Deceptive Updating and Minimal Information Methods

Haim Gaifman
Columbia University
hg17@columbia.edu

Anubav Vasudevan
Columbia University
av2137@columbia.edu

Abstract

The technique of minimizing information (infomin) has been commonly employed as a general method for both choosing and updating a subjective probability function. We argue that, in a wide class of cases, the use of infomin methods fails to cohere with our standard conception of rational degrees of belief. We introduce the notion of a *deceptive* updating method, and argue that non-deceptiveness is a necessary condition for rational coherence. Infomin has been criticized on the grounds that there are no higher order probabilities that 'support' it, but the appeal to higher order probabilities is a substantial assumption that some might reject. The elementary arguments from deceptiveness do not rely on this assumption. While deceptiveness implies lack of higher order support, the converse does not, in general, hold, which indicates that deceptiveness is a more objectionable property. We offer a new proof of the claim that infomin updating of any strictly-positive prior with respect to conditional-probability constraints is deceptive. In the case of expected-value constraints, infomin updating of the uniform prior is deceptive for some random variables, but not for others. We establish both a necessary condition and a sufficient condition (which extends the scope of the phenomenon beyond cases previously considered) for deceptiveness in this setting. Along the way, we clarify the relation which obtains between the strong notion of higher order support, in which the higher order probability is defined over the full space of first order probabilities, and the apparently weaker notion, in which it is defined over some smaller parameter space. We show that under certain natural assumptions, the two are equivalent. Finally, we offer an interpretation of Jaynes, according to which his own appeal to infomin methods avoids the incoherencies discussed in this paper.

1. Introduction

Let Pr be a subjective probability function defined over a finite Boolean algebra, $\mathcal{B}$. An updating method U is a rule which tells an agent how to modify Pr so as

to satisfy a given condition or *constraint*. We express constraints in terms of the value of a parameter λ, and we consider families of constraints $\{C_\lambda\}_{\lambda\in\Lambda}$, where Λ is the set of possible values of λ. For each $\lambda \in \Lambda$, the result of applying U under the constraint C_λ is an updated probability $U(Pr; C_\lambda)$, which satisfies this constraint. We shall denote the updated probability by 'Pr_λ' whenever U and Pr can be inferred from the context. It is assumed throughout that if Pr satisfies the constaint C_λ, then the updating causes no change, i.e., $Pr_\lambda = Pr$. For the time being, we may view an updating method as an abstract mathematical entity. Below we will make explicit the assumptions that are required in order to grant to an updating method normative significance.

With no loss of generality, we assume that $\mathcal{B}$ is the algebra of all subsets of Ω, where Ω is some finite set. If $A \subseteq \Omega$, then we write '$\overline{A}$' for the complement of A and '$|A|$' for the cardinality of A.

The following list provides a few examples of the sorts of constraints to which updating methods are typically applied (constraints are expressed using the schematic letter 'P' to refer to probabilities defined over $\mathcal{B}$):

(i) For an event $A \in \mathcal{B}$, such that $0 < Pr(A) < 1$, the truth-value of A is given, i.e., the agent is informed whether A or its complement is the case. The constraints are $P(A) = \lambda$, where $\lambda \in \{0, 1\}$.

The accepted method for updating under such constraints is conditionalization. Let $A^1 = A$ and $A^0 = \overline{A}$, then $Pr_\lambda = Pr(\ \ |A^\lambda)$. This is the simplest and least problematic of all updating methods.[1]

(ii) For an event $A \in \mathcal{B}$, such that $0 < Pr(A) < 1$, the probability of A is given (this is a straightforward generalization of (i)). The constraints are $P(A) = \lambda$, where $\lambda \in [0, 1]$.

A well-known method of updating under such constraints is Jeffrey conditionalization. For all $B \in \mathcal{B}$:

$$Pr_\lambda(B) = \sum_{\omega\in B\cap A} Pr(\omega|A)\lambda + \sum_{\omega\in B\cap\overline{A}} Pr(\omega|\overline{A})(1-\lambda)$$

Jeffrey conditionalization also applies to the more general case of updating under constraints of the form $P(A_i) = \lambda_i$, for $i = 1, \ldots, m$, where $\{A_i\}_i$ is a partition of Ω and the λ_i's are non-negative reals whose sum is 1.

(iii) For a random variable X over $\mathcal{B}$, the expected value of X is given. The

[1] While our framing of the issue represents the truth of an event in $\mathcal{B}$ as a constraint and conditionalization as a particular updating method, this may not correctly reflect Jaynes's view of the matter (see §7).

constraints are $E(X) = \lambda$, where:

$$E(X) = \sum_{\omega \in \Omega} P(\omega)X(\omega)$$

and where λ ranges over some interval determined by X.

(iv) The constraints are $P(B|A) = \lambda$, where $\lambda \in [0, 1]$. Here A and B are fixed events, and $Pr(A) > 0$.

We shall refer to constraints of the form (iii) as expected-value constraints, and constraints of the form (iv) as conditional-probability constraints. All of the above mentioned constraints are special cases of *linear* constraints, that is, constraints which can be expressed as one or more linear equations in the expected values of certain random variables. For example, a constraint of the form $P(B|A) = \lambda$ can be expressed:

$$-\lambda E_P(X) + E_P(Y) = 0$$

where X and Y are the characteristic functions of A and $A \cap B$, respectively. In general, λ can be a vector $(\lambda_1, \ldots, \lambda_n)$, where the λ_i's appear as coefficients in the equation.[2]

In the last fifty years or so, a great deal of attention has been paid to the general problem of how to update under linear constraints. By far, the most widely considered method is the technique of minimizing information, or what we shall call 'infomin'.

The first to apply infomin methods to the assignment of subjective probabilities was Jaynes, who in (1957) advocated the principle of maximum entropy as a general rule by which to determine the uniquely rational prior probability satisfying a system of expected-value constraints.[3] The principle recommends that one choose, among all the probabilities satisfying the constraint, that which minimizes the Shannon information[4] (or, equivalently, maximizes the Shannon entropy):

$$S(P) = \sum_{\omega \in \Omega} P(\omega) \log P(\omega)$$

Since in the absence of any constraints, the uniform probability minimizes information, later authors construed Jaynes's principle as providing a method

[2]Since all linear constraints are linear equations in the expected value of random variables, we should make it clear that by 'expected-value' constraints, we mean the special case, where the constraint is of the form $E(X) = \lambda$.

[3]Historically, the use of maximum entropy methods in physics dates back to the work of Maxwell and Boltzmann, and is most famously exemplified in the statistical mechanical derivation of the probability distribution of the velocities of the particles in a gas at thermal equilibrium. However, it was only in Jaynes (1957) that the technique was first put forward as a method for selecting a subjective prior probability.

[4]Shannon (1948).

for updating the uniform prior given new information in the form of a constraint. The technique was later generalized to an updating method that can applied to non-uniform priors, by appeal to a measure of relative-information (or information gain) first introduced in Kullback & Leibler (1951). The general prescription in this case is to choose the probability function satisfying the constraint, which minimizes the Kullback-Leibler (KL) divergence:

$$D_{\mathrm{KL}}(Pr, P) = \sum_{\omega \in \Omega} P(\omega) \log \left(\frac{P(\omega)}{Pr(\omega)} \right),$$

where Pr is the agent's prior. Infomin updating is a general method which can be applied to arbitrary linear constraints. When applied to constraints of the form (ii), it reduces to Jeffrey conditionalization (and *a fortiori* when applied to constraints of the form (i), to Bayesian conditionalization). The use of infomin updating in the context of both expected-value constraints and conditional-probability constraints has come under attack from Bayesians, who have pointed out that it is incompatible with the methodology of applying Bayesian conditionalization to higher order probabilities (Friedman & Shimony (1971), Shimony (1973), Seidenfeld (1987)).

The criticisms of Shimony et al. have traditionally been expressed in terms of a conflict between infomin methodologies and Bayesianism.[5] We have chosen to avoid this characterization, on the grounds that Bayesianism itself is not a precisely defined view,[6] and so, the question of whether and to what extent infomin comports with "Bayesianism" is largely a matter of semantics.[7] On the

[5] "... [T]he anomaly that has been presented is almost a demonstration that PME [read: infomin] is inconsistent with Bayesian probability theory." (Shimony 1985, p.41)

[6] See, e.g., Good (1972)

[7] A Bayesian position can perhaps be minimally characterized as one which subscribes to a methodology that appeals to prior probabilities, which are then updated via conditionalization, but the details of the view then depend upon how rich a field of events these probabilities are defined over. The foundational aspiration, which lay at the bottom of Carnap's project, of deriving all probabilities from a single, all-encompassing prior, has been shown to be untenable in a way which points to the essential limitations of any theory of inductive inference based solely on the updating of prior probabilities. This point, which was first established in Putnam (1963) and further developed in Gaifman & Snir (1982), is often underappreciated. A prior probability is a mathematically defined function, which can itself be used to construct a hypothesis with respect to which the prior behaves badly. The construction of such confounding hypotheses uses a diagonalization technique analogous to that used in the proof of Gödel's incompleteness results. In Putnam (1963), a prior of the type considered by Carnap is used to define a satisfiable universal hypothesis such that given any finite data that confirms the hypothesis, the conditional probability for the next case is $\leq 1/2$. The problem is not that of conditionalizing on events of probability 0 (such conditionalization can be handled in models based on a two-argument conditional probability function). It is rather that, due to epistemic limitations, we cannot construct a probability that behaves as desired with respect to all hypotheses that might arise in the course of an inquiry. Suppose that what appears as an initial segment of a random sequence displays previously unnoticed correlations, on the basis of which, one would like to assign probability 0.8 to some universal hypothesis whose prior probability (given one's background knowledge) is 0. This abrupt change cannot be viewed as an act of conditionalization – no matter whether one uses a one-place or two-place

other hand, what it means for an updating method to be "supported" within a framework of higher order probabilities can be given a precise characterization. Before we proceed to do so, a few remarks are in order concerning higher order probabilities.

First order probabilities are probabilities defined over the Boolean algebra $\mathcal{B}$. A *higher order* probability is a probability over first order probabilities. Such a probability assigns values to certain higher order events, construed as sets of first order probabilities. If $\mathcal{B}$ has n atoms, then every first order probability determines and is determined by a point, $\overline{p} = (p_1, \ldots, p_n) \in \mathbb{R}^n$, where p_i is the probability of the i^{th} atom. The space of first order probabilities can thus be identified with the $(n-1)$ dimensional simplex Δ consisting of all points $\overline{p}$ with non-negative coordinates whose sum is 1. A higher order probability is any probability function that assigns values to some subsets of Δ.

Higher order probabilities can be interpreted in various ways. The standard interpretation is as a subjective probability over some parameter space, where the parameter determines the objective chances of events in $\mathcal{B}$. This sort of interpretation is exemplified in such straightforward statistical examples as tossing a coin with unknown bias, or drawing at random from an urn containing several kinds of objects in unknown proportions. Alternatively, higher order probabilities can be interpreted as subjective probabilities over other subjective probabilities. As was first pointed out in Savage (1954), a (non-trivial) higher order probability and the first order probabilities over which it is defined cannot represent the degrees of belief of a single agent at a given time, but there remain other possibilities. For example, a higher order probability can be interpreted as an agent's subjective probability over his own beliefs at some later time (at which he occupies an improved epistemic state); or they can represent the agent's beliefs concerning the subjective probabilities of an expert agent (Gaifman 1986). However this might be, the crucial feature of a higher order probability is that the agent's subjective probabilities over $\mathcal{B}$ should be obtained by integration of the first order values with respect to the higher order measure.

It is often natural to define higher order probabilities not on Δ itself, but on some other more restricted space. For example, given a family of constraints $\{C_\lambda\}_{\lambda\in\Lambda}$, one may consider a higher order probability μ over a σ-field of subsets of Λ.[8] In this case, $\mu(\Theta)$ is the probability that the constraint satisfied by P belongs to the set $\{C_\lambda\}_{\lambda\in\Theta}$.

If an agent possesses higher order probabilities over Λ, then clearly the agent's current probabilities should match his expected posterior probabilities, i.e., for

function – unless the probability was designed to take account of this possibility. When the prior probability function can itself be used to state hypotheses, this is, in principle, not always possible. The limitation is a "probabilistic kin" of the limitations concerning provability and truth that are due to the incompleteness results. See Gaifman (1983) pp. 338-342, and Gaifman (2004) pp. 115-6 for concrete examples.

[8] In this paper probabilities are, by definition, countably additive.

all $A \in \mathcal{B}$:

$$Pr(A) = \int_{\Lambda} Pr_{\lambda}(A)\ d\mu(\lambda) \tag{1.1}$$

Also, from the higher order point of view, an agent's updating method should not be vacuous, that is, the higher order probability assigned to the event that the updated probability differs from Pr should be non-zero:

$$\mu(\{\lambda : Pr_{\lambda} \neq Pr\}) > 0 \tag{1.2}$$

For a given probability Pr, we will say that μ *supports* the updating method U, and that U is *supported* by μ, if (1.1) and (1.2) hold.

The criticism of infomin developed in Friedman & Shimony (1971) was based on the observation that for certain families of expected-value constraints, any higher order probability μ that satisfies (1.1), violates (1.2). In other words, the updating lacks higher order support. This result was extended in Shimony (1973) so as to apply to all cases of expected-value constraints in which the random variable takes at least three distinct values.

The application of infomin updating to conditional-probability constraints was first considered by van Fraassen in his (1981) and its follow-up. In that paper, van Fraassen construes a scene from the film *Private Benjamin* as an updating problem, in which an agent (Private Benjamin) has a prior probability Pr, such that $Pr(A) = 1/2$ and $Pr(B|A) = 1/2$. Private Benjamin then receives new information requiring that she update her probabilities so as to satisfy the constraint $P(B|A) = 2/3$. Applying the infomin method to this problem, van Fraassen noticed the "glaring feature" that the probability of A decreased as a result of the update. Moreover, the updated probability of A is strictly less than its initial value of $1/2$, whenever the constraint prescribes a conditional probability other than $Pr(B|A)$. In an appendix to Seidenfeld (1987), this phenomenon was shown to be a general one: in any application of infomin updating to conditional-probability constraints, if the updating leads to any change in the agent's prior probabilities, then the updated probability assigns to A a value strictly smaller than $Pr(A)$. From this fact it follows immediately that the updating lacks higher order support, which allows for the line of criticism initiated by Friedman and Shimony to be extended to conditional-probability constraints.

In spite of these objections, a number of systematic attempts have been made to derive the infomin methodology from principles of rationality belonging to the framework of subjective probabilities. One attempt of this kind is in Shore & Johnson (1980), in which infomin updating with respect to expected-value constraints is derived from general axioms alleged to apply to any rational updating method. A more recent attempt, presented in Paris & Vencovská (1997), proposes a proof that infomin yields the rationally mandated prior probability satisfying given linear constraints (the latter are specifically meant to include

conditional-probability constraints).[9]

The above-mentioned arguments against infomin, which rely on its lack of higher order support, presuppose that an updating method should fit within a framework of higher order probabilities. These objections could thus be avoided by rejection this presupposition. Indeed, the assumption that an agent possesses a higher order probability is a quite substantial one, implying, for instance, that the agent has subjective probabilities over such events as that "the (first order) probability yields an expected value of X in the range (α, β)." In their (1971) paper, Friedman and Shimony acknowledged this point and suggested that a promising way of responding to their criticism would be to simply reject the framework of higher order probabilities.[10] Jaynes, in his brief reply to Friedman and Shimony, seemed to endorse this response.[11]

A main goal of this paper is to avoid any recourse to higher order probabilities by finding fault with infomin updating on more elementary grounds. To this end, we focus on the property first discovered in the context of van Fraassen's *Judy Benjamin* problem, which we call *deceptiveness*:

Definition 1.1. An updating method U is *deceptive* for a given probability Pr and for a family of constraints $\{C_\lambda\}_{\lambda\in\Lambda}$, if there exists an event $A \in \mathcal{B}$, such that for all $\lambda \in \Lambda$, either $Pr_\lambda(A) < Pr(A)$ or $Pr_\lambda = Pr$.

Deceptiveness is a very simple property, defined without any appeal to higher order probabilities. As we shall argue, it implies that the use of the method for updating subjective probabilities is incoherent. It also (trivially) implies that the method lacks higher order support. As we will see, however, the reverse implication does not hold. Hence, as a property of an updating method, deceptiveness is considerably worse than lack of higher order support.

Though our arguments make no appeal to higher order probabilities, they do presuppose a certain minimal reflective capacity on the part of the agent. In particular, the agent must be able to recognize that the updating method is

[9] An informal presentation of the argument is given in Paris (1998)

[10] "[Another response] is to deny that the probabilities $F(\hat{d_\varepsilon}|b)$ are capable of being well-defined, even though each $\hat{d_\varepsilon}$ is well-defined. A defense along these lines seems promising to us. However, to make it convincing one would need criteria for deciding when a proposition can and when it cannot be assigned a reasonable degree of belief on given evidence, which in turn presupposes a deep and systematic analysis of the concept of reasonable degree of belief." (p. 384). Here, $\hat{d_\varepsilon}$ is a certain higher order event, and $F(\hat{d_\varepsilon}|b)$ its probability, given the background knowledge b.

[11] "[Friedman and Shimony] suggest that a possible way of resolving all this is to deny that the probability of $\hat{d_\varepsilon}$ can be well-defined. Of course it cannot be; however, to understand the situation we need no 'deep and systematic analysis of the concept of reasonable degree of belief.' We need only raise our standards of exposition to the same level that is required in any other application of probability theory; i.e., we must define our propositions and sample space with enough precision to make a determinate mathematical problem."(Jaynes 1983, p. 41)

deceptive, and he must acknowledge (in some qualitative sense) the possibility of acquiring new information which will require a change in his current beliefs.

A similar criticism to that which is based on deceptiveness can be addressed to certain applications of infomin in the context of choosing rather than updating a prior. A failure to properly account for this distinction has, in the past, led to misplaced criticisms. However, even when infomin is utilized for the purpose of choosing a prior, the constraints to which it is applied cannot be arbitrary. If, for example, we choose a prior under the constraint $E(X) = \lambda$, the random variable X must have special significance relative to the agent's background knowledge, if the selected prior is to be coherently interpreted as the agent's degrees of belief.

We return briefly to the issue of higher order support in section 4, where we clarify, in an abstract general setting, the relation which obtains between a full-fledged probability defined over the simplex Δ and a more restricted probability defined over Λ, induced by the first. We show that under certain natural continuity assumptions, which hold for all updating methods considered in the literature, a higher order support defined on Λ can be "lifted" into a support over Δ, in which all the measure is concentrated on a homeomorphic copy of Λ inside Δ. This shows that the condition of full higher order support on Δ is not stronger than the condition of higher order support on Λ.

The new technical results of the paper (discussed in sections 5 and 6) are the following: In the case of conditional-probabilty constraints, we provide an alternative proof of the deceptiveness result first established by Seidenfeld, which gives new insight into the underlying 'cause' of the phenomenon. Our proof, which relies on general features of the way in which information is measured, applies not only to infomin updating, but to other information-based methods as well.

With regard to expected-value constraints, we establish both a necessary condition and a sufficient condition (on the random variable X) for the deceptiveness of infomin updating of the uniform prior. The sufficiency condition extends the scope of the phenomenon beyond those cases previously considered. The necessity condition, combined with the general result of Shimony (1973), implies that there are many cases in which the updating is not deceptive but still lacks higher order support (a fact whose significance was noted above). There remains a gap between the sufficient and the necessary conditions.

The last section of the paper is devoted to a brief discussion of Jaynes's own intepretation of the infomin methodology. This is an intricate subject, the difficulty of which is not least of all due to Jaynes's brusque and sometimes dismissive style. Still, we think that Jaynes's position has interesting foundational implications, and, if properly construed, manages to avoid the incoherence arguments developed in this paper.

2. Deceptiveness and updating a subjective probability

Mathematically, an updating method is a function that chooses, for a given probability function and a constraint, another probability function that satisfies the constraint. The implementation of the method involves a diachronic step: the agent who is informed of the constraint replaces the initial (subjective) probability by the updated one. The extent to which such a shift can be mandated by rational norms has been a subject of philosophical debate. This is a foundational issue deserving a separate treatment, but some clarifications are due concerning the presuppositions of the updating step.

The question of why a given constraint should be enforced, i.e., why the probability should be modified so as to satisfy C_λ, is a question that has different answers depending on the nature of the case. Suppose Pr is the agent's probability function defined over the possible outcomes of a random drawing from an urn containing objects of various kinds in unknown proportions. Any information about these proportions is a constraint that should be enforced. In this setting, the proportions play the role of the "true" or objective probabilities. Had the agent known them, they would constitute his degrees of belief. Arguably, the very notion of subjective probabilities requires an appeal to some such model.[12] It appears quite compelling that if the objective probability is known to satisfy C_λ, then the agent should switch to a probability that satisfies that constraint provided that no other relevant information has been added.

In other cases, the constraint can represent the opinion of an expert, or some authority to whom the agent defers. The Private Benjamin scenario mentioned in the introduction is such a case. The constraint can also issue from the agent himself, as in the example suggested by Jeffrey (1965, c. 11) of an agent who decides, after perceiving a color in dim light, that the probability that the color is blue is 0.7.

The updating implicitly presupposes that no other relevant information has been obtained. In practice, learning that a certain constraint holds always involves the acquisition of some additional information; e.g., one learns of the constraint via e-mail, hence one learns that such and such an e-mail was sent. At some point, such additional items must be held to be irrelevant, otherwise there will be an infinite regress. It is also assumed that nothing has happened that might provide the agent with a good reason to abandon the updating method. Such *ceteris paribus* clauses accompany any practical application of a theory. To allay any worries on this front, we may regard the updating as a *default* rule: the agent should update unless he can provide a good reason why the constraint should not be satisfied. *How* he should update is, of course, a separate question

[12] The definition of subjective probabilities in terms of utilities or preferences relies on very strong axioms, whose justification requires an appeal to fair lotteries, that is, to the urn model. This is true of Savage's system, not to speak of the axioms presupposed by Ramsey. We suspect that one cannot advance beyond partially ordered qualitative probabilities, without the fair lottery assumption.

(it is with this latter question that the present paper is concerned).

Let us consider again the deceptiveness property defined above. We assume throughout a fixed updating method U, an initial probability Pr, and a constraint family $\{C_\lambda\}_{\lambda\in\Lambda}$. It is convenient to introduce the following related property of events.

Definition 2.1. An *always decreasing* event A is an event such that, for all $\lambda \in \Lambda$, either $Pr_\lambda = Pr$ or $Pr_\lambda(A) < Pr(A)$. An *always increasing* event is defined similarly.

Obviously, A is always decreasing iff $\overline{A}$ is always increasing. An updating method is deceptive, with respect to the constraint family $\{C_\lambda\}_{\lambda\in\Lambda}$, if there exists an always decreasing event (or, equivalently, there exists an always increasing event).

Thus far, deceptiveness has been viewed as a mathematical property of an updating method. To see that there is something incoherent about employing a deceptive method (so that the term "deceptive" is justified), consider a rational agent, Ann, whose current subjective probability function is Pr, who employs a deceptive method U.

Suppose that A is always decreasing. This is a mathematical feature of U, and we may assume that Ann is already aware of this fact (if not, we can show her the proof). Now, λ is a parameter that does not depend on what Ann believes or chooses to do. In fact, we may assume that its value has already been determined and is written on a piece of paper, so that all Ann has to do is to take a look. Prior to her doing so, Ann already knows that no matter what value is written on the paper, either her posterior degree of belief in A will be $< Pr(A)$ or there will be no change in her subjective probabilities. Let $\Lambda^<$ be the set of all λ for which the first of these alternatives obtains, i.e.:

$$\Lambda^< = \{\lambda \in \Lambda : Pr_\lambda \neq Pr\}$$

If Ann thinks that the possibility that $\lambda \in \Lambda^<$ is not to be ignored, then her current degree of belief in A must be $< Pr(A)$. Hence, Pr is not her subjective probability function. On the other hand, if Ann thinks that the possibility can be ignored, then there is no point in her looking at the paper, since she is already certain that the information it contains will require no revision of her current beliefs. To make this more concrete, assume that Ann is asked to pay something for the piece of paper. If she is willing to pay any small enough amount, then her current degree of belief in A must be $< Pr(A)$. If, on the other hand, Ann refuses to pay anything for the paper, it must be that she is already certain that the information it contains will not require any updating of her beliefs. In this case 'updating' is a deceptive name.

This incoherence can be also cashed out in terms of betting odds.[13] Ann is aware that her current commitments oblige her to accept the following two bets:

1. A bet staking \$1 on A at odds $(1 - Pr(A))/Pr(A) \ : \ 1$.
2. A bet, whose odds are determined by the presently unknown parameter λ, staking $\$(1-Pr(A))/Pr(A)$ against A at odds $Pr_\lambda(A)/(1-Pr_\lambda(A)) \ : \ 1$.

Ann knows that in accepting these two bets, she cannot, under any circumstances, earn a positive return, but she will lose money if $\lambda \in \Lambda^-$ and A does not occur.

The same reasoning can be applied to the updating of an *imprecise* prior (modeled as a set of 'admissible' probability functions), provided that for every admissible prior, A is always decreasing under U. This is because, for each prior in the set, Ann can never win and she might lose. Here, we assume that the same prior that is used to evaluate the first bet is also used to evaluate the second.[14] This assumption cannot be made, and hence the argument does not go through, in the case of *indeterminate* probabilities.[15]

Note that the argument requires only a minimal meta-reflective capacity on the part of the agent. We have only assumed that Ann is capable of both recognizing that her updating method is deceptive and of acknowledging the possibility that $\lambda \in \Lambda^<$. There is no need to assume that this latter acknowledgement is an expression of any more detailed estimate on Ann's part of how likely it is that the updating will be non-trivial.

Conceivably, Ann could refuse to allow herself to reflect at all on the effects of her coming to know λ, shutting her mind to all such considerations and stubbornly adhering to both her current probabilities and her commitment to U. While such systematic myopia can perhaps figure within a useful practical methodology, we cannot pretend that, in such a case, the probabilities represent an agent's *degrees of belief*. One's actual beliefs are not so insulated from the effects of minimal self-reflection.

The betting argument given above is known as a *diachronic* dutch book argument. Some philosophers have objected to such arguments, claiming that rationality only places coherence constraints on an agent's beliefs at a given time (the objections were first made in the context of Bayesian conditionalization, but they apply equally well to other diachronic updating methods). Though we

[13] We are not endorsing the position that subjective probabilities should be defined in terms of betting odds or within a framework based on utilities or preferences. Nevertheless, we do hold that very simple betting situations can help to sharpen our basic intuitions concerning probabilistic reasoning. They can perhaps also be used in the elicitation of subjective probabilities. The same applies to called-off bets and their relation to conditional probabilities.

[14] The argument also applies if the constraint itself is 'imprecise', i.e., if instead of a single λ, the information consists of a set of λ's, giving rise to an imprecise posterior.

[15] For the distinction between imprecise and indeterminate probabilities, see Levi (1985)

took up the issue briefly at the beginning of the section, it may be worthwhile to elaborate the point. Arguments against appeals to diachronic dutch books generally derive from worries that additional relevant information is somehow smuggled in. Such worries can be dispelled by treating the updating method as a default rule: refusal to update carries with it the burden of proof; one should justify the refusal, by pointing to some relevant information acquired which goes beyond the constraint itself. One cannot rationally refuse to apply the rule simply because a certain amount of time has passed. We can therefore redefine the second bet as a *default* bet: the agent will win or lose according to the odds determined by λ, unless there is adequate justification for not updating according to the method. "Adequate justification" is of course a vague notion, but so are all clauses of the ceteris paribus type that mediate between a theory and its applications. For our purposes, it suffices to acknowledge that in all cases there is a rational obligation to justify one's deviations from the rule.

While a diachronic setup is presupposed by the very subject of this paper, we note that 'updating' can be given a purely synchronic interpretation in terms of betting odds, by using called-off bets. These bets were first suggested by de Finetti as a way of giving operational significance to conditional probabilities.[16] In that case, the synchronic rule is as follows: If the inequality $P(A|D) \leq \alpha$ is true of the agent's degrees of belief, the agent should accept a bet against A with odds $\alpha : 1-\alpha$, conditional on D (that is, the bet is called-off if D does not occur). The generalization of this rule from the case of Bayesian conditionalization to an arbitrary updating method can be stated as follows:

(†) Suppose that an agent, who is about to be informed of the value of λ, has degrees of beliefs given by the probability function Pr. Then, if $\Lambda_1 \subseteq \Lambda$, and if, for all $\lambda \in \Lambda_1$, $Pr_\lambda(A) \leq \alpha$, the agent should accept a bet against A with odds $\alpha : 1 - \alpha$, conditional on the event '$\lambda \in \Lambda_1$'.

Any finite system of bets satisfying (†) should be accepted by the agent. Now suppose, as before, that A is always decreasing. Let $\Lambda^< = \{\lambda : Pr_\lambda \neq Pr\}$ and $\Lambda_0 = \Lambda - \Lambda^<$. Then $Pr_\lambda < Pr(A)$ for all $\lambda \in \Lambda^<$. Assume first that there is an $\varepsilon > 0$, such that for all $\lambda \in \Lambda^<$, $Pr_\lambda(A) \leq Pr(A) - \varepsilon$. Then we can construct a synchronic dutch book, which simulates the diachronic one, consisting of the following three bets:

1. An unconditional bet on A, at odds $(1 - Pr(A))/Pr(A)$: 1.
2. A bet against A at odds $Pr(A)/(1 - (Pr(A))$: 1, conditional on $\lambda \in \Lambda_0$.
3. A bet against A at odds $(Pr(A) - \varepsilon)/(1 - (Pr(A) - \varepsilon))$: 1, conditional on $\lambda \in \Lambda^<$.

[16] See de Finetti (1974).

Ann stakes \$1 on the first bet, and \$ $(1 - Pr(A))/Pr(A)$ on each of the other two. Then, if $\lambda \in \Lambda_0$, the first two bets return nothing, and bet (3) is called-off; and if $\lambda \in \Lambda^<$, bet (2) is called-off, and (just as in the diachronic case) Ann has nothing to gain and she loses money if A does not occur.

If, on the other hand, the values $\{Pr_\lambda(A)\}_{\lambda \in \Lambda^-}$ are not bounded strictly below $Pr(A)$, let $\varepsilon_n = 1/(n+1)$, $n = 0, 1, 2 \ldots$, and consider a partition of $\Lambda^<$ into subsets Λ_n, $n = 1, 2, \ldots$, where:

$$\Lambda_n = \{\lambda : Pr(A) - \varepsilon_{n-1} < Pr_\lambda(A) \leq Pr(A) - \varepsilon_n\}$$

Let Bet_n be the bet against A at odds $(Pr(A) - \varepsilon_n)/(1 - (Pr(A) - \varepsilon_n)) \ : \ 1$, conditional on $\lambda \in \Lambda_n$. A synchronic dutch book can now be constructed, provided that the following rather weak continuity assumption holds:

(‡) Let $D_1, \ldots D_n, \ldots$ be disjoint events and let Bet_n be a bet against A, conditional on D_n, such that for all Bet_n the agent stakes the same fixed amount where the odds are in some fixed interval $[\alpha, \beta]$, $0 < \alpha < \beta < 1$. If for each k, the agent accepts $Bet_1, \ldots, Bet_k$, then the agent should accept all the bets $Bet_1, \ldots, Bet_n, \ldots$.

The synchronic dutch book now consists of bets (1) and (2) and the (countable) system of bets, Bet_n, obtained by replacing in (3), $\Lambda^<$ by Λ_n and ε by ε_n.

The above incoherence arguments are relevant for assessing infomin because in a broad class of cases infomin is deceptive. This follows from results presented in sections 5 and 6. These results concern updating under two kinds of constraints: (i) for a given random variable X, the family of expected-value constraints $\{E(X) = \lambda\}_\lambda$, and (ii) for any two events A and B, such that $Pr(A) > 0$, the family of conditional-probability constraints $\{P(B|A) = \lambda\}_\lambda$.

We indicated already that for conditional-probability constraints, $\{P(B|A) = \lambda\}_\lambda$, A is an always-decreasing event, hence infomin updating is always deceptive (see §5). With regard to expected-value constraints the situation is a bit more complex. Infomin updating of the uniform prior is not always deceptive, but in a broad class of cases it is. The results in section 6 are as follows: If there is no event A, such that the average value of X over A equals the average value of X over Ω, then the updating is non-deceptive. On the other hand, if there is such an event A and if, moreover, the event is 'central' in the sense that $X(A) = X(\Omega) \cap [a', b']$, where $\Lambda = [a, b]$ and $a < a' \leq b' < b$, then A is always decreasing.

3. Choosing a Prior and Shiftiness

There is a simple mathematical fact relating infomin prior selection to infomin updating: for any given constraint, the probability (satisfying the constraint)

that minimizes the Shannon information is equal to the probability that minimizes the information relative to the uniform prior, as measured by the KL divergence.[17] This equivalence has led some to suggest that infomin prior selection should be thought of as a special case of infomin updating. However, this assimilation overlooks the important fact that in the context of prior selection the uniform probability serves merely as technical device, utilized at a stage at which the agent does not yet have a subjective probability function over $\mathcal{B}$.[18]

Consider an agent Abe, who, like Ann, is about to receive information reporting the true value of λ. At this stage, there is no probability function representing Abe's credal state, but he is going to choose $U(Pr_0; C_\lambda)$ as his prior, where Pr_0 is the uniform distribution. Suppose that A is always decreasing under U with respect to Pr_0. Since we may assume that Abe is aware of this fact, he is already in a position to infer that, regardless of the value of λ, the prior he will choose will satisfy the inequality $Pr(A) \leq |A|/n$, where $n = |\Omega|$. Moreover, he knows that this inequality will be strict unless $Pr = Pr_0$. Note, however, that in this case Abe's knowing this fact does not lead to incoherence of the sort described in the previous section, since there can be no conflict with his current subjective probability, as the latter does not exist. As we will see, however, Abe's situation is not much better.

Suppose that the prior to be chosen is defined over events relating to a single random drawing from a collection of objects of various kinds, and that the constraint consists of partial information concerning the relative frequencies of the kinds in the collection. [19] Specifically, imagine a bag containing a large number of apples and pears of two kinds: expensive and inexpensive. Let A be the event that the next drawn object will be an apple and let B be the event that it will be an expensive apple. The constraint $P(B|A) = \lambda$ means that the relative frequency of expensive apples among apples is λ, because, under the assumptions of our scenario, this is the true value of $P(B|A)$ (or, if you like, it is the ratio of objective chances).[20] Now, suppose that A is always decreasing under U with respect to Pr_0 and the family of constraints $\{P(B|A) = \lambda\}_\lambda$. Abe thus finds himself in a situation where the mere knowledge that he will be informed of the relative frequency of expensive apples among apples implies a non-trivial bound on his probability that the next fruit drawn will be an apple. In fact, on the basis of this knowledge alone he is led to prefer a bet on a pear to a bet on an apple.

This remarkable outcome is so counterintuitive as to cast serious doubt on the

[17] The relationship between infomin prior selection and infomin updating is discussed in Hobson & Cheng (1973).

[18] In this paper, we do not presuppose the view according to which an agent is always guided by a probability distribution.

[19] Such a scenario is similar to that discussed in Paris (1998), where a doctor has to infer probabilities concerning a patient (the "next drawn object"), on the basis of linear constraints on the probability function, which are intended to reflect the doctor's background knowledge.

[20] The assumptions of the scenario justify the non-problematic application of the urn model. One can, if one wishes, add the further detail that the bag is shaken before the drawing.

legitimacy of Abe's method of choosing a prior. We stress that the problem is *not* due to any intuition that $P(B|A)$ should be independent of $P(A)$ – perhaps there should be some correlation between the two. Rather, the outcome is bizarre because the mere knowledge that Abe will be informed of the value of a certain objective parameter implies a substantial bound on how likely it is that a particular event will occur. The very fact that Abe will shortly know the ratio of expensive apples to apples makes it more plausible that the bag contains fewer apples than pears!

There are indeed circumstances in which mere knowability can have substantial implications. For instance, that a certain piece of information may be known can show that it is not top secret, or that it can be cheaply bought, or that it lacks significance (in some suitable sense of the term). Some such logic underlies the methodology of infomin. The updated probability is chosen so that, in retrospect, it is minimally surprising that the value of λ that determines the constraint turned out to be what it was. A prior commitment to infomin thus licenses the assumption that if new information should come to light, which requires some change in one's current probabilities, then the change will minimize the significance of that information. Now, in Abe's case, information relating the ratio of expensive apples to apples is, roughly speaking, less significant the smaller is the proportion of apples in the bag. Consequently, when Abe learns that he will soon come to know of this ratio, in order to minimize the significance of this future knowledge, he concludes *a priori* that the bag contains fewer apples than pears. This effect is responsible for the always-decreasing phenomenon of infomin updating on conditional probability constraints.[21]

Equally bizarre outcomes will occur in other cases of deceptive updating. In particular, an analogous objection can be raised against certain uses of infomin methods in the case of expected-value constraints (see §6). These phenomena show that the rationale underlying infomin is badly suited to many common scenarios of choosing or updating subjective probabilities.[22]

[21] To be sure, the finer details have to be worked out (see §5), because the deviation of $P(A)$ from 1/2 has its own price in terms of information.

[22] In his argument in support of infomin, Paris (1998) insists that the method only be applied to constraints reflecting the whole of the agent's knowledge. Are the counterintuitive consequences described above due to a failure to meet this requirement? We do not think so. The requirement that the constraints reflect the agent's total knowledge is justified with regard to explicit items in the scenario, but, as a general condition, it can never be fully met. In every case we assume that certain information is acquired by the agent, without going into the way it was acquired, or into the agent's reason for trusting it. Otherwise we are in for an infinite regress. Abe takes it for granted that the right value of $P(B|A)$ is written on a piece of paper. If pushed further he might say that the written number is the ratio of expensive apples to apples in the bag (we take it for granted that the drawing is random, and hence that Abe is justified in assuming that the conditional probability is that ratio). Now counting the apples and computing the ratio might be easier if there are fewer apples, hence the availability of this information suggest a smaller number of apples. This seems to be the rationale of infomin. But should speculations about how this ratio was known to the person who wrote it enter into the story, the story is endless.

The problems are enhanced if we consider the possibility of an agent's choosing between information from one of two distinct constraint families. In the apples-and-pears scenario let $A' = \overline{A}$, that is, the event that the next drawn fruit is a pear, and let B' be the event that the next drawn fruit will be an expensive pear. Now, suppose Abe is given the choice between learning the true value of $P(B|A)$, or the true value of $P(B'|A')$. If he opts for the first, then he knows that he will assign a probability ≤ 0.5 to A (and that the inequality will be strict unless his prior turns out to be uniform). If he opts for the option, he knows that he will assign a probability ≥ 0.5 to A. Thus, Abe finds himself in a situation in which he can determine certain features of his prior, and consequently whether he will prefer a bet on an apple to a bet on a pear, merely by choosing the kind of information he will receive.

The possibility of manipulating the chosen prior in this way also exists in the case of expected-value constraints. As we will see, for any event A, such that $2 \leq |A| \leq n-2$, there exist two random variables X and Y, such that, under infomin updating of the uniform prior, A is always decreasing with respect to $\{E(X) = \lambda\}_\lambda$, and always increasing with respect to $\{E(Y) = \kappa\}_\kappa$.

Call an updating method *shifty* if there are two families of conditional-probability constraints, or two families of expected-value constraints, such that for some event A, A is always decreasing with respect to one family and always increasing with respect to the other. When a choice between these constraint families is viewed as arbitrary, shiftiness prevents us from interpretating the updated probabilities as the agent's degrees of belief. Suppose, in the apples-and-pears scenario, that the prior Abe chooses on the basis of the constraint $P(B|A) = \lambda$ yields a probability of 0.45 for A. Can this express Abe's degree of belief, given that he knows with certainty that had he updated on information of the form $P(B'|A') = \kappa$, he would have assigned to A a probability ≥ 0.5?

Situations may arise in which an agent must choose between one of two constraint families (say, the information costs money, or time, and resources are limited), and where a random choice is to be preferred to no choice at all, or to the adoption of the uniform prior. In such a situation, the agent's choice will most likely be determined by pragmatic considerations which will vary from one context to another. If, for Abe's purposes, the most significant information is the ratio of expensive apples to apples, then this is the information he should acquire. Still, while some values of the chosen prior may express Abe's subjective probabilities, we should not pretend that this is true of the function as a whole.[23]

Of course, choosing a prior by appeal to a shifty updating method can only be criticized if the agent views the choice between the constraints as arbitrary. If,

[23] Classical statistics provides various prescriptions for choosing a statistical hypothesis, but it is wrong to interpret the adopted hypothesis as a subjective probability function, and even in the case of non-Bayesian methodologies, shiftiness, with the possibilities it provides for ad-hoc manipulation, would be considered a serious defect.

for example, in the case of expected-value constraints, one of the two random variables is of particular significance with respect to an agent's background knowledge, then the mere fact that the method is shifty does not constitute an objection (see §7).

4. Higher Order Support

A higher order support for an updating method is a probability μ defined over Λ, which satisfies the following two conditions:

(1.1) $Pr(A) = \int_\Lambda Pr_\lambda(A)\ d\mu(\lambda)$, for all $A \in \mathcal{B}$,

(1.2) $\mu(\{\lambda : Pr_\lambda \neq Pr\}) > 0$,

where it is a part of condition (1.2) that the set $\{\lambda : Pr_\lambda \neq Pr\}$ is measurable. We then have the following theorem, whose proof is trivial.

Theorem 4.1. If an updating method is supported by some higher order probability, then it is not deceptive.

Proof. Assume for contradiction that A is an always decreasing event. Let $\Lambda_0 = \{\lambda : Pr_\lambda = Pr\}$, and let $\Lambda_1 = \Lambda - \Lambda_0$. Then $Pr_\lambda(A) \leq Pr(A)$ for all $\lambda \in \Lambda$, and $Pr_\lambda(A) < Pr(A)$ for all $\lambda \in \Lambda_1$. By (1.2), $\mu(\Lambda_1) > 0$, hence, for some $\varepsilon > 0$, $\mu(\{\lambda : Pr_\lambda(A) < Pr(A) - \varepsilon\}) > 0$. But this implies that the integral on the right-hand side of (1.1) is strictly less than $Pr(A)$. □

A full-fledged higher order probability is a probability, m, defined on the Borel algebra of Δ, where Δ is the $(n-1)$ dimensional simplex consisting of all points representing first order probabilities, i.e., $\Delta = \{\overline{p} \in [0,1]^n : \sum_i p_i = 1\}$. We refer to the Borel subsets of Δ as higher order events. A higher order probability m determines through integration a first order probability Pr over Δ. In other words, if $P_{\overline{p}}$ is the probability represented by $\overline{p}$, then for all $A \in \mathcal{B}$:

$$Pr(A) = \int_\Delta P_{\overline{p}}(A)\ dm(\overline{p}) \tag{4.1}$$

A constraint on the first order probability is represented as a higher order event, $\Gamma \subseteq \Delta$, consisting of all points $\overline{p}$ such that $P_{\overline{p}}$ satisfies the constraint. Usually constraints are Borel sets. The m-weighted average of the probabilities satisfying the constraint Γ is obtained by means of the conditionalized measure $m(\ \ |\Gamma)$, i.e., for all $A \in \mathcal{B}$:

$$Pr_\Gamma(A) = \frac{1}{m(\Gamma)} \int_\Gamma P_{\overline{p}}(A)\ dm(\lambda) \tag{4.2}$$

This requires that $m(\Gamma) > 0$. If $m(\Gamma) = 0$ (as is often the case), the conditionalization can still be defined as a limit, provided Γ is sufficiently smooth and m is a non-pathological measure which assigns values > 0 to open sets containing Γ. For example, if Γ is the event $\{\overline{p} : E_{P_{\overline{p}}}(X) = \lambda\}$, then Γ is the intersection of Δ with a certain hyperplane. In this case, the conditional probability can usually be defined by conditionalizing on events of the form $\{\overline{p} : \lambda - \varepsilon < E_{P_{\overline{p}}}(X) < \lambda + \varepsilon\}$ and taking the limit as $\varepsilon \to 0$.

Given a family of constraints $\{C_\lambda\}_{\lambda \in \Lambda}$, we now establish in an abstract general setting how the space Δ is related to the parameter space Λ. We will say that $\overline{p}$ satisfies C_λ, if $P_{\overline{p}}$ does. We assume that (i) every C_λ is satisfied by some $\overline{p}$ (the λ's for which C_λ is unsatisfiable should be dropped from Λ); and (ii) every $\overline{p}$ satisfies at most one constraint C_λ. Let $\Delta' = \{\overline{p} \in \Delta : \overline{p}$ satisfies some $C_\lambda, \lambda \in \Lambda\}$. We do not assume that every $\overline{p}$ satisfies some constraint; for instance, in the case of conditional-probability constraints $\{P(B|A) = \lambda\}_\lambda$, $\Delta' = \{\overline{p} \in \Delta : P_{\overline{p}}(A) > 0\}$. We assume that Δ' is a Borel set. Its topology is, by definition, the topology induced by Δ.

For $\overline{p} \in \Delta'$, let $\pi(\overline{p})$ be the unique $\lambda \in \Lambda$ such that $\overline{p}$ satisfies C_λ. Then the function $\pi : \Delta' \to \Lambda$ is onto. We put on Λ the minimal topology for which π is continuous, i.e., a set $\Theta \subseteq \Lambda$ is open iff $\pi^{-1}(\Theta)$ is an open subset of Δ'. Henceforth, we use 'Λ' to refer to this topological space.

By a probability on Λ we mean a probability defined over the Borel field of Λ. (Note that, in most cases, $\Lambda \subseteq \mathbb{R}^k$ for some k, and the above topology is the one induced by the Euclidean space. It is instructive, however, to develop the framework in the context of a completely abstract parameter space Λ).

Definition 4.1. A probability m over Δ *induces* the probability μ over Λ if (i) $m(\Delta') > 0$; and (ii) $\mu(\Theta) = m(\pi^{-1}(\Theta)|\Delta')$, for every Borel subset $\Theta \subseteq \Lambda$.

It is sufficient to require that condition (ii) hold for all open subsets of Λ. Obviously every m such that $m(\Delta') > 0$ induces a unique μ over Λ, and if $m' = m(\ \ |\Delta')$, then $m'(\Delta') = 1$ and m' induces the same probability μ.

Definition 4.2. For a given family of constraints $\{C_\lambda\}_{\lambda \in \Lambda}$, a probability m over Δ *supports* the updating method U, if $m(\Delta') > 0$, and the probability which m induced on Λ, supports U.

Every updating method U can be associated with a function σ which selects for each $\lambda \in \Lambda$ a point in the set $\pi^{-1}(\lambda)$, corresponding to the updated probability:

$$P_{\sigma(\lambda)} = Pr_\lambda = U(Pr; C_\lambda)$$

We call the function $\sigma : \Lambda \to \Delta'$ the *selection function*. Obviously, σ is an embedding of Λ into Δ' such that $\pi(\sigma(\lambda)) = \lambda$ for all $\lambda \in \Lambda$. Thus σ is 1-1 and

its inverse is the restriction of π to $\sigma(\Lambda)$. Since π is continuous, the inverse of σ is continuous (where the topology of $\sigma(\Lambda)$ is induced by Δ').

Definition 4.3. An updating method is *continuous* if its corresponding selection function is continuous.

Roughly speaking, an updating is continuous if small changes in λ result in small changes in the updated probability. All the methods considered in the literature are continuous. Indeed, continuity is essential for the application of an updating method, when the parameter λ is reported with some margin of error.

If an updating method is continuous, then both σ and its inverse are continuous. Hence, σ is a homeomorphism from Λ onto $\sigma(\Lambda)$. Given any probability μ over Λ, we obtain by means of σ a corresponding probability μ' on its homeomorphic copy. We refer to μ' as the *copy* of μ, under the homeomorphism σ. Now, define the probability m_μ over Δ by:

$$m_\mu(\Gamma) =_{df} \mu'(\Gamma \cap \sigma(\Lambda))$$

where Γ ranges over the Borel subsets of Δ. It is easy to check that m_μ induces μ on Λ. This establishes the following claim:

Theorem 4.2. For a continuous updating method, and a family of constraints $\{C_\lambda\}_{\lambda\in\Lambda}$, every probability μ on Λ is induced by a probability m_μ on Δ, which is concentrated on a subset of Δ that is a homeomorphic copy of Λ. The restriction of m_μ to this homeomorphic copy is the copy of μ under the homeomorphism.

This shows that, for continuous updating methods, being supported by a higher order probability over Δ is not stronger than being supported by a higher order probability over Λ. As an illustration, consider a continuous updating method applied to the family of expected-value constraints $\{E(X) = \lambda\}_{\lambda\in\Lambda}$. Here Λ is a real closed interval, say $[a, b]$. Each $\lambda \in [a, b]$ defines a hyperplane that intersects Δ. The updating method chooses a point $\pi(\lambda)$ in the intersection, so that $[a, b]$ is mapped to a homeomorphic curve inside Δ. Any probability μ on $[a, b]$ can be induced by a probability over Δ that assigns measure 1 to that curve, and whose restriction to the curve is the copy of μ.

5. Updating on conditional-probability constraints

In this section we provide a proof of the following claim:

Theorem 5.1. (Seidenfeld 1987) Let $A, B \in \mathcal{B}$ be such that $A \neq \Omega$ and B is a non-empty proper subset of A. If $Pr(A) > 0$, A is always decreasing under infomin updating of Pr with respect to the constraint family $\{P(B|A) = \lambda\}_\lambda$.

This claim was first established in Seidenfeld (1987).[24] We present an alternative proof of the theorem, which gives new insight into the underlying cause of the "always-decreasing" phenomenon for conditional-probability constraints. Our proof, which relies on general features of the way in which information is measured, applies not only to infomin, but to other information-based methods as well.

We noted in section 3 that for conditional-probability constraints the reason for the "always decreasing" phenomenon is that the total cost of altering the conditional probability $Pr(\ \ |A)$ is lessened if we decrease the probability of A. To extract from this observation a general proof requires a more careful analysis of the competing costs involved in the update. We now proceed to provide such an analysis.

Assume that D is a relative information measure, i.e., a binary function that maps every pair (P, P^*) of probability functions over $\mathcal{B}$ to a real number $D(P, P^*)$. We will prove that certain general conditions on D, which are satisfied by the KL divergence, imply that updating by minimizing D-information under conditional-probability constraints is deceptive.

None of the updating methods discussed in the literature alter the values of 0-events (i.e., events whose prior probability is 0). This is presupposed throughout this paper. Hence, we can restrict the set of probabilities satisfying a constraint to those which assign probability 0 to all 0-events. We can therefore take as our universal set, the subset obtained by removing from Ω all ω for which $Pr(\omega) = 0$. Thus, without loss of generality, we assume that Pr is strictly positive, i.e., $Pr(\omega) > 0$ for all $\omega \in \Omega$ (of course, strict positivity is not required of the updated probabilities).

We consider constraints of the form $P(B|A) = \lambda$, where A and B are fixed events. Obviously, we can assume that $B \subseteq A$. We also assume that $\emptyset \subset B \subset A$, for otherwise there is only one possible constraint (either $P(B|A) = 1$ or $P(B|A) = 0$) which is trivially satisfied. We should also assume that $A \neq \Omega$, for, otherwise, the constraints reduce to $P(B) = \lambda$, in which case infomin yields the non-deceptive Jeffrey conditionalization. Finally the updated probability must satisfy $P(A) > 0$, in order for $P(B|A)$ to be defined. Let Δ^+ be the set of all strictly positive probabilities (over the subsets of Ω) and let Δ_A be the set of all probabilities that assign to A a non-zero value. Then, the following assumptions are made, where Pr is the prior probability: (i) $Pr \in \Delta^+$; (ii) $\emptyset \subset B \subset A \subset \Omega$; and (iii) there is a unique probability which minimizes D (relative to Pr) and this probability is in Δ_A.

[24] p. 283, Corollary 1, Appendix B. We thank an anonymous referee of an earlier draft of this paper for calling our attention to this fact. Our initial motivation for investigating deceptiveness in the context of conditional probability constraints, was a remark made in Grove & Halpern (1997).

For every probability function $P \in \Delta_A$, we write $P|A$ for the conditional probability function $P(\ \ |A)$ restricted to the algebra of subsets of A.

Any probability $P \in \Delta_A$ is uniquely determined by $P(A)$, $P|A$ and $P|\overline{A}$. Therefore, for any $P, P^* \in \Delta_A$, we can decompose the update from P to P^* into three successive steps: first, we change the probability of A from $P(A)$ to $P^*(A)$, leaving unchanged the conditional probabilities $P|A$ and $P|\overline{A}$; next, we change $P|A$ to $P^*|A$ leaving $P|\overline{A}$ unchanged; and finally, we change $P|\overline{A}$ to $P^*|\overline{A}$. Fully spelt out, the three steps are:

(1) $P \to P'$, where $P'(A) = P^*(A)$, $P'|A = P|A$, $P'|\overline{A} = P|\overline{A}$.

(2) $P' \to P''$, where $P''(A) = P'(A)$, $P''|A = P^*|A$, $P''|\overline{A} = P'|\overline{A}$.

(3) $P'' \to P^*$

We shall formulate three conditions on D. The first two are about the information costs for changes of type (1) and type (2) (obviously, (2) and (3) are changes of the same type), and the third concerns the way in which these separate costs combine to determine the total cost of the update from P to P^*.

Before proceeding, we note that if $P^*(A) = 1$, then $P'|\overline{A}$ is undefined. We do not rule out this possibility, but interpret the update, in this case, by dropping from (1) and (2) the equalities $P'|\overline{A} = P|\overline{A}$ and $P''|\overline{A} = P'|\overline{A}$, and omitting (3). Some of the following definitions will require obvious adjustments to account for this possibility, but the proof will carry through. We leave these adjustments to the reader.

The condition concerning changes of type (1) is referred to as *unimodality*. For convenience we represent the change as a change from P to P^* (i.e., we put $P' = P^*$). The condition requires that the cost depend only on the values of $P(A)$ and $P^*(A)$, that it is strictly decreasing as $P^*(A)$ approaches $P(A)$ from either side, and that it is a differentiable function of $P^*(A)$.

> **Unimodality** (UNI): Let D_1 be the restriction of D to the set of all $(P, P^*) \in \Delta_A \times \Delta_A$, such that $P|A = P^*|A$ and $P|\overline{A} = P^*|\overline{A}$. Then D_1 is independent of both $P|A$ and $P|\overline{A}$ and is strictly decreasing as $P^*(A)$ approaches $P(A)$ from the left, as well as from the right. Moreover, D_1 is a differentiable function of $P^*(A)$ in $(0, 1)$:

(If $P^*(\overline{A}) = 0$, then, by definition, the above set of pairs consist of all pairs $(P, P^*) \in \Delta^+ \times \Delta_A$, such that $P^*(A) = 1$ and $P|A = P^*|A$.)

Note that (UNI) implies that the derivative of D_1 with respect to $P^*(A)$ equals 0 at $P(A)$.

The condition concerning changes of type (2) is referred to as *conditional monotonicity*. Again, for convenience, we represent the change as a change from P to P^* (i.e., we put $P' = P$ and $P'' = P^*$). The condition requires that the cost not depend on $P|\overline{A}$; and that, for $P|A \neq P^*|A$, it is strictly increasing as a function of $P(A)$ and satisfies the differentiability requirement specified below.

> **Conditional Monotonicity** (CM) Let D_2 be the restriction of D to the set of all $(P, P^*) \in \Delta_A \times \Delta_A$ such that $P(A) = P^*(A)$ and $P|\overline{A} = P^*|\overline{A}$. Then D_2 is independent of $P|\overline{A}$. If $P^*|A \neq P|A$, then D_2 is differentiable with respect to $P^*(A)$ in $(0, 1)$, and its derivative is bounded below by some positive number (which can depend on $P|A$ and $P^*|A$).

The third and final condition relates the total cost $D(P, P^*)$ to the three costs $D(P, P')$, $D(P', P'')$, $D(P', P^*)$. Note that for $D = D_{KL}$, the total cost is simply the sum of the three costs:[25]

(ADD) $$D(P, P^*) = D(P, P') + D(P', P'') + D(P', P^*)$$

The property (ADD) often appears in arguments in support of D_{KL} as a measure of relative information, and it is closely linked with the use of log-ratios. It turns out, however, that the following much weaker requirement is all that is required in order to ensure deceptiveness (we assume the notation used in (1), (2), (3)):

> **Cost Combination** (CCO): For any $P \in \Delta_0$ there is a function $F(t, u, v, w)$ such that, for all $P^* \in \Delta_A$:
>
> $$D(P, P^*) = F(P^*(A), D(P, P'), D(P', P''), D(P'', P^*)),$$
>
> and the following hold:
>
> (i) F is strictly increasing in each of u, v, w.
>
> (ii) F has a total differential.
>
> (iii) $\partial F(t, u, v, 0)/\partial t \geq 0$.
>
> (iv) for some $c > 0$, $\partial F(t, u, v, 0)/\partial v > c$.

For $D = D_{KL}$, (CCO) holds trivially and it is also easy to check that (UNI) and (CM) are satisfied.[26] An example of an information measure satisfying (UNI)

[25] This is only true if the change from P to P^* is carried out in the order given by (1), (2) and (3). Additivity does not, in general, hold if the steps are carried out in a different order, say, by first adjusting $P|A$ and then $P(A)$.

[26] For $D = D_{KL}$ we have (assuming for each of the cases the corresponding notations for (1) and (2) used above): $D_1(P, P^*) = P^*(A) \log\left(\frac{P^*(A)}{P(A)}\right) + P^*(\overline{A}) \log\left(\frac{P^*(\overline{A})}{P(\overline{A})}\right)$ and $D_2(P, P^*) = P^*(A)\left(\sum_{\omega \in A} P^*(\omega) \log\left(\frac{P^*(\omega)}{P(\omega)}\right)\right)$. The first implies (UNI) and the second implies (CM).

and (CM) that does not satisfy (ADD), yet satisfies (CCO), is $D^\dagger$, obtained by substituting the simple ratio for the log-ratio in D_{KL}:

$$D^\dagger(P, P^*) = \left(\sum_{\omega \in \Omega} P^*(\omega) \left(\frac{P^*(\omega)}{P(\omega)} \right) \right) - 1$$

Hence, the following theorem implies that $D^\dagger$-infomin is deceptive.[27]

Theorem 5.2. Assume that D satisfies (UNI), (CM) and (CCO). Let $Pr \in \Delta^+$ and let Pr_λ be the probability that minimizes $D(Pr, P)$ under the constraint $P(B|A) = \lambda$. Then, for all $\lambda \in (0, 1)$, if $Pr_\lambda \neq Pr$, then $Pr_\lambda(A) < Pr(A)$.

Proof. First note that $Pr_\lambda|\overline{A} = Pr|\overline{A}$; otherwise, condition (i) in (CCO) implies that we can further reduce the cost by changing $Pr_\lambda|\overline{A}$ to $Pr|\overline{A}$. Hence, the probability Pr^* that minimizes the total cost, minimizes the function $F(P^*(A), D(P, P'), D(P', P''), 0)$, where $P = Pr$. That is, Pr^* minimizes $f(P^*(A), D(P, P'), D(P', P''))$, where $f(t, u, v) =_{df} F(t, u, v, 0)$.

Now, if $Pr^*(A) > Pr(A)$, then we can get a further reduction of cost, by updating Pr to Pr^{**}, where $Pr^{**}(A) = Pr(A)$, $Pr^{**}|A = Pr^*|A$ and $Pr^{**}|\overline{A} = Pr^*|\overline{A}$. This follows from the fact that $f(t, u, v)$ is non-decreasing in t (since $\partial f/\partial t \geq 0$) and is strictly increasing in u and v; by (UNI) there is a reduction in cost of the first step (i.e., a reduction in u), and – by (CM) – a reduction in cost of the second step (i.e., a reduction in v).

It remains to show that if $\lambda \neq Pr(B|A)$, then $Pr^*(A) \neq Pr(A)$. Since the conditional probabilities relative to A and to $\overline{A}$ are held fixed, the total cost is a function of $Pr^*(A)$. It is therefore sufficient to show that $df/dt > 0$ at the point $t = Pr(A)$. We have:

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial u}\frac{\partial u}{\partial t} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial t}$$

By (iii) of (CCO) the first term is ≥ 0; by (UNI) the second term is 0; and by (iv) of (CCO) and (CM) the third term is > 0. □

6. Infomin updating on expected-value constraints

Throughout this section, we take X to be a fixed random variable and the prior probability to be uniform, unless otherwise indicated. The following theorem

[27] Seidenfeld's proof of theorem 5.1, is altogether different from that developed in this section. He first proves the result for the case where the prior is uniform. He then extends the result to arbitrary rational-valued priors by refining Ω, and noting that infomin is preserved under such refinements. Finally, he obtains the general result by taking limits. As it turns out, this method can also be applied to $D^\dagger$. We have not, in this work, investigated the relative scopes of the two methods.

settles the question of higher order support for infomin updating under expected-value constraints:

Theorem 6.1. (Shimony 1973) If the range of X contains at least three distinct values, there is no higher order probability that supports infomin updating under the constraints $\{E(X) = \lambda\}_\lambda$.

A weaker version of this theorem was proven in Friedman & Shimony (1971).[28]

Theorem 6.1 has served as the basis for many of the past criticisms of infomin methods.[29] As we have noted, however, arguments from lack of higher order support are based on the substantial assumption of a higher order probability, which is open to rejection, as acknoweldged by Friedman and Shimony (see footnote 9). Arguments from deceptiveness, on the other hand, appeal only to straightforward coherence conditions, which require no more than a minimal capacity for self-reflection. The extent to which lack of higher order support implies deceptiveness is therefore worth investigating. It turns out that for a wide variety of random variables infomin is deceptive, but for many others it is not.

We adopt the following notation: If $E(X) = E_P(X)$ is the expected value of X for the probability P and if $P(A) > 0$, then we write '$E(X|A)$' for the conditional expected value of X, i.e. $E(X|A) = 1/P(A) \sum_{\omega \in A} X(\omega)$. Obviously, if $0 < P(A) < 1$, then:

$$E(X) = P(A) \cdot E(X|A) + P(\overline{A}) \cdot E(X|\overline{A}) \tag{1}$$

We put $E_0(X) = E_{P_0}(X)$, $E_0(X|A) = E_{P_0}(X|A)$, where P_0 is the uniform probability over Ω.

We shall state a condition on X that is necessary for the deceptiveness of infomin updating (of the uniform prior) and which shows that, for any Ω such that $|\Omega| \geq 2$, there are random variables for which the updating is not deceptive. The condition is based on a necessary condition for an event A to be always decreasing.

Some conditions on A are obviously necessary. If M and m are, respectively, the maximum and minimum values of X, and if $\{\omega : X(\omega) = M\} \subseteq A$, then, trivially, A is not always decreasing since the constraint $E(X) = M$ is satisfied only if the probability assigns to $\{\omega : X(\omega) = M\}$ (and hence to A) the value 1. Similarly, A is not always decreasing if $\{\omega : X(\omega) = m\} \subseteq A$. A deeper

[28] The weaker version made the additional assumption that for some $\omega \in \Omega$, $X(\omega)$ was equal to the mean value of X. The "always decreasing" phenomena was utilized in the proof in a roundabout way: it was shown that $\{\omega\}$ is always-decreasing, but this fact was derived under the presupposition of higher order support. Shimony's proof in (1973) is altogether different and it obviates the additional assumption.

[29] See, e.g., Shimony (1985) and Seidenfeld (1979).

condition is required in order to establish the existence of random variables for which no event is always decreasing.

Definition 6.1. A is a *mean* event if $A \neq \emptyset$ and $E_0(X|A) = E_0(X)$.

Obviously, for $\emptyset \subset A \subset \Omega$, (1) implies that A is a mean event iff $\overline{A}$ is a mean event.

Theorem 6.2. If A is always decreasing, then A is a mean event.

Note that if $A \neq \Omega$, the theorem implies that if A is not a mean event, it is also not always increasing; because, if A is a not a mean event, neither is $\overline{A}$. Hence, $\overline{A}$ is not always decreasing, which implies that A is not always increasing.

The proof of the theorem, given in Appendix A, shows that if $|E_0(X|A) - E_0(X)| > 0$, then for some $\varepsilon > 0$, $Pr_\lambda(A) \geq P_0(A)$, for all λ between $E_0(X)$ and $E_0(X|A)$ satisfying $|\lambda - E_0(X)| < \varepsilon$.

As a corollary of the theorem we have:

Corollary. A necessary condition for the deceptiveness of infomin updating of the uniform prior under expected-value constraints is the existence of a non-trivial (i.e., $\neq \Omega$) mean event.

For any non-empty Ω, it is easy to define random variables over Ω for which there are no non-trivial mean events; for example, if $\Omega = \{\omega_i : 1 \leq i \leq 6\}$, put $X(\omega_i) = i$, for $i = 1, \ldots, 5$, and $X(\omega_6) = 7$. Infomin updating on the expected value of such a random variable is not deceptive, though it lacks higher order support.

We now state a sufficient condition for deceptiveness.

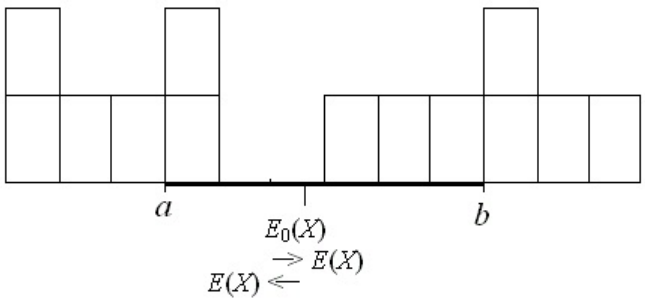
Definition 6.2. An *interval* event is a non-empty event of the form:

$$\{\omega \in \Omega : a \leq X(\omega) \leq b\}$$

Theorem 6.3. If $A \neq \Omega$ is a mean event and is also an interval event, then A is always decreasing.

Hence, a sufficient condition for the deceptiveness of infomin updating is the existence of a mean, interval event $\neq \Omega$.

The proof of theorem 6.3 is given in Appendix B. The following illustration, which is a histogram of X, clarifies the heuristics underlying both theorems 6.2 and 6.3. The probabilities (which are plotted on the vertical axis) are of the



form k/n, where $n = |\Omega|$ and k is the number of atoms for which X assumes the given value. The interval event A is $X^{-1}([a, b])$.

If $\lambda \neq E_0(X)$, the constraint $E(X) = \lambda$ requires that we move the expected value to the right (left), provided $\lambda > E_0(X)$ ($\lambda < E_0(X)$). This shift is subject to an obvious torque effect, so that a change in the probability of an atom leads to a greater change in the value of $E(X)$, the farther away the atom is from the "center" at $E_0(X)$. If $E_0(X|A) = E_0(X)$, and the required shift in expected value is sufficiently small, then the minimal information shift will involve first moving some mass from A to some further removed atom in its complement, where it can have a greater effect, and then adjusting the mass distributions inside A and $\overline{A}$. If, however, $E_0(X|A) > E_0(X)$, then sufficiently small increases in $E(X)$ can be more cheaply accomplished by moving the center of gravity of A to the right and, possibly, increasing its weight. To show that the costs balance in this way, we appeal to the strict convexity of the infomin update of the uniform prior. A further promising line of investigation is to consider whether the cost-balancing intuitions can be used to generalize the results to priors which are non-uniform.

Note that theorem 6.3 is not valid if the assumption that A is an interval event is omitted. This is because one can define a random variable X for which there is a mean event $A \neq \Omega$ such that $X^{-1}(M) \subseteq A$, in which case, the constraint $E(X) = M$ requires that A be assigned a probability of 1. Let A be an *interior* event if $m < X(\omega) < M$, for all $\omega \in A$. Then a possible strengthening of theorem 6.3 (which considerably narrows the gap between the necessary and sufficient conditions, and is therefore worth investigating) is obtained by replacing the assumption that A is an interval event with the assumption that it is an interior event.

7. A Brief Apology for Jaynes

We have argued that in a wide variety of cases infomin should be rejected, on the grounds that it is incoherent when used as a method for either updating

or choosing a subjective probability function. We suggest that on a certain charitable reading of Jaynes (one which focuses on his particular applications of the method rather than his pronouncements concerning its *a priori* status) his own appeal to infomin avoids the incoherencies which result from its unqualified application. The crucial observation is that, on Jaynes's account, infomin can only be applied when the problem of choosing a probability satisfying the given constraints is 'well-posed', or, in Jaynes's words, when the constraints to which the method is applied reflect '*all* the physical constraints actually operating in the random experiment.' As we will see, this requirement rules out the use of infomin as a method for updating probabilities, and, in the context of prior selection, limits its correct usage to those scenarios in which an agent possesses substantial background knowledge concerning the stochastic mechanism at work in the setup.

For Jaynes, the 'constraints' that figure in an infomin analysis are not merely abstract conditions on an agent's subjective probabilities. They are meant to reflect the agent's knowledge concerning the actual stochastic process underlying the scenario in which infomin is to be applied. The prior probability that is chosen under these constraints represents the agent's best guess as to the physical probabilities that characterize this process. Interpreting the chosen probability this way (as opposed to merely viewing it as an agent's degrees of belief at a given time) leads to a sharp distinction between the act of *choosing* a prior – a choice that is always based on a physical hypothesis, within the background of some physical theory – and that of *updating* a prior, which amounts to Bayesian conditionalization on some event in $\mathcal{B}$.[30]

On Jaynes's view, once a prior has been selected by infomin, an agent's probabilties are updated by Bayesian conditionalization, until a point at which the agent receives new information necessitating a change in his prior probabilities. The agent then appeals to infomin to select a new prior probability, and a new series of conditionalizations begins. Thus, on Jaynes's account, the revision of an agent's subjective probabilities proceeds in fits and starts, with periods of conditionalization punctuated by abrupt alterations of the prior.

An abrupt change of prior is the kind of change that takes place when an agent receives new information which leads him to reject the hypothesis that "... the information incorporated into the ... analysis includes all the constraints actually operating in the random experiment..."[31] When the evidence suggests that this is not the case, the agent must start again from square one, selecting a new prior via the principle of maximum entropy (i.e., infomin) under a different

[30] Jaynes clearly distinguishes between evidence reporting the truth of an event in $\mathcal{B}$, and constraints on the probability over $\mathcal{B}$: "If a statement referring to a probability distribution... is testable (for example, if it specifies a mean value... for some random variable defined on Ω) ... then it can be used as a constraint in [infomin]; but it cannot be used as a conditioning statement in Bayes' theorem because it is not a statement about any event in $\mathcal{B}$ or any other space ...." (Jaynes 1983, p. 262)

[31] Jaynes (2003, p. 371).

set of constraints than those which led to the choice of his previous prior.[32] Obviously, this process of starting over cannot be decribed as "updating" in any meaningful sense.[33]

Now, what exactly it means for a set of constraints to express 'all the constraints operating in a physical experiment' is not entirely clear, but it is an essential part of Jaynes's view that, in practice, scientists are often in a position to make such judgments. In other words, they are often in a position to assess whether or not a given problem is *well-posed*. To provide a satisfactory account of what it means for a problem to be well-posed is a difficult task. For our present purposes, however, it suffices to note that a well-posed problem provides an agent with sufficient information to determine its solution, and is also such that any information not included in the problem statement can be ignored. According to Jaynes, then, infomin is a rule that tells us which prior probabilities we ought to assign, given the assumption that the problem of choosing a probability satisfying the given constraints is well-posed. Insofar as this assumption is justified, we are justified in using the selected prior as the basis for subsequent conditionalization. If new information gives us reason to reject this assumption, the chosen prior is to be thrown out, and we must start from scratch.

As an illustrative example, consider the case of coin tossing with outcomes 0 and 1, and assume that $Pr(0) = 1/3$ and $Pr(1) = 2/3$, where Pr is an agent's prior. Let X be some random variable on the space $\{0, 1\}$. Then the prior expectation of X is given by $\gamma = \frac{1}{3}X(0) + \frac{2}{3}X(1)$. Now, suppose that the coin is tossed 100 times and let X^* be a random variable (on the extended sample space), such that for any sequence s of length 100, if $m(s) =$ number of 0's in s, then:

$$X^*(s) = \frac{1}{100}[m(s)X(0) + (100 - m(s))X(1)]$$

Suppose the evidence gives to the agent the observed value, λ, of X^*. Then the only "update" available in Jaynes's framework is that obtained by conditional-

[32] "... [S]uppose an experiment fails to confirm the maximum entropy prediction... Then, since by hypothesis the [original] data were true if incomplete, we will conclude that the physical mechanism of the experiment must contain some additional constraint which was not taken into account in the maximum entropy calculation... In this way, [one] can discover empirically that his information was incomplete..."(Jaynes 2003, p. 370).

[33] It is quite clear that Jaynes never proposed that infomin techniques be used for the purpose of updating prior probabilities. In fact, on the only occassion in which Jaynes makes explicit reference to the KL divergence, he introduces it as an alternative to the χ^2 statistic, as a measure of the 'goodness of fit' between a statistical hypothesis and the empirical measure obtained from a data sample. This appeal to the KL divergence is clearly very different from its use in updating probabilities, where it is taken as a measure of the distance between statistical hypotheses themselves. The notion of 'relative' information is discussed by Jaynes, but the relativization he envisioned was not with respect to a prior probability, but rather to an underlying measure reflecting the physical symmetries in the problem. This measure was introduced so as to ensure that the continuous form of the Shannon information would be invariant under a change of variables (see Jaynes (1968), sec. 6-8). Since the problem of invariance does not arise in the context of discrete spaces, the notion of relative information does not appear at all in Jaynes's discussion of the application of infomin methods to finite random experiments, such as those considered here.

izing on the event $X^* = \lambda$ (Note that, the tosses being construed as a Bernoulli sequence, the probability of events relating to future tosses is unaffected by this evidence). If, however, $|\lambda - \gamma|$ is not sufficiently small, the agent may deduce that the real value of $E(X)$ is not γ, but λ (assuming that 100 tosses is taken to be a sufficiently large sample). In that case, he should discard Pr and select a new prior by means of infomin applied to a different constraint than that which was used to obtain Pr. In the simplest case the agent might choose a new prior under the constraint $E(X) = \lambda$, but, conceivably, the evidence might indicate that there are other more complicated physical factors at work in the setup.

Note that this paper is based on an altogether different perspective on updating than that of Jaynes. We assume that additional information can, in many cases, lead to rationally justified alterations of one's prior, and that conditionalization can be viewed as the simplest of a broad spectrum of updating methods (this point was elaborated somewhat at the beginning of section 2). We also think that one's choice of a prior can in many cases be justified wihtout presupposing a physical hypothesis that determines the underlying stochastic mechanism of the setup. In short, we assume a Bayesian view that is more comprehensive than that which figures in Jaynes's system. Nonetheless, we also think that, in light of the sorts of limitations discussed in footnote 6, there is always the possibility of new evidence pointing to hypotheses not previously envisaged, and that the abrupt changes of prior which result from this information cannot be viewed as the result of any act of "updating". Thus, our position can be described as falling somewhere "between" that of Jaynes and the extreme Bayesian, who believes that human knowledge acquisition can in principle be reduced to a sequence of conditionalizations applied to an "original prior".

The requirement that infomin only be applied to well-posed problems also allows Jaynes to avoid the difficulties discussed in §3. Recall, for example, that shiftiness occurs when an agent is faced with an arbitrary choice between different constraint families, e.g., the expected value of one variable versus the expected value of another. If the problem is well-posed, however, an agent's background knowledge will determine which of the constraints has physical significance.

To take a concrete example, consider the case of tossing a six-sided die. In this case, $\Omega = \{\omega_1, \ldots, \omega_6\}$. Let $X(\omega_i) = i$, and let $A = \{\omega_2, \ldots, \omega_5\}$. If an agent, Judy, is about to choose a prior based on the value of $E(X)$, then she knows already that she will assign to A a probability $\leq 2/3$, because theorem 6.3 implies that A is always decreasing (A is obviously a mean interval event). Now suppose that Judy can also choose a prior on the basis of the value of $E(Y)$, where:

$$Y(\omega_i) = X(\omega_i) + 3 \text{ (mod 6)}$$

Now, with respect to Y, $\overline{A}$ satisfies the conditions of theorem 6.3. Hence if she chooses a prior under this information, she will end up assigning a prior probability to A which is $\geq 2/3$. As we have noted, this situation is unacceptable if the decision between $E(X)$ and $E(Y)$ is arbitrary. On the other hand, if

certain assumptions that Judy accepts endow one of the variables, say X, with a particular physical significance, then there is nothing problematic about her choosing a prior on the basis of the value of $E(X)$. Such background knowledge already implies that the probability of A is $\leq 2/3$.[34]

This is an important and often overlooked fact concerning the applicability of infomin methods. The agent, before employing the method, must already possess a sufficient theoretical knowledge of the situation to support the choice of the family of constraints, which are used in applying the method; such knowledge can itself determine certain features of the prior.

As it turns out, this was a point upon which Jaynes himself was perfectly clear. In responding to an objector, who asked rhetorically whether there is anything in the physics of throwing dice to suggest the plausibility of the infomin prior, Jaynes remarked:

> ... [T]he most obvious imperfection (of a die) is that different faces have different numbers of spots. This affects the center of gravity, because the weight of ivory removed from a spot is obviously not (in any die I have seen) compensated by the paint then applied. Now, the numbers of spots on opposite faces add up to seven. Thus, the center of gravity is moved towards the "3" face, away from "4", by a small distance x corresponding to the one spot discrepancy. The effect of this must be a slight probability difference which is surely, for very small x, proportional to x... But the (2-5) face direction has a discrepancy of three spots, and the (1-6) of five. Therefore we anticipate the ratios: $(p_4 - p_3):(p_5 - p_2):(p_6 - p_1) = 1 : 3 : 5$. But this says ... that the spot frequencies vary linearly with i... This is the most obvious "physical constraint"...[35]

This passage makes it clear that in spite of the fact that Jaynes considered the example of the die to be merely illustrative, he clearly had a very detailed physical experiment in mind.[36] Indeed, additional assumptions pertaining to how the die was manufactured led Jaynes to further suppose that the probability distribution should have a convex form. All these considerations lead

[34] The problem of shiftiness bears a close conceptual connection to the well-known Bertrand paradoxes associated with the principle of indifference. In both cases, the problem is that the agent has too much freedom to arbitrarily manipulate his prior, and in both cases the difficulties are resolved by attributing to the agent further knowledge in the form of judgments concerning what is and what is not relevant information. See (Keynes 1920, ch. 4)

[35] Jaynes (1983), p. 259

[36] Consider again the case of drawing at random from an urn containing several kinds of objects in unknown proportions. Without knowing anything about the physical factors determining these proportions, the problem of choosing a probability is not well-posed. If, on the other hand, we know that the urn was selected at random from a large collection of urns all of which agree as to $E(X)$, then it may be a well-posed problem to choose a probability (and hence one would be justified in applying infomin) under the constraint $E(X) = \lambda$.

to the identification of $E(X)$ as the parameter that captures the information underlying the physical process, and this knowledge, in itself, implies that the probability of A should be $\leq 2/3$. As Jaynes puts it: "...[I]t is enough if we can recognize...what are the systematic influences at work that represent the 'physical constraints'. If by any means we can recognize these, infomin then takes over and supplies the rest of the solution."[37]

[37] Jaynes (1983), pp. 266-67

Appendix A

We prove theorem 6.2: If an event A is always decreasing under infomin updating of the uniform prior with respect to the family $\{E(X) = \lambda\}_\lambda$, then A is a mean event.

The infomin update of the uniform prior is the Maxwell-Boltzmann distribution:

$$Pr_\lambda(\omega) = Ce^{\beta X(\omega)} \tag{2}$$

where $C = (\sum_{\omega\in\Omega} e^{\beta X(\omega)})^{-1}$ is a normalizing factor, and $\beta = \beta(\lambda)$ is a constant depending on λ, which, as a function of λ, is continuous, strictly monotone and equals 0 for $\lambda = E_0(X)$.

For each $\lambda \in [m, M]$, where M and m are the maximum and minimum of X, respectively, let:

$$\varphi_\lambda(u) = Ce^{\beta u}$$

We assume that λ ranges over $[m, M]$. Note that if $\lambda \neq E_0(X)$ (i.e., $\beta \neq 0$), φ_λ is strictly convex. Thus, for all $\lambda \neq E_0(X)$ and for any $A \neq \emptyset$:

$$\frac{Pr_\lambda(A)}{|A|} \geq \varphi_\lambda(E_0(X|A)) \tag{3}$$

Claim: Let $y \in [m, M]$ be any number such that $y \neq E_0(X)$. Then there exists $\varepsilon > 0$, such that for every $\lambda \neq E_0(X)$ which is on the same side of $E_0(X)$ as y, if $|\lambda - E_0(X)| < \varepsilon$, then:

$$\varphi_\lambda(y) \geq \frac{1}{n}$$

Proof. The Taylor expansion of $e^{\beta u}$ (about $u = 0$) is given by:

$$e^{\beta u} = 1 + \beta u + R_\beta(u)$$

with remainder term:

$$R_\beta(u) = \frac{(\beta u)^2 e^{\beta c}}{2} \tag{4}$$

where c is a constant between 0 and u. For all $\omega \in \Omega$, let $x_\omega = X(\omega)$. The inequality holds just in case:

$$\begin{aligned} 1 + \beta y + R_\beta(y) \geq \frac{1}{Cn} &= \frac{1}{n}\sum_{\omega\in\Omega}(1 + \beta x_\omega + R_\beta(x_\omega)) \\ &= 1 + \beta E_0(X) + \frac{1}{n}\sum_{\omega\in\Omega} R_\beta(x_\omega) \end{aligned}$$

In other words, the required inequality is:

$$\beta(y - E_0(X)) \geq \frac{1}{n}\left(\sum_{\omega \in \Omega} R_\beta(x_\omega)\right) - R_\beta(y)$$

By assumption, $y - E_0(X) \neq 0$. Choose λ on the same side of $E_0(X)$ as y. Letting:

$$F(\gamma) =_{df} \frac{1}{n}\left(\sum_{\omega \in A} \frac{R_\gamma(x_\omega)}{|\gamma|}\right) - \frac{R_\gamma(y)}{|\gamma|} \tag{5}$$

it follows that the inequality is satisfied just in case:

$$|y - E_0(X)| \geq F(\beta) \tag{6}$$

But from (4) and (5), for all y:

$$\lim_{\gamma \to 0} F(\gamma) = 0$$

Hence, (6) holds if $|\lambda - E_0(X)| > 0$ is sufficiently small. □

The proof now proceeds as follows. Suppose that A is *not* a mean event. Then $E_0(X|A) \neq E_0(X)$. The claim implies that for some $\varepsilon > 0$, if λ is between $E_0(X)$ and $E_0(X|A)$, and $|\lambda - E_0(X)| < \varepsilon$, then:

$$\varphi_\lambda(E_0(X|A)) \geq \frac{1}{n}$$

Hence, (3) implies that for all such λ's:

$$Pr_\lambda(A) \geq \frac{|A|}{n}$$

Thus, A is not always decreasing.

Appendix B

We assume the notation introduced in Appendix A. We prove theorem 6.3: If $A \neq \Omega$ is a mean, interval event, then A is always decreasing under infomin updating with respect to the uniform prior and the family of constraints $\{E(X) = \lambda\}_\lambda$.

Recall that a function f is strictly convex iff, for all $a < c$, and all $0 < \alpha < 1$:

$$\alpha f(a) + (1-\alpha) f(c) > f(\alpha a + (1-\alpha)c)$$

The following elementary generalization is implied by strict convexity:

Claim: For a strictly convex f, if $a < b_i < c$, $i = 1, \ldots, k$, and $0 < \alpha < 1$ is such that $\alpha a + (1-\alpha)c = \frac{1}{k}(b_1 + \ldots + b_k)$, then:

$$\alpha f(a) + (1-\alpha) f(c) > \frac{1}{k}(f(b_1) + \ldots + f(b_k))$$

(Note that the condition defining strict convexity is obtained if $k = 1$.)

Proof. Let $b^* = 1/k \sum_i b_i$. Thus, $\alpha = (c - b^*)/(c - a)$ and $(1 - \alpha) = (b^* - a)/(c - a)$. Then, for all $y \in (a, c)$:

$$f(y) < \left(\frac{c-y}{c-a}\right) f(a) + \left(\frac{y-a}{c-a}\right) f(c)$$

Hence:

$$\begin{aligned} \frac{1}{k}\sum_{i=1}^{k} f(b_i) &< \frac{1}{k}\sum_{i=1}^{k}\left(\left(\frac{c-b_i}{c-a}\right) f(a) + \left(\frac{b_i-a}{c-a}\right) f(c)\right) \\ &= \left(\frac{c-b^*}{c-a}\right) f(a) + \left(\frac{b^*-a}{c-a}\right) f(c) = \alpha f(a) + (1-\alpha) f(c) \end{aligned}$$

□

The proof now proceeds as follows: Let $A = \{\omega_1, \ldots, \omega_k\}$ be a mean, interval event, and let $\lambda \in [m, M]$ be any number $\neq E_0(X)$. We have to show that $Pr_\lambda(A) < |A|/n$.

Consider the events:

$$A^- = \{\omega \in \Omega : x_\omega < \min_{\omega \in A} x_\omega\}$$

$$A^+ = \{\omega \in \Omega : x_\omega > \max_{\omega \in A} x_\omega\}$$

Since A is an interval event, $A \cup A^- \cup A^+ = \Omega$, and since A is a mean event, distinct from Ω, both A^- and A^+ are non-empty. Put $x^- = E_0(X|A^-)$ and

$x^+ = E_0(X|A^+)$. The numbers $x^-, x_{\omega_1}, \ldots, x_{\omega_k}, x^+$ satisfy the conditions of the claim with:

$$\alpha = \frac{|A^-|}{|A^-| + |A^+|}$$

Hence:

$$\frac{|A^-|\varphi_\lambda(x^-) + |A^+|\varphi_\lambda(x^+)}{|A^-| + |A^+|} > \frac{1}{k}\sum_{\omega \in A} \varphi_\lambda(x_\omega) = \frac{Pr_\lambda(A)}{|A|} \tag{7}$$

Again appealing to the convexity of φ_λ, we get:

$$\frac{Pr_\lambda(\overline{A})}{|\overline{A}|} = \frac{Pr_\lambda(A^-) + Pr_\lambda(A^+)}{|A^-| + |A^+|} > \frac{|A^-|\varphi_\lambda(x^-) + |A^+|\varphi_\lambda(x^+)}{|A^-| + |A^+|} \tag{8}$$

From (7) and (8), we have:

$$\frac{Pr_\lambda(A)}{|A|} < \frac{Pr_\lambda(\overline{A})}{|\overline{A}|}$$

from which it follows that $Pr_\lambda(A) < |A|/n$. Hence, A is always decreasing.

References

de Finetti, B. (1974), *Theory of Probability, Volume 1*, John Wiley and Sons.

Friedman, K. & Shimony, A. (1971), 'Jaynes's maximum entropy prescription and probability theory', *Journal of Statistical Physics* **3**(4), 381.

Gaifman, H. (1983), 'Paradoxes of infinity and self-applications, I', *Erkenntnis* **20**, 131–155.

Gaifman, H. (1986), A theory of higher order probabilities, *in* J. Halpern, ed., 'Theoretical Aspects of Reasoning about Knowledge', Morgan Kaufmann Publishers Inc.

Gaifman, H. (2004), 'Reasoning with limited resources and assigning probabilities to arithmetical statements', *Synthese* **140**, 97–119.

Gaifman, H. & Snir, M. (1982), 'Probabilities over rich languages, testing and randomness', *Journal of Symbolic Logic* **47**(3), 495–548.

Good, I. J. (1972), '46,656 varieties of bayesians', *American Statistician* **25**, 62–63.

Grove, A. & Halpern, J. (1997), Probability update: Conditioning vs. cross entropy, *in* 'Proceedings of the Thirteenth Annual Conference on Uncertainty in Artifical Intelligence'.

Hobson, A. & Cheng, B. (1973), 'A comparison of the shannon and kullback information measures', *Journal of Statistical Physics* **7**(4), 301–310.

Jaynes, E. T. (1957), 'Information theory and statistical mechanics, 1', *Physical Review* **106**, 620–630.

Jaynes, E. T. (1968), Prior probabilities, *in* R. Rosenkrantz, ed., 'E.T. Jaynes: Papers on Probability, Statistics and Statistical Physics', D. Reidel Publishing Company, Boston, U.S.A., pp. 116–130.

Jaynes, E. T. (1983), Where do we stand on maximum entropy?, *in* R. Rosenkrantz, ed., 'E.T. Jaynes: Papers on Probability, Statistics and Statistical Physics', D. Reidel Publishing Company, Boston, U.S.A., pp. 210–314.

Jaynes, E. T. (2003), *Probability Theory: The Logic of Science*, Cambridge University Press.

Jeffrey, R. (1965), *The Logic of Decision*, McGraw Hill.

Keynes, J. M. (1920), *A Treatise on Probability*, 2006 edn, Cosimo, Inc., New York, NY.

Kullback, S. & Leibler, R. (1951), 'On information and sufficiency', *Annals of Mathematical Statistics* **22**(1), 79–86.

Levi, I. (1985), 'Imprecision and indeterminacy in probability judgment', *Philosophy of Science* **52**(3), 390–409.

Paris, J. (1998), 'Common sense and maximum entropy', *Synthese* **117**(1), 75–93.

Paris, J. & Vencovská, A. (1997), 'In defense of the maximum entropy inference process', *International Journal of Approximate Reasoning* **17**(1), 77–103.

Putnam, H. (1963), Degree of confirmation and inductive logic, *in* P. Schilpp, ed., 'The Philosophy of Rudolf Carnap', The Open Court Publishing Co.,, pp. 761–784.

Savage, L. (1954), *The Foundations of Statistics*, John Wiley and Sons.

Seidenfeld, T. (1979), 'Why I am not an objective bayesian', *Theory and Decision* **11**, 413–440.

Seidenfeld, T. (1987), Entropy and uncertainty (revised), *in* I. MacNeill & G. Humphreys, eds, 'Foundations of Statistical Inference', D. Reidel Publishing Co., pp. 259–287.

Shannon, C. E. (1948), 'A mathematical theory of communication', *The Bell System Technical Journal* **27**, 379–423.

Shimony, A. (1973), 'Comment on the interpretation of inductive probabilities', *Journal of Statistical Physics* **9**(2), 187–191.

Shimony, A. (1985), 'The status of the principle of maximum entropy', *Synthese* **63**, 35–53.

Shore, J. & Johnson, R. (1980), 'Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy', *IEEE Trans. on Info. Theory* **IT-26**(1), 26–37.

van Fraassen, B. (1981), 'A problem for relative information minimizers in probability kinematics', *The British Journal for the Philosophy of Science* **32**(4), 375–379.