V_a and a point v_b on the boundary of V_b such that the line segment $L = \{\eta v_a + (1-\eta)v_b \colon \eta \in [0, 1]\}$ is contained in both V_b and U_s. Let $H_1(\cdot)$ denote one-dimensional Hausdorff measure in $\mathbb{R}^d$. By virtue of (19)

$$2\epsilon \le \|v_a - v_b\| = H_1(L). \tag{21}$$

The definition of L_n ensures that U_s is the union of $k \le b_n$ disjoint sets $U_1, \ldots, U_k$, each of which is a terminal region of T_n. In conjunction with (20), this implies that

$$H_1(L) = \sum_{j=1}^{k} H_1(L \cap U_j) \le \sum_{j=1}^{k} \operatorname{diam}(U_j) \le k\,\frac{\epsilon}{b_n} \le \epsilon.$$

However, this contradicts (21), so that $U_s = \bigcup_{j=1}^{k} U_j$ must be contained in V_b. The inequality above then shows that $\operatorname{diam}(U_s) \le \epsilon$, and therefore $\max\{\operatorname{diam}(T'_n[x]) \colon x \in V_a\} \le \epsilon$. It follows that

$$\limsup_{n\to\infty} P\{x\colon \operatorname{diam}(T'_n[x]) > \epsilon\} \le P(V_a^c) \le \delta$$

for every choice of $\epsilon, \delta > 0$. Relabeling the trees T'_n if necessary, Lemma 1 ensures that $R(T'_n) \to 0$. The consistency of the complexity pruned subtrees $\hat{T}_n$ follows immediately from Corollary 1.

References

[1] A. R. Barron, "Complexity regularization with application to artificial neural networks," in *Nonparametric Functional Estimation and Related Topics*. ser. NATO ASI, G. Roussas, Ed. Dordrecht, The Netherlands: Kluwer Academic, 1991, pp. 561–576.

[2] A. Barron, L. Birgé, and P. Massart, "Risk bounds for model selection via penalization," *Probab. Theory Related Fields*, vol. 113, pp. 301–413, 1999.

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.

[4] A. Ciampi, C.-H. Chang, S. Hogg, and S. McKinney, "Recursive partition: A versatile method for exploratory data analysis in biostatistics," in *Biostatistics*, I. B. MacNeil and G. J. Umphrey, Eds. Dordrecht, The Netherlands: Reidel, 1987, pp. 23–50.

[5] P. Chou, T. Lookabaugh, and R. M. Gray, "Optimal pruning with applications to tree-structured source coding and modeling," *IEEE Trans. Inform. Theory*, vol. 37, pp. 31–42, Jan. 1989.

[6] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.

[7] D. L. Donoho, "CART and best-ortho-basis: A connection," *Ann. Statist.*, vol. 25, pp. 1870–1911, 1997.

[8] S. B. Gelfand, C. S. Ravishankar, and E. J. Delp, "An iterative growing and pruning algorithm for classification tree design," *IEEE Trans. Pattern Anal. Machine Intel.*, vol. 13, pp. 163–174, Feb. 1991.

[9] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Dordrecht, The Netherlands: Kluwer, 1991.

[10] S. Gey and E. Nedelec, "Model selection for CART regression trees," Lab. Math., Univ. Paris-Sud, Orsay, France, Prépublication 2001-56, 2001.

[11] L. Gordon and R. Olshen, "Asymptotically efficient solutions to the classification problem," *Ann. Statist.*, vol. 6, pp. 515–533, 1978.

[12] ——, "Almost sure consistent nonparametric regression from recursive partitioning schemes," *J. Multivariate Anal.*, vol. 15, pp. 147–163, 1984.

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag, 2001.

[14] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, pp. 13–30, 1963.

[15] G. Lugosi and K. Zeger, "Concept learning using complexity regularization," *IEEE Trans. Inform. Theory*, vol. 42, pp. 48–54, Jan. 1996.

[16] G. Lugosi and A. B. Nobel, "Consistency of data-driven histogram methods for density estimation and classification," *Ann. Statist.*, vol. 24, pp. 687–706, 1996.

[17] A. B. Nobel and R. A. Olshen, "Termination and continuity of greedy growing for tree-structured vector quantizers," *IEEE Trans. Inform. Theory*, vol. 42, pp. 191–205, Jan. 1996.

[18] A. B. Nobel, "Vanishing distortion and shrinking cells," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1303–1305, July 1996.

[19] ——, "Histogram regression estimation using data-dependent partitions," *Ann. Statist.*, vol. 24, pp. 1084–1105, 1996.

[20] ——, "Recursive partitioning to reduce distortion," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1122–1133, July 1997.

[21] B. D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[22] V. N. Vapnik, *Estimation of Dependencies Based on Empirical Data*. New York: Springer-Verlag, 1982.

[23] V. N. Vapnik and A. Ya. Chervonenkis, *Theory of Pattern Recognition* (in Russian). Moscow, U.S.S.R.: Nauka, 1974.

[24] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

[25] L. C. Zhao, P. R. Krishnaiah, and X. R. Chen, "Almost sure L_r-norm convergence for data-based histogram density estimates," *Theory of Probab. its Applic.*, vol. 35, pp. 396–403, 1990.

[26] V. N. Vapnik and A. Ya. Chervonenkis, *Theorie der Zeichenerkennung* (in German). Berlin, Germany: Akademie Verlag, 1979.

Maximum Entropy Versus Minimum Risk and Applications to Some Classical Discrete Distributions

Flemming Topsøe

Abstract—**The game which can be taken to lie behind the maximum-entropy principle is studied. Refining previous techniques, new theoretical results are obtained. These results are illustrated by concrete examples pertaining to well-known classical models.**

Index Terms—**Binomial distribution, code length game, convergence in divergence, empirical distribution, geometric distribution, maximum-entropy attractor, maximum-entropy distribution, multinomial distribution, Nash equilibrium code, Poisson distribution.**

I. Introduction and Background Information

Let $\mathbb{A}$, the *alphabet*, be a finite or countably infinite set. The notion of (*idealized*) *codes* will play an important role in the sequel. For the purpose of motivation, we remind the reader of the usual notion of a *binary prefix-free code*. This is a map which to any $a \in \mathbb{A}$ assigns a binary codeword in such a way that no codeword which appears in this way is a prefix of another such codeword. An example is shown in Table I which displays a codebook for the first six letters of the English alphabet. We may use the code for identification of an unknown letter from $\mathbb{A}$ or we may conceive the code as a strategy for observation, assuming that an observation is broken down in units of binary questions, and that any such question is feasible. For the example shown, the code points to the question "is the letter one of a, b, c, d, or f?"

Manuscript received March 6, 2001; revised January 24, 2002. This work was supported by the Danish Natural Science Research Council and by INTAS, Project 00-738. The material in this correspondence was presented at the Information Theory Workshop, Banach Center, Warsaw, Poland, October 2001 and at the Conference on Information Theory, Cryptography and Statistics, Budapest, Hungary, October 2001.

The author is with the Department of Mathematics, University of Copenhagen, DK-2100 Copenhagen, Denmark (e-mail: topsoe@math.ku.dk).

Communicated by P. Narayan, Associate Editor for Shannon Theory.

Publisher Item Identifier 10.1109/TIT.2002.800479.

TABLE I
A CODEBOOK

$\mathbb{A}$	codeword	κ
a	100	3
b	1110	4
c	101	3
d	110	3
e	0	1
f	1111	4

as our first question (equivalent to the question "is the first binary digit in the codebook a 1?"). Continuing in this way, enquiring about the further binary digits until identification is possible, we realize that the *code length function* κ which is shown in Table I gives us the number of bits needed for identification.

By $M_+^1(\mathbb{A})$ we denote the set of probability distributions over $\mathbb{A}$. If, besides the code given in Table I, we also know the true distribution P, then the *average code length*, for which we use the bracket notation $\langle \kappa, P\rangle$, can be computed. According to what was said above, we can interpret $\langle \kappa, P\rangle$ as *mean observation time* (basing observations on the given code and assuming that P is the true distribution). On the coding side, we realize that the detailed structure of the code is immaterial for the calculation of mean observation time—only the code length, as given by the function κ, is important. It is, therefore, essential to note the following result.

Theorem 1 Kraft's Inequality: A necessary and sufficient condition for a function

$$\kappa: \mathbb{A} \to \mathbb{N}_0 = \{0, 1, 2, \ldots\}$$

to be a code-length function for a binary prefix-free code is that the following inequality holds:

$$\sum_{i \in \mathbb{A}} 2^{-\kappa(i)} \leq 1. \tag{1}$$

It is convenient to expand this slightly by allowing the value $\kappa(i) = \infty$ corresponding to the "impossible," infinitely long codeword, a codeword which has no finite codeword as prefix and is, typically, used in situations where you believe that the letter in question cannot possibly occur. Allowing $\kappa(i) = \infty$ does not change the validity of Kraft's result, quoted above. In order to be systematic, we should also allow the empty codeword with length 0. In practice, this is only used when you feel certain what the outcome will be.

A simple proof of Theorem 1, which works equally well for finite and for infinite alphabets, depends on the natural 1–1 correspondence between codewords and binary intervals. For this correspondence, the empty codeword corresponds to $[0, 1]$ and if $\varepsilon_1 \cdots \varepsilon_k$ corresponds to the interval I, then $\varepsilon_1 \cdots \varepsilon_k 0$ corresponds to the left half, and $\varepsilon_1 \cdots \varepsilon_k 1$ to the right half of I (and the "impossible" codeword corresponds to the empty set $\emptyset$). For instance, for the code given in Table I, you find that the set of corresponding binary intervals is as shown in Fig. 1.

Note also that the case of equality in (1) corresponds to the case of a "maximally compressed" code in the sense that no binary prefix-free code κ^* has a code length function which satisfies $\kappa^*(i) \leq \kappa(i)$ for all $i \in \mathbb{A}$ with strict inequality for one or more $i \in \mathbb{A}$. The reader will find more details in [7].

In spite of how well known the above facts are, they are still needed as motivation for the game we shall study.

Apart from focusing on the code length function κ (and not on the full code) we decide, first, to pay attention only to maximally compressed codes, i.e., to the case of equality in (1), and, second, to idealize by allowing arbitrary nonnegative numbers as codeword lengths.

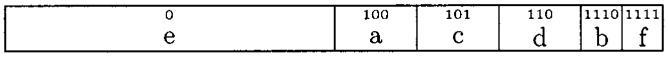

Fig. 1. Binary intervals corresponding to the code in Table I.

This idealization is motivated by the wish to avoid somewhat arbitrary effects caused by the choice of the binary alphabet $\{0, 1\}$ as reference alphabet, and can be justified in various ways, e.g., by pointing to block coding and the noiceless coding theorem or by the fact that any idealized code length function will be at most one bit away from an integer-valued code length function.[1] A final modification of the notion of a code length function is purely technical and a matter of mathematical convenience. It consists in changing the base for logarithms and exponentiation from 2 to e.

With the above remarks in mind, we now define the set $K(\mathbb{A})$ of *idealized code-length functions* or, as we shall simply say in the sequal, of *codes*, as the set of mappings $\kappa: \mathbb{A} \to [0, \infty]$ such that

$$\sum_{a \in \mathbb{A}} e^{-\kappa(a)} = 1.$$

If $\kappa \in K(\mathbb{A})$ and $P \in M_+^1(\mathbb{A})$, we say that (κ, 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 κ is *adapted to* P, or that P is the distribution *matching* κ.

As above, we use $\langle \cdot, P\rangle$ to denote mean value with respect to (w.r.t.) P, and we use $H = H(\cdot)$ to denote entropy and $D = D(\cdot\|\cdot)$ to denote information divergence. For any $P \in M_+^1(\mathbb{A})$ and any $\kappa \in K(\mathbb{A})$

$$\langle \kappa, P\rangle = H(P) + D(P\|Q) \tag{2}$$

where Q is the distribution matching κ. This is the *linking identity*.

A slight variation of concepts is often natural. If P is a distribution and κ a code, we introduce the *redundancy of P given κ*, or the *redundancy of κ assuming P*, which may be thought of as the unavoidable redundancy which results from using κ in order to code events which are governed by the true distribution P. This quantity is denoted by $D(P\|\kappa)$ and defined to be equal to $D(P\|Q)$ where Q is the distribution matching κ. Thus, we may rewrite the linking identity in the form

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

By the *usual topology* on $M_+^1(\mathbb{A})$ we shall mean the topology of pointwise convergence and topological notions such as closure, continuity, and semicontinuity are understood to be with respect to this topology. For instance, the entropy function $P \curvearrowright H(P)$ is continuous if $\mathbb{A}$ is finite but only lower semicontinuous for an infinite alphabet. Note that the usual topology is metrizable by total variation. This follows from Scheffé's theorem, cf. [4]. We use $V(P, Q)$ to denote the *total variation* between P and Q, i.e., $V(P, Q) = \sum_i |p_i - q_i|$, and we write $P_n \xrightarrow{V} P$ if $(P_n)_{n\geq 1}$ converges in total variation to P.

A sequence $(P_n)_{n\geq 1} \subseteq M_+^1(\mathbb{A})$ *converges in divergence* to $P \in M_+^1(\mathbb{A})$ if $D(P_n\|P) \to 0$. We express this by writing $P_n \xrightarrow{D} P$. Convergence in divergence is stronger than convergence in total variation as follows from Pinskers inequality: $D(P\|Q) \geq \frac{1}{2}V(P, Q)^2$. At times we find it convenient to say that P_n *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.

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 *information space* is a pair $(\mathbb{A}, \mathcal{P})$, where

[1] If $\kappa: \mathbb{A} \to [0, \infty]$ satisfies $\sum_{i \in \mathbb{A}} 2^{-\kappa_i} \leq 1$ and we put $\kappa^* = \lceil \kappa \rceil$ then $\langle \kappa, P\rangle \leq \langle \kappa^*, P\rangle \leq \langle \kappa, P\rangle + 1$ for all $P \in M_+^1(\mathbb{A})$, and there exists a binary prefix-free code with κ^* as code length function.

$\mathbb{A}$—the alphabet as above—is a countable set and $\mathcal{P}$ is an arbitrary subset of $M_+^1(\mathbb{A})$. We shall mostly use the relatively neutral terminology *model* for the set $\mathcal{P}$. If you have applications to quantum physics in mind, it would be better to call $\mathcal{P}$ the *preparation space*—and distributions in $\mathcal{P}$ *individual preparations*—whereas, if you think in terms of statistical concepts, it would be natural to refer to $\mathcal{P}$ as a *statistical model* and, perhaps, to parametrize the distributions in $\mathcal{P}$. The concept has of course been studied extensively in one form or another. The view which we favor is forcefully put forward by Jaynes, cf., e.g., [15], where he stresses the distinction between distributions as the "truth" about "reality" and as a means of expressing our *knowledge* about reality.

The distributions in $\mathcal{P}$ are referred to as *consistent distributions*. A distribution $P \in M_+^1(\mathbb{A})$ is *essentially consistent* if there exists a sequence of consistent distributions which converges to P in divergence.

We shall exploit a game, the *code-length game*, which is closely related to the *maximum entropy principle*. This game was introduced by the author in [24], cf. also [25], and is defined as the two-person zero-sum game with *code length*, which maps $(\kappa, P) \in K(\mathbb{A}) \times \mathcal{P}$ into $\langle \kappa, P \rangle$, as cost function. In more detail, the set $\mathcal{P}$ is the strategy set for *the system* ("Player I") and $K(\mathbb{A})$ the strategy set for *the observer* ("Player II"). It is the objective of the observer to minimize 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 κ 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 (2), 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* $H_{\max} = H_{\max}(\mathcal{P})$. We also refer to the game as the $H_{\max}/R_{\min}$-game.[2]

Clearly, $H_{\max} \le R_{\min}$. 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 κ with $R(\kappa) = R_{\min}$. A *maximum entropy distribution* ($H_{\max}$*-distribution*) is an essentially consistent distribution P with $H(P) = H_{\max}$. We emphasize that a maximum entropy distribution is only required to be essentially consistent, not necessarily consistent. The results to follow—and comments in Section VII—constitute arguments in favor of this departure from usual practice. In our terminology, the usual concept is a consistent $H_{\max}$-distribution which, in game-theoretical terms, is the same as an optimal strategy for the system.

Further concepts are important. First, a sequence $(P_n)_{n \ge 1}$ of consistent distributions is *asymptotically optimal* if $H(P_n) \to H_{\max}$ and, second, $P^* \in M_+^1(\mathbb{A})$ is the *maximum-entropy attractor* (the $H_{\max}$*-attractor*) if $P_n \xrightarrow{D} P^*$ for every asymptotically optimal sequence $(P_n)_{n \ge 1}$. 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 $H(P^*) \le H_{\max}$. Therefore, it must be the unique $H_{\max}$-distribution if $H(P^*) = H_{\max}$.

Basic information about the $H_{\max}/R_{\min}$-game is contained in the following result which may be derived directly from [24, Theorems 1–3], cf. also [25, Theorem 2]. Note the use of "co" for "convex hull."

Theorem 2: The information space $(\mathbb{A}, \mathcal{P})$ is in equilibrium if and only if $H_{\max}(\mathrm{co}\,(\mathcal{P})) = H_{\max}(\mathcal{P}) < \infty$. If this condition is fulfilled, there exists a unique minimum risk code κ^* as well as a, likewise unique, maximum entropy attractor, P^*, and (κ^*, P^*) is a matching pair.

In particular, if the condition of the theorem holds then there is a unique distribution to which any attempt of finding a maximum-entropy distribution must converge, even in a rather strong sense. Though Theorem 2 is sufficient for most purposes, the existence of the $H_{\max}$-attractor can be established under weaker conditions, cf. Section VII.

For a model in equilibrium, we refer to the matching pair, the existence of which is ensured by Theorem 2, as the *optimal matching pair* associated with the model.

We warn the reader that in Theorem 2, the equality $H(P^*) = H_{\max}$ need not hold, thus the maximum-entropy distribution may not exist. In the more typical case when $H(P^*) = H_{\max}$ does hold, we say that the model is *entropy-continuous*. Any model with a finite alphabet $\mathbb{A}$ is entropy-continuous by continuity of the entropy function. In the case of an infinite alphabet, the entropy function is only lower semicontinuous. Thus, for a convergent sequence $P_n \xrightarrow{V} P$, we can only assert that $\liminf_{n \to \infty} H(P_n) \ge H(P)$. This is why we can only conclude that the inequality $H(P^*) \le H_{\max}$ holds in Theorem 2.

II. Criteria for Optimality

Theorem 2 is an existence result and does not give much of a clue as to how one finds the optimal 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, borrowed from mathematical economics, cf. [1], for example, turns out to be particularly useful. The code κ^* is the *Nash equilibrium code* for $(\mathbb{A}, \mathcal{P})$ if the distribution P^* which matches κ^* is essentially consistent and $R(\kappa^*) = H(P^*) < \infty$. In the two theorems to follow, we shall see that the Nash equilibrium code is unique and that, typically, the Nash equilibrium code *does* exist. Note that, in principle, it is possible to check if a code is a Nash equilibrium code without knowing $H_{\max}$ or $R_{\min}$, whereas a direct check if a given distribution is the $H_{\max}$-attractor or a $H_{\max}$-distribution requires that $H_{\max}$ be known.

For a number of naturally occurring models, the Nash equilibrium code is also *stable*, i.e., $\langle \kappa^*, P \rangle$ is finite and independent of P for every consistent distribution P (cf. [25]). There may be many stable codes. If a stable code has a consistent-matching distribution, it must be the Nash equilibrium code. Often, the Nash equilibrium code can be found in this way, i.e., by first searching for stable codes—Section IV contains some illustrative examples of this approach. We stress that the Nash equilibrium code need not be stable and also, it may have an inconsistent matching distribution.

Generalizing [25, Theorem 2] we obtain the following.

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

$$H(P) + D(P \| P^*) \le H_{\max}(\mathcal{P}) \tag{3}$$

$$R_{\min}(\mathcal{P}) + D(P^* \| \kappa) \le R(\kappa). \tag{4}$$

[2] 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.

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

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

It follows that the sequence $(H(P_n))_{n\geq 1}$ is convergent and that

$$\lim_{n\to\infty} H(P_n) = R(\kappa^*) = H(P^*).$$

In particular

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

As $H_{\max} \leq R_{\min}$ always holds, $H(P^*) = H_{\max} = R_{\min} = R(\kappa^*)$. Thus, $\mathcal{P}$ is in equilibrium, (P_n) is asymptotically optimal, κ^* is a minimum risk code, and P^* a maximum entropy distribution.

If κ is any code, then

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

where, in the last step, we used the lower semicontinuity of $D(\cdot\|\kappa)$ and the fact that $P_n \xrightarrow{D} P^*$, hence $P_n \xrightarrow{V} P^*$. Thus, (4) holds and κ^* is the unique minimum risk code (uniqueness because $D(P^*\|\kappa) = 0$ implies that κ is the code adapted to P^*).

For $Q \in \mathcal{P}$

$$H(Q) + D(Q\|P^*) = \langle \kappa^*, Q\rangle \leq R(\kappa^*) = H_{\max}$$

thus (3) holds. Therefore, P^* is the $H_{\max}$-attractor as well as the unique maximum entropy distribution (uniqueness because $D(Q\|P^*) = 0$ implies $Q = P^*$). □

The proof shows that if the Nash equilibrium code exists and $(P_n)_{n\geq 1}$ is a sequence of consistent distributions, then the conditions that $(P_n)_{n\geq 1}$ converges in divergence to P^* and that $(P_n)_{n\geq 1}$ is asymptotically optimal are equivalent.

The theorem points to a possible approach in the search for the optimal matching pair in cases when a search for stable codes does not lead to the goal. This approach is illustrated by examples in Sections V and VI.

If $(\mathbb{A}, \mathcal{P})$ is in equilibrium and entropy-continuous, any asymptotically optimal sequence of distributions does of course converge in entropy to the $H_{\max}$-attractor. This points to the information spaces which are in equilibrium and entropy-continuous 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 (κ^*, P^*) the optimal matching pair. Then the following conditions are equivalent.

i) $(\mathbb{A}, \mathcal{P})$ is entropy-continuous.

ii) $(\mathbb{A}, \mathcal{P})$ has a $H_{\max}$-distribution (necessarily P^*).

iii) $(\mathbb{A}, \mathcal{P})$ has a Nash equilibrium code (necessarily κ^*).

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

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

For our last theoretical result, we point out that any result which asserts the existence of the $H_{\max}$-attractor can be viewed as a limit theorem. In what follows, we further emphasize this aspect (note the use of "$\overline{\mathrm{co}}$" for "closed convex hull").

Theorem 5: Let $(\mathbb{A}, \mathcal{P}_n)_{n\geq 1}$ be a sequence of information spaces and assume that they are all in equilibrium, say with $H_{\max}$-attractors $P_n^*; n \geq 1$. Assume that $\sup_{n\geq 1} H_{\max}(\mathcal{P}_n) < \infty$ and that the models are nested in the sense that $\mathrm{co}(\mathcal{P}_1) \subseteq \mathrm{co}(\mathcal{P}_2) \subseteq \cdots$.

Then all models $\mathcal{P}$ with

$$\bigcup_{n\geq 1}\mathcal{P}_n \subseteq \mathcal{P} \subseteq \overline{\mathrm{co}}\left(\bigcup_{n\geq 1}\mathcal{P}_n\right) \tag{5}$$

are in equilibrium and have the same $H_{\max}$-attractor, P^*. Furthermore, $P_n^* \xrightarrow{V} P^*$ and, in case all models $\mathcal{P}_n$ are entropy-continuous, convergence even takes place in divergence: $P_n^* \xrightarrow{D} P^*$.

Proof: Put $h = \sup\limits_{n\geq 1} H_{\max}(\mathcal{P}_n)$. Then, for any $\mathcal{P}$ satisfying (5)

$$\begin{aligned} h &\leq H_{\max}(\mathcal{P}) \leq H_{\max}(\mathrm{co}(\mathcal{P})) \leq H_{\max}\left(\overline{\mathrm{co}}\bigcup_{n\geq 1}\mathcal{P}_n\right) \\ &= H_{\max}\left(\mathrm{co}\bigcup_{n\geq 1}\mathcal{P}_n\right) = H_{\max}\left(\bigcup_{n\geq 1}\mathrm{co}(\mathcal{P}_n)\right) \\ &= \sup_{n\geq 1} H_{\max}(\mathrm{co}(\mathcal{P}_n)) = h \end{aligned}$$

where the first equality follows by lower semicontinuity of the entropy function. By Theorem 2, we now see that $H_{\max}(\mathcal{P}) = h$ and that $\mathcal{P}$ is in equilibrium.

Again, let $\mathcal{P}$ satisfy (5) and let P^* be the $H_{\max}$-attractor of $\mathcal{P}$. We shall prove that P_n^* converges to P^* in total variation. This will show that the attractor is independent of $\mathcal{P}$ as long as $\mathcal{P}$ satisfies (5). For each $n \geq 1$, choose $P_n \in \mathcal{P}_n$ such that $H(P_n) \geq H_{\max}(\mathcal{P}_n) - \frac{1}{n}$ and such that $V(P_n, P_n^*) \leq \frac{1}{n}$. Then, $(P_n)_{n\geq 1}$ is asymptotically optimal for $\mathcal{P}$, hence $P_n \xrightarrow{D} P^*$, in particular, $P_n \xrightarrow{V} P^*$. Clearly then, $P_n^* \xrightarrow{V} P^*$.

In case all the $\mathcal{P}_n$ are entropy-continuous, we consider a closed model $\mathcal{P}$ satisfying (5). Then $(P_n^*)_{n\geq 1}$ is asymptotically optimal for $\mathcal{P}$ and $P_n^* \xrightarrow{D} P^*$ follows. □

III. Some Classical Models and Associated Distributions

We shall study some of the classical distributions based on information-theoretical considerations. Without being comprehensive we mention earlier research in this direction: [19], [5], [8], and [2]. However, our approach is also based on games. The 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 $H_{\max}/R_{\min}$-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 $\mathbb{A}$ and the *uniform distribution* over $\mathbb{A}$ which we define as the maximum entropy distribution for $\mathcal{P} = M_+^1(\mathbb{A})$. 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

$$\mathbb{A}_n = \{0, 1, 2, \ldots, n\}, \qquad n \geq 1$$

and

$$\mathbb{A}^* = \{0, 1, 2, \ldots\}.$$

We now use $E(P)$ for the mean value of a random variable with distribution P.

For $0 \leq \lambda \leq n$, $B_n(\lambda) \subseteq M_+^1(\mathbb{A}_n)$ is the set of distributions of sums of n independent Bernoulli variables for which the sum has mean value λ. Recall that Bernoulli variables are random variables that can only assume the two values, 0 and 1. Note that we do not require that the Bernoulli variables are identically distributed, only that they are independent.

Further, $G_n(\lambda)$ is the set of all $P \in M_+^1(\mathbb{A}_n)$ with mean value λ: $E(P) = \lambda$. Using the natural embedding of the sets $M_+^1(\mathbb{A}_n)$ in $M_+^1(\mathbb{A}^*)$, 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 $0 \leq p \leq 1$

$$B_1(p) = G_1(p) = \{\mathrm{BIN}(1, p)\}$$

$\mathrm{BIN}(1, p)$ denoting the Bernoulli distribution with parameter (success probability) p.

By $B_\infty(\lambda)$ we denote the set of distributions in $M_+^1(\mathbb{A}^*)$ of infinite sums of independent $\mathrm{BIN}(1, p_n)$-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 $P \in M_+^1(\mathbb{A}^*)$ with mean value λ. We shall use the notation $X(I)$, where X could stand for B_n, B^*, B_∞, G_n, G^*, or G_∞ 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 $P \in M_+^1(\mathbb{A}^*)$ with mean value at most λ.

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

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

IV. 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 $\mathbb{A}^*$ determined by the equation

$$P_x(k) = P_x(0) \cdot x^k, \qquad k \geq 0. \tag{6}$$

The matching codes κ_x are given by

$$\kappa_x(k) = -\ln P_x(0) - k \ln x, \qquad k \geq 0 \tag{7}$$

and are, therefore, 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

$$P_x(0) = \frac{1}{1+\lambda}, \qquad x = \frac{\lambda}{1+\lambda}. \tag{8}$$

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

$$\begin{aligned} H_{\max}(G_\infty(\lambda)) &= H_{\max}(G_\infty([0, \lambda])) \\ &= \ln(1+\lambda) + \lambda \ln\left(1 + \frac{1}{\lambda}\right). \end{aligned} \tag{9}$$

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

$$P_{n,x}(k) = P_{n,x}(0) \cdot x^k, \qquad 0 \leq k \leq n. \tag{10}$$

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

$$\kappa_{n,x}(k) = -\ln P_{n,x}(0) - k \ln x, \qquad 0 \leq k \leq n \tag{11}$$

and are, therefore, 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 $x \curvearrowright E(P_{n,x})$ is strictly increasing in x, a fact that also follows from continuity of this map and from Theorems 2 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 $\mathrm{GEO}(n, \lambda)$ has been identified as the distribution $P_{n,x}$. It may be noted that for $0 \leq x \leq \infty$, $x \neq 1$

$$\begin{aligned} E(P_{n,x}) &= \frac{x}{1-x} - (n+1)\frac{x^{n+1}}{1-x^{n+1}} \\ &= n - \left(\frac{n+1}{1-x^{n+1}} - \frac{1}{1-x}\right). \end{aligned} \tag{12}$$

By (11), 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 Nash equilibrium 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 6: For fixed n and $0 \leq \lambda \leq n$, $G_n(\lambda)$ is in equilibrium and the $H_{\max}$-distribution, $\mathrm{GEO}(n, \lambda)$, is well defined and characterized as the distribution in $G_n(\lambda)$ determined by (10). If $\lambda \leq n/2$, this distribution is also the $H_{\max}$-distribution of $G_n([0, \lambda])$ and if $n/2 \leq \lambda \leq n$, it is the $H_{\max}$-distribution of $G_n([\lambda, n])$.

For $0 \leq \lambda < \infty$, the model $G_\infty(\lambda)$ is in equilibrium and the $H_{\max}$-distribution, $\mathrm{GEO}(\lambda)$, is well defined and characterized by (6) and (8). This distribution is also the $H_{\max}$-distribution of any of the models $G_\infty([0, \lambda])$, $G^*(\lambda)$, and $G^*([0, \lambda])$. The maximum entropy value $H_{\max}$ is given by (9).

The models considered are entropy-continuous and, for $0 \leq \lambda < \infty$, the distributions $\mathrm{GEO}(n, \lambda)$ converge in divergence as well as in entropy to $\mathrm{GEO}(\lambda)$.

V. The Binomial and Poisson Distributions

In this section, we agree to use P_X to denote the distribution of X, whether X is a random variable or a random vector. The key to the results of this section is a combination of our game-theoretical results with an inequality due to Hoeffding, cf. [12, Theorem 3]. We begin with a statement of the inequality we need. For the convenience of the reader, we also include a brief proof. Note the use of "$*$" for "convolution."

Theorem 7 (Hoeffding's Inequality): Let $P_1, P_2, \ldots, P_n$ be distributions in $M_+^1(\mathbb{A}^*)$ and put $\overline{P} = \frac{1}{n}\sum_1^n P_k$. Then, in case $P_1, P_2, \ldots, P_n$ are all supported by $\{0, 1\}$, the inequality

$$\langle g, P_1 * P_2 * \cdots * P_n \rangle \leq \left\langle g, \overline{P}^{*n} \right\rangle \tag{13}$$

holds for any "integer convex" function g: $\mathbb{A}^* \to \mathbb{R}$, i.e., for any function $k \curvearrowright g_k$ such that $2g_{k+1} \leq g_k + g_{k+2}$ for $k \in \mathbb{A}^*$.

Proof: Let X_k; $1 \le k \le n$ be independent P_k-distributed random variables and put $S_n = \sum_1^n X_k$. Put $p_k = P_k(1)$, $1 \le k \le n$. Fix $\lambda = \sum_1^n p_k$ and, for a while, also $p_3, \ldots, p_n$. Put $2\alpha = \lambda - \sum_3^n p_k$. Thus, for an x with $|x| \le \alpha$, $p_1 = \alpha - x$ and $p_2 = \alpha + x$. Put $S' = X_1 + X_2$ and $S'' = \sum_3^n X_k$. For each ν, one can split the probability $P(S_n = \nu)$ into three terms according to the value of S'. This then leads to the following expression for $\langle g, P_{S_n}\rangle$:

$$\langle g, P_{S_n}\rangle = c - x^2 \sum_{\nu=0}^{n-2} (g_\nu - 2g_{\nu+1} + g_{\nu+2}) P(S'' = \nu)$$

with

$$c = \sum_{\nu=0}^{n-2} \big((1-\alpha)^2 g_\nu + 2\alpha(1-\alpha) g_{\nu+1} + \alpha^2 g_{\nu+2}\big) P(S'' = \nu).$$

By our reasoning, c is to be considered as a constant. Therefore, the convexity assumption shows that $\langle g, P_{S_n}\rangle$ is maximal for $x = 0$, i.e., for $p_1 = p_2$. Here, $p_3, \ldots, p_n$ were fixed. Repeating the argument with other values fixed we realize that as long as $\lambda = \sum_1^n p_k$ is kept fixed, $\langle g, P_{S_n}\rangle$ is largest when all the p_i's are equal. The result follows. □

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

$$P^*(k) = \binom{n}{k} p^k q^{n-k}, \qquad 0 \le k \le n \tag{14}$$

and κ^* the matching code

$$\kappa^*(k) = -\ln \binom{n}{k} - k \ln p - (n-k) \ln q, \qquad 0 \le k \le n. \tag{15}$$

As is well known and classical, $P^* \in B_n(\lambda)$. We shall show that κ^* is the Nash equilibrium 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 $\mathrm{BIN}(n, p)$. It was proved independently by Mateev [21] and by Shepp and Olkin [23], cf. also Marshall and Olkin [20], that P^* is indeed the $H_{\max}$-distribution. For a recent treatment, see [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

$$\langle \kappa^*, P\rangle \le \langle \kappa^*, P^*\rangle \tag{16}$$

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

$$\boldsymbol{\kappa}^*(k_1, k_2) = -\ln n! + \ln k_1! + \ln k_2! - k_1 \ln p - k_2 \ln q \tag{17}$$

where $0 \le k_1 \le n$ and $k_1 + k_2 = n$. Then (16) is equivalent to the inequality

$$\langle \boldsymbol{\kappa}^*, \boldsymbol{P}\rangle \le \langle \boldsymbol{\kappa}^*, \boldsymbol{P}^*\rangle. \tag{18}$$

We find that

$$\langle \boldsymbol{\kappa}^*, \boldsymbol{P}\rangle = -\ln n! - nH(p, q) + \langle \ln k!, P_{S_n}\rangle + \langle \ln k!, P_{T_n}\rangle. \tag{19}$$

By Theorem 7, each of the two averages here is maximized when the underlying probabilities $p_1, \ldots, p_n$ are equal. Thus, (18), hence also (16), hold.

Now fix $\lambda > 0$ and consider the model $B^*(\lambda)$. It is convenient also to consider the model

$$\mathrm{co}(B^*(\lambda)) = \bigcup_{n \ge \lambda} \mathrm{co}(B_n(\lambda)).$$

As $\mathrm{co}(B^*(\lambda)) \subseteq G^*(\lambda)$

$$H_{\max}(B^*(\lambda)) \le H_{\max}(\mathrm{co}(B^*(\lambda))) < \infty.$$

Then Theorem 5 applies. It follows that $\mathrm{BIN}(n, \lambda/n)$ converges in divergence to the $H_{\max}$-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

$$P^*(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k \ge 0.$$

We can now summarize the findings.

Theorem 8: The $H_{\max}$-distribution for the models $B_n(\lambda) = B_n(np)$ is the classical binomial distributions $\mathrm{BIN}(n, p)$, and the $H_{\max}$-distribution for the models $B^*(\lambda)$, $B_\infty(\lambda)$ and $\overline{\mathrm{co}}(B_\infty(\lambda))$ is the classical Poisson distribution $\mathrm{POI}(\lambda)$. The models considered are entropy-continuous and, for each $\lambda \ge 0$, $\mathrm{BIN}(n, \lambda/n)$ converges in total variation, in divergence as well as in entropy to $\mathrm{POI}(\lambda)$.

For the convergence in entropy we refer the reader to [9, Theorem 8] where property v) of Theorem 4 is verified.[3]

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 9: For fixed n, $\mathrm{BIN}(n, \frac{1}{2})$ is the unique maximum-entropy distribution among all binomial distributions $\mathrm{BIN}(n, p)$, $0 \le p \le 1$.

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

Theorem 10: For $n \ge 2$, consider the model of all binomial distributions $\mathrm{BIN}(n, p)$; $0 \le p \le 1$. Then

$$H_{\max} = H(\mathrm{BIN}(n, 1/2)) < R_{\min} = \ln(n+1).$$

Proof: Let $\mathcal{P}$ denote the model in question and consider $\mathrm{co}(\mathcal{P}) = \overline{\mathrm{co}}(\mathcal{P})$. We base the proof on the general equalities

$$R_{\min}(\mathcal{P}) = R_{\min}(\mathrm{co}(\mathcal{P})) = H_{\max}(\mathrm{co}(\mathcal{P})). \tag{20}$$

The first equality follows directly from the definitions of risk and minimum risk, and the second equality is part of Theorem 2. The fact that $H_{\max}(\mathrm{co}(\mathcal{P})) = \ln(n+1)$ follows as the uniform distribution over $\{0, 1, 2, \ldots, n\}$ belongs to $\mathrm{co}(\mathcal{P})$. Indeed, this distribution is the uniform mixture over $x \in [0, 1]$ of the binomial distributions $\mathrm{BIN}(n, x)$, a direct consequence of the classical formula for the beta function since, by that formula

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

which equals $\frac{1}{n+1}$ for all $0 \le k \le n$. □

[3] In fact, a simple proof—which, however, relies on more theory—amounts to a check that the Poisson distribution is not "hyperbolic," cf. Harremoës and Topsøe [10, Theorem 8.4]. Intuitively, the requirement is that the tails must not be too large.

VI. 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 $\mathbb{N} = \{1, 2, \ldots\}$ (representing the various levels of "success").

Denote by Ω^n the set of infinite vectors $\boldsymbol{k} = (k_1, k_2, \ldots)$ with the integers k_i all nonnegative and $\sum_1^\infty k_i = n$, and denote by Ω^* the set of similar vectors but with the looser requirement $\sum_1^\infty k_i < \infty$. Then Ω^* is countable and decomposed into the sets $\Omega^0, \Omega^1, \ldots$.

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

Consider the model $\mathcal{P} = \mathcal{P}(n, Q)$ constructed as follows. For each finite set $P_1, \ldots, P_n$ in $M_+^1(\mathbb{N})$ with $\overline{P} = Q$ we consider independent random variables $X_1, \ldots, X_n$ such that X_ν has distribution P_ν; $1 \le \nu \le n$ and then we consider the random vector $\boldsymbol{S}_n = (S_{n1}, S_{n2}, \ldots)$ where S_{ni} denotes the number of $1 \le \nu \le n$ with $X_\nu = i$; $i = 1, 2, \ldots$. By definition, the model $\mathcal{P}$ consists of all distributions of random vectors $\boldsymbol{S}_n$ that arise in this way. A typical element of $\mathcal{P}$ is denoted $\boldsymbol{P}$, i.e., $\boldsymbol{P} = P_{\boldsymbol{S}_n}$. Since $\boldsymbol{P} \in \mathcal{P}$ depends on $P_1, \ldots, P_n$ (with $\overline{P} = Q$), we may write $\boldsymbol{P} = \boldsymbol{P}(P_1, \ldots, P_n)$.

Theorem 11: If $H(Q) < \infty$, then $\mathcal{P}(n, Q)$ is in equilibrium and entropy-continuous. The maximum entropy distribution is

$$\boldsymbol{P}_0 = \boldsymbol{P}(Q, \ldots, Q).$$

Proof: Let $\boldsymbol{\kappa}_0$ be the code adapted to $\boldsymbol{P}_0$ and let $q_i, i \ge 1$ be the point probabilities for Q. Then $\boldsymbol{P}_0$ is given by

$$\boldsymbol{P}_0(k_1, k_2, \ldots) = n! \prod_{i=1}^\infty \frac{q_i^{k_i}}{k_i!} \tag{21}$$

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

$$\boldsymbol{\kappa}_0(k_1, k_2, \ldots) = -\ln n! - \sum_{i=1}^\infty k_i \ln q_i + \sum_{i=1}^\infty \ln k_i!. \tag{22}$$

In order to establish the theorem, we shall prove that $\boldsymbol{\kappa}_0$ is the Nash equilibrium code, i.e., for any $\boldsymbol{P} = P_{\boldsymbol{S}_n} \in \mathcal{P}$ with $\boldsymbol{S}_n = (S_{n1}, S_{n2}, \ldots)$ we shall prove that $\langle \boldsymbol{\kappa}_0, \boldsymbol{P}\rangle$ is maximal for $P_1 = \cdots = P_n = Q$. By (22)

$$\begin{aligned}\langle \boldsymbol{\kappa}_0, \boldsymbol{P}\rangle &= -\ln n! - \sum_{i=1}^\infty nq_i \ln q_i + \sum_{i=1}^\infty E(\ln S_{ni}!)\\ &= -\ln n! + nH(Q) + \sum_{i=1}^\infty E(\ln S_{ni}!).\end{aligned}$$

From the investigation in Section V, we realize that the individual expectations $E(\ln S_{ni}!)$ are upper-bounded by the corresponding expectations when all ith-point probabilities of $P_1, \ldots, P_n$ are equal. Thus,

$$\begin{aligned}\langle \boldsymbol{\kappa}_0, \boldsymbol{P}\rangle &\le -\ln n! + nH(Q) + \sum_{i=1}^\infty \sum_{k=0}^n \ln k! \binom{n}{k} q_i^k (1-q_i)^{n-k}\\ &= -\ln n! + nH(Q) + \sum_{k=2}^n \ln k! \binom{n}{k} \sum_{i=1}^\infty q_i^k (1-q_i)^{n-k}\\ &= H(\boldsymbol{P}_0).\end{aligned}$$

As we also find that $H(\boldsymbol{P}_0) < \infty$, the proof is complete. □

Let us agree that $H_{\max}(n, Q)$ denotes the maximum entropy value of the model $\mathcal{P}(n, Q)$. Then $H_{\max}(n, Q)$ 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 11.

In [23, Theorem 3], a concavity result was proved for the entropy of multinomial distributions. In what follows, we generalize this result to one concerning the $H_{\max}$-distributions of the models $\mathcal{P}(n, Q)$ for distributions which are not required to have finite support. The computations needed for the proof are the same as those given in [23] and are, therefore, only briefly indicated.

Theorem 12: For all $n \ge 1$, the map $Q \curvearrowright H_{\max}(n, Q)$ is strictly concave, in fact, for any infinite convex combination $\Sigma_1^\infty \alpha_\nu Q_\nu$ of measures in $M_+^1(\mathbb{N})$

$$H_{\max}\left(n, \sum_{\nu=1}^\infty \alpha_\nu Q_\nu\right) \ge \sum_{\nu=1}^\infty \alpha_\nu H_{\max}(n, Q_\nu) \tag{23}$$

and, if the right-hand side is finite, the inequality is strict unless all Q_ν with $\alpha_\nu > 0$ are identical.

Proof: By Theorem 11 and its proof

$$H_{\max}(n, Q) = -\ln n! + \sum_{i=1}^\infty f_n(q_i) \tag{24}$$

with the functions $f_n\colon [0, 1] \to \mathbb{R}_+$ defined by

$$f_n(q) = -nq \ln q + \sum_{k=0}^n \ln k! \cdot \binom{n}{k} q^k (1-q)^{n-k}. \tag{25}$$

A straightforward 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 $f_n''(q) < -\frac{n}{q}(1-q)^{n-1}$, hence each function f_n is strictly concave and the result follows by (24). □

For $n = 1$, Theorem 12 reduces to the usual concavity property of the entropy function.

Finally, let us consider the model $\mathcal{P}(n, Q)$ in case Q has finite support. Then the $H_{\max}$-distribution, identified in Theorem 11 as the empirical distribution $\boldsymbol{P}(Q, \ldots, Q)$ is nothing but the multinomial distribution determined by Q, denoted by $\mathrm{MULT}(n, Q)$, say. We may now combine the concavity of $H_{\max}(n, \cdot)$ established in Theorem 12 and the obvious symmetry of this function with Theorem 11. In this way, we obtain the following result, generalizing parts of Theorems 8 and 9.

Theorem 13: Let $r \ge 2$ and $n \ge 2$ 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 $\mathrm{MULT}(n, Q)$ 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.

VII. Discussion

A. Theory

This correspondence and previous research demonstrates that the $H_{\max}/R_{\min}$-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 twofold: Identification of interesting distributions and associated limit theorems, possibly accompanied by certain inequalities, facilitated by, e.g., (3) and (4). Future extensions may involve extra structure, say in the form of Markov kernels, side information, or symmetry.

Theorem 3, which was overlooked in [25], is a key result. Note that the proof given does not depend on previous results. Therefore, this correspondence is largely self-contained. Previous research consists of the author's papers [24] and [25], and then, we point to [17] and, for games covering also the continuous case, [11]. In [10], further results, actually developed after the first submission of the present manuscript, can be found.

Regarding the controversial definition of an $H_{\max}$-distribution, requiring besides maximal entropy only essential consistency, note that it follows from Theorem 5 (with $\mathcal{P}_n$'s independent of n) that it does not really matter if, to a given model, you add the essentially consistent distributions or even all distributions in the closure of the model. And if you do that, the normally accepted definition is of course all you need. On the other hand, it is awkward only to work with models which are closed. Indeed, the most frequently studied models are those given by moment constraints and, typically, these models are not closed.

A fundamental phenomenon is the possibility that the equality $H(P^*) = H_{\max}$ may not hold for any consistent or essentially consistent distribution. As examples show, cf. [13], [24] or, for a more conclusive study, [10], this situation may occur.

This also explains the importance of the $H_{\max}$-attractor as it does allow for a discontinuity or loss of entropy $(H(P^*) < H_{\max})$ and yet, if a maximum-entropy distribution exists, this is the one to search for. Originally, the notion of $H_{\max}$-attractor was defined differently, cf. [24], requiring only normal convergence (i.e., convergence in total variation) rather than convergence in divergence for asymptotically optimal sequences. But as the stronger convergence property does, in fact, hold for the main category of models—those in equilibrium—and as convergence in divergence appears to be the right kind of convergence to work with for information-theoretical investigations, the chosen definition appears justified.

If we turn our 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 [7] for an illuminating discussion.

Further comments on continuity considerations concern Theorem 5. In the case of entropy-continuous models $\mathcal{P}_n$, we cannot in general assert that the limit model $\mathcal{P}$ is also entropy-continuous, thus, P_n^* may not converge in entropy to P^*. An example to illustrate this point can be extracted from [24, Theorem 21, case (d)] (start with the limit model and consider approximating models by restricting the support of the distributions to larger and larger finite sets). Another comment is that probably $P_n \xrightarrow{D} P$ does not hold generally in Theorem 5. However, the author is not aware of an example to illustrate this.

One may consider the concept of $H_{\max}$-attractor as a key object of study, quite independently of the game-theoretical setting. In this connection, we emphasize that existence of the $H_{\max}$-attractor can be established in a number of cases not covered by Theorem 2. Simple examples with a finite alphabet point to this (consider, for example, the model consisting only of the two distributions with point masses $(1, 0)$ and $(\frac{1}{4}, \frac{3}{4})$, respectively).

In order to state more general existence results for the $H_{\max}$-attractor, we introduce the notation

$$h = H_{\max}(\mathcal{P}) \quad \text{and} \quad \mathcal{P}_t = \{P \in \mathcal{P} | H(P) > t\}, \qquad \text{for } t < h.$$

It follows from Theorem 2 that if there exists $t < h$ such that $H_{\max}(\text{co}\,(\mathcal{P}_t)) \le h$ (in which case equality must hold), then the $H_{\max}$-attractor exists. If we weaken the condition and assume only that

$$\lim_{t \to h} H_{\max}(\text{co}(\mathcal{P}_t)) \le h \tag{26}$$

the existence of a "weak" $H_{\max}$-attractor P^* follows, one for which $P_n \xrightarrow{V} P^*$ for every asymptotically optimal sequence $(P_n)_{n \ge 1}$. A strengthening of this result to one asserting the existence of the usual "strong" $H_{\max}$-attractor (one with $P_n \xrightarrow{D} P^*$) is not possible, even for a finite alphabet, as simple examples will show. For finite $\mathbb{A}$, (26) is also necessary for the existence of a weak $H_{\max}$-attractor.

B. Applications

The type of problems treated in Section IV are well known, even classical. This also concerns more general models defined by moment conditions. From the reference list we may quote [14], [15], [13], [24], and [25], but there are many other sources from physics, chemistry, statistics (exponential families), and information theory. Reference [16] contains a comprehensive bibliography. The game-theoretical approach, however, is not standard. It leads more directly to an understanding of such models than other approaches (typically based on optimization via the introduction of Lagrange multipliers). The search for stable codes is instrumental in this respect.

Regarding Section V, we again stress the game-theoretical treatment. For a more direct approach, see [9] where one also finds detailed approximations relating binomial distributions to the limiting Poisson distribution.

We also want to emphasize the conjecture, going back to [23],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 Bernoulli sum of independent Bernoulli variables with success probabilities, respectively, $p_1, \ldots, p_n$. For partial results in this direction, see [23] and [9].

In Section VI, 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 λ and n with $0 \le \lambda \le n$ 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 [21, Corollary 2]) and to Shepp and Olkin [23], was an important part of Theorem 8. It is equivalent to the following inequality (with notation as in Theorem 7):

$$H\left(\overline{P}^{*n}\right) \ge H(P_1 * P_2 * \cdots * P_n) \tag{27}$$

valid for distributions $P_1, P_2, \ldots, P_n$ which are all supported by $\{0, 1\}$. To realize the stated equivalence, simply note that all P_k are of the form $P_k = \text{BIN}(1, p_k)$, that $\overline{P} = \text{BIN}(1, \overline{p})$ with $\overline{p} = \frac{1}{n} \sum_1^n p_k$, 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 (27) 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 $P_1 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \ldots)$, $P_2 = (\frac{1}{2}, 0, \frac{1}{2}, \ldots)$ (in terms of the point probabilities associated with the elements in $\mathbb{A}^* = \{0, 1, 2, \ldots\}$). One finds that

$$P_1 * P_2 = \left(\frac{1}{6}, \frac{1}{6}, \frac{2}{6}, \frac{1}{6}, \frac{1}{6}, \ldots\right)$$

and

$$\overline{P} * \overline{P} = \left(\frac{25}{144}, \frac{20}{144}, \frac{54}{144}, \frac{20}{144}, \frac{25}{144}, \ldots\right)$$

hence

$$H(P_1 * P_2) = \frac{2}{3} \ln 2 + \ln 3 \approx 1.5607$$

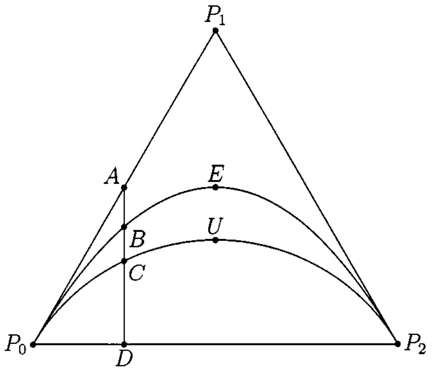

Fig. 2. Illustration of models with $\mathbb{A} = \{0, 1, 2\}$.

and

$$H(\overline{P} * \overline{P}) = \frac{221}{72}\ln 2 + \frac{7}{8}\ln 3 - \frac{35}{36}\ln 5 \approx 1.5241.$$

Therefore, (27) does not hold in this case.

As we have seen, Hoeffding's inequality, Theorem 7, may be considered to lie behind (27). It is noteworthy that whereas (27) itself appears difficult to generalize, far-reaching generalizations of Hoeffding's inequality (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. [3], [6] and [22]. How, or if, these results can be exploited in information-theoretical studies is unclear.

Shepp and Olkin's paper [23] contains a reference to Hoeffding's inequality. Really, the reference is only one of analogy. Shepp and Olkin do not make any use of the inequality, only 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 for the discussion of models defined in terms of Bernoulli sums.

As noted several times, the results concerning binomial and multinomial distributions owes much to Mateev and to Shepp and Olkin. The results, as far as they involve maximum-entropy considerations, are quite natural and were perhaps considered by several 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.

Finally, we shall illustrate some of the findings regarding the binomial and geometric distributions by looking at the case $n = 2$. The simplex $M_+^1(\mathbb{A}_2)$ together with various models are shown in Fig. 2. 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 $G_2(\lambda)$ for some $\lambda < 1$, whereas AB represents the model $B_2(\lambda)$. Note that for higher values of n, $B_n(\lambda)$ is not convex (but still connected). The points B and C represent the $H_{\max}$-distributions of $B_2(\lambda)$ and $G_2(\lambda)$, respectively. The curve P_0BEP_2 represents the model of all binomial distributions $\mathrm{BIN}(2, p)$; $0 \leq p \leq 1$, 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.

Acknowledgment

The author has had useful discussions with Peter Harremoës of all parts of the manuscript. In fact, the present research was triggered by a study of Harremoës, [9].

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 an informal report, based on the technical report from 1978 which preceded the publication of [23].

The essential importance of Hoeffding's inequality for a key argument in Section V was discussed with Paul Ressel who also drew the authors attention to the papers [3], [6], and [22]. The discussion of (27) and of Hoeffding's inequality owes much to the exchange of ideas with Paul Ressel.

After presentations of the material in October 2001 in Warsaw and in Budapest, discussions with Imre Csiszár led to several improvements, in particular, an error in a previous version of Theorem 5 was corrected.

References

[1] J.-P. Aubin, *Optima and Equilibria. An Introduction to Nonlinear Analysis*. Berlin, Germany: Springe-Verlag, 1993.

[2] A. R. Barron, "Entropy and the central limit theorem," *Ann. Probab.*, vol. 14, pp. 336–342, 1986.

[3] P. J. Bickel and W. R. van Zwet, "On a theorem of Hoeffding," in *Asymptotic Theory of Statistical Tests and Estimation*, I. M. Chakravarti, Ed. New York: Academic, 1980, pp. 307–324.

[4] P. Billingsley, *Convergence of Probability Measures*. New York: Wiley, 1968.

[5] N. N. Čencov, *Statistical Decision Rules and Optimal Decisions* (in Russian). Moscow, U.S.S.R.: Nauka, 1972.

[6] J. P. R. Christensen and P. Ressel, "A probabilistic characterization of negative definite and completely alternating functions," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 57, pp. 407–417, 1981.

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

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

[9] P. Harremoës, "Binomial and Poisson distributions as maximum entropy distributions," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2039–2041, July 2001.

[10] P. Harremoës and F. Topsøe, "Maximum entropy fundamentals," *Entropy*, vol. 3, pp. 191–226, 2001. [Online], Available: http://www.mdpi.org/entropy/.

[11] D. Haussler, "A general minimax result for relative entropy," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1276–1280, July 1997.

[12] W. Hoeffding, "On the distribution of the number of successes in independent trials," *Ann. Math. Statist.*, vol. 27, pp. 713–721, 1956.

[13] R. S. Ingarden and K. Urbanik, "Quantum informational thermodynamics," *Acta Phys. Polonica*, vol. 21, pp. 281–304, 1962.

[14] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106/108, pp. 620–630/171–190, 1957.

[15] ——, "Clearing up mysteries—The original goal," in *Maximum Entropy and Bayesian Methods*, J. Skilling, Ed. Dordrecht, The Netherlands: Kluwer, 1989.

[16] J. N. Kapur, *Maximum Entropy Models in Science and Engineering*. New York: Wiley, 1993. First edition: 1989.

[17] D. Kazakos, "Robust noiceless source coding through a game theoretic approach," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 576–583, July 1983.

[18] B. Lindström, "Determining subsets by unramified experiments," in *A Survey of Statistical Design and Linear Models*. Amsterdam, The Netherlands: North-Holland, 1975, pp. 407–418.

[19] Yu. V. Linnik, "An information-theoretic proof of the central limit theorem with the Lindeberg condition," *Theory of Probab. its Appl.*, vol. 4, pp. 288–299, 1959.

[20] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications*. New York: Academic, 1979.

[21] P. Mateev, "On the entropy of the multinomial distribution," *Theory of Probab. its Appl.*, vol. 23, pp. 188–190, 1978.

[22] P. Ressel, "A general Hoeffding type inequality," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 61, pp. 223–235, 1982.

[23] L. A. Shepp and J. Olkin, "Entropy of the sum of independent Bernoulli random variables and of the multidimensional distribution," in *Contributions to Probability*. New York-London: Academic, 1981, pp. 201–206.

[24] F. Topsøe, "Information theoretical optimization techniques," *Kybernetika*, vol. 15, pp. 1–27, 1979.

[25] ——, "Game theoretical equilibrium, maximum entropy and minimum information discrimination," in *Maximum Entropy and Bayesian Methods*, A. Mohammad-Djafari and G. Demoments, Eds. Dordrecht, Boston, London: Kluwer Academic, 1993, pp. 15–23.