Information Theory at the Service of Science

Flemming Topsøe [*]
Department of Mathematics
University of Copenhagen
topsoe@math.ku.dk

Abstract

Information theory is becoming more and more important for many fields. This is true for engineering- and technology-based areas but also for more theoretically oriented sciences such as probability and statistics.

Aspects of this development is first discussed at the non-technical level with emphasis on the role of information theoretical games. The overall rationale is explained and central types of examples presented where the game theoretical approach is useful.

The final section contains full proofs related to a subject of central importance for statistics, the estimation or updating by a posterior distribution which aims at minimizing divergence measured relative to a given prior.

Keywords. Description length, entropy, divergence, description gain, game theoretical equilibrium, maximum entropy, minimum information discrimination, capacity, redundancy, I-projection, Pythagorean inequality.

1 Introduction, Background

Information Theory is of importance for a number of disciplines from the very applied engineering oriented ones to more theoretical fields. One of the strongest interfaces is to probability and statistics. One can see a line in monographs such as Čencov [8], Kullback [35], Dembo and Zeituni [15] and Cover and Thomas [9], and in recent and ongoing research of authors

[*] Research supported by INTAS, project 00-738, and by the Danish Natural Science Research Council.

including Amari [1], Barron, Csiszár and Matúš [13], Harremoës [19] and others, much of it in the publishing process. General purpose textbooks which aim at making the new tools available are starting to appear, though still not taking full advantage of recent improvements and extensions of the techniques (cf. Jessop [29] and Applebaum [2]).

Information theory offers technical tools and a conceptual base which contribute to an understanding of many of the fundamental concepts and methods of probability and statistics. This also leads to a better understanding of the basic notion of probability. In this connection recall what de Finetti wrote in the foreword to his monograph [14] after a life long study of these fields: "Probabilities do not exist!" That there is something strange about the concept of probability is nowadays realized by all workers in theoretical probability and statistics and the last word on the concept of probability, on randomness, has certainly not been said. Is it something inherent in the real world or only something going on in our head, in our *perception* of the real world?

The change of paradigm which is underway will to begin with stick to probability theory with the set-up going back, basically, to Kolmogorov [34] as the primary theoretical basis (another main contributor is Hausdorff). However, more drastic changes may be underway, either based on information theory proper or on the neighbouring and inter-related discipline of complexity theory. In this connection we refer to Shafer and Vovk [42].

In order to understand what information theory has to offer which could change, if not revolutionize, probabilistic and statistical thinking, we point to three factors: The basic concepts have natural interpretations and as such are "just the right ones" and, secondly, you can often use these concepts to model *conflict situations* which are common in many areas, such as biology, economics, physics and then also in probability and, more pronouncedly so, statistics. As a last reason we point to the technical tools of information theory, especially to the powerful inequalities.

As to the first feature pointed to, we owe to Shannon, cf. his pioneering article [43] from 1948, the definition of "the right concepts" such as entropy, conditional entropy, mutual information and so on. Especially important for the interface to statistics is the introduction in 1951 by Kullback and Leibler [36], of a further quantity, now mainly called (*information*) *divergence*, cf. also Kullback [35].

The ability of information theory to model conflict situations can be said to go back to Shannon himself. As an early source we also mention Kelly [32]. However, we want to emphasize the comprehensive study of the Maximum Entropy Principle (MaxEnt) by Jaynes who from 1957 onwards has worked to put this principle on a firm information theoretical basis and also discussed the qualitative and philosophical aspects at great length, cf. [26], [27], [25] and [28]. MaxEnt dictates that if $\mathcal{P}$ is a model consisting of probability distributions, then the least biased one, and hence the one best suited for predictions, is that distribution in the model – if such a unique distribution exists – which has maximal entropy. Jaynes further stresses the view – which is at the same time a guiding principle – that one should not think of the model as a set of distributions, one of which is the "true" distribution. Rather, Jaynes maintains that $\mathcal{P}$ models our *knowledge* about the system studied.

Some years ago the author pointed out that a principle of *Game Theoretical Equilibrium* (GTE) can be taken to be basic, cf. [50] and [46]. Thus, from GTE, you are led to MaxEnt as well as to a principle going back to Kullback, the *Minimum Information Discrimination Principle*, which dictates that you select that distribution in the model which has the smallest divergence to a given or suitably chosen *prior* distribution. As the term "divergence" now appears to be more common than "discrimination" (or "information distance" or "relative entropy" which are also terms in use), we refer to Kullback's principle as MinDiv, the *minimum divergence principle.* The geometrically-flavoured structure involved in MinDiv was studied by Csiszár [10], who introduced the concept of *I-projection* for the optimal distribution of MinDiv. Further studies of Csiszár demonstrated the significance of information theory for statistics, cf. [11].

The success of information theory at the service of other sciences depends on the ability to derive key results based on information theoretical principles in a way which is felt natural and technically convenient within the sciences in question. The sources mentioned contain results in this direction for the fields of probability and statistics. Some additional references include the difficult paper by Linnik [37] regarding the Central Limit Theorem, Barron's follow-up paper [4] (with more underway, I understand), and we may also mention Harremoës [22] and Topsøe [48]. The forthcoming paper [18] by Grünwald and Dawid is very much in line with the approach adopted here. The reader should also watch out for forthcoming work by Boyd and Chiang.

2 Game Theoretical Equilibrium, the idea

In this section we shall reveal the over all idea of GTE. In principle, this is simple: GTE dictates that you use information theoretical concepts to view a problem as a game, typically as a two-person zero-sum game. The search for optimal strategies, i.e. the attempt to approach equilibrium in the game theoretical sense, will then lead to useful, even completely satisfactory insight into the original problem. The power of GTE then depends on the ability to make meaningful transformations of interesting problems to the game theoretical setting. Below follow some qualitative, partly philosophical remarks on this issue.

Information theory provides concepts and tools for expressing the role of an "observer" or "decision maker" (the statistician, the physicist, the investment planner or what the case may be). As a brief indication of this aspect of information theory, we need only point to coding as a way to *represent* or to *describe* data and, at the same time, as a means to *identify* outcomes, hence to express strategies for making *observations*.

When we use information theory to model a part of "*reality*", we should acknowledge – based on a Platonic or neo-Platonic view – that all we can do is to model our *ignorance* about reality or, put positively, to model our *knowledge* about reality. In our modelling we should be open to any eventuality. Possibly motivated by the experience that "what can go wrong *does* go wrong", this points to applying a safe (cowardice!) strategy of minimizing maximal risk, hence this leads to by now classical "mini-max thinking" of game theory. In its simplest form, we are thus led to consider two-person zero-sum games. This modelling is possible in many specific situations via the identification of an *objective function* (a *cost* or a *pay-off*, depending on the sign).

Qualitatively, the above considerations are in consistency with basic Bayesian thinking in statistics. And other sciences acknowledge similar ideas. In physics we point to the Copenhagen interpretation of quantum physics and the emphasis of Niels Bohr on the interplay, *complementarity*, between the system under study and the observer looking at it through whatever glasses are available. A system behaves not only in accordance with what is known because of insight gained by previous observations. The behaviour depends on *all* factors – including those we *could* have taken into account. Regarding quantum physics it may well be that the Copenhagen interpretation is ripe for replacement or thorough revision, largely because the notion of probability is not well explained or motivated and a change which would involve, among other elements, Jaynes' ideas which combine information theory and the interpretation of probability concepts is desirable. For the field of biology, we point to Maynard Smith's principle of *evolutionary stability*. For a recent application of this principle, see Broom [7]. We may also point to the social sciences, to economy, cf. von Neumann's pioneering work in [51], or more modern texts such as Aubin [3], and also, we can point to recent applications to finance where Delbaen et. al. [41] and Bellini and Frittelli [5] serve as entrance points to the relevant literature where information theoretical considerations come into play.

3 Information theoretical concepts

3.1 The general idea

In order to stress the underlying rationale, thereby pointing to the wider applicability of basic concepts from information theory, we introduce these concepts in a *context* which is quite abstract and freed from any reference to standard information theory. However, we do use terminology borrowed from that theory.

The nature of the context does not really concern us here. Anyhow, it may be helpful to think of it as some well defined part of "reality" which may have been isolated by some process of *preparation*. As an indication of what we have in mind, consider the case of an ideal gas submerged in a heat bath. Through the preparation, a "system" or "context" suitable for study emerges.

We shall characterize any specific context by three entities. Firstly a set $\mathcal{P}$, called the *knowledge base*. This set reflects the available knowledge. Elements of $\mathcal{P}$ are referred to as *consistent instances*, i.e. as instances which are consistent with the preparation of the system. In specific situations when we know the nature of the elements of $\mathcal{P}$, we may reflect that in our terminology. For instance, in all of our examples, $\mathcal{P}$ will be a set of probability distributions and we may then talk about *consistent distributions*.

Further, we assume that we have access to a set K of *description strategies*. Various interpretations are possible. Either the strategies can be used for making observations from the system or for *representing* consistent instances or, in the terminology we shall prefer below, for *describing* such instances.

The third entity we have in mind is a function, the *objective function*, which to $\kappa \in K$ and $P \in \mathcal{P}$ associates a non-negative (possibly infinite) number, denoted $\langle \kappa, P \rangle$. This is the *description length* (with κ as description strategy and P as the consistent instance).

In our set-up we have focused on three elements: Our *knowledge*, the available *tools for description* and our *objective*. Further elements – such as side information, conditioning, symmetry etc. – may be brought into the picture but the chosen framework appears to constitute an adequate playground for basic information theoretical considerations. Instead of the neutral reference to a context, one could refer to the triple $(\mathcal{P}, K, \langle \cdot, \cdot \rangle)$ as an *information space*.

Sometimes (see Section 5), it is not possible to suggest in a meaningful way an objective function which is non-negative. Often, one can then instead work with an objective function which can also assume negative values. This will, typically, correspond to situations where you find it natural to measure performance relative to some chosen reference. One may, therefore, work with two notions, one of *absolute information spaces* as considered up to now and one of *relative information spaces*. It is only the former which we have in mind in this section.

Now then, let us demonstrate that the set-up as introduced allows the introduction of key quantities for further study.

First define the *entropy* of a consistent instance P as the minimum description length:

$$H(P) = \min_{\kappa \in K} \langle \kappa, P \rangle \,. \tag{1}$$

If there exists $\kappa \in K$ with $\langle \kappa, P \rangle = H(P)$ we say that κ is *adapted* to P. Often, such strategies are uniquely determined from P.

Then we can define *redundancy* $D(P\|\kappa)$ as the unnecessary part of the description length when using $\kappa \in K$ as strategy instead of an optimal strategy adapted to $P \in \mathcal{P}$, i.e.

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

This quantity may also be conceived as the saving in description length inherent in the information obtained if we are told what P is. For this interpretation it is understood that the strategy κ is used for description of P before we are told what P is and that we switch to a strategy best adapted to P as soon as we are told what P is.

Note that (2) may not be adequate as a definition in all cases. Sometimes, as in the case of information theory proper, it is possible to suggest a meaningful definition which also covers the indeterminate case when (2) leads to the form $D(P\|\kappa) = \infty - \infty$. Anyhow, the *linking identity*

$$\langle \kappa, P \rangle = H(P) + D(P\|\kappa). \tag{3}$$

always holds (as we agree that the right hand side is $+\infty$ when $H(P) = \infty$).

A further quantity of interest is the *description gain* inherent in a change of strategy from κ to ρ in the description of P. This quantity can even be taken to be more fundamental than entropy as well as redundancy and it is to be expected that it will play a significant role in future research (it already does so but in a less dominant and more implicit way). A natural defining relation would be

$$D(R\|\kappa \rightsquigarrow \rho) = \langle \kappa, P \rangle - \langle \rho, P \rangle.$$

However, this is very likely to lead to the indeterminate quantity $\infty - \infty$ and it is better to exploit the linking identity (3) for $\langle \kappa, P \rangle$ as well as for

$\langle \rho, P\rangle$ and use the relation

$$D(P\|\kappa \rightsquigarrow \rho) = D(P\|\kappa) - D(P\|\rho) \tag{4}$$

as definition. Of course, this quantity may be negative or even $-\infty$. Normally, situations with $D(P\|\kappa \rightsquigarrow \rho) \geq 0$, i.e. with a genuine gain, will have our main interest and other cases can be ignored. From (4) we obtain the relation

$$D(P\|\kappa) = D(P\|\kappa \rightsquigarrow \rho) + D(P\|\rho)\,, \tag{5}$$

which could be called the *linking identity of the second kind* (then (3) is of the first kind). However – unlike (3) – (5) is not necessarily valid in all cases but does hold, e.g. whenever $D(P\|\rho) < \infty$.

Further concepts may be introduced but this will, typically, require extra structure, and we shall only look into that in the context of information theory proper.

3.2 Proper information theoretical concepts

Information theory provides the key example where the above definitions apply and have natural interpretations as suggested by the terminology used. Let us briefly go through this. In so doing we shall here restrict the scope and only have the discrete case in mind. In the more technical final section we extend the scope to encompass also the continuous case.

As starting point we shall then take a discrete set $\mathbb{A}$, i.e. a finite or countably infinite set. This is the *alphabet*. As $\mathcal{P}$ we take a set of probability distributions on $\mathbb{A}$ and as K the set of *codes* or, more precisely, *idealized code length functions*. The way we think about $\kappa \in K$ is as an allocation of code words to elements in $\mathbb{A}$, however only paying attention to the lengths of the code words. More precisely, $\kappa \in K$ is a map which to each $i \in \mathbb{A}$ associates a number $\kappa(i) \in [0,\infty]$ in such a way that Krafts equality in idealized form, i.e. $\sum \exp(-\kappa(i)) = 1$, holds (the idealization refers to the acceptance of arbitrary real numbers as values of κ, and to the choice of the natural base rather than the base 2 for the exponentiation). The interpretations related to this definition are well known, see e.g. [48].

There is an important natural $1-1$ correspondance, written $\kappa \leftrightarrow Q$, between codes and distributions for which $P(i) = \exp(-\kappa(i))\,, \kappa(i) = -\log P(i)$ (we use "log" to denote natural logarithms).

In order for the definitions of Section 3.1 to apply we also have to specify the objective function $(\kappa, P) \curvearrowright \langle \kappa, P\rangle$. For this we take expected code length, i.e. we put

$$\langle \kappa, P\rangle = \sum \kappa(i)P(i).$$

The definitions of entropy, redundancy and description gain then make good sense. As is classical, $H(P) = \langle \kappa, P \rangle$ with $\kappa \leftrightarrow P$.

We use the correspondance between codes and distributions to define *divergence* (from P to Q) by

$$D(P\|Q) = D(P\|\kappa) \tag{6}$$

with $\kappa \leftrightarrow Q$ and also to define the following quantity:

$$D(P\|Q \rightsquigarrow R) = D(P\|\kappa \rightsquigarrow \rho) \tag{7}$$

with $\kappa \leftrightarrow Q$ and $\rho \leftrightarrow R$. Note that $D(P\|Q) = D(P\|Q \rightsquigarrow P)$. We can think of $D(P\|Q \rightsquigarrow R)$ as reflecting an estimation- or prediction- or updating situation with Q as *prior* and R as *posterior distribution.* We may, therefore, call $D(P\|Q \rightsquigarrow R)$ the *estimation gain* or the *updating gain* associated with this situation.

Much clarity results by realizing that redundancy (related to description via coding) conceptually preceedes divergence (related to distributions).

It is well known that whereas entropy can only be finite for essentially discrete distributions, redundancy and divergence has a much wider scope of applicability which matches the requirement in probability and statistics to encompass basic continuous distributions. Indeed, in the case of divergence, one is led to the well known formula

$$D(P\|Q) = \int \log \frac{dP}{dQ} dP \tag{8}$$

(when $P \ll Q$), which may be used as a definition which then covers all eventualities (with $D(P\|Q) = \infty$ in case $P \not\ll Q$). Technical details can be found, e.g. in [47] where you will also find important extensions of concepts introduced, especially involving conditioning and datareduction.

4 Instances of GTE

The conceptual base developed in the previous section gives rise to a number of specific two-person zero-sum games of which we mention three. The results of the first two subsections are of particular relevance for applications to probability and statistics. Possible applications lead to interesting characterizations of key distributions, often accompanied by appropriate limit theorems. Many such results can be conceived as instances of the GTE principle as discussed here. Kapur [30] contains a great number of such results. We may also mention Haussler [23], Kazakos [31], Harremoës [22] and Topsøe

[48] which are based on the game theoretical approach. The first papers using this approach are Pfaffelhuber [39] and Topsøe [50]. [*] For further work in this direction see Harremoës and Topsøe [20] as well as forthcoming work by several authors, including Grünwald and Dawid [18].

4.1 The $H_{\max}$-game

The first game is the *maximum entropy game*, the $H_{\max}$-*game*, which is related to entropy and code length in the discrete case and refers to a given set $\mathcal{P}$ of distributions over the alphabet $\mathbb{A}$. One may also refer to this game as the *code length game*. For the $H_{\max}$-game, Player I ("the system") chooses a consistent distribution $P \in \mathcal{P}$ and Player II (the statistician, say) chooses a code $\kappa \in K$. Description length $\langle \kappa, P \rangle$ is taken as the cost, seen from the point of view of Player II.

We define $H_{\max} = H_{\max}(\mathcal{P})$, the *maximum entropy value* as

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

For $\kappa \in K$, we define $R(\kappa) = R(\kappa|\mathcal{P})$, the *risk* associated with κ, to be

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

and we put

$$R_{\min} = \inf_{\kappa \in K} R(\kappa).$$

Then $H_{\max} \leq R_{\min}$ and if $\mathcal{P}$ is convex and $H_{\max} < \infty$, it can be shown that the game is in *equilibrium*, i.e. $H_{\max} = R_{\min}$. Therefore, *maximum entropy equals minimum risk*. As the risk was defined via description length, we may also say that *maximum entropy equals minimum description length*.

A technically simple result which is very powerful for the applications is that if there exists a *Nash equilibrium code* adapted to $\mathcal{P}$, i.e. – copying concepts from game theory – a code κ^* such that $R(\kappa^*) \leq H(P^*) < \infty$ with $\kappa^* \leftrightarrow P^*$ and $P^* \in \mathcal{P}$, then κ^* and P^* are unique optimal strategies for the players. In particular, P^* is the unique maximum entropy distribution. For a quick introduction to these results, see Topsøe [48]. More details are in Harremoës and Topsøe [21].

[*] The paper [39] was only discovered by the author shortly before submission of the final manuscript. Apparently the research of Pfaffelhuber was carried out in 1975 whereas the main work of the author stems from 1976 with the first formal publication in Danish in 1978. In any case the author has to share the introduction of the game theoretical approach with Pfaffelhuber.

4.2 The $D_{\min}$-game

The second game we shall study is the $D_{\min}$-*game* which we here describe in rather qualitative and motivating terms and again only having the discrete case in mind. We do, however, follow-up in Section 5 with concrete technical details which apply to a more general situation, covering also the continuous case.

The setting is as in Section 4.1, except that now we have also given a *reference code* κ_0 or, equivalently, a *prior distribution* P_0 $(\kappa_0 \leftrightarrow P_0)$. Again, Player I chooses a $P \in \mathcal{P}$ and Player II chooses a code $\kappa \in K$ or, equivalently, a *posterior distribution* R (with $\kappa \leftrightarrow R$). But now, description gain, $D(P\|\kappa_0 \rightsquigarrow \kappa)$ or, equivalently, *estimation* or *updating gain*, $D(P\|P_0 \rightsquigarrow R)$, is taken as pay-off seen from the point of view of Player II. According to the zero-sum character which we insist on, $D(P\|\kappa_0 \rightsquigarrow \kappa)$ is considered as a cost to Player I.

If Player I chooses $P \in \mathcal{P}$ one readily sees that fixing this distribution, the largest cost which Player I risks is $D(P\|\kappa_0)$. Therefore, an optimal strategy for Player I is a $P^* \in \mathcal{P}$ for which $D(P^*\|\kappa_0) = D_{\min}$ with the latter quantity defined by

$$D_{\min} = \inf_{P \in \mathcal{P}} D(P\|P_0).$$

The game thus leads to the MinDiv-principle.

In the theoretical discussion of this game, it is advantegeous to consider also the other side of the game. A main result states that under mild regularity conditions ($\mathcal{P}$ convex, $D_{\min} < \infty$), there exists κ^* such that

$$D(P\|\kappa_0 \rightsquigarrow \kappa^*) \geq D_{\min} \tag{9}$$

for every $P \in \mathcal{P}$.

In fact, with $\kappa_0 \leftrightarrow P_0$ and $\kappa^* \leftrightarrow P^*$, the inequality (9) is equivalent to the celebrated *Pythagorean inequality* of Csiszàr:

$$D(P\|P_0) \geq D(P\|P^*) + D_{\min} \tag{10}$$

for all $P \in \mathcal{P}$. In Csiszàr's terminology, P^* is the (generalized) *I-projection of* P_0 *on* $\mathcal{P}$. With an interpretation as indicated above, P^* is the *optimal estimator* or the *optimal updating strategy* based on the prior P_0. In the coding terminology, κ^* is the *optimal adjustment* of κ_0. Note that if $(P_n)_{n\geq 1} \subseteq \mathcal{P}$ satisfies $D(P_n\|P_0) \to D_{\min}$, then $D(P_n\|P^*) \to 0$. This is a rather strong type of limit theorem, in particular stronger than setwise convergence of P_n to P^*.

4.3 The $I_{\max}$-game

The third instance we shall mention where GTE applies is of main interest for information theory proper and is called the *capacity–redundancy game* or, for reasons explained below, the *maximum information transmission game*, the $I_{\max}$-*game*. Again, it involves a set $\mathcal{P}$ of probability distributions but now no prior code or distribution is given. If Player II ("the receiver") can choose a code and redundancy $D(P\|\kappa)$ is taken as cost, we are led to consider *minimal redundancy* defined as

$$R_{\min} = \inf_{\kappa} \sup_{P\in\mathcal{P}} D(P\|\kappa)\,. \tag{11}$$

Clearly, if the other side of the game is taken to involve only $P \in \mathcal{P}$ as permissable strategies for Player I, the game will normally not be interesting as

$$\sup_{P\in\mathcal{P}} \inf_{\kappa} D(P\|\kappa) = 0\,. \tag{12}$$

However, if we – "à la von Neumann" – consider mixed strategies (i.e. randomization corresponding to convex combinations $\sum \alpha_\nu P_\nu$ of members of $\mathcal{P}$) and choose the associated *average redundancy*, $\sum \alpha_\nu D(P_\nu\|\kappa)$, as the quantity the players should worry about (the objective function), an interesting game is obtained. It is the simplest to consider the case when $\mathcal{P}$ consistes of a finite or countably infinite set of distributions P_ν ($1 \le \nu \le n$ or $1 \le \nu < \infty$). This game may be interpreted as one involving a *channel* with $\alpha = (\alpha_\nu)$ as input distribution and then the quantity replacing (12) is

$$\sup_{(\alpha_\nu)} \inf_{\kappa} \sum \alpha_\nu D(P_\nu\|\kappa) \tag{13}$$

which is the *capacity* of the channel (see below). The quantity (11) is not affected as divergence has strong convexity properties, in particular, it is convex in the first argument. By applying the principle of GTE one then proves an important result of information theory, the *Gallager–Ryabko* theorem which equates capacity with minimum redundancy, cf. Ryabko [40].

Let us go a little more into the details. First we point out the following identity, the *compensation identity*, cf. Theorems 6.1 and 9.1 of [47]. With notation as above, it states that

$$\sum \alpha_\nu D(P_\nu\|\kappa) = \sum \alpha_\nu D(P_\nu\|\kappa_0) + D(P_0\|\kappa) \tag{14}$$

where $P_0 = \sum \alpha_\nu P_\nu$ and $\kappa_0 \leftrightarrow P_0$. This shows that the infimum in (13) can be identified as

$$I(\alpha) = \sum \alpha_\nu D(P_\nu\|\kappa_0)\,. \tag{15}$$

This quantity is the *information transmission rate* associated with the input distribution α. The maximum information transmission rate, the $I_{\max}$-value, is the *capacity* of the channel and this then is the quantity appearing in (13). The intuitive content of (15) is perhaps best understood if the code κ_0 is replaced by the matching distribution P_0: With probability α_ν a "letter" is sent through the channel and this changes the distribution at the output side from the á priori distribution P_0 to P_ν. The *redundancy removed*, hence the *information received* by this change is $D(P_\nu \| P_0)$. Thus, the information transmission rate is defined as average information obtained at the output side. By (15) we may write (14) in the form

$$\sum \alpha_\nu D(P_\nu \| \kappa) = I(\alpha) + D(P_0 \| \kappa) \tag{16}$$

which is a *linking identity*, now of the third kind. [†]

Intuitively it is to be expected that optimal usage of the channel must be achieved for an input distribution α for which the *Kuhn–Tucker conditions*

$$D(P_\nu \| P_0) = C \quad \text{for all } \nu \text{ with } \alpha_\nu > 0 \tag{17}$$

$$D(P \| P_0) \leq C \quad \text{for all } P \in \mathcal{P} \tag{18}$$

are fulfilled with C some constant. In fact it is nothing but a simple exercise to use the compensation identity to realize that (17) and (18) are sufficient conditions for optimality. The wide applicability of this result has been established in collaboration with Peter Harremoës (unpublished). For a simple proof with $\mathcal{P}$ a finite set of distributions, see Topsøe [45].

When (17) and (18) hold, it follows directly from the compensation identity (14) and from the definitions involved that $I(\alpha) = R_{\min} = C$, hence proving the Gallager–Ryabko theorem.

The analysis of the $I_{\max}$-game applies in many settings with $D(P\|Q)$ replaced by other measures of "divergence". The reason for this is that the crucial compensation identity (14) holds in a variety of cases.

Firstly, (14) holds for squared Euclidean distance. In that case, the problem suggested by (11) belongs to *location theory*, cf. [17]. The specific problem goes back to Sylvester [44] who wrote "It is required to find the least circle which shall contain a given system of points in a plane" – in fact, this is the full text of [44]!

The compensation identity also holds for *Bregman divergences*, (regarding these, see Bregman [6] and Csiszàr [12]). [‡]

[†] Note that in all cases, a *linking identity* is an identity revealing a basic structural relation which involves an objective function for one of the games considered.

[‡] This remark is due to Csiszàr, oral communication at the conference "Information Theory, Cryptography and Statistics", Balatonlelle, October 2000.

GTE	Player I	Player II
$H_{\max}$-game	Max Ent	Min Length
$D_{\min}$-game	Min Div	Max Saving
$I_{\max}$-game	Max Inf	Min Redundancy

Table 1: Instances of the principle of GTE

As a final example of the wide applicability of the Kuhn-Tucker criterion for variants of the $I_{\max}$-game, we mention the analogue of this game in the setting of quantum information theory with $\operatorname{Trace} P(\log P - \log Q)$ as replacement of classical divergence (here, P and Q are density operators, cf. Holevo [24] or Ohya and Petz [38]). [§] In the setting of quantum information theory, the compensation identity is often referred to as *Donalds identity*, cf. Donalds [16]. In the classical case, the identity possibly first appeared in [49].

In Table 1 we summarize the three games discussed in this section together with an indiction of the principles this leads to when seen from the point of view of each of the two players in the game.

5 Technical discussion of the $D_{\min}$-game

Let $(X, \mathbb{B})$ be a *Borel space*, i.e. a set provided with a σ-algebra and denote by $M_+^1(X, \mathbb{B})$ the set of probability distributions on $(X, \mathbb{B})$. Let $\mathcal{P} \subseteqq M_+^1(X, \mathbb{B})$ and $P_0 \in M_+^1(X, \mathbb{B})$ be given.

Consider the $D_{\min}$-game with $\mathcal{P}$ as the strategy set for Player I, with P_0 as prior distribution and with the set of all probability distributions on $(X, \mathbb{B})$ as the strategy set for Player II. Thus we choose to consider estimators or posterior distributions rather than codes as the available strategies for Player II. This will give the results a more conventional flavour which is more likely to appeal to the reader interested mainly in applications to statistics. And of course, the concept of codes really only applies meaningfully to the discrete case. Thus, focusing on distributions enables us to present a completely general version of the $D_{\min}$-game.

We realize that the setting fits into the framework outlined in Section 3.1 of an information space. The knowledge base is $\mathcal{P}$, the set K of description strategies is $M_+^1(X, \mathbb{B})$ and the objective function is the map $(P, R) \curvearrowright$

[§] This observation is joint work (unpublished) with Peter Harremoës, made during the MaPsySto workshop, October, 1999, University of Aarhus.

$D(P\|P_0 \rightsquigarrow R)$. As this function may assume negative values, the proper setting is that of a *relative* information space.

The defining relations (8) (for divergence) and (7) combined with (4) (for estimation gain) are now taken as the basic definitions. Written explicitly, the definition we use for estimation gain is as follows:

$$D(P\|Q \rightsquigarrow R) = D(P\|Q) - D(P\|R). \tag{19}$$

As there are some subtleties connected with (19) (the $\infty - \infty$ problem), it is not entirely clear if all aspects of the $D_{\min}$-game make sense. However, with appropriate conventions this can in fact be achieved. We simply agree that if a suppremum is considered involving indeterminate numbers we interpret the result in the "least favourable way" as $+\infty$ and, likewise, if an infimum involves an indeterminate number, we interpret the result as giving the quantity $-\infty$. With these conventions it is clear that each of the players can assign a performance index to any specific strategy available to the player. And for Player I we can easily identify the performance index. Indeed, if Player I chooses the strategy $P \in \mathcal{P}$, the associated performance index is nothing but the divergence $D(P\|P_0)$ as

$$\sup_R D(P\|P_0 \rightsquigarrow R) = D(P\|P_0)$$

(here, it is understood that R ranges over all strategies available to Player II, i.e. over all probability distributions on $(X, \mathbb{B})$, and that the conventions just introduced are in force). Player I is then led to consider the quantity

$$D_{\min} = \inf_{P \in \mathcal{P}} D(P||P_0)$$

and the notion of an optimal strategy for Player I makes good sense as a distribution $P \in \mathcal{P}$ with $D(P\|P_0) = D_{\min}$.

As for Player II, the performance index associated with the strategy R is given by

$$\Pi(R) = \inf_{P \in \mathcal{P}} D(P\|P_0 \rightsquigarrow R).$$

This is the *pay-off* or, more specifically, the *estimation gain* associated with the strategy. Player II is then led to consider the *maximal pay-off* given by

$$\Pi_{\max} = \sup_R \Pi(R)$$

and the notion of an optimal strategy for Player II then makes good sense as a distribution R with $\Pi(R) = \Pi_{\max}$. Clearly, $\Pi_{\max} \leq D_{\min}$. If equality holds, the common value is the *value* of the game and the game is then said to be in *equilibrium.*

The first result, really Theorem 8 of [50], can be formulated in a very standard way without reference to the $D_{\min}$-game at all. For the proof we need *Pinsker's inequality* which states that for any two distributions P and Q,

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

with V denoting total variation. We also need to know that $(P,Q) \curvearrowright D(P||Q)$ is jointly lower semi-continuous, even with respect to the relatively weak topology of setwise convergence. For these facts, see [47].

Theorem 1. *Assume that $\mathcal{P}$ is convex and that $D_{\min} < \infty$. Then there exists a unique distribution P^* such that, for every $P \in \mathcal{P}$, the Pythagorean inequality holds, i.e.*

$$D(P||P_0) \geq D(P||P^*) + D_{\min}. \tag{20}$$

Proof. Let $(P_n)_{n\geq 1}$ be a sequence of distributions in $\mathcal{P}$ such that the convergence $D(P_n||P_0) \to D_{\min}$ holds. Then, by the compensation identity, we find that for every $n \geq 1, m \geq 1$, and with $P_{n,m}$ denoting the distribution $\frac{1}{2}P_n + \frac{1}{2}P_m$,

$$\begin{aligned} D_{\min} &+ \frac{1}{2}D(P_n||P_{n,m}) + \frac{1}{2}D(P_m||P_{n,m}) \\ &\leq D(P_{n,m}||P_0) + \frac{1}{2}D(P_n||P_{n,m}) + \frac{1}{2}D(P_m||P_{n,m}) \\ &= \frac{1}{2}D(P_n||P_0) + \frac{1}{2}D(P_m||P_0). \end{aligned}$$

As this last quantity converges to $D_{\min}$, we conclude by Pinskers inequality – which shows that both $D(P_n||P_{n,m})$ and $D(P_m||P_{n,m})$ are lower bounded by $\frac{1}{8}V(P_n,P_m)$ – that $(P_n)_{n\geq 1}$ is a Cauchy-sequence w.r.t. total variation. By completeness of total variation, there exists a probability distribution P^* such that $P_n \to P^*$ in total variation, i.e. $V(P_n,P^*) \to 0$. By a standard argument involving the mixing of two sequences, we realize that P^* is independent of the particular sequence $(P_n)_{n\geq 1} \subseteq \mathcal{P}$ with $D(P_n||P_0) \to D_{\min}$.

Now choose a sequence $(P_n)_{n\geq 1} \subseteq \mathcal{P}$ which "converges fast" in the sense that

$$n(D(P_n||P_0) - D_{\min}) \to 0.$$

We shall use this auxiliary sequence to establish (20). To do so, fix $P \in \mathcal{P}$ and consider the distributions $Q_n, n \geq 1$ given by

$$Q_n = (1 - \tfrac{1}{n})P_n + \tfrac{1}{n}P.$$

Clearly, $D(Q_n||P_0) \geq D_{\min}$. Again appealing to the compensation identity, we find that

$$\begin{aligned} D_{\min} + \frac{1}{n}D(P\|Q_n) &\leq D(Q_n||P_0) + (1-\frac{1}{n})D(P_n||Q_n) + \frac{1}{n}D(P||Q_n) \\ &= (1-\frac{1}{n})D(P_n||P_0) + \frac{1}{n}D(P||P_0) \end{aligned}$$

and therefore,

$$D(P||P_0) + n(D(P_n||P_0) - D_{\min}) \geq D(P||Q_n) + D(P_n||P_0)\,.$$

Now then, (20) follows from the "fast convergence" of (P_n) and from the lower semi-continuity of $Q \curvearrowright D(P||Q)$ (as Q_n converges setwise, even in total variation, to P^*).

To finish the proof we note that the Pythagorean inequality uniquely characterizes P^*. Indeed, if $(P_n)_{n\geq 1} \subseteq \mathcal{P}$ satisfies $D(P_n||P_0) \to D_{\min}$, then $D(P_n||P^*) \to 0$, i.e. P_n *converges in divergence* to P^*. In particular, P_n converges in total variation to P^*. □

Note the somewhat peculiar aspect of the above proof: First one establishes as an auxiliary result a not-so-strong form for convergence of P_n with $(P_n) \subseteq \mathcal{P}$ and $D(P_n||P_0) \to D_{\min}$ and then, this is strengthened in the last part of the proof to a much stronger, and information theoretically more appropriate form of convergence. The fact that convergence in divergence (for ordinary sequences as here considered) is a topological notion of convergence follows from a well known fact of general topology (Kisynski's theorem, cf. [33]) but a more direct and intrinsicly information theoretical analysis of this important observation is possible as demonstrated in Harremoës [19].

Following Csiszàr [11], we say that P^* of Theorem 1 is the *generalized I-projection of P_0 on $\mathcal{P}$* (previous terminology of the author spoke about P^* as the *relative centre of attraction*, cf. [50]).

In many cases we can write the Pythagorean inequality in the form $D(P||P_0 \rightsquigarrow P^*) \geq D_{\min}$. For instance, if $D(P||P_0) < \infty$, this is so and by standard considerations involving Radon Nikodym derivatives it follows that then

$$D(P||P_0 \rightsquigarrow P^*) = \int \log \frac{dP^*}{dP_0} dP. \tag{21}$$

We shall now strengthen the assumptions $D_{\min} < \infty$ and use Theorem 1 to derive a key property of the $D_{\min}$-game. The main part of the result was proved as Theorem 9 of [50].

Theorem 2. *Assume that $\mathcal{P}$ is convex and that $D(P||P_0) < \infty$ for every $P \in \mathcal{P}$. Then the $D_{\min}$-game is in equilibrium and P^*, defined as the generalized*

I-projection of P_0 on $\mathcal{P}$, can also be characterized as the unique optimal estimator for Player II. Furthermore, for every distribution R the inequality

$$\Pi(R) + D(P^*||R) \leq \Pi_{\max} \tag{22}$$

holds when properly interpreted[¶] *. Finally, if $R \in \mathcal{P}$, then*

$$V(P^*, R) \leq \sqrt{D(R||P_0) - \Pi(R)}\,. \tag{23}$$

Proof. By (20), $\Pi(P^*) \geq D_{\min}$, and as the reverse inequality is trivial, the game is in equilibrium, and P^* is an optimal estimator. Let $(P_n)_{n\geq 1} \subseteq \mathcal{P}$ be an auxiliary sequence chosen such that $D(P_n||P_0) \to D_{\min}$. Consider a distribution R with and assume that $\Pi(R) > -\infty$ (otherwise, (22) is trivial). Then,

$$\begin{aligned}
\Pi(R) &\leq \limsup_{n\to\infty} D(P_n||P_0 \rightsquigarrow R) \\
&= \limsup_{n\to\infty}(D(P_n||P_0) - D(P_n||R)) \\
&= D_{\min} - \liminf_{n\to\infty} D(P_n||R) \\
&\leq D_{\min} - D(P^*||R)
\end{aligned}$$

by lower semi-continuity. As $\Pi(R) > -\infty$ by assumption, (22) follows.

The uniqueness of P^* viewed as an optimal estimator follows from (22).

Combining (20) and (22) and using Pinsker's inequality, both for $D(R||P^*)$ and for $D(P^*||R)$, (23) follows. □

We end by discussing some special situations. Quite often, e.g. for information spaces specified by moment constraints, the optimal estimator P^* is very robust in the sense that the estimation gain $D(P||P_0 \rightsquigarrow P^*)$ is independent of which distribution $P \in \mathcal{P}$ is chosen by Player I. In this connection we define some more specialized concepts. The distribution P^* is *pay-off stable* if $D(P||P_0 \rightsquigarrow P^*)$ is well defined and finite for all $P \in \mathcal{P}$ and, furthermore, independent of $P \in \mathcal{P}$. So we demand that $D(P||P_0)$ and $D(P||P^*)$ are finite for all $P \in \mathcal{P}$ and that, for some constant π, $D(P||P_0 \rightsquigarrow P^*) = \pi$ for all $P \in \mathcal{P}$. Expressed briefly, the requirement is that

$$\int \log \frac{dP^*}{dP_0} dP = \pi \;\text{ for all } P \in \mathcal{P}\,.$$

There may be many pay-off stable distributions. They constitute the *exponential family* associated with $(\mathcal{P}, P_0)$. This definition of exponential

[¶] By this we mean that either $\Pi(R) = -\infty$ or else $\Pi(R) \in \mathbb{R}$ and (22) holds. Thus the inequality implies that if $\Pi(R) > -\infty$, then $\Pi(R)$ and $D(P^*||R)$ are both finite.

families leads to well known families for classical models specified by moment constraints.

The first important concept we wish to introduce is that of a *consistent pay-off stable distribution.* Sometimes a weaker concept, here only defined for consistent distributions, is useful: A consistent distribution P^* is a *Nash equilibrium estimator* if $\Pi(P^*) \geq D(P^*\|P_0)$ and $D_{\min} < \infty$. The main condition amounts to the inequality

$$D(P||P_0 \rightsquigarrow P^*) \geq D(P^*||P_0) \text{ for all } P \in \mathcal{P}. \tag{24}$$

Implicit in (24) is the requirement that $D(P||P^*)$ is finite for all $P \in \mathcal{P}$. To see that the last concept introduced is more general than the first, assume that P^* is consistent and pay-off stable, say $D(P||P_0 \rightsquigarrow P^*) = \pi$ for $P \in \mathcal{P}$. Then $D(P^*||P_0) = \pi$ as P^* is consistent and (24) even holds with equality. Thus, a consistent pay-off stable estimator is indeed a Nash equilibrium estimator.

We now give a direct proof – independent of the proofs of Theorems 1 and 2 – which shows that all the conclusions of these therems hold if only P^* is known to be a Nash equlibrium estimator. And we obtain an extra desirable property. In the situation covered by Theorems 1 and 2 there is a risk that $D(P^*||P_0) \leq D_{\min}$ holds with strict inequality. This cannot happen if P^* has the Nash equilibrium property.

Theorem 3. *Let $\mathcal{P}$ be any set of distributions and P_0 any prior distribution. Assume that P^* is a consistent Nash equilibrium estimator. Then the $D_{\min}$-game for $(\mathcal{P}, P_0)$ is in equilibrium and P^* is the unique optimal strategy for Player I as well as the unique optimal strategy for Player II. In other words, P^* has minimum divergence to P_0 among all consistent distributions and, considered as an estimator, P^* achieves the maximal estimation gain. Furthermore, the inequalities of Theorems 1 and 2 hold and, in case P^* is pay-off stable, the Pythagorean inequality* (20) *holds with equality for every consistent distribution P.*

Before the proof we point out that $\mathcal{P}$ is not assumed to be convex for this result.

Proof. As $\Pi(P^*) \leq D(P^*\|P_0)$ holds generally, we conclude from the inequality $\Pi(P^*) \geq D(P^*\|P_0)$ that $\Pi(P^*) = D(P^*\|P_0) = \Pi_{\max} = D_{\min}$. The inequality (20) follows as $\Pi(P^*) \geq D_{\min}$.

And (22) follows as, for any estimator R with $D(P^*\|R) < \infty$,

$$\Pi(R) + D(P^*\|R) \leq D(P^*\|P_0 \rightsquigarrow R) + D(P^*\|R) = D(P^*\|P_0) = D_{\min}.$$

(Note that for the first equality sign we need the assumption $D(P^*\|R) < \infty$).

The uniqueness assertions are now trivial and the case of equality in (20) when P^* is pay-off stable is also easily derived. □

As an important trivial example, indicative of the scope of GTE for the $D_{\min}$-game, Harremoës [19] points out that if P_0 and A with $P_0(A) > 0$ are given, then the conditional distribution $P^* = P_0(\cdot|A)$ is the I-projection of P_0 on the family of distributions supported by A. This follows as $P_0(\cdot|A) \in \mathcal{P}$ is pay-off stable. Thus we should think of information projections as a natual and far reaching generalization and companion of normal conditioning in probability theory.

Recent research, cf. Harremoës [19], has pointed to the possibility and also the need for a more thorough analysis and the present exposition should therefore be considered as a quick introduction to the basic results. What is required is a more subtle definition of estimation gain, aiming more at (21) than relying on the simpler equation (19). Furthermore, one should extend the notion of consistent distributions and work more with *essentially consistent* distributions, distributions P^* for which there exists a sequence $(P_n)_{n\geq 1}$ of consistent distributions which converges in divergence to P^*.

Acknowledgements

The author has discussed questions of terminology with Peter Harremoës.

References

[1] S. I. Amari. Information geometry on hierarchy of probability distributions. *IEEE Trans. Inform. Theory*, 47:1701–1711, 2001.

[2] D. Applebaum. *Probability and Information. An integrated approach.* Cambridge Univ. Press, Cambridge, 1996.

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

[4] A. R. Barron. Entropy and the central limit theorem. *Ann. Probab.*, 14(1):336–342, 1986.

[5] F. Bellini and M. Frittelli. On the existence of minimax martingale measures. *Mathematical Finance*, 12:1–21, 2002.

[6] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex

programming. *USSR Comput. Math. and Math. Phys.*, 7:200–217, 1967. Translated from russian.

[7] M. Broom. Using game theory to model the evolution of information: An illustrative game. *Entropy*, 4:35–46, 2002. ONLINE at http://www.unibas.ch/mdpi/entropy/.

[8] N. N. Čencov. *Statistical decision rules and optimal inference.* Nauka, Moscow, 1972. In russian, translation in "Translations of Mathematical Monographs", 53. American Mathematical Society, 1982.

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

[10] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3:146–158, 1975.

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

[12] I. Csiszár. Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. *Ann. Stat.*, 19:2032–2066, 1991.

[13] I. Csiszár and F. Matúš. Information projections revisited. In *Proceedings ISIT2000*, page 490. IEEE, 2000.

[14] B. de Finetti. *Theory of Probability.* Wiley, London, 1974. Italian original 1970.

[15] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Jones and Bartlett Publishers International, Boston, 1993.

[16] M. J. Donald. On the relative entropy. *Commun. Math. Phys.*, 105:13–34, 1985.

[17] Z. Drezner and H. Hamacher, editors. *Facility location. Applications and Theory.* Springer, Berlin, 2002.

[18] Peter D. Grünwald and A. Philip Dawid. Game theory, maximum entropy, minimum discrepancy, and robust bayesian decision theory. *Annals of Mathematical Statistics*, xx:xxx–xxx, 2003.

[19] P. Harremoës. Information topologies with applications. 2002. submitted for publication.

[20] P. Harremoës and F. Topsøe. Unified approach to optimization techniques in shannon theory. In *Proceedings, 2002 IEEE International Symposium on Information Theory*, page 238. IEEE, 2002.

[21] P. Harremoës and F. Topsøe. Maximum entropy fundamentals. *Entropy*, 3:191–226, Sept. 2001. http://www.unibas.ch/mdpi/entropy/ [ONLINE].

[22] Peter Harremoës. Binomial and Poisson distributions as maximum entropy distributions. *IEEE Trans. Inform. Theory*, 47(5):2039–2041, July 2001.

[23] D. Haussler. A general minimax result for relative entropy. *IEEE Trans. Inform. Theory*, 43:1276–1280, 1997.

[24] Alexander S. Holevo. *Statistical Structure of Quantum Theory.* Springer, Berlin, 2001.

[25] E. T. Jaynes. Webpage maintained by L. Brethorst, dedicated to Jaynes work, available ONLINE from http://bayes.wustl.edu.

[26] E. T. Jaynes. Information theory and statistical mechanics, I and II. *Physical Reviews*, 106 and 108:620–630 and 171–190, 1957.

[27] E. T. Jaynes. Clearing up mysteries – the original goal. In J. Skilling, editor, *Maximum Entropy and Bayesian Methods.* Kluwer, Dordrecht, 1989.

[28] E. T. Jaynes. *Probability Theory - The Logic of Science.* Cambridge University Press, Cambridge, 2003.

[29] A. Jessop. *Informed Assessments, an Introduction to Information, Entropy and Statistics.* Ellis Horwood, New York, 1995.

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

[31] D. Kazakos. Robust noiceless source coding through a game theoretic approach. *IEEE Trans. Inform. Theory*, 29:577–583, 1983.

[32] J. L. Kelly. A new interpretation of information rate. *Bell System Technical Journal*, 35:917–926, 1956.

[33] J. Kisynski. Convergence du typè $\updownarrow$. *Colloq. Math.*, 7:205–211, 1960.

[34] A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer, Berlin, 1933.

[35] S. Kullback. *Informaton Theory and Statistics.* Wiley, New York, 1959.

[36] S. Kullback and R. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22:79–86, 1951.

[37] Yu. V. Linnik. An information-theoretic proof of the central limit theorem with lindeberg condition. *Theory Probab. Appl.*, 4:288–299, 1959.

[38] Masanori Ohya and Dénes Petz. *Quantum Entropy and Its Use.* Springer, Berlin Heidelberg New York, 1993.

[39] E. Pfaffelhuber. Minimax information gain and minimum discrimination principle. In I. Csiszár and P. Elias, editors, *Topics in Information Theory*, volume 16 of *Colloquia Mathematica Societatis János Bolyai*, pages 493–519. János Bolyai Mathematical Society and North-Holland, 1977.

[40] B. Ya. Ryabko. Comments on "a source matching approach to finding minimax codes". *IEEE Trans. Inform. Theory*, 27:780–781, 1981. Including also the ensuing Editor's Note.

[41] F. Delbaen P. Grandits T. Rheinlaender D. Samperi M. Schweizer and C. Stricker. Exponential hedging and entropic penalties. *Mathematical Finance*, 12:99–123, 2002.

[42] G. Shafer and V. Vovk. *Probability and finance. It's only a game!* Wiley, Chichester, 2001.

[43] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423 and 623–656, 1948.

[44] J. J. Sylvester. A question in the geometry of situation. *Quarterly Journal of Pure and Applied Mathematics*, 1:79, 1857.

[45] F. Topsøe. A new proof of a result concerning computation of the capacity for a discrete channel. *Z. Wahrscheinlichkeitstheorie verw. Geb.*, 22:166–168, 1972.

[46] F. Topsøe. Game theoretical equilibrium, maximum entropy and minimum information discrimination. In A. Mohammad-Djafari and G. Demoments, editors, *Maximum Entropy and Bayesian Methods*, pages 15–23. Kluwer Academic Publishers, Dordrecht, Boston, London, 1993.

[47] F. Topsøe. Basic concepts, identities and inequalities – the toolkit of information theory. *Entropy*, 3:162–190, 2001. http://www.unibas.ch/mdpi/entropy/ [ONLINE].

[48] F. Topsøe. Maximum entropy versus minimum risk and applications to some classical discrete distributions. *IEEE Trans. Inform. Theory*, 48:2368–2376, 2002.

[49] Flemming Topsøe. An information theoretical identity and a problem involving capacity. *Studia Scientiarum Mathematicarum Hungarica*, 2:291–292, 1967.

[50] Flemming Topsøe. Information theoretical optimization techniques. *Kybernetika*, 15(1):8–27, 1979.

[51] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior.* Princeton University Press, Princeton, 1947. 2nd. edition.