1st International Symposium on Imprecise Probabilities and Their Applications, Ghent, Belgium, 29 June - 2 July 1999

On the distribution of natural probability functions

J.B.Paris*, **P.N.Watton** † **and G.M.Wilmers***
Department of Mathematics
University of Manchester
Manchester M13 9PL
UK
email jeff@ma.man.ac.uk

Abstract

The purpose of this note is to describe the underlying insights and results obtained by the authors, and others, in a series of papers aimed at modelling the distribution of 'natural' probability functions, more precisely the probability functions on $\{0,1\}^n$ which we encounter naturally in the real world as subjects for statistical inference, by identifying such functions with large, random, sentences of the propositional calculus. We explain how this approach produces a robust parameterised family of priors, J_n, with several of the properties we might have hoped for in the context, for example marginalisation, invariance under (weak) renaming, and an emphasis on multivariate probability functions exhibiting high interdependence between features.

Keywords. Prior probability, imprecise probability, random rentences, probabilistic reasoning, uncertain reasoning.

1 Motivation

The motivation for the research described in this paper can, at least partly, be traced back to our experiences in the 1980's with so called expert systems, or knowledge based systems, and the underlying 'theories of uncertain reasoning, or imprecise probabilities' on which they were founded. By that time there were already numerous different approaches to the problem of producing an expert system on the basis of some knowledge base, and, not unnaturally perhaps, the builders of these systems generally praised the capabilities and accuracy of their inventions. However it seemed to us difficult in practice to objectively test the appropriateness of the methodology underlying their construction, firstly because expert systems were usually customized to one, very specific, situation with carte blanche to introduce whatever ad hockery seemed beneficial, and secondly because in any case the 'correct answers' against which to measure the systems were not known.

With this perception of the state of the art in expert system engineering at that time we considered whether it might not be possible to *objectively test* expert system methodologies, in a similar way, for example, to the way that we might test the effectiveness and accuracy of a computer algorithm or a statistical decision procedure, by determining their performance over the range, and distribution, of their expected or intended applications.

Since we were predominantly interested in expert systems which adopted various theories of 'imprecise probabilities' in an attempt to infer *probabilities* from a knowledge base consisting of probabilistic constraints, the test material we sought was a family, or more precisely a *distribution*, of probability functions (ideally ones from which we would be able to compute basic probabilities) similar in distribution and structure to those encountered by expert system builders. We shall refer to such probability functions as 'natural', and to the posited distribution of such functions as the 'natural distribution'.

Whilst evaluation of expert system methodologies was one of our main incentives for this research, the general problem of selecting a prior in situations of ignorance is one with a long history in the foundations of probability and has much wider implications within the study of 'imprecise probabilities'. In particular the assumption of an initial prior before anything is known is a central step in Bayesian Inference, although the significance of this choice is commonly dismissed through the observation that provided there is a reasonably large supply of data the ultimate influence of the prior is generally minimal. Nevertheless, even here, given that the demands of AI often require squeezing useful information from small samples (such

*This research was partially supported by a joint British Council/University of Athens Research Grant.

†This research was supported by an EPSRC Advanced Course Studentship, grant reference number 96419705.

as we are apparently capable of learning from), the prior problem is still very much of contemporary relevance.

It would not be appropriate in the present paper to enter into a detailed discussion of the history of the prior problem. We do however feel it necessary to draw attention to a key respect in which our approach to solving this problem differs fundamentally from that of some of the more influential earlier approaches. Previous justifications by logical probabilists of particular solutions to the prior problem have tended to concentrate on purely *logical* or *analytic* arguments. Thus, whatever their differences such authors as Laplace [9], Carnap [1], or Jaynes [7] or [8], have all emphasised the fundamental rôle of symmetry arguments, of which the prototype was Laplace's principle of indifference (cf. Rosenkrantz [16] for a history of the problem). Despite the obviously unsatisfactory nature of the solutions given by the logical probabilists, a rather curious feature of the debate between 'logical' and subjective Bayesians has been a lack of serious discussion of the possibility of a coherent notion of *objective* prior probability derived from considerations which *transcend the purely logical.* Our own analysis, sketched out in this paper, involves positing a cognitive model of certain aspects of the process by which random natural phenomena are presented to us. [Such a model could, at least in theory, be partly justified (or refuted) by empirical testing.] The intuition here is that the phenomena presented to us 'by nature' are logically complex in their underlying structure and that this logical complexity itself induces dispositions of prior probability. (For further discussion see [15].)

It follows that the prior(s) that we shall describe in this paper have a rather different origin from those espoused by the tradition of logical probability, being based on the idea of attempting to *model* the uncertainty that we encounter in the real world. As regards the problem of choosing a prior in the absence of any information, one could imagine a justification for our priors along the following lines: *If I knew that I would receive data from one of some, say, ten, possible probability functions, with each of which I was well acquainted, then I could certainly come up with an insightful prior based on the (ten point) distribution of these functions in some $\mathbb{D}_n$. Now of course the reality of my ignorance is that I do not know that the data will come from one of some ten probability functions. However, what in practice I might claim to know, or at least feel justified in believing, despite my apparent ignorance, is that the data I shall receive will come from some real world 'experiment', some natural probability function, it will not simply have been made up. And in this case, according to my modelling, I do have still have a prior distribution for such functions.*

In order to make the ideas behind our construction more precise we need to introduce some mathematical notation. Let $L_n = \{q_1, q_2, ..., q_n\}$ be a propositional language with propositional variables $q_1, q_2, ..., q_n$ and let SL_n be the sentences of L_n built up from these propositional variables using some set of connectives including at least $\neg, \vee, \wedge$ (with their usual meanings). Within this framework a *probability function* on SL_n is a function $w : SL_n \longrightarrow [0, 1]$ such that for all $\theta, \phi \in SL_n$,

(*i*) if $\models \theta$ then $w(\theta) = 1$,
(*ii*) if $\models \neg(\theta \wedge \phi)$ then $w(\theta \vee \phi) = w(\theta) + w(\phi)$.

All the standard properties of probability functions follow from this definition, for example for any $\theta, \phi \in SL_n$,

$$w(\theta) + w(\neg\theta) = 1,$$
$$w(\theta \vee \phi) = w(\theta) + w(\phi) - w(\theta \wedge \phi).$$

[For these and other basic facts, see, for example, [11].]

Such a probability function w is determined by its values on the *atoms* of L_n, that is on those 2^n sentences $\alpha_1, \alpha_2, ..., \alpha_{2^n}$ of L_n of the form

$$\pm q_1 \wedge \pm q_2 \wedge ... \wedge \pm q_n.$$

Indeed, since for a probability function w on SL_n we must have that

$$w(\alpha_1) + w(\alpha_2) + ... + w(\alpha_{2^n}) = 1,$$

and

$$w(\theta) = \sum_{\alpha_i \models \theta} w(\alpha_i),$$

it is straightforward to show that the correspondence between w and the 2^n-tuple

$$< w(\alpha_1), w(\alpha_2), ..., w(\alpha_{2^n}) >$$

provides a one to one correspondence, or identification, between the probability functions w on SL_n and the points in the polyhedron

$$\mathbb{D}_n = \{ < x_1, x_2, ..., x_{2^n} > \quad | \quad x_1, x_2, ..., x_{2n} \geq 0 \text{ and } \sum_{i=1}^{2^n} x_i = 1 \}.$$

The importance of this identification is that it easily allows us to say what we mean by a *probability distribution* on the set of all probability functions on SL_n, namely it is just a countably additive normalised measure on the (Borel subsets of the) polyhedron $\mathbb{D}_n$, in the sense of, say, [5] pages 30, 171. [As here, we shall endeavour to use the expression *probability function* when the domain is a set of sentences of a language and *probability distribution*, or *measure*, when the domain is the Borel subsets of a subset of Euclidean space.]

Returning now to our main theme, what we sought was a distribution $\mathcal{P}$ on $\mathbb{D}_n$ such that for a Borel subset I of $\mathbb{D}_n$, $\mathcal{P}(I)$ somehow reflected the 'probability of a random, natural, probability function w being in I'. Expressed another way this is a version of the 'Prior Problem', that is the problem of picking a prior distribution in a situation of ignorance.

It is worth re-emphasising here our intended status for $\mathcal{P}$. Although there are various alternative interpretations which one might give (see, for example, [15]), for the purpose of this paper the intention is that the likelihood that $\mathcal{P}$ assigns to any particular probability function reflects the likelihood of that probability function being encountered, in nature, via an objective, recognisably random, independently repeatable experiment. The method we describe in this paper for approaching this goal involves attempting to provide a general model of such 'random, independently repeatable experiments', and then arguing, essentially by an appeal to symmetry, about their distribution.

The reader may very well question at this point whether such a notion makes any sense at all. Without doubt we are on thin ice here, and certainly this notion of 'natural' is very much bound up with the categories we use to decribe our world. An argument however that might be advanced that this notion is not devoid of meaning is to consider a very large probability function, let us say based on all recorded medical conditions, signs, and symptoms (so these are what the propositional variables stand for) and look at all the marginals of this function on sublanguages of, say 4 propositional variables. These marginals correspond to points in the polyhedron $\mathbb{D}_4$ and by giving these points all equal probability, or measure (summing to 1), we could indeed *claim* that we have obtained such a distribution, at least on medical probability functions on SL_4.

An alternative 'argument', which we shall largely pursue in this note, is to press on regardless to produce a tentative model which generates such a distribution and then argue in favour of this model as capturing at least some of the aspects of 'natural probability functions', that is the sort of probability functions which we encounter in the real world.

There is, of course, one very natural candidate for this distribution $\mathcal{P}$ which springs to mind immediately, namely the *uniform distribution*, first proposed by Laplace, which, in effect gives all points in $\mathbb{D}_n$ equal likelihood of being encountered. More precisely $\mathcal{P}$ is just the standard normalised Lebesgue measure on $\mathbb{D}_n$. Unfortunately this choice for $\mathcal{P}$ suffers from the serious criticism that it does not *marginalise*. That is, if we assume that all probability functions on SL_n are equally likely to be encountered and then subsequently restrict our attention to sentences from the sublanguage SL_{n-1} we will not retain the uniform distribution, the probability functions on SL_{n-1} will no longer be all equally likely. Instead they will be distributed as a certain *Dirichlet distribution*. From this it follows that if we assume the uniform distribution for SL_n, and not unreasonably demand also marginalisation, then we shall have to settle for (different) Dirichlet distributions for any subsequent marginalisations (and similarly if we wish to enlarge the language). Hence if we want marginalisation and the uniform distribution in some SL_n we are forced to assume a *family* of Dirichlet distributions. Indeed this is well known to be equivalent to Carnap's λ-continuum of inductive methods for the fixed value of $\lambda = 2^n$. This is clearly unsatisfactory in the context of natural probability distributions, firstly because there is usually no clear, fixed, number of propositional variables under consideration (for example in a medical context new conditions could appear, or even disappear, at any time) so it would seem perverse for this parameter to play such a crucial role, and secondly because it would seem to be hard to justify the uniform distribution for any one particular language whilst not so doing for other languages. [For an alternative 'solution' to this problem see [13].]

A second criticism of assuming a uniform distribution (as already remarked in [12]), at least in the sort of contexts where expert systems would normally be seen as being applicable, is that for probability functions w on SL_n representative of such situations we would expect the $w(q_i)$ to be rather variable and the distinct q_i, q_j to be, at least somewhat, dependent i.e. $(w(q_i \wedge q_j) - w(q_i)w(q_j))^2$ should be relatively large. However for the uniform distribution on the w we have the expected values

$$E((w(q_i) - 1/2)^2) = \frac{1}{4(2^n+1)},$$

$$E((w(q_i \wedge q_j) - w(q_i)w(q_j))^2) = \frac{2^n}{16(2^n+3)(2^n+1)}.$$

Essentially then, for all but very small n, a random probability function w chosen according to this uniform distribution we can expect the $w(q_i)$ to be *very* close to $1/2$ and q_i, q_j to be practically independent. This is certainly not the sort of fertile ground on which expert systems germinate and thrive. As such then the uniform distribution seems to provide an inappropriate likelihood of encountering natural probability functions, at least as far as the objective testing of expert systems is concerned.

Instead of simply opting for the uniform distribution therefore we proceeded in [15], [12], [19] to develop other candidates for $\mathcal{P}$ based on modelling natural probability functions themselves. The plan of the rest of this paper is as follows. In the next section we shall describe this modelling for *univariate* natural probability functions, that is where $n = 1$ and the language L_n has just a single propositional variable. This required us to make one key assumption about the general structure of such functions. In the following section we shall describe how this modelling was extended to cover *multivariate* natural probability functions. To achieve this required making a second key assumption as to how these arise and how correllations between the individual propositional variables originate. Throughout we shall avoid going too deeply into the technical mathematical details. The interested reader may find, up to straightforward generalizations, proofs of all the results we state in [15], [12], [19].

2 Univariate natural probability functions

In this section we restrict our attention to the case where the language L_n has just one propositional variable (i.e. $n = 1$), which we shall denote by q. Of course this case is hardly directly relevant to the issue of objectively testing expert system methodologies, where we would expect n to be relatively large. However treating the univariate case is an essential first step towards the multivariate case.

Notice that in the univariate case the atoms of L_1 are just q and $\neg q$ so, since $w(\neg q) = 1 - w(q)$ for w a probability function on SL_1, we can more simply identify w with the point $w(q) \in [0, 1]$, rather than the vector $< w(q), w(\neg q) > \in \mathbb{D}_1$. With this revision then our goal was to develop a natural probability distribution $\mathcal{P}$ on the real interval $[0, 1]$.

It is interesting to consider at this stage what we might expect $\mathcal{P}$, or more precisely, successively finer histograms of $\mathcal{P}$ to look like. Three such features seem, to our experience, evident. The first is symmetry about $1/2$. In practice whether or not we have choosen to denote a particular feature, or its negation, by any one particular propositional variable is, one feels, entirely contingent, and in consequence $\mathcal{P}$ should be invariant under renaming of propositional variables and transpositions of a propositional variable and its negation.

A second property of $\mathcal{P}$ that experience might lead us to anticipate is that its histogram(s) should rise rapidly around zero and one. The reason for this is that, in our everyday lives, a seemingly disproportunate number of the probabilities we encounter are close to 0 or 1. We expect the train to run, we expect three leafed clovers, we expect not to win the lottery. Indeed the importance attached to default logic, the study of reasoning with statements which are *usually* true, would seem to confirm the ubiquity of such knowledge in our everyday world.

Finally, a third feature of everyday probabilities that we might expect to be reflected in $\mathcal{P}$ is that there seems also to be a clustering of probabilities around $1/2$, for example the toss of a coin, the sex of a baby. One partial explanation for this might be that linguistic categories that divide the population roughly in half are descriptively more efficient, although it is hard to see that that can really account for the two examples cited above! [Another possible explanation for such clustering is given in [15].]

Putting together these three impressions then we might anticipate that a histogram of such a $\mathcal{P}$ should be symmetric about $1/2$, exhibit a bump at $1/2$ and sharp rises at 0 and 1.

Returning now to our modelling, notice that in this case where $n = 1$ we can think of a natural probability function w as a actual process which randomly outputs 1's (corresponding to q being true) and 0's (corresponding to q being false, i.e. $\neg q$ being true) with probability $w(q)$ of outputing 1 (so $w(q)$ is also the expected value of the output). Armed with this picture the key idea, or assumption, we adopted at this point is that, in general, natural 0-1 random processes in our everday macroscopic world are actually very complicated affairs in which any genuine randomness is hidden at a very deep level. For example the sex of a baby is not simply decided 'randomly' at the point at which the baby first enters the light of day, or is scanned[1]. The sex of the baby is (largely) determined when sperm meets ovum, but the crucial issue of which sperm meets the ovum first is the main influencing source of randomness and this lies at a deeper level still. Even in the proverbial coin toss the randomness is hidden deep, determined by the inner chemistry of the tosser's thumb muscles, the detailed physical contours of the surface it lands on, etc, etc., and does not simply 'happen' at the point at which we first observe the outcome.

With such a picture in mind we set about trying to

[1] For this paper we adopt this viewpoint, which we take to be widely acceptable.

provide a general model of such processes. [This was not our only 'picture', see [15].] The model of a natural random 0-1 process that we alighted upon was as the truth value of a very large sentence θ of the propositional calculus where the only randomness is at the level of the propositional variables, which themselves were randomly, and independently, assigned truth values 1 (i.e. *true*) or 0 (i.e. *false*).

To give a toy example of what we mean here, if θ is the sentence

$$(\neg p_2 \vee (p_1 \wedge p_3)) \vee \neg p_1$$

and the propositional variables p_1, p_2, p_3 are independently distributed with expected (truth) values 1/2, 1/3, 2/3 respectively (so that the atom $p_1 \wedge \neg p_2 \wedge p_3$ has expected truth value $1/2 \cdot (1 - 1/3) \cdot 2/3 = 2/9$ etc) then a straightforward calculation based on the observation that θ is true just if one of the incompatable $\neg p_1$, $p_1 \wedge \neg p_2$, $p_1 \wedge p_2 \wedge p_3$ are true shows that θ has expected value 17/18.

This choice of model for a natural 0-1 random process is clearly only one amongst many possible models, and in part reflects our own familiarity with logical calculi and our predeliction to explain the world in logical terms. Nevertheless we would suspect that somewhat similar conclusions to those we have obtained would hold for a wide range of similarly motivated models, although these remain to be investigated.

Having made our choice of the basic model numerous other choices now present themselves. Firstly how should we choose the distribution of the expected values of the propositional variables? Isn't this surely the very problem our model was intended to answer for us?! We shall return to this question later, but for the present let us suppose for simplicity that the expected truth values of the propositional variables are all 1/2. In other words the distribution, or measure, here just gives measure 1 to the single point 1/2 (more correctly to the singleton subset $\{\frac{1}{2}\}$ of $[0,1]$) and no measure anywhere else.

A second choice is which connectives to allow in our sentences. In [15] we chose to include only the connectives $\neg, \vee, \wedge$. This led to some surprising consequences, which we shall describe later, and which suggested this was an inappropriate choice given the goal we had in mind. In a subsequent paper, [12], we expanded this list of connectives to $\neg, \vee, \wedge, \leftrightarrow, \updownarrow$, where $\updownarrow$ is the dual of $\leftrightarrow$, that is $\theta \updownarrow \phi$ is true just if θ, ϕ have different truth values. With hindsight this choice again seemed not perhaps the most appropriate so in [19] the results were generalised to cover various combinations of binary connectives. From these many choices the most justified, in our opinion, and the one which we shall make for this paper, is the set of all binary connectives which genuinely depend on both arguments[2]. For future reference let us denote the set of such connectives by $\mathcal{C}$. The reason for excluding the remaining binary connectives is that they would be either essentially redundant in the random processes we have in mind, or even, in the case of the identically true or identically false, connectives, have the power to remove the randomness altogether. The justification for not going beyond binary to ternary and so on, is based on our intuition that in nature basic interactions almost invariably take place between pairs of interagents.

Finally we need to resolve the question of what measure of size we are going to put on a sentence θ in order to say what we mean by 'very large'. Here there is an obvious solution. For the language L with propositional variables $p_1, p_2, p_3, \ldots$ and set of sentences SL of L built up using just the binary connectives from $\mathcal{C}$ the size of a sentence $\theta \in SL$ will be measured by the number of connectives occuring in θ. In fact however, our conclusions appear robust under sensible perturbations of this definition; see for example lemma 2.7 of [15].

We assume that, as models of 0-1 random processes, all sentences of a fixed size are equally likely to be encountered. This may be regarded as another form of the principle of indifference, although applied to a different 'model' to that of Laplace. [See [15] for further discussion.]

To sum up then, our initial model of a natural random 0-1 process is the truth value of a random, very large sentence $\theta \in SL$ when the truth values of the propositional variables are randomly, and independently, assigned truth values 1 (i.e. *true*) or 0 (i.e. *false*) with expected value 1/2.

Carrying through the intention described above this modelling now yields a family D_n of probability distributions on $[0,1]$ defined as follows: For I a Borel subset of $[0,1]$,

$$D_n(I) = Lim_{m\to\infty}\left[\frac{|\{\ \theta \in A_n^m \mid E(\theta) \in I\ \}|}{|A_n^m|}\right],$$

where A_n^m is the set of sentences of SL containing exactly n connectives and only mentioning propositional variables p_i for $1 \leq i \leq m$, and $E(\theta)$ is the expected value of θ when the truth value of each proposition variable is identically and independently distributed with expected value 1/2.

Our only remaining immediate concern now is what we mean by n being 'large'. The central result which

[2]Notice that, effectively, negation is still present in the language, albeit now as a derived, or defined, connective.

resolves this question is that for I a Borel subset of $[0,1]$, the weak limit, $D(I)$, as $n \to \infty$ of the $D_n(I)$ exists and defines a countably additive measure on $[0,1]$. [This limit is rather robust in that there is considerable flexibility in the definition of D. We could, for example, allow n and m to tend to infinity independently, and/or replacing A_n^m by, say, the set of sentences with *at most* n connectives.]

This measure D could now be taken as our natural distribution of the expected values of 0-1 random processes, equivalently probability functions on SL_1. Certainly it has the right shape, as discussed earlier, in that its histograms rises at 0 and 1 and have a discernible bump at 1/2.

One major criticism however, which we have already noted, is that D has been constructed under the assumption that the propositional variables all have expected truth value 1/2. Fortunately there seems now to be a self evident way to address this deficiency. Namely, since D currently represents our best attempt at a natural probability distribution, we should 'improve' our original assumption that all the propositional variables had expected value 1/2 by assuming instead that all the propositional variables have expected values distributed acording to D.

If we do this, and repeat the whole process again we find that the weak limit, $D^{(2)}$ say, again exists and is a countably additive measure with the the 'right shape'. But now our best natural distribution is $D^{(2)}$ so we should start over again with $D^{(2)}$ in place of D to obtain $D^{(3)}$ and so on and so on. In this way we obtain a sequence of better and better candidates, $D, D^{(2)}, D^{(3)}, D^{(4)}, \ldots$ to our 'natural probability distribution' and these again have a weak limit which we denote as J (after Jítka Vilímová).

J is, within this modelling of (univariate) natural probability functions, our sought after 'natural probability distribution'. It is continuous, and so countably additive, and again has the 'right shape', see Figure 1 for the relative frequency histogram approximation computed using the method of Bernstein polynomials based on the first hundred moments.

It is, furthermore, impervious to further improvement in the sense that if we again repeat the process that yielded $D^{(2)}$ from D but with J in place of D then we get back J again. But more spectacularly, if we repeat the whole process which yielded J not from the initial distribution on the expected values of the propositional variables which gave them all value 1/2 but with this distribution replaced by *any* symmetric (about 1/2) countably additive measure T then, with just one exception, we will again obtain J. That one exception is when T is the measure U which spreads

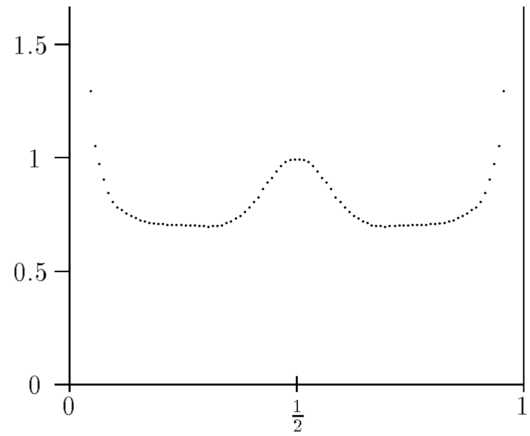

Figure 1:

all measure equally on the points 0 and 1. In that case there is no randomness about the initial propositional variables and the process just gives back U again.

It is interesting to compare this result with the conclusion we arrived at in [15] where we took instead only the binary connectives which, with arguments p,q, have the truth tables of

$p \wedge q,\ p \wedge \neg q,\ \neg p \wedge q,\ \neg p \wedge \neg q,\ p \vee q,\ p \vee \neg q,\ \neg p \vee q,\ \neg p \vee \neg q.$

[This is equivalent to taking sentences built up from $\{\pm p_i \mid i = 1,2,3,\ldots\}$ using only the connectives $\wedge, \vee$ which is how it was actually done in [15].] In that case the final weak limit of the $D, D^{(2)}, D^{(3)}, D^{(4)}, \ldots$ again exists, and is U! In other words, in the limit all randomness has disappeared! Furthermore we obtain U no matter what countably additive measure T we start from.

As *the* choice for the natural univariate prior under ignorance J has some very pleasant properties, as we shall see in the next section. Furthermore we 'know' what J is in the sense that we know all its moments. Precisely, the expected values with respect to J are given recursively by $E(x^0) = 1$,

$$\begin{aligned} E(x^k) &= \frac{2}{5}E(x^k)^2 + \sum_{r=0}^{k} \frac{2}{5}(-1)^r E(x^r)^2 \binom{k}{r} \\ &\quad + \sum_{r=0}^{k} \frac{1}{5} E(x^r(1-x)^{k-r})^2 \binom{k}{r}. \end{aligned}$$

We conclude this section by mentioning an alternative justification for the choice of J as a natural prior probability distribution: Consider some natural probability function (on SL_1), or equivalently some random 0-1 process in the real world. As we have already argued our experience of such natural probability functions is that this process is ultimately a very complex combination of other processes and that any true randomness is hidden deep down at the microscopic level. Now it seems reasonable to suppose that on closer inspection this process will be seen to be the result of a simple combination of some few other natural processes, where, just as in the initial process, the true randomness is again hidden deep down at the microscopic level. Clearly then, if we suppose that our initial natural probability function was distributed according to a natural prior K then we would seem obliged to afford the same status to the these 'few other natural probability functions' whose simple combination yielded our initial function. If we now further agree that the possible 'simple combinations' are just those in $\mathcal{C}$ (and each of these are equally likely here) this imposes a fixed point condition on K - whose only solutions are J and U! [See [12], Theorem 13.] Discounting U for reasons already given leads us again then to the distribution J.

3 Multivariate natural probability functions

In the previous sections we derived the prior probability distribution J on the probability functions on SL_1 (where $L_n = \{q_1, ..., q_n\}$). In short the justification for considering J as a prior on natural probability functions was based on the assumptions that randomness in nature is the result of a very deep combination of 'microscopic' random events in processes and that such processes can be adequately modelled as very large sentences in SL.

In this section we turn our attention to extending J from natural probability functions on SL_1 to natural probability functions on SL_n, that is those probability functions on SL_n which we encounter in the real world and which knowledge engineers endeavour to approximate via expert systems.

The particular feature of such probability functions on SL_n is that the propositional variables q_i correspond to natural, and rather interdependent, features. For example the relevant features a doctor might use in diagnosing, say, types of tumours. Indeed the existence of such interdependencies is precisely what gives expert systems their hopes of success. In order to extend J we needed to somehow model these interrelationships.

It was here that we made our second key assumption. We assumed that the dependencies we find between such observable features arise because these features are themselves combinations of certain other 'basic', independent, features and that the dependencies between the observable features arise as a result of having these 'basic' features in common.

In consequence of the conclusions stated in the previous section these 'basic' features were assumed to be (independently) distributed according to J whilst the observable features were assumed to correspond to sentences built up from these basic features, using again the connectives in $\mathcal{C}$.

More precisely we assumed that to a 'natural' probability function, w say, on SL_n there corresponded sentences $\theta_1, ..., \theta_n \in B^k_m$, where B^k_m is the set of sentences built up from the propositional variables $p_1, p_2, p_3, ..., p_m$ *without repetions* and using exactly k occurrences of connectives from $\mathcal{C}$ (here k, m are fixed parameters), such that,

$$\begin{aligned} w(q_i) &= \textit{Expected truth value of } \theta_i \\ &= \textit{Probability that } \theta_i \textit{ is true.} \end{aligned}$$

In particular then, for an atom $\bigwedge_{i=1}^n \pm q_i$ of SL_n,

$$\begin{aligned} w(\bigwedge_{i=1}^n \pm q_i) &= \textit{Expected truth value of } \bigwedge_{i=1}^n \pm\theta_i \\ &= \textit{Probability that } \bigwedge_{i=1}^n \pm\theta_i \textit{ is true,} \end{aligned}$$

when the truth values of the p_i are themselves determined by some random, natural, 0-1 process.

The natural distribution, or measure, J_n on these probability function was then defined by assuming that for such w all the (finitely many) n-tuples $< \theta_1, \theta_2, ..., \theta_n >$ of sentences in B^k_m were equally likely to occur and that the expected truth values of the p_i, $i = 1, 2, ..., m$, were independently and identically distributed according to the distribution J. There is thus a continuous aspect to J_n, corresponding to the expected truth values of the propositional variables p_i, and a discrete aspect corresponding to the particular combinations θ_j of these propositional variables.

There are clearly a number of assumptions here that deserve comment. Firstly, the choice to identify the combinations of 'basic' independent features with sentences using connectives in $\mathcal{C}$ is just the natural extension of the same idea which was developed in order to derive J. The choice of taking the same, *fixed*, k and m for all the θ was however made partly on pragmatic grounds, to allow the consequent mathematical analysis to run more smoothly, and transparently, and partly because these provide natural parameters with which to tweak the resulting distribution. (Strictly these parameters should have appeared as additional arguments in J_n but we suppress them simple to avoid a surfeit of subscripts and superscripts.)

As far as this parametric aspect is concerned, for fixed k having m small has the effect of making the correlations between the q_i more variable. The effect of k is rather more subtle, basically having k large means that overall there are more dependencies between the q_i. Having such parameters seems an unavoidable part of what we are attempting here in the distribution J_n. For if we were to consider the sort of natural probability functions which would be encountered as suitable cases for expert systems then these clearly have much stronger interdependencies between the q_i than when we considered natural probability functions where the q_i were just randomly chosen features from the natural world. [Indeed in some approaches, for example Walley's theory of imprecise probabilities based on upper and lower probability, see [17], [18], admitting a range of parameters here may be positively advantageous in that it accommodates also a degree of ignorance as to the actual degree and nature of interdependence between the variables present.]

A second, and more substantial, point is why we chose sentences *without repeated propositional variables.* Again it must be acknowledged that in part this choice was made both to simplify the ensuing mathematics and because it *gives the right answers*! However there is an argument for our choice which touches on a rather fundamental problem in regard to the goal of constructing a 'natural probability distribution'. For suppose we had allowed our sentences to have contained repeated propositional variables. In such a case we would have been allowing the definite possibility of choosing θ_i to be a tautology (or contradiction) in which case q_i would not be random at all. Now in the context of constructing expert systems we (apparently) never do this. That is, we do not include amongst our variables features which are actually not variable at all. In this sense then we preselect our features, and our choice of sentences should reflect this preselection. Of course not *all* sentences with repeated propositional variables are tautologies or contradictions. Nevertheless we could see no natural, and workable, compromise which allows repeats whilst avoiding the problem of tautologies and contradictions. Our choice then of modelling, or approximating, this preselection by the cavalier barring of repeated variables is clearly not ideal (at least in the absence of further justification), but was one we made nevertheless in order to facilitate further progress.

We should point out here that, in effect, preselection of 'random processes' was already implicit in our development of J since our model only covered processes which genuinely had an element of randomness. [It might be conjectured that had we allowed into J entirely non-random processes then our subsequent difficulties with the multivariate extensions of J would vanish. Unfortunately however this appears not to be the case.]

With the above definition in place it turns out that the J_n have many of the properties we would have hoped for. Firstly the J_n marginalise. That is if we take $m > n$ and a Borel subset A of $\mathbb{D}_n$ then

$$J_n(A) = J_m(\{w \in \mathbb{D}_m | w \text{ restricted to } SL_n \text{ is in } A\}).$$

Secondly, the J_n satisfy *weak renaming.* That is, if σ is a permutation of the literals $\pm q_1, \pm q_2, ..., \pm q_n$ of L_n such that if $\sigma(q_i) = \pm q_j$ then $\sigma(\neg q_i) = \mp q_j$ and we extend σ to a bijection of SL_n in the obvious way (i.e. replacing $\pm q_i$ everywhere by $\sigma(\pm q_i)$) then for a Borel subset A of $\mathbb{D}_n$,

$$J_n(A) = J_n(\{\ w\sigma \mid w \in A\ \}).$$

This is a clearly desirable property in this context because, informally, it amounts to saying that the likelihood of encountering a particular probability function on SL_n does not depend on which names (i.e. $\pm q_i$) we had used to denote the features in question.

On the other hand the stronger property of *full renaming* does not hold for the J_n. Full renaming here would amount to saying that for any Borel subset A of $\mathbb{D}_n$ and permutation τ of $1, 2, ..., 2^n$,

$$J_n(A) = J_n(\{\vec{x}_\tau | \vec{x} \in A\}$$

where, for $\vec{x} =< x_1, x_2, ..., x_{2^n} >$,

$$\vec{x}_\tau =< x_{\tau(1)}, x_{\tau(2)}, ..., x_{\tau(2^n)} > .$$

It is not difficult to see that weak renaming can also be expressed in a similar way, but for a restricted class of permutations τ. On the face of it then full renaming, which does hold for Dirichlet Priors, might appear to be a very desirable symmetry condition. However there are consequences of this property for *induction*, see for example [6], which would cause us to question this desirability in distributions with the pretentions of these J_n. In consequence we actually view the failure of this principle with some satisfaction! [In [18] it is suggested that 'recategorisation' is also a desirable property in the general context of priors under ignorance. Within the framework of this paper, however, the conditions under which recategorisation applies would themselves be enough to invalidate the assumption of ignorance. For a detailed explanation of this point see [14].]

A third property of the J_n which in our view is essential is that $J_1 = J$. It is interesting to note that this is independent of the particular choice of k and m. Clearly this is an important result as far as our modelling of natural distributions is concerned. For it would have been unfortunate, to say the least, if having settled on J as our prior distribution on natural probability functions on SL_1 we were then to argue for a prior on SL_n which was different in the case $n = 1$. [It is primarily this property which rests on our choice of 'unrepeated variables' discussed earlier.]

We now turn to the question of the variability of the $P(q_i) - 1/2)^2$ and $(P(q_i \wedge q_j) - P(q_i)P(q_j))^2$ for P a probability function on SL_n random according to J_n. Recall that the figures given earlier for these quantities for the uniform distribution were argued to be too low, certainly in the context of P representing the underlying probabilities in a specialised area such as expert systems frequently attempt to approximate. In the case of J_n, for random P on SL_n,

$$E((P(q_i) - 1/2)^2) = 1/8$$

whilst the values of $E((P(q_i \wedge q_j) - P(q_i)P(q_j))^2)$ $(i \neq j)$ are given for some small sample values of k and m by:

m	$k = 0$	$k = 1$	$k = 3$	$k = 5$
1	0.02500			
2	0.01250	0.01255		
3	0.008333	0.005652		
4	0.006250	0.003561	0.005198	
5	0.005000	0.0025776	0.001749	
6	0.004167	0.002012	0.0009597	0.0031175
8	0.003125	0.001393	0.0005142	0.0003562
11	0.002273	0.0009497	0.0003289	0.0001142

By comparison for the marginalising Dirichlet priors corresponding to a particular value of λ (for Carnap's λ-continuum) the value of $E((P(q_i \wedge q_j) - P(q_i)P(q_j))^2)$ has a maximum of 0.00837 when $\lambda = \sqrt{3}$ (see [10]).

The corresponding values given by the uniform distribution on probability functions on SL_n (i.e. when $\lambda = 2^n$) are:

$n = 1$	$n = 2$	$n = 3$	$n = 4$
0.008333	0.007143	0.005051	0.003096

$n = 6$	$n = 8$	$n = 10$	$n = 12$
0.0009185	0.0002432	0.00006080	0.00001524

What sort of meaningful conclusions can be drawn from these figures we shall not consider further here, except to say that they clearly do confirm our expectation that for small values of m and k propositional variables distributed according to the J_n are more interdependent than in the family of uniform distributions, at least for moderately large n, and may therefore provide (in addition to their explanatory power) a better general model of such specialised areas as expert systems frequently endeavour to approximate.

One final pleasing property possessed by J_n (for $n > 1$) is that, as priors, they give non-zero probability to $P(\alpha) = 0$ for α an atom of SL_n, and in consequence may give non-zero probability to non-tautological universal sentences of SL_n. The failure of the Dirichlet priors (i.e. Carnap's continuum of inductive methods) in this regard has sometimes been viewed as undesirable. For an interesting discussion of this property of Dirichlet priors with respect to universal sentences, called *dogmatism* by Gaifman and Snir in [3], see section 3 of [4]. The fact that J_1 is in this sense also dogmatic, is, we would argue, an excuseable consequence of 'preselection'.

Having so readily pointed out what we view as the J_n's good features we should at least mention a minor problematic aspect to actually using these distributions at present. Namely, apart from the case $n = 1$, we currently know of no *simple* general formulae for their moments.

4 Conclusion

The work described in this note demonstrates, we believe, that the distribution J and its multivariate extensions J_n have a number of desirable properties as candidates for 'natural' priors, not least that they are based on a clear *model* of random processes in the real world.

Of course, this model depends on several debateable assumptions, in particular:

1. That large sentences with the randomness hidden deep down provided adequate models for natural 0-1 random processes.

2. That, as models of 0-1 random processes, all sentences with the same number of connectives (from $\mathcal{C}$) are equally likely to be encountered.

3. That the dependencies we find between observable features in nature arise because those features are themselves certain simple Boolean combinations (each with equal numbers of connectives and without repeated variables) of overlapping 'basic' features.

4. That the problem of 'preselection' of observable features is adequately addressed by the restriction to non-overlapping variables referred to in 3 above.

On the other hand since the very existence of the object we seek, i.e. the distribution of natural probability functions, has not been proven, making some speculative assumptions seems unavoidable. And at least, having made these assumptions, we could argue that the model gains some credibility through possessing many of the properties we might have expected, or hoped for, in a distribution on the natural probability functions.

Whether or not these factors alone are enough to conclude that 'the distribution of natural probability functions', has some meaning, and may even have been approached, clearly remains problematic. Perhaps at this stage the best interpretation we can put on these results is that they provide a model which *could* be correct in *some* universe. As a final remark, we point out that in such a universe there would be no need to explain, or seek the ultimate source of, random phenomena. Randomness would have no beginning.

Acknowledgements

We would like to thank Alena Vencovská for her contribution to the research on which this paper is based, to Costas Dimitracopoulos for his encouragement and practical support and to Peter Walley and the two referees for their interest and advice.

References

[1] R. Carnap. *The Continuum of Inductive Methods,* University of Chicago Press, Chicago, 1952.

[2] W. Feller. *An Introduction to Probability Theory and its Applications, Vol.II*, 2'nd edition, John Wiley, New York, pp251, 1957.

[3] H. Gaifman and M. Snir. Probabilities over Rich Languages, Testing and Randomness. *Journal of Symbolic Logic,* 47:495-548, 1982.

[4] H. Gaifman. *Towards a Unified Concept of Probability.* Logic, Methodology and Philosophy of Science VII, eds. Barcan Marcus et alii, Elsevier, pp319-350, 1986.

[5] P. R. Halmos. *Measure Theory.* Van Nostrand, 1950.

[6] M. J. Hill. *Some Aspects of Induction and the Principle of Duality.* Ph.D. thesis, Manchester University, submitted January 1999.

[7] E. T. Jaynes. The Well-Posed Problem. *Found. Phys.,* 3:477-493, 1973.

[8] E. T. Jaynes. Where do we stand on maximum entropy?. *The Maximum Entropy Formalisation*, R. D. Levine and M. Tribus, eds., MIT Press, Cambridge, MA, 1978.

[9] P. S. Laplace. *Essai Philosophique sur les Probabilités.* Bachelier, Paris, 1840.

[10] J. Lawry. *Natural Distributions in Inexact Reasoning.* Ph.D thesis, Manchester University, 1994.

[11] J. B. Paris. *The Uncertain Reasoner's Companion - A Mathematical Perspective.* Cambridge University Press, 1994.

[12] J. B. Paris, C. Dimitracopoulos, A. Vencovská and G. M. Wilmers. A Multivariate Natural Prior Probability Distribution Based on the Propositional Calculus. *Technical Report of the Manchester Centre for Pure Mathematics*, Department of Mathematics, University of Manchester Institute of Science and Technology. 1998.

[13] J. B. Paris and A. Vencovská. A Method of Updating that Justifies Minimum Cross Entropy. *International Journal of Approximate Reasoning,* 7:1-8, 1992.

[14] J. B. Paris and A. Vencovská. In Defence of the Maximum Entropy Inference Process. *International Journal of Approximate Reasoning,* 17:77-103, 1997.

[15] J. B. Paris, A. Vencovská, and G. M. Wilmers. A Natural Prior Probability Distribution Derived from the Propositional Calculus. *Annals of Pure and Applied Logic*, 70:243-285, 1994.

[16] R. D. Rosenkrantz. *Foundations and Applications of Inductive Probability.* Ridgeview Publ. Co., 1981.

[17] P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London, 1991.

[18] P. Walley. Inferences from Multinomial Data: Learning about a Bag of Marbles. *J. Royal Statistical Society, Series B*, 58:3-57, 1996.

[19] P. N. Watton. *Natural Prior Probability Distributions Derived from the Propositional Calculus.* MSc dissertation, Manchester University, 1997.