ON THE BAYESIAN "OCCAM FACTORS" ARGUMENT FOR OCCAM'S RAZOR

to appear in *Computational Learning and Natural Learning Systems Vol. 3*, T. Petsche et al. (Ed.), MIT Press, 1994 (hopefully).

by David H. Wolpert

The Santa Fe Institute, 1660 Old Pecos Trail, Suite A, Santa Fe, NM, 87501 (dhw@santafe.edu)

Abstract: This paper discusses some of the problematic aspects of the Bayesian first-principles "proof" of Occam's razor which involves Occam factors. Although it is true that the posterior for a model is reduced due to Occam factors if that model is capable of expressing many functions, the phenomenon need not have anything to do with Occam's razor. This paper shows this by i) performing *reductio ad absurdum* on the argument that the Occam factors effect implies Occam's razor; ii) presenting an alternative Bayesian approach which explicitly does not result in Occam's razor; and finally iii) disentangling the underlying problem with viewing the Occam factors argument as a proof or "automatic embodiment" of Occam's razor.

INTRODUCTION

This paper concerns the problem of inductive inference of input-output functions, sometimes also known as (supervised) machine learning. Special cases of the problem are regression and classification. For current purposes the problem can be formulated as follows: We have an *input space* X and an *output space* Y. There is an unknown single-valued function from X to Y which will be referred to as the *target* function f. One is given a *training set* L which consists of a set of m samples of the target function, perhaps made with observational noise. (The set of the X components of the elements of L is written as L_X, and similarly for the Y components.) One is then given a value from the input space as a *question*. The problem is to use the training set to guess what output space value corresponds to the given question. Such a guessed function from questions to outputs is known as a *hypothesis* function h. An algorithm which produces a hypothesis function as a guess for a target function, basing the guess only on the training set of m $(X \times Y)$ vectors read off of that target function, is called a *generalizer*. The quality of such a generalizer (according to some appropriate measure) is here called the generalization *error*, or *cost*.

It is commonly assumed that Occam's razor works well in supervised learning problems. The question naturally arises of *wh*y it works well. One approach to this issue is to derive sufficiency conditions for Occam's razor to work [Wolpert 1990]. Another more common approach has been to try, in essence, to formally justify Occam's razor from first principles [Blumer et al. 1987, Sorkin 1983]. One recently popular attempt to do this [MacKay 1991, Berger et al. 1992, Loredo 1990, Jeffreys 1939, Gull 1989a, Garrett 1991] has relied on using "Occam factors" in the context of conventional Bayesian analysis [Skilling 1989, 1992, Gull 1989ab, Loredo 1990, MacKay 1991, Buntine and Weigend 1991 and references therein, Wolpert and Stolorz 1992]. This paper discusses some of the difficulties, in the context of supervised learning, with using this Bayesian "Occam factors" argument to infer Occam's razor.

Section 1 synopsizes the Occam factors argument. Section 2 then presents some *reductio ad absurdum* difficulties with that argument. Section 3 goes on to present an alternative Bayesian

analysis which explicitly does not result in Occam's razor. Finally, section 4 discusses the underlying causes for the difficulties with viewing the Occam factors argument as a proof or "automatic embodiment" of Occam's razor.

Throughout this paper I will use the notation P(.) to mean either a probability function or a probability density function; the context should make the meaning clear.

1. OUTLINE OF THE OCCAM FACTOR "PROOF" OF OCCAM'S RAZOR

The Bayesian Occam factor-based "proof" of Occam's razor can be stated in several ways. Some of them are explicitly based on the "evidence procedure". This procedure has recently come under attack. In particular, it has recently become appreciated that the evidence procedure has nontrivial sufficiency conditions [Strauss1992, Wolpert 1992ab]. It appears that these conditions do not hold very often, and in any case, it would be a very odd kind of Occam's razor which holds only when those sufficiency conditions are met.

Without explicitly using the evidence procedure, the Occam factor argument can be summarized as follows (see [MacKay 1991, Berger et al. 1992, Loredo 1990, Jeffreys 1939, Gull 1989a, Garrett 1991]). Consider a parameter space C. Define a "model", or a "theorist", as a mapping from any $c \in C$ to a target function from X to Y. (This is essentially the same as what is called a "method" in [Wolpert 1990] or an "interpreter" in [Pearl 1978].) As an example, if X is the real numbers, **R**, as is Y, and if C is the set of possible quintuples of real numbers, the 4th order polynomial series using those five parameters is a model: the model is the mapping $\{p_0, p_1, p_2, p_3, p_4\} \rightarrow \Sigma_{i=0}^{4} p_i x^i$. (Note that this example could be easily modified so that either X and/or Y is not infinite.) Another example of a model, which uses the same C but in a nonlinear manner, is the following 5th order series of Legendre polynomials: $\{p_0, p_1, p_2, p_3, p_4\} \rightarrow \Sigma_{i=0}^{4} L_i(p_i x)$. Note that the image space of C (i.e., the set of functions from X to Y which are expressible with some $c \in C$)

differs for the two models. Together, a particular model and a particular set of parameter values define a particular target function. Accordingly, I will often write (m, c) as shorthand for the function given by parameter c and model m.

Now consider two models, m_1 and m_2, with associated parameter spaces C_1 and C_2. For simplicity, assume that both C_1 and C_2 are subsets of the same Euclidean vector space and have the same dimension. Assume further that $C_1 \subset C_2$. (For example, C_1 might be the interior of one hypercube in $\mathbf{R}^n$, and C_2 the interior of a larger hypercube, properly surrounding C_1.) Let c_1 refer to elements of C_1, and similarly for c_2. Our event space consists of triples {data, model, parameter value from the parameter space associated with that model}. So for example P(data = L, model = m_1, C_2 parameter value = c_2) is undefined.

Now in general, the posterior for a model, $P(m_i \mid L)$, equals $P(L \mid m_i) \times P(m_i) / P(L)$. In turn, $P(L \mid m_i) = \int dc_i\, P(L \mid m_i, c_i) \times P(c_i \mid m_i)$. Examine two particular models, m_1 and m_2. Since we have no way of choosing between the two models, by the "principle of indifference" [Loredo 1990], we might wish to take $P(m_1) = P(m_2)$. Using this gives

$$P(m_1 \mid L) / P(m_2 \mid L) = P(L \mid m_1) / P(L \mid m_2)$$
$$= \int dc_1\, P(L \mid m_1, c_1) \times P(c_1 \mid m_1) \;\; / \;\; \int dc_2\, P(L \mid m_2, c_2) \times P(c_2 \mid m_2).$$

This is the so-called "Bayes factor" for model m_1 over model m_2. Dividing it by the ratio of maximum likelihood values, $\{\max_{c_1} [P(L \mid m_1, c_1)]\} \;/\; \{\max_{c_2}[P(L \mid m_2, c_2)]\}$, we get the so-called "Occam factor" [Loredo 1990].

To see why this might have something to do with Occam's razor, for simplicity assume that the ratio $\{\int dc_1\, P(L \mid m_1, c_1)\} \;/\; \{\int dc_2\, P(L \mid m_2, c_2)\}$ can be well approximated by the ratio $\{\max_{c_1} [P(L \mid m_1, c_1)]\} \;/\; \{\max_{c_2} [P(L \mid m_2, c_2)]\}$. (This might be reasonable, for example, if $P(L \mid m_i, c_i)$ is peaked as a function of c_i, for both i = 1 and i = 2.) Also assume the "uninformative"

form for $P(c_i \mid m_i)$, namely a uniform density: $P(c_i \mid m_i) = 1 / [\int_{C_i} dc_i\ 1] \equiv [V(C_i)]^{-1}$. These conditions give

$$\frac{P(m_1 \mid L)}{P(m_2 \mid L)} = \frac{V(C_2) \times \max_{c_1} [P(L \mid m_1, c_1)]}{V(C_1) \times \max_{c_2} [P(L \mid m_2, c_2)]}.$$

Dividing the right-hand side by the ratio of maximum likelihoods, we see that the Occam factor for model 1 over model 2 is simply the (inverse) of the ratio of volumes of the associated parameter spaces.

To clarify the discussion, assume that in addition to $P(c_i \mid m_i) = 1 / [\int_{C_i} dc_i\ 1]$, we also have $\int dc_1\ [P(L \mid m_1, c_1)] = \int dc_2\ [P(L \mid m_2, c_2)]$ (whether or not the ratio of those integrals equals the ratio of the respective maximum likelihoods). Under this assumption, the ratio of $P(m_1 \mid L)$ to $P(m_2 \mid L)$ is just the ratio of volumes of the parameter spaces. So everything else being equal, the "bias" favoring m_1 over m_2 is given by (the reciprocal of) the ratio of the volume of C_1 to the volume of C_2. Models with a large a priori range of possible parameter values are penalized. This is the basis for the conventional "Occam factor" argument for why Occam's razor must hold *a priori* ([MacKay 1991, Jeffreys 1939, Berger 1992, Loredo 1990, Gull 1988, Garrett 1991]), for the case where C_1 and C_2 have the same dimension but different volumes.

Now in supervised learning it is almost always the case that what we are *ultimately* interested in isn't "models" at all, but rather input-output functions. After all, almost always the objects associated with real world costs are those functions. Moreover, it is not even clear that probabilities over models have physical meaning. For example, how could we ever observe that one particular model has probability 1? (As opposed to observing that one particular input-output function, which one can fit with (usually) many models, has probability 1.) What could such a statement mean?[1]

To circumvent this issue and also deal directly with the ultimate objects of interest, we must

extend the conventional argument given above to deal with functions rather than models. In making this extension we must be very careful to define our probabilities precisely. For example, if X is finite, and Y is **R**, then f lives in a Euclidean vector space, just as C does. In such a scenario transforming between probability densities over C (the realm of Occam factors) to probability densities over f (the realm of that which directly interests us) involves multiplying by Jacobians, determining the single-valuedness of the mapping between C and f's, etc. If both X and Y are uncountable, then the math becomes even more messy, since the cardinality of C differs from that of the set of all f's.

This potentially crucial fact is ignored in the standard treatments of Occam factors, which content themselves with calculating probabilities of models rather than functions. So to parallel those treatments, and also to keep the analysis relatively simple, here I will take both X and Y finite (although C is still a subset of a Euclidean vector space). As demonstrated below, doing this means there are no Jacobian factors or the like introduced when we extend the analysis from models to functions.

Using such a finite X and Y, let $F_m(c')$ indicate the set of $c \in C$ such that (m, c) = (m, c'). All the c in $F_m(c')$ code for the same function, (m, c'). So if f is only expressible with the single model m, the probability of the function f = (m, c') is given by $P(m, c : c \in F_m(c'))$, or just $P(m, F_m(c'))$ for short. This probability is the integral over all $c \in F_m(c')$ of the density function P(m, c). In other words, f is a (function-valued) random variable defined over the event space of models and associated parameter values; if f is only expressible with the single model m, then $P(f) = \int dc\, P(c, m)\, \delta((c, m), f)$.

Let f_1 be the function expressible as (m_1, c_1), and let f_2 be the (different) function expressible as (m_2, c_2). We are interested in $P(f_i \mid L)$ for $i \in \{1, 2\}$. For convenience, assume that f_1 can not be expressed with m_2, and f_2 can not be expressed with m_1. (The parameter spaces of the two models overlap - recall that $C_1 \subset C_2$ - but there is no a priori assumption that the models' image spaces of possible target functions overlap.) Then *if we assume that the only possible models are* m_1 *and*

m_2 (i.e., all other models have zero probability), we can write $P(f_i \mid L) = P(m_i, F_{m_i}(c_i) \mid L) = P(F_{m_i}(c_i) \mid m_i, L) \times P(m_i \mid L)$.

This is important because in most real-world scenarios, it is $P(f \mid L)$ which is of primary interest, not $P(m \mid L)$. Yet as mentioned above, the arguments in the Occam factor literature usually concern themselves exclusively with $P(m \mid L)$. This is rather unfortunate, since the preceding analysis shows that in general the object of primary interest, $P(f_i \mid L)$, is not equal to $P(m_i \mid L)$ up to an overall proportionality constant, even in those simple cases where we don't have to worry about Jacobians and the like. In particular, the most likely function (given the data) need not even be expressible by the most likely model (given the data). So stopping the analysis short by only analyzing posteriors over models can give misleading results.

On the other hand, $P(f \mid L)$ also exhibits an Occam factor effect, just like $P(m \mid L)$. Moreover, although it is not true in general, under certain assumptions $P(f \mid L)$ *is* proportional to $P(m \mid L)$, so that it reflects Occam factors in the exact same manner as $P(m \mid L)$. In particular, this is the case for the assumptions made in this section (together with new ones introduced below).

More precisely, use Bayes' theorem to re-express $P(F_m(c_i) \mid m_i, L)$ in our formula for $P(f_i \mid L)$, thereby getting $P(f_i \mid L) \propto P(L \mid m_i, F_m(c_i)) \times P(F_{m_i}(c_i) \mid m_i) \times P(m_i)$. An important point to note is that $P(m_i \mid L)$ has dropped out. (Note also that $P(m_i \mid L)$ is simply the sum over distinct sets $F_m(.)$ of $P(f_i \mid L) = P(m_i, F_m(c_i) \mid L)$.)

Using this result, and making the same assumption of uniform $P(m_i)$ used to calculate $P(m_i \mid L)$, we can write

$$\frac{P(f_1 \mid L)}{P(f_2 \mid L)} = \frac{P(L \mid m_1, F_{m_1}(c_1)) \times P(F_{m_1}(c_1) \mid m_1)}{P(L \mid m_2, F_{m_2}(c_2)) \times P(F_{m_2}(c_2) \mid m_2)}.$$

Now assume that both f_1 and f_2 have the same likelihood value, $P(L \mid f_i)$. Also assume that

$P(F_{m_i}(c_i) \mid m_i) = \kappa / [\int_{C_i} dc_i\ 1]$ for some constant κ. (This would occur for example if $P(c_i \mid m_i)$ were uniform over all allowed c_i, and if the C_i-space volume corresponding to any function f expressible with m_i were a constant, independent of f or i.) Then we get

$$P(f_1 \mid L) / P(f_2 \mid L) = [\int_{C_2} dc_2\ 1] / [\int_{C_1} dc_1\ 1].$$

Just like before, everything else being equal, if one's model has a large volume, one gets penalized.

Note that this argument doesn't say that, everything else (e.g., the likelihood) being equal, one should use only the model m_i with the smallest parameter-space volume, $V(C_i)$. In fact, it doesn't even say that, everything else being equal, if one must use either model m_1 or model m_2 one should use the model with smaller volume. (After all, $P(f_1 \mid L) > P(f_2 \mid L)$ does not necessarily imply that the expected loss using f_1 is less than the expected loss using f_2 - see [Wolpert and Stolorz, 1992].) What it <u>does</u> say is that, everything else being equal, the model m_i with smallest volume will contribute the most to the posterior average, i.e., to the (quadratic loss function) Bayes-optimal guess, $\Sigma_f f \times P(f \mid L)$. This is the sense in which the argument recounted above can be viewed as related to Occam's razor.

Finally, note that we are examining $P(F_{m_i}(c_i), m_i \mid L)$, not (for example) $P(F_{m_i}(c_i), m_i \mid m_i, L) = P(F_{m_i}(c_i) \mid m_i, L)$. In other words, we are <u>not</u> talking about something like "the probability of f *given* model m". (After all, if m were given, there would be no point in calculating things like Bayes factors.) This is despite the fact that we got from posteriors over models to posteriors over functions by multiplying by $P(F_{m_i}(c_i) \mid m_i, L)$.

2. FLAWS DUE TO FREEDOM TO REDEFINE MODELS

There exist a number of ways to see that the traditional Occam factor argument recounted

above must be based on ultimately ad hoc assumptions. For example, if I only give you two functions from X to Y, say $y = x^2$ and $y = \sin(x)$, it would be an amazing piece of inference to deduce *a priori* which is more likely without making any ad hoc assumptions. Which are more likely in the universe, parabolas or sine waves? However if one reasons using "models" in the Occam-factor-type manner, one would think that one might be able to answer this question from first principles. For example, one might let sin(x) be the function (m_1, c_1), and x^2 the function (m_2, c_2), for some essentially arbitrary models m_1 and m_2. As above, assume that sin(x) can not be expressed with model 2, and similarly for x^2. Then $P(\sin(x)) / P(x^2) = P(m_1, F_{m_1}(c_1)) / P(m_2, F_{m_2}(c_2))$, which equals $[P(F_{m_1}(c_1) \mid m_1) / P(F_{m_2}(c_2) \mid m_2)] \times [P(m_1) / P(m_2)]$ = the Occam factor for model m_1 over model m_2, if we make the assumptions for $P(F_{m_i}(c_i) \mid m_i)$ and $P(m_i)$ made in the "proof" of Occam's razor recounted above.

This illustrates what is perhaps the most important shortcoming of the Occam factor argument: it evaluates probabilities by using models, but there is no a priori protocol for how one should constrain the space of possible models. An example of the problems this can cause is provided by MacKay, in the setting where the image spaces of (functions expressible by) two models overlap, so that there are functions they can both express. In [MacKay 1991] he writes "... imagine that for two models, the most probable interpolants happen to be identical [i.e., the most probable function expressible with model 1 is the same as the most probable function expressible with model 2]. In this case the generalization error for the two solutions [i.e., for the two most probable functions] must be the same. But the .... [Occam-factor-based posterior] will not in general be the same: typically, the model that was a priori more complex will suffer a larger Occam factor ..." In other words, a particular target function f expressible by both models might be considered either complex or simple - and penalized accordingly - depending solely on the whim of the researcher as to what model to use. This means that when using Occam factor type arguments, which model the researcher uses can affect the function (s)he will guess. However almost always when engaged in supervised learning one is ultimately concerned *solely* with the generalization error. And the gen-

eralization error of a particular f is independent of what model is used to express that f (obviously). The conclusion is that the correlation between how Occam factor type arguments advise us to guess and *that which we are interested in* (generalization error) can be poor.[2] In other words, if Occam's razor were simply a ramification of Occam factors and nothing more, we would be forced to conclude that Occam's razor has little to say concerning generalization error.

In fact, any argument assigning posteriors by measuring the volume of allowed parameter space, without any concern for the model being used, has a number of problems. (No such concern is present in the Occam factor argument; it is precisely this lack of concern which allows MacKay to make his statement). One such problem follows from the fact that one can always bijectively map a space of parameters into a subset of that space. In fact, one can always bijectively map an n-dimensional Euclidean parameter space with arbitrarily large volume into a 1-dimensional Euclidean parameter space with arbitrarily small volume (although in the case where the dimensions of the spaces differ, such a map will not be differentiable, in general). Without explicit consideration of what model one is using, there is no formal way to distinguish between the pre-mapping and post-mapping models. Therefore without such consideration, one can not derive single-valued results. (See the discussion in [Wolpert 1990] concerning Occam invariances.)

One doesn't have to go to the trouble of bijectively mapping entire parameter spaces to run into problems however. As an example, modify the Occam factor argument by introducing two new models, m_1' and m_2', with the same parameters spaces as m_1 and m_2, namely C_1 and C_2 respectively. Let $m_1' = m_1$ except for one small change: whereas the target function f_1 is expressed by any element $(m_1, F_{m_1}(k))$ (k being a particular member of C_1 such that $f_1 = (m, k)$), any element $(m_1', F_{m_1}(k))$ does not express f_1, but rather expresses f_2. Other than that single exception, $m_1' = m_1$, i.e., $(m_1', c_1) = (m_1, c_1) \; \forall \; c_1 \in C_1$ such that $c_1 \notin F_{m_1}(k) = F_{m_1'}(k)$. Similarly, let $m_2' = m_2$ except for one small change: the target function $(m_2', F_{m_2}(t))$ does not equal f_2, but rather equals f_1. (In [Wolpert 1990], similar modifications of the model are called "Occam transformations".)

Neither m_1' nor m_2' are "illegal" models. Furthermore, there is no a priori reason not to set the

ratio $P(m_1') / P(m_2')$ equal to one, just like the ratio $P(m_1) / P(m_2)$. However if we accept this ratio, and then go through the exact same reasoning used in the Occam factor argument (only using m_1' and m_2' rather than m_1 and m_2), we come to the exact *opposite* conclusion: $P(f_1 \mid L) / P(f_2 \mid L) = [\int_{C_1} dc_1\ 1] / [\int_{C_2} dc_2\ 1]$. So again we see that the conclusions of Occam factor type arguments are not single-valued, in general.

One might try to respond to the foregoing by arguing that the Occam factor argument should only be made for "reasonable" models, and that, for example, the models m_1' and m_2' discussed above aren't "reasonable". Such an argument completely negates all claims to formal rigor however. Indeed, we have absolutely no assurances that we won't accidentally stumble across such "unreasonable" models in real life.

In addition, such an argument invites a whole host of difficulties: How does one formally set the dividing line between "reasonable" and "unreasonable" models? Isn't it true that in practice almost all of the models which are excluded as "unreasonable" are complicated (in which case by restricting ourselves to "reasonable" models we're already assuming Occam's razor and are therefore being circular)? For that matter, is the dividing line between "reasonable" and "unreasonable" models really sharp, or is it instead gradual, in which case all models, even "reasonable" ones, have different priors, depending on their "reasonableness"? If it's gradual, then to have any confidence in one's results, mustn't one rigorously prove that the approximation of ignoring all models except those which are most "reasonable" doesn't introduce large errors? etc., etc.[3]

There are other peculiar aspects to performing the analysis over models rather than over functions, in addition to those mentioned above. Perhaps the most obvious of these is the fact that the very term "Occam's razor" is used in a somewhat idiosyncratic fashion in model-based Occam factor arguments; it is being taken to mean a bias in favor of simpler *models* over more complex ones. This contrasts with the more common usage in supervised learning where it means a bias in favor of simpler *functions* over more complex ones [Wolpert 1990, Blumer et al. 1987, Sorkin 1983, Pearl 1976].

One might worry that this could result in a sharp difference between the predictions of an analysis based on favoring simpler functions and an analysis like the Occam factor argument which favors simpler models. In this regard, note that although the Occam factor calculations presented above are for posterior probabilities, they hold just as well for prior probabilities ($P(f) = P(F_m(c) \mid m) \times P(m) \propto 1 / V(C_m)$). In other words, the Occam factor argument can be viewed as, at its base, an indirect argument for how to set prior probabilities over functions, P(f). However it does not result in a P(f) similar to those usually considered when one performs Bayesian analysis over functions directly. For example, it does not result in a uniform prior, or a prior favoring smooth f, or a prior penalizing "complex f". In fact, it might very well result in a prior *favoring* "complex f", depending on the models used and on one's notion of what a "complex" function is. (At a minimum, there is no guarantee that it won't do this.) This worrisome aspect of the Occam factor argument will be returned to later.

3. INTERNAL CONSISTENCY VERSUS THE PRINCIPLE OF INDIFFERENCE

One way around these problems is to ignore "models" altogether and consider probabilities over target functions directly. It is hard to find counter-arguments to the view that, if it is at all rational, Occam's razor must hold for a complexity measure which depends solely on the input-output function, and not (directly) on how the function is expressed (see the quotation from MacKay above and [Wolpert 1990]). After all, in general one can more easily justify assigning particular priors (e.g., uniform priors) to functions rather to than models.

If, nonetheless, one does insist on considering probabilities over models, then it is imperative that one be formal and rigorous. Without such formality one isn't performing a proper Bayesian analysis, and to quote John Skilling, there is only one "valid defense of using non-Bayesian methods, namely incompetence" [Skilling, 1992].

Following this admonition of Skilling, we should set all distributions - and in particular the

priors $P(m_i)$ - in an axiomatic, formal, and consistent manner. Unfortunately there is no currently known desideratum which allows us to set the priors over models in such an axiomatic manner, for any and all scenarios. In particular, the principle of indifference, which would suggest setting uniform priors over models (as in the conventional Occam factor arguments recounted above), in fact does **not** apply to the setting of P(m). This is because the principle of indifference is predicated on there being no relevant distinction between the objects under consideration. Accordingly, it could only apply if there were no relevant pre-data distinction between the models. But the whole point of the Occam factor argument is to show that there <u>is</u> a relevant distinction between the models (namely, the volumes of their parameter spaces). So Occam factors, conventionally justified by using the principle of indifference, also demonstrate the inapplicability of the principle of indifference.

As it turns out, in those scenarios in which one can apply a desideratum to fix the $P(m_i)$, often the resultant $P(m_i)$ are <u>not</u> those assumed in the Occam factor argument. For example, consider the situation where we have two possible sets of models, M and M', consisting of models $\{m_1, ..., m_N\}$ and $\{m'_1, ..., m'_{N'}\}$ respectively, and associated with two different event spaces. (For example, M might consist of decision trees of various types, and M' might consist of Walsh polynomial expansions of various types.) Scientist A uses the model set M, and scientist B uses the model set M'. Just as in the original version of the Occam factor argument presented above, assume that any function f which is expressible with a model from M is expressible with only one model m from M. Make the same assumption for M'. Also make the assumption made in the Occam factor argument, that $P(F_{m_i}(c_i) \mid m_i) = \kappa / [\int_{C_i} dc_i \; 1]$ for all $m_i \in$ M, and again, similarly for the models in M'.

For this two-scientist scenario, one desideratum one might want to apply is that of “internal consistency”. It says that “if a conclusion can be reasoned out in more than one way, every possible way must lead to the same result” ([Loredo 1990]). In other words, the two scientists must set priors over their models in such a way that the difficulties in the Occam factor argument discussed above *can not* arise. It particular, they must set priors over their models in such a way that it can

not matter which model set they use (the two possible ways in which "a conclusion can be reasoned out").

Interestingly, it turns out that this desideratum is exactly equivalent to assuming a uniform prior P(f) (in contrast to the non-uniform P(f) arising from a scientist's assuming uniform P(m)). To see this, let f_1 be a function expressible with both model m_1 and model m'_1, and let f_2 be a function expressible with both model m_2 and model m'_2. The desideratum of internal consistency says that the ratio of posteriors, $P(f_1 \mid L) / P(f_2 \mid L)$, must depend only on the expressed quantities, f_1, f_2, and L. In other words, that ratio must be independent of which model set one uses, of whether one is scientist A or scientist B. If we define parameters c_1, c_2, c'_1, and c'_2 by the two equations $(m_1, F_{m_1}(c_1)) = (m'_1, F_{m'_1}(c'_1)) = f_1$ and $(m_2, F_{m_2}(c_2)) = (m'_2, F_{m_2}(c'_2)) = f_2$, this requirement means that

$$\frac{P(m_1, F_{m_1}(c_1) \mid L)}{P(m_2, F_{m_2}(c_2) \mid L)} = \frac{P(m'_1, F_{m'_1}(c'_1) \mid L)}{P(m'_2, F_{m'_2}(c'_2) \mid L)}.$$

If we now expand $P(m_1, F_{m_1}(c_1) \mid L) = P(L \mid m_1, F_{m_1}(c_1))\, P(m_1, F_{m_1}(c_1)) / P(L)$, and similarly for the other three probabilities, then divide both sides of our equality by $P(L \mid f_1)$, multiply both sides by $P(L \mid f_2)$, and as in the Occam factor argument assume that $P(F_m(c) \mid m) = \kappa / V(C)$ for all $m \in$ M or M' and for all $c \in C_m$, we get

$$\frac{P(m_1)\, V(C_2)}{P(m_2)\, V(C_1)} = \frac{P(m'_1)\, V(C'_2)}{P(m'_2)\, V(C'_1)}.$$

The only way this equality can hold in general, for any model sets, is if for any model m and associated parameter space C, $P(m) / V(C_m)$ equals some constant which is the same for all models in

a given model set.[4] In other words, to satisfy our desideratum, whether we are scientist A or scientist B, we must set the prior of a model as proportional to the associated volume in parameter space.

It is easily verified that this scheme for fixing priors ensures that the problems with the Occam factor argument recounted above - and in particular the problem arising when one redefines models to get the exact opposite guess - do not occur. With this scheme, the freedom to redefine models will not result in any contradictions. Nor will the possibility of having two models guess the same "interpolant", as in MacKay's comments, cause any trouble. This scheme for choosing P(m) also has the (very reasonable) ramification - absent from the uniform P(m) case - that so long as $\Sigma^{N}_{i=1} V(C_{m_i}) = \Sigma^{N'}_{i=1} V(C_{m'_i})$, P(f) has the same value whether f is expressed with model set M or model set M'. ($P(f) = P(m, F_m(c)) = P(F_m(c) \mid m) \times P(m) = \kappa \times P(m) / V(C_m)$, where due to normalization of P(m), the condition on the volume sums, and the constancy of the ratio for either model set, the ratio has the same value for both model sets.)

On the other hand, the model priors of this scheme *exactly cancel the ratio of volumes in the Occam factor*; there is no more automatic Occam's razor. Indeed, in addition to assuming the model priors of this scheme, make the Occam factor argument's assumptions that i) $P(F_m(c) \mid m) = \kappa / V(C)$; and ii) any function expressible with at least one model from our model set is expressible with only one such model. Then the ratio of posteriors is given by the ratio of likelihoods, $P(f_1 \mid L) / P(f_2 \mid L) = P(L \mid f_1) / P(L \mid f_2)$. In other words, according to the principle of internal consistency, the assumptions of the Occam factor argument imply that P(f) is uniform over all f expressible by any of the models in the scientist's set of models. This means in particular that even if the principle of internal consistency results in wildly varying P(m), one should not consider this "unreasonable" unless one also considers it "unreasonable" to have a uniform P(f).)

This contrasts with the case with the principle of indifference, for which the assumptions that $P(F_m(c) \mid m) = \kappa / V(C)$ and that there is only one model to a function result in little in the way of generally obeyed laws concerning P(f). Combined with the principle of indifference, those assumptions only imply that P(f) is uniform over those f all expressible with the same particular mod-

el; the precise form of P(f) as one ranges over all possible f depends on the boundaries between regions of f's all expressible with the same model. (This implication follows from the fact that P(f) = $P(F_m(c), m) = P(F_m(c) \mid m) \times P(m)$.) As an aside, note that there are some other interesting implications of the principle of internal consistency, in particular for the distribution P(f | m, L). See footnote 5.

It is not the purpose of this paper to argue strongly that one should use the principle of internal consistency and adopt its implications for priors over models. Rather the point is that by using that principle one can make an argument which is at least as formally rigorous as the traditional Occam factor argument, and yet reach the opposite conclusion. In particular, note that those who would claim to have proven Occam's razor using Occam factors must disagree with the conclusion of the internal consistency argument concerning the ratio of the posteriors of f_1 and f_2. They must claim that a priori, for first principles reasons alone, so long as the volumes of the associated models differ it is theoretically impossible that the ratio of posteriors of two functions equals the ratio of likelihoods. In other words, they must claim that it is theoretically impossible that the prior over functions f is flat. It is hard to see how this claim can be substantiated; they must argue that their arguments assuming a uniform P(m) over some ad hoc restricted set of "reasonable" models are a priori more sensible than simply assuming a uniform P(f). What makes the Occam factor argument so odd is this very claim that uniform P(f) can be ruled out by first-principles reasoning.

There is an element of "smoke and mirrors" in the use of models, at least as far as supervised learning is concerned: It is perfectly feasible to perform Bayesian analysis in supervised learning without considering anything other than functions and training sets; there is no a priori need for models. Moreover, use of models can be quite obscuring, by (for example) leading one away from so straight-forward a prior as P(f) = constant.

Of course, if one doesn't use models in one's Bayesian analysis, Occam factors never arise. In essence, the "automatic embodiment" of Occam's razor in Bayesian analysis can be viewed as a side-effect of using models to implicitly set the prior over f. In general the effect does not arise if one performs the analysis directly in the objects of interest, namely the functions from X to Y.

4. FLAWS DUE TO NOT INTEGRATING OVER ALL MODELS

Even if one accepts the presumption of the Occam factor argument that one should perform the Bayesian analysis using models rather than functions, there are still a number of difficulties with the claim that one has inferred Occam's razor. This section discusses some of those difficulties, by highlighting the flaws in the logical argument presented in section 1.

First, note that in reality, $P(f \mid L)$ isn't set by a single model and/or single parameter value, as assumed heretofore. Rather $P(f \mid L) = S_m \int dc_m \{\delta((m, c_m), f) \times P(m, c_m \mid L)\}$, where the Kronecker delta function restricts the integral to those models m and associated parameters c_m which give the function f. The integrals each run over the (in general varying) parameters spaces associated with the models m. As discussed below, by "S_m" is meant a sum of some sort, running over models.

This equation for $P(f \mid L)$ raises a major problem for any model-based analysis of $P(f \mid L)$: Let me go about constructing models, and for each one let me ask you if it's completely, one hundred percent impossible, or if it instead has some non-zero prior probability, however small. In almost all conceivable situations, you will never answer that the probability *exactly* equals zero.[6] This means that in almost all conceivable situations, just to express $P(f \mid L)$ in terms of models requires that S_m be some kind of "super-integral", which runs over all possible models with image functions from X to Y. The math for this has never been worked out. (Indeed, it seems to be non-trivial simply to define the space of all possible models, since it implies all possible parameter spaces.) In other words, for almost all conceivable situations, the math for calculating that which we are interested in ($P(f \mid L)$) <u>in terms of models</u> has never been worked out.[7]

The Occam factor argument circumvents this problem rather disingenuously, by implicitly assuming not only that there are countably many models with non-zero prior, but that there are in fact finitely many such models. In other words, it takes S_m to be a sum over a finite number of m. In

this section, I will make the slight relaxation of this assumption by allowing the mere *possibility* that there are countably infinite models with non-zero prior. In other words, I will write $P(f \mid L) = \Sigma_m \int dc_m \{\delta((m, c_m), f) \times P(m, c_m \mid L)\}$, where the sum is over a countably infinite set. The points I wish to make in this section can be made using this slight relaxation; for current purposes, there is no reason to consider the full-blown possibility of allowing all models.

For simplicity's sake, assume $P(c_m \mid m) = 1 / V(C_m)$. Then up to an overall constant we can rewrite our sum-integral as $P(L \mid f) \times \Sigma_m \int dc_m \{\delta((m, c_m), f) \times P(m) / V(C_m)\}$. First, two points of nomenclature: i) Define the integral in this expression as the "Occam bias". It is the bias in favor of a function which arises independent of the likelihood, due solely to the fact that the function can be expressed by certain models. ii) Refer to the assumption that for all models m $P(m) \in \{0, k\}$ for some constant k - the assumption made in the conventional Occam factor argument for Occam's razor - as the assumption of "flat" P(m). The Occam factor argument for Occam's razor says, loosely speaking, that if all functions can only be expressed by a single model, and if the volume of parameter values corresponding to a function doesn't vary with the function or model, and if P(m) is flat, then the Occam bias favors those functions f which can be expressed with a "simple" model.

Note that if P(m) is non-zero for enough models, then in fact there will be more than one model which can express a given f, and our Occam bias is found by averaging the Occam factors over all those models which are capable of expressing f (loosely speaking). In such a situation, even if P(m) is flat, the fact {f_1 can be expressed with a simple model whereas f_2 can not} does not imply that the Occam bias favors f_1. One has to consider *all* the models which can express the two functions. Indeed, if f_1 can be expressed with one model and f_2 with two, then even if f_1's model is simpler than either of f_2's, it will often be the case that f_2 would have the higher posterior.

Moreover, consider the case where for any function there are *many* models with non-zero prior which can express that function (a situation which almost always holds in supervised learning when X and Y are finite). In this case, to find the posterior one has to sum over many Occam bias

terms. Some such terms might be relatively large, but the fact that we're summing over many terms would be expected to hide this somewhat; if you have two sums, Σ_1 and Σ_2, both consisting of many terms, the fact that the largest term in Σ_1 happens to be greater than the largest term in Σ_2 doesn't give you strong reason to believe that $\Sigma_1 > \Sigma_2$. So for this scenario, with flat P(m), there is little bias favoring a function just because it can be expressed with a particularly simple model. In other words, as soon as one admits of enough possible models, even if one accepts the assumption of flat P(m), the Occam factor "proof" of Occam's razor (in the sense of a bias favoring functions which can be expressed with simple models) dissipates.

So the first crucial assumption in the Occam factor proof of Occam's razor is that the set of models with non-zero prior is extremely small. In other words, all models other than a select few are assumed completely impossible, with exactly zero prior probability. Note that this <u>is</u> an assumption, and is <u>not</u> based on first principles reasoning or any particular desiderata. One is tempted to use a favorite phrase of Occam factor proponents and say that this assumption is in fact highly "unreasonable". After all, it's hard to imagine how one could make a strong argument in favor of restricting the support of P(m) so drastically.

This point can be summarized as follows: The contraction in parameter space described by Occam factors is a real phenomenon. But if one allows enough models, or even allows the changing of the set of allowed models, the ramifications of that contraction need not have anything to do with Occam's razor.

In addition to having the set of models with non-zero prior be extremely small, the Occam factor argument for Occam's razor makes another crucial assumption. This second crucial assumption is that P(m) is "flat" and has the same value for those m for which it doesn't equal 0. As was noted previously, the principle of indifference does not justify this assumption. Yet if P(m) can be a function of m, all bets are off. In particular, as was shown above, if P(m) is proportional to $V(C_m)$, then no Occam's razor ensues. In fact, independent of arguments like the principle of consistency, one might argue that a flat P(m) is, explicitly, extremely peculiar. Why should <u>all</u> models have prior of 0 or k? Why can't any models have any other prior values? And if some models *can* have other

values, then what's to prevent those "other values" from negating the Occam factor effect, as when P(m) is proportional to $V(C_m)$?

One *could* say that even if P(m) is non-uniform, since the Occam factor effect biasing the posterior in favor of simpler models (for few enough allowed models) is "automatic", Bayesian analysis "automatically embodies" Occam's razor. This is rather disingenuous however. One particular case where P(m) is non-uniform is when the principle of internal consistency holds. For such a scenario, would we want to say that, since the ratio of P(m)'s automatically favors more "complicated" models, Bayesian analysis "automatically embodies" the *opposite* of Occam's razor?

CONCLUSION

In [Wolpert 1992c] the distribution P(generalization error = E | function guessed by the researcher h, data L) is considered (an off-training-set generalization error being used). It is proven there that if the universe is maximum entropy (i.e., if P(f) is flat), then this conditional probability is independent of h. Therefore, unless one can prove that the universe is not maximum entropy, one can not prove that the function one guesses affects likely generalization error, and in particular one can not prove that choosing h according to Occam's razor affects likely generalization error.

This means that there must be some rather problematic aspects to any "proof" or "automatic embodiment" of Occam's razor. In particular, there must be such aspects to the Occam factors-based "automatic embodiment" of Occam's razor.

One such aspect is the fact that the Occam factor argument concerns models. The first problem with this is that the mathematics for performing a fully rigorous analysis in terms of models has never been worked out, and appears to be highly non-trivial. To circumvent this difficulty, i.e., to allow the mere possibility of performing an analysis in terms of models, the Occam factor argument must make an implicit (and very hard to justify) assumption: all but a countable number of models have *exactly* zero probability.

Given this assumption, the Occam factor argument results in favoring simpler models. However in the real world we're almost always ultimately interested in *functions* expressed in terms of those models. And in general, biases in favor of simple models need not correspond to biases in favor of simple functions. So the Occam factors "Occam's razor" can differ significantly from the "Occam's razor" which is more usually invoked in supervised learning, namely that one should favor simpler f.

Indeed, when you set things up in terms of functions, you find the Occam factor effect again, but as a term in the *prior* over f's. Stated differently, the Occam factor argument is, at root, just another way (one of maybe hundreds) to set a prior over functions. It has nothing to do with data. And the innocent-seeming assumption of a flat prior over models m - an assumption which lies at the core of the Occam factor argument - corresponds to a hugely non-flat prior over the ultimate object of interest, functions f. In addition, if rather than assuming a flat prior over models one uses the principle of internal consistency to set the prior, one arrives at a non-flat prior, one which *exactly cancels the Occam factor*, and therefore results in a uniform P(f). (In general, flat P(m) corresponds to non-flat P(f) and flat P(f) corresponds to non-flat P(m).)

Another problematic aspect to the Occam factors argument for Occam's razor is that it depends not only on having a countable number of models with non-zero prior; in fact the reasoning which results in Occam's razor depends crucially on having a flat P(m) which is non-zero for only a small number of models. As it turns out, if one allows that many models are possible, and/or that probabilities over models can vary, then the Occam factor effect disappears.

Alternatively, even if one is willing to restrict the prior over models to be flat with very small support, then one is still in trouble if one can't justify one's choice of the support set. In particular, if one sticks to Occam factor type arguments, then by changing that set of allowed models it's possible to get contradictions in the calculations of probabilities of functions.

Certainly if one does, somehow, have prior knowledge of a flat prior over models which is non-zero for only a small number of models, then the Occam factors "automatic embodiment" of Occam's razor holds (i.e., everything else being equal, P(f) will be smaller for those f expressible only

with models which can express many f). However this is a huge amount of prior knowledge (i.e., the Shannon information of such a distribution is extremely close to maximal). It is extremely difficult to see how that much knowledge could be *derived*, from first principles (as opposed to assuming it, in an ad hoc manner). This is especially true in the common real-world supervised learning scenarios, where *any* substantial prior knowledge concerning models is rare.

Phrased differently, the Occam factor argument is most naturally viewed as an equating of one "reasonable", but ad hoc, assumption (flat prior over models with only a small number of "reasonable" models allowed) with another "reasonable", but ad hoc, assumption (everything else being equal, functions expressible with simple models are more likely). The Occam factor argument shows that the "prior knowledge" contained in the one assumption is equivalent to the "prior knowledge" contained in the other. It is not a proof of either one. It is a re-*casting* of Occam's razor rather than a *derivation* of it.

Ultimately, the entire issue is arriving at the prior over models. If one can't justify one's choice of that prior on first-principles grounds, then one can not justify conclusions based on that choice, on first principles grounds. Even if that conclusion is the model-based version of Occam's razor.

All of this notwithstanding, it is important to note that this paper makes no claims concerning how well Occam's razor works in practice. This paper certainly does not argue that Occam's razor is false, and it also does not argue that the particular model-based form of Occam's razor equivalent to the assumption {flat prior over models with only a small number of models allowed} is false. What this paper <u>does</u> argue is that this assumption is crucial to the Occam factor argument, and that it has never been justified. Indeed, if one is going to try to argue things from "first principles", then the choice of prior over models which exactly negates Occam factors (and thereby induces a uniform P(f)) is at least as easy to justify as a choice of uniform prior over models. Accordingly, the claim that the Occam factor argument somehow "proves" or "automatically embodies" Occam's razor in a rigorous manner is seen to fall short. Why Occam's razor works, either for models and/or functions, is still a deep mystery.

FOOTNOTES

1. Especially, what could $P(m_i)$ mean if - as is often the case in the real world - I know that none of the m_i at hand can express the true f exactly? In general, it's one thing to consider the probability that this physical black box, sitting here in front of me, maps its input to its output via the function f (and none other). It's quite another thing to consider the probability that the input-output function of this black box, sitting on the table in front of me, is in some sense associated with one and only one particular non-physical "model" (and none other). (Especially if that model can't even express the f of the black box!) f is a real, physical property, of a real, physical object. It's hard to see how a model can be viewed the same way.

2. Note that this kind of argument also serves as a potent challenge to the common practice of setting (hyper)priors over hyperparameters without any regard as to how those hyperparameters get mapped to input-output functions. To see this, let f be a function which is parameterized in terms of two different basis sets, B_1 and B_2. Assume that for both basis sets the probabilities involved concern the same hyperparameters. (E.g., for both basis sets we have conditional priors P(expansion coefficients | α), where α is a hyperparameter.) Then setting hyperpriors over those hyperparameters without concern for the basis set used will in general result in different posteriors for f, depending on whether B_1 or B_2 is used. This means that the correlation between {how the hyperparameter procedure advises us to guess} and {that which we are interested in, i.e., generalization error} can be extremely poor.

3. Note that even if the approximation in the last question is valid, so long as the "most reasonable" models are allowed to have different priors from one another, one will not arrive at Occam's razor in general (see below). In sum, if we were content to not prove things rigorously and instead rely on notions of what is and is not "reasonable", then we might as well simply start with (the em-

inently “reasonable”) Occam’s razor directly, and not pretend to “prove” it or “automatically embody” it.

4. To see this, replace the function f_1 with the function f_3. Just like f_1, f_3 can be expressed with m_1. Assume, however, that it can not be expressed with m'_1, but that it can be expressed instead with m'_3. Now enforce our desideratum that both the ratio of the posterior of f_3 to the posterior of f_2 and the ratio of the posterior of f_1 to the posterior of f_2 be independent of whether one is scientist A or scientist B. The result is that $P(m'_1) / V(C'_1) = P(m'_3) / V(C'_3)$. Except for very specially constructed model sets B, we will be able to continue this way to force $P(m') / V(C_{m'})$ to be the same for all models in set B. In general, we will also be able to use this trick to force $P(m) / V(C_m)$ to be the same for all models in set A.

5. All of the arguments presented so far consider distributions of the form $P(m_i, F_{m_i}(c_i) \mid L)$ rather than $P(F_{m_i}(c_i) \mid m_i, L)$. The reason for this is that, generically speaking, the condition-side of a conditional probability distribution should include only that which is known, and we certainly don’t know what the “true” model is. Nonetheless, it is straight-forward to calculate $P(f \mid m, L)$ under the assumption of internal consistency. To do this, let c^* be a parameter value such that $f = (m, c)$ for any $c \in F_m(c^*)$. We are invited to calculate $P(m, F_m(c^*) \mid m, L) = P(F_m(c^*) \mid m, L)$. This can be rewritten as the ratio $P(F_m(c^*), m \mid L) / P(m \mid L)$. The numerator is the kind of distribution we have been discussing in the text. In particular, by the consistency principle the numerator is proportional to the likelihood $P(L \mid f)$ (assuming $P(F_m(c) \mid m) \propto 1 / V(C)$ and that the only model in our model set which can express f is m, so that the consistency principle results in uniform P(f)). Therefore this numerator is independent of m. The denominator is a marginalization of the numerator. Different choice of model m (and in particular different “parameterizations”) mean different $P(m \mid L)$; everything else in the expression for $P(f \mid m, L)$ is independent of m.

6. Of course, in stating this I'm implicitly assuming some notion of what it means to say "probability of model m = value". But as was pointed out before (see footnote 1), this problem of giving a precise definition to P(m) is highly non-trivial. All I'm saying here is that whatever one's definition of P(m), it will almost always be the case that $P(m) \neq 0.0000 \ldots$ for any model m.

7. Note that for the finite X and finite Y case, there is no such "super-integral" difficulty if one performs the analysis directly in terms of f (so that rather than priors over models, one directly considers priors over functions). However there is a (less severe) version of this difficulty for an analysis performed directly in f space, if both X and Y are uncountable.

8. This issue of the multiplicity of models has other implications, besides those concerning Occam's razor. For example, we can use it to characterize the technique of "model selection" as the approximation $P(f \mid L) = \Sigma_m \int dc_m \{\delta((m, c_m), f) \times P(c_m \mid m, L) \times P(m \mid L)\} \cong \int dc_{m^*} \{\delta((m^*, c_{m^*}), f) \times P(c_{m^*} \mid m^*, L)\}$, up to an overall proportionality constant, where m* is the model maximizing P(m | L). Viewed this way, model selection is seen to be essentially equivalent to the "evidence approximation" (Gull 1989ab; MacKay 1991); rather than the noise or a regularization constant being the unknown hyperparameter, here it's the model. One might expect this approximation to be quite good when P(m | L) is strongly peaked about m*. It turns out though that such peakedness is not a sufficient condition for the evidence approximation to hold; the evidence approximation can be quite poor even when P(m | L) is strongly peaked. (See (Wolpert 1992ab, Strauss 1992).) Accordingly, one should use model selection with great care.

ACKNOWLEDGMENTS

This work was done in part under the auspices of the Department of Energy, and in part at The Santa Fe Institute and under NLM grant F37 LM00011. I would like to thank Paul Stolorz, C. E.

M. Strauss, and Richard Silver for interesting discussions. I would also like to thank Harry Martz for reading over a copy of this paper. Finally, I would like to thank Radford Neal, an anonymous referee of an early draft of this paper, and especially Wray Buntine for helpful comments.

REFERENCES

Berger, J.O., and Jefferys, W.H. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, **80**, 64-72.

Blumer, A., et alia (1987). Occam's razor. *Information Processing Letters*, **24**, 377-380.

Buntine, W., Weigend, A. (1991). Bayesian back-propagation. *Complex Systems*, **5**, ~~6~~03-643.

Garret, A.J.M., (1991). Ockham's razor. In "Maximum entropy and Bayesian methods",W.T. Grandy and L.H. Schick (Eds.). Kluwer Academics publishers.

Gull, S.F. (1989a). Bayesian inductive inference and maximum entropy. In "Maximum entropy and Bayesian methods in Science and Engineering, Vol. 1: Foundations". Ed. G. Erickson, Kluwer Academic publishers.

Gull, S.F. (1989b). Developments in maximum entropy data analysis. In "Maximum entropy and Bayesian methods", J. Skilling (Ed.). Kluwer Academics publishers.

Jeffreys, H. (1939). Theory of Probability. Oxford University Press, Oxford.

Loredo, T.J. (1990). From Laplace to supernova SN 1987A: Bayesian inference in astrophysics. In "Maximum entropy and Bayesian methods", P. F. Fougere (Ed.). Kluwer Academics publishers.

MacKay D. (1991). "Bayesian Interpolation", and "A practical Bayesian framework for backprop networks". Companion papers presented at *Neural Networks for Computing* conference, Snowbird, Utah.

Pearl, J. (1978). On the connection between the complexity ad credibility of inferred models", *International Journal of General Systems*, **4**, 255-264.

Skilling, J. (1989). Classic maximum entropy. In "Maximum-entropy and Bayesian methods", J.

Skilling (Ed.). Kluwer Academics publishers.

Skilling, J. (1991). Fundamentals of MaxEnt in data analysis. In “Maximum Entropy in Action”, Brian Buck and Vincent A. Macaulay (Ed.), Clarendon Press, Oxford, England.

Sorkin, R. (1983). A quantitative Occam’s razor. *International Journal of Theoretical Physics*, **22**, 1091-1104.

Strauss, C.E., Wolpert, D.H., Wolf, D.R. Alpha, evidence, and the entropic prior. In *Maximum Entropy and Bayesian Analysis*. Ed. A. Mohammed-Djafari. Kluwer Academic Press. In press.

Wolpert, D., and Stolorz, P. (1992). On the implementation of Bayes-optimal generalizers. SFI TR 92-03-012.

Wolpert, D. (1990). The relationship between Occam’s razor and convergent guessing. *Complex Systems*, **4**, 319-368.

Wolpert, D (1992a). A rigorous investigation of “evidence” and “Occam factors” in Bayesian reasoning. Santa Fe Institute TR 92-03-013.

Wolpert (1992b). On the use of evidence in neural networks. In *Advances in Neural Information Processing Systems - 5*, Ed. S. Hanson et al. Morgan Kauffmann, in press.

Wolpert, D. (1992c). On the connection between in-sample testing and generalization error. *Complex Systems*, **6**, 47-94.