A RIGOROUS INVESTIGATION OF "EVIDENCE" AND "OCCAM FACTORS" IN BAYESIAN REASONING

by David H. Wolpert

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

Abstract: This paper first reviews the reasoning behind the Bayesian "evidence" procedure for setting parameters in the probability distributions involved in inductive inference. This paper then proves that the evidence procedure is incorrect. More precisely, this paper proves that the assumptions going into the evidence procedure do not, as claimed, "let the data determine the distributions". Instead, those assumptions simply amount to an implicit replacement of the original distributions, containing free parameters, with new distributions, none of whose parameters are free. For example, as used by MacKay [1991] in the context of neural nets, the evidence procedure is a means for using the training set to determine the free parameter α in the distribution $P(\{w_i\}) \propto \exp(\alpha \Sigma_{i=1}^{N} w_i^2)$, where the N w_i are the N weights in the network. As this paper proves, in actuality the assumptions going into MacKay's use of the evidence procedure do not result in a distribution $P(\{w_i\}) \propto \exp(\alpha \Sigma_{i=1}^{N} w_i^2)$ for some α, but rather result in a parameter-less distribution, $P(\{w_i\}) \propto (\Sigma_{i=1}^{N} w_i^2)^{-(N/2 + 1)}$. This paper goes on to prove that if one makes the assumption of an "entropic prior" with unknown parameter value, in addition to the assumptions used in the evidence procedure, then the prior is completely fixed, but in a form which can not be entropic. (This calls into question the self-consistency of the numerous arguments purporting to derive an entropic prior "from first principles".) Finally, this paper goes on to investigate the Bayesian first-principles "proof" of Occam's razor involving Occam factors. This paper proves that that "proof" is flawed.

INTRODUCTION

This paper concerns the problem of inductive inference, sometimes also known as (supervised) machine learning. For current purposes this problem can be formulated as follows: We have an *input space* X and an *output space* Y. There is an unknown function from X to Y which will be referred to as the *target* function. (This function is sometimes instead called the "parent" function, or the "generating" function.) One is given a set of m samples of the target function (the *training set*), perhaps made with observational noise. 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 on the target function corresponds to the given question. Such a guessed function from questions to outputs is known as a *hypothesis* function. 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*.

One popular framework for addressing the problem of inductive inference is conventional "Bayesian" analysis [Skilling 1989, 1992, Gull 1989, Loredo 1990, MacKay 1991, Buntine and Weigend 1991 and references therein]. Recently there has been work done in this framework investigating a technique which involves using the "evidence" to set various parameters. (e.g., regularizing parameters, noise parameters) [MacKay 1991, Gull 1989, Loredo 1990, Skilling 1992]. This paper proves that this technique is unjustified and incorrect. There has also recently been work done in the Bayesian framework purporting to prove Occam's razor from first principles [MacKay 1991, Gull 1989, Loredo 1990]. This paper demonstrates that that "proof" is flawed.

Section 1 of this paper presents the mathematical background for the rest of the paper. This section motivates conventional Bayesian analysis by introducing the concept of "Bayes optimality". Section 2 proves that the use of evidence is incorrect. Some of the arguments presented in this section are similar to arguments which have been known for a while to main-stream statisticians ([Lindley and Smith 1972, Morris 1983]). Section 3 demonstrates that the purported first-principles "proof" of Occam's razor using Occam factors is flawed.

Those who wish to should be able to skip directly to sections 2 and 3 without too much trouble. Some relevant notational points: $f \in F$ refers to a target function, $h \in H$ refers to a hypothesis function, and θ refers to a training set.

1. Background

This section presents the mathematical background necessary for the rest of the paper and discusses "Bayes optimality". For a more detailed exposition of this material, see [Wolpert 1992a] and [Wolpert and Stolorz 1992].

Consider the situation where the input space, X, is finite and discrete, consisting of n elements. The output space, Y, is also finite and discrete and consists of r elements. There are four fundamental quantities involved in inductive inference:

1) Single-valued target functions f from X to Y. The space of all such functions is denoted by F.

2) Single-valued hypothesis functions h from X to Y. The space of all such functions is denoted by H. H is identical to F.

3) X values at which the target function is sampled to create the training set, $\theta_X(i)$.

4) The corresponding Y values of the training set, $\theta_Y(i)$.

The (ordered) set of all the values $\theta_X(i)$ is written as θ_X. The (ordered) set of all the values $\theta_Y(i)$ is written as θ_Y. The full training set $\{\theta_X, \theta_Y\}$ is written as θ, and m is the number of input-output pairs in θ.

Since f, h, θ_X, and θ_Y are the four quantities of interest, *all* aspects of any relevant problem can be expressed in terms of a joint probability distribution over them. In particular,

I) The *sampling assumption* is the rule for how one creates training sets θ from target functions f. In other words, the sampling assumption is the conditional probability distribution $P(\theta \mid f)$. As an example, if one assumes that the sampling is noise free, then $P(\theta \mid f) \propto \delta(\theta \subset f)$, where the delta

function is defined as 1 if θ lies on f, 0 otherwise. (The set notation is motivated by viewing both θ and f as sets of input-output pairs.) If the sampling assumption is noise-free i.i.d. sampling, then $P(\theta \mid f) \propto \delta(\theta \subset f) \times \prod_{i=1}^{m} \pi(\theta_X(i))$, where π(.) is sometimes called the "sampling distribution" and is usually unknown. (The proportionality constant in this formula for P(θ | f) is set by normalization.)

II) The *generalizer* is the probability of guessing a certain hypothesis function, given a certain training set. In other words, the generalizer is P(h | θ). As an example, one generalizer which has recently received a good deal of attention is the "Gibbs learning algorithm" [Haussler et al. 1991, Opper and Haussler 1991], also sometimes known as the "exhaustive learning generalizer" [Schwartz et al. 1990, Hertz et al. 1991]: "Exclude all hypothesis functions which don't agree with the training set, and then pick randomly from those hypothesis functions remaining", where "random" means according to some distribution T(h). More formally, this generalizer is defined by $P(h \mid \theta) = k(\theta) \times T(h) \times \delta(\theta \subset h)$, where k(θ) is a normalization factor, $T(h) \geq 0 \; \forall \; h$, and $\forall \; \theta \; \exists$ at least one h agreeing with θ for which T(h) > 0. (This last requirement ensures that the generalizer is defined for all θ.) The Gibbs generalizer can be viewed as a zero-temperature limit of the scenarios analyzed in the "statistical mechanics" supervised learning framework (see [Seung et al. 1991a, 1991b, Tishby et al. 1989]).

III) The generalization error function is a mapping from $(f, h, \theta_X, \theta_Y)$ to **R**. It measures how "good" h is as a guess for f. One rather popular choice is the i.i.d. error function: $Er(f, h, \theta) \equiv \sum_{x \in X} \pi(x) \times [1 - \delta(f(x), h(x))]$, where π(.) is the same distribution used to define the sampling assumption, and the δ function here is a Kronecker delta. This error function gives the average (according to π(.)) number of agreements between a hypothesis function h and a target function f. It is appropriate for symbolic output spaces.

One can view the analysis of probability distributions across the event space $(f, h, \theta_X, \theta_Y)$ as an extension of conventional Bayesian analysis, to have it distinguish target functions from hypothesis functions. The failure to make this distinction is one of the major shortcomings of conventional

Bayesian analysis. For example, by analyzing distributions across (f, h, θ_X, θ_Y), one can prove the intuitive notion that how "aligned" one's generalizer is with "the universe's generalizer" determines how well one generalizes. More formally, one can prove that P(generalization error = E | training set θ) is given by the following (non-Euclidean) inner product:

$$1) \quad P(E \mid \theta) = \Sigma_{\{f, h\}} P(h \mid \theta) \times P(f \mid \theta) \times M_{E,\theta}(h, f),$$

where $M_{E,\theta}(h, f)$ is a symmetric matrix whose precise form is determined by the error function one uses. (See [Wolpert 1992a] for a proof of (1) and a discussion of its significance.) Given (1), it is straight-forward to prove, for example, that if the output space is symbolic (and one uses an appropriate error function) and if P(f | θ) is uniform over those f agreeing with θ, then as far as off-training-set error is concerned, *it is completely irrelevant how one guesses*. Similarly, if P(h | θ) is uniform over those h agreeing with θ, then as far as off-training-set error is concerned, *it is completely irrelevant what the true P(f) is*. Since it does not distinguish target functions from hypothesis functions, *by construction* conventional Bayesian analysis is incapable of deriving such results. A discussion of some of the other issues beyond the scope of conventional Bayesian analysis can be found in [Wolpert 1992a].

There exist many different scenarios of interest in supervised machine learning. For example, both the PAC framework ([Valiant 1984, Blumer et al. 1987, Blumer et al. 1989]) and the statistical mechanics framework concentrate on the analysis of the distribution P(E | f, m) (for some fixed but usually unknown target function f) when i) the sampling assumption is i.i.d., and ii) the error function is the i.i.d. error function. The difference between the two frameworks arises from the fact that they concentrate on different types of generalizers (see [Wolpert 1992c]). Since f is fixed in both frameworks, both frameworks derive results which are independent of P(f | θ), i.e., both reach conclusions which are essentially independent of the real world.

In contrast, in conventional Bayesian analysis f is never fixed, which leads such analysis to

concentrate directly on the distribution $P(f \mid \theta)$. In particular, the "Bayes-optimal" generalizer is the generalizer $P(h \mid \theta)$ which minimizes the expectation value $\langle E \rangle|_\theta \equiv \Sigma_E [E \times P(E \mid \theta)]$.[1] Since $P(E \mid \theta)$ leaves f unfixed, to find the Bayes-optimal generalizer, one must know (or assume) $P(f \mid \theta)$.

More formally, assume we have an error function of the form $Er(f, h) = \Sigma_{x \in X} [d(f(x), h(x)) \times \pi(x)]$, where $d(., .)$ is a metric and $\pi(x) \geq 0 \; \forall \; x$. (It is implicitly assumed that there exists at least one x for which $\pi(x) > 0$.) The set of problems using such an error function is extremely broad. For example, the conventional i.i.d. error functions found in PAC and the statistical mechanics school are of this form, where $\pi(.)$ is the sampling distribution used to choose the elements of θ_X. In general though, the $\pi(.)$ appearing in our assumed error function need not have anything to do with the sampling process. Indeed, the analysis presented below goes through even if $\pi(.)$ is determined by θ, as is the case for example if one wishes to use an error function which measures only off-training-set generalization. Define $\langle E \rangle|_\theta \equiv \Sigma_E [E \times P(E \mid \theta)]$. In ([Wolpert and Stolorz 1992]) the following is proven:

2) i) The $P(h \mid \theta)$ which minimizes $\langle E \rangle|_\theta$ is $\delta(h, \text{argmin}_{h \in H} [\langle E \rangle|_{h,\theta}])$, where

$\langle E \rangle|_{h,\theta} \equiv \Sigma_E [E \times P(E \mid h, \theta)]]$).

ii) If the error function can be written as $Er(f, h) = \Sigma_{x \in X} [d(f(x), h(x)) \times \pi(x)]$, then the h which minimizes $\langle E \rangle|_{h,\theta}$ is the function $h^* \equiv \{x \in X \rightarrow \text{argmin}_{y \in Y} \Omega(y, x)\}$, where

$\Omega(y, x) \equiv \Sigma_f [d(f(x), y) \times P(f \mid \theta)]$.

iii) The resultant value of $\langle E \rangle|_\theta$ is $\Sigma_{x \in X} \{\pi(x) \times \Omega(h^*(x), x)\}$.

Intuitively, (2) says that for any question x, one should choose the y which minimizes $\Omega(y, x)$, the average distance from y to f(x) (the average is over all f(.), according to the distribution $P(f \mid \theta)$). Note that this result holds regardless of what (if any) noise process is present; any such noise is taken care of automatically, in the $P(f \mid \theta)$ term. Note also that, technically speaking, h^* is a function of θ; for notational simplicity, that dependence will be explicitly expressed only when

it is important.

Example 1: Let the metric be $d(a, b) = (a - b)^2$. (Strictly speaking, this function is not a metric, since it doesn't obey the triangle inequality. However in this paper I will be a bit loose with the terminology and refer to it as a metric. (In this paper the triangle inequality is never used.)) One might want to use this metric when the output space is a set of real numbers, as for example in time-series analysis. (Note that use of such a metric does not necessarily imply anything about the noise; in general, it only need concern solely how we wish to measure "error". In particular, such a squaring metric does not necessarily imply that noise is Gaussian.) Using this metric, $h^* = \{x \in X \rightarrow \text{argmin}_{y \in Y} \Omega(y, x)\} = \{x \in X \rightarrow \Sigma_f [P(f \mid \theta) \times f(x)]$, i.e., $h^*(x) = \Sigma_f [P(f \mid \theta) \times f(x)]$. This metric is implicit in most conventional "Bayesian analysis" ([Skilling 1989, 1992, Gull 1989, Loredo 1990, MacKay 1991, Buntine and Weigend 1991 and references therein.'s]) where once one figures out the posterior of interest (which here is $P(f \mid \theta)$) one guesses the average according to that posterior. Note that even if we want to restrict things so that (for example) $P(f) = 0$ unless $f(x) \in \{0, 1\} \; \forall \; x \in X$, with the squaring metric we will usually implicitly want Y to be (some interval in) the reals. If this weren't the case, but rather (for example) $Y = \{0, 1\}$, then in general $h^* \notin H$.

Example 2: Let the metric be $d(a, b) = 1 - \delta(a, b)$, where $\delta(., .)$ is the Kronecker delta function. One might want to use this metric for classification problems, where Y is a space of abstract symbols; with this metric, the error function becomes the average number of times h(x) and f(x) disagree. Using this metric, $h^* = \{x \in X \rightarrow \text{argmin}_{y \in Y} \Omega(y, x)\} = \{x \in X \rightarrow \text{argmax}_{y \in Y} \Sigma_f [\delta(f(x), y) \times P(f \mid \theta)]$. This is the hypothesis function referred to as "Bayes-optimal" in [Haussler et al. 1991] and [Opper and Haussler, 1991]. Note that unlike the h^* in the previous example, the h^* here might not be unique (i.e., there might exist more than one y minimizing $\Omega(y, x)$ for some particular x). In fact, as was implied by the discussion following (1), if P(f) is flat and there is no noise, then there is no minimum to $\Omega(y, x)$ for $x \notin \theta_X$; all guesses result in the same (expected) error.

Note that the h* defined in (2) is independent of $\pi(.)$. On the other hand, the proof of (2) makes clear that if $\pi(x) = 0$ for any particular x value x', changing h(x'), even if it changes the value of $\Omega(h(x'), x')$, does not affect $\langle E \rangle|_{\theta,h}$ (so long as $h(x \neq x')$ is left unchanged). (This result is independent of the metric used.) In other words, (1) only sets the behavior of the optimal hypothesis function over those x such that $\pi(x) \neq 0$; behavior over those x for which $\pi(x)$ does equal 0 is completely free. So for example if one uses an off-training-set error function (i.e., if $\pi(x) = 0$ for all $x \in \theta_X$), then the behavior of h for $x \in \theta_X$ is completely independent of the expected generalization error. In this case, if I tell you that of two hypothesis functions one perfectly reproduces θ and the other does not, and if I tell you nothing else, then I have <u>not</u> provided you with a means for predicting which of those two hypothesis functions has the smaller expected generalization error (see [Wolpert 1992a]). This is true even if $P(f \mid \theta)$ strongly favors those f which reproduce θ.

Note also that the function h* is critically dependent on the metric used. For example, assume that outputs $y \in Y$ have p components (i.e., are p-dimensional). In this case one might wish to use the Euclidean metric, $d(a, b) = [\ \Sigma_{k=1}^{p}\ (a_k - b_k)^2\ ]^{1/2}$. The resultant h* will in general be different from $\Sigma_f [P(f \mid \theta) \times f(x)]$, the h* associated with the (multi-dimensional version of the) metric of example 2, $d(a, b) = \Sigma_{k=1}^{p}\ (a_k - b_k)^2$. For example, when p = 1, the Euclidean metric leads one to choose the median rather than the mean. Of course, in practice one often assumes that distributions are so peaked that any reasonable scheme for getting a single number from the distribution $P(E \mid h, \theta)$ will give essentially the same number. However for many scenarios (e.g., when one has a small data set), this assumption is likely to be a bad one, and our choice of metric will become critical.[2]

Finally, note that there exist many reasonable error functions which are not of the form assumed in (2). For example, one might be interested in the error function, $Er(f, h) \equiv 1 - \delta([\Sigma_{x \in X}\ \delta(f(x), h(x))], 0)$. This error function is not of the type assumed in (2). However minimizing $\Sigma_E [E \times P(E \mid h, \theta)]$ for this error function is equivalent to maximizing $P(E = 0 \mid h, \theta)$ for an error function of the type which is of the assumed in (2) (see [Wolpert and Stolorz 1992]).

3. The use of "evidence" in conventional Bayesian analysis

This section proves that the use of "evidence" in conventional Bayesian analysis is incorrect.

Despite the fact that there are perfectly reasonable alternative choices (see end of section 2), for the most part conventional Bayesian analysis restricts itself to situations where the error functions is of the form assumed in (2), with a squaring metric (as in example 1). For such situations, to guess in a Bayes-optimal fashion, one must know (or assume) $P(f \mid \theta)$.

Often one knows $P(\theta \mid f) \; \forall \; \theta, f$. Given such knowledge, one way to deduce $P(f \mid \theta) \; \forall \; \theta, f$ is to know (or assume) the r^n values of $P(f \mid \theta)$ for any single θ. In other words, knowing $P(f \mid \theta)$ for any (!) particular θ, together with $P(\theta \mid f) \; \forall \; \theta, f$, is sufficient to fix $P(f \mid \theta) \; \forall \; \theta, f$ (see [Wolpert and Stolorz 1992]). This is not the usual line of attack however. Given $P(\theta \mid f) \; \forall \; \theta, f$, a more common approach is to assume the r^n values of $P(f)$. Using Bayes' theorem, this fixes $P(f \mid \theta) \; \forall \; \theta, f$. Recently a variation of this second approach has been explored, in which by using "evidence" one (tries to) circumvent the need to directly assume $P(f)$ and $P(\theta \mid f)$. The procedure is as follows:

First, one introduces a dummy variable, γ. γ is often thought of as a "regularizing parameter", or as a "noise level". In this exposition, γ will be taken to signify both. (This means that γ is 2-dimensional.) In some cases, γ might also be taken to delineate different "models" (which can loosely be thought of as different generalizers.) For simplicity of the exposition, such cases will not be explicitly considered here, although the arguments presented apply just as well to those sorts of γ.

Implicitly extend the framework presented in section 1 to have Y be the set of all real numbers. One starts the evidence procedure by specifying three distributions: $P(f \mid \gamma)$, $P(\theta \mid f \, \gamma)$, and $P(\gamma)$. One wants to derive $P(f \mid \theta)$. To do this one writes

$$P(f \mid \theta) = \int d\gamma \, P(f \mid \theta \, \gamma) \, P(\gamma \mid \theta)$$

$$= \{P(\theta)\}^{-1} \times \int d\gamma \, \{P(f \mid \theta \, \gamma) \times [\, P(\gamma) \, \Sigma_f \, (P(\theta \mid f \, \gamma) \, P(f \mid \gamma)) \,]\}.$$

The sum equals $P(\theta \mid \gamma)$, and is called "the evidence" for parameter γ. Due to the fact that we are given the three distributions inside the [], we can evaluate both the evidence and the full term inside the []. Often when one does this one finds that the [] is sharply peaked as a function of γ, so one approximates it with a Dirac delta function, $\delta(\gamma - \gamma')$. (This approximation amounts to the assumption that, given only a training set, the parameters in γ (e.g., noise level, regularizing constant) are almost exactly specified.) For many $P(\gamma)$ (e.g., uniform $P(\gamma)$), γ' can be well approximated by $\text{argmax}_{\gamma}(P(\theta \mid \gamma))$. Whether or not one can find γ' so easily, once it is found one can write

$$P(f \mid \theta) = P(f \mid \theta\, \gamma') / P(\theta)$$

$$= [P(\theta \mid f\, \gamma')\, P(f \mid \gamma')] \;\; / \;\; [P(\theta)\, P(\theta \mid \gamma')]$$

$$\propto P(\theta \mid f\, \gamma')\, P(f \mid \gamma').$$ (N.b. the normalization constant depends on θ.)

This sets how one guesses; one finds γ', plugs it into the distributions $P(\theta \mid f\, \gamma)$ and $P(f \mid \gamma)$, and then evaluates the product of those two distributions to get the posterior $P(f \mid \theta)$.

Example Ia: In [MacKay 1991],

i) One sets $P(f \mid \gamma)$ to $\exp(-\gamma_1 \times E_W(f)) / Z_W(\gamma_1)$, where $Z_W(.)$ is a normalization constant, γ_1 means the first component of γ, and $E_W(f)$ is $\Sigma_i\, w_i^2$, where the w_i are the values of a set of parameters specifying f in some particular parameterization scheme. The idea behind this form for $P(f \mid \gamma)$ is to build in a bias towards "simpler" f.[3]

ii) To set $P(\theta \mid f, \gamma)$ one equates it with $P(\theta_X \mid f, \gamma) \times P(\theta_Y \mid \theta_X, f, \gamma)$, assumes that $P(\theta_X \mid f, \gamma)$ is independent of f (and can therefore be absorbed into an overall normalization constant), and sets $P(\theta_Y \mid \theta_X, f, \gamma)$ to $\exp(-\gamma_2 \times E_D(\theta_Y, \theta_X, f)) / Z_D(\gamma_2)$, where $Z_D(.)$ is a normalization constant, γ_2 means the second component of γ, and $E_D(\theta_Y, \theta_X, f) = \Sigma_{i=1}^{m}\; [\theta_Y(i) - f(\theta_X(i))]^2 / 2$ (Note that $Z_D(.)$ turns out to be independent of f, since one integrates over all possible θ_Y.) This $P(\theta_Y \mid \theta_X, f, \gamma)$ corresponds to assuming Gaussian noise.

iii) $P(\gamma)$ is assumed to be uniform. (One suspects there are some implicit limits on the range of γ, although such limits are never stated, and in particular never occur in the relevant

integrals.)

Example IIa: In [Gull 1989],

i) One sets $P(f \mid \gamma)$ to $\exp(-\gamma_1 \times E_W(f)) / Z_W(\gamma_1)$, where $Z_W(.)$ is a normalization constant, γ_1 means the first component of γ, and $E_W(f)$ is the entropy of f, $-\Sigma_{x \in X} [f(x) \ln(f(x))]$ (an integral for continuous valued X).[4] This is the so-called "entropic prior" which, it has been argued, is *a priori* the optimal form when f represents a positive additive distribution [Skilling 1989, 1992].

ii) As in [MacKay 1991], one uses $P(\theta \mid f, \gamma) = P(\theta_X \mid f, \gamma) \times P(\theta_Y \mid \theta_X, f, \gamma)$, assumes that $P(\theta_X \mid f, \gamma)$ is independent of f (and can therefore be absorbed into an overall normalization constant), and sets $P(\theta_Y \mid \theta_X, f, \gamma)$ to $\exp(-\gamma_2 \times E_D(\theta_Y, \theta_X, f)) / Z_D(\gamma_2)$, where $Z_D(.)$ is a normalization constant, and γ_2 means the second component of γ. Although the analysis in [Gull 1989] isn't restrictedto a particular form for $E_D(., ., .)$, a quadratic form (like in [MacKay 1991]) is explicitly presented as a reasonable choice for $E_D(., ., .)$.

iii) $P(\gamma_1) \propto 1 / \gamma_1$ across some "sensible range" [a, b]. $P(\gamma_2)$ is not explicitly specified, but a similar choice for it is sensible. It is implicitly assumed that $P(\gamma_1, \gamma_2) = P(\gamma_1) \times P(\gamma_2)$.

Proponents of the "evidence" approach claim that with it, "the emphasis of the modern Bayesian approach is not on the inclusion of priors into inference, as the old myth states ... it is not necessary to work out 'the right prior', assign it, and stick with it ... [we can consider alternative situations which] assign different priors, and we can compare these alternatives in the light of the data by evaluating the evidence" [MacKay 1991]. The idea is to "let the data determine the distributions", roughly speaking.[5] Of course, this assessment of the evidence procedure is not completely accurate. Even in the evidence procedure, one <u>does</u> "work out 'the right prior', assign it, and stick with it"; it's just that those priors are over γ rather than over F and θ directly.

Note that, although at the end of the day <u>all</u> that concerns us is whether or not the evidence pro-

cedure results in a more accurate estimate of $P(f \mid \theta)$, no argument is offered by its proponents that it does in fact result in more accurate $P(f \mid \theta)$. In other words, no argument is provided saying that setting priors over γ in a fixed manner, without any free parameters, necessarily results in more accurate $P(f \mid \theta)$ than fixing priors over f and θ directly, without any free parameters. No argument is provided telling us that the evidence procedure will result in better generalization.

Moreover, when one thinks about it, the fact that with the evidence procedure γ (usually) goes through a strongly non-linear mapping to get to where it says something about how to guess makes it a very odd thing to claim that it should have a simple prior (e.g., the uniform prior used in [MacKay 1991]). Generically, one would expect the distribution over hypotheses resulting from a simple $P(\gamma)$ to be extremely non-simple and informative (!). At a minimum, one almost certainly would want to choose different $P(\gamma)$ depending on the form for $P(f \mid \gamma)$ and $P(\theta \mid f\ \gamma)$. The conventional evidence procedure as outlined above ignores this point entirely. Indeed, I(i) explicitly ignores the mapping taking a set of w_i to target functions. Loosely speaking, since it doesn't distinguish amongst them, it operates in the "limit" in which all parameterizations of target functions (e.g., neural nets, radial basis functions, etc.) are the same. It is hard to see how this position can be defended. (Remember, the fact that the mathematics involved in setting some priors correctly might be intractable <u>in no way</u> justifies ignoring that mathematics and simply assuming some "reasonable" priors. Not if one is claiming that one's procedure is *a priori* superior.)

There's a more profound issue here however. Formally, all that it is done in the evidence procedure is the introduction of dummy variables along with associated (assumed) priors. In other words, the advice of Occam's razor is being violated, in the sense that extra variables are being introduced into the analysis. Furthermore, it is claimed that by using the evidence to sort things out, the introduction of those dummy variables will result in a more accurate $P(f \mid \theta)$. The obvious question is, if this claim is true, than why not introduce hundreds of new variables, assign them all priors, and then use evidence to set them all, thereby getting a even more "accurate" $P(f \mid \theta)$? Is it somehow most "robust" (in a statistical sense) to only introduce two new variables?

The natural way to address this issue is to investigate what implications carrying out the evi-

dence procedure has for $P(f \mid \theta)$. In doing this, one discovers a startling fact: Starting with parameterless priors over f and θ, and then introducing dummy parameters along with priors over those parameters, is **exactly equivalent** to simply using different parameterless priors over f and θ to start with. In other words, even if one believes that a particular choice for $P(f \mid \gamma)$, $P(\theta \mid f\ \gamma)$, and $P(\gamma)$ is correct, there is no reason whatsoever to find γ' (i.e., "maximize the evidence") and then plug that in to get $P(f \mid \theta)$. Formally,

3) i) $P(f) = \int d\gamma \{P(f \mid \gamma) \times P(\gamma)\}$. This form for is P(f) is <u>independent of the data</u>, as a proper prior for f *must be*.

ii) $P(\theta \mid f) = \int d\gamma \{P(\theta \mid f\ \gamma) \times P(f \mid \gamma) \times P(\gamma) / P(f)$. Therefore, the correct posterior is

iii) $P(f \mid \theta) \propto \int d\gamma \{P(\theta \mid f\ \gamma) \times P(f \mid \gamma) \times P(\gamma)$

This result is <u>exact</u>. Unlike in the evidence procedure, no approximations are assumed. Given that to those who have previously championed the evidence procedure, "any inference must be based on strict adherence to the laws of probability theory, because any deviation automatically leads to inconsistency" [Gull, 1992], this should finish the matter. To do things correctly one should not use evidence. The same conclusion holds when instead of different real-valued parameters one is trying to decide amongst different basis sets, different neural net architectures, different regularizers, or anything else. It is more correct to evaluate the integrals (and sums, when one is comparing models) of 3(iii) directly than to use the evidence procedure.

Mathematical consistency and accuracy aside however, one might ask how the evidence procedure and equation 3 compare in other respects. The information necessary to use either the evidence procedure or equation 3 are the same; to calculate $P(f \mid \theta)$, both need the three distributions $P(f \mid \gamma)$, $P(\theta \mid f\ \gamma)$, and $P(\gamma)$. Therefore the only remaining relevant distinction between the two procedures is ease of use. Here too, equation 3 usually wins; it is often a *lot* simpler to use equation 3 than the evidence procedure. The conclusion is immediate: To estimate $P(f \mid \theta)$, there is no need to "evaluate evidence for different values of the parameters" and in fact one <u>should not do this</u>, since

it is a less accurate procedure which usually entails more work.[6]

Example Ib: Return to the distributions assumed in [MacKay 1991]. Evaluating 3(i) for his $P(f \mid \gamma)$ and $P(\gamma)$, one sees that MacKay is implicitly assuming that P(f), which equals $\int d\gamma \{P(f \mid \gamma) \times P(\gamma)\}$, equals $\int d\gamma_1 \{\exp(-\gamma_1 \times E_W(f)) / Z_W(\gamma_1)\}$. It is straight-forward to evaluate $Z_W(\gamma_1)$; the result is $(2\pi / \gamma_1)^{N/2}$, where N is the number of w_i. Therefore $P(f) = \int d\gamma_1 \{\exp(-\gamma_1 \times \Sigma_i w_i^2) \times (\gamma_1 / 2\pi)^{N/2} \propto [\Sigma_i (w_i)^2]^{-(N/2 + 1)} \equiv [W(\mathbf{w})]^{-(N/2 + 1)}$ ($\mathbf{w}$ being the vector of all the w_i). Note that as N increases, this prior becomes more and more biased towards small W. Note also that unless one imposes an infinitesimal "hole" of no allowed $\mathbf{w}$ values for $\Sigma_i (w_i)^2$ close to 0, P(f) is improper (i.e., its integral diverges). This reflects the un-normalizability of $P(\gamma)$. Normalized $P(\gamma)$ would, of course, give a different result.

Similarly, for the distributions assumed in [MacKay 1991], $P(\theta \mid f) = \int d\gamma \{P(\theta \mid f\, \gamma) \times P(f \mid \gamma) \times P(\gamma) / P(f)$. Integrating over γ_1, one gets $\int d\gamma_2 \{P(\theta \mid f\, \gamma_2) \times P(\gamma_2)\}$. Evaluating, this is proportional to $\int d\gamma_2 \{\exp(-\gamma_2 \times E_D(\theta_Y, \theta_X, f)) / Z_D(\gamma_2)\}$. Now it is straight-forward to rewrite $Z_D(\gamma_2)$ as $\Pi_{i=1}^{m} [\int d\{\theta_Y(i)\} \exp\{-\gamma_2 \times [\theta_Y(i) - f(\theta_X(i))]^2 / 2\}] \propto [\gamma_2]^{-m/2}$. Having done this, $P(\theta \mid f) \propto \{\Sigma_{i=1}^{m} [\theta_Y(i) - f(\theta_X(i))]^2\}^{-(m/2 + 1)}$. The implication is that the more elements in the training set, the more likely there is to be a small χ^2 value.

It is illuminating to compare the correct form for $P(f \mid \theta)$ with the form assumed in [MacKay 1991]; loosely speaking, the approximations of the evidence procedure as used in [MacKay 1991] are accurate only to the degree that one can fit the function $W^{-(N/2 + 1)} (\chi^2)^{-(m/2 + 1)}$ with an exponential of the form $\exp(-[\gamma'_1 W + \gamma'_2 \chi^2])$.

On the other hand, one should note that Bayesians often don't calculate the full average according to the posterior $P(f \mid \theta)$, but instead assume the distribution is sufficiently peaked so that they

can approximate that average with the mode. This is the approach used in [MacKay 1991]; rather than evaluating the full average, one simply finds the weight vector $\mathbf{w}^*$ which minimizes $\gamma'_1 W(\mathbf{w}) + \gamma'_2 \chi^2(\mathbf{w})$. This $\mathbf{w}^*$ satisfies the vector equality $\gamma'_1 \nabla W(\mathbf{w}) + \gamma'_2 \nabla \chi^2(\mathbf{w}) = 0$ (the delta functions being in $\mathbf{w}$ space). Now examine the case where $\gamma'_1 / \gamma'_2 = \chi^2(\mathbf{w}^*) / W(\mathbf{w}^*)$. In this case, the $\mathbf{w}^*$ found by using the evidence procedure is a solution to the equation $\chi^2(\mathbf{w}) \times \nabla W(\mathbf{w}) + W(\mathbf{w}) \times \nabla \chi^2(\mathbf{w}) = 0$. However, when m = N, the mode of the correct posterior, $W^{-(N/2 + 1)} \times (\chi^2)^{-(m/2 + 1)}$, is also given by the solution to the equation $\chi^2(\mathbf{w}) \times \nabla W(\mathbf{w}) + W(\mathbf{w}) \times \nabla \chi^2(\mathbf{w}) = 0$. So if one is content to approximate the mode with the mean, and if $N \approx m$, then the evidence procedure will give approximately the same answer as using the correct priors, assuming it arrives at parameter values having the ratio $\gamma'_1 / \gamma'_2 = \chi^2(\mathbf{w}^*) / W(\mathbf{w}^*)$.

Example IIb: Rather than perform the integrals for the entropic prior of [Gull 1989], it's illuminating to consider the following issue: As mentioned previously, many authors have argued that for any positive additive function f, the prior *must* be of the entropic form $P(f) = \exp(-\alpha \times S(f)) / Z_W(\alpha)$ for some constant α, where $Z_W(.)$ is a normalization factor, and S(f) is the entropy of f, $-\Sigma_{x \in X} [f(x) \ln(f(x))]$ (an integral for continuous valued X). Therefore the question arises: Is there *any* nowhere-negative function $P(\gamma_1)$, even one which isn't a proper probability distribution (i.e., isn't normalizable), such that with an entropic $P(f \mid \gamma)$, our prior $P(f) = \int_a^b d\gamma_1 \{P(f \mid \gamma_1) \times P(\gamma_1)\}$ is also entropic?

The answer to this question is yes, but the only such function $P(\gamma_1)$ is a Dirac delta function. To see this, for expository simplicity replace γ_1 with z and make the a and b limits implicit. Having done this, rewrite our proposed equality as $\exp(-\alpha \times S) = \int dz \{P(z) \times \exp(-z \times S)$ (the normalization factors having all been absorbed into P(z)). We must find an α and a function P(z) such that this equality holds for all S. First, evaluate both sides of the proposed equality for S = 0. The result is $\int dz P(z) = 1$; since P(z) is everywhere non-negative, this means that P(z) is a proper probability density. Now take the derivative with respect to S of both sides of the proposed equality, and evaluate

for S = 0. The result is $\alpha = \int dz(z \times P(z))$. (Since P(z) is a probability, this means that $a \leq \alpha \leq b$.) More generally, by taking derivatives k times we find that $\alpha^k = \int dz((z)^k \times P(z))$ for any positive integer k. But just by raising the single-derivative case to the k'th power we also know that $\alpha^k = [\int dz(z \times P(z))]^k$. Therefore we have $\int dz((z)^k \times P(z)) = [\int dz(z \times P(z))]^k = \alpha^k$. Now let the $L_i(z)$ be the (appropriately renormalized and rescaled) Legendre polynomials for the interval [a, b] of the integral. Since those polynomials are orthonormal, we can write $P(z) = \Sigma_{i=0}^{\infty} P_i L_i(z)$, where $P_i = \int dz\, P(z) L_i(z)$. If we write the j'th Legendre polynomial $L_j(z)$ as $\Sigma_{k=0}^{j} L_k \times z^k$, then $P_i = \Sigma_{k=0}^{i} L_k \times \alpha^k = L_i(\alpha)$. Plugging in to our expansion for P(z), we get $\Sigma_{i=0}^{\infty} [L_i(z) \times L_i(\alpha)]$. By the closure relation, this means that $P(z) = \delta(z - \alpha)$. QED.

The disquieting implication is that, in a certain sense, the arguments used to justify the entropic prior are not self-consistent.[7]

Sometimes one isn't interested directly in $P(f \mid \theta)$, but instead in (for example) the noise level. Amongst other things, proponents of the evidence procedure view it as a means for estimating the noise level; in examples I and II, according to the evidence procedure the ("most likely") noise level is given by γ'_2. This raises an obvious question: Although the evidence procedure might not be strictly correct, since equation 3 apparently doesn't provide any way of estimating noise from the data, isn't the use of the evidence procedure for that purpose (somewhat) justified?

The answer is no. It is true that 3(ii) is independent of the data, and therefore doesn't provide a means for inferring anything about noise levels from the data. However there are other distributions which one can calculate properly (i.e., by marginalizing out γ) which both depend on the data and reflect noise levels. For example, consider the distribution$P(\chi^2 = k \mid \theta)$. This is the probability, given the data, of a particular misfit (which can be loosely interpreted as a noise level) between the data and the truth. This distribution is given by $P(\chi^2 = k \mid \theta) = \Sigma_f P(f \mid \theta) \times \delta(\chi^2(f, \theta), k)$ (where for

clarity the dependence of χ^2 on the data and the target function is explicitly indicated). As an example, in the case studied in [MacKay 1991], $P(\chi^2 = k \mid \theta) = \int d\mathbf{w}\ P(\mathbf{w} \mid \theta) \times \delta(\chi^2(f_{\mathbf{w}}, \theta) - k)$. In general, for most parameterizations of target functions $f_{\mathbf{w}}$, it is difficult to evaluate this integral, even numerically. Fortunately, we are usually only interested in the moments of this distribution, and those moments *are* (relatively) straight-forward to evaluate. So for example, for the case investigated in [MacKay 1991], the expectation value of χ^2, given the data, is

$$\begin{aligned}
\langle \chi^2 \rangle|_\theta &= \int d\mathbf{w}\ P(\mathbf{w} \mid \theta) \times \chi^2(f_{\mathbf{w}}, \theta) \\
&= \int d\mathbf{w}\ [W(\mathbf{w})^{-(N/2 + 1)} \times (\chi^2(f_{\mathbf{w}}, \theta))^{-(m/2 + 1)} \times \chi^2(f_{\mathbf{w}}, \theta)] \ \ / \\
&\qquad\qquad \int d\mathbf{w}\ [W(\mathbf{w})^{-(N/2 + 1)} \times (\chi^2(f_{\mathbf{w}}, \theta))^{-(m/2 + 1)}]. \\
&= \int d\mathbf{w}\ [W(\mathbf{w})^{-(N/2 + 1)} \times (\chi^2(f_{\mathbf{w}}, \theta))^{-(m/2)}] \ \ / \\
&\qquad\qquad \int d\mathbf{w}\ [W(\mathbf{w})^{-(N/2 + 1)} \times (\chi^2(f_{\mathbf{w}}, \theta))^{-(m/2 + 1)}].
\end{aligned}$$

To proceed further, one needs to know the precise parameterization of target functions $f_{\mathbf{w}}$. However for many such parameterizations, the ratio of integrals is straight-forward to evaluate. For example, if target functions are parameterized linearly in terms of some basis with the $\mathbf{w}$ being the expansion coefficients (i.e., if $f_{\mathbf{w}}(x) = \Sigma^N_{i=1}\ w_i\ \phi_i(x)$ for some basis of functions $\{\phi_i(x)\}$), then the integrands are simply (products of) powers of the components of $\mathbf{w}$.

As a final note on the evidence procedure, it is important to realize that this paper does <u>not</u> claim that the evidence procedure doesn't work in the real world. The evidence procedure, used with a particular choice of $P(f \mid \gamma)$, $P(\theta \mid f\ \gamma)$, and $P(\gamma)$, is simply a rule for going from training sets to hypothesis functions, i.e., it is simply a distribution $P(h \mid \theta)$. Therefore (1) tells us that the quality of the guesses accruing from using the evidence procedure is determined by how "aligned" it is with $P(f \mid \theta)$. No proof has been offered in this paper that this "alignment" is poor. What this paper <u>does</u> say is that traditionally, people have tried to justify the evidence procedure by using Bayes' theorem in concert with the 3 distributions $P(f \mid \gamma)$, $P(\theta \mid f\ \gamma)$, and $P(\gamma)$, and that these attempted justifications are erroneous; one can not use the evidence procedure without knowing those distri-

butions, and if one knows those distributions (and believes them), there is no reason to use the evidence procedure rather than integrate out the posterior directly. If one truly wants to let "the data decide one's prior", then in general one must use a meta-generalization technique, like cross-validation ([Efron 1979]) or stacked generalization ([Wolpert 1992b]). See [Wolpert 1992a].

3. The Bayesian "proof" of Occam's razor

This section points out some flaws in the purported first principles "proof" of Occam's razor based on Occam factors.

The Bayesian "proof" of Occam's razor can be stated in several ways. Some of them are explicitly based on the use of evidence, which, given the results of the previous section, will be considered no further. Without using evidence, the argument can be summarized as follows (see [Loredo 1990, Jeffreys 1939, Gull 1989, MacKay 1991]).

Consider a parameter space P. Define a "model", or a "theorist", as a mapping from any $p \in P$ to a target function from X to Y. As an example, with P being the set of all possible quintuples of real numbers, the 4th order polynomial expansion using those five parameters is a model: the model is the mapping $\{\{p_0, p_1, p_2, p_3, p_4\} \rightarrow \Sigma_{i=0}^{4} \; p_i(x)^i\}$. Together, a model and a set of parameter values define a target function.

Now consider two models, m_1 and m_2, with associated parameter spaces P_1 and P_2. For expository simplicity, assume that both P_1 and P_2 are subsets of the same Euclidean vector space and have the same dimension. Assume further that $P_1 \subset P_2$. Let p_1 be a particular element of P_1, and similarly for p_2. Let f_1 be the function (m_1, p_1), and let f_2 be the function (m_2, p_2). $P(f_i \mid \theta) = P(m_i, p_i \mid \theta) = P(\theta \mid m_i, p_i) \times P(p_i \mid m_i) \times P(m_i) / P(\theta)$. Therefore

$$P(f_1 \mid \theta) / P(f_2 \mid \theta) = [P(\theta \mid m_1, p_1) \times P(p_1 \mid m_1) \times P(m_1)] / [P(\theta \mid m_2, p_2) \times P(p_2 \mid m_2) \times P(m_2)].$$

Now assume that both f_1 and f_2 have the same likelihood value, $P(\theta \mid f_i)$. Since we have no way of choosing between the two models, by the "principle of indifference" ([Loredo 1990]), $P(m_1) = P(m_2)$. Now assume the "uninformative" form for $P(p_i \mid m_i)$, namely a uniform density: $P(p_i \mid m_i) = 1 / [\int_{P_i} dp_i\ 1]$. Combining our results,

$$P(f_1 \mid \theta) / P(f_2 \mid \theta) = [\int_{P_2} dp_2\ 1] / [\int_{P_1} dp_1\ 1].$$

The ratio on the right-hand side of this equality is called the "Occam factor". It says that everything else being equal, the bias favoring f_1 over f_2 is given by (the reciprocal of) the ratio of the volume of P_1 to the volume of P_2. So models with a large a priori range of possible parameter values are 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, $V(P_i)$. 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 $P(f_1 \mid \theta) > P(f_2 \mid \theta)$ does not necessarily imply that $\langle E \rangle|_{h=f_1,\theta} < \langle E \rangle|_{h=f_2,\theta}$.) 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 Bayes-optimal guess, $\{x \in X \rightarrow \text{argmin}_{y \in Y}\ \Omega(y, x)\}$. This is the sense in which the argument recounted above can be viewed as an a priori "proof" of Occam's razor.

It is trivial to demonstrate that this "proof" of Occam's razor is flawed. The rest of this paper consists of several such demonstrations. The first demonstration involves simply examining the case where the function (m_1, p_1) equals the function (m_2, p_2). The Occam factor argument for this case says that if $V(P_1) \neq V(P_2)$, then $P(f_1 \mid \theta) \neq P(f_2 \mid \theta)$, despite the fact that $f_1 = f_2$!

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

above must be flawed. For example, if I just 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 common in the universe, parabolas or sine waves? 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. (Let $\sin(x)$ be the function (m_1, p_1), and let x^2 the function (m_2, p_2), for some arbitrary models m_1 and m_2. Then $P(\sin(x)) / P(x^2) = P(m_1, p_1) / P(m_2, p_2) = [P(p_1 \mid m_1) / P(p_2 \mid m_2)] \times [P(m_1) / P(m_2)]$ = the Occam factor for model m_1 over model m_2, if we make the assumptions made in the "proof" of Occam's razor recounted above.)

In fact, any argument assigning posteriors by engaging in parameter counting or by measuring the volume of allowed parameter space, <u>without any concern whatsoever for the model being used</u>, is patently silly (no such concern is present in the Occam factor argument). One can always bijectively map a space of parameters into a smaller 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. Without explicit consideration of what model one is using, there is nothing to distinguish between the pre-mapping and post-mapping scenarios. Therefore without such consideration, one can not possibly derive single-valued results. (See the discussion in [Wolpert 1990] concerning Occam invariances.)

The fundamental difficulty in the Occam factor argument is that it evaluates probabilities by using models, and there is *no reason whatsoever* to constrain the space of possible models. To see why this is a problem, modify the Occam factor argument by having two new models, m_1' and m_2', with the same parameters spaces as m_1 and m_2, namely P_1 and P_2 respectively. Let $m_1' = m_1$ except for one small change: the target function (m_1', p_1) does not equal f_1, but rather equals f_2. Other than that single exception, $m_1' = m_1$, i.e., $(m_1', p) = m_1', p) \; \forall \; p \in P_1$ such that $p \neq p_1$. Similarly, let $m_2' = m_2$ except for one small change: the target function (m_2', p_2) does not equal f_2, but rather equals f_1. (In [Wolpert 1990], similar modifications of the model are known as "Occam transformations".)

Neither m_1' and m_2' are "illegal" models. Furthermore, even if one wishes to view them as "unreasonable" models and therefore assign them low priors, there is no a priori reason not to set the ratio $P(m_1') / P(m_2')$ equal to $P(m_1) / P(m_2)$. However if we accept this ratio, and then go through the exact same reasoning used to derive the Occam factor (only using m_1' and m_2' rather than m_1 and m_2), we come to the exact *opposite* conclusion: $P(f_1 \mid \theta) / P(f_2 \mid \theta) = [\int_{P_1} dp_1\ 1] / [\int_{P_2} dp_2\ 1]$. Therefore the "Occam factor" argument, in its traditional form, is meaningless; it does not give a single unique answer.

The proper way around these problems is to ignore "models" altogether and consider probabilities over target functions directly. (After all, target functions are what we're ultimately interested in.) If one insists on considering probabilities over models however, then it is imperative that one be formal and rigorous.[8] 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. Only by following this strategy can we expect the problems recounted above to disappear. Fortunately, as it turns out, there is a well-known axiom which allows us to set the priors over models. This axiom is not the principle of indifference; the principle of indifference only applies if there is no relevant distinction between the models, and the whole point of the Occam factor argument is to show that there <u>is</u> a relevant distinction between the models. Rather than the principle of indifference, the appropriate principle 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, priors must be set in such a way that the difficulties in the Occam factor argument *can not* arise.

To see what this means, consider again the case where one has two models, m_1 and m_2, with associated parameter spaces P_1 and P_2. Assume that for both m_i, the mapping from the associated P_i to the space of target functions is one-to-one. Let M_1 be the mapping of m_1, taking an element

of P_1 to F. Similarly for M_2. Indicate by an asterisk (*) the corresponding inverse mappings. Let the set of target functions $M_1(p \in P_1)$ be written as F_1, and similarly for F_2.

The principle of consistency says that for any function f contained in both F_1 and F_2, and for any θ, $P(f \mid \theta)$ must be the same regardless of which model one uses, i.e., $P(m_1, M_1^*(f) \mid \theta)$ must equal $P(m_2, M_2^*(f) \mid \theta)$. (It doesn't matter what function we choose for f, so long as it is contained in both F_1 and F_2.) By Bayes' theorem, the principle of consistency means that $P(\theta \mid m_1, M_1^*(f)) / P(\theta \mid m_2, M_2^*(f)) = \{P(M_2^*(f) \mid m_2) / P(M_1^*(f) \mid m_1)\} \times \{P(m_2) / P(m_1)\}$. This result is exact; no proportionalities are involved.

Now for almost any noise process, since the target function $(m_1, M_1^*(f))$ equals the target function $(m_2, M_2^*(f))$ (both equal f), $P(\theta \mid m_1, M_1^*(f)) = P(\theta \mid m_2, M_2^*(f))$. For such processes, from the principle of consistency, we get $P(m_2) / P(m_1) = P(M_1^*(f) \mid m_1) / P(M_2^*(f) \mid m_2)\}$.[9] This ratio of the priors of the models exactly cancels the Occam factor. No problems arise from the freedom to redefine models. On the other hand, there is no more automatic Occam's razor: $P(f_1 \mid \theta) / P(f_2 \mid \theta) = P(\theta \mid f_1) / P(\theta \mid f_2)$.

Strictly speaking, the argument recounted above only holds if $F_1 \cap F_2 \neq \varnothing$. However even if F_1 and F_2 have zero intersection, we can always find a third model m_3 such that F_3 intersects both F_1 and F_2. We can now apply the above argument twice, once to go from $P(m_1)$ to $P(m_3)$, and once to go from $P(m_3)$ to $P(m_2)$. In this way we can use the principle of consistency to negate the Occam factor even when F_1 and F_2 have zero overlap.

Note how reasonable this negating of the Occam factor argument is. It is extremely 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 is a function solely of input-output mappings, and not (directly) a function of how such mappings are parameterized. After all, what is really at issue is not deciding between theorists, but rather deciding between theories. Moreover, in general one can more easily justify giving particular priors (e.g., uniform priors) to theories rather to than theorists. Indeed, as the arguments pre-

sented above illustrate, assigning priors to theorists, which consist of completely arbitrary sets of theories, rather than to the theories themselves, is just asking for paradoxes.

Just as it makes no claims concerning the real-world utility of the evidence procedure, this paper also makes no claims concerning how well Occam's razor works in practice. That issue is governed by equation (1). In fact, this paper would not even claim that it's a bad idea to use uniform $P(m_i)$ in the real world. What this paper does claim is that if one is going to try to argue things from "first principles", then the choice of $P(m_i)$ which exactly negates Occam factors is much easier to justify than a choice of uniform $P(m_i)$. Accordingly, the claim that the Occam factor argument somehow "proves" Occam's razor in a rigorous manner is seen to be flawed. At best, that argument simply allows one to equate the *assumption* of Occam's razor with the *assumption* of (relatively) uniform $P(m_i)$. This is about as rigorous as the one-line "proof" of Occam's razor which consists of identifying simpler functions as those with higher prior. By no stretch of the imagination is Occam's razor rigorously proven in any of these kinds of arguments.

Finally, it should be noted that there are a number of subtle issues in both the traditional Occam factor argument and its "principle of consistency" version given above. For example, strictly speaking, $P(f \mid \theta)$ is given by $\int dmdp \{\delta((m, p) = f) \times P(m, p \mid \theta)\}$ (where the delta function restricts the integral to those models m and parameters p giving the function f). Now there is no reason to believe that $P(m, p \mid \theta)$ is strongly peaked about one particular model and associated parameter value - in fact, even if one doesn't fully accept the principle of consistency argument, it's hard to escape its implication that $P(m, p \mid \theta)$ is not strongly peaked. Therefore there is no reason to think that we can approximate $P(f \mid \theta)$ by $P(m_a, M_a^*(f) \mid \theta)$ for some particular model m_a, an approximation which is implicit in both versions of the Occam factor argument. (A simple solution to this difficulty, of course, is to not ask for trouble by considering distributions across models but rather concentrate on distributions across target functions directly.)

Another subtle issue is that both the traditional and "principle of consistency" versions of the Occam factor argument consider distributions of the form $P(m_i, M_i^*(f) \mid \theta)$ rather than $P(M_i^*(f) \mid$

m_i, θ). Unlike the problem with summing over models however, the concentration on $P(m_i, M_i^*(f) \mid \theta)$ rather than $P(M_i^*(f) \mid m_i, \theta)$ is formally correct. This is because 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 "correct" model is.

Nonetheless, it is straight-forward to calculate $P(M_i^*(f) \mid m_i, \theta)$. By Bayes' theorem, this distribution equals $P(m_i, M_i^*(f) \mid \theta) / P(m_i \mid \theta)$. By the consistency principle, the numerator is independent of i. The denominator is a marginalization of the numerator: $P(m_i \mid \theta) = \int p_i \, P(m_i, p_i \mid \theta) = \int df \, \{d(p_i) / df \times P(m_i, M_i^*(f) \mid \theta)\} = \int df \, \{d(M_i^*(f)) / df \times P(m_i, M_i^*(f) \mid \theta)\}$. Different choice of models (i.e., different "parameterizations") mean different $d(M_i^*(f)) / df$; everything else is independent of i. This contrasts with the case of the distribution $P(m_i, M_i^*(f) \mid \theta)$, which by the principle of consistency is completely independent of i.

FOOTNOTES

1. Some authors use the phrase "Bayes-optimal" differently. See for example [Buntine and Weigend 1991].

2. So in particular, unless one penalizes mistakes exactly according to the square of the error a uniform prior does not imply that Laplace's law of succession is the Bayes-optimal way of guessing the bias of a coin based on the observation of a small number of flips of that coin.

3. Strictly speaking, one should require that the space of all functions parameterized by vectors of w_i values (i.e., by vectors **w**) spans the space F. Otherwise, one is implicitly assuming that those $f \in F$ which aren't in span(**w**) are impossible, with strictly zero probability; an assumption which is manifestly absurd. This difficulty is ignored in [MacKay 1991].

4. Strictly speaking, there are other terms in the entropy functional, involving "default models" and the like. Such terms are irrelevant to the present discussion.

5. Formally, one can write

$$P(f) = \Sigma_\theta P(f \mid \theta) P(\theta)$$

$$= \Sigma_\theta \{P(\theta \mid f\, \gamma') \times P(f \mid \gamma') / P(\theta \mid \gamma')\} \text{ (given the evidence procedure's } P(f \mid \theta))$$

$$= \Sigma_\theta \{[P(\theta \mid f\, \gamma') \times P(f \mid \gamma')] / [\Sigma_f (P(\theta \mid f\, \gamma') \times P(f \mid \gamma')]\}$$

$$= P(f \mid \gamma') \times \Sigma_\theta \{[P(\theta \mid f\, \gamma')] / [\Sigma_f (P(\theta \mid f\, \gamma') \times P(f \mid \gamma')]\}$$

This could be calculated exactly, if one wanted to. In general, the answer will depend on γ', Since γ' depends on θ, this means that P(f) isn't a proper "prior", and the evidence procedure "avoids setting priors". (Using this equation for P(f) and the fact that in the evidence procedure $P(f \mid \theta) \propto P(\theta \mid f\, \gamma') \times P(f \mid \gamma')$, one can write down immediately that $P(\theta \mid f) \propto P(\theta \mid f\, \gamma') /$

$[\Sigma_\theta \{[P(\theta \mid f\, \gamma')] / [\Sigma_f (P(\theta \mid f\, \gamma') \times P(f \mid \gamma')]\}]$. Just like P(f), this also depends on γ', in general.)

6. In the context of neural nets, this result means that once should just calculate 3(iii), and plug that in as the cost function to the gradient descent procedure. Full stop. No "evaluating the evidence", as in [MacKay 1991].

7. This is perhaps not too surprising, given that there are a number of other reasons for questioning the correctness of using an *exactly* entropic prior. One such reason is provided by all the non-Bayesian arguments for MaxEnt (e.g., [Tikochinsky et al. 1984ab, Shore and Johnson 1980 and references therein]). If one accepts the axioms/desiderata of those arguments, then the entropic prior can not be *exactly* correct if (as is conventional with Bayesians) one assumes a squaring metric in the error function, so that one should form estimates by taking means. This is because in general the entropic prior only gives the MaxEnt result in the approximation that the mean of the posterior equals its mode. Even if the posterior is very peaked, so that the mode is an excellent approximation to the mean, in general there is still a non-zero difference between the two. Since the "proper" Bayesian thing to do is take the mean, and since that mean disagrees with MaxEnt, we have a contradiction. At best, either the set of "intuitively obvious" axioms/desiderata of the original MaxEnt proofs are only approximately valid, or the "intuitively obvious" axioms/desiderata leading to the entropic prior are only approximately valid. (At worst, both sets of axioms/desiderata are only approximately valid, or perhaps not even particularly valid at all - see [Donoho et al. 1992], and in particular the comments of Gelman.)

8. One *might* try to argue that the Occam factor argument should only be made for "reasonable" models, and that m_1' and m_2' aren't "reasonable". Such a strategy completely negates all claims to formal rigor however, and doesn't resolve problems like those which arise when $f_1 = f_2$. Moreover, this strategy invites a whole host of objections: How does one define the dividing line between "reasonable" and "unreasonable" models? Is that dividing line sharp, or gradual? If the latter, then

shouldn't all models, even "reasonable" ones, have different priors, depending on their "reasonableness"?, etc., etc. As long as we're going to be content not to prove things rigorously and instead rely on notions of what is and is not "reasonable", we might as well simply start with (the eminently "reasonable") Occam's razor directly, and not pretend to "prove" it.

9. As an aside, note that if both f and f' $\in (F_1 \cap F_2)$, and if $P(M_1^*(f) \mid m_1) / P(M_2^*(f) \mid m_2) \neq P(M_1^*(f') \mid m_1) / P(M_2^*(f') \mid m_2)$, the principle of consistency can not be satisfied. (In addition, if this inequality holds, then the traditional Occam factor "proof" of Occam's razor explicitly falls apart.) If one wishes, this can be used to set restrictions on the possible forms of $P(p \in P_i \mid m_i)$.

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 and C. E. M. Strauss for interesting discussions. I would also like to thank Harry Martz for reading over a copy of this paper and providing the Lindely and Morris references.

REFERENCES

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

Blumer, A., et al. (1989). Learnability and Vapnik-Chervonenkis dimension. *Journal of the ACM*, **36**, 929-965.

Buntine, W., Weigend, A. (1991). Bayesian back-propagation. *Complex Systems*, **5**, 603-643.

Donoho et al., (1992). Maximum entropy and the nearly black object. Journal of the Royal Statistical Society B, **54**, 41-81.

Efron, B. (1979). Computers and the theory of statistics: thinking the unthinkable, *SIAM REVIEW*, **21**, 460-480.

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

Gull, S.F. (1992). Commentary at end of Donoho, D., et al., "Maximum entropy and the nearly black object", J.R. Statist. Soc. B, **54**, No. 1.

Haussler, D., et al. (1991). Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. In *COLT '91*, Morgan Kauffman, San Mateo, CA.

Hertz J., et al. (1991). *Introduction to the theory of neural computation*. Addison-Wesley, NY, NY.

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

Lindley, D.V., and Smith, A.F.M. (1972). Journal of the Royal Statistical Society B, **34**, 1-41.

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.

Morris, C. (1983). Journal of the American Statistical Association, **78**, 47-65.

Opper, M., and Haussler, D. (1991). Generalization performance of Bayes optimal classification algorithm for learning a perceptron. *Physical Review Letters*, **66**, 2677-2680.

Schwartz D., et al. (1990). Exhaustive Learning. *Neural Computation*, **2**, 374-385.

Seung H., et al. (1991). Statistical mechanics of learning from examples I, II. Submitted.

Shore, J.E., and Johnson, R. W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Transactions on information theory, vol. IT-26, 26-37.

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.

Tikochinsky et al., (1984a) Consistent inference of probabilities for reproducible experiments. Physical Review Letters. **52**, 1357-1360.

Tikochinsky et al., (1984b). Alternative approach to maximum-entropy inference. Physical Review A, **30**, 2638-2644. See also the follow-up discussion in Physical Review Letters, **55**, 336-337, 1985.

Tishby N., et al. (1989). Consistent inference of probabilities in layered networks: predictions and generalization. In *International Joint Conference on Neural Networks*, Vol. II, 403-409, IEEE, New York.

Valiant, L. (1984). A theory of the learnable. *Communications of the ACM*, **27**, 1134-1142.

Wolpert, D., and Stolorz, P. (1992). On the implementation of Bayes-optimal generalizers. Submitted.

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

Wolpert, D. (1992a). On the connection between in-sample testing and generalization error. *Complex Systems*, in press.

Wolpert, D. (1992b). Stacked Generalization. *Neural Networks*, in press.

Wolpert, D. (1992c). The relationship between the "statistical mechanics" supervise learning framework and PAC. Submitted.