OPEN ACCESS

1st International Electronic Conference on Entropy and Its Applications
3 - 21 November 2014

www.sciforum.net/conference/ecea-1

Conference Proceedings Paper – Entropy

A Quantitative Theory of Cognition with Applications

Flemming Topsøe

University of Copenhagen, Department of Mathematical Sciences, Universitetsparken 5, 2100 Copenhagen, Denmark
E-mail: topsoe@math.ku.dk

Received: 20 August 2014 / Accepted: 26 October 2014 / Published: 3 November 2014

Abstract: *Cognition* is based partly on *belief* and strives for *knowledge* of *"truth"*. *Description* and the derived notion of *control* lead to notions of *entropy* and *divergence* and help to clarify what can be known, the *"knowable"*, by setting limits to *information*. Via game theoretical considerations, a notion of *core* is introduced which, in classical settings, is related to that of *exponential families*. It facilitates the process of *inference* in concrete cases of interest.
The basic setting is abstract. When we turn to probabilistic modelling, *interaction* between truth, belief and knowledge is essential for an interpretation of *Tsallis entropy*.

Keywords: truth, belief, knowledge, description, entropy, divergence, proper effort functions, fundamental inequality, core, Tsallis entropy.

PACS classifications: 05.90.+m, 89.70.Cf

MSC classifications: 81P05, 94A15, 94A17

Introduction

The driving force behind the present study has been to overcome the difficulties you encounter when trying to extend the clear and convincing operational interpretations associated with classical information theory as developed by Shannon [1] and followers, to the theory promoted by Tsallis for statistical physics and thermodynamics, cf. [2], [3]. That there are difficulties is witnessed by the fact that some physicists do not recognize the new theory as sound - despite the apparent success of Tsallis and hes followers. Evidence of this attitude may be found in Gross [4] and in Shalizi's notes [5].

As it will turn out, if you accept a certain kind of *interaction* between *truth, belief* and *knowledge*, you are led in a natural way to the family of *Tsallis entropies*, cf. Section 21. Further study revealed that the philosophical elements of the indicated approach make sense in a much wider setting than originally intended. One does not achieve the same degree of clarity as in classical Shannon theory, where *coding* provides a solid reference. However, we shall demonstrate that the extension to a more abstract framework is meaningful and opens up for new areas of research. In addition, known results are consolidated and unified.

Our study falls into two parts. In Part I, an abstract theory of *information without probability* is presented. It is based on somewhat speculative considerations which, taken together, constitute possible *paradigms of cognition*. Inspiration from Shannon Theory and from the theory of inference within statistics and statistical physics is apparent. However, the ideas are here presented as an independent theory.

Previous endeavours in the direction taken includes research by Ingarden and Urbanik [6] who wrote *"... information seems intuitively a much simpler and more elementary notion than that of probability ... [it] represents a more primary step of knowledge than that of cognition of probability ..."*. We also point to Kolmogorov, cf. [7] and [8] who in the latter reference (but going back to 1970 it seems) stated *"Information theory must preceed probability theory and not be based on it"*. The ideas by Ingarden and Urbanik were taken up by Kampé de Fériet, see the survey [9]. The work of Kampé de Fériet is rooted in logic. Logic is also a key ingredient in comprehensive studies over some 40 years by Jaynes, collected posthumously in [10]. Though many philosophically oriented discussions are contained in the work of Jaynes, the situations dealt with are probabilistic in nature and intended mainly for a study of statistical physics.

In *complexity theory* as developed by Solomonoff, Kolmogorov and others, cf. the recent survey [11] by Rathmanner and Hutter, we have a highly theoretical discipline which aims at inference not necessarily tied to probabilistic modelling. The *Minimum Description Length Principle* may be considered an important spin-off of this theory. It is mainly directed at problems of statistical inference and was developed, primarily, by Rissanen and by Barron and Yu, cf. [12]. We also point to the treatise [13] by Grünwald. There you find discussions of many of the issues dealt with here, including a discussion of the work of Jaynes.

Still other areas of research have a bearing on "information without probability", e.g. *semiotics*, *philosophy of information*, *pragmatism*, *symbolic linguistics*, *placebo research*, *social information* and *learning theory*. Many areas within *psychology* are also of relevance. Some specific works of interest include Jumarie [14], Shafer and Vovk [15], Gernert [16], Bundesen and Habekost [17], Benedetti [18] and Brier [19]. The handbook [20] edited by Adriaans and Bentham and the encyklopedia article [21] by Adriaans collect views on the very concept of information. Over the years, an overwhelming amount of thoughts has been devoted to that concept in one form or another. Most of this bulk of material is entirely philosophical and not open to quantitative analysis. Part of it is impractical and presently mainly of theoretical interest. And some is far from Shannon's theory which we hold as a corner stone of quantitative information theory. In fact, we consider it a requirement of any quantitative theory of information to be downward compatible with basic parts of Shannon theory. This requirement is largely respected in the present work. But not entirely. For example, it is doubtful if one can meaningfully lift

the concept of *coding* as known from Shannon theory to a more abstract level. Likewise, the notion of *conditioning* and the concept of *mutual information* may best be studied in an abstract setting after introducing more structure.

We thus attempt *"to go beyond Shannon"*. So does e.g. Brier in his development of *cybersemiotics*, cf. [22], [19]. Brier goes deeper into some of the philosophical aspects than we do and also attempts a broad coverage by incorporating not only the exact natural sciences but also life science, the humanities and the social sciences. On the other hand, our study aims at more concrete results by basing the study more directly on quantitative elements. Both studies emphasize the role of the individual in the cognitive process. Further studies of what appears to be an intensified field of research may lead to a certain unification and general agreement.

A special feature of our development is the appeal to basic game theoretical considerations, cf. especially Sections 10 and 11. To illuminate the importance we attach to this aspect we quote from Jaynes preface to [10] where he comments on the *maximum entropy principle*, the central principle of inference promoted by Jaynes:

"*... it [maximum entropy] predicts only observable facts (functions of future or past observations) rather than values of parameters which may exist only in our imagination ... it protects us against drawing conclusions not warranted by the data. But when the information is extremely vague, it may be difficult to define any appropriate sample space, and one may wonder whether still more primitive principles than maximum entropy can be found. There is room for much new creative thought here.* "
This is where game theory comes in. It represents a main addition, we claim, to Jaynes' work [1]. The merits of game theory in relation to information theoretical inference were presented in the probabilistic, Shannon-like setting, independently of eachother, by Pfaffelhuber [24] and the author [25]. These works were often overlooked by subsequent authors[2]. More recent references include Harremoës and Topsøe [26], Grünwald and Dawid [27], Friedman et al [28] (a utility-based work) and Dayi [29]. As sources of background material, [30], [31] and [32] may be helpful.

Apart from introducing game theory into the picture, a main feature of the present work lies in its abstract nature with a focus on interpretations rather than on axiomatics which was the emphasis of many previous authors, including Jaynes.

Part II is devoted to applications and may be viewed as a justification of the partly speculative deliberations of Part I. The applications come from combinatorial geometry, probabilistic information theory, statistics and statistical physics. For most of them, we focus on providing the key notions needed for the theory to work, thus largely leaving concrete applications aside. The aim is to provide enough details in order to demonstrate that our modelling can be applied in quite different contexts. For the case of discrete probabilistic models we do, however, embark on a more thorough analysis. The reason is, firstly, that this is what triggered the research reported on and, secondly, with a thorough discussion of

---

[1] At the conference "Maximum Entropy and Bayesian Methods", Paris 1993, the author had much hoped to discuss the impact of game theoretical reasoning with professor Jaynes. Unfortunately, Jaynes, who died in 1998, was too ill at the time to participate and thus did not take into account arguments such as those in [23] which support the theory developed by Jaynes.

[2] admittedly, also the present author only some 15 years ago became aware of [24]!

modelling in this context, virtually all elements introduced in Part I have a clear and natural meaning. In fact, full appreciation of the abstract theory may only be achieved after reading the material in Section 21.

Our treatment is formally developed independently of previous research. However, unconsciously or not, it depends on earlier studies as referred to above and on the tradition developed over time. More specifically, we mention that our focus on *description effort*, especially the notion of *properness*, cf. Section 6, is closely related to ideas first developed for areas touching on meteorology, statistics and information theory. In Sections 6 and 15 we comment on this in more detail.

Finally, we mention that [33], [34], [35] and [36] are forerunners of the present work.

Part I
Information without Probability

1. The world and you

By Ω we denote the *world*, more precisely the *actual world*, perhaps one among several *conceivable worlds*. Two fictive persons play a major role in our modelling, *"Nature"* and *"Observer"*. The interplay between the two takes place in relation to studies of *situations* from the world. Nature is seen as an expression of the world itself and reflects the *rules of the world*. Observer seeks *knowledge* about situations studied. It may be helpful to think of Observer as "you". Somewhat stereotypical, we take Nature to be female, Observer male.

The *knowledge* sought by Observer aims at *inference* concerning particular situations under study. A higher form of inference may also be possible if Observer does not know the rules of the world, in other words, does not know which world he is placed in. Then, having a reservoir of *conceivable worlds* in mind, and based on experience from the study of several situations, Observer may attempt to infer which one is the actual world.

We think of Ω as limited in some sense, a *partial world*. This appears to be the most realistic. In principle, one could consider all kinds of phenomena at the same time, say of a statistical, physical, social, psychological or other nature. However, the rules of the world may vary from context to context and – if you do not take these rules as absolutes – even from one Observer to another. A finer modelling than here considered may bring the notion of *context* more prominently into the picture.

The notions introduced are left as loose indications. They will take more shape as the modelling progresses. The terminology chosen here and later on is intended to provoke associations to common day experiences of the cognitive process. In addition, the terminology is largely consistent with usage in philosophy.

2. Truth and Belief

Nature, as an expression of the fixed rules of the world, does not have a mind. She is the holder of *truth*. Observer seeks the truth but is relegated to *belief*. However, Observer possesses a conscious and creative mind which can be exploited with the goal to obtain *knowledge* as effortlessly as possible.

We introduce two non-empty sets, X, the *state space*, and Y, the *belief reservoir*. Elements of X, generically denoted by x, are *truth instances* or *states of truth* or just *states*, whereas elements of Y, generically denoted by y, are *belief instances*. Typically, in any situation, we imagine that Nature chooses a state and that Observer chooses a belief instance. This leads to the introduction of certain games which will be studied systematically at a later point, starting with Section 10.

We assume that $Y \supseteq X$. Therefore, in any situation, it is conceivable that Observer actually believes what is true. Often, $Y = X$ will hold. Then, whatever Observer believes, *could* be true.

Though there may be no such thing as *absolute truth*, it is tempting to imagine that there is and to think of Natures choice as an expression of just that. We shall not attempt to model the mechanisms behind Natures choice. Later on, we open up for the possibility that somehow Observer may influence Natures choice. In any case, the interplay between Nature and Observer is a key to our modelling.

In any specific situation, Natures choice is not free within all of X but restricted to a non-empty subset $\mathcal{P}$ of X, the *preparation*. This set depends on the particular situation studied. The idea is that Observer, perhaps a physicist, can "prepare" a situation, thereby forcing Nature to restrict her choice of state accordingly. For instance, by placing a gas in a heat bath, Nature is restricted to states which have a mean energy consistent with the prescribed temperature.

A situation is normally characterized by specifying a preparation. However, further details, especially regarding Observers behaviour may also be included in the modelling of "a situation". A state x is *consistent* – viz. consistent with the preparation $\mathcal{P}$ of the situation – if $x \in \mathcal{P}$. Later on, we shall consider *preparation families* which are sets, generically denoted by $\mathbb{P}$, whoes members are preparations.

Faced with a specific situation with preparation $\mathcal{P}$, Observer speculates about the state of truth chosen by Nature. He may express his opinion by assigning a belief instance to the situation. If he insists on choosing this instance from the preparation $\mathcal{P}$, Observer will only believe what *could* be true. Sometimes, Observer may prefer to assign a belief instance in $Y \setminus \mathcal{P}$ (or even in $Y \setminus X$) to the situation. Then this instance cannot possibly be one chosen by Nature. Nevertheless, it may be an adequate choice if an instance in $\mathcal{P}$ would contradict Observers subjective beliefs. Therefore, the chosen instance may be the "closest" to the actual truth instance in some subjective sense. Anyhow, Observers choice of belief instance is considered a subjective choice which takes available information into account such as general insight and any *prior knowledge*. Qualitatively, these thoughts agree with Bayesian thinking, and as such enjoy the merits, but are also subject to the standard criticism, which applies to this line of thought, cf. [11] and [37].

Our modelling may involve a set of *certain beliefs*, a subset Y_{det} of Y. Beliefs from Y_{det} are chosen by Observer if he is quite determined on what is going on – but, of course, Observer could be wrong. If we do not find it appropriate to work with certain beliefs, we formally put $Y_{\mathrm{det}} = \emptyset$.

3. A tendency to act, a wish to control

Two ways will lead us to new and important structural elements. These elements will be considered identical for the present modelling. Finer modelling may later change that.

First, we point to the mantra that *belief is a tendency to act*. This is a rewording taken from Good [38] who suggested this point of view as a possible interpretation of the notion of *belief*. In daily life, action

appears more often than not to be a spontaneous reaction to situations man is faced with, rather than a result of rational considerations or, the reaction depends on psychological factors or brain activity largely outside conscious control. Here, we shall appeal to rational thinking based on quantitative considerations. Precise details will have to wait until Section 6. Right now we introduce the basic structural elements which will facillitate the further modelling. To do so, we shall work with a set $\hat{Y}$, the *action space*, and a map from Y into $\hat{Y}$, referred to as *response*. Elements of $\hat{Y}$ are *actions*. We use the notation $y \mapsto \hat{y}$ to indicate the action which is Observers response in situations where hes belief is represented by the belief instance y.

Response need not be injective, thus it is in general not possible to infer Observers beliefs from hes actions. Elements of Y with the same response are said to be *response-equivalent*, notationally written $y_1 \sim y_2$. Response need not either be surjective, though for most applications it will be so. Elements not in the range are idle for the actual model under discussion but may become relevant if the setting is later expanded.

In order to simplify the exposition we have taken response to be an ordinary map. However, for some models, cf. Section 17, it would be appropriate to work with set-valued maps. This will enable Observer to take situations into account where he considers several possible actions to be equally attractive.

Let us turn to another tendency of man, the wish to control. This makes us introduce a set W, the set of *controls*. For the present modelling, we take W and $\hat{Y}$ to be identical: $W = \hat{Y}$. The point of view is that in order to exercise control, Observer has to act, typically by setting up appropriate experiments, and in a rough mathematical model as here suggested we simply identify the two aspects. Later elaborations may change that and lead to a clear distinction between action and the more passive concept of control.

The simplest models are obtained when response is an injection or even a bijection. And simplest among these models are the cases when $Y = \hat{Y} = W$ and response is the identity map. This corresponds to a further identification of belief, action and control. Even then it makes a difference if you think about elements as expressions of belief or as expressions of actions necessary to obtain control.

Sometimes, it is technically convenient to assume that W contains a special element, $w_\emptyset$, the *empty action*. This reflects on situations where Observer sees no reason to take any action or to exercise any control. If the modelling involves a non-empty set Y_{det}, we assume that $w_\emptyset \in W$ and that $\hat{y} = w_\emptyset$ for every $y \in Y_{\mathrm{det}}$.

Though many models do not need the introduction of $\hat{Y}$ (and W), the further development will mainly refer to $\hat{Y}$-related concepts. Technically, this results in greater generality, as response need not be injective. Belief-type concepts, often indicated by pointing to the "Y-domain" will primarily be derived from action- or control-based concepts, often indicated by pointing to the "$\hat{Y}$-domain". The qualifying indication may be omitted if it is clear whether we work in the one domain or the other.

4. Atomic situations, Controlability and Visibility

The two closely connected relations to be introduced in this section constitute refinements which may be disregarded at a first reading. This can be done by taking the relations to be the *diffuse relations* (in notation below, $X \otimes \hat{Y} = X \times \hat{Y}$ and $X \otimes Y = X \times Y$).

Elements of $\hat{Y}$ will below mainly be conceived as controls.

Pairs of states and belief instances or of states and controls are key ingredients in situations from the world. However, not all pairs will be allowed. Instead, we imagine that offhand, Observer has some limited insight into Natures behaviour and therefore, Observer takes care not to associate "completely stupid" belief instances or controls, as the case may be, with situations of interest.

We express these ideas in the $\hat{Y}$-domain by introducing a relation from X to $\hat{Y}$, called *controlability* and denoted $X \otimes \hat{Y}$. Thus $X \otimes \hat{Y}$ is a subset of the product set $X \times \hat{Y}$. Elements of $X \otimes \hat{Y}$ are *atomic situations* (in the $\hat{Y}$-domain). An atomic situation (x, w) is an *adapted pair* if w is *adapted* to x in the sense that $w = \hat{x}$. If (x, w) is an atomic situation, we write $w \succ x$ and say that w *controls* x or that x can be *controled* by w.

Let $\mathcal{P}$ be a preparation. We write $w \succ \mathcal{P}$, and call w a *control point* (or just a *control*) of $\mathcal{P}$, if $w \succ x$ for every $x \in \mathcal{P}$. By $\hat{}[\mathcal{P}]$ we denote the set of all control points of $\mathcal{P}$. We write $\hat{}[x]$ if $\mathcal{P}$ is the singleton set $\{x\}$. For $w \in \hat{Y}$, $]w[$ denotes the *control region* of w, the set of $x \in X$ for which $w \succ x$. Clearly, $w \in \hat{}[\mathcal{P}]$, $w \succ \mathcal{P}$ and $\mathcal{P} \subseteq]w[$ are equivalent statements. Sometimes we consider the *restriction* $\mathcal{P} \otimes \hat{Y}$ which consists of all atomic situations (x, w) with $x \in \mathcal{P}$.

We assume that the following conditions hold:

$$\forall x \in X : \hat{x} \succ x\,, \tag{1}$$

$$\forall w \in \hat{Y} :]w[\neq \emptyset \tag{2}$$

and, normally also that

$$\exists y \in Y : \hat{y} \succ X. \tag{3}$$

The first condition is essential and the second rather innocent. The third condition is introduced when we want to ensure that X is not "too large". Models where (3) does not hold are considered unrealistic, beyond what man (Observer) can grasp. If response is surjective, it amounts to the condition $\hat{}[X] \neq \emptyset$.

Corresponding to controlability, we consider the derived relation of *visibility* in the Y-domain, denoted $X \otimes Y$ and given by

$$X \otimes Y = \{(x, y) \in X \times Y | \hat{y} \succ x\}\,. \tag{4}$$

We use the same sign, $\succ$, for visibility as for controlability. The context will have to show if we work in the $\hat{Y}$- or in the Y-domain. We find that $y \succ x$ if and only if $\hat{y} \succ x$. If this condition holds, we say that y *covers* x or that x is *visible* from y. Pairs in $X \otimes Y$ are *atomic situations* (in the Y-domain). An atomic situation (x, y) is an *adapted pair* if $(x, \hat{y})$ is so in the $\hat{Y}$-domain, i.e. if $y \sim x$. And (x, y) is a *perfect match* if $y = x$. The two notions coincide if response is injective. An atomic situation (x, y) is a *situation of certainty* if $y \in Y_{\text{det}}$.

By (1), $x \succ x$ for all $x \in X$, thus $X \otimes Y$ contains the diagonal $X \times X$. The *outlook* (or *view*) from $y \in Y$ is the set $]y[= \{x | y \succ x\}$. Clearly, $]y[=]\hat{y}[$. By definition (4) and by (2), this set is non-empty and, when (3) holds, for at least one belief instance, the outlook is all of X.

For a preparation $\mathcal{P}$ we write $y \succ \mathcal{P}$, and call y a *view point* of $\mathcal{P}$, if $y \succ x$ for every $x \in \mathcal{P}$. The set of all view points of $\mathcal{P}$ is denoted $[\mathcal{P}]$. We write $[x]$ if $\mathcal{P}$ is the singleton $\mathcal{P} = \{x\}$. By $\mathrm{ctr}(\mathcal{P})$, the *centre of* $\mathcal{P}$, we denote the set of view points in the preparation, i.e. $\mathrm{ctr}(\mathcal{P}) = \mathcal{P} \cap [\mathcal{P}]$. This set may be empty.

Restrictions $\mathcal{P} \otimes Y = \{(x, y) \in X \otimes Y | x \in \mathcal{P}\}$ are at times of relevance.

In any situation, Observer should ensure that from his chosen belief instance, every state which could conceivably be chosen by Nature is visible. Therefore, in a situation where the preparation $\mathcal{P}$ is known to Observer, Observer should only consider belief instances in $[\mathcal{P}]$. Indeed, if Observer chooses a belief instance $y \in Y \setminus [\mathcal{P}]$, there is a risk that Natures choice will be a truth instance which is not visible from y – and, guided as we shall be, by the cautious principle that "what can go wrong *does* go wrong" – this will not be acceptable to Observer.

In the sequal we shall often consider bivariate functions, generically denoted by either $\hat{f}$ ($\hat{Y}$-domain) or by f (Y-domain). The $\hat{f}$-type functions are defined either on $X \otimes \hat{Y}$ or on some subset of $X \otimes \hat{Y}$ of the form $\mathcal{P} \times \hat{} [\mathcal{P}]$ for some preparation $\mathcal{P}$. The range of $\hat{f}$ may be arbitrary, a subset of the extended real line or some abstract set. Given $\hat{f}$, it is understood that f denotes the *derived function* defined by $f(x, y) = \hat{f}(x, \hat{y})$ on pairs (x, y) for which $(x, \hat{y})$ is in the domain of definition of $\hat{f}$. Clearly, the domain of definition of the derived function is either $X \otimes Y$ or the set $\mathcal{P} \times [\mathcal{P}]$ if $\hat{f}$ is defined on $\mathcal{P} \times \hat{} [\mathcal{P}]$.

Every derived function is *response-only dependent*, i.e. the value does not change if the belief instance entering in the definition is changed to an response-equivalent one. If response is a surjection, there is a natural one-to-one relation between $\hat{Y}$-type functions and response-only dependent Y-type functions.

Consider an f-type function defined on all of $X \otimes Y$. For $y \in Y$, f^y denotes the *marginal function given* y, defined on $]y[$ by $f^y(x) = f(x, y)$. Occasionally, we also need the *marginal function given* $x \in X$. Notation and defining relation is $f_x(y) = f(x, y)$ for $y \in [x]$. We write $f^y < \infty$ on $\mathcal{P}$ to express, firstly, that $y \succ \mathcal{P}$ so that f^y is well defined on all of $\mathcal{P}$ and, secondly, that this marginal function is finite on $\mathcal{P}$. We write $f^y < \infty$ if $f^y < \infty$ on X.

5. Knowledge, Perception and Interaction

Observer strives for *knowledge*, conceived as the *synthesis of extensive experience*. Referring to probabilistic thinking, we could point to situations where accidental experimental data are smoothed out over time as you enter the regime of the law of large numbers. However, Observers endeavours may result in less definitive insight, a more immediate reaction which we refer to as *perception*. It reflects how Observer *perceives* situations from the world or, with a different focus, how situations from the world are *presented* to Observer.

In the same way as we have introduced truth- and belief instances, we shall also consider *knowledge instances*, also referred to as *perceptions*. Typically, they are denoted by z and taken from a set denoted Z, the *knowledge base* (or *perception base*).

A central and simplifying assumption for our modelling is that the rules of the world Ω contain a special function, $\hat{\Pi}$ which maps $X \otimes \hat{Y}$ into Z, generically, $z = \hat{\Pi}(x, w)$. The derived function, Π, then maps $X \otimes Y$ into Z. Both functions are referred to as the *interactor*. The context will show which one we have in mind, $\hat{\Pi}$ or Π.

Thus knowledge can be derived deterministically from truth and belief alone, and as far as belief is concerned, we only have to know the associated response. In terms of perception, Observers perception z of an atomic situation (x, y) is given by the formula $z = \Pi(x, y)$.

In the present study, we consider the world as characterized by the associated interactor and we may thus talk about the *world with interactor* Π, $\Omega = \Omega_\Pi$. The rules of the world may contain other elements

than the interactor, but such further elements are not specified in the present study. Other elements which could be considered in future developments include *context*, *noise from the environment*, and *dynamics*. Such features can to some extent be expressed by defining X, Y and Z appropriately and by introducing suitable interpretations.

In case response is a bijection and Z contains X as well as Y we consider two special conceivable worlds by introducing the interactors Π_1 and Π_0 defined by $\Pi_1(x, y) = x$, respectively $\Pi_0(x, y) = y$. The associated worlds are $\Omega_1 = \Omega_{\Pi_1}$ and $\Omega_0 = \Omega_{\Pi_0}$. In Ω_1, *"what you see is what is true"*, whereas in Ω_0, *"you only see what you believe"*. The world Ω_1 is the *classical world* where, optimisticly, *truth can be learned*, whereas, in Ω_0, you cannot learn anything about truth. We refer to Ω_0 as a *black hole*. It is a narcisistic world, a world of extreme skepticism, only reflecting Observers beliefs and bearing no trace of Nature. If Z is provided with some linear structure, we can introduce a parameter q and consider further interactors Π_q by putting $\Pi_q(x, y) = qx + (1 - q)y$. Worlds associated with these interactors are denoted Ω_q.

The simplest world to grasp is the classical world, but also the worlds Ω_q and even a black hole contain elements which are familiar to us from daily experiences, especially in relation to certain psychological phenomena. In this connection we point to *placebo effects*, cf. Benedetti [18], and to *visual attention*, cf. Bundesen and Habekost [17]. The relevance of our modelling in relation to these phenomena is, presently, purely qualitative.

6. Effort and Description

We turn to the introduction of the key quantitative tool we shall work with. In so doing, we will be guided by the view that *perception requires effort*. Expressed differently, *knowledge is obtained at a cost*. Since, according to the previous section, knowledge can be derived from truth and action, effort can be modelled by a bivariate function defined on $X \otimes \hat{Y}$, the *effort function*. The rules of the world Ω may not point directly to an effort function which Observer can favourably work with. Or there may be several sensible functions to choose from. The actual selection is considered a task left to Observers ingenuity.

As a further speculation, we imagine that effort is derived from *description*. Description is intended to aid Observer in his encounters with situations from the world. Logically, description comes before effort. Effort arises when specific ideas about description are developed into a *method of description*. Such methods we may think of as synonymous with *experiments*. The implementation of a method of description or the performance of the corresponding experiment involves a cost which is specified quantitatively by the effort function.

In order to develop these ideas further, it appears desirable to study more closely the nature of description. We shall not enter into that here, only remark that it seems that description is essentially quantitative. Fact is that though we often think of description in loose qualitative terms, a closer view will show that in order to develop precise concepts which can be communicated among humans, quantitative elements will inevitably be involved. This may be based on a finite set of *descriptors*, real-valued functions defined on X.

Imagine now that somehow Observer has chosen all elements needed – response, actions, experiments – and settled for an effort function, $\hat{\Phi}$. Let us agree on what a "good" effort function should mean. Generally speaking, Observer should of course aim at experiments with a low associated effort. To reach more detailed criteria of "goodness" , consider a fixed truth instance x and the various possible actions w, in principle free to be any action which controls x. It appears desirable that the action adapted to x is the one preferred by Observer. Thus effort should be minimal in this case, i.e. $\hat{\Phi}(x, w) \geq \hat{\Phi}(x, \hat{x})$ should hold. Further, if the inequality is sharp except for the adapted action, this will have a *training effect* and hopefully over time encourage Observer to choose the optimal action, $\hat{x}$.

Formally, we define a $\hat{Y}$*-effort function* as a function $\hat{\Phi}$ on $X \otimes \hat{Y}$ with values in $]-\infty, +\infty]$ such that

$$\hat{\Phi}(x, w) \geq \hat{\Phi}(x, \hat{x}) \text{ for all } (x, w) \in X \otimes \hat{Y} \,. \tag{5}$$

If $w_\emptyset \in \hat{Y}$, we also require that $\hat{\Phi}(x, w_\emptyset) = 0$ when $w_\emptyset \succ x$. The effort function is *proper*, if equality holds in (5) only if either $\hat{\Phi}(x, \hat{x}) = \infty$ or else w is adapted to x, $w = \hat{x}$.

Note that effort may be negative (but not $-\infty$). This flexibility will later be convenient as it will allow us to pass freely between notions of effort and notions of utility by a simple change of sign. But normally, effort functions will be non-negative.

We define a Y*-effort function* as a function $\Phi : X \otimes Y \mapsto]-\infty, \infty]$ such that

$$\Phi(x, y) \geq \Phi(x, x) \text{ for all } (x, y) \in X \otimes Y \,. \tag{6}$$

If $Y_{\text{det}} \neq \emptyset$, we require that Φ vanishes for any atomic situation of certainty. The effort function is *proper* if equality holds in (6) only if either $\Phi(x, x) = \infty$ or else there is a perfect match, $y = x$. These notions are defined directly with reference to the Y-domain. However, it lies nearby also to consider functions which can be derived from $\hat{Y}$-effort functions $\hat{\Phi}$. They are *derived effort functions* and, in case $\hat{\Phi}$ is proper, *proper derived effort functions*. The two strategies for definitions, intrinsic and via derivation, give somewhat different concepts. In case response is injective, the resulting notions are equivalent. In general, derived effort functions are response-only dependent and, in the other direction, for a proper derived effort function, you can only conclude response-equivalence, $y \sim x$, if $\Phi(x, y) = \Phi(x, x)$ and $\Phi(x, x) < \infty$. We shall talk about effort functions without a qualifying prefix, $\hat{Y}$ or Y, if it is clear from the context what we have in mind. We shall always point out if we have derived functions in mind.

The effort functions introduced determine *net effort*. However, the implementation of the method of description – which we imagine lies behind – may, in addition to a specific cost, entail a certain *overhead* and, occasionally, it is appropriate to include this overhead in the effort. We refer to Section 21, (109) for an important instance of theis.

Two effort functions $\hat{\Phi}_1$ and $\hat{\Phi}_2$, or Φ_1 and Φ_2, which only differ from each other by a positive scalar factor are *scalarly equivalent*. There may be many non-scalarly equivalent effort functions for Observer to choose from. The choice among scalarly equivalent ones amounts to a determination of a *unit of effort*. If an effort function is proper, so is every scalarly equivalent one.

We imagine that the choice of effort function involves considerations related to knowledge and to the rules of the world. However, once $\hat{\Phi}$, hence also Φ are fixed, these other elements are only present indirectly. They will not appear for the remainder of Part I. The ideas of Section 5 have thus mainly

served as motivation for the further abstract development. The ideas will be taken up again when in Section 21 we turn to a study of probabilistic models.

The author was led to consider proper effort functions in order to illuminate certain aspects of statistical physics, cf. [33], [36]. However, the ideas have been around for quite some time, especially among statisticians. For them it has been more natural to work with functions taken with the reverse sign by looking at "score" rather than effort. Our notion of proper effort functions, when specialized to a probabilistic setting, matches the notion of *proper scoring rules* as you find it in the statistical literature. As to the literature, Csiszár [39] comments on the early sources, including Brier [40], a forerunner of research which followed, and Good [38], Savage [41] (see e.g. Section 9.4) and Fischer [42]. See also the reference work [43] by Gneiting and Raftery. For research of Dawid and collaborators – much in line with what you find here and in [25] – see [44], [45] and [46].

Regarding terminology we shall at times say that a Y-effort function satisfies the *perfect match principle* if it is proper. Fact is that the word "proper" does not say that much. The word was chosen to fit in with previous and current usage in the statistical literature just pointed to.

7. Basic Concepts of Information, Information Triples

Information in any particular situation concerns truth. If $\mathcal{P}$ is a preparation, "$x \in \mathcal{P}$" signifies that the true state is to be found among the states in $\mathcal{P}$. If $\mathcal{P}$ is a singleton, we talk about *full information* and use the notation "x" rather than "$x \in \{x\}$"; otherwise, we talk about *partial information*.

We shall not be concerned with how information can be obtained – if at all. Perhaps, Observer only speculates about the potential possibility of acquiring information, either through his own activity or otherwise, e.g. via the involvement of an aid or a third party, an *informer*[3].

We shall connect information with quantitative considerations and take as base a proper effort function $\hat{\Phi}$. Following Shannon we disregard semantic content. Instead, we focus on the possibility for Observer to benefit from information by a saving of effort. Accordingly, we view $\hat{\Phi}(x, w)$ as the information content of "x" in an atomic situation with x as truth instance and w as action – indeed, if you are told that x is the true state, you need not allocate the effort $\hat{\Phi}(x, w)$ to the situation which you were otherwise prepared to do. The somewhat intangible and elusive concept of "information" is thus measured by the more concrete and physical notion of effort. The unit of information is, therefore, the same as the unit used for effort.

There is a huge literature elucidating what information really "*is*". Suffice it here to refer to [20] and, as an example of a discussion more closely targeted on our main themes, we refer to Caticha [47] who maintains that "*Just as a force is defined as that which induces a change in motion, so information is that which induces a change in beliefs*". One may just as well, we find, talk about a change of action.

The undisputed central concept of the theory developed by Shannon is that of *entropy*. In our general abstract setting, entropy also makes sense. One possible interpretation is as *guaranteed saving of effort*. With effort given by $\hat{\Phi}$ we are led to define, for any state x, the *entropy of* x – understood as the entropy

---

[3] For example, at the airport, you may speculate about the departure time of your flight when you hear the announcement that "the flight to Copenhagen departs at 4 p.m."

associated with the information "x" – as the minimum over w of $\hat{\Phi}(x,w)$. By the defining property (5), this equals $\hat{\Phi}(x,\hat{x})$. Denoting entropy by H, we therefore have

$$\mathrm{H}(x) = \hat{\Phi}(x,\hat{x})\,. \tag{7}$$

The motivating consideration makes most sense if, one way or another, Observer eventually obtains full information about the true state. However, if instead you view entropy as *necessary allocation of effort* understood as the effort you have to reserve in order to have any chance to obtain full information, it does not appear important actually to obtain that information.[4] As yet a third route to entropy we suggest to view it as a quantitative expression of the *complexity* of the various states. To evaluate this, Observer may suggest to use *minimal accepted effort*, the effort he is willing to allocate to the various states.

Entropy may also be obtained with reference only to the Y-domain. Indeed, with Φ the derived effort function, for each state x,

$$\mathrm{H}(x) = \Phi(x,x)\,.$$

Whichever route to entropy you take – including the game theoretical route of Section 10 – subjective elements will be involved, typically through Observers choice of description and associated experiments. If, modulo scalar equivalence, the actual world only allows one proper effort function, entropy, and notions related to entropy, are of a more objective nature. We shall later see examples of such worlds but even then, subjective elements may enter through inference by Observer regarding which world is the actual one.

Entropy as a notion derived from effort should not be considered in isolation. Apart from effort itself, we turn to the introduction of two other basic concepts which make sense in our abstract setting, viz. *redundancy* for the $\hat{Y}$-domain and its counterpart, *divergence*, for the Y-domain.

We start with redundancy and consider an atomic situation $(x,w) \in X \otimes \hat{Y}$. Then *redundancy* $\hat{\mathrm{D}}$ between x and w is measured by the difference between actual and minimal effort, i.e., ideally, as

$$\hat{\mathrm{D}}(x,w) = \hat{\Phi}(x,w) - \mathrm{H}(x)\,. \tag{8}$$

Assume, for a moment, that H is finite-valued. Then redundancy in (8) is well defined and, by a trivial rewriting of this equation, the three basic quantities, $\hat{\Phi}$, H and $\hat{\mathrm{D}}$ are connected by the *linking identity*

$$\hat{\Phi}(x,w) = \mathrm{H}(x) + \hat{\mathrm{D}}(x,w)\,, \tag{9}$$

valid for any atomic situation $(x,w) \in X \otimes \hat{Y}$. Furthermore, redundancy is non-negative and only vanishes under perfect adaptation: $\hat{\mathrm{D}}(x,w) \geq 0$ and $\hat{\mathrm{D}}(x,w) = 0 \Leftrightarrow w = \hat{x}$. These facts we refer to as the *fundamental inequality* of abstract information theory ($\hat{Y}$-domain).

The linking identity is a technically important way of rewriting (8), partly as it allows us to circumvent the difficulty regarding possible indeterminacy of redundancy and partly as it opens up for an independent axiomatic treatment of concepts of information. What we shall do, rather than assuming that entropy is

---

[4] Instead of "entropy" one could have suggested a more neutral terminology such as "necessity". This may be considered less awquard when we consider other applications of the abstract theory than classical Shannon theory.

always finite, is to assume that the function $\hat{\mathrm{D}}$ can be defined on all of $X \otimes \hat{Y}$ as a non-negative extended real-valued function in such a way that the linking identity as well as the fundamental inequality hold. We express this assumption by saying that $(\hat{\Phi}, \mathrm{H}, \hat{\mathrm{D}})$ is an *information triple*, for clarity an *information triple over* $X \otimes \hat{Y}$. Thus, the main conditions imposed are, that $\hat{\Phi} = \mathrm{H} + \hat{\mathrm{D}}$ [5] and that $\hat{\mathrm{D}}(x, w) \geq 0$ with equality if and only if $w = \hat{x}$. We emphasize that the extended real-valued functions $\hat{\Phi}$ and H are not allowed to assume the value $-\infty$. Note also that the effort function of an information triple is automatically proper.

Normally, there is a natural way to extend the redundancy function as defined by (8) when $\mathrm{H}(x) < \infty$, so that an information triple emerges. In this way the problem of indeterminacy of redundancy disappears, and the slightly strengthened assumption that redundancy can be defined "appropriately" on all of $X \otimes \hat{Y}$ will, as it turns out, present no difficulty in concrete cases of interest.

Information triples occur frequently in the sequal. Sometimes one does not need a full triple $(\hat{\Phi}, \mathrm{H}, \hat{D})$ but only the redundancy function. Formally, a function $\hat{D} : X \otimes \hat{Y} \mapsto [0, \infty]$ is a *general redundancy function* if it satisfies the fundamental inequality with $w = \hat{x}$ as the condition for vanishing redundancy. From such a function you may obtain a full information triple by adding any function on X with values in $]-\infty, \infty]$, taking this function as the entropy function.

We turn to the definition of Y*-type information triples*. They are triples $(\Phi, \mathrm{H}, \mathrm{D})$ with $\Phi : X \otimes Y \mapsto]-\infty, \infty]$, $\mathrm{H} : X \mapsto]-\infty, \infty]$ and $\mathrm{D} : X \otimes Y \mapsto [0, \infty]$, such that the *linking identity*

$$\Phi(x, y) = \mathrm{H}(x) + \mathrm{D}(x, y) \tag{10}$$

holds for all $(x, y) \in X \otimes Y$ and such that the *fundamental inequality* holds for D, i.e., for all $(x, y) \in X \otimes Y$,

$$\mathrm{D}(x, y) \geq 0 \tag{11}$$

with equality if and only if there is a perfect match, $y = x$. The function Φ is the effort function of the triple, H the entropy function and D the *divergence function*. Automatically, Φ is proper.

A triple $(\Phi, \mathrm{H}, \mathrm{D})$ is a *derived information triple* if there exists $\hat{\Phi}$ and $\hat{\mathrm{D}}$ such that $(\hat{\Phi}, \mathrm{H}, \hat{\mathrm{D}})$ is a $\hat{Y}$-information triple and Φ and D are the functions derived from, respectively, $\hat{\Phi}$ and $\hat{\mathrm{D}}$. If response is injective, the two types of information triples for the Y-domain are equivalent. In general, D, hence also Φ, of a derived triple are response-only dependent and the condition for equality in the fundamental inequality is one of response-equivalence ($y \sim x$) rather than one of equality.

Just as for the $\hat{Y}$-triples, one may at times take divergence as the basic concept. A *general divergence function* D on $X \otimes Y$ – or just a *divergence function* – as a function $\mathrm{D} : X \otimes Y \mapsto [0, \infty]$ which satisfies the fundamental inequality with $y = x$ as the condition for equality in (11). And a *general derived divergence function* is one which can be derived from a general redundancy function.

Redundancy and divergence are emphasized in the following definition: Two information triples are *equivalent* if they have the same redundancy (applies to the $\hat{Y}$-domain) or the same divergence (applies to the Y-domain)[6]. Despite the chosen terminology, equivalent triples may have quite different properties

---

[5] correctly: $\hat{\Phi} = \mathrm{H} \circ \hat{\mathrm{pr}} + \hat{\mathrm{D}}$ with $\hat{\mathrm{pr}}$ the projection from $X \otimes \hat{Y}$ to X.

[6] a stronger form of equivalence requires in addition that the two entropy functions simultaneously assume the value ∞

and one may search for an equivalent triple with good properties. Note that among triples equivalent to a given triple, say $(\Phi, \mathrm{H}, \mathrm{D})$, we always have the special triple $(\mathrm{D}, 0, \mathrm{D})$. This triple we may often want to modify. One way to do that is via a process of *randomization*. This relates to results of Kuhn-Tucker type with concrete applications in information theory (channel capacity) and in location theory (Sylvesters problem). Though quite important, we shall not include that in the present write-up.

Instead of taking triples as introduced above as the basis, it is sometimes more natural to start out with a triple of the "opposite nature". This refers to situations where it is appropriate to focus on a positively oriented quantity such as *utility* or *pay-off* rather than on *effort*. This is often the case for studies of economy, meteorology and statistics where one meets the notion of "score" as previously indicated. In order to distinguish the two types of triples from eachother, we may refer to them as *effort-based*, respectively *utility-based* information triples. To be precise, let us focus on the Y-domain and define a *utility-based information triple* as a triple $(\mathrm{U}, \mathrm{M}, \mathrm{D})$ for which $(-\mathrm{U}, -\mathrm{M}, \mathrm{D})$ is an effort-based information triple. For a *derived utility-based information triple* we require that $(-\mathrm{U}, -\mathrm{M}, \mathrm{D})$ is a derived effort-based information triple. The function $\mathrm{U} = \mathrm{U}(x, y)$ defined on $X \otimes Y$ is *utility*, the function $\mathrm{M} = \mathrm{M}(x)$ defined on X *max-utility*, and $\mathrm{D} = \mathrm{D}(x, y)$ defined on $X \otimes Y$ is, as before, *divergence*. The linking identity for a utility-based triple takes the form $\mathrm{U} = \mathrm{M} - \mathrm{D}$ ($\mathrm{U} = \mathrm{M} \circ \mathrm{pr} - \mathrm{D}$) which can never result in the indeterminate form $\infty - \infty$ since, by definition, U, hence also M, can never assume the value $+\infty$.

For the $\hat{Y}$-domain, $(\hat{\mathrm{U}}, \mathrm{M}, \hat{\mathrm{D}})$ is a *utility-based information triple* if $(-\hat{\mathrm{U}}, -\hat{\mathrm{M}}, \hat{\mathrm{D}})$ is an effort-based information triple.

In view of the main examples we have in mind, we have found it most illuminating to take effort rather than utility as the basic concept to work with, and hence to develop the main results for effort-based quantities. Anyhow, even if you are primarily interested in considerations based on effort, you are easily led to consider also utility-based quantities as we shall see in Section 8.

The concept of information triples is, except for minor technical details, equivalent to the concept of proper effort functions. We find that apart from a slight technical advantage, the triples constitute a preferable base for information theoretical investigations as the three truly basic notions of information are all emphasized and their basic interrelationship – the linking identity – focused on. Historically, the notions arose for classical probabilistic information theoretical models, cf. Section 19. Effort functions go back to Kerridge [48] who coined the term *inaccuracy*, entropy to Shannon [1] and divergence to Kullback [49]. The term "redundancy" which we have used for another side of divergence, corresponds to one usage in information theory, though the term is there used in several other ways which are not expressed in our abstract setting.

Our way to information triples was through effort and one may ask why we did not go directly to the triples. For one thing, triples lead to a smooth axiomatic theory, for the beginnings of which see Topsøe [50]. However, though axiomatization can be technically attractive, we find that a focus on *interpretation* as in our more philosophical and speculative approach, is of primary importance and contributes best to an understanding of central concepts of information. Axiomatics only comes in after basic interpretations are in place.

8. Relativization, Updating

In this section we shall work entirely in the Y-domain. We start by considering an effort-based information triple $(\Phi, \mathrm{H}, \mathrm{D})$ on $X \otimes Y$. Often, it is natural to measure effort relative to some standard performance rather than by Φ itself. An especially important instance of this kind of *relativization* concerns situations where Observer originally fixed a *prior*, say $y_0 \in Y$, but now wants to *update* his belief by replacing y_0 with a *posterior* y. Perhaps, Observer – through his own actions or via an informer – has obtained the information "$x \in \mathcal{P}$" for some preparation $\mathcal{P}$. If $y_0 \notin \mathcal{P}$, Observer may want to replace y_0 by a posterior $y \in \mathcal{P}$. The associated *updating gain* is, in a first attempt of a reasonable definition, given by the quantity $\mathrm{U}_{|y_0}$ obtained by comparing performance under the posterior with performance under the prior:

$$\mathrm{U}_{|y_0}(x, y) = \Phi(x, y_0) - \Phi(x, y)\,. \tag{12}$$

A difficulty with (12) concerns the possible indeterminate form $\infty - \infty$. If we ignore this difficulty and apply the linking identity (10) to both terms in (12), entropy $\mathrm{H}(x)$ cancels out and we find the expression

$$\mathrm{U}_{|y_0}(x, y) = \mathrm{D}(x, y_0) - \mathrm{D}(x, y)\,, \tag{13}$$

which is less likely to be indeterminate. When not of the indeterminate form $\infty - \infty$, we therefore agree to use (13) as the formal definition of updating gain, more precisely of *relative updating gain with* y_0 *as prior*. For the present study, we shall only work with updating gain when D^{y_0} is finite on some preparation $\mathcal{P}$ under consideration. Assuming that this is the case, we realize that

$$(\mathrm{U}_{|y_0}, \mathrm{D}^{y_0}, \mathrm{D}) \tag{14}$$

is a utility-based information triple on $\mathcal{P} \otimes Y$. Max-utility is identified as the marginal function D^{y_0} on $\mathcal{P}$ and divergence is the original divergence function restricted to $\mathcal{P} \otimes Y$.

It is important to note that the triples which occur in this way by varying y_0 and $\mathcal{P}$ do not require the full effort function Φ for their definition. It suffices to start out with a general divergence function on $X \otimes Y$ in order for the construction to make sense. When the construction is based on a general divergence function D, we refer to (14) as the updating triple *generated by* D *and with* y_0 *as prior*. For these updating triples, we take y_0 as the only certain belief instance. The triples just introduced can be identified in a simple manner among all utility-based information triples. We formulate the result corresponding to the full preparation $\mathcal{P} = X$:

Proposition 1. *Let* $(\mathrm{U}, \mathrm{M}, \mathrm{D})$ *be a utility-based information triple on* $X \otimes Y$ *and assume that there exists a certain belief instance in* X *from which all of* X *is visible. Then there can only be one such belief instance, say* y_0*, and in this case* $\mathrm{D}^{y_0} < \infty$ *and* $(\mathrm{U}, \mathrm{M}, \mathrm{D}) = (\mathrm{U}_{|y_0}, \mathrm{D}^{y_0}, \mathrm{D})$ *on* $X \otimes Y$.

Proof. Assume that $y_0 \in X$ is a certain belief instance and that $]y_0[= X$. Then, for every $x \in X$, $(x, y_0) \in X \otimes Y$ and $0 = \mathrm{U}(x, y_0) = \mathrm{M}(x) - \mathrm{D}(x, y_0)$, hence $\mathrm{D}^{y_0} < \infty$ and $\mathrm{M} = \mathrm{D}^{y_0}$. If $y_1 \in X$ has similar properties, also $\mathrm{M} = \mathrm{D}^{y_1}$ holds. In particular, as $y_0 \in X$, $\mathrm{D}^{y_0}(y_0) = \mathrm{D}^{y_1}(y_0)$, i.e. $0 = \mathrm{D}(y_0, y_1)$, hence $y_1 = y_0$. The result follows. $\square$

Though rather trivial, the observations regarding updating gain are important as they imply that results in that setting may be derived from results based on effort. To emphasize this, we introduce, based only on a general divergence function D, the effort-based information triple *associated with* (14) as the triple

$$(\Phi_{|y_0}, -\mathrm{D}^{y_0}, \mathrm{D}) \tag{15}$$

with $\Phi_{|y_0}$, given by

$$\Phi_{|y_0}(x, y) = \mathrm{D}(x, y) - \mathrm{D}(x, y_0)\,. \tag{16}$$

This is a perfectly feasible effort-based triple on $\mathcal{P} \otimes Y$ whenever D^{y_0} is finite on $\mathcal{P}$. As with the utility-based triple (14) we take y_0 as the only certain belief instance.

In Sections 11 and 14 we shall derive results about minimum divergence (information projections) from results about maximum entropy by exploiting the simple facts here uncovered.

As we have seen, natural information triples may be derived from a general divergence function by a simple process of *relativization*. While we are at it, we note that in case $Y = X$, also *reverse divergence* $(x, y) \mapsto \mathrm{D}(y, x)$ defines a genuine divergence function on $X \otimes Y$[7]. Therefore, if $\mathrm{D}_{y_0} < \infty$ and we put $\Phi^r_{|y_0}(x, y) = \mathrm{D}(y, x) - \mathrm{D}(y_0, x)$,

$$\big(\Phi^r_{|y_0}(x, y), -\mathrm{D}(y_0, x), \mathrm{D}(y, x)\big) \tag{17}$$

defines a genuine information triple (when restricting the variables x and y appropriately). These triples are, however, not found to be that significant.

9. Feasible Preparations

We claim that *description is the key to what can be known,* a key to the "knowable". Not every possible information "$x \in \mathcal{P}$" for any odd preparation $\mathcal{P}$ can be expected to reflect a realistic situation. The questions we ask are "*what can Observer know?"* or "*what kind of information can Observer hope to obtain?"*. We thus want to investigate "limits to knowledge" and "limits to information". In order to provide an answer, we shall identify classes of preparations which represent *feasible information*. These classes will be defined with reference to an effort function $\hat{\Phi}$. For this section, $\hat{\Phi}$ need not be proper.

Given $w \in \hat{Y}$ and a *level* $h < \infty$, we define the *level set* $\mathcal{P}^w(h)$ and the *sublevel set* $\mathcal{P}^w(h^{\downarrow})$ by

$$\mathcal{P}^w(h) = \{\hat{\Phi}^w = h\}\,;\ \mathcal{P}^w(h^{\downarrow}) = \{\hat{\Phi}^w \leq h\}\,, \tag{18}$$

i.e. as the set of states which are controled by w, either at the *level* h or at the *maximum level* h. These sets are genuine preparations whenever they are non-empty. When w is the response of a state $x \in X$, the sublevel set is non-empty whenever $h \geq \mathrm{H}(x)$. As level- and sublevel sets for other functions will appear later on, cf. Section 12, we may for clarity refer to $\mathcal{P}^w(h)$ and to $\mathcal{P}^w(h^{\downarrow})$ as, respectively, $\hat{\Phi}^w$*-level sets* and $\hat{\Phi}^w$*-sublevel sets*.

The preparations in (18) are *primitive strict*, respectively *primitive slack preparations*. A *general strict*, respectively a *general slack preparation* is a finite non-empty intersection of primitive strict,

---

[7] in contrast, reverse description effort need not define a genuine effort function.

respectively primitive slack preparations. The *genus* of these preparations is the smallest number of primitive preparations (either strict or slack as the case may be) which can enter into the definition just given. Thus primitive preparations are of genus 1.

If $\mathbf{w} = (w_1, \cdots, w_n)$ are elements of $\hat{Y}$ and $\mathbf{h} = (h_1, \cdots, h_n)$ are real numbers, the sets

$$\mathcal{P}^{\mathbf{w}}(\mathbf{h}) = \bigcap_{i \leq n} \mathcal{P}^{w_i}(h_i) \text{ and } \mathcal{P}^{\mathbf{w}}(\mathbf{h}^{\downarrow}) = \bigcap_{i \leq n} \mathcal{P}^{w_i}(h_i^{\downarrow}) \tag{19}$$

define strict, respectively slack preparations whenever they are non-empty. When using these expressions, it is natural to assume that n is the genus of the preparations considered. If $\mathcal{P}^{\mathbf{w}}(\mathbf{h}) \neq \emptyset$ this set is the *corona* of $\mathcal{P}^{\mathbf{w}}(\mathbf{h}^{\downarrow})$.

The preparations introduced are those we consider to be feasible and we formally refer to them as the *feasible preparations*. They provide the answer to the question about what can be known. They are the key ingredients in situations which Observer can be faced with. In any such situation a main problem concerns *inference*, an issue we shall take up in the next section.

Of special interest are families of feasible preparations. Given $\mathbf{w} = (w_1, \cdots, w_n)$, we denote by $\mathbb{P}^{\mathbf{w}}$, respectively $\mathbb{P}^{\mathbf{w}\downarrow}$, the families which consist of all preparations $\mathcal{P}^{\mathbf{w}}(\mathbf{h})$, respectively $\mathcal{P}^{\mathbf{w}}(\mathbf{h}^{\downarrow})$, which can be obtained by varying $\mathbf{h}$.

Clearly, the feasible preparations can also be expressed by reference to the derived effort function Φ rather than $\hat{\Phi}$. We use the notation $\mathcal{P}^y(h)$ and $\mathcal{P}^y(h^{\downarrow})$ for, respectively, the Φ^y*-level set* $\{\Phi^y = h\}$ and the Φ^y*-sublevel set* $\{\Phi^y \leq h\}$. If $\hat{y} = w$, $\mathcal{P}^y(h) = \mathcal{P}^w(h)$ and $\mathcal{P}^y(h^{\downarrow}) = \mathcal{P}^w(h^{\downarrow})$[8]. For finite sequences $\mathbf{y} = (y_1, \cdots, y_n)$ of elements of Y and $\mathbf{h} = (h_1, \cdots, h_n)$ of real numbers, the sets $\mathcal{P}^{\mathbf{y}}(\mathbf{h})$ and $\mathcal{P}^{\mathbf{y}}(\mathbf{h}^{\downarrow})$ are defined in the obvious manner as are the families of preparations $\mathbb{P}^{\mathbf{y}}$, respectively $\mathbb{P}^{\mathbf{y}\downarrow}$.

From a formal point of view, it does not matter if we use $\mathcal{P}^w$-type sets or $\mathcal{P}^y$-type sets as the basis for the definition of feasible preparations. However, entering into more speculative interpretations, the $\mathcal{P}^w$-type sets with w a control seem preferable. Individual controls $w \in \hat{Y}$ or a collection of such controls point to experiments which Observer may perform. An experimental setup identifies a certain preparation, the preparation of states *consistent* with the setup, and thus determines what is known to Observer. Determining all preparations which can arise in this way, we are led to the class of feasible preparations as defined above.

As to the nature of the various controls, we imagine that they are derived from description. To control a situation, you must be able to describe it, and with a description you have the key to control. We might imagine that, corresponding to a control w, Observer can realize a certain experimental setup consisting of various parts – measuring instruments and the like. In particular, there is a special handle which is used to fix the level of effort. If the level, perhaps best thought of as a kind of *temperature*, is fixed to be h, the states available to Nature are those in the appropriate feasible preparation. Several experiments can be carried out with the same equipment by adjusting the setting of the handle. If Observer wants to constrain the states by other means, he can add equipment corresponding to another control w' and choose a level h' for the experimental setup constructed based on w'. The result is a restriction of the available states to the intersection of the two preparations involved.

---

[8] note that for an expression such as $\mathcal{P}^q(h)$, the nature of q determines if this is a $\hat{\Phi}$- or a Φ-level set.

If the preparation is $\mathcal{P}^w(h^{\downarrow})$ and the actual state is not inside this preparation, you may imagine that the result is overheating and breakdown of the experimental setup! Thus you must keep the state inside the preparation and this may well be what requires an effort as specified by $\hat{\Phi}$.

10. Inference via Games

For this section, $(\hat{\Phi}, \mathrm{H}, \hat{D})$ is an effort-based information triple on $X \otimes \hat{Y}$ and $(\Phi, \mathrm{H}, \mathrm{D})$ the derived triple on $X \otimes Y$.

Consider partial information "$x \in \mathcal{P}$". In practice, $\mathcal{P}$ will be a feasible preparation, but we need not assume so for this section.

The process of *inference* concerns the identification of "sensible" states in $\mathcal{P}$ – ideally only one such state, the *inferred state*. In many cases, this can be achieved by game theoretical methods involving a two-person zero-sum game. As it turns out, this will result in "double inference" where also either control instances or belief instances will be identified – ideally, only one such instance, the *inferred control* or the *inferred belief instance* as the case may be.

An inferred state, say x^*, brings Observer as close as possible to the truth in a way specified in some sense by the method applied. On the other hand, focusing on control, an inferred control instance w^* is more of an instruction to Observer on how to act regarding the setup of experiments and performance of subsequent observations. You may say that actions by Observer as dictated by the control w^* is what is needed for Observer in order to justify the inference x^* about truth. In short, double inference gives Observer information both about *what* can be inferred about truth and *how*.

Given $\mathcal{P}$, we shall study two closely related two-person zero-sum games, $\hat{\gamma}(\mathcal{P})$ and the *derived game* $\gamma(\mathcal{P})$. If need be, we may write $\hat{\gamma}(\mathcal{P}\,|\hat{\Phi})$ and $\gamma(\mathcal{P}\,|\Phi)$. The games have Nature and Observer as players and $\hat{\Phi}$, respectively Φ as *objective function*. Nature is understood to be a *maximizer*, Observer a *minimizer*. For both games, *strategies* involve the choice by Nature of a state in $\mathcal{P}$. Observer strategies for $\hat{\gamma}(\mathcal{P})$, respectively $\gamma(\mathcal{P})$ are controls from which every state in $\mathcal{P}$ can be controled, respectively belief instances from which every state in $\mathcal{P}$ is visible. Pairs of *permissable strategies* for the two games are either pairs $(x, w) \in X \otimes \hat{Y}$ with $x \in \mathcal{P}$, $w \succ \mathcal{P}$ or pairs $(x, y) \in X \otimes Y$ with $x \in \mathcal{P}$, $y \succ \mathcal{P}$. In consistency with the discussion in Section 2, an Observer strategy may be thought of as a strategy which is not "completely stupid" whatever the strategy of Nature as long as that strategy respects the requirement $x \in \mathcal{P}$. The choice of strategy for Observer may be a real choice, whereas, for Nature, it may be more appropriate to have a fictive choice in mind, reflecting Observers thoughts about what the truth could be.

Following standard philosophy of game theory, Observer should always be prepared for a choice by Nature which is least favourable for him. One can argue that in our setting anything else would mean that Observer would not have used all available information. The line of thought goes well with Jaynes thinking as collected in [10], though there you find no reference to game theory.

In order for our exposition to be self-contained and also because our games are slightly at variance with what is normally considered [9], we shall here give full details regarding definitions and proofs. As a general reference to game theory we point to [31].

Let us introduce basic notions for the $\hat{\gamma}$-game and then only comment briefly about the corresponding notions for the γ-game.

The two *values* of $\hat{\gamma}(\mathcal{P})$ are, for Nature,

$$\sup_{x\in\mathcal{P}} \inf_{w\succ x} \hat{\Phi}(x,w) \tag{20}$$

and, for Observer,

$$\inf_{w\succ\mathcal{P}} \sup_{x\in\mathcal{P}} \hat{\Phi}(x,w)\,. \tag{21}$$

Note the slight deviation from usual practice in that w in the infimum in (20) varies over $\hat{}[x]$ and not just over $\hat{}[\mathcal{P}]$ or some other set independent of x. Philosophically, one may argue that Nature does not know of the restriction to $\mathcal{P}$ – this is something Observer has arranged – and hence cannot know of any restriction besides the natural one $w \succ x$. As the infimum in (20) is nothing but the entropy $\mathrm{H}(x)$, the value for Nature, denoted $\mathrm{H}_{\max}(\mathcal{P})$, is the *maximum entropy value*

$$\mathrm{H}_{\max}(\mathcal{P}) = \sup_{x\in\mathcal{P}} \mathrm{H}(x)\,, \tag{22}$$

also referred to as the MaxEnt-*value*.

Problems on the determination of this value and associated strategies reaching the value (if any) are *maximum entropy problems*, for short MaxEnt-problems. The archetypical concrete problems of this nature are discussed in Section 19.

As to the value for Observer, we identify the supremum in (21) with the *risk* associated with the strategy w and denote it by $\hat{\mathrm{Ri}}(w|\,\mathcal{P})$:

$$\hat{\mathrm{Ri}}(w|\,\mathcal{P}) = \sup_{x\in\mathcal{P}} \hat{\Phi}(x,w)\,. \tag{23}$$

The value for Observer then is the *minimal risk* of the game, also referred to as the MinRisk-*value*:

$$\hat{\mathrm{Ri}}_{\min}(\mathcal{P}) = \inf_{w\succ\mathcal{P}} \hat{\mathrm{Ri}}(w|\,\mathcal{P})\,. \tag{24}$$

An *optimal strategy for Nature* is a strategy $x^*\in\mathcal{P}$ with $\mathrm{H}(x^*) = \mathrm{H}_{\max}(\mathcal{P})$ and an *optimal strategy for Observer* is a strategy $w^*\succ\mathcal{P}$ with $\hat{\mathrm{Ri}}(w^*|\,\mathcal{P}) = \hat{\mathrm{Ri}}_{\min}(\mathcal{P})$.

The reader will easily verify the general validity of the *minimax inequality*:

$$\mathrm{H}_{\max}(\mathcal{P}) \le \hat{\mathrm{Ri}}_{\min}(\mathcal{P})\,. \tag{25}$$

If this inequality holds with equality and defines a finite quantity, the game is said to be in *game theoretical equilibrium*, or just in *equilibrium*, and the common value of $\mathrm{H}_{\max}(\mathcal{P})$ and $\hat{\mathrm{Ri}}_{\min}(\mathcal{P})$ is the *value* of the game.

---

[9] due to assumptions of properness of the effort functions considered and due to the refinement introduced regarding the relations between control, visibility and response

We need one more notion of equilibrium which we associate with the name of Nash[10]. A pair of permissable strategies (x^*, w^*) is a *Nash equilibrium pair for* $\hat{\gamma}(\mathcal{P})$ if, with these strategies, none of the players have an incentive to change strategy – provided the opponent does not do so either. This means, for Nature, that

$$\forall x \in \mathcal{P} : \hat{\Phi}(x, w^*) \leq \hat{\Phi}(x^*, w^*)\,, \tag{26}$$

and, for Observer, that

$$\forall w \succ \mathcal{P} : \hat{\Phi}(x^*, w) \geq \hat{\Phi}(x^*, w^*)\,. \tag{27}$$

The inequalities (26) and (27) constitute (a special case of) the celebrated *saddle-value inequalities* of game theory. Note that, in our case, one of these inequalities, (27), is automatic if (x^*, w^*) is an adapted pair. Clearly, since the pair should also be a pair of permissable strategies, this requires that $x^* \in \mathrm{ctr}(\mathcal{P})$.

We find that the game $\hat{\gamma}(\mathcal{P})$ is the most natural to consider in view of the associated interpretations. However, we shall also formulate results for $\gamma(\mathcal{P})$. Thereby one avoids direct consideration of controls. The values of $\gamma(\mathcal{P})$ are $\sup_{x\in\mathcal{P}} \inf_{y\succ x} \Phi(x, y)$ and $\inf_{y\succ\mathcal{P}} \sup_{x\in\mathcal{P}} \Phi(x, y)$ and notions of strategies and optimal strategies are defined in an obvious manner. We use the notation Ri for *risk* in this game, defined, for $y \succ \mathcal{P}$, as $\mathrm{Ri}(y|\mathcal{P}) = \sup_{x\in\mathcal{P}} \Phi(x, y)$. Clearly, $\mathrm{Ri}(y|\mathcal{P}) = \hat{\mathrm{Ri}}(\hat{y}|\mathcal{P})$. Therefore, if $y_1 \succ \mathcal{P}$ and $y_2 \succ \mathcal{P}$ are response-equivalent, the associated risks are the same. Further, the values of $\gamma(\mathcal{P})$ agree with those of $\hat{\gamma}(\mathcal{P})$, in particular $\mathrm{Ri}_{\min}(\mathcal{P}) = \hat{\mathrm{Ri}}_{\min}(\mathcal{P})$. We leave it to the reader to transform the notions of equilibrium and the form of the saddle-value inequalities to the game $\gamma(\mathcal{P})$.

Important results of game theory are non-constructive by nature and aim at securing equilibrium and existence of optimal strategies for wide classes of games. For our setting, this will be taken up in Section 14. However, for the present section we shall focus on the possibility to identify optimal strategies. This leads to problems which are easy to handle technically and yet, it may be argued that from an applied point of view such results are the more impoertant ones.

In our first result, we point out an important property of the optimal strategies for the key case we shall deal with, that of a game in equilibrium for which both players have optimal strategies.

Theorem 1. *[Basics] The game $\hat{\gamma}(\mathcal{P})$ is in equilibrium and both players have optimal strategies, if and only if these properties also hold for the game $\gamma(\mathcal{P})$. Assume now that this is the case.*

Then there is only one optimal strategy, say w^*, for Observer in $\hat{\gamma}(\mathcal{P})$ and all optimal strategies for Nature in $\hat{\gamma}(\mathcal{P})$ are response-equivalent with w^* as response and lie in the centre of $\mathcal{P}$.

The optimal strategies for Nature in $\gamma(\mathcal{P})$ are the same as the optimal strategies for Nature in $\hat{\gamma}(\mathcal{P})$ and a belief instance $y \succ \mathcal{P}$ is an optimal strategy for Observer in $\gamma(\mathcal{P})$ if and only if it has w^* as response.

If response is injective, all optimal strategies considered are unique and the two optimal strategies for $\gamma(\mathcal{P})$ coincide.

Proof. The first statement follows directly from the definitions involved. In order to prove the remaining parts, we concentrate on the game $\hat{\gamma}(\mathcal{P})$ and assume that this game is in equilibrium and that optimal

[10] It should, however, be said that for the relatively simple case here considered (two players, zero sum), the ideas we need originated with von Neumann, see [51] and [52] and, for a historical study, Kjeldsen [53].

strategies for both players exist. Let x^*, w^* be any set of such optimal strategies. By the defining relations (5) and (7), by the assumptions of equilibrium and optimality of x^* and of w^*, and by the definition (23) of risk, we find that

$$\hat{\Phi}(x^*, w^*) \geq \hat{\Phi}(x^*, \hat{x^*}) = \mathrm{H}(x^*) = \mathrm{H}_{\max}(\mathcal{P}) = \hat{\mathrm{Ri}}_{\min}(\mathcal{P}) = \hat{\mathrm{Ri}}(w^* | \, \mathcal{P}) \geq \Phi(x^*, w^*)\,, \tag{28}$$

hence $\hat{\Phi}(x^*, w^*) = \mathrm{H}(x^*)$ and, since $\hat{\Phi}$ is proper and since $\mathrm{H}(x^*) = \mathrm{H}_{\max}(\mathcal{P}) < \infty$, it follows that $w^* = \hat{x^*}$. Then $\hat{x^*} = w^* \succ \mathcal{P}$ and, as $\hat{x^*} \succ \mathcal{P}$ is equivalent with $x^* \succ \mathcal{P}$, we conclude that $x^* \in \mathrm{ctr}(\mathcal{P})$.

Since x^* above was an arbitrary optimal strategy for Nature and w^* an arbitrary optimal strategy for Observer, we conclude from $w^* = \hat{x^*}$ that the optimal Observer strategy is unique and that all optimal strategies for Nature are response-equivalent, lie in $\mathrm{ctr}(\mathcal{P})$ and has the optimal Observer strategy as response.

We leave it to the reader to establish the results for $\gamma(\mathcal{P})$, say by noting that $y \succ \mathcal{P}$ is equivalent with $\hat{y} \succ \mathcal{P}$ and that $\mathrm{Ri}(y|\mathcal{P}) = \hat{\mathrm{Ri}}(\hat{y}|\mathcal{P})$ and by using the first facts established.

Assume now that response is injective. Then uniqueness of optimal strategies, say of (x^*, w^*) for $\hat{\gamma}(\mathcal{P})$ and of (x^*, y^*) for $\gamma(\mathcal{P})$ follows readily and the identity of x^* and y^* follows as these belief instances are response-equivalent. □

Warning: It is not true that all Nature strategies with the optimal Observer strategy as response have to be optimal. Simple examples, say with "collapse of response", i.e. with $\hat{Y}$ a singleton, will demonstrate that.

When the games are in equilibrium and optimal strategies exist, we refer to any optimal strategy for Nature as a *a bi-optimal strategy*. The bi-optimality refers to the fact that also Observer-optimality is secured. Indeed, $\hat{x^*}$ is optimal for Observer in $\hat{\gamma}(\mathcal{P})$ and any $y^* \succ \mathcal{P}$ which is equivalent to x^* (including x^* itself) is optimal for Observer in $\gamma(\mathcal{P})$. If response is injective, there is only one such state, *the* bi-optimal state.

Whereas it may be difficult to find optimal strategies, it is often easy to check if given candidates are in fact optimal.

Theorem 2. *[Identification] Let (x^*, w^*) be permissable strategies for $\hat{\gamma}(\mathcal{P})$ with $x^* \in \mathrm{ctr}(\mathcal{P})$ and $\mathrm{H}(x^*) < \infty$.*

Then a necessary and sufficient condition that $\hat{\gamma}(\mathcal{P})$ is in equilibrium with (x^*, w^*) as optimal strategies is that (x^*, w^*) is a Nash equilibrium pair. When this is true, w^* is adapted to x^*.

Proof. The sufficiency follows since (26) is equivalent with the condition $\hat{\mathrm{Ri}}(w^* | \, \mathcal{P}) \leq \hat{\Phi}(x^*, w^*)$ and, under the assumption $x^* \succ \mathcal{P}$, (27) is equivalent with the condition $\Phi(x^*, w^*) \leq \mathrm{H}(x^*)$. Thus, when (x^*, w^*) is a Nash equilibrium pair, $\hat{\mathrm{Ri}}(w^* | \, \mathcal{P}) \leq \mathrm{H}(x^*)$, hence, by the minimax inequality, x^* and w^* are optimal strategies and $\mathrm{H}_{\max}(\mathcal{P}) = \hat{\mathrm{Ri}}_{\min}(\mathcal{P})$. As we assumed that $\mathrm{H}(x^*) < \infty$, $\hat{\gamma}(\mathcal{P})$ is in equilibrium.

The necessity and the last part of the theorem follow from Proposition 1 and the above noticed equivalent forms of the saddle-value inequalities. □

Elaborating slightly, we obtain the following corollary:

Corollary 1. *Let $\mathcal{P}$ be a preparation and consider strategies x^*, w^* with $x^* \in \mathrm{ctr}(\mathcal{P})$, w^* adapted to x^* and $\mathrm{H}(x^*) < \infty$. Then $\hat{\gamma}(\mathcal{P})$ is in equilibrium with x^* as bi-optimal strategy if and only if,*

$$\forall x \in \mathcal{P} : \hat{\Phi}(x, w^*) \leq \mathrm{H}(x^*)\,. \tag{29}$$

Proof. Under the conditions stated, (27) is automatic and (29) is a reformulation of (26). Thus (29) implies that (x^*, w^*) is a Nash equilibrium pair and the result then follows from Theorem 2. □

A main consequence of the existence of a bi-optimal strategy is the validity of the *Pythagorean inequalities*. The *direct Pythagorean inequality*, or just the *Pythagorean inequality*, is the inequality $\mathrm{H}(x) + \hat{\mathrm{D}}(x, w^*) \leq \mathrm{H}(x^*)$, typically considered for $x \in \mathcal{P}$. This is nothing but a trivial rewriting of (29). When it holds, $\mathrm{H}(x^*) = \mathrm{H}_{\max}(\mathcal{P})$ and the inequality for an individual state $x \in \mathcal{P}$ is, therefore, a sharper form of the trivial inequality $\mathrm{H}(x) \leq \mathrm{H}_{\max}(\mathcal{P})$. The *dual Pythagorean inequality* is the inequality $\hat{\mathrm{Ri}}(w^*|\mathcal{P}) + \hat{\mathrm{D}}(x^*, w) \leq \hat{\mathrm{Ri}}(w|\,\mathcal{P})$, typically considered for $w \succ \mathcal{P}$. When it holds, $\hat{\mathrm{Ri}}(w^*|\,\mathcal{P}) = \hat{\mathrm{Ri}}_{\min}(\mathcal{P})$, and the inequality for an individual strategy $w \succ \mathcal{P}$ is, therefore, a sharper form of the trivial inequality $\hat{\mathrm{Ri}}_{\min}(\mathcal{P}) \leq \hat{\mathrm{Ri}}(w|\,\mathcal{P})$.

Theorem 3. *[Pythagorean inequalities] If $\hat{\gamma}(\mathcal{P})$ is in equilibrium with x^* as a bi-optimal strategy then, with $w^* = \hat{x^*}$, the direct as well as the dual Pythagorean inequality holds:*

$$\forall x \in \mathcal{P} : \mathrm{H}(x) + \hat{\mathrm{D}}(x, w^*) \leq \mathrm{H}(x^*)\,, \tag{30}$$

$$\forall w \succ \mathcal{P} : \hat{\mathrm{Ri}}(w^*|\,\mathcal{P}) + \hat{\mathrm{D}}(x^*, w) \leq \hat{\mathrm{Ri}}(w|\,\mathcal{P})\,. \tag{31}$$

Proof. As to (30), this follows from Corollary 1. Also (31) must hold since, for $w \succ \mathcal{P}$,

$$\hat{\mathrm{Ri}}(w^*|\,\mathcal{P}) + \hat{\mathrm{D}}(x^*, w) = \mathrm{H}(x^*) + \hat{\mathrm{D}}(x^*, w) = \hat{\Phi}(x^*, w) \leq \hat{\mathrm{Ri}}(w|\,\mathcal{P})\,.$$

□

The simple translation of results to games of type $\gamma(\mathcal{P})$ rather than type $\hat{\gamma}(\mathcal{P})$ is left to the reader.

For the last results of this section we find it more natural to work in the Y-domain.

First we point to an extra property of bi-optimal strategies which follows from (30). In order to formulate this in a convenient way we need some definitions. A sequence (x_n) of states *converges in divergence* to the state x, written $x_n \xrightarrow{\mathrm{D}} x$, if $\lim_{n\to\infty} \mathrm{D}(x_n, x) = 0$. This requires that $(x_n, x) \in X \otimes Y$ for all n. If $x_n \in \mathcal{P}$ for all n, we say that (x_n) is *asymptotically optimal*, more precisely *asymptotically optimal for Nature in the game $\gamma(\mathcal{P})$*, if $\mathrm{H}(x_n) \to \mathrm{H}_{\max}(\mathcal{P})$ as $n \to \infty$. Finally, a state x (not necessarily in $\mathcal{P}$) is a *maximum entropy attractor for $\mathcal{P}$*, or an *$\mathrm{H}_{\max}$-attractor*, if $x_n \xrightarrow{\mathrm{D}} x$ for every asymptotically optimal sequence. We can now state a trivial corollary to Theorem 3 (transformed to the Y-domain):

Corollary 2. *Any bi-optimal strategy x^* of a game $\gamma(\mathcal{P})$ in equilibrium, is a maximum entropy attractor for $\mathcal{P}$.*

One can establish existence of the attractor in many cases, even if the bi-optimal strategy does not exist. We shall return to this issue in Section 14.

Dual versions of the notions and results indicated above could be introduced, depending on (31) rather than on (30). However, it seems that the notions related to the direct pythagorean inequality are the more useful ones.

The pythagorean flavour of (30) is more pronounced when one turns to models of updating, cf. Sections 11 and 18.

For the corollary to follow we need an abstract version of *Jeffrey's divergence* given, for two states x_1 and x_2 by

$$\mathrm{J}(x_1, x_2) = \mathrm{D}(x_1, x_2) + \mathrm{D}(x_2, x_1)\,. \tag{32}$$

Corollary 3. *[transitivity inequality] If $\gamma(\mathcal{P})$ is in equilibrium with x^* as a bi-optimal strategy, then, for every state $x \in \mathcal{P}$ and every belief instance $y \succ \mathcal{P}$, the inequality*

$$\mathrm{H}(x) + \mathrm{D}(x, x^*) + \mathrm{D}(x^*, y) \leq \mathrm{Ri}(y|\,\mathcal{P}) \tag{33}$$

holds. In particular, for every $x \in \mathrm{ctr}(\mathcal{P})$,

$$\mathrm{H}(x) + \mathrm{J}(x, x^*) \leq \mathrm{Ri}(x|\,\mathcal{P}). \tag{34}$$

Proof. First note that also $\hat{\gamma}(\mathcal{P})$ is in equilibrium with x^* as bi-optimal strategy. Then, putting $w^* = \hat{x^*}$, (30) and(31) hold. Therefore, and as $\mathrm{H}(x^*) = \hat{\mathrm{Ri}}(w^*|\mathcal{P})$, for $x \in \mathcal{P}$ and $w \succ \mathcal{P}$,

$$\mathrm{H}(x) + \hat{\mathrm{D}}(x, w^*) + \hat{\mathrm{D}}(x^*, w) \leq \hat{\mathrm{Ri}}(w|\mathcal{P})\,. \tag{35}$$

To a given belief instance y with $y \succ \mathcal{P}$ we then apply (35) with $w = \hat{y}$. As $\hat{\mathrm{D}}(x, w^*) = \mathrm{D}(x, x^*)$, $\hat{\mathrm{D}}(x^*, w) = \mathrm{D}(x^*, y)$ and $\hat{\mathrm{Ri}}(w|\mathcal{P}) = \mathrm{Ri}(y|\mathcal{P})$, (33) follows. □

We refer to (33) as the *transitivity inequality*. It is a sharper version of the trivial inequality $\mathrm{H}(x) \leq \mathrm{Ri}(y|\,\mathcal{P})$. It combines both Pythagorean inequalities and these are easily derived from it. If $\mathrm{Ri}(y|\,\mathcal{P}) < \infty$, the inequality holds with equality if and only if both Pythagorean inequalities (30) and (31) hold with equality .

As to the last part of Corollary 3, we note that if you put $r = \mathrm{Ri}(x|\,\mathcal{P}) - \mathrm{H}(x)$, then the bi-optimal strategy has Jeffrey divergence at most r from x.

11. Games based on Utility, Updating

In the previous section we investigated games related to an effort-based information triple. Similar notions and results apply when we start-out with a utility-based triple. Let us work in the Y-domain and start out with a utility based information triple $(\mathrm{U}, \mathrm{M}, \mathrm{D})$ over $X \otimes Y$. Then, given a preparation $\mathcal{P}$, the associated game $\gamma(\mathcal{P}\,|\,\mathrm{U})$ has Observer as maximizer and Nature as minimizer [11] and the two values of the game are, for Nature, the *minimax utility* $\mathrm{M}_{\min}(\mathcal{P})$:

$$\mathrm{M}_{\min}(\mathcal{P}) = \inf_{x\in\mathcal{P}} \sup_{y\succ x} \mathrm{U}(x, y) = \inf_{x\in\mathcal{P}} \mathrm{M}(x)\,, \tag{36}$$

and, for Observer, the corresponding *maximin value*

$$\sup_{y\succ\mathcal{P}} \inf_{x\in\mathcal{P}} \mathrm{U}(x, y)\,. \tag{37}$$

---

[11] thus, if T in $\gamma(\mathcal{P}\,|T)$ is declared as a utility function, this convention applies, whereas if T is a declared effort function, the players swop roles with Observer as minimizer and Nature as maximizer as in the previous section.

For $y \succ \mathcal{P}$, the infimum occurring here is the *guaranteed utility* associated with the strategy y. We denote it $\mathrm{Gtu}(y|\,\mathcal{P})$. The maximin value (37) is also referred to as the *maximal guaranteed utility*. We denote it $\mathrm{Gtu}_{\max}(\mathcal{P})$:

$$\mathrm{Gtu}_{\max}(\mathcal{P}) = \sup_{y \succ \mathcal{P}} \mathrm{Gtu}(y|\,\mathcal{P}) = \sup_{y \succ \mathcal{P}} \inf_{x \in \mathcal{P}} \mathrm{U}(x,y)\,. \tag{38}$$

Notions and results, e.g. related to equilibrium, to optimal or bi-optimal strategies etc. are developed in an obvious manner, either by following Section 10 in parallel or by applying the results of Section 10 to the effort-based triple $(-\mathrm{U}, -\mathrm{M}, \mathrm{D})$. We leave this for the interested reader to do. However, for the important case of updating, cf. Section 8, we shall be more explicit.

We take as starting point a general divergence function D on $X \otimes Y$, a preparation $\mathcal{P}$ and a prior y_0 with $\mathrm{D}^{y_0} < \infty$ on $\mathcal{P}$. The game associated with the utility-based information triple $(\mathrm{U}_{|y_0}, \mathrm{D}^{y_0}, \mathrm{D})$ we denote $\gamma(\mathcal{P}; y_0)$. Following Section 10, the value for Nature in $\gamma(\mathcal{P}; y_0)$ is $\inf_{x \in \mathcal{P}} \mathrm{D}^{y_0}(x)$, also denoted $\mathrm{D}_{\min}(\mathcal{P}; y_0)$ and referred to as the the *minimum divergence value* or the MinDiv-*value*:

$$\mathrm{D}_{\min}(\mathcal{P}; y_0) = \inf_{x \in \mathcal{P}} \mathrm{D}(x, y_0)\,. \tag{39}$$

An optimal strategy for Nature is here called a D-*projection of* y_0 *on* $\mathcal{P}$. If Nature has a unique optimal strategy, it is *the* D-projection of y_0 on $\mathcal{P}$. Consider an Observer strategy $y \succ \mathcal{P}$, i.e. a possible posterior. We use the same notation as in the general case, "Gtu" , to indicate Observers evaluation of the performance of the posterior. Incidentally, the letters can here be taken to stand for "guaranteed updating (gain)" . Thus

$$\mathrm{Gtu}(y|\,\mathcal{P}; y_0) = \inf_{x \in \mathcal{P}} \mathrm{U}_{|y_0}(x,y) = \inf_{x \in \mathcal{P}} \Big(\mathrm{D}(x, y_0) - \mathrm{D}(x,y) \Big) \tag{40}$$

is the *guaranteed updating gain* associated with the choice y of posterior, and

$$\mathrm{Gtu}_{\max}(\mathcal{P}; y_0) = \sup_{y \succ \mathcal{P}} \mathrm{Gtu}(y|\,\mathcal{P}; y_0) \tag{41}$$

is Observers value of the game, the *maximum guaranteed updating gain*, or the MaxGtu-*value* of $\gamma(\mathcal{P}; y_0)$.

The basic results for the updating game may be summarized as follows:

Theorem 4. *Let* D *be a general divergence function on* $X \otimes Y$*,* $\mathcal{P}$ *a preparation and* y_0 *a belief instance with* $\mathrm{D}^{y_0} < \infty$ *on* $\mathcal{P}$*. Consider the updating game* $\gamma = \gamma(\mathcal{P}; y_0)$*.*

If $x^* \in \mathrm{ctr}(\mathcal{P})$*, then* γ *is in equilibrium with* x^* *as bi-optimal strategy if and only if the Pythagorean inequality*

$$\mathrm{D}(x, y_0) \geq \mathrm{D}(x, x^*) + \mathrm{D}(x^*, y_0) \tag{42}$$

holds for every $x \in \mathcal{P}$*. And if this condition is satisfied,* x^* *is the* D*-projection of* y_0 *on* $\mathcal{P}$*. Furthermore, the dual pythagorean inequality*

$$\mathrm{Gtu}(y|\,\mathcal{P}; y_0) + \mathrm{D}(x^*, y) \leq \mathrm{Gtu}(x^*|\,\mathcal{P}; y_0) \tag{43}$$

holds for every $y \succ \mathcal{P}$*.*

The proof can be carried out by applying Corollary 1 and Theorem 3 to the effort function $\Phi_{|y_0}$ associated with the updating game considered, cf. (16). Details are left to the reader.

The concept of an attractor, cf. Corollary 2, also makes sense for updating games. Then the relevant notion is that of a *relative attractor given* y_0, also called the $\mathrm{D}^{y_0}_{\min}$*-attractor*, which is defined as a state x^* such that, for every sequence (x_n) in $\mathcal{P}$ with $\mathrm{D}(x_n, y_0) \to \mathrm{D}_{\min}(\mathcal{P}; y_0)$ it holds that $x_n \overset{\mathrm{D}}{\to} x^*$. In the situation covered by Theorem 4 – assuming also that limit states for convergence in divergence are unique – the relative attractor exists and coincides with the bi-optimal strategy.

The Pythagorean inequality originated with Chentsov [54] and Csiszár [55] where updating in a probabilistic setting was considered. Further versions, still probabilistic in nature can be found in Csiszár and Matus [56]. In [57] these authors present a general abstract study. We also mention Glonti et al [58] where you find a *reversed Pythagorean inequality* which results by applying the (direct) Pythagorean inequality to a system as given by the triple (17).

12. Formulating results with a geometric flavour

The results of Section 10 are formulated analytically, based on properties of the effort function. In this section we make a translation to results which have a certain geometric flavour. We shall work entirely in the Y-domain and assume throughout the section that $(\Phi, \mathrm{H}, \mathrm{D})$ is an effort-based information triple.

In the previous sections, we had a fixed preparation in mind. Here, we shall also discuss to which extent you can change a preparation without changing an optimal strategy.

Sublevel sets of the form $\{\Phi^y \leq a\}$ play a a key role. These sets appeared before as primitive feasible preparations. Here they have a different role and we prefer to use the bracket notation as above.

Proposition 2. *Let x^* be a state with finite entropy $h = \mathrm{H}(x^*)$. Then, given a preparation $\mathcal{P}$, the necessary and sufficient condition that the game $\gamma(\mathcal{P})$ is in equilibrium with x^* as bi-optimal strategy is that $\mathcal{P}$ is squeezed in between $\{x^*\}$ and $\{\Phi^{x^*} \leq h\}$, i.e. that $x^* \in \mathcal{P} \subseteq \{\Phi^{x^*} \leq h\}$. In particular, $\{\Phi^{x^*} \leq h\}$ is the largest such preparation.*

This follows directly from Theorem 2 and Corollary 1.

For a fixed preparation $\mathcal{P}$, we can express the two values of $\gamma(\mathcal{P})$, $\mathrm{H}_{\max}(\mathcal{P})$ and $\mathrm{Ri}_{\min}(\mathcal{P})$, in a geometrically flavoured way. This will be done whether or not the game is in equilibrium and the result can thus be used to check if the game is in fact in equilibrium. It is convenient to introduce some preparatory terminology.

Firstly, a subset of X is an *entropy sublevel set* if it is a (non-empty) set of the form $\{\mathrm{H} \leq a\}$. The *size* of such a set is the smallest number a which can occur in this representation (clearly equal to the MaxEnt-value associated with the preparation $\{\mathrm{H} \leq a\}$). Given a preparation $\mathcal{P}$, the associated *enveloping entropy sublevel set* is the smallest entropy sublevel set containing $\mathcal{P}$.

Secondly, and quite analogously in view of (22) and (23), we introduce the *size* of the Φ^y-sublevel set $\{\Phi^y \leq a\}$ as the smallest number a which can occur in this representation. And we define the *enveloping Φ^y-sublevel set* associated with $\mathcal{P}$ to be the smallest Φ^y-sublevel set containing $\mathcal{P}$.

Proposition 3. *Consider the game $\gamma(\mathcal{P})$ associated with a preparation $\mathcal{P}$. Then:*

(i) The MaxEnt-value $\mathrm{H}_{\max}(\mathcal{P})$ is the size of the enveloping entropy sublevel set associated with $\mathcal{P}$;

(ii) For fixed $y \succ \mathcal{P}$, $\mathrm{Ri}(y|\,\mathcal{P})$ is the size of the enveloping Φ^y-sublevel set associated with $\mathcal{P}$.

(iii) The MinRisk-value $\mathrm{Ri}_{\min}(\mathcal{P})$ is the infimum over $y \succ \mathcal{P}$ of the sizes of the enveloping Φ^y-sublevel sets associated with $\mathcal{P}$.

In view of (22), (23) and (24), this is obvious. Some comments on the result are in order. In (i) it is understood that the size is infinite if no entropy sublevel set exists which contains $\mathcal{P}$. A similar convention applies to (ii). Also note that the result gives rise to a simple geometrically flavoured proof of the minimax inequality (25) by noting that for each $y \succ \mathcal{P}$ and each h, $\{\Phi^y \leq h\} \subseteq \{\mathrm{H} \leq h\}$.

There are two families of sets involved in Proposition 3, the entropy sublevel sets and the Φ^y-sublevel sets. As the proposition shows, both families give valuable information about the games we are interested in. From the second family alone, one can in fact obtain rather complete information. Indeed, if $\{\Phi^y \leq a\}$ contains a given preparation for appropriately chosen y and a, the associated game is well behaved:

Proposition 4. *Given a preparation $\mathcal{P}$, a necessary and sufficient condition that $\gamma(\mathcal{P})$ is in equilibrium and has a bi-optimal strategy is that $\{\Phi^y \leq a\} \supseteq \mathcal{P}$ for some (y, a) with $y \in \mathcal{P}$ and $a = \mathrm{H}(y)$. When the condition is fulfilled, a is the value of the game and y the bi-optimal strategy.*

The simple proof is left to the reader. It is the sufficiency which is most useful in practical applications.

The results above translate without difficulty to results about games associated with a utility-based information triple $(\mathrm{U}, \mathrm{M}, \mathrm{D})$. For this, *superlevel sets* of the form $\{\mathrm{U}^y \geq k\}$ as well as *strict sublevel sets* of the form either $\{\mathrm{M} < a\}$ or $\{\mathrm{U}^y < a\}$ play an important role. The notion of *size* of these latter sets, those defined by strict inequality, is defined as the largest value of a which can occur in the representations given (note: *largest* rather than *smallest* as was the case before).

We shall consider the largest sets of the form $\{\mathrm{M} < a\}$, respectively $\{\mathrm{U}^y < a\}$, which are contained in the complement $\complement\,\mathcal{P}$ or, as we shall consistently prefer to say below, which are *external to $\mathcal{P}$*.

Either directly – or as corollaries to Propositions 2, 3 and 4 applied to the effort-based triple $(-\mathrm{U}, -\mathrm{M}, \mathrm{D})$ – one derives the following results:

Proposition 5. *Let $(\mathrm{U}, \mathrm{M}, \mathrm{D})$ be a utility-based information triple and consider a state x^* with $k = \mathrm{M}(x^*) > -\infty$. Then, for any preparation $\mathcal{P}$, the game $\gamma(\mathcal{P} \,|\, \mathrm{U})$ is in equilibrium with x^* as bi-optimal strategy if and only if $x^* \in \mathcal{P} \subseteq \{\mathrm{U}^{x^*} \geq k\}$. In particular, the largest such preparation is the superlevel set $\{\mathrm{U}^{x^*} \geq k\}$.*

Proposition 6. *Let $(\mathrm{U}, \mathrm{M}, \mathrm{D})$ be a utility-based information triple and consider a preparation $\mathcal{P}$ and the associated game $\gamma(\mathcal{P} \,|\, \mathrm{U})$. Then:*

(i) The value $\mathrm{M}_{\min}(\mathcal{P})$ is the size of the largest strict sublevel set $\{\mathrm{M} < a\}$ which is external to $\mathcal{P}$.

(ii) For fixed $y \succ \mathcal{P}$, $\mathrm{Gtu}(y|\,\mathcal{P})$ is the size of the largest strict sublevel set $\{\mathrm{U}^y < a\}$ which is external to $\mathcal{P}$.

(iii) The value $\mathrm{Gtu}_{\max}(\mathcal{P})$, as the supremum of $\mathrm{Gtu}(y|\,\mathcal{P})$, is the supremum of all sizes of sets of the form $\{\mathrm{U}^y < a\}$ with $y \succ \mathcal{P}$ which are external to $\mathcal{P}$.

Proposition 7. *Let $(\mathrm{U}, \mathrm{M}, \mathrm{D})$ be a utility-based information triple and consider a preparation $\mathcal{P}$. Then a necessary and sufficient condition that $\gamma(\mathcal{P} \,|\, \mathrm{U})$ is in equilibrium and has a bi-optimal strategy is that $\{\mathrm{U}^y < a\}$ is external to $\mathcal{P}$ for some (y, a) with $y \in \mathcal{P}$ and $a = \mathrm{M}(y)$. When the condition is fulfilled, a is the value of the game and y the bi-optimal strategy.*

We also note that the minimax inequality $\mathrm{Gtu}_{\max}(\mathcal{P}) \leq \mathrm{M}_{\min}(\mathcal{P})$ follows from Proposition 6 by applying the fact that, generally, $\{\mathrm{M} < a\} \subseteq \{\mathrm{U}^y < a\}$.

Let us look specifically at models of updating, cf. Section 11.

Given is a general divergence function D on $X \otimes Y$ and we consider preparations $\mathcal{P}$ and priors y_0 for which $\mathrm{D}^{y_0} < \infty$ on $\mathcal{P}$. The sets we shall focus on related to the games $\gamma(\mathcal{P}; y_0)$ are of two types, which we associate with, respectively *"balls"* and *"half-spaces"*. Firstly, for $r > 0$, consider the *open divergence ball with radius* r *and centre* y_0, defined as the D^{y_0}-sublevel set

$$\mathrm{B}(r|y_0) = \{\mathrm{D}^{y_0} < r\}\,. \tag{44}$$

In case $r = \mathrm{D}(x^*, y_0)$ for some state x^*, we write this set as $\mathrm{B}(x^*|y_0)$:

$$\mathrm{B}(x^*|y_0) = \mathrm{B}(\mathrm{D}(x^*, y_0)|y_0) = \{x|\, \mathrm{D}(x, y_0) < \mathrm{D}(x^*, y_0)\}\,. \tag{45}$$

And, secondly, we consider sets – all referred to as *half-spaces* – of one of the following forms

$$\sigma^+(y, a|y_0) = \{x|\, \mathrm{U}_{|y_0} < a\} = \{x|\, \mathrm{D}(x, y_0) - \mathrm{D}(x, y) < a\} \tag{46}$$

$$\sigma^-(y, a|y_0) = \{x|\, \mathrm{U}_{|y_0} \geq a\} = \{x|\, \mathrm{D}(x, y_0) - \mathrm{D}(x, y) \geq a\} \tag{47}$$

$$\sigma^+(y|y_0) = \{x|\, \mathrm{U}_{|y_0} < \mathrm{D}(y, y_0)\} = \{x|\, \mathrm{D}(x, y_0) - \mathrm{D}(x, y) < \mathrm{D}(y, y_0\} \tag{48}$$

$$\sigma^-(y|y_0) = \{x|\, \mathrm{U}_{|y_0} \geq \mathrm{D}(y, y_0\} = \{x|\, \mathrm{D}(x, y_0) - \mathrm{D}(x, y) \geq \mathrm{D}(y, y_0)\} \tag{49}$$

Associated with the sets introduced we define certain "boundary sets", respectively *peripheries* and *hyper-spaces*. Notation and definition for the former type of sets is given by

$$\partial\,\mathrm{B}(r|y_0) = \{x|\, \mathrm{D}(x, y_0) = r\} \text{ and}$$
$$\partial\,\mathrm{B}(x^*|y_o) = \{x|\, \mathrm{D}(x, y_0) = \mathrm{D}(x^*, y_0)\}$$

and for the latter type we use

$$\partial\sigma(y, a|y_0) = \{x|\, \mathrm{D}(x, y_0) - \mathrm{D}(x, y) = a\} \text{ and}$$
$$\partial\sigma(y|y_0) = \{x|\, \mathrm{D}(x, y_0) - \mathrm{D}(x, y) = \mathrm{D}(y, y_0)\}\,.$$

When translating basic parts of Propositions 5, 6 and 7 to the setting we are now considering, we find the following result:

Proposition 8. *Let* D *be a general divergence function on* $X \otimes Y$ *and consider a belief instance* $y_0 \succ X$ *such that* $\mathrm{D}^{y_0} < \infty$. *Then the following results hold for the associated updating games with* y_0 *as prior:*

(i) For any $x^* \in X$, *the largest preparation* $\mathcal{P}$ *for which* $\gamma(\mathcal{P}; y_0)$ *is in equilibrium with* x^* *as bi-optimal strategy, hence with* x^* *as the* D*-projection of* y_0 *on* $\mathcal{P}$, *is the half-space* $\sigma^-(x^*|y_o)$.

(ii) For a fixed updating game $\gamma(\mathcal{P}\ \ y_0)$, *the MinDiv-value* $\mathrm{D}_{\min}(\mathcal{P}; y_0)$ *is the size of the largest strict divergence ball* $\mathrm{B}(r|y_0)$ *which is external to* $\mathcal{P}$, *and the maximal guaranteed updating gain* $\mathrm{Gtu}_{\max}(\mathcal{P}; y_0)$ *is the supremum of* a *for which there exists* $y \succ \mathcal{P}$ *such that the half-space* $\sigma^+(y, a|y_0)$ *is external to* $\mathcal{P}$.

(iii) An updating game $\gamma(\mathcal{P}; y_0)$ *is in equilibrium and has a bi-optimal strategy if and only if, for some* $y \in \mathcal{P}$, *the half-space* $\sigma^+(y|y_0)$ *is external to* $\mathcal{P}$. *When this condition holds,* y *is the bi-optimal strategy, hence the* D*-projection of* y_0 *on* $\mathcal{P}$.

13. Robustness and Core

Let $(\hat{\Phi}, \mathrm{H}, \hat{\mathrm{D}})$ be an effort-based information triple over $X \otimes \hat{Y}$ and $(\Phi, \mathrm{H}, \mathrm{D})$ the derived triple.

We shall study special circumstances under which the crucial condition (29) holds. Consider a preparation $\mathcal{P}$ and let w^* be a permissable Observer-strategy for the game $\hat{\gamma}(\mathcal{P})$, i.e. $w^* \succ \mathcal{P}$. This strategy is *robust* for $\hat{\gamma}(\mathcal{P})$ if the effort with that strategy for Observer is finite and independent of Natures strategy, i.e. if, for some finite constant h, $\hat{\Phi}(x, w^*) = h$ for all $x \in \mathcal{P}$. The constant h is the *level of robustness*. Similarly, $y^* \succ \mathcal{P}$ is *robust* for $\gamma(\mathcal{P})$ at the *level* h if $\Phi(x, y^*) = h$ for all $x \in \mathcal{P}$.

Theorem 5. *[Robustness theorem] Let (x^*, w^*) be an adapted pair of permissable strategies for $\hat{\gamma}(\mathcal{P})$ and assume that w^* is robust with level of robustness h. Then $\hat{\gamma}(\mathcal{P})$ is in equilibrium with h as value and with x^* as bi-optimal strategy. Furthermore, for any $x \in \mathcal{P}$, the Pythagorean inequality holds with equality:*

$$\mathrm{H}(x) + \hat{\mathrm{D}}(x, w^*) = \mathrm{H}_{\max}(\mathcal{P})\,. \tag{50}$$

Similarly, if (x^*, y^*) are permissable strategies for $\gamma(\mathcal{P})$, if y^* is response-equivalent to x^* (hence $y^* = x^*$ if response is injective) and if y^* is robust for $\gamma(\mathcal{P})$ with level of robustness h, then $\gamma(\mathcal{P})$ is in equilibrium with x^* as bi-optimal strategy and, for $x \in \mathcal{P}$,

$$\mathrm{H}(x) + \mathrm{D}(x, y^*) = \mathrm{H}_{\max}(\mathcal{P})\,. \tag{51}$$

The result follows directly from Theorem 2 and the linking identity. The equality (50) or (51) for $x \in \mathcal{P}$ is the *Pythagorean equality*, here in an abstract version. A more compact geometry flavoured formulation of the first part of Theorem 5 à la Corollary 1 runs as follows:

Corollary 4. *If h is finite and $x^* \in \mathcal{P} \subseteq \mathcal{P}^{\hat{x^*}}(h)$, then $h = \mathrm{H}(x^*)$ and $\hat{\gamma}(\mathcal{P})$ is in equilibrium with x^* as bi-optimal strategy.*

Whereas Theorem 2, Corollary 1 and Proposition 2 demonstrate the significance of sublevel sets, Theorem 5 and Corollary 4 does the same but for level sets.

In case response is injective, the second part of Theorem 5 really only involves one element, x^*, as the other element, y^*, has to be identical to x^*. The two essential conditions are one on x^* as a strategy for Nature, viz. that it is consistent, and one on x^* as a strategy for Observer, viz. that it is robust. There can only be one such element. If we drop the condition of consistency, there may be many more such elements. They form what we shall call the *core of* $\gamma(\mathcal{P})$.

The core is defined both for the Y- and for the $\hat{Y}$- domain, and whether or not response is injective, by the formulas

$$\mathrm{core}(\mathcal{P}) = \{y \in Y | \exists h < \infty : \mathcal{P} \subseteq \mathcal{P}^y(h)\}\,, \tag{52}$$

$$\hat{\mathrm{core}}(\mathcal{P}) = \{w \in \hat{Y} | \exists h < \infty : \mathcal{P} \subseteq \mathcal{P}^w(h)\,. \tag{53}$$

By definition, $y \succ \mathcal{P}$ if $y \in \mathrm{core}(\mathcal{P})$ and $w \succ \mathcal{P}$ if $w \in \hat{\mathrm{core}}(\mathcal{P})$.

If need be, we write $\mathrm{core}(\mathcal{P}|\Phi)$ and $\hat{\mathrm{core}}(\mathcal{P}|\hat{\Phi})$.

For a family $\mathbb{P}$ of preparations the *core*, is defined as the intersection of the individual cores:

$$\mathrm{core}(\mathbb{P}) = \bigcap_{\mathcal{P}\in\mathbb{P}} \mathrm{core}(\mathcal{P}) \tag{54}$$

$$\mathrm{core\hat{}}(\mathbb{P}) = \bigcap_{\mathcal{P}\in\mathbb{P}} \mathrm{core\hat{}}(\mathcal{P})\,. \tag{55}$$

The notion is particularly useful for preparation families consisting of strict feasible preparations. Consider a typical such family, $\mathbb{P}^{\mathbf{y}}$, specified by a set $\mathbf{y}= (y_1, \cdots, y_n)$ of elements of Y, cf. (19). From the definitions introduced and from the robustness theorem you derive the following simple, but useful result:

Theorem 6. *Consider a preparation family $\mathbb{P}^{\mathbf{y}}$ with $\mathbf{y} = (y_1, \cdots, y_n)$. Let x^* be a state, put $y^* = x^*$ and assume that $y^* \in \mathrm{core}(\mathbb{P}^{\mathbf{y}}|\Phi)$. Further, put $\mathbf{h} = (h_1, \cdots, h_n)$ with $h_i = \Phi(x^*, y_i)$ for $i = 1, \cdots, n$ and assume that these constants are finite. Then $\mathcal{P}^{\mathbf{y}}(\mathbf{h}) \in \mathbb{P}^{\mathbf{y}}$ and $\gamma\big(\mathcal{P}^{\mathbf{y}}(\mathbf{h})\big)$ is in equilibrium and has x^* as bi-optimal strategy. In particular, x^* is the MaxEnt strategy for $\mathcal{P}^{\mathbf{y}}(\mathbf{h})$.*

We leave it to the reader to formulate analagous results as above for the $\hat{Y}$-domain.

The notions robustness and core also make sense for games defined in terms of utility-based information triples. If $(\mathrm{U}, \mathrm{M}, \mathrm{D})$ is such a triple, we simply apply the above definitions to the associasted effort-based triple $(-\mathrm{U}, -\mathrm{M}, D)$.

The notion of robustness has not received much attention in a game theoretical setting. It is implicit in [55] and in [25] and perhaps first formulated in [23]. Apparently, the existence of suitable robust strategies is a strong assumption. However, for typical models appearing in applications, the assumption is often fulfilled when optimal strategies exist. Results from[26] point in this direction.

14. Adding convexity

It has been recognized since long that notions of convexity play an important role for basic properties of Shannon theory and optimization theory in general. We have deliberately postponed the introduction of this element in our abstract modelling until this late moment. Thereby we demonstrate that a large number of concepts and results, especially those related to games of information, can be formulated quite abstractly and do not require convexity considerations. Our late introduction of convexity also emphasizes exactly where this notion comes in. In this connection note the results starting with Theorem 7 below.

Throughout the section we assume that X is a *convex topological space*, i.e. that X is convex and provided with a Hausdorff topology which renders the algebraic operations continuous. The convex hull of a preparation $\mathcal{P}$ is denoted $\mathrm{co}(\mathcal{P})$ and the closed convex hull is denoted $\overline{\mathrm{co}}(\mathcal{P})$. We assume that the relation of visibility is adapted to the convex structure in the sense that, firstly, for every $y \in Y$, $]y[$ is convex and closed and, secondly, y covers a convex combination, say $y \succ \overline{x} = \sum \alpha_i x_i$, if and only if y covers every x_i with $\alpha_i > 0$. In particular, for every convex combination $\overline{x} = \sum \alpha_i x_i$, it holds that $\overline{x} \succ x_i$ for all i with $\alpha_i > 0$. In the foregoing, as in the sequal, a *convex combination* is understood to be a finite convex combination, often written as above without introducing any special notation for the relevant index set. The topology is referred to as the *reference topology* and convergence of sequences in

this topology is denoted $x_n \to x$ or simillar. We leave it to the interestred reader to keep track of results – possibly suitably reformulated – which only require the sufficiency part in the above requirement for the condition $y \succ \overline{x}$.

For Lemma 1 and Theorem 7 below, an effort-based information triple $(\Phi, \mathrm{H}, \mathrm{D})$ over $X \otimes Y$ is in the background, whereas the last results, Theorem 8 and Theorem 9, are based on a general divergence function D.

For the functions involved in our modelling, emphasis will be on properties of *concavity, convexity* and *affinity*. For the effort function Φ, we will consider such properties for the y-marginals Φ^y – either all of them or only those with $y \in X$. Concavity of Φ^y means that if $\sum \alpha_i x_i$ is a convex combination such that $y \succ \sum \alpha_i x_i$, then $\Phi(\sum \alpha_i x_i, y) \geq \sum \alpha_i \Phi(x_i, y)$. For convexity the inequality sign is turned around and for affinity it is replaced by equality.

For states $x \in X$, conditions of *sequential lower semi-continuity on* X for D^x as well as for D_x will be of significance. Let (x_n) be a convergent sequence in X, say $x_n \to x^*$. Then, for D^x the condition is that $\mathrm{D}(x^*, x) \leq \liminf_{n\to\infty} \mathrm{D}(x_n, x)$ and, for D_x, that $\mathrm{D}(x, x^*) \leq \liminf_{n\to\infty} \mathrm{D}(x, x_n)$. The latter condition is to be understood in the sense that if $\liminf_{n\to\infty} \mathrm{D}(x, x_n) < \infty$ (which presupposes that $x_n \succ x$, eventually), then $x^* \succ x$ and the stated inequality holds.

Basic properties of entropy and divergence under added conditions about the marginals Φ^y are contained in the following result:

Lemma 1. *(i) Assume that all marginals Φ^y with $y \in X$ are concave. Then, for every convex combination $\overline{x} = \sum \alpha_i x_i$ of elements in X,*

$$\mathrm{H}\left(\sum \alpha_i x_i\right) \geq \sum \alpha_i \, \mathrm{H}(x_i) + \sum \alpha_i \, \mathrm{D}(x_i, \overline{x}) \,. \tag{56}$$

In particular, H *is strictly concave on X. If the marginals Φ^y with $y \in X$ are even affine, equality holds in* (56).

(ii) Assume that all marginals Φ^y with $y \in Y$ are affine. Then, for every convex combination $\overline{x} = \sum_i \alpha_i x_i$ of elements in X with $\mathrm{H}(\overline{x}) < \infty$, and for every $y \in Y$ with $y \succ \overline{x}$,

$$\sum \alpha_i \, \mathrm{D}(x_i, y) = \mathrm{D}\left(\sum \alpha_i x_i, y\right) + \sum \alpha_i \, \mathrm{D}(x_i, \overline{x}) \,. \tag{57}$$

In particular, for $y \in Y$, the restriction of D^y to convex preparations $\mathcal{P}$ with $\mathrm{H}_{\max}(\mathcal{P}) < \infty$ is strictly convex.

The result is a slight variation over Theorem 1 of [50]. The proof is straightforward – first establishing (56) and then deriving (57) from (56).

Referring to terminology from [59], we refer to the general validity of (57) as the *compensation identity* with the last term in (57) as *compensation term*. Often, (57) is only needed for elements $y \in X$. Then one need only assume affinity of Φ^y for $y \in X$.

For an even mixture $\overline{x} = \frac{1}{2}x_1 + \frac{1}{2}x_2$, the compensation term is the (abstract) *Jensen-Shannon divergence* between x_1 and x_2 for which we use the notation

$$\mathrm{JSD}(x_1, x_2) = \frac{1}{2}\,\mathrm{D}(x_1, \overline{x}) + \frac{1}{2}\,\mathrm{D}(x_2, \overline{x}) \text{ with } \overline{x} = \frac{1}{2}x_1 + \frac{1}{2}x_2 \,. \tag{58}$$

We turn to a continuation of our study of games of information. An easy consequence of the strict concavity of H in Lemma 1 is that the MaxEnt-strategy for games $\gamma(\mathcal{P}\,|\Phi)$ with a convex preparation is unique – if only $\mathrm{H}_{\max}(\mathcal{P}) < \infty$. A similar remark applies to uniqueness of optimal strategies for Nature in games based on utility.

From our previous study in Section 10 we have realized the central importance of (29), equivalent to $\mathrm{Ri}(y^*|\,\mathcal{P}) \leq \mathrm{H}(x^*)$. From that condition, assuming also $x^* \in \mathcal{P}$, you can conclude equilibrium of $\gamma(\mathcal{P})$ and also identify the bi-optimal strategy. In particular, you can conclude that $\mathrm{H}(x^*) = \mathrm{H}_{\max}(\mathcal{P})$. Adding conditions of convexity, (29) actually follows from the formally weaker condition $\mathrm{H}(x^*) = \mathrm{H}_{\max}$ as we shall now see:

Theorem 7. *Assume that the marginal functions Φ^y with $y \in X$ are concave and that the marginal functions D_x with $x \in X$ are sequentially lower semi-continuous on X. Let $\mathcal{P}$ be a convex preparation and let $x^* \in \mathcal{P}$ have finite entropy. Then, the condition $\mathrm{H}(x^*) = \mathrm{H}_{\max}(\mathcal{P})$ is not only necessary, but also sufficient for (29) to hold, hence for $\gamma(\mathcal{P})$ to be in equilibrium with x^* as bi-optimal strategy.*

Proof. In order to establish (29), consider an element $x \in \mathcal{P}$ and apply (56) to a convex combination of the form $y_n = (1 - \frac{1}{n})x^* + \frac{1}{n}x$. We find that $\mathrm{H}(x^*) \geq \mathrm{H}(y_n) \geq (1 - \frac{1}{n})\,\mathrm{H}(x^*) + \frac{1}{n}\,\mathrm{H}(x) + \frac{1}{n}\,\mathrm{D}(x, y_n)$ from which we conclude that $\mathrm{H}(x) + \mathrm{D}(x, y_n) \leq \mathrm{H}(x^*)$. By sequential lower semi-continuity of D_{x^*}, $x^* \succ x$ and $\mathrm{H}(x) + \mathrm{D}(x, x^*) \leq \mathrm{H}(x^*)$ follows. As $x \in \mathcal{P}$ was arbitrary, (29) holds. The result then follows from Corollary 1. □

One may criticize the result as you cannot apply it to feasible preparations in case the marginals Φ^y are strictly concave, since then the feasible preparations will, typically, not be convex. Rather than reacting negatively towards this observation, we take it as a strong indication that really useful modelling requires that the marginals Φ^y are in fact affine.

The kind of reasoning in the above proof can be expanded, roughly speaking by replacing the occuring optimal strategy by an asymptotically optimal sequence, and then leads to results about existence of the maximum entropy attractor, cf. [50] as pointed to before. We shall not go into that for games based on an effort function but will do so below when we turn to games of updating, cf. Theorem 9.

Translating Theorem 7 to a setting based on utility and formulating it with an assumption of affinity instead of concavity, one finds the following result:

Corollary 5. *Let $(\mathrm{U}, \mathrm{M}, \mathrm{D})$ be a utility-based information triple. Assume that all marginals U^y with $y \in X$ are affine and that all marginals D_x with $x \in X$ are sequentially lower semi-continuous on X. Let $\mathcal{P}$ be a convex preparation and x^* a state in $\mathcal{P}$ with $\mathrm{M}(x^*)$ finite. Then, if $\mathrm{M}(x^*) = \mathrm{M}_{\min}(\mathcal{P})$, $x^* \in \mathrm{ctr}(\mathcal{P})$ and the game $\gamma(\mathcal{P}\,|\,\mathrm{U})$ is in equilibrium and has x^* as bi-optimal strategy. In particular, the pythagorean inequality $\mathrm{M}(x) \geq \mathrm{D}(x, x^*) + \mathrm{M}(x^*)$ holds for every $x \in \mathcal{P}$.*

Finally, we shall investigate updating games under convexity. For this, D is a general divergence function on $X \otimes Y$. We shall say that the compensation identity holds for D if (57) holds for every convex combination $\overline{x} = \sum \alpha_i x_i$ of states and every $y \succ \overline{x}$.

Theorem 8. *Assume that the compensation identity holds for the divergence function D on $X \otimes Y$ and that the marginal functions D_x with $x \in X$ are sequentially lower semi-continuous on X. Consider a*

convex preparation $\mathcal{P}$ and an associated prior y_0. Then, if the D*-projection of y_0 on $\mathcal{P}$ exists, say equal to x^*, it holds that $x^* \succ \mathcal{P}$ and that the updating game $\gamma(\mathcal{P} \,|\, \mathrm{U}_{|y_0})$ is in equilibrium with x^* as bi-optimal strategy. In particular, the pythagorean inequality (42) holds for all $x \in \mathcal{P}$.*

Proof. We shall apply Corollary 5 to the utility-based triple $(\mathrm{U}_{|y_0}, \mathrm{D}^{y_0}, \mathrm{D})$ on $\mathcal{P} \times [\mathcal{P}]$. Consider any $y \in Y$ and any convex combination $\overline{x} = \sum \alpha_i x_i$ of states in $\mathcal{P}$. As $\mathrm{D}^{y_0} < \infty$ on $\mathcal{P}$, the sum $\sum \alpha_i \, \mathrm{D}(x_i, y_0)$ is finite. By the compensation identity, so is the sum $\sum \alpha_i \, \mathrm{D}(x_i, \overline{x})$. For $y \in Y$, we find that

$$\begin{aligned}
\mathrm{U}_{|y_0}(\overline{x}, y) &= \mathrm{D}(\overline{x}, y_0) - \mathrm{D}(\overline{x}, y) \\
&= \Big(\sum \alpha_i \, \mathrm{D}(x_i, y_0) - \sum \alpha_i \, \mathrm{D}(x_i, \overline{x}) \Big) - \Big(\sum \alpha_i \, \mathrm{D}(x_i, y) - \sum \alpha_i \, \mathrm{D}(x_i, \overline{x}) \Big) \\
&= \sum \alpha_i \, \mathrm{D}(x_i, y_0) - \sum \alpha_i \, \mathrm{D}(x_i, y) \\
&= \sum \alpha_i \, \mathrm{U}_{|y_0}(x_i, y) \, .
\end{aligned}$$

Thus the condition of affinity from Corollary 5 is fulfilled. The result follows. □

It lies nearby to search for conditions which ensure existence of the D-projection. This requires extra properties. A sequence (x_n) of states is a JSD- *Cauchy sequence* if

$$\lim_{n,m \to \infty} \mathrm{JSD}(x_n, x_m) = 0 \, . \tag{59}$$

And X is JSD-*complete* if every JSD-Cauchy sequence converges in the reference topology. This notion is adapted from [50] [12]. Let us collect the key results about updating games in one theorem:

Theorem 9. *Assume that the compensation identity holds for* D*, that X is* JSD*-complete, and that, for $x \in X$, the marginal functions D_x as well as D^x are sequentially lower semi-continuous on X. Consider a preparation $\mathcal{P}$ and an associated prior y_0. Then the following holds:*

(i) Observer strategies for $\gamma(\mathrm{co}(\mathcal{P}); y_0)$ and for $\gamma(\mathcal{P}; y_0)$ coincide, i.e. $[\mathrm{co}(\mathcal{P})] = [\mathcal{P}]$, and for every such strategy y, $\mathrm{Gtu}(y|\, \mathrm{co}(\mathcal{P}); y_0) = \mathrm{Gtu}(y|\, \mathcal{P}; y_0)$, hence

$$\mathrm{Gtu}_{\max}(\mathrm{co}(\mathcal{P}); y_0) = \mathrm{Gtu}_{\max}(\mathcal{P}; y_0). \tag{60}$$

(ii) Without adding extra conditions, Observer has a unique optimal strategy, y^*, in the game $\gamma(\mathcal{P}; y_0)$.

(iii) If $\mathcal{P}$ is convex, the game $\gamma(\mathcal{P})$ is in equilibrium and the $\mathrm{D}^{y_0}_{\min}$-attractor exists. This attractor, say x^*, is identical to the optimal Observer strategy y^* from (ii); it is the D*-projection of y_0 on $\mathcal{P}$ if and only if $x^* \in \mathcal{P}$.*

(iv) If $\mathcal{P}$ is closed and convex, the D*-projection of y_0 on $\mathcal{P}$ exists.*

(v) The game $\gamma(\mathcal{P} \,|y_0)$ is in equilibrium if and only if

$$\mathrm{D}_{\min}(\mathrm{co}(\mathcal{P}); y_0) = \mathrm{D}_{\min}(\mathcal{P}; y_0) \, . \tag{61}$$

---

[12] In fact, we only need the condition that a JSD-Cauchy sequence has a convergent subsequence; however, this condition is very close to the stated condition which follows from it under an added assumption of joint lower semi-continuity of D.

Proof. The results follow by adapting Theorem 2 and Corollary 1 of [50] to the present setting. Let us indicate the details.

The proof of (i) is trivial. Now assume that $\mathcal{P}$ is convex and let us analyze the game $\gamma = \gamma(\mathcal{P}; y_0)$. Consider a sequence (x_n) of states in $\mathcal{P}$ such that $\mathrm{D}(x_n, y_0) \to \mathrm{D}$ with $\mathrm{D} = D_{\min}(\mathcal{P}; y_0)$. Put $\delta_n = \mathrm{D}(x_n, y_0) - \mathrm{D}$ and assume that $\delta_n < 1$. Appealing to convexity of $\mathcal{P}$ and to the compensation identity, it is seen that (x_n) is a JSD-Cauchy sequence. Then $x_n \to x^*$, say. We may assume that all δ_n are positive (otherwise we are in a situation which can be covered by Theorem 8). Given $x \in \mathcal{P}$, consider (y_n) given by $y_n = \sqrt{\delta_n}x + (1 - \sqrt{\delta_n})x^*$. Apply the fact that $\mathrm{D} \leq \mathrm{D}(y_n, y_0)$, use the compensation identity, throw away one term and divide by $\sqrt{\delta_n}$. This shows that $\mathrm{D}(x, y_n) + \mathrm{D}(x_n, y_0) \leq \mathrm{D}(x, y_0) + \sqrt{\delta_n}$. Going to the limit and exploiting the semi-continuity property of D_x, we find that $x^* \succ x$ and that $\mathrm{D} \leq \mathrm{U}_{|y_0}(x, x^*)$. As this holds for every $x \in \mathcal{P}$, we conclude that $y^* = x^*$ is in $[\mathcal{P}]$ and that $\mathrm{D} \leq \mathrm{Gtu}(y^* | \mathcal{P}; y_0)$. Thus y^* is an optimal Observer strategy and also, we see that x^* is the $\mathrm{D}^{y_0}_{\min}$-attractor. From these observations and from (i), the remaining properties claimed are easily derived. □

Part II
Applications

15. Protection against misinformation

We start out with a very general type of application which deals with a theme that has been important for the development of the notion of proper score functions.

In a sense, what we shall discuss here is what happens if Nature can communicate. Then we speak instead of *Expert*. And Observer becomes *Customer*. Expert holds the truth, x, or rather, x represents Experts best evaluation of what the truth is. Customer wants to know what Expert thinks about a certain situation and asks Expert for advice – against payment, to be agreed upon. For despicable reasons, Expert may be tempted to advice against better knowing, i.e. to give as advice y, instead of the honest advice x. Misinformation could either be due to the difficulty Expert may have in reaching a true expert opinion or it could be out of self-interest, with Expert taking advantage of false information given to Customer. Or Expert may try to mislead Customer in order to hide a bussiness secret.

We assume that truth will be revealed to both Expert and Customer soon after Expert has given advice to Customer and further, that a proper effort function $\Phi = \Phi(x, y)$ is known to both Expert and Customer. We shall device a payment scheme which will protect Customer against misinformation. The idea is simple. At the time of signing a contract – before advice is given – Customer pays a flat sum to Expert and further, Expert and Customer agree on an insurance scheme stipulating a penalty to be payed by Expert to Customer proportional to $\Phi(x^*, y)$ where x^* represents what really happened and y is the advice given. If Expert is confident that he knows what will happen, he will assume that $x^* = x$ will hold and it will be in his own interest to give to Customer the honest advice $y = x$.

In the literature this scheme is mainly considered based on a *proper score function*, the same as a proper utility function. This gives an obvious variation of the payment scheme with the score function determining payment from Customer to Expert. The most often treated situation is probably that

of weather forecasting with Brier [40] the first and Weijs and Giesen [60] the presently most recent contribution. But also situations from economy and statistics have been studied frequently. Apart from sources just cited we refer to the sources pointed to in Section 6 and to McCarthy [61] as well as to the recent contribution [62] by Chambers. As a final reference we point to Hilden [63] where applications to diagnostics is discussed.

Works cited and their references will reveal a rich literature. With access to our abstract modelling, further meaningful applications, not necessarily tied to probabilistic modelling may emerge.

16. Cause and Effect

We continue with one more section where the basic interpretations are changed. For this we assume that $Y = X$ and define $\hat{X} = \hat{Y}$, equivalently, $\hat{X} = W$. Elements of X are now interpreted as *causes* and response, considered as a map defined on X, as the transformation of a cause into its associated consequence. This change moves the focus from Observers thoughts as discussed in Section 3 to a reflection of causality in Nature, a basic mechanism of the world. The set-up is in this way conceived as a model of *cause and effect*.

Previously we considered possible choices of Observer (for γ- or $\hat{\gamma}$-type games). Now it is more pertinent to focus on consequences – elements of W – as possible observations by Observer of the effect of the actual cause. For $x \in X$ and $w \in W$, $\hat{\Phi}(x, w)$ may now be interpreted as the cost to Observer if he has observed (or belives to have observed) the effect w when the actual cause is x.

Consider the game $\hat{\gamma}$, say with preparation $\mathcal{P} = X$. With the new interpretation in mind it appears particularly pertinent to consider Observers risk associated with the various possible observations.

Concrete situations where the change of interpretation makes sense, involves information theoretical problems of capacity (to be included in a later write-up).

17. Atomic triples and generation by integration

We shall define natural building blocks for information triples with a focus on effort-based triples with effort functions satisfying the perfect match principle. If nothing is said to the contrary, state space and belief reservoir are identical and visibility is the diffuse relation. Except for some final indications of possible expansions of the material in this section as well as of the general theory, we do not involve response or controls.

An *atomic effort-based information triple* $(\phi, \mathrm{h}, \mathrm{d})$ over $I \times I$ with I a subinterval of $]-\infty, \infty[$ consists of three real-valued functions, $\phi = \phi(s, u)$ defined on $I \times I$, $\mathrm{h} = \mathrm{h}(s)$ defined on I and $\mathrm{d} = \mathrm{d}(s, u)$ defined on $I \times I$, such that, for all $(s, u) \in I \times I$,

$$\phi(s, u) = \mathrm{h}(s) + \mathrm{d}(s, u)\,, \tag{62}$$

$$\mathrm{d}(s, u) \geq 0\,, \tag{63}$$

$$\mathrm{d}(s, u) = 0 \Leftrightarrow u = s\,. \tag{64}$$

Clearly, triples defined in this way are indeed effort-based information triples. We may allow that the functions take the value $+\infty$ at eventual endpoints of I.

From Section 14 we know that it is important for the effort function to have affine marginals, fixing the second argument. For this to be the case, there must exist two real-valued functions on I, η and ξ, such that, for $(s,u) \in I \times I$,

$$\phi(s,u) = s\eta(u) + \xi(u)\,. \tag{65}$$

There is a simple way to generate a multitude of information triples of the type just described. The method is inspired by Bregman, [64], who used the construction for other purposes. Given is a *Bregman generator* h which is here understood to be a continuous, strictly concave function on I which is sufficiently smooth on the interior of the interval, say continuously differentiable. We take this function as entropy function. Defining effort and divergence by

$$\phi(s,u) = \mathrm{h}(u) + (s-u)\,\mathrm{h}'(u) \tag{66}$$

$$\mathrm{d}(s,u) = \mathrm{h}(u) - \mathrm{h}(s) + (s-u)\,\mathrm{h}'(u)\,, \tag{67}$$

the information triple $(\phi, \mathrm{h}, \mathrm{d})$ has an effort function with affine marginals, given u.

As two examples of triples constructed this way, we point to the triple

$$\phi(s,u) = u^2 - 2su\,, \tag{68}$$

$$\mathrm{h}(s) = -s^2\,, \tag{69}$$

$$\mathrm{d}(s,u) = (s-u)^2 \tag{70}$$

over $]-\infty, +\infty[$ [13] and the triple

$$\phi(s,u) = u - s + s\ln\frac{1}{u}\,, \tag{71}$$

$$\mathrm{h}(s) = s\ln\frac{1}{s}\,, \tag{72}$$

$$\mathrm{d}(s,u) = u - s + s\ln\frac{s}{u} \tag{73}$$

over $[0,\infty]$.

The first triple leads to basic concepts of real Hilbert space by a natural process of *integration* and by a similar process, the second leads to basic concepts of Shannon information theory. Before returning to that, we note that there is a natural way to generalize the second example without distroying the essential property of affinity of the marginals ϕ^u. One simply replaces logarithms with *deformed logarithms* $\ln_q$, where q is a real parameter. Following [65], the deformed logarithms are given by the expression

$$\ln_q t = \begin{cases} \ln t \text{ if } q = 1 \\ \frac{1}{1-q}\left(t^{1-q} - 1\right) \text{ otherwise}\,. \end{cases} \tag{74}$$

---

[13] the reader will realize that it is more natural to consdier the triple $(-\phi, -h, d)$ which corresponds to a change of focus from effort-based triples to utility-based triples. This then requires convexity rather than concavity conditions on the generator. Though we find it desirable to state concepts and results for both versions, we shall leave this to the interested reader

The *deformed generator* h_q defined by

$$\mathrm{h}_q(s) = s \ln_q \frac{1}{s}\,, \tag{75}$$

is a genuine Bregman generator on $[0, \infty]$ for all positive values of q. The atomic triples obtained from these generators, denoted $(\phi_q, \mathrm{h}_q, \mathrm{d}_q)$, constitute the *Tsallis family* of atomic information triples. For $q = 1$, we find the expressions given by (71) – (73) and for other values of $q > 0$ we find the expressions

$$\phi_q(s,u) = u^q + \frac{qu^{q-1} - 1}{1-q} s\,, \tag{76}$$

$$\mathrm{h}_q(s) = \frac{1}{1-q}\big(s^q - s\big)\,, \tag{77}$$

$$\mathrm{d}_q(s,u) = u^q + \frac{1}{1-q}\big(qsu^{q-1} - s^q\big)\,. \tag{78}$$

Here $q > 0$ but, in fact, we may consider the case $q = 0$ as a degenerate case. It gives the triple $(1-s, 1-s, 0)$ which is not a genuine information triple because the divergence function is identically 0. For all other values, i.e. for $q > 0$, the essential fundamental inequality, i.e. the statement that $\mathrm{d}_q(s,u) \geq 0$ with equality only if $u = s$, holds.

Let us return to the process of integration hinted at above. In fact, this process may be applied to any family of information triples and gives us new triples to work with. And by the linearity of integration, the essential property of affinity of marginals (the Φ^y's) is preserved. Thus divergence functions constructed this way will, according to Lemma 1, satisfy the compensation identity.

Consider, as the key case, integration of one and the same atomic triple $(\phi, \mathrm{h}, \mathrm{d})$ over some interval I with Bregman generator h. Let T be a set provided with a Borel structure and an associated measure μ. Let $X = Y$ be the function space consisting of all measurable functions $x : T \mapsto I$ for which the integral defining $\mathrm{H}(x)$ below converges. Define the full triple $(\Phi, \mathrm{H}, \mathrm{D})$ by integration, i.e.

$$\Phi(x,y) = \int_T \phi\big(x(t), y(t)\big) d\mu(t)\,, \tag{79}$$

$$\mathrm{H}(x) = \int_T \mathrm{h}\,\big(x(t)\big) d\mu(t) \tag{80}$$

$$\mathrm{D}(x,y) = \int_T \mathrm{d}\,\big(x(t), y(t)\big) d\mu(t)\,. \tag{81}$$

If the generator is non-negative, we may enlarge the function space and consider all measurable functions $x : T \mapsto I$. Clearly, $(\Phi, \mathrm{H}, \mathrm{D})$ is a well defined information triple over $X \times Y$. The divergence functions which can be obtained in this way are *Bregman divergencies*. Note that with this construction, the essential fundamental inequality even holds "coordinatewise". This follows by (63) and (64) applied to the atomic divergence function d. For this reason, we refer to (63) and (64) as the *pointwise fundamental inequality*. We add that Bregman divergence makes sense for any pair of measurable functions $x \in X$ and $y \in Y$ as the integrand in (81) is non-negative. Bregman divergence may be used to modify visibility. Indeed, one may take $X \otimes Y$ to consist of all pairs $(x,y) \in X \times Y$ with $\mathrm{D}(x,y) < \infty$.

For the Bregman generator $\mathrm{h}(s) = -s^2$ over $I =]-\infty, +\infty[$, cf. (68)-(70), the construction leads to Hilbert space quantities with $\mathrm{H}(x) = -\|x\|^2$ and $\mathrm{D}(x,y) = \|x - y\|^2$.

For the Bregman generator $\mathrm{h}(s) = s \ln \frac{1}{s}$, cf. (71)-(73), the construction has a few variants. Most classical, we may consider $I = [0,1]$ and a countable set $\mathbb{A}$, the *alphabet*, which is chosen for T and provided with counting measure. If further we take as $X = Y$ the space of probability distributions over $\mathbb{A}$, Φ becomes Kerridge inaccuracy, H Shannon entropy and D Kullback-Leibler divergence. If we generalize to cover non-discrete settings, entropy will only be finite for distributions with countable support. But the generalization of divergence makes sense more generally. For instance, we may consider the generator on $I = [0,\infty]$ and consider an arbitrary measure space with T as basic set, provided with some measure μ. As $X = Y$ we can then, as one possibility, take the set of measures absolutely continuous with respect to μ and with finite-valued Radon-Nikodym derivatives with respect to μ. For two such measures, say $P = pd\mu$ and $Q = qd\mu$ we find the *generalized Kullback-Leibler divergence* given by

$$\mathrm{D}(P,Q) = \int \Big(p(t) \ln \frac{p(t)}{q(t)} + q(t) - p(t) \Big) d\mu(t) \,. \tag{82}$$

If we restrict attention to finite measures P and Q with the same total mass, this reduces to the standard expression $\int p \ln \frac{p}{q} d\mu$. The standard expression also gives a divergence measure if the two measures are finite and $Q(T) \leq P(T)$ and, moreover, the important compensation identity also holds in this case since the additional terms (stemming from $u - s$ in (94)) are integrable and affine.

When extending these constructions to cover also integration of the Tsallis family $(\phi_q, \mathrm{h}_q, \mathrm{d}_q)$, we obtain the triples $(\Phi_q, \mathrm{H}_q, \mathrm{D}_q)$ defined over appropriate function spaces, typically representing probability distributions. Here H_q – in the discrete case – is *Tsallis entropy*.

Whereas there is no need to comment on the case $q = 1$ which leads to classical concepts of Shannon theory, we shall comment on the extension to Tsallis type concepts, especially on Tsallis entropy. Tsallis' paper [2] is from 1988 but, originally, the notion goes back to Havrda and Charvát [66], to Daróczy [67] and to Lindhard and Nielsen [68], [69] who all, independantly of eachother, found the notion of interest. Characterization via functional equations was studied in Aczél and Daróczy [70], see also the reference work [71] as well as [39]. Regarding the physical literature, there is a casual reference to Lindhard's work in one of Jaynes' papers, [72]. But only after the publication of Tsallis 1988-paper mathematicians and, especially, physicists took an interest in the "new" entropy measure. We refer to the database maintained by Tsallis with more than 2000 references. We shall not comment on that except for a reference to Naudts, [73] who also emphasized the convenient approach via Bregman generators.

As a final comment on the process of integration, we note that if the divergence function in (81) is allowed to vary with t: $\mathrm{d} = \mathrm{d}_t$, then not all these functions need be genuine divergence functions. They should all be non-negative but the implication $\mathrm{d}_t(x,y) = 0 \Rightarrow y = x$ need only hold for some set of positive μ-measure. Of course, this observation also applies if you work with integration of more general divergencies than atomic ones. We shall right away see an example where this remark is relevant, viz. for the process of *relativization*, cf. Section 8. This process may be viewed as a special case of generation by integration. Indeed, assuming that Φ^{y_0} is finite, the *relativized triple* (15) may be obtained from the two triples $(\Phi, \mathrm{H}, \mathrm{D})$ and $(-\Phi^{y_0}, -\Phi^{y_0}, 0)$ by simple addition. Note that the second triple is not a genuine information triple as divergence is identically zero. Anyhow, in agreement with the remark above, the resulting triple is a genuine information triple. If the first triple has affine marginals, so does the second

and hence also the resulting triple. As in Section 8, we may consider the relativized triple if only D^{y_0}, rather than Φ^{y_0}, is assumed to be finite.

We end the section by indicating how the Bregman construction of information triples can trigger ideas leading to an expansion of not only the atomic triples that arise in this way but of the whole theoretical set-up. To see what we have in mind, it appears essential to introduce a control space. The tangent lines in the standard construction may be taken as controls. In fact any line *controlling* h in the sense that it lies above the subgraph of h may be considered. Response in the setting above with a strictly concave generator is then, given $u \in I$, the *best control at* u in an obvious sense. But really, why restrict h to be strictly concave? The triples generated are related to problems of global optimization, especially maximization, and for such problems much more general functions can be allowed. If we just allow cracks on the graph of the generator, we realize that then response should, sensibly, be replaced by a set-valued map. Going further, we suggest to the class of "controllable", e.g. upper bounded, usc-functions, *upper semi-continuous* functions, the best controls at a point u will not pass through $(u, \mathrm{h}(u))$ but through a point on an "upper envelope" of h, defined appropriately.

On philosophical grounds the indicated approach appears preferable. An appeal to methods depending on global control (via half-spaces corresponding to the straight lines or via other geometrical figures) seems better adjusted to the type of problems under discussion than methods depending on local behaviour of functions with the differential calculus as a central tool. Our claim is put forward in spite of the frightening efficiency of these tools – as also witnessed by the ease of definition of our two concrete examples, cf. (68) – (73). For basic theoretical investigations an expansion of the theory is desirable. This poses certain challanges, especially a new form of the fundamental inequality is essential. The author plans to return to this in the near future.

18. A geometric model

In this section $X = Y$ denotes a real Hilbert space. As visibility we take the diffuse relation $X \otimes Y = X \times Y$. Consider the divergence function

$$\mathrm{D}(x, y) = \|x - y\|^2 \tag{83}$$

on $X \times Y$. As we saw in Section 17, D may be constructed by integration from the Bregman generator in (70) and as such it satisfies the compensation identity (57). In this case, the identity is of central importance for classical least squares analysis[14].

If $y_0 \in Y$ is a prior and the preparation $\mathcal{P}$ is convex and closed, the D-projection x^* of y_0 on $\mathcal{P}$ exists; it is the unique point in $\mathcal{P}$ which is closest in norm to y_0[15]. As standard convexity- and continuity assumptions are also in place, Theorem 8 applies. It follows that the game $\gamma(\mathcal{P}; y_0)$ is in equilibrium with the D-projection x^* as bi-optimal strategy. The updating gain for this game is given by (13), i.e.

$$\mathrm{U}_{|y_0}(x, y) = \|x - y_0\|^2 - \|x - y\|^2 . \tag{84}$$

---

[14] apparenttly, the identity has no special name in this setting – it would not be unjustified to attach Gauss' name to it.

[15] though classical, the reader may appreciate to note that this existence result is derived with ease and some elegance from the compensation identity and completeness of Hilbert space.

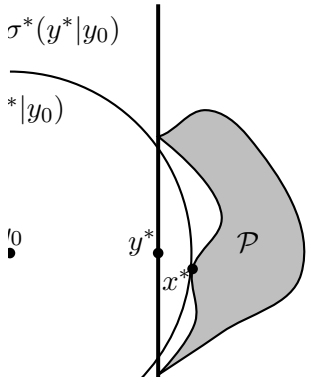
In this case the Pythagorean inequality reduces to the classical inequality

$$\|x - y_0\|^2 \geq \|x - x^*\|^2 + \|x^* - y_0\|^2 , \tag{85}$$

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

Combining Proposition 8 and Theorem 9 we obtain rather complete information about the updating games, also for preparations which are not necessarily convex. For instance, Figure 1 illustrates a case with unique optimal strategies for both players and yet, the game is not in equilibrium. Figure 2 illustrates a typical case with a game in equilibrium. For both figures, x^* denotes the optimal strategy for Nature and y^* the optimal strategy for Observer. Indicated on the figures you also find the largest strict divergence ball $\mathrm{B}(x^*|y_0)$ and the largest half-space $\sigma^+(y^*|y_0)$ which is external to $\mathcal{P}$. The two values of the game can then be determined from the figures, $\|x^* - y_0\|^2$ for Nature, respectively $\|y^* - y_0\|^2$ for Observer.

Figure 1. Game not in equilibrium



It is easy to identify the feasible preparations. The strict ones are affine subspaces and the slack ones are convex polyhedral subsets. We shall determine the core of families of strict preparations:

Figure 2. Game in equilibrium

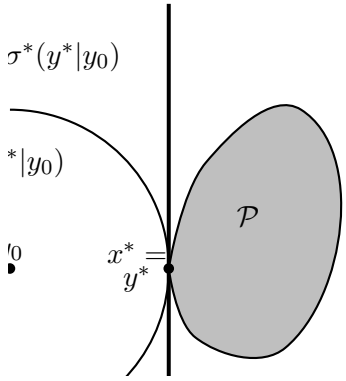
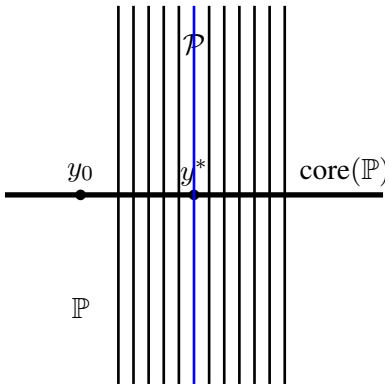


Figure 3. Preparation family and its core



Proposition 9. *Consider a family $\mathbb{P} = \mathbb{P}^{\mathbf{y}}$ of strict feasible preparations determined by finitely many points $\mathbf{y} = (y_1, \cdots, y_n)$ in X. The core of this family consists of all points in the affine subspace through y_0 generated by the vectors $y_i - y_0$; $i = 1, \cdots, n$, i.e.*

$$\mathrm{core}(\mathbb{P}) = \left\{ y_0 + \sum \alpha_i (y_i - y_0) \middle| (\alpha_1, \cdots, \alpha_n) \in \mathbb{R}^n \right\}. \tag{86}$$

Proof. An individual member $\mathcal{P}$ of $\mathbb{P}$ is determined by considering all $x \in X$ for which the values of $\mathrm{U}_{|y_0}(x, y_i)$; $i = 1, \cdots, n$ have been fixed. Note that fixing these values is the same as fixing the inner products $\langle x - y_0, y_i - y_0 \rangle$ or, equivalently, the inner products $\langle x, y_i - y_0 \rangle$. If y^* is of the form given by (86), $y^* = y_0 + \sum \alpha_i (y_i - y_0)$, then $\langle x, y^* - y_0 \rangle = \sum \alpha_i \langle x, y_i - y_0 \rangle$ and we realize that this is independent of x if x is restricted to run over some preparation in $\mathbb{P}$. Then also $\mathrm{U}_{|y_0}(x, y^*)$ is independent of x when x is so restricted. We conclude that $y^* \in \mathrm{core}(\mathbb{P})$. This proves the inclusion "$\supseteq$" of (86).

To prove the other inclusion, assume, as we may, that $y_0 = 0$ and that the y_i forms an orthonormal system. Consider a point $y^* \in \mathrm{core}(\mathbb{P})$. Determine $\mathcal{P} \in \mathbb{P}$ such that $y^* \in \mathcal{P}$. By Theorem 5, y^* is the bi-optimal strategy of $\gamma(\mathcal{P}; y_0)$. Let c_i; $i = 1, \cdots, n$ denote the common values of $\langle x, y_i \rangle$ for $x \in \mathcal{P}$. Then $x^* = \sum c_i y_i$ is the orthogonal projection of $y_0 = 0$ on $\mathcal{P}$, hence $y^* = x^*$. This argument shows that the core is contained in the subspace generated by the y_i. This is the result we want as we assumed that $y_0 = 0$. □

In order to determine the projection of y_0 on a specific preparation $\mathcal{P} = \mathcal{P}^{\mathbf{y}}(\mathbf{h}) \in \mathbb{P}$, we simply intersect $\mathrm{core}(\mathbb{P})$ with $\mathcal{P}$. If you do this analytically, one may avoid trivial cases and assume that $y_i - y_0$; $i = 1, \cdots, n$ are linearly independent. In Figure 3 we have illustrated the situation in the simple case when $n = 1$.

19. Maximum Entropy Problems

Terminology and results of e.g. Sections 7, 10 and 13, are evidently inspired by maximum entropy problems of classical information theory. We shall now see how these problems can be handled as applications of the abstract theory. The problems concern inference of probability distributions over some finite or countably infinite *alphabet* $\mathbb{A}$, typically with preparations given in terms of certain constraints, typically *"moment constraints"*. We shall leave it to the interested reader to go through specific examples in detail. Examples are numerous, from information theory proper, from statistics,

from statistical physics or elsewhere. The variety of possibilities may be grasped from the collection of examples in Kapur's monograph [76]. The abstract results developed in Part I can favourably be applied to all such examples. This then has a unifying effect. However, for many concrete examples, the main work consists in actually verifying the validity of concrete instances of Nash's inequality (29) or of appropriate calculations related to robustness and core, cf. Theorems 5 and 6.

As state space and belief reservoir we may take $X = Y = M_+^1(\mathbb{A})$, the set of probability distributions over $\mathbb{A}$. Modelling can then be based on the information triple $(\Phi, \mathrm{H}, \mathrm{D})$ given by Kerridge inaccuracy, Shannon entropy and Kullback-Leibler divergence:

$$\Phi(P, Q) = \sum_{a \in \mathbb{A}} P(a) \ln \frac{1}{Q(a)} \tag{87}$$

$$\mathrm{H}(P) = \sum_{a \in \mathbb{A}} P(a) \ln \frac{1}{P(a)} \tag{88}$$

$$\mathrm{d}(P, Q) = \sum_{a \in \mathbb{A}} P(a) \ln \frac{P(a)}{Q(a)} \,. \tag{89}$$

However, we find it more illuminating to introduce the action space $\hat{Y} = K(\mathbb{A})$ consisting of all *code length sequences* κ, in short *codes*, which are functions $\kappa : \mathbb{A} \mapsto [0, \infty]$ satisfying *Kraft's equality*

$$\sum_{a \in \mathbb{A}} \exp(-\kappa(a)) = 1 \,. \tag{90}$$

As response we take the bijection $Q \mapsto \hat{Q}$ from Y to $\hat{Y}$ given by

$$\hat{Q}(a) = \ln \frac{1}{Q(a)} \; ; a \in \mathbb{A} \,. \tag{91}$$

The interpretation of code length sequences is well known from information theory. We have merely replaced binary logarithms with natural ones and allowed values which are not necessarily integers. The information triple to work with in the $\hat{Y}$-domain is $(\hat{\Phi}, \mathrm{H}, \hat{\mathrm{D}})$ given by

$$\hat{\Phi}(P, \kappa) = \sum_{a \in \mathbb{A}} P(a) \kappa(a) \tag{92}$$

$$\mathrm{H}(P) = \sum_{a \in \mathbb{A}} P(a) \ln \frac{1}{P(a)} \tag{93}$$

$$\hat{\mathrm{D}}(P, \kappa) = \sum_{a \in \mathbb{A}} P(a) (\kappa(a) - \hat{P}(a)) \,. \tag{94}$$

Standard results from information theory, cf. also Section 17, show that $(\Phi, \mathrm{H}, \mathrm{D})$ and $(\hat{\Phi}, \mathrm{H}, \hat{\mathrm{D}})$ are genuine information triples and the machinery of the abstract theory applies.

Various extra elements of the modelling may be introduced. For instance one may take the deterministic distributions as belief instances of certainty. And control $\kappa \succ P$ could mean that the implication $\kappa(a) = \infty \Rightarrow P(a) = 0$ holds. Then visibility $Q \succ P$ amounts to absolute continuity of P with respect to Q. Note that though discrete alphabets with more than enumerably many elements in principle could be considered, that would contradict the sensible requirement (3). Another variation will be to allow Observer to choose also incomplete distributions (with point masses summing up to a

number less than one). From an interpretation point of view this is perfectly sensible and may at times be technically convenient.

Let us also illuminate the role of the feasible preparations. Thinking of states P as determining the distribution of a random element ξ over $\mathbb{A}$, it is often desirable to consider preparations corresponding to the prescription of one or more mean values of ξ. A typical preparation consists of all $P \in X$ such that

$$\sum_{a\in\mathbb{A}} P(a)\lambda(a) = c \tag{95}$$

with c a given constant and $\lambda = (\lambda(a))_{a\in\mathbb{A}}$ a given function. This is a strict feasible preparation if and only if the *partition function* (a special *Dirichlet series*),

$$Z(\beta) = \sum_{a\in\mathbb{A}} \exp(-\beta\lambda(a)) \tag{96}$$

has a finite abscissa of convergence, i.e. converges for some finite constant β, cf. [25] (or monographs on Dirichlet series). However, for the most important part to us, the "if"-part, this is clear. Indeed, if the condition is fulfilled, there exist constants α_0 and β_0 such that the function κ_0 given for $a \in \mathbb{A}$ by

$$\kappa_0(a) = \alpha_0 + \beta_0\lambda(a) \tag{97}$$

defines a code. Then $\mathcal{P} = \mathcal{P}^{\kappa_0}(k)$ for some constant k, hence it is a strict feasible preparation of genus 1. It is a member of the preparation family $\mathbb{P} = \mathbb{P}^{\kappa_0}$. Consider, for any β with $Z(\beta) < \infty$, the code κ_β given for $a \in \mathbb{A}$ by

$$\kappa_\beta(a) = \ln Z(\beta) + \beta\lambda(a)\,. \tag{98}$$

Then this code is a member of $\mathrm{core}\hat{}(\mathbb{P}^{\kappa_0})$ as is easily seen. In fact all members of the core are of this form (fact not proved here). If we can adjust the parameter β such that the corresponding distribution P_β given by

$$P_\beta(a) = \frac{\exp(-\beta\lambda(a))}{Z(\beta)} \text{ for } a \in \mathbb{A} \tag{99}$$

is a member of the original preparation $\mathcal{P}$, this must be the maximum entropy distribution of $\mathcal{P}$, as follows from Theorem 6 (translated to the $\hat{Y}$-domain).

Schematically: In searching for the MaxEnt distribution of a given preparation, first identify the preparation as a feasible preparation (of genus 1 or higher), calculate if possible the appropriate partition function and adjust parameters to fit the original constraint(s). This gives you the MaxEnt distribution searched for. If calculations are prohibitive, you may use numerical and/or graphical methods instead.

The literature very often solves MaxEnt-problems of the type considerd by the introduction of *Lagrange multipliers*. As shown, this is not necessary. The approach building on the abstract theory of Part I appears preferable. For one thing, the fact that you obtain a maximum for the entropy function (and not just a stationary point) is automatic – it is all hidden in the fundamental inequality. And, for another, the quantities you work with when appealing to the abstract theory, have natural interpretations. The Lagrange multipliers in the standard approach, especially within statistical physics, are of significance. However, they also come up as natural quantities to consider if you tackle MaxEnt-problems as here suggested.

20. Determining D-projections

The setting is basicly the same as in the previous section, especially we again consider a preparation $\mathcal{P}$ given by (95). The problem we shall consider is how to update a given prior $Q_0 \in M_+^1(\mathbb{A})$. Then, the triple $(\Phi, \mathrm{H}, \mathrm{D})$ given by (87), (88) and (89) is no longer relevant but should be replaced by the triple $(\mathrm{U}_{|Q_0}, \mathrm{D}^{Q_0}, \mathrm{D})$ as defined in Section 8, cf. (13). This makes good sense if D^{Q_0} is finite on $\mathcal{P}$. The update we seek is the D-projection of Q_0 on $\mathcal{P}$ as defined in Section 11 in connection with (39).

We shall apply much the same strategy as in the previous section. However, we choose not to introduce response and an action space in this setting [16]. Instead, we work directly in the Y-domain and seek a representation of $\mathcal{P}$ as a primitive strict preparation, now to be understood with respect to $\mathrm{U}_{|Q_0}$. Analyzing what this amounts to, we find that if the partition function, now defined by

$$Z(\beta) = \sum_{a \in \mathbb{A}} Q_0(a) \exp(-\beta\lambda(a)\,, \tag{100}$$

converges for some $\beta < \infty$, a representation as required is indeed possible. Assuming that this is the case we realize that for each β with $Z(\beta) < \infty$, the distribution Q_β defined by

$$Q_\beta(a) = \frac{Q_0(a)\exp(-\beta\lambda(a))}{Z(\beta)} \text{ for } a \in \mathbb{A} \tag{101}$$

is a member of the core of $\mathcal{P}$. Then it is a matter of adjusting β such that Q_β is consistent, and we have found the sought update.

The cancellation that takes place from (12) to (13) allows an extension of the discussion of updating from the discrete setting to general measurable spaces. Indeed, as is well known, cf. also Section 17, the definition of Kullback-Leibler divergence makes good sense in the general case. Thus updating and strategies for the calculation of D-projections as presented above in the discrete case extends without difficulty to the general case. For instance, one may consider instead of $\mathbb{A}$, a general measurable space provided with a σ-finite *reference measure* μ and then work with distributions that have densities with respect to μ. If the prior has density q_0, the partition function should be $Z(\beta) = \int \exp(-\beta\lambda) q_0 d\mu$. Further details and consideration of concrete examples are left to the interested reader.

21. Tsallis worlds

We turn to a study of worlds defined by probabilistic considerations. Only discrete probabilities will be considered. The main result, Theorem 10 was presented in a different form in [35] and, less formally, in [34]. Proofs were only indicated in these sources. [17]

The key point is that so far our introduction of entropy measures of information theory in Section 17 was dictated by a seemingly arbitrary consideration of various Bregman generators, the h_q's. This

---

[16] this can be done with controls consisting of *code improvements* which are code length functions measured relative to the code κ_0 associated with Q_0 but is less convincing, especially for extensions beyond the discrete case

[17] To prevent any misunderstanding, large parts of the material from [35] have been copied or only slightly changed for the present submission, thereby improving readability and making the manuscript self contained. For an eventual final publication, this material may be omitted or only included in condensed form.

does not in itself give rise to an acceptable interpretation. Of course, we know how to motivate the introduction of Shannon entropy, most convincingly via coding. But despite some attempts to extend this to the more general entropy measures, cf. [77], [78] and references there as well as [79], this has not yet been really successful. And it appears that the supporters of the new entropy measures had no and still has no convincing interpretation.

The interpretation offered here points to rules of interaction for the physical world around us as a possible key. It appears especially appealing for $0 < q \leq 1$. However, general acceptance among researchers of statistical physics must be based on physical evidence. The mathematical evidence provided here is only indicative that perhaps there is some physical explanation out there.

By $\mathbb{A}$ we denote a discrete set, the *alphabet*, of *basic events*. The events are identified by an *index*, typically denoted by i. Sensible indexing depends on the concrete application. The semiotic assignment of indices should facilitate technical handling and catalyze semantic awareness. As we have no concrete application in mind, we shall not introduce any extra structure related to the choice of indices.

The state space X is taken to be identical to the belief reservoir Y and equal to $M_+^1(\mathbb{A})$, the set of probability distributions over $\mathbb{A}$. Generically, $x = (x_i)_{i\in\mathbb{A}}$ will denote a state and $y = (y_i)_{i\in\mathbb{A}}$ a belief instance. Thus x and y are characterized by their point probabilities. As set of certain belief instances we take the subset $Y_{\mathrm{det}} \subseteq M_+^1(\mathbb{A})$ of deterministic distributions. A knowledge instance $z = (z_i)_{i\in\mathbb{A}}$ will be a sequence of real numbers over $\mathbb{A}$, not necessarily a probability distribution. The interpretation of z_i is as the *weight* with which the basic event indexed by i will be presented to Observer. We do not need the action space and the response function in this section.

Visibility $y \succ x$ means that x is absolutely continuous w.r.t. y or, expressed differently, that the *support* of x – the set of i with $x_i > 0$ – is contained in the support of y.

The interaction between x, y and z is given by an interactor Π, cf. Section 5. We assume that Π acts *locally*, i.e. that there exists a function π, the *local interactor*, defined on $[0,1] \times [0,1]$ such that, when $z = \Pi(x,y)$, $z_i = \pi(x_i, y_i)$ for all $i \in \mathbb{A}$. The world defined this way is denoted Ω_π. From now on, we talk about *the interactor* when we in fact mean the local interactor.

The interactor is *sound* if $\pi(s,s) = s$ for every $s \in [0,1]$. All interactors we will deal with will be sound. Regarding regularity conditions, we assume that π is finite on $[0,1]\times]0,1]$, continuous on $[0,1]\times[0,1]\setminus\{(0,0)\}$ and continuously differentiable on $]0,1[\times]0,1[$. The interactor is *weakly consistent* if $\sum_{i\in\mathbb{A}} z_i = 1$ whenever x and $y \succ x$ are probability distributions over $\mathbb{A}$ and $z = \Pi(x,y)$. If we can even conclude that z is a probability distribution, π is *strongly consistent*. For $q \in \mathbb{R}$, the interactor π_q is given by

$$\pi_q(s,t) = qs + (1-q)t \text{ for } (s,t) \in [0,1] \times [0,1]\,. \tag{102}$$

These interactors are weekly consistent and, for $0 \leq q \leq 1$, even strongly consistent. The corresponding worlds are denoted Ω_q. This is consistent with the notation introduced in Section 5.

From the essential condition that interaction takes place locally and an added condition of weak consistency, we are left with the worlds Ω_q:

Lemma 2. *Consider a world Ω_π with atomic situations as described above, involving discrete probability distributions over the alphabet $\mathbb{A}$. Assume that $\mathbb{A}$ is countably infinite and that π is weekly consistent. Then $\pi = \pi_q$ with $q = \pi(1,0)$. In particular, the interactor is sound.*

Proof. By weak consistency, $\pi(s,t)+\pi(1-s,1-t)=1$ for all $(s,t)\in[0,1]\times[0,1]$. In particular, π is finite valued. Also, $\pi(0,1)=1-q$. Consider $(x_0,y_0)=(0,1)$ and $(x_i,y_i)=(\frac{1}{n},0)$ for $i=1,\cdots,n$ and apply weak consistency. We find that $\pi(\frac{1}{n},0)=q\frac{1}{n}$. Then consider as (x_i,y_i) the vectors $(0,1),(\frac{1}{n},0),\cdots,(\frac{1}{n},0),(\frac{p}{n},0)$. Using what we already know, and again appealing to weak consistency, we conclude that $\pi(s,0)=qs$ for all rational s. By continuity, this formula holds for all $s\in[0,1]$. From the first step of the proof, $\pi(0,t)=(1-q)t$. Finally, $\pi=\pi_q$ follows by weak consistency applied to $(s,t),(1-s,0),(0,1-t)$. □

We shall search for proper effort functions for the worlds Ω_π. For this we introduce a (local) *descriptor* as any continuous function $\kappa:[0,1]\to[0,\infty]$ which is finite, strictly decreasing and continuously differentiable on $]0,1]$, vanishes at $t=1$ and which satisfies the following condition of *normalization*:

$$\kappa'(1)=-1\,. \tag{103}$$

Note that this definition does not depend on π.

The *description effort generated by π and κ* is the function defined for atomic situations by

$$\Phi_\pi(x,y|\kappa)=\sum_{i\in\mathbb{A}}\pi(x_i,y_i)\kappa(y_i)\,. \tag{104}$$

Some comments on the interpretations are in order. For $t\in[0,1]$, $\kappa(t)$ is the effort, when using the descriptor κ, which Observer must allocate to any basic event which he believes has probability t. This effort has to be multiplied with the force with which the basic event in question is presented to Observer. Accumulating the local contributions $\pi(x_i,y_i)\kappa(y_i)$, you obtain the description effort as given by (104).

The condition $\kappa(1)=0$ reflects the fact that if you feel certain that a basic event will occur, there is no reason why you should allocate any effort at all to such an event. Also, it is to be expected that an event with low probability is more difficult to describe than one with high probability, therefore, we may just as well assume from the outset that κ is decreasing. The condition (103) is a condition of normalization which allows one to compare entropy, divergence and other quantities corresponding to different descriptors and even across different worlds. The unit defined by this condition we call the *natural information unit*, the "nat".

The reader may wish to note that if only π is sound, description effort vanishes for any atomic situation of certainty.

Denote by $\delta_{\pi\kappa}$ the function on $[0,1]\times[0,1]$ given by the expression

$$\delta_{\pi,\kappa}(s,t)=\big(\pi(s,t)\kappa(t)+t\big)-\big(\pi(s,s)\kappa(s)+s\big)\,. \tag{105}$$

We say that $\delta_{\pi,\kappa}$ satisfies the *pointwise fundamental inequality* if, for every $(s,t)\in[0,1]\times[0,1]$, $\delta_{\pi,\kappa}(s,t)\geq 0$ with equality only if $t=s$.

Lemma 3. *If the pointwise fundamental inequality holds for $\delta_{\pi,\kappa}$, the effort function $\Phi_\pi(\cdot,\cdot|\kappa)$ is proper.*

Proof. For every (x,y) with $y\succ x$,

$$\begin{aligned}\Phi_\pi(x,y|\kappa)+1&=\sum\big(\pi(x_i,y_i)\kappa(y_i)+y_i\big)\\&\geq\sum\big(\pi(x_i,x_i)\kappa(x_i)+x_i\big)\\&=\Phi_\pi(x,x|\kappa)+1\,,\end{aligned}$$

The result follows. □

Incidentally, we note that by replacing the first equality above with an inequality ("≥") we can expand the setting by allowing incomplete probability distributions for belief instances.

Lemma 4. *Assume that the alphabet $\mathbb{A}$ has at least three elements. Let π be a local interactor and denote by χ the function on $]0,1[$ defined by*

$$.\chi(t) = \frac{\partial \pi}{\partial t}(t,t)\,. \tag{106}$$

Under the assumption that χ is bounded in the vicinity of $t = 1$, there can only exist one descriptor κ such that description effort given by (104) defines a proper effort function. Indeed, κ must be the unique solution in $]0,1[$ to the differential equation

$$\chi(t)\kappa(t) + t\kappa'(t) = -1 \tag{107}$$

for which $\kappa(1) = \lim_{t\to 1}\kappa(t) = 0$.

Proof. Assume that κ exists with $\Phi_\pi(\cdot,\cdot|\kappa)$ proper. For $0 < t < 1$ put

$$f(t) = \chi(t)\kappa(t) + t\kappa'(t)\,.$$

Consider a, for the time, fixed probability vector $x = (x_1, x_2, x_3)$ with positive point probabilities. Then the function F given by

$$F(y) = F(y_1, y_2, y_3) = \sum_1^3 \pi(x_i, y_i)\kappa(y_i)$$

on $]0,1[\times]0,1[\times]0,1[$ assumes its minimal value for the interior point $y = x$ when restricted to probability distributions. As standard regularity conditions are fulfilled, there exists a Lagrange multiplier λ such that

$$\frac{\partial}{\partial y_i}\big(F(y) - \lambda\sum_1^3 y_i\big) = 0 \text{ for } i = 1,2,3$$

when $y = x$. This shows that $f(x_1) = f(x_2) = f(x_3)$.

Using this with $(x_1, x_2, x_3) = (\frac{1}{2}, x, \frac{1}{2} - x)$ for a value of x in $]0, \frac{1}{2}[$, we conclude that f is constant on $]0, \frac{1}{2}]$. Then consider a value $x \in]\frac{1}{2}, 1[$ and the probability vector $(x, \frac{1}{2}(1-x), \frac{1}{2}(1-x))$ and conclude from the first part of the proof that $f(x) = f(\frac{1}{2}(1-x))$. As $0 < \frac{1}{2}(1-x) < \frac{1}{2}$, we conclude that $f(x) = f(\frac{1}{2}))$. Thus f is constant on $]0,1[$. By letting $t \to 1$ in (107) and appealing to the technical boundedness assumption, we conclude that the value of the constant is -1. □

We can now formulate the main theorem on the worlds Ω_π pertaining to situations involving discrete probability distributions. Motivated by Lemma 2 we find it justified only to study the worlds Ω_q [18].

---

[18] there may, however, be interesting worlds to study if we restrict attention only to distributions over a two-element set.

Theorem 10. *If $q \leq 0$, there is no descriptor which, together with π_q, generates a proper effort function.*

If $q > 0$ there exists a unique descriptor κ_q such that π_q and κ_q generates a proper effort function. This descriptor is given by

$$\kappa_q(s) = \ln_q \frac{1}{s} \tag{108}$$

and the information triple determined by the generated effort function is the Tsallis triple $(\Phi_q, \mathrm{H}_q, \mathrm{D}_q)$. The function $\delta = \delta_{\pi_q,\kappa_q}$ from (105) satisfies the pointwise fundamental inequality and is identical to the pointwise divergence function d_q from (78).

Proof. By Lemma 4 we see that κ_q given by (108) is the only descriptor which, together with π_q, could possibly generate a proper effort function. For $q < 0$, this is not the case as the reader can easily verify by considering atomic situations with $x = (1-\varepsilon, \varepsilon)$ and $y = (\frac{1}{2}, \frac{1}{2})$ and letting ε tend to 0.

For $q > 0$ one finds that π_q and κ_q generate the previously studied proper effort function Φ_q obtained by integration of ϕ_q from (76). A simple calculation shows that δ from (105) is identical to the atomic divergence function d_q from (78) which is known to satisfy the pointwise fundamental inequality. □

We note that for the degenerate case $q = 0$, a black hole, $\Phi_0(x, y) = \mathrm{H}_0(x) = n - 1$ with n the size of the number of basic events in the support of x. As noted before, divergence vanishes identically in this case.

Thinking more over the reasoning which led to the main result, we realize, from the proof of Lemma 3, that perhaps the definition (104) is not the most natural one. It appears sensible to replace it by the definition of *gross description effort* given by (dropping π and κ from the notation in (104))

$$\tilde{\Phi}(x, y) = \sum_{i \in \mathbb{A}} \big(\pi(x_i, y_i)\kappa(y_i) + y_i\big) . \tag{109}$$

Similarly, *gross entropy* is defined by

$$\tilde{\mathrm{H}}(x) = \sum_{i \in \mathbb{A}} \big(x_i\kappa(x_i) + x_i\big) . \tag{110}$$

The added terms are interpreted as an *overhead* related to the handling of the event with index i, and this term is proportional to the believed point probability, either y_i in (109) or x_i in (110). According to a frequential interpretation, these added terms are thus proportional to the occurrence of the event in question. The total added overhead in an atomic situation (x, y) is $\sum y_i = 1$ and in an atomic situation (x, x) it is also $\sum_i x_i = 1$. If we allow incomplete distributions as belief instances, the overhead be less than 1 in the first case, as is only natural. We may say that the normalization (103) corresponds to choosing the overhead cost as the unit to work with. This makes good sense in the Shannon world since, apart from the necessary adjustment from nats to bits, the overhead in that case corresponds to taking the cost of having access to a binary memory cell as the basic unit.

Note that gross entropy is always bounded below by the overhead cost, 1 nat.

We have noted that the descriptor is uniquely determined from the interactor. Therefore, in principle, only the interactor needs to be known. Examples will show that quite different interactors may well determine the same descriptor. Thus, knowing only the descriptor, you cannot know which world you operate in, in particular, you cannot determine divergence or description effort. But you *can* determine

the entropy function. This emphasizes again the general thesis, that *entropy should never be considered alone*. Experience says that even when entropy can be considered by itself in interesting connections full understanding and easy technical handling is always accomplished by introducing further basic quantities, typically description effort.

It is instructive to consider the family $(\kappa_q)_{0\leq q<\infty}$ of descriptors. This is a descending family of decreasing functions on $[0,1]$. The largest descriptor, $\kappa_0(x) = \frac{1}{x} - 1$, is associated with a black hole. For $0 \leq q \leq 1$, the descriptors are convex and assume the value ∞ for $x = 0$. For $q = 1$, we find the descriptor $\kappa_1(x) = \ln \frac{1}{x}$ associated with the classical world. Then, for $1 < q < 2$ the descriptors are convex and finite valued, also for $x = 0$. The special descriptor $\kappa_2(x) = 1 - x$ is affine. For $2 < q < \infty$ we find descriptors which are concave with $\kappa_q'(0) = 0$. The zero function is not a descriptor covered by Theorem 1. It may be conceived as a limiting case corresponding to $q = \infty$. A world corresponding to this value of q would lead to situations with no outstanding issues, a world of wisdom (*paradise* or *hell* according to personal taste).

22. conclusions

An abstract theory of basic elements of cognition is initiated with an emphasiz on interpretations. Justification of the theory lies both in the philosophically motivated considerations per se and in the wide range of applications.

23. Acknowledgements The research presented spans a number of years, and is actually rooted in [25] from 1979. However, the realization that the methods applied has a much wider applicability than originally intended only matured slowly from around 2003. The author is thankful to organizers of workshops and conferences where he has presented aspects of the ideas. Acknowledgement goes to Peter Harremoës for several discussions of the subject matter. Finally, a stipend from the San Cataldo Foundation, December 2012, allowed the author to start collecting the material in a more comprehensive and coherent form.

References

1. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423 and 623–656.
2. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Physics* **1988**, *52*, 479–487.
3. Tsallis, C. *Introduction to Nonextensive Statistical Mechanics*; Springer: Berlin Heidelberg, 2009.
4. Gross, D. Comment on: "Nonextensivity: from low-dimensional maps to Hamiltonian systems" by Tsallis et al. arXiv:0210448[cond-mat.stat-mech], 2002.
5. Shalizi, C.R. Tsallis Statistics, Statistical Mechanics for Non-extensive Systems and Long-Range Interactions. Technical report, 2007. Informal notes from the authors homepage.
6. Ingarden, R.S.; Urbanik, K. Information without probability. *Colloq. Math.* **1962**, *9*, 131–150.

7. Kolmogorov, A.N. Logical basis for information theory and probability theory. *IEEE Trans. Inform. Theory* **1968**, *14*, 662–664.
8. Kolmogorov, A.N. Combinatorial foundations of information theory and the calculus of probabilities. *Russian Mathematical Surveys* **1983**, *38*, 29–40. (from text prepared for the International Congress of Mathematicians, 1970, Nice).
9. de Fériet, K. La theorie génerélisée de l'information et la mesure subjective de l'information. In *Théories de l'information (Colloq. Iiformation et Questionnaires, Marseille-Luminy, 1973*; Springer: Berlin, 1974; pp. 1–35.
10. Jaynes, E.T. *Probability Theory - The Logic of Science*; Cambridge University Press: Cambridge, 2003.
11. Rathmanner, S.; Hutter, M. A Philosophical Treatise of Universal Induction. *Entropy* **2011**, *13*, 1076–1136.
12. Barron, A.; Rissanen, J.; Yu, B. The Minimum Description Length Principle in Coding and Modeling. *IEEE Trans. Inform. Theory* **1998**, *44*, 2743–2760.
13. Grünwald, P.D. *the Minimum Description Length principle*; MIT Press: Cambridge, Massachusetts, 2007.
14. Jumarie, G. *Maximum Entropy, Information Without Probability and Complex Fractals – Classical and Quantum Approach*; Kluwer: Dordrecht, 2000.
15. Shafer, G.; Vovk, V. *Probability and finance. It's only a game!*; Wiley: Chichester, 2001.
16. Gernert, D. Pragmatic Information: Historical Exposition and General Overview. *Mind and Matter* **2006**, *4*, 141–167.
17. Bundesen, C.; Habekost, T. *Principles of Visual Attention*; Oxford University Press: Oxford, 2008.
18. Benedetti, F. *Placebo effects. Understanding the mechanisms in health and disease*; Oxford University Press: Oxford, 2009.
19. Brier, S. Cybersemiotics: An Evolutionary World View Going Beyond Entropy and Information into the Question of Meaning. *Entropy* **2010**, *12*, 1902–1920.
20. van Benthem, J.; Adriaans, P., Eds. *Handbook on the Philosophy of Information*; Vol. 8, *Handbook of the Philosophy of Science*, Elsivier, 2007.
21. Adriaans, P. Information. In *Stanford Encyclopedia of Philosophy*; 2012; p. 43 pages. available from http://plato.stanford.edu/entries/information/.
22. Brier, S. *Cybersemiotics: Why information is not enough*; Toronto University Press, 2008.
23. Topsøe, F. Game Theoretical Equilibrium, Maximum Entropy and Minimum Information Discrimination. In *Maximum Entropy and Bayesian Methods*; Mohammad-Djafari, A.; Demoments, G., Eds.; Kluwer Academic Publishers: Dordrecht, Boston, London, 1993; pp. 15–23.
24. Pfaffelhuber, E. Minimax Information Gain and Minimum Discrimination Principle. Topics in Information Theory; Csiszár, I.; Elias, P., Eds. János Bolyai Mathematical Society and North-Holland, 1977, Vol. 16, *Colloquia Mathematica Societatis János Bolyai*, pp. 493–519.
25. Topsøe, F. Information Theoretical Optimization Techniques. *Kybernetika* **1979**, *15*, 8 – 27.
26. Harremoës, P.; Topsøe, F. Maximum Entropy Fundamentals. *Entropy* **2001**, *3*, 191–226.
27. Grünwald, P.D.; Dawid, A.P. Game Theory, Maximum Entropy, Minimum Discrepancy, and Robust Bayesian Decision Theory. *Annals of Mathematical Statistics* **2004**, *32*, 1367–1433.

28. C. Friedman, J.H.; Sandow, S. A Utility-Based Approach to Some Information Measures. *Entropy* **2007**, *9(1)*, 1–26.
29. Dayi, H. Game Analyzing based on Strategic Entropy. *Chinese Journal of Management Science* **2009**, *17*, 133–138. (chinese).
30. Harremoës, P.; Topsøe, F. The Quantitative Theory of Information. In *Handbook on the Philosophy of Information*; van Benthem, J.; Adriaans, P., Eds.; Elsivier, 2008; Vol. 8, *Handbook of the Philosophy of Science*, pp. 171–216.
31. Aubin, J.P. *Optima and equilibria. An introduction to nonlinear analysis*; Springer: Berlin, 1993.
32. Cesa-Bianchi, N.; Lugosi, G. *Prediction, learning and games*; Cambridge University Press: Cambridge, 2006.
33. Topsøe. Interaction between Truth and Belief as the key to entropy and other quantities of statistical physics. arXiv:0807.4337v1[math-ph], 2008.
34. Topsøe, F. Truth, Belief and Experience – a route to Information. *Journal of Contemporary Mathematical Analysis – Armen. Acad. Scienc.* **2009**, *44*, 105–110. Dedicated to Klaus Krickeberg on the occasion of his 80th year birthday.
35. Topsøe, F. On truth, belief and knowledge. 2009 IEEE International Symposium on Information Theory; IEEE, , 2009; pp. 139–143.
36. Topsøe, F. Towards operational interpretations of generalized entropies. In *Journal of Physics: Conference Series*; H. Suyari, A. Ohara, T.W., Ed.; 2010; Vol. 201, p. 15 pages.
37. Wikipedia. Bayesian probability — Wikipedia, The Free Encyklopedia, 2009. [Online; accessed 31-January-2011].
38. Good, I.J. Rationel Decisions. *J. Royal Statist. Soc., Series B* **1952**, *14*, 107–114.
39. Csiszár, I. Axiomatic Characterizations of Information Measures. *Entropy* **2008**, *10*, 261–273.
40. Brier, G.W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **1950**, *78*, 1–3.
41. Savage, L.J. Elicitation of Personal Probabilities and Expectations. *Journal of the American Statistical Association* **1971**, *66*, 783–801.
42. Fischer, P. On the Inequality $\sum p_i f(p_i) \geq \sum p_i f(q_i)$. *Metrika* **1972**, pp. 199–208.
43. Gneiting, T.; Raftery, A.E. Strictly Proper Scoring Rrules, Prediction, and Estimation. *Journal of the American Statistical Association* **2007**, *102*, 359–378.
44. Dawid, P. The geometry of proper scoring rules. *Annal of the Institute of Statistical Mathematics* **2007**, *59*, 77–93. doi:10.1007/s10463-006-0099-8.
45. Dawid, A.P.; Musio, M. Theory and Applications of Proper Scoring Rules. arXiv:1401.0398[math.ST], 2014.
46. A. Philip Dawid, M.M.; Ventura, L. Minimum Scoring Rule Inference. arXiv:1403.3920[math.ST], 2014.
47. Caticha, A. Information and Entropy. Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 27th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Emginering; et al., K.H.K., Ed., 2007, Vol. 954, *AIP Conference Proceedings*, pp. 11–22.
48. Kerridge, D.F. Inaccuracy and inference. *J. Roy. Stat. Soc. B.* **1961**, *23*, 184–194.

49. Kullback, S. *Information Theory and Statistics*; Wiley: New York, 1959.
50. Topsøe, F. Game theoretical optimization inspired by information theory. *J. Global Optim.* **2009**, *43*, 553–564.
51. von Neumann, J. Zur Theorie der Gesellschaftsspiele. *Math. Ann.* **1928**, *100*, 295–320.
52. von Neumann, J. Uber ein okonomische Gleichungssystem und eine Veralgemeinerung des Brouwerschen Fixpunktsatzes. *Ergebnisse eines Mathematisches Kolloquims. Wien* **1937**, *8*, 73–83. Translation in *Review of Economic studies*, vol.13, pp. 1–9, 1945.
53. Kjeldsen, T.H. John von Neumann's Conception of the Minimax Theorem: A Journey Through Different Mathematical Contexts. *Arch. Hist. Exact Sci.* **2001**, pp. 39–68.
54. Čencov, N.N. *Statistical Decision Rules and Optimal Inference.*; Nauka: Moscow, 1972. In russian, translation in "Translations of Mathematical Monographs", 53.AmericanMathematical Society, 1982.
55. Csiszár, I. I-Divergence Geometry of Probability Distributions and Minimization Problems. *Ann. Probab.* **1975**, *3*, 146–158.
56. Csiszár, I.; Matús, F. Information projections revisited. *IEEE Trans. Inform. Theory* **2003**, *49*, 1474–1490.
57. Csiszár, I.; Matús, F. Generalized minimizers of convex integral functionals, Bregman distance, Pythagorean identities. *Kybernetika* **2012**, *48*, 637–689.
58. Glonti, O.; Harremoës, P.; Khechinashili, Z.; Topsøe, F. Nash Equilibrium in a Game of Calibration. *Theory of Probability and its Applications* **2007**, *51*, 415–426.
59. Topsøe, F. Basic Concepts, Identities and Inequalities – the Toolkit of InformationTheory. *Entropy* **2001**, *3*, 162–190. http://www.unibas.ch/mdpi/entropy/ [ONLINE].
60. Weijs, S.V.; van de Giesen, N. Accounting for Observational Uncertainty in Forecast Verification: An Information-Theoretical View on Forecasts, Observations, and Truth. *Monthly Weather Review* **2011**, *139*, 2156–2162.
61. McCarthy, J. Measures of the Value of Information. *Proc. Nat. Acad. Sci.* **1956**, *42*, 654–655.
62. Chambers, C.P. Proper scoring rules for general decision models. *Games and Economic Behavior* **2008**, *63*, 32–40.
63. Hilden, J. Scoring Rules for Evaluation of Prognosticians and Prognostic Rules. first version 1999, updated 2008, 2008.
64. Bregman, L.M. 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.* **1967**, *7*, 200–217. Translated from Russian.
65. Tsallis, C. What are the numbers that experiments provide? *Quimica Nova* **1994**, *17*, 468.
66. Havrda, J.; Charvát, F. Quantification method of classification processes. Concept of structural$a$-entropy. *Kybernetika* **1967**, *3*, 30–35. Review by I. Csiszár in MR, vol. 34, no.8875.
67. Daróczy, Z. Generalized Information Functions. *Information and Control* **1970**, *16*, 36–51.
68. Lindhard, J.; Nielsen, V. Studies in Statistical Dynamics. *Mat. Fys. Medd. Dan. Vid. Selsk.* **1971**, *38*, 1–42.
69. Lindhard, J. On the Theory of Measurement and its Consequences in Statistical Dynamics. *Mat. Fys. Medd. Dan. Vid. Selsk.* **1974**, *39*, 1–39.

70. Aczél, J.; Daróczy, Z. *On measures of information and their characterizations*; Academic Press: New York, 1975.
71. B. Ebanks, P.S.; Sander, W. *Characterizations of Information Measures*; World Scientific: Singapore, 1998.
72. Jaynes, E.T. Where do we Stand on Maximum Entropy? In *The Maximum Entrropy Formalism*; Levine, R.; Tribus, M., Eds.; M.I.T. Press: Cambridge, MA, 1979; pp. 1–104.
73. Naudts, J. Generalised exponential families and associated entropy functions. *Entropy* **2008**, *10*, 131–149.
74. Sylvester, J.J. A Question in the Geometry of Situation. *Quarterly Journal of Pure and Applied Mathematics* **1857**, *1*, 79.
75. Drezner, Z.; Hamacher, H., Eds. *Facility location. Applications and Theory*; Springer: Berlin, 2002.
76. Kapur, J.N. *Maximum Entropy Models in Science and Engineering*; Wiley: New York, 1993. first edition 1989.
77. van der Lubbe, J.C.A. On certain coding theorems for the information of order α and oftypeβ. Trans. Eighth Prague Conf. on Inform. Theory, Statist. Decision Functions,RandomProcesses; Czech. Acad. Science, Academia Publ.: Prague, 1978. Prague, 1979.
78. Ahlswede, R. Identification Entropy. In *General Theory of Information Transfer and Combinatorics*; et al, A., Ed.; Springer: Berlin, 2006; Vol. 4123, *Lecture Notes in Computer Science*, pp. 595–613.
79. Suyari, H. Tsallis entropy as a lower bound of average description length for the q-generalized code tree. Proceedings ISIT 2007. IEEE, 2007, pp. 901–905.