geq 0.
\]  

(4)

The matching codes \( \kappa_x \) are given by

\[
\kappa_x(k) = -\ln P_x(0) - k \ln x; \quad k \geq 0,
\]

(5)

and are, therefore, all stable for all models \( G_\infty(\lambda); 0 \leq \lambda < \infty \). It is easy to determine \( P_x(0) \) as well as \( x = x(\lambda) \) explicitly such that \( E(P_x) = \lambda \). Not surprisingly, one finds the well known expressions:

\[
P_x(0) = \frac{1}{1 + \lambda}, \quad x = \frac{\lambda}{1 + \lambda}.
\]  

(6)

Thus, Theorem 3 applies. In particular, \( P_x \) (with \( x = x(\lambda) \)) is the \( H_{\text{max}} \)-distribution of \( G_\infty(\lambda) \) as well as of \( G_\infty([0, \lambda]) \) and

\[
H_{\text{max}}(G_\infty(\lambda)) = H_{\text{max}}(G_\infty([0, \lambda])) = \ln(1 + \lambda) + \lambda \ln(1 + \frac{1}{\lambda}).
\]  

(7)
Then fix $n$. For each $0 \leq x \leq \infty$, let $P_{n,x}$ be the distribution in $M_1^+(\mathbb{A}_n)$ for which the point probabilities are given by

$$P_{n,x}(k) = P_{n,x}(0) \cdot x^k; \quad 0 \leq k \leq n.$$  \hfill (8)

The cases $x = 0$ and $x = \infty$ are conceived as singular cases with $P_{n,0} = \delta_0$ and $P_{n,\infty} = \delta_n$ (point distributions concentrated in 0 and in $n$, respectively).

The matching codes $\kappa_{n,x}$ are given by

$$\kappa_{n,x}(k) = -\ln P_{n,x}(0) - k \ln x; \quad 0 \leq k \leq n,$$  \hfill (9)

and are, therefore, all stable for all models $G_n(\lambda)$, $0 \leq \lambda \leq n$, indeed that is how they were determined. The mean value $E(P_{n,x})$ varies from 0 (for $x = 0$) to $n$ (for $x = \infty$) with intermediate value $n/2$ (for $x = 1$). It is clear that $x \mapsto E(P_{n,x})$ is strictly increasing in $x$, a fact that also follows from continuity of this map and from Theorems 1 and 3.

To a given $0 \leq \lambda \leq n$, let $x = x_n(\lambda)$ denote that value of $x$ with $E(P_{n,x}) = \lambda$. Then Theorem 3 applies. In particular, the geometric distribution GEO($n,\lambda$) has been identified as the distribution $P_{n,x}$. It may be noted that for $0 \leq x \leq \infty$, $x \neq 1$,

$$E(P_{n,x}) = \frac{x}{1-x} - (n+1) \frac{x^{n+1}}{(1-x)^{n+1}}$$

$$= n - \left( \frac{n+1}{1-x^{n+1}} - \frac{1}{1-x} \right).$$

By (9) and as $x = x_n(\lambda) \leq 1$ for $\lambda \leq n/2$, we see that if $\lambda \leq n/2$, $\kappa_{n,x}$ is also the central code for the model $G_n([0,\lambda])$. If $n/2 \leq \lambda \leq n$, an analogous result is obtained for the model $G_n([\lambda,n])$.

Our discussion and Theorems 3 and 4 now lead to the following result:

**Theorem 5.** For fixed $n$ and $0 \leq \lambda \leq n$, $G_n(\lambda)$ is in equilibrium and the $H_{\text{max}}$-distribution, GEO($n,\lambda$), is well defined and characterized as the distribution in $G_n(\lambda)$ determined by (8). If $\lambda \leq n/2$, this distribution is also the $H_{\text{max}}$-distribution of $G_n([0,\lambda])$ and if $n/2 \leq \lambda \leq n$, it is the $H_{\text{max}}$-distribution of $G_n([\lambda,n])$.

For $0 \leq \lambda < \infty$, the model $G_\infty(\lambda)$ is in equilibrium and the $H_{\text{max}}$-distribution, GEO($\lambda$), is well defined and characterized by (4) and (6). This distribution is also the $H_{\text{max}}$-distribution of any of the models $G_\infty([0,\lambda])$, $G^*(\lambda)$ and $G^*([0,\lambda])$. The maximum entropy value $H_{\text{max}}$ is given by (7).

The models considered preserve entropy and, for $0 \leq \lambda < \infty$, the distributions GEO($n,\lambda$) converge in divergence as well as in entropy to GEO($\lambda$).
5 The binomial and Poisson distributions

In this and in the next section we shall save on the use of parenthesis by agreeing that expressions of the form $\ln k! \cdot B$ mean $(\ln(k!)) \cdot B$. As a further notational convention we agree to use $P_X$ to denote the distribution of $X$, whether $X$ is a random variable or a random vector.

First consider the model $B_n(\lambda)$ for $0 < \lambda < n$ (the cases $\lambda = 0$ and $\lambda = n$ are singular, trivial cases). Put $p = \frac{\lambda}{n}$ and $q = 1 - p$ and let $P^*$ be the distribution given by

$$P^*(k) = \binom{n}{k} p^k q^{n-k}; \quad 0 \leq k \leq n,$$

and $\kappa^*$ the matching code:

$$\kappa^*(k) = -\ln \binom{n}{k} - k \ln p - (n - k) \ln q; \quad 0 \leq k \leq n.$$  

As is well known and classical, $P^* \in B_n(\lambda)$. We shall show that $\kappa^*$ is the central code of $B_n(\lambda)$. Then Theorem 3 will apply, in particular it will follow that $P^*$ is the $H_{\max}$-distribution and in this way we will have identified $\text{BIN}(n,p)$. It was proved independently by Mateev [21] and by Shepp and Olkin [23] that $P^*$ is indeed the $H_{\max}$-distribution. For a recent treatment, see Harremoës [9]. We shall also present a proof as the availability of Theorem 3 gives rise to some simplifications and as the game theoretical approach leads to a more informative result.

What we have to prove is that for any $p_1, \ldots, p_n$ with $\sum_{i=1}^n p_i = \lambda$, the inequality

$$\langle \kappa^*, P \rangle \leq \langle \kappa^*, P^* \rangle$$

holds where $P = P_{S_n}$ with $S_n = \sum_{k=1}^n X_k$, the sum of $n$ independent $\text{BIN}(1,p_k)$-distributed random variables. It is convenient to reformulate this by introducing the random variable $T_n = n - S_n$, the number of “failures”. By $P$ we denote the distribution of the vector $(S_n, T_n)$ where $P_{S_n} = P$ and similarly for $P^*$ (when $P_{S_n} = P^*$). By $\kappa^*$ we denote the code adapted to $P^*$, i.e.

$$\kappa^*(k_1, k_2) = -\ln n! + \ln k_1! + \ln k_2! - k_1 \ln p - k_2 \ln q$$

where $0 \leq k_1 \leq n$ and $k_1 + k_2 = n$. Then (12) is equivalent with the inequality

$$\langle \kappa^*, P \rangle \leq \langle \kappa^*, P^* \rangle.$$
We find that
\[ \langle \kappa^*, P \rangle = -\ln n! - nH(p, q) + \langle \ln k!, P_{S_n} \rangle + \langle \ln k!, P_{T_n} \rangle. \] (15)

We treat the two averages here separately and show that each of them is maximized when the underlying probabilities \( p_1, \ldots, p_n \) are equal. This will of course prove (14).

By symmetry we need only consider the first average in (15). The desired result follows from a general inequality of Hoeffding, cf. Theorem 3 of [11].

For the sake of completeness let us briefly indicate the proof in the present case (the reasoning is similar to that in Harremoës [9] and was also used by Hoeffding, loc.cit.).

Fix \( p_3, \ldots, p_n \). Put \( 2\alpha = \lambda - \sum_{k=3}^{n} p_k \). Then \( p_1 + p_2 = 2\alpha \). Let us parametrize so that \( p_1 = \alpha - x, p_2 = \alpha + x \). Then \( |x| \leq \alpha \).

Let \( S' = X_1 + X_2, S'' = \sum_{k=3}^{n} X_k \). For \( 0 \leq \nu \leq n \), one finds that
\[
P(S_n = \nu) = \left( (1 - \alpha)^2 - x^2 \right) P(S'' = \nu) + (2\alpha - 2\alpha^2 + 2x^2) P(S'' = \nu - 1) + (\alpha^2 - x^2) P(S'' = \nu - 2).
\]

This leads to an expression for \( \langle \ln k!, P_{S_n} \rangle \) which, when rearranged, gives
\[
\langle \ln k!, P_{S_n} \rangle = c - x^2 \sum_{k=0}^{n-2} \ln \frac{k + 2}{k + 1} \cdot P(S'' = k)
\]
for some constant \( c \) that does not depend on \( p_1 \) or \( p_2 \). We conclude that \( \langle \ln k!, P_{S_n} \rangle \) is largest for \( x = 0 \), i.e. for \( p_1 = p_2 \). Repeating the argument, we find that the overall maximum is attained for \( p_1 = \cdots = p_n \), i.e. for \( P = P^* \).

The second average in (15) is treated in the same way. All in all, we realize that (14), hence also (12) is proved.

Now fix \( \lambda > 0 \) and consider the model \( B^*(\lambda) \). It is convenient also to consider
\[
\text{co}(B^*(\lambda)) = \bigcup_{n \geq \lambda} \text{co}(B_n(\lambda)).
\]

As \( \text{co}(B^*(\lambda)) \subseteq G^*(\lambda) \),
\[
H_{\text{max}}(B^*(\lambda)) \leq H_{\text{max}}(\text{co}(B^*(\lambda))) < \infty.
\]

Then Theorem 1 applies. It follows that \( \text{BIN}(n, \lambda/n) \) converges in divergence to the \( H_{\text{max}} \)-attractor of \( B^*(\lambda) \), for which we again use the notation \( P^* \). In
particular, the point probabilities converge. Then, by well known reasoning we conclude that

$$P^*(k) = \frac{\lambda^k}{k!} e^{-\lambda}; \quad k \geq 0.$$ 

Now Theorem 3 applies and we are ready to summarize the findings:

**Theorem 6.** The $H_{\text{max}}$-distribution for the models $B_n(\lambda) = B_n(np)$ is the classical binomial distributions $\text{BIN}(n,p)$ and the $H_{\text{max}}$-distribution for the models $B^*(\lambda), B^\infty(\lambda)$ and $\overline{\text{co}}(B^\infty(\lambda))$ is the classical Poisson distribution $\text{POI}(\lambda)$. The models considered preserve entropy and, for each $\lambda \geq 0$, $\text{BIN}(n, \lambda/n)$ converges in total variation, in divergence as well as in entropy to $\text{POI}(\lambda)$.

For the convergence in entropy we refer the reader to Theorem 8 of Harremoës [9] where property (v) of Theorem 4 is verified.\(^2\)

Mateev [21] and Shepp and Olkin [23], proved the following further result, cf. also Marshall and Olkin, [20] (and results in the next section):

**Theorem 7.** For fixed $n$, $\text{BIN}(n, \frac{1}{2})$ is the unique maximum entropy distribution among all binomial distributions $\text{BIN}(n,p)$, $0 \leq p \leq 1$.

The model considered here is an interesting example of a naturally occurring model which does not behave well seen from our information theoretical point of view in the sense that the value of the associated game does not exist:

**Theorem 8.** For $n$ fixed, consider the model of all binomial distributions $\text{BIN}(n,p)$; $0 \leq p \leq 1$. Then $H_{\text{max}} = H(\text{BIN}(n,1/2)) < R_{\text{min}} = \ln(n + 1)$.

**Proof.** Let $\mathcal{P}$ denote the model in question and consider $\text{co}(\mathcal{P}) = \overline{\text{co}}(\mathcal{P})$. Then $H_{\text{max}}(\text{co}(\mathcal{P})) = R_{\text{min}}(\text{co}(\mathcal{P})) = \ln(n + 1)$. This follows as the uniform distribution over $\{0, 1, 2, \ldots, n\}$ belongs to $\text{co}(\mathcal{P})$. In fact, this distribution is the uniform mixture over $x \in [0, 1]$ of the binomial distributions $\text{BIN}(n,x)$, a direct consequence of the classical formula for the beta function since, by that formula,

$$\int_0^1 \binom{n}{k} x^k (1-x)^{n-k} \, dx = \binom{n}{k} \frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} = \frac{1}{n+1}$$

for all $0 \leq k \leq n$. \(\square\)

\(^2\)In fact, a slightly more streamlined approach than above is to use this result in connection with a direct investigation of the code adapted to the Poisson distribution. In view of the previous findings involving mean values of $\ln k!$ it follows that the code is central.
6 Multinomial distributions, empirical distributions

The basis of considerations in the previous section was Bernoulli variables. They only assume two values corresponding to “success” and “failure”. Now let us consider a more general situation but still within discrete probability theory. Without loss of generality we may then confine the study to random variables taking values in the natural numbers \( \mathbb{N} = \{1, 2, \cdots \} \) (representing the various levels of “success”).

Denote by \( \Omega^n \) the set of infinite vectors \( k = (k_1, k_2, \cdots) \) with the integers \( k_i \) all non-negative and \( \sum_1^\infty k_i = n \), and denote by \( \Omega^* \) the set of similar vectors but with the looser requirement \( \sum_1^\infty k_i < \infty \). Then \( \Omega^* \) is discrete (i.e. countable) and decomposed into the sets \( \Omega_0, \Omega_1, \cdots \).

We now fix \( n \in \mathbb{N} \) and \( Q \in M_1^+(\mathbb{N}) \). We also agree that if \( P_1, \cdots, P_n \) are distributions in \( M_1^+(\mathbb{N}) \), then \( P \) denotes the average \( \frac{1}{n} \sum_1^n P_k \).

Consider the model \( P = P(n, Q) \) constructed as follows. For each finite set \( P_1, \cdots, P_n \) in \( M_1^+(\mathbb{N}) \) with \( P = Q \) we consider independent random variables \( X_1, \cdots, X_n \) such that \( X_\nu \) has distribution \( P_\nu \); \( 1 \leq \nu \leq n \) and then we consider the random vector \( S_n = (S_n_1, S_n_2, \cdots) \) where \( S_n_i \) denotes the number of \( 1 \leq \nu \leq n \) with \( X_\nu = i \); \( i = 1, 2, \cdots \). By definition, the model \( P \) consists of all distributions of random vectors \( S_n \) that arise in this way. A typical element of \( P \) is denoted \( P \), i.e. \( P = P(S_n) \).

**Theorem 9.** If \( H(Q) < \infty \), then \( P(n, Q) \) is in equilibrium and preserves entropy. The maximum entropy distribution is \( P_0 = P(Q, \cdots, Q) \).

**Proof.** Let \( \kappa_0 \) be the code adapted to \( P_0 \) and let \( q_i, i \geq 1 \) be the point probabilities for \( Q \). Then \( P_0 \) is given by

\[
P_0(k_1, k_2, \cdots) = n! \prod_{i=1}^\infty \frac{q_i^{k_i}}{k_i!}
\]  

for all \( (k_1, k_2, \cdots) \in \Omega^n \) and for these vectors,

\[
\kappa_0(k_1, k_2, \cdots) = -\ln n! - \sum_{i=1}^\infty k_i \ln q_i + \sum_{i=1}^\infty \ln k_i!.
\]  

In order to establish the theorem we shall prove that \( \kappa_0 \) is the central code, i.e. for any \( P = P(S_n) \in P \) with \( S_n = (S_{n1}, S_{n2}, \cdots) \) we shall prove that \( \langle \kappa_0, P \rangle \)
is maximal for \( P_1 = \cdots = P_n = Q \). By (17),

\[
\langle \kappa_0, P \rangle = -\ln n! - \sum_{i=1}^{\infty} n q_i \ln q_i + \sum_{i=1}^{\infty} E(\ln S_{ni}!)
\]

\[
= -\ln n! + n H(Q) + \sum_{i=1}^{\infty} E(\ln S_{ni}!).
\]

From the investigation in the previous section we realize that the individual expectations \( E(\ln S_{ni}!) \) are upper bounded by the corresponding expectations when all \( i \)'th point probabilities of \( P_1, \cdots, P_n \) are equal. Thus

\[
\langle \kappa_0, P \rangle \leq -\ln n! + n H(Q) + \sum_{k=2}^{\infty} \sum_{i=1}^{n} k! \cdot \binom{n}{k} q_i^k (1-q_i)^{n-k}
\]

\[
= H(P_0).
\]

As we also find that \( H(P_0) < \infty \), the proof is complete. \( \square \)

Let us agree that \( H_{\text{max}}(n, Q) \) denotes the maximum entropy value of the model \( P(n, Q) \). From the above calculations it is easy to deduce an important structural property of the function \( (n, Q) \mapsto H_{\text{max}}(n, Q) \):

**Theorem 10.** Given \( n \in \mathbb{N} \), define \( f_n : [0,1] \to \mathbb{R}_+ \) by

\[
f_n(q) = -n q \ln q + \sum_{k=0}^{n} \ln k! \cdot \binom{n}{k} q^k (1-q)^{n-k}.
\]

Then \( f_n \) is strictly concave and, for any \( Q = (q_1, q_2, \cdots) \in M_+^1(\mathbb{N}) \),

\[
H_{\text{max}}(n, Q) = -\ln n! + \sum_{i=1}^{\infty} f_n(q_i).
\]

**Proof.** The validity of (19) with \( f_n \) given by (18) was shown above. Our proof that \( f_n \) is strictly concave will follow the proof of Lemma E.2.a. in Chapter 13 of Marshall and Olkin [20]. Let us briefly indicate the details. An interesting calculation shows that

\[
f''_n(q) = -\frac{n}{q} + n(n-1) \sum_{k=0}^{n-2} \binom{n-2}{k} q^k (1-q)^{n-k-2} \ln \frac{k+2}{k+1}
\]
and upper-bounding the logarithmic term by \( \frac{1}{k+1} \) gives the inequality

\[
f''_n(q) < -\frac{n}{q} + \frac{n}{q} \sum_{k=0}^{n-2} \left( \frac{n-1}{k+1} \right) q^{k+1} (1-q)^{n-k-2} = -\frac{n}{q} (1-q)^{n-1},
\]

hence \( f''_n(q) < 0 \) and the proof is complete.

As a corollary we see that \( Q \searrow H_{\text{max}}(n, Q) \) is strictly concave: For any infinite convex combination \( \Sigma \alpha_\nu Q_\nu \) of measures in \( M_+^1(\mathbb{N}) \),

\[
H_{\text{max}} \left( n, \sum_{\nu=1}^{\infty} \alpha_\nu Q_\nu \right) \geq \sum_{\nu=1}^{\infty} \alpha_\nu H_{\text{max}}(n, Q_\nu)
\]

and, if the right hand side is finite, the inequality is strict unless all \( Q_\nu \) with \( \alpha_\nu > 0 \) are identical.

For \( n = 1 \), (20) reduces to the usual concavity property of the entropy function.

Note that \( H_{\text{max}}(n, Q) \) is the entropy of the empirical distribution corresponding to \( n \) independent random variables \( X_1, \ldots, X_n \), all with distribution \( Q \). The qualifying “max” signals that this empirical distribution has maximal entropy among all “generalized empirical distributions” corresponding to independent random variables subject to the condition that the average of the associated distributions coincide with \( Q \). This is the content of Theorem 9.

In case \( Q \) has finite support, the empirical distribution just discussed is nothing but the multinomial distribution determined by \( Q \), denoted by, say \( \text{MULT}(n, Q) \). Combining the concavity of \( H_{\text{max}}(n, \cdot) \) with the obvious symmetry of this function and referring to Theorem 9, we obtain the following result, generalizing parts of Theorem 6 and Theorem 7:

**Theorem 11.** Let \( r \geq 2 \) and \( n \geq 2 \) be natural numbers. Among all generalized empirical distributions corresponding to independent random variables \( X_1, \ldots, X_n \) with values in \( \{1, \ldots, r\} \), the multinomial distribution \( \text{MULT}(n, Q) \) with \( Q \) the uniform distribution on \( \{1, \ldots, r\} \) has maximal entropy.

For the case when only multinomial distributions are considered in the model, this result was proved, again both by Mateev and by Shepp and Olkin.
7 Discussion

The $H_{\text{max}}/R_{\text{min}}$-game, results

This correspondence and previous research demonstrates that the $H_{\text{max}}/R_{\text{min}}$-game is useful when setting up natural models reflecting our knowledge in a given situation. Typically, the kind of results one can expect from this approach are two-fold: Identification of interesting distributions and associated limit theorems (possibly accompanied by certain inequalities, facilitated by e.g. (2) an (3)). Of course, our setting is quite simple. Extensions may involve extra structure, say in the form of Markov kernels, side information or symmetry.

As emphasized already, the point of the present correspondence is the improved theoretical results regarding the $H_{\text{max}}/R_{\text{min}}$-game. In fact it appears that for this game – the simplest among several possibilities – all basic theoretical questions have been given a satisfactory answer. The useful Theorem 3 which has a (surprisingly?) simple proof was overlooked in [25] (compare with Theorem 2 of that paper). The observation needed is that one may replace a demand of equality associated with the notion of stable codes by a less restrictive demand of inequality.

The $H_{\text{max}}/R_{\text{min}}$-game, terminology

The terminology adopted differs somewhat from previous practice. Most pronounced, and perhaps a bit controversial, is the definition of a maximum entropy distribution where, besides the natural requirement $H(P^*) = H_{\text{max}}$, we only require that $P^*$ be essentially consistent, not necessarily consistent. The author feels quite strongly that the definition here put forward is “the right one”. Arguments in favour of this choice lie in the theoretical results and in typical examples as those presented here. Formulation of results is facilitated and, as follows from Theorem 2 (with $P_n$’s independent of $n$), it does not really matter if you add the essentially consistent distributions to a model.

Another matter is that we changed the terminology from “centre of attraction” (cf. [24]) to the present “maximum entropy attractor”. This is not very significant, and only motivated by the wish to avoid a possible misunderstanding which the word “central” might lead to, viz. that the distribution concerned is consistent.

Possible loss of entropy

A fundamental phenomenon is the possibility that the equality $H(P^*) = H_{\text{max}}$ may be impossible for any consistent or essentially consistent distribution. As examples show, cf. Ingarden and Urbanik, [12] (or [24]) this situation
may occur. This also explains the importance of the $H_{\text{max}}$-attractor. It always exists (for models in equilibrium), and does allow a possible drop of entropy ($H(P^*) < H_{\text{max}}$). Originally, cf. [24], the notion was defined differently, requiring only normal convergence rather than convergence in divergence for asymptotically optimal sequences. But as the stronger convergence property does in fact hold, and as convergence in divergence appears to be just the right kind of convergence to work with for investigations as here discussed, the suggested definition appears to be justified.

If we turn the attention to differential entropy and the associated maximum entropy principle, the same phenomenon of loss of entropy may take place. The reader is referred to Cover and Thomas [5] for an illuminating discussion.

A fine topology of information theory

As already stated, convergence in divergence is a key concept. Note that this notion is topological in that there exists a unique strongest topology on $M_1^+(\mathcal{A})$ for which convergence for sequences is identical with convergence in divergence. This may be seen directly but is clear also in view of a general topological result due to Kisyński, cf. [17]. The topology can also be defined in the continuous case and appears to be of quite some significance for the subject matter of the present correspondence and for the discussion of limit theorems more generally. We propose to call this topology the information topology. It is a fine topology of information theory, strictly finer than the usual topology, even for finite alphabets. The existence of this topology does not contradict well known research of Csiszár, cf. [6], [7], who showed, basically, that a definition based on “neighbourhoods” of the form $\{P \mid D(P\|P^*) < \varepsilon\}$ is not feasible. For finite $\mathcal{A}$, convergence in divergence (hence in the information topology) of $P_n$ to $P$ amounts to normal convergence and the eventual inclusion of the support of the $P_n$’s in that of $P$, whereas for infinite $\mathcal{A}$, convergence in divergence is a much stronger property.

Stable codes, exponential families

Some remarks on the significance of the notion of stable codes are in order. Consider, for example, the family of stable codes given by (9) in the study of geometric distributions. This family we conceive as the exponential family associated with the given model. Normally, it is the family of matching distributions — in the present case given by (8) — which constitutes the exponential family in statistical literature. However, we maintain the view that it is more natural to focus on the codes and the property of stability. The author hopes soon to return to this point of view in greater detail.

Further references

Regarding the game-theoretical approach, the reader will find interesting material in references already pointed to and in Kazakos [16] and, for games
covering also the continuous case, in Haussler [10]. However, these authors were unaware of the present authors paper [24] (which offers a number of theoretical simplifications).

**Geometric distributions and beyond**

The type of problems treated in Section 5 are, apart from the game theoretical approach, well known, even classical. This also concerns more general moment-type conditions, possibly involving several such conditions. In statistical thermodynamics, the *partition function* occupies a central position for such investigations and this function also appears in other literature. From the reference list we may quote [13], [14], [12], [24] and [25], but there are many other sources from physics, chemistry, statistics (exponential families) and information theory.

The game theoretical approach leads directly to the partition function in a way which is more natural than in other studies (which are, typically, based on Lagrange multipliers). The search for stable codes is instrumental in this respect.

**Applications to binomial and Poisson distributions**

In relation to the treatment of the binomial and Poisson distributions, we note that much more accurate results than those given here are contained in Harremoës [9], cf. also the references listed there.

We also want to emphasize the conjecture, going back to Shepp and Olkin [23], cf. also Harremoës [9], that \( H(p_1, \ldots, p_n) \) is concave in \( (p_1, \ldots, p_n) \), where \( H(p_1, \ldots, p_n) \) denotes the entropy of the sum of independent Bernoulli variables with success probabilities, respectively, \( p_1, \ldots, p_n \). The best results in this direction appear to be Lemma 4 of Harremoës, [9], where the case when all except two success probabilities are kept fixed is considered, and then the result of Mateev and Shepp and Olkin, reproduced here as Theorem 7, which considers the case of equal success probabilities. Note that Shepp and Olkin proved the somewhat weaker property of Schur-concavity, see also Marshall and Olkin, [20], Theorem E.1 of Chapter 13. That this condition is indeed weaker than normal concavity follows from the discussion in [20], cf. Proposition C.2 of Chapter 3, where it is proved that concavity in conjunction with symmetry implies Schur-concavity.

**A reformulation of one of the Mateev- and Shepp and Olkin results**

In section 6 we generalized one of Mateev and Shepp and Olkin’s results to a vector valued setting. We now consider the possibility of a generalization in another direction.

Let \( \lambda \) and \( n \) with \( 0 \leq \lambda \leq n \) be given and put \( p = \frac{\lambda}{n} \). Then \( \text{BIN}(n, p) \) is the \( H_{\max} \)-distribution of the model \( B_n(\lambda) \). This result, due to Mateev (somewhat
put away in the proof of Corollary 2 of [21]) and to Shepp and Olkin [23], was an important part of Theorem 6. It is equivalent to the following inequality:

**Theorem 12.** Let $P_1, P_2, \cdots, P_n$ be distributions in $M_1^+(\mathbb{A}^*)$ and put

$$\bar{P} = \frac{1}{n} \sum_{k=1}^{n} P_k. \tag{21}$$

Then, if $P_1, P_2, \cdots, P_n$ are all supported by $\{0, 1\}$,

$$H(\bar{P} * n) \geq H(P_1 * P_2 * \cdots * P_n), \tag{22}$$

where “*” denotes convolution of distributions.

To realize the stated equivalence, simply note that all $P_k$ are of the form $P_k = \text{BIN}(1, p_k)$, that $\bar{P} = \text{BIN}(1, \bar{p})$ with $\bar{p} = \frac{1}{n} \sum_{1}^{n} p_k$ and recall that the distribution of a sum of independent random variables is the convolution of the corresponding individual distributions.

It is natural to inquire if the above inequality holds under less stringent conditions on the support of the distributions $P_k$. Though interesting results in this direction may hold, it seems that Theorem 12 is a very special and perhaps in some sense unique instance of such results. This is illustrated by the simple example for which $n = 2$ and

$$P_1 = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, \cdots\right),$$
$$P_2 = \left(\frac{1}{2}, 0, \frac{1}{2}, 0, 0, \cdots\right)$$

(in terms of the point probabilities associated with the elements in $\mathbb{A}^* = \{0, 1, 2, \cdots\}$). One finds that

$$P_1 * P_2 = \left(\frac{1}{6}, \frac{1}{6}, \frac{2}{6}, \frac{1}{6}, \frac{1}{6}, 0, 0, \cdots\right),$$
$$\bar{P} * \bar{P} = \left(\frac{25}{144}, \frac{20}{144}, \frac{54}{144}, \frac{20}{144}, \frac{25}{144}, 0, 0, \cdots\right)$$

and

$$H(P_1 * P_2) = \frac{2}{3} \ln n 2 + \ln n 3 \approx 1.5607,$$
$$H(\bar{P} * \bar{P}) = \frac{321}{72} \ln n 2 + \frac{7}{8} \ln n 3 - \frac{35}{36} \ln n 5 \approx 1.5241,$$

hence (22) does not hold in this case.

One may also view (22) as a kind of data-reduction inequality related to product measures (rather than to convolution of measures) where the data-reduction is effectuated by the sum-map $(k_1, \cdots, k_n) \leadsto k_1 + \cdots + k_n$.  

19
Comments on Hoeffding’s inequality

With conditions and notation as in Theorem 12, the inequality of Hoeffding which we used to prove that result (Theorem 3 of [11]) can be stated as follows: For any convex function \( g : \mathbb{A}^* \rightarrow \mathbb{R} \), the inequality

\[
\langle g, \mathcal{P}^*n \rangle \geq \langle g, P_1 * P_2 * \cdots * P_n \rangle
\]

holds. The argument given in section 5 actually proves this result, though we only needed the special case \( g(k) = \ln k! \).

It may be worth while to notice that whereas Theorem 12 itself appears difficult to generalize, far reaching generalizations of (23) (weakening the requirement on the support of the \( P_k \)'s and generalizing the whole setting to one based on abstract semigroup theory) have appeared, cf. Bickel and van Zwet [2], Christensen and Ressel [4] and Ressel [22]. How, or if, these results can be exploited in the context of information theory is unclear.

Looking back on the results for binomial- and polynomial distributions

The key results are Theorems 6, 7, 9 and 11. As noted several times, most of this is due to Mateev and to Shepp and Olkin (apart from the game theoretical view, the new results concern Theorem 9 and parts of Theorem 11). If we just focus on maximum entropy type of results, really, there are no surprises - unless you consider it a surprise that the proofs are relatively complicated in the sense that they need a proper structural treatment before they become tractable. Perhaps these problems were considered by many mathematicians before they were settled. Both Mateev and Shepp and Olkin refer to sources of inspiration from others – M.B. Malyutov, B. Lindström [18] and A.D. Wyner.

In Shepp and Olkin’s paper [23] we find a reference to Hoeffding’s inequality from [11], discussed above. Really, the reference is only one of analogy. Shepp and Olkin do not make any use of Hoeffding’s inequality, but note its qualitative similarity with problems and results they are led to consider. In this context it is interesting that with the present approach, Hoeffding’s inequality is the central tool needed - of course supplied with the general theoretical findings related to the \( H_{\max}/R_{\min} \)-game.

In comparison to previous methods, the proofs presented in this correspondence appear to be shorter, intrinsic (no recourse to Lagrange multipliers or other general analytical tools) and they give more (via the game theoretical approach). Needless to say, in developing the results, the author has been inspired by previous research here referred to.

Illustrations of some models
Finally, we shall illustrate some of the findings regarding the binomial and geometric distributions by looking at the case $n = 2$. The simplex $M^+_1(A_2)$ together with various models are shown in Figure 1. The points $P_0$, $P_1$ and $P_2$ represent the deterministic distributions concentrated in 0, 1 and 2, respectively. The line $ABCD$ represents the model $G_2(λ)$ for some $λ < 1$, whereas $AB$ represents the model $B_2(λ)$. Note that for higher values of $n$, $B_n(λ)$ is not convex (but still connected). The points $B$ and $C$ represent the $H_{\text{max}}$-distributions of $B_2(λ)$ and $G_2(λ)$, respectively. The curve (parabola) $P_0BEP_2$ represents the model of all binomial distributions $\text{BIN}(2, p)$; $0 \leq p \leq 1$, and $E$ the associated maximum entropy distribution. Similarly, the curve $P_0CUP_2$ is the model of geometric distributions and $U$, the uniform distribution, the corresponding maximum entropy distribution.

Fig. 1. Illustration of models with $A = \{0, 1, 2\}$.

**Acknowledgements**

The author thanks Peter Harremoës for useful discussions of all parts of the manuscript. In fact, the present research was triggered by a study of the recent contribution Harremoës [9] which, indirectly, pointed to possibilities for refinements of the theoretical basis.
The intricacies of entropy estimation in the binomial distribution was first pointed out to the author quite some years ago by Ole Groth Jørsboe who wrote the report [15] based on [23]. The essential importance of Hoeffding’s inequality for a key argument in section 5 was discussed with Paul Ressel who also drew the authors attention to the papers [2], [4] and [22]. The discussion of Theorem 12 and of Hoeffding’s inequality owes much to the exchange of ideas with Paul Ressel.

References


