Inductive Influence

Jon Williamson

Draft of March 14, 2006

Abstract

Objective Bayesianism has been criticised for not allowing learning from experience: it is claimed that an agent must give degree of belief $\frac{1}{2}$ to the next raven being black, however many other black ravens have been observed. I argue that this objection can be overcome by appealing to *objective Bayesian nets*, a formalism for representing objective Bayesian degrees of belief. Under this account, previous observations exert an *inductive influence* on the next observation. I show how this approach can be used to capture the Johnson-Carnap continuum of inductive methods, as well as the Nix-Paris continuum, and show how inductive influence can be measured.

Contents

§1 Introduction

To what extent should I believe it will rain tomorrow? Objective Bayesianism is a theory which puts forward precise answers to questions like this.[1] In common with other Bayesians, objective Bayesians argue that an agent's degrees of belief should be probabilities. But objective Bayesians go further by isolating a single probability function as a candidate for an agent's degrees of belief.[2] This probability function is objectively determined by the extent of the agent's background knowledge.

Background knowledge isolates the most appropriate probability function in two ways. First, the agent's degrees of belief should make the commitments that are warranted by her background knowledge: those probability functions that do not satisfy constraints imposed by background knowledge should be eliminated from consideration. Knowledge of long-run frequencies, for instance, constrains degrees of belief. Thus if the agent knows only that $\mathit{freq}_a(B(a)) = x$ where the frequency is found by repeatedly sampling individuals a, then the agent should set degree of belief $p(B(a_1)) = x$. Probability functions that do not satisfy this constraint should be disregarded.

Second, the agent should not believe things to a greater or lesser extent than is warranted by background knowledge: the agent should select a probability function, from all those remaining, that embodies the most middling degrees of belief, those furthest from the extremes of 0 and 1.[3] Distance from the extremes is measured by entropy $H = -\sum_{\omega\in\Omega} p(\omega)\log p(\omega)$; hence the Maximum Entropy Principle: an agent should adopt as her belief function, from all the probability functions that satisfy constraints imposed by background knowledge, that which has maximum entropy.

Objective Bayesianism faces a number of challenges,[4] not least the charge that learning from experience becomes impossible on the objective Bayesian account (§2). In §3 I shall argue that this charge is a mistake, attributable to a misapplication at the first stage of the objective Bayesian method: the constraints imposed by background knowledge have not been correctly assessed. In order to elucidate these constraints I introduce the machinery of *objective Bayesian nets* in §4. These nets offer a way of representing maximum entropy probability functions that renders probabilistic dependence and independence relationships perspicuous. They are useful here, I claim, because when learning from experience past observations exert an *inductive influence*—a type of dependence relationship—on future observations (§5).

When objective Bayesian nets are applied to the problem of learning from experience, the resulting formalism yields the Johnson-Carnap continuum of inductive methods as a natural special case (§6). In §7 we see that the Nix-Paris continuum of inductive methods emerges as another special case—though arguably a less central special case. The question now arises as to which point in the Johnson-Carnap continuum yields the most appropriate inductive method

[1](Rosenkrantz, 1977; Jaynes, 2003)

[2]I will only be considering finite probability spaces in this paper. The extension of objective Bayesianism to the infinite case is steeped in controversy and arguably proceeds at the expense of uniqueness of the most appropriate probability function—see Williamson (2006b, §19).

[3](Williamson, 2006a)

[4](Williamson, 2006b, Part III)

from the objective Bayesian perspective. In §8 I reject the idea that the classification efficiency of the agent's language might provide the answer to this question. Instead in §9 I show how frequency considerations can be used to isolate the optimal inductive method.

§2
The Problem

Consider an agent whose language contains a large number of variables B_1, $B_2, \ldots, B_k$ each of which takes one of two possible values, *true* or *false*. We shall write b_n^1 for the assignment $B_n = \mathit{true}$ and b_n^0 for $B_n = \mathit{false}$.

This agent, we shall suppose, has no background knowledge that links these variables. In that case, the maximum entropy principle will yield a probability function that gives each outcome the same probability and that renders all variables probabilistically independent:

$$p(b_n^1) = p(b_n^0) = p(b_{101}^1 \mid b_1^1 \cdots b_{100}^1) = 1/2.$$

But this can seem counter-intuitive. Suppose $B_n = \mathit{true}$ if and only if the n'th raven to be observed is black. Then $p(b_{101}^1|b_1^1 \cdots b_{100}^1) = 1/2 = p(b_1^1)$ represents a failure to learn from experience: an agent who observes a hundred ravens, all black, should not give any more credence to the next raven being black than she did before collecting this evidence.

Many have argued that this failure to learn from experience reveals a flaw in objective Bayesianism, and that the Maximum Entropy Principle should be duly rejected.[5]

§3
Diagnosis

The inference of §2—that independence in the face of ignorance leads to a failure to learn from experience—is, I claim, too hasty. True, the Maximum Entropy Principle does yield probabilistic independence when the agent has no background knowledge. But in the learning problem there is background knowledge that has not been taken into account by the above analysis. When this knowledge is taken into account, the conclusion does not follow: there is, after all, no problematic failure to learn from experience.

What is this background knowledge that has been overlooked? In our raven example it is the explicit supposition: $B_n = \mathit{true}$ if and only if the n'th raven to be observed is black. It is known that *the variables are all instantiations of the same predicate.* In fact in the raven example B_n is better written $B(a_n)$, where a_n denotes the n'th observed raven and B is the predicate 'is black'. So the variables have predicate B in common—there is a known connection between all the variables.

[5] For instance, Gillies (2000, pp. 45–46); Howson and Urbach (1989, pp. 65–66) and Earman (1992, p. 17) criticise the Principle of Indifference, which is a special case of the Maximum Entropy Principle, on the basis of the problem of learning from experience. See also Paris (1994, p. 178).

Herein lies a difficulty: this type of qualitative background knowledge tends to be overlooked when the Maximum Entropy Principle is applied. The reason for this is the following. The first step of the objective Bayesian method—eliminating from consideration those probability functions that do not satisfy constraints imposed by background knowledge—requires some procedure for deciding whether a probability function is compatible with background knowledge. If background knowledge consists of a set of quantitative constraints on degrees of belief then we can test to see the probability function satisfies those constraints. But if the knowledge is qualitative, as it is in the learning case, it is hard to see exactly what constraints such knowledge imposes. One of the challenges for objective Bayesianism is to clarify the ways in which qualitative knowledge constrains degrees of belief.[6]

So there is qualitative background knowledge that has not been taken into account. If we are to resolve the problem we must somehow convert this qualitative knowledge into quantitative constraints on degrees of belief. To do that we shall need to apply the machinery of *objective Bayesian nets*.[7]

§4
Objective Bayesian Nets

An objective Bayesian net is a representational tool. It is a way of representing the degrees of belief that an agent should adopt under the objective Bayesian account. In this section I shall briefly sketch the key aspects of objective Bayesian nets in this section—see Williamson (2005b) for a fuller account.[8] In §5 we shall see how objective Bayesian nets can be applied to the problem at hand, learning from experience.

A *Bayesian net* consists of a directed acyclic graph whose nodes are variables, together with the probability distribution of each variable conditional on its parents in the graph. Assuming the *Markov condition*, which says that each variable is probabilistically independent of its non-descendants conditional on its parents, the graph and conditional distributions suffice to determine the joint probability distribution over all the variables in the graph. A Bayesian net is a good representational device because it is relatively compact (if the graph is sparse then relatively few probabilities need to be specified to determine the joint distribution) and because it perspicuously represents the independencies that the probability function satisfies.[9]

As we have seen, according to objective Bayesianism an agent should adopt as her belief function the probability function, from those compatible with her background knowledge, that has maximum entropy. An *objective Bayesian net* is just a Bayesian net representation of this entropy-maximising probability function.

Given a set of quantitative constraints involving the variables in an agent's language, an objective Bayesian net can be constructed as follows. First construct an undirected *constraint graph* by taking the variables as nodes and link-

[6](Williamson, 2006b, §18)

[7]Note that Paris (1994, pp. 198–199) offers a similar diagnosis but a different resolution. See also Paris and Vencovská (2003).

[8]See also Williamson (2002) and Williamson (2005a, §§5.6–5.7).

[9](Pearl, 1988; Neapolitan, 1990)

Figure 1: No background knowledge.

ing two variables by an edge if they occur in the same constraint. The following key property holds: if a set Z of variables separates sets X and Y of variables in the graph, then the maximum entropy probability function renders X and Y probabilistically independent conditional on Z, written $X \perp\!\!\!\perp Y \mid Z$. Given this property, one can easily transform this graph into a directed acyclic graph for which the Markov condition holds. Finally, one can maximise entropy to find the probability distribution of each variable conditional on its parents in the graph. This yields an objective Bayesian net.

Certain types of qualitative information can be handled as follows. A relation R is an *influence relation* if learning of new variables that are known not to be R-influences of current variables provides no reason to change one's degrees of belief concerning the current variables. For example causality is an influence relation: while learning of a new common cause can lead one to render two variables more probabilistically dependent than they were, learning of non-causes provides no reason to change one's degrees of belief.[10] Knowledge of qualitative influence relationships can be converted into quantitative constraints on degrees of belief: for sets of variables U and V, if $U \subseteq V$ is closed under knowledge of influences (i.e. any variable in V that is not ruled out as an influence of a variable in U is itself in U), and an agent has no quantitative information that rules otherwise, then $p^V_{\beta_V \downharpoonright U} = p^U_{\beta_U}$: the belief function $p^V_{\beta_V}$ on V formed from full background knowledge β_V should, when restricted to U, match the belief function $p^U_{\beta_U}$ on U formed from the background knowledge β_U that pertains to U. One special case will be important for our purposes: if the agent possesses full knowledge of influences and some knowledge concerning their strengths then the graph in her objective Bayesian net will just be the influence graph—the graph whose arrows depict the direct influence relationships. Moreover in this case the probability distributions in the net can be determined iteratively: first find the probability distribution of the root variables by maximising entropy, then find those of their children, then their grandchildren, and so on. This iterative approach can greatly simplify the entropy-maximisation task.

§5
Resolution

Now let us apply objective Bayesian nets to the problem of learning from experience.

Consider our starting point: an agent has a domain of binary variables $V = \{B_1, \ldots B_k\}$, but no background knowledge. To construct an objective Bayesian net we first link each pair of variables that occur in the same constraint. But there is no knowledge here—so no constraints, and no edges in the constraint graph. To construct the objective Bayesian net—the Bayesian net

[10]Other examples of influence relations are discussed in Williamson (2005b, Part II).

that represents the maximum entropy probability function—we must convert this graph into a directed acyclic graph that satisfies the Markov Condition, and determine the probability distribution of each variable conditional on its parents in this graph. Since there are no edges in the constraint graph, there are no arrows in the graph of the objective Bayesian net (Fig. 1)—all variables are probabilistically independent. No variable has any parents in the graph, so all probability distributions in the objective Bayesian net are unconditional. The probability values furthest from the extremes of 0 and 1 are of course $p(b_n^{\varepsilon_n}) = 1/2, n = 1, \ldots, k$ where $\varepsilon_n = 0$ or 1. This Bayesian net determines the conditional distribution

$$p_\varepsilon =_{df} p(b_{n+1}^1 | b_1^{\varepsilon_1} \cdots b_n^{\varepsilon_n}) = 1/2$$

for all $\varepsilon_1, \ldots, \varepsilon_n \in \{0, 1\}$, and where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$. This distribution clearly represents an inability to learn from experience, but since the variables are not known to be related in any way, that is by no means unreasonable.

Now suppose instead that it is known that these variables are all instantiations of the same predicate. This knowledge provides a connection that links the variables. Moreover, suppose the agent wishes to predict the value of B_{n+1}, after observing $b_1^{\varepsilon_1}, \ldots, b_n^{\varepsilon_n}$ in that order: variable B_i is observed before variable B_j for $i < j$.

Consider this relation *observed before*. This is an influence relation: coming to learn of a variable that has not been observed before any others does not provide any grounds to change one's degrees of belief concerning the others. We shall say that B_i is an *inductive influence* of B_j if B_i is observed before B_j. Qualitative knowledge of inductive influence—knowledge of the order of observation—then translates into equality constraints on degrees of belief, as discussed in §4.

Moreover if, as is the case here, the relata of *observed before* are instantiations of the same predicate, then one expects some kind of dependence between observations: one's degree of belief in a new instance would be higher given a positive past instance than given a negative past instance. (One would expect each positive past instance to make the same difference to the degree of belief in the new instance. Similarly for negative past instances. One would also expect that the greater the number of past observations, the smaller the difference each observation would make.)

We shall suppose—just for the sake of argument—that the agent has some quantitative knowledge about the strength of inductive influence, namely that

$$p_\varepsilon \geqslant p_{\varepsilon'} + \tau_n$$

if ε' has fewer positive instances than ε, i.e. if $\sum_{j=1}^n \varepsilon_j > \sum_{j=1}^n \varepsilon'_j$, where τ_n is some small non-negative real number and $n \geqslant 1$. We shall call τ_n the n-th *inductive influence threshold*.

With full knowledge of inductive influence relationships and some knowledge of their strengths, we have the special case mentioned at the end of §4. Consequently the graph in the objective Bayesian net is just the influence graph, with an arrow from B_i to B_j just if B_i is observed before B_j, i.e. iff $i < j$, as depicted in Fig. 2. Each variable B_{n+1} has as its parents all previous variables $B_1, \ldots, B_n$. The least extreme values for the conditional distributions, found

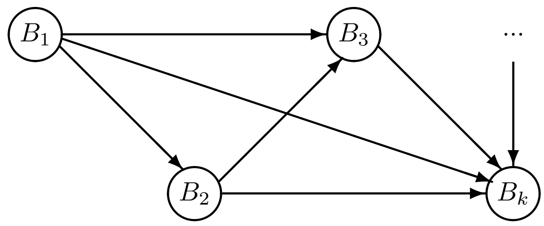

Figure 2: Knowledge of influences.

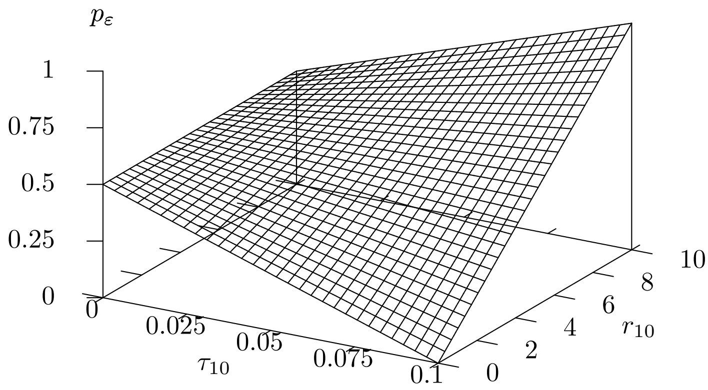

Figure 3: Inductive influence: r_{10} positive instances; inductive influence threshold τ_{10}.

by maximising entropy, are:

$$p_\varepsilon =_{df} p(b^1_{n+1}|b_1^{\varepsilon_1} b_2^{\varepsilon_2} \cdots b_n^{\varepsilon_n}) = \frac{1}{2} + \tau_n \left(\sum_{j=1}^{n} \varepsilon_j - \frac{n}{2} \right),$$

for $n = 0, \ldots, k-1$. Equivalently,

$$p_\varepsilon = 1/2 + \tau_n(r_n - \frac{n}{2}),$$

where $r_n =_{df} \sum_{j=1}^{n} \varepsilon_j$ is the number of observed positive instances. Equivalently,

$$p_\varepsilon = \frac{1 + \tau_n(r_n - s_n)}{2},$$

where $s_n =_{df} n - r_n$ is the number of observed negative instances. For ten observations, p_ε is plotted in Fig. 3.

In this case there is learning from experience as long as $\tau_n > 0$: $p(b_1^1) = 1/2$ but $p(b^1_{101} \mid b^1_1 \cdots b^1_{100}) = 1/2 + 50\tau_{100}$. The key question is how to determine the inductive influence thresholds τ_n.

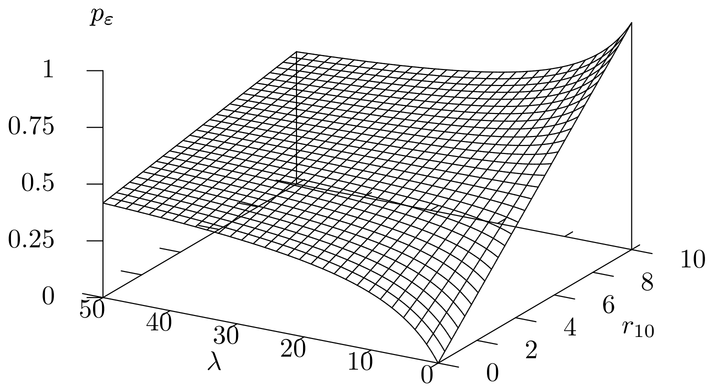

Figure 4: The Johnson-Carnap continuum of inductive methods.

§6
The Johnson-Carnap Continuum

There is one obvious constraint on the inductive influence thresholds: it must be the case that $\tau_n \leqslant 1/n$, for otherwise there exist $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ such that $p(b^1_{n+1}|b_1^{\varepsilon_1} b_2^{\varepsilon_2} \cdots b_n^{\varepsilon_n}) > 1$, in violation of the axioms of probability. So we shall write

$$\tau_n = \frac{1}{n + \lambda_n}$$

where $\lambda_n \in [0, \infty]$.

If $\lambda_n = \lambda$, a constant, then $\tau_n = 1/(n + \lambda)$ and we get what is known as the Johnson-Carnap inductive method with parameter $\lambda \in [0, \infty]$:[11]

$$p_\varepsilon = \frac{r_n + \lambda/2}{n + \lambda}.$$

A portion of this class of inductive methods is depicted in Fig. 4. (There is one qualification to make here: if $n = \lambda = 0$ then p_ε is undefined with the Johnson-Carnap method; τ_n is also undefined, but $p_\varepsilon = 1/2$ under the inductive-influence approach.)

There are some important special cases of this family of inductive methods. If $\lambda = 0$, $p_\varepsilon = \frac{r_n}{n}$: the agent's degree of belief in the next raven being black is just the observed frequency of black ravens. If $\lambda = 1$, $p_\varepsilon = \frac{r_n + 1/2}{n+1}$: this is the Jeffreys-Perks rule of succession.[12] If $\lambda = 2$, $p_\varepsilon = \frac{r_n + 1}{n+2}$, Laplace's rule of succession. If $\lambda = \infty$, $p_\varepsilon = \frac{1}{2}$: this is the case of no learning from experience.

[11](Johnson, 1932; Carnap, 1952)
[12](Good, 1965, p. 18)

We see, then, that the Johnson-Carnap continuum emerges as a special case of the inductive-influence approach. The extra generality of the inductive-influence framework is of key importance for the following reason. In their derivation of their continuum, Johnson and Carnap make a crucial assumption: a kind of *exchangeability* assumption. This is the assumption that the probability p_ε depends only on n and the number of observed positive instances r_n, and not on the order in which these observations occur (this principle is known as Johnson's Sufficientness Postulate—see §7). As Gillies points out, this is quite reasonable in cases where the underlying process exhibits objective independence—for example when observing ravens or when tossing a coin.[13] But in other cases, cases where the underlying process is a dependent (Markovian) process, this assumption is clearly unreasonable—for example in the game of *red or blue*, where a tally is kept of the number of heads and tails in a coin-tossing experiment and a blue signal is output when the number of heads is greater than or equal to the number of tails, otherwise a red signal is output.[14] So exchangeability and the Johnson-Carnap continuum are appropriate only in certain circumstances.

On the inductive-influence approach, exchangeability is not a blanket assumption. As to whether exchangeability holds depends on background knowledge: it depends on whether the λ_n are constant, i.e., it depends on the inductive influence thresholds. And there is nothing to prevent an inductive influence threshold τ_n depending on the previously observed evidence ε. An agent may start out with constant λ_n for low n, but as n increases the evidence may indicate a dependent process such as the game of red or blue, and the λ_n may vary accordingly. Thus the inductive-influence approach is more flexible than the Johnson-Carnap approach, and overcomes a key objection to the latter approach, namely its insistence on exchangeability.

Note that the Johnson-Carnap continuum is a special case ($\alpha = \beta$) of the following rule

$$p_\varepsilon = \frac{r_n + \alpha + 1}{n + \alpha + \beta + 2}$$

which is induced by the beta distribution.[15] (The beta distribution with parameters α, β, has density function $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(b)} x^{\alpha-1}(1-x)^{\beta-1}$.) This rule can be modelled in the inductive-influence framework if we set

$$\tau_n = \frac{2r_n - n + \alpha - \beta + 1}{(2r_n - n)(n + \alpha + \beta + 1)}.$$

Before discussing the measurement of the inductive influence thresholds, we shall take a look at a connection with another continuum of inductive methods.

§7
The Nix-Paris Continuum

The Johnson-Carnap continuum is not the only family of inductive methods that has been put forward in the literature. Also of interest is the Nix-Paris

[13] (Gillies, 2000, pp. 77–83)

[14] (Feller, 1950, pp. 67–95; Popper, 1957; Gillies, 2000, pp. 77–83)

[15] See Good (1965, p. 17) for example, or Zabell (1982).

continuum of inductive methods with parameter δ, which is characterised by

$$p_\varepsilon = \frac{1}{2}\left(\frac{1-\delta}{2}\right)^k \left[\left(\frac{1+\delta}{1-\delta}\right)^{r_k} + \left(\frac{1+\delta}{1-\delta}\right)^{k-r_k}\right],$$

where $\delta \in [0,1)$.[16] Note that if $\delta = 1$ then Nix and Paris (2006) set $p(b_1^{\varepsilon_1} \cdots b_k^{\varepsilon_k}) = 1$ if $k = 0$, $p(b_1^{\varepsilon_1} \cdots b_k^{\varepsilon_k}) = 1/2$ if $r_k = 0$ or k, and $p(b_1^{\varepsilon_1} \cdots b_k^{\varepsilon_k}) = 0$ otherwise. The δ-continuum differs from the λ-continuum except at the extreme values: $\delta = 1$ corresponds to $\lambda = 0$ and $\delta = 0$ corresponds to $\lambda = \infty$.[17]

The δ-continuum is the set of probability functions that satisfy the following constraints (where θ, ϕ, ψ are quantifier-free sentences of a monadic first order predicate language containing predicate B):[18]

Regularity $p(\theta) = 0$ iff $\models \neg\theta$.

Constant Exchangeability If θ' is obtained from θ by permuting constant symbols then $p(\theta') = p(\theta)$.

Predicate Exchangeability If θ' is obtained from θ by permuting predicate symbols then $p(\theta') = p(\theta)$.

Strong Negation If θ' is obtained from θ by negating each occurrence of a predicate then $p(\theta') = p(\theta)$.

Generalised Principle of Instantial Relevance If $\theta \models \phi$ and $\phi(a_{i+1}) \wedge \psi$ is consistent then $p(\theta(a_{i+2}) | \phi(a_{i+1}) \wedge \psi) \geqslant p(\theta(a_{i+1}) | \psi)$.[19]

On the other hand, the λ-continuum is the set of probability functions that satisfy Regularity, Constant Exchangeability and

Johnson's Sufficientness Postulate $p_\varepsilon = p(b_{n+1}^1 \mid b_1^{\varepsilon_1} \cdots b_n^{\varepsilon_n})$ depends only on n and $r_n = \sum_{j=1}^n \varepsilon_j$.

Let $s_n = n - r_n$ as before, and $\beta = (1+\delta)/(1-\delta)$, where $\delta \neq 1$. Then

$$p_\varepsilon = \frac{\beta^{r_n - s_n + 1} + 1}{(\beta^{r_n - s_n} + 1)(\beta + 1)}.$$

A portion of this family of inductive methods is depicted in Fig. 5.

For $\delta \neq 1$ we can model this rule under the inductive-influence approach if we let

$$\tau_n = \frac{\delta(\beta^{r_n - s_n} - 1)}{(\beta^{r_n - s_n} + 1)(r_n - s_n)}.$$

However it should be noted that the δ-continuum suffers in an important respect: $p_\varepsilon = p(b_{n+1}^1 \mid b_1^{\varepsilon_1} \cdots b_n^{\varepsilon_n})$ depends only on the difference $r_n - s_n$ between positive and negative instances, not on their absolute values. Thus one positive instance out of one past instance yields the same degree of belief in the next instance being positive as 501 positive instances out of 1001 past instances.

[16] (Nix and Paris, 2006, Theorem 14)
[17] (Nix, 2005, Proposition 4.2)
[18] (Nix and Paris, 2006, Theorem 24)
[19] This principle is discussed in Wilmers et al. (2002).

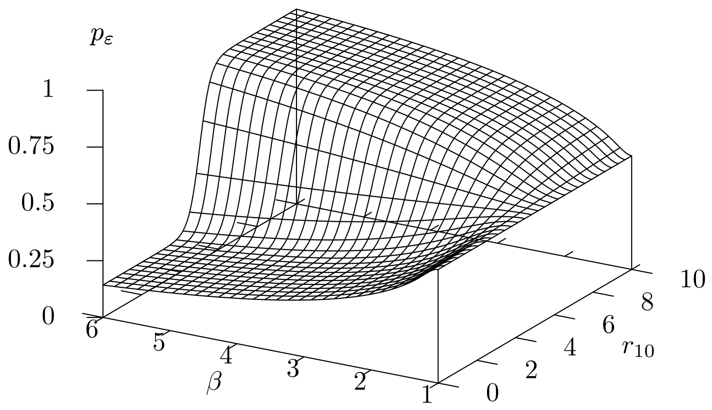

Figure 5: The Nix-Paris continuum of inductive methods.

In contrast, Johnson's Sufficientness Postulate ensures that (apart from the extreme case $\lambda = \infty$) degrees of belief become calibrated with frequencies in the long run:

$$\lim_{n \longrightarrow \infty} \left(p(b_{n+1}^1 \mid b_1^{\varepsilon_1} \cdots b_n^{\varepsilon_n}) - \frac{r_n}{n} \right) = 0.$$

This is surely a desirable characteristic of any inductive method, yet it contradicts the Generalised Principle of Instantial Relevance.[20] Thus the δ-continuum fails to satisfy this property. This provides grounds, then, for preferring the λ-continuum over the δ-continuum.

§8
Linguistic Slack

We now return to the question of how to determine the inductive influence thresholds τ_n, or equivalently the λ_n introduced in §6. I suggested there that the λ_n might depend on observed evidence as well as previous $\lambda_i, i < n$: if the evidence is compatible with an independent process then constant $\lambda_n = \lambda$ may be appropriate, otherwise the evidence will guide appropriate choice of λ_n. Exactly how the evidence will guide this choice is a question for future research; here I would like to focus on the former case, the choice of λ when the evidence does not indicate a dependent process. In this section we shall examine a proposal for setting λ by appealing to features of the agent's language. I shall argue that the resulting method is ultimately unsatisfactory. In §9 I shall put forward what I think is the right proposal.

[20](Wilmers et al., 2002, Theorem 3)

Predicates or property terms have two key roles in language. First, *classification*: they are used to describe individuals and to efficiently classify them by means of definite descriptions. For instance, the property terms 'female', 'Kentish' and 'logician' might be used in the sentence 'Bertha is the female Kentish logician' to communicate the identity of the individual Bertha. Second, *conceptualisation*: property terms are used to capture natural kinds or concepts. 'Female', 'Kentish' and 'logician' are natural concepts in that they latch on to categories about which we communicate and admit, albeit weakly, generalisations.

These two roles typically pull a language in different directions. The game of twenty questions shows us that for a language to be optimal with respect to classification, each predicate should bisect the population of individuals: the proportion of individuals that instantiate a conjunction of j property terms should be $1/2^j$ for $j \geqslant 1$. In a language that is optimal for classification, 20 property terms suffice to uniquely classify a million individuals; twenty questions suffice to isolate each such individual. But it is rare that natural concepts neatly bisect the population. While about half of all individuals are female, a far smaller proportion are Kentish, and fewer still are logicians. Thus a natural language tends to be non-optimal with respect to classification efficiency—from the point of view of classification there is redundancy or slack.

Plausibly, knowledge of linguistic slack has a bearing on an agent's degrees of belief. If a language has no slack—i.e. is optimal with respect to classification efficiency—then each property has frequency $\frac{1}{2}$. Objective Bayesianism advocates setting degrees of belief to frequencies where known, so an agent who knows that there is no linguistic slack should give degree of belief $\frac{1}{2}$ that a property will hold of the next individual to be observed (in the absence of further knowledge that constrains this degree of belief). So in this case an agent's degrees of belief should not be permitted to vary from $\frac{1}{2}$ on the basis of observed evidence; there should be no learning from experience, $\lambda = \infty$. On the other hand, if a language does have slack then it is likely that the frequency of some property is not $\frac{1}{2}$. Knowledge of this slack should lead the agent to be less cautious about changing her degrees of belief on the basis of observed evidence: her degrees of belief should be permitted to vary from $\frac{1}{2}$, and the more slack the more variation.

One can quantify linguistic slack as follows. Given a language, let EQ be the expected number of single-predicate questions required to identify an individual. If the individuals are sampled uniformly at random then $EQ \geqslant \log_2 n$ where n is the number of individuals. Let the *slack* of the language (lack of classification efficiency) be measured by $\sigma = EQ - \log_2 n$. $1/\sigma$ can then be used as a measure of the classification efficiency of the language (if $1/\sigma$ is high then the properties have frequency near $\frac{1}{2}$). Thus $1/\sigma$ is a natural candidate for the inductive parameter λ:

$$\lambda = \frac{1}{EQ - \log_2 n}$$

Consider some toy examples. Suppose we have four individuals, Auberon and Bertha who are logicians, and Cuthbert and Doreen who are not. In language 1 there are two natural property terms *female* and *logician*. Here $EQ = 2, \lambda = \infty$ and $p(\mathit{female}(\mathit{Bertha})|\neg \mathit{female}(\mathit{Auberon})) = 1/2$. In language 2 there is one natural property term *female*, and an unnatural property term *random*, which holds of Auberon and Doreen. Again, $EQ = 2, \lambda = \infty$ and $p(\mathit{female}(\mathit{Bertha})|\neg\mathit{female}(\mathit{Auberon})) = 1/2$: naturalness of the predicates need not impact on classification efficiency. In language 3 there are two natural property terms *female* and *human*. Since all the individuals are human, this language does not have the capacity to isolate individuals by their properties, $EQ = \infty, \lambda = 0$ and $p(\mathit{female}(\mathit{Bertha})|\neg\mathit{female}(\mathit{Auberon})) = 0$. Finally language 4 has three property terms *female*, *logician* and *human*. In this case $EQ = 8/3, \lambda = 3/2$ and $p(\mathit{female}(\mathit{Bertha})|\neg\mathit{female}(\mathit{Auberon})) = 3/10$. Natural languages will of course be most like language 4 in that they will have natural property terms, some of which are redundant, and hence some slack.

Note that EQ (and thus λ) can be estimated by performing 20-question type games. This procedure is also readily generalisable to individuals sampled according to some distribution q which need not be uniform. In this case the slack $\sigma = EQ - EQ^*$ where EQ^* is the optimum EQ; information theory tells us that this optimum EQ is determined by an optimum coding, e.g. Huffman coding, and that $H(q) \leqslant EQ^* < H(q) + 1$ where H is entropy.

While this procedure gives an objective way of determining the inductive influence thresholds

$$\tau_n = \frac{\sigma}{n\sigma + 1}$$

before the arrival of empirical observations, it suffers from a number of problems. First, there is an implicit assumption here that λ_n is a constant λ. It would be nice to have some justification for this assumption. Second, the procedure is somewhat arbitrary—why set λ to $1/\sigma$ rather than some other function inversely proportional to σ? Third, although unlikely in a natural language, it is quite possible to construct a language that has a large amount of slack and for which all properties have frequency $\frac{1}{2}$ because they all apply to the same half of the population. In this case an increase in slack fails to motivate an increase in amount to which past observations can change degrees of belief—ideally degrees of belief should not budge from the known frequency $\frac{1}{2}$. Thus the link between slack and degrees of belief is not a strong as might be thought. Finally, the linguistic slack may simply not be known, in which case the question of the choice of λ_n remains open.[21]

In view of the above problems, I think that this method for setting the λ_n is untenable. We must continue our quest to identify the inductive influence thresholds.

§9
Frequencies and Degrees of Belief

In this section I shall put forward what I think is the right way to determine the inductive influence thresholds τ_n—equivalently the λ_n.

Suppose, just for the sake of argument, that our agent knows that

$$\mathit{freq}_{F,a}(F(a)) = x,$$

i.e. knows that the frequency of an arbitrary individual instantiating an arbitrary property term in the language is some value x. Here the reference class ranges

[21] It might also be objected that the procedure makes induction language-relative. I argue in Williamson (2005a, Chapter 12) and Williamson (2006b, §16) that this is no bad thing.

over both individual terms a and property terms F in the agent's language. In the toy languages of §8, for example, x is $\frac{1}{2}$ for languages 1 and 2, $\frac{3}{4}$ for language 3, and $\frac{2}{3}$ for language 4. (Note that one could take $|1/2-x|$ to be an alternative measure of the slack in a language—however the concept of linguistic slack does not play a part in the proposal being put forward here.)

If this is all the background knowledge that the agent has, then according to objective Bayesianism this frequency information should directly constrain the agent's prior degrees of belief, $p(b_n^1) = x$ for all n. So set $p(b_1^1) = x$. Now

$$\begin{aligned} p(b_2^1) &= p(b_2^1|b_1^1)p(b_1^1) + p(b_2^1|b_1^0)p(b_1^0) \\ &= x\frac{1+\tau_1}{2} + (1-x)\frac{1-\tau_1}{2} \\ &= \tau_1(x-1/2) + 1/2 \end{aligned}$$

and this is equal to x if and only if $\tau_1 = 1$ for $x \neq 1/2$. When $x = 1/2$ continuity considerations would motivate setting $\tau_1 = 1$ as well. Now $\tau_1 = 1/(1+\lambda_1)$ so $\lambda_1 = 0$.

One can show inductively that for general n,

$$p(b_{n+1}^1) = n\tau_n(x-1/2) + 1/2$$

and this is equal to x if and only if $\tau_n = 1/n$ (for $x \neq 1/2$, and, appealing to continuity considerations, for $x = 1/2$ too). $\tau_n = 1/(n+\lambda_n)$ so $\lambda_n = 0$.

In sum, then, in the absence of further knowledge e.g. that the B_n are produced by a Markov process, whatever the value of x the agent should simply set her degrees of belief according to the observed frequencies:

$$p_\varepsilon = \frac{r_n}{n}.$$

Equivalently (when $n \geqslant 1$), she should set her degrees of belief according to the Johnson-Carnap inductive method with parameter $\lambda = 0$.

This argument began with the supposition that the frequency $\mathit{freq}_{F,a}(F(a))$ is known. But this supposition is not essential. All that is required is indifference: as long as the initial background knowledge does not warrant giving different prior degrees of belief to b_i^1 and b_j^1 for some i and j, then by the Principle of Indifference (which is a special case of the Maximum Entropy Principle) these degrees of belief must be the same, $p(b_1^1) = p(b_2^1) = \cdots = p(b_k^1) = x$ say. Whence by the above argument, $\lambda_n = 0$ for all $n = 1, \ldots, k$.

In the absence of indifference let j be the smallest index such that background knowledge differentiates between b_j^1 and the b_i^1 that come before it, so $p(b_1^1) = p(b_2^1) = \cdots = p(b_{j-1}^1) \neq p(b_j^1)$. Then $\lambda_1 = \lambda_2 = \cdots = \lambda_{j-2} = 0$ and background knowledge will guide the determination of subsequent λ_n.

In sum, objective Bayesianism advocates setting degrees of belief to observed frequencies in the short run as well as the long run.

§10
Conclusion

I hope to have shown that learning from experience is, after all, possible under objective Bayesianism. We saw in §5 that objective Bayesian nets provide a new way of framing the problem of learning from experience: *observed before* is an influence relation and earlier observations exert an inductive influence on later observations. We then bootstrapped a quantitative solution as follows. First we supposed known inductive influence thresholds τ_n. The resulting objective Bayesian net has parameters

$$p(b_{n+1}^1|b_1^{\varepsilon_1}b_2^{\varepsilon_2}\cdots b_n^{\varepsilon_n}) = \frac{1+\tau_n(r_n - s_n)}{2}.$$

In §6 we saw that the axioms of probability force $\tau_n \leqslant 1/n$. Then in §9 we saw that indifference forces $\tau_n = 1/n$. So we do, in fact, know the inductive influence thresholds. Consequently, in the absence of further relevant knowledge one should set one's degree of belief in the next raven being black to the observed frequency of black in past observations of ravens:

$$p(b_{n+1}^1|b_1^{\varepsilon_1}b_2^{\varepsilon_2}\cdots b_n^{\varepsilon_n}) = \frac{r_n}{n}.$$

There is, of course, more to do. I suggested in §6 that background knowledge and observed evidence might override the default inductive influence thresholds; it would be interesting to explore this possibility in more detail. Another important task is to extend the formalism to cover multiple predicates and also relations.

References

Carnap, R. (1952). *The continuum of inductive methods.* University of Chicago Press, Chicago IL.

Earman, J. (1992). *Bayes or bust?* MIT Press, Cambridge MA.

Feller, W. (1950). *Introduction to probability theory and its applications.* Wiley, New York, third (1971) edition.

Gillies, D. (2000). *Philosophical theories of probability.* Routledge, London and New York.

Good, I. J. (1965). *The esimation of probabilities: an essay on modern Bayesian methods.* MIT Press, Cambridge MA.

Howson, C. and Urbach, P. (1989). *Scientific reasoning: the Bayesian approach.* Open Court, Chicago IL, second (1993) edition.

Jaynes, E. T. (2003). *Probability theory: the logic of science.* Cambridge University Press, Cambridge.

Johnson, W. E. (1932). Probability: the deductive and inductive problems. *Mind*, 41(164):409–423.

Neapolitan, R. E. (1990). *Probabilistic reasoning in expert systems: theory and algorithms.* Wiley, New York.

Nix, C. (2005). *Probabilistic induction in the predicate calculus.* PhD thesis, University of Manchester.

Nix, C. J. and Paris, J. B. (2006). A continuum of inductive methods arising from a generalised principle of instantial relevance. *Journal of Philosophical Logic*. To appear.

Paris, J. B. (1994). *The uncertain reasoner's companion*. Cambridge University Press, Cambridge.

Paris, J. B. and Vencovská, A. (2003). The emergence of reasons conjecture. *Journal of Applied Logic*, 1(3–4):167–195.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, San Mateo CA.

Popper, K. R. (1957). Probability magic or knowledge out of ignorance. *Dialectica*, 11(3–4):354–374.

Rosenkrantz, R. D. (1977). *Inference, method and decision: towards a Bayesian philosophy of science*. Reidel, Dordrecht.

Williamson, J. (2002). Maximising entropy efficiently. *Electronic Transactions in Artificial Intelligence Journal*, 6. www.etaij.org.

Williamson, J. (2005a). *Bayesian nets and causality: philosophical and computational foundations*. Oxford University Press, Oxford.

Williamson, J. (2005b). Objective Bayesian nets. In Artemov, S., Barringer, H., d'Avila Garcez, A. S., Lamb, L. C., and Woods, J., editors, *We Will Show Them! Essays in Honour of Dov Gabbay*, volume 2, pages 713–730. College Publications, London.

Williamson, J. (2006a). Motivating objective Bayesianism: from empirical constraints to objective probabilities. In Harper, W. L. and Wheeler, G. R., editors, *Probability and Inference: Essays in Honour of Henry E. Kyburg Jr.* College Publications, London.

Williamson, J. (2006b). Philosophies of probability: objective Bayesianism and its challenges. In Irvine, A., editor, *Handbook of the philosophy of mathematics*. Elsevier. Handbook of the Philosophy of Science volume 9.

Wilmers, G. M., Hill, M. J., and Paris, J. B. (2002). Some observations on induction in predicate probabilistic reasoning. *Journal of philosophical logic*, 31:43–75.

Zabell, S. L. (1982). W.E. Johnson's "sufficientness" postulate. *The Annals of Statistics*, 10(4):1090–1099.