Complex Systems 4 (1990) 319–368

The Relationship Between Occam's Razor and Convergent Guessing

David H. Wolpert*
Theoretical Division and Center for Nonlinear Studies,
Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Abstract. Occam's razor, the principle of parsimony, is a tool that finds application in many areas of science. Its validity has never been proven in full generality from first principles. Convergent guessing is the property that as more examples of an input–output mapping are provided, one would expect the modelling of that mapping to become more and more accurate. It too is widely used and has not been proven from first principles. In this paper it is shown that Occam's razor and convergent guessing are not independent — if convergent guessing holds, so does Occam's razor. (The converse of this statement is also true, providing some extra conditions are met.) Therefore, if you have reason to believe that your guesses are getting more accurate as you are fed more data, you also have reason to believe that application of Occam's razor will likely result in better guesses. Rather than attributes concerning how an architecture *works* (e.g., its coding length, or its number of free parameters), this paper is concerned exclusively with how the architecture *guesses* (which is, after all, what we're really interested in). In this context Occam's razor means that one should guess according to the "simplicity" of an architecture's guessing behavior (as opposed to according to the simplicity of how the architecture works). This paper deduces an optimal measure of the "simplicity" of an architecture's guessing behavior. Given this optimal simplicity measure, this paper then establishes the aforementioned relationship between Occam's razor and convergent guessing. This paper goes on to elucidate the many other advantages, both practical and theoretical, of using the optimal simplicity measure. Finally, this paper ends by exploring the ramifications of this analysis for the question of how best to measure the "complexity" of a system.

*Electronic mail address: dhw@coot.lanl.gov.

Introduction

This paper is concerned with the question of how best to generalize from a set of input–output examples (the *learning set*) to a full mapping from the input space to the output space. Many of the processes of inferring mathematical theories from data can be viewed as such a generalization problem [24]. More prosaically, much of the research in machine learning, neural nets, and artificial intelligence is concerned with this problem [3,5,12,14,15,17,18,21,22,25–30].

For the purposes of this paper, Occam's razor is taken to mean that, given a learning set of samples of an input–output mapping, choosing between two generalizers of that learning set according to which is simpler (according to an appropriate simplicity measure) will, on average, result in better generalizing (i.e., better guessing of what output should correspond to a novel input). In other words, Occam's razor is a way to build (hopefully) optimal generalizers. Convergent guessing refers to the shrinking of guessing error, on average, when the cardinality of the learning set increases. Much of machine learning is devoted to the problem of trying to maximize the rate of convergent guessing in certain severely limited contexts. Both convergent guessing and Occam's razor share the property that their utility is ultimately dependent on the kinds of generalization problems one is likely to encounter in our universe. As such, it might be impossible to provide a rigorous first principles proof of either one.

Rather than trying to prove either Occam's razor or convergent guessing from first principles, it is shown here that one of them necessitates the other (for large enough learning sets). In this sense, the validity of Occam's razor does not follow from the utility of limiting the number of degrees of freedom of the guesser, or some vague observation that nature works parsimoniously, or anything of this sort. Rather, it is a simple consequence of convergent guessing; if convergent guessing applies, so does Occam's razor, and vice versa. This relationship between convergent guessing and Occam's razor shouldn't be too surprising. One would expect both convergent guessing and Occam's razor to apply if (and only if) the function to be guessed that created the learning set has some "regularity" in its structure.[1]

This paper concentrates on the guessing distribution of architectures, that is, how they guess in response to a novel input. It completely ignores how the architectures are implemented. All implementation considerations (e.g., coding length, number of free parameters) are irrelevant to the issue of generalization accuracy. Minimizing such variables cannot *directly* gain you anything in generalization accuracy, since they are not directly concerned with generalization accuracy. Such accuracy, and its relation to Occam's razor and convergent guessing, is instead determined *in toto* by the guessing distribution. Only by concentrating on the guessing distribution can whole

[1]Indeed, one can use either the usefulness of Occam's razor when guessing, or the validity of convergent guessing, as a *definition* of "regularity" in the structure of the function that produced the learning set.

categories of architectures be treated at once (as opposed to on a case-by-case basis). Moreover, even when working with a single particular architecture, it is only by concentrating on the guessing distribution that one can hope to *prove* anything concerning the efficacy of Occam's razor and its relationship to convergent guessing.

The complete ignoring of implementation issues is one of the more important distinctions between the approach of this paper and previous work on the subject of Occam's razor (e.g., [16, 1]). This exclusive concentration on the guessing distribution is essential if one wishes to make broad statements concerning Occam's razor. In other words, it is essential if one wishes to address one of the most important issues of modern science: why does Occam's razor, which underlies the entire scientific enterprise, work so well in so many contexts?

Viewed in terms of guessing behavior, convergent guessing means that the average guess of the guessing distribution gets closer and closer to the "correct" answer as more and more input–output examples are made known. It holds if there is a correlation between the architecture and the "correct" input–output function that generates the learning sets. Often a simplicity value is associated with a particular implementation of an architecture. Occam's razor says that, on average, using the simplest implementation will give guesses close to the "correct" ones. This paper deduces an optimal measure of the simplicity of a given implementation of an architecture in terms of the architecture's guessing behavior (not in terms of how the architecture works). The minimal such simplicity value turns out to occur at the guess made more often than any other by the possible implementations of the architecture. Therefore Occam's razor says to make this guess. Roughly speaking, convergent guessing implies Occam's razor because if the average of the guessing is getting closer to the "correct" answer, then the mode of the guessing should be getting closer to it as well.

Section 1 of this paper is a presentation of a mathematical formalism for addressing Occam's razor and convergent guessing in an architecture-independent manner. Many if not all of the conventional means of applying Occam's razor are subsumed under this formalism (e.g., minimizing the number of axioms in a theory, minimizing the coding length of a computing device, minimizing the number of free parameters in a neural net).

Section 2 is an elucidation of some of the problems with the conventional ways of applying Occam's razor and how to remedy them. It is in this section that the optimal simplicity measure is derived. This section goes on to elucidate the many advantages of using this optimal simplicity measure. One such advantage is the fact that because this optimal measure is determined solely by the guessing distribution, it allows Occam's razor to be used even when the architecture possesses no obvious "handle" to be minimized (like a number of free parameters). Being independent of implementation considerations, Occam's razor as formulated in this paper is not restricted by such considerations. This section ends by relating the optimal simplicity measure

to those measures contained in the "conventional ways of exploiting Occam's razor," summarized in section 1.

Section 3, with the help of appendices A and B, proves the relationship between Occam's razor and convergent guessing. This relationship is a direct consequence of using Occam's razor with the optimal simplicity measure deduced in section 2. As such this relationship is yet another reason (in addition to all those mentioned in section 2) for using the optimal simplicity measure rather than any of the conventional simplicity measures.

Occam's razor is intimately related to many of the suggested measures of a system's "complexity" (e.g., algorithmic information complexity [4], logical depth [2], thermodynamic depth [13]). Accordingly the analysis of this paper has ramifications for the question of how best to measure a system's complexity. Section 4, which can be profitably read without first reading sections 1 through 3, describes these ramifications in some detail.

1. Occam's razor and convergent guessing

In this section the basics of a formalism for investigating Occam's razor and convergent guessing is presented. This formalism is architecture-independent, i.e., it can be applied to any computational architecture capable of supporting Occam's razor. Special cases of how this formalism applies to Occam's razor are minimizing the coding length of a tape fed into a universal Turing machine (UTM) [9], minimizing the number of weights in a neural net (NN) [21], striving for parsimony of rules in a classifier system (CS) [8], and Rissanen's minimum description length scheme (MDL) [20]. This section starts by defining architectures and simplicity measures, the central concepts in the analysis of this paper. This section ends by giving Occam's razor and then convergent guessing rigorous definitions in terms of these concepts.

1.1 Architectures and simplicity measures

In investigating Occam's razor we are interested in functions from inputs to outputs. In this paper we will assume that the inputs $\in X \subseteq \mathbb{R}^m$ and the outputs $\in Y \subseteq \mathbb{R}^n$. Although the analysis of this paper is phrased for $n = 1$, it carries over for any n. A function $f(x)$ from X to Y is called a *generating* or *parent* function. A finite set of samples of a parent function is called a *learning set*, and is usually indicated by the symbol θ: $\theta \in \{(\{X\} \times \{Y\})^p; p \in \mathbb{Z}^+\}$. For mathematical convenience, no learning set is allowed to contain more than one element with the same x value; it is assumed that there are no inconsistencies or duplications in the learning set. A *question*, usually indicated by the symbol q, is any element of the input space X. Our problem is how best to guess the output corresponding to an arbitrary question, given only the learning set, i.e. how best to generalize the full function from only some instances of it.

To apply Occam's razor, any function from inputs to outputs must be broken up into two parts. The first part is a set of elements called the *defin-*

ing set and is usually indicated by the symbol ϕ. The second part, called a *method*, is usually indicated by the symbol M and maps the defining set to functions from inputs to outputs. For example, the weights and connectivity in a neural net could serve as a defining set, with the method being the mapping of such weights into a function taking inputs to outputs. A particular method M together with a set of defining sets Φ and input and output spaces X and Y is called an *architecture*, $\{M, \Phi, X, Y\}$. Formally, $M : (\phi \in \Phi) \rightarrow \{f : X \rightarrow Y\}$. ϕ uniquely *defines* the guessed function in all of its details, while M is the *method* by which this defining is achieved. A defining set is loosely equivalent to what Pearl calls a "sentence," whereas a method is loosely equivalent to what he calls an "interpreter" [16].

If M is a method and ϕ is a defining set, (M, ϕ) is defined to be the set of all pairs, {question, output to the question guessed using M and ϕ}. It is the guessed function defined by ϕ. The value of this function for an input q is written as $(M, \phi)(q)$.

To apply Occam's razor, it is necessary to have a measure of the "simplicity" of a function (M, ϕ). As commonly used (see examples below), such a measure is a function only of the defining set. A simplicity measure of a defining set ϕ is written as $S(\phi) : \{S : (\phi \in \Phi) \rightarrow D \subseteq \mathbf{R}^+\}$. S is surjective over the set D. Along with D, Φ, the set of allowed arguments of a simplicity measure S, is implicit when writing "$S(\phi)$". To avoid infinities, in practice D is bounded from above. To normalize different simplicity measures and different architectures, we require that for any simplicity measure S and any architecture $\{M, \Phi, X, Y\}$ $\exists\, \phi \in \Phi$ such that $S(\phi) = 0$. We want such a normalization point to correspond to as *uncomplicated* a defining set as possible, so higher $S(\phi)$ means a more complicated defining set ϕ.

The term "simplicity" is used because "information content" already means something [23], and "complexity" doesn't necessarily mean the quantity that must be minimized to apply Occam's razor; the term carries other connotations as well (see, for example, [13], and chapter 12 of [9]).

Some examples of methods, defining sets, and simplicity measures appear in table 1. In example 5 in table 1, $k \equiv$ the smallest possible encoded TM length, given a particular coding scheme for the UTM. It is subtracted from the simplicity measure so that the measure meets the requirement that $0 \in D$. Similarly, in example 4 we subtract 1 when calculating the simplicity measure to take care of the fact that every TM must have a start state [9].

Occam's razor means use the defining set with the smallest simplicity measure subject to some restriction (usually that of reproducing the learning set). For example, in example 1 of table 1 Occam's razor says that, given a set of basis functions, you should generalize from a learning set by finding the smallest subset of that set of basis functions such that a linear combination of the functions of that subset goes through all the points in the learning set. That linear combination of basis functions is your guess for the parent function. In example 3, Occam's razor says that you should generalize from a learning set with the smallest neural net that can reproduce the learning set. A more formal statement of Occam's razor occurs later in this section.

	Method	Defining set	Simplicity measure
1.	input = x; output = $\sum_{i=1}^{n} a_i h_i(x)$	The set $\{a_i\}$.	n, the cardinality of $\{a_i\}$.
2.	Using math and established science, use the set of axioms to create a theory (i.e. a mapping from inputs to outputs).	The set of axioms.	Vague; approximately given by the number of axioms.
3.	The input–output algorithm of conventional feedforward neural nets.	The architecture of a particular net, suitably encoded.	Number of hidden neurons (or alternatively the number of weights) in a net.
4.	The input–output algorithm of Turing machines.	The state transition table of a particular TM.	The number of states in the transition table of a particular TM, -1.
5.	The input–output algorithm of a universal Turing Machine.	The code of a particular Turing machine, encoded on a UTM's tape.	Length of the code of the Turing machine encoded on a UTM's tape, $-k$.
6.	Rissanen's MDL scheme.	The parameterized conditional distribution $f_{\Theta,k}(x_{t+1}\|x_t)$ along with the learning set $\{x_n\}$.	See [20].[2]

Table 1: Some examples of methods, defining sets, and simplicity measures.

Nothing in the list of examples making up table 1 is supposed to be unique. For example, some might prefer to measure the simplicity of a Turing machine as the number of steps it takes to reproduce the learning set. Using such a measure rather than the one given in the list above is akin to measuring the complexity of a sequence of numbers in terms of its logical depth [2] rather than in terms of its algorithmic information complexity [4].

[2]Note the peculiarity of Rissanen's scheme, which states that since only conditional probability distributions are used, no guessing can be done without providing some data points, i.e., the defining set must contain the elements of the learning set.

Although "simplicity" isn't synonymous with Shannon's information, Shannon's information can be used as a simplicity measure for certain methods (as with Rissanen's MDL scheme — see [7]). Similarly, although they are not necessarily identical, simplicity and complexity are intimately related with one another. For example, the algorithmic information complexity of a sequence of numbers is simply the minimal simplicity value of a defining set that produces the sequence, using the method and simplicity measure of example 5.[3,4] The connection between simplicity measures and complexity is examined in more detail in section 4.

1.2 Required properties of architectures and simplicity measures

We do not want to allow information from a learning set to be hidden in a method, thereby allowing us to reproduce the learning set using defining sets with simplicity measure arbitrarily (and meaninglessly) close to 0. Therefore it is required that any method must be able to reproduce any learning set exactly. This means that the method cannot have any data "hidden" in it pertinent to some particular learning set. Formally,

(1.1) For all architectures $\{M, \Phi, X, Y\}$ and for all learning sets $\theta = \{(\{X\} \times \{Y\})^p;\ p \in \mathbb{Z}^+\}$ of finite cardinality, there exists a defining set $\phi \in \Phi$ such that $\theta \subseteq (M, \phi)$.

For example 1 from the list above, requirement (1.1) will be met if the $h_i(x)$ are a complete basis for the space $X \times Y$ of allowed input–output functions. For the Turing machine examples, (1.1) is roughly equivalent to the requirement of computational universality. For the conventional feedforward neural net of example 3, it is equivalent to the requirement that for every

[3]Strictly speaking, to make the correspondence between algorithmic information complexity and minimal simplicity measure we need to make a correspondence between a finite sequence of numbers and an input–output mapping. Perhaps the simplest way to do this is to use the same simplicity measure as in example 5, but to modify the defining sets to be the entire contents of the tape fed to the UTM and to modify the method so that the domain of the input–output mapping is a single (arbitrary and unspecified) number. The sequence of numbers generated by running the UTM is then viewed as the decimal expansion of the output of the architecture, where this output is a rational number between 0.0 and 1.0 whose decimal expansion consists of a finite number of digits (assuming the UTM halts). The defining set now includes all of the tape fed in to the UTM. Under this scheme, all learning sets have cardinality 1, and equation (1.1) (see below) is immediate. Since (1.2) (see below) is a property of defining sets (i.e., a property of the rules for running the TM mapping the input to the output) and not of how inputs and outputs are interpreted, it too holds, simply because it holds for the TM examples in the list above.

[4]TMs, whether used in simplicity measures or in algorithmic information complexity, suffer from the Halting Problem [9]. As a result, the function $(M, \phi)(q)$ in examples 4 and 5 might not be defined for all $q \in X$. This difficulty can be avoided, if so desired: simply set a maximum on the number of possible steps of the TM. Essentially, this is equivalent to reducing the TM to a deterministic finite automata. Another way to avoid the problem of a TM that never halts is to simply expand the output space to include a new symbol representing "no answer."

finite learning set there is a set of weights such that the resultant neural net reproduces the learning set.[5,6]

Since the simplicity is supposed to correspond to Occam's razor as humans usually use it, we make two additional requirements of $S(\phi)$, the second requirement being a logical consequence of the first. We state these requirements here for the case where we desire exact reproduction of the learning set and where Φ is finite. (Φ is finite, for example, whenever the system is being emulated on a finite digital computer.) Nonfinite Φ will be dealt with in section 2.

(1.2) Take any architecture $\{M, \Phi, X, Y\}$, defining set $\phi \in \Phi$, simplicity measure S, and learning set $\theta \subseteq (M, \phi)$. Then for all simplicity values $s' \in D$ where $s' > S(\phi)$, there exists a defining set $\phi' \in \Phi$ such that $\theta \subseteq (M, \phi')$ and $S(\phi') = s'$.

(1.3) We are given an architecture $\{M, \Phi, X, Y\}$, a learning set $\theta \in \{(\{X\} \times \{Y\})^p;\ p \in \mathbb{Z}^+\}$ of finite cardinality, and a simplicity measure S. Let Φ_σ be the set of all defining $\phi \in \Phi$ sets such that $\theta \subseteq (M, \phi)$ and $S(\phi) = \sigma$. Then the cardinality of Φ_σ is a nondecreasing function of σ over the range D.

In (1.3), Φ_σ is implicitly a function of M and θ as well as σ.

Intuitively speaking, (1.2) is the requirement that if you can reproduce a learning set with a simple system, then you can reproduce it with a more complicated system as well. If need be, the extra complexity can be designed to have no effect on the guessing.

Equation (1.3) is the requirement that the more complicated a defining set ϕ, the higher the number of functions with its simplicity (i.e., the higher the number of functions in the set $\{M, \phi'\};\ \phi' \in \Phi_{S(\phi)}$). In other words, the more complicated a defining set is required to be, the less of a constraint that requirement imposes on its allowed guessing. A property similar to (1.3) also appears in Pearl's work on parsimony [16]. Although (1.2) implies (1.3), the reverse is not true. Together, (1.2) and (1.3) relate the simplicity of a set of

[5] In [30] it is shown that the two variations of (1.1) also hold: "for all defining sets $\phi \in \Phi$ and for all learning sets $\theta \in \{(\{X\} \times \{Y\})^p;\ p \in \mathbb{Z}^+\}$, there exists a method M such that $\theta \in (M, \phi)$" and "for all methods M and all defining sets $\phi \in \Phi$, there exists a learning set θ such that $\theta \in (M, \phi)$." Other aspects of the mathematics of methods and defining sets are investigated in [30] as well. For example, it is shown there how to perform arithmetic operations over the space of methods.

[6] Sometimes (1.1) has to be amended so as to not run astray of cardinality arguments. For example, for a method taking defining sets consisting of a finite number of integers, (1.1) cannot be obeyed exactly for learning sets consisting of finite numbers of reals. In such cases, (1.1) is assumed to be modified to read "Over an arbitrarily large region, for all methods M, and for all learning sets θ, there exists a defining set ϕ such that a set of numbers that constitute an arbitrarily good (but not necessarily perfect) approximation to the elements of θ is contained in (M, ϕ)." ("Arbitrarily good," for example, could mean that the sum of the squares of the differences between θ and (M, ϕ) is less than some pre-set threshold, δ, over some region of questions.) For these cases the variations of (1.1) mentioned in footnote 5 are modified in a similar way.

defining sets to the number of degrees of freedom that set has, as measured by its ability to reproduce learning sets.

For example 1 from the list above, (1.2) follows from the fact that a_{n+1} can always equal zero, so any function that can be represented using n coefficients can also be represented using $n+1$ coefficients. For example 3, (1.2) follows from the fact that the values of additional weights can always be set to zero. Similarly, for the Turing machine examples it follows from the fact that the state transition table can be designed so that additional states are never reached. Every simplicity measure I know of that is used for generalizing from a learning set either obeys (1.2) and (1.3) directly or can be modified slightly so as to obey them.

1.3 Restrictions imposed on defining sets by learning sets

In applying Occam's razor, one does not work with the entire set Φ. In general, one works with a subset of Φ that is determined by the learning set. For example, one might only consider those defining sets that, in concert with the method, reproduce the learning set. Such a restriction on Φ is reflected in the phrasing of (1.1) through (1.3). It is desirable, however, to expand the formalism to allow more general kinds of restrictions than those of (1.1) through (1.3). This entails modifying (1.1) through (1.3).

Formally, restrictions on Φ are expressed via mappings from Φ to Φ; given an architecture (M, Φ, X, Y) and a learning set θ, a *restriction* $R_{M,\theta}(\Phi)$ is a mapping, determined by M and θ, from Φ to a subset of itself: $\{R_{M,\theta} : \Phi \rightarrow \Phi' \subseteq \Phi\}$. Requiring reproduction of the learning set is a restriction — given the method M and learning set θ, it replaces the set of all $\phi \in \Phi$ with the set of all $\phi \in \Phi$ such that $(M, \phi) \supset \theta$. If the data is noisy, it might be desired to restrict our attention to some set of defining sets other than those which perfectly reproduce the learning set. The definition of restriction given here is broad enough to accommodate such situations.

Let $\theta = \{(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)\}$. Then the *$\theta$-restricted input space*, X_θ, is defined as the set X with θ's input values removed; $X - \{x_1\} - \{x_2\} - \ldots - \{x_N\}$. In this paper, to allow the analysis to be as broad as possible, a "restricted" form of (1.1) will be used:

(1.1′) For all architectures $\{M, \Phi, X, Y\}$, for all learning sets $\theta \in \{(\{X\} \times \{Y\})^p;\ p \in \mathbb{Z}^+\}$ of finite cardinality, for all pairs $(x \in X_\theta, y \in Y)$, and for all restrictions R, there exists a defining set $\phi \in R_{M,\theta}(\Phi)$ such that $(M, \phi)(x) = y$.

The restricted form of (1.1) says that if you have a learning set and resultant restriction on what defining set you are allowed to consider, you can still find a defining set to guess any input–output pair, so long as that pair lies outside of the learning set that determined the restriction. If the restriction is reproduction of the learning set, then the restricted form of (1.1) holds for a particular architecture iff the unrestricted form of (1.1) holds for that architecture. Whether the restriction is reproduction of the learning

set or something else, no restriction-architecture pairs will be allowed if they do not obey the restricted form of (1.1). From now on, whenever "equation (1.1)" is referred to in this paper, what is really meant is the restricted form of (1.1), (1.1').

In a similar manner, whenever "equation (1.2)" is referred to what is really mean is the restricted form of (1.2):

(1.2') Take any architecture $\{M, \Phi, X, Y\}$, learning set $\theta \in \{(\{X\} \times \{Y\})^p; p \in \mathbb{Z}^+\}$ of finite cardinality, restriction R, defining set $\phi \in R_{M,\theta}(\Phi)$, simplicity measure S, and pair $(x \in X_\theta, (M, \phi)(x))$. Then for all simplicity values $\sigma \in D$ where $\sigma > S(\phi)$, there exists a defining set $\phi' \in R_{M,\theta}(\Phi)$ such that $(M, \phi')(x) = (M, \phi)(x)$ and $S(\phi') = \sigma$.

Similarly, from now on "equation (1.3)" means the restricted form of (1.3):

(1.3') We are given an architecture $\{M, \Phi, X, Y\}$, a learning set $\theta \in \{(\{X\} \times \{Y\})^p; p \in \mathbb{Z}^+\}$, a restriction R, and a simplicity measure S. Let Φ_σ be the set of all defining sets $\in R_{M,\theta}(\Phi)$ with simplicity measure σ. Then the cardinality of Φ_σ is an increasing function of σ over the range of D.

Note that Φ_σ is implicitly a function of M and θ as well as σ. As before, we are assuming that Φ is finite. For the restriction of reproducing the learning set, (1.1) through (1.3) are logically equivalent to (1.1') through (1.3'); for this case the restricted forms imply the unrestricted forms and vice versa.

For the rest of this paper, all architectures, simplicity measures, and restrictions are assumed to obey the restricted forms of (1.1) through (1.3). Note that by performing the analysis in the context of restrictions, we are implicitly concerned with the guessing distribution. The analysis has been narrowed to consideration of the behavior of the generalizer in response to a particular question, given that the learning set is reproduced (or given that some other restriction is met). Moreover, by considering θ-restricted input spaces we are making the analysis completely independent of the irrelevant and trivial issue of reproducing the learning set (or meeting some other restriction). Anyone who can write a lookup table can reproduce a learning set — we are after bigger game here.

1.4 Occam's razor

We now have the terminology necessary for giving a precise formulation of Occam's razor. For the purposes of this paper Occam's razor is the following statement: "Of two possible explanations for a set of data, with no way to choose between the two explanations except according to their simplicity, choosing that explanation that is simpler gives better guessing, on average." Translating, "set of data" means a learning set θ generated from a parent function, said function constituting the "correct" explanation of the data. (The cardinality of that learning set will be indicated by the symbol n.) "Simpler" means lower simplicity measure. An "explanation" for a learning

set is an element of $R_{M,\theta}(\Phi)$. "Better guessing" refers to guessing behavior for questions chosen from X_θ — through the restriction the "set of data" itself fixes the guessing behavior for all other questions. Translated in full, Occam's razor is the following statement: "Assume you are given an architecture, (M, Φ, X, Y), a parent function $f : X \to Y$, a learning set cardinality n, a simplicity measure S, and a restriction R. Randomly pick a learning set θ with cardinality n such that $f(x) \supset \theta$. Take two defining sets ϕ_1 and ϕ_2, both contained in $R_{M,\theta}(\Phi)$, where $S(\phi_1) < S(\phi_2)$. Then, on average, generalizing from θ using (M, ϕ_1) gives guesses closer to $f(x \in X_\theta)$ than generalizing with (M, ϕ_2)." "On average" means averaged over all such pairs of defining sets, over all n-element learning sets θ chosen from $f(x)$, and over all questions from X_θ. It is implicitly assumed that the cardinality of X exceeds n so that X_θ is not empty. For mathematical convenience, we will here only consider defining sets ϕ_1 and ϕ_2 where $S(\phi_1)$ is the largest simplicity value smaller than $S(\phi_2)$.

To define Occam's razor in a completely rigorous manner, first make the notational definition

$$\langle \text{stuff} \rangle_{(\text{constraints})} \equiv \frac{\sum_{\text{constraints}} \text{stuff}}{\sum_{\text{constraints}} 1}$$

($\int$'s replace $\sum$'s for continuous constraints.)

We will say that "Occam's razor applies" for an architecture (M, Φ, X, Y), a cardinality n, a parent function $f(x)$, a simplicity measure S, and a restriction R, if the *Occam error*,

$$E_{\text{Occam}} \equiv \left\langle \sum_{S''} \{ \langle |(M, \phi_{ij})(q) - f(q)| \rangle_{\phi_{ij} \in R_{M,\theta_i}(\Phi), S(\phi_{ij}) = S'} - \langle |(M, \phi'_{ij})(q) - f(q)| \rangle_{(\phi'_{ij} \in R_{M,\theta_i}(\Phi), S(\phi'_{ij}) = S'')} \} \right\rangle_{(\theta_i \subset f(x), q \in X_{\theta_i})}$$

is less than zero. In this definition of E_{Occam} the averaging over learning sets θ_i comes first, i.e., the sum over all possible learning sets is the outermost sum going into the average. i indexes these learning sets, all of which are assumed to have cardinality n. The summing over S'' is over all simplicity measure values meeting the following requirement: for the learning set θ_i and the method M, there is both at least one defining set from the restricted set of defining sets with that simplicity measure value S'', and there is at least one defining set from the restricted set of defining sets with simplicity measure value smaller than S''. S' is the largest such simplicity measure value $< S''$. (We sum over S'' rather than average over it to make the mathematical bookkeeping simpler.) ϕ_{ij} is the set of defining sets $\in R_{M,\theta_i}(\Phi)$. For a given θ_i, these defining sets are indexed by j.

We could modify the definition of the Occam error to make use of the squares of the guessing errors rather than the absolute values of those errors, but for the purposes of this paper absolute values will be easier to work with. We could also modify the definition of E_{Occam} to concern differences between

a simplicity measure value and *all* simplicity measure values smaller than it (not just the largest simplicity smaller than it). Similarly, rather than working with the Occam error we could work with the covariance between $S(\phi_{ij})$ and $|(M, \phi_{ij})(q) - f(q)|$, or analyze the correlation between the two. In general, however, it is easier to work with the Occam error as given here than with any of these alternative measures of the efficacy of Occam's razor.

Because $(a_0 - a_1) + (a_1 - a_2) + \ldots + (a_{n-1} - a_n) = a_0 - a_n$, the Occam error can be rewritten as

$$E_{\text{Occam}} = \Big\langle \langle |(M, \phi_{ij})(q) - f(q)| \rangle_{(\phi_{ij} \in R_{M,\theta_i}(\Phi), S(\phi_{ij}) = S_{\min}(\theta_i))} - \langle |(M, \phi'_{ij})(q) - f(q)| \rangle_{(\phi'_{ij} \in R_{M,\theta_i}(\Phi), S(\phi'_{ij}) = S_{\max}(\theta_i))} \Big\rangle_{(\theta_i \subset f(x), q \in X_\theta)} \tag{1.4}$$

$S_{\max}(\theta_i)$ is the maximum simplicity measure value such that, for the given learning set, there is a defining set from $R_{M,\theta i}(\Phi)$ with this value, and $S_{\min}(\theta_i)$ is the minimum such value.[7] We can just as easily use (1.4) as the original definition of E_{Occam} to determine whether or not Occam's razor applies. When the values of S are continuous (i.e., when the set S has nonzero Lebesgue measure) the largest simplicity measure value $< S''$ is not well defined, so the original definition of Occam error given above is meaningless. For such cases we define E_{Occam} via equation (1.4).

As an example consider example 1 from table 1 of the linear combination of basis functions $h_i(.)$, with the restriction being reproduction of the learning set. You are given a learning set θ of cardinality n generated from a parent function f(.). Let m be an integer such that there exists an m-element linear combination of the $h_i(.)$ that reproduces θ. Compare the average error (as determined by $f(.)$) when guessing the response to a question q with an m-element linear combination to the average error when guessing with an $(m+1)$-element linear combination. If, averaged over all n-element learning sets, all m, and all q's, the error is smaller with the m-element linear combination, then the Occam error < 0 and Occam's razor applies.

The following simple-minded model illustrates why Occam's razor is reasonable. Assume that our restriction is reproduction of the learning set, so the larger the learning set, the fewer the allowed defining sets: if $\theta_2 \supset \theta_1$, then $R_{M,\theta_1}(\Phi) \supset R_{M,\theta_2}(\Phi)$. Furthermore, assume that we have a finite number of possible defining sets and therefore (due to (1.1')) a finite X. Let f_Φ be the set of defining sets compatible with the full parent function $f(.)$;

[7] Note that the Occam error being negative does not imply that the average error magnitude of those defining sets with the lowest simplicity measure is smaller than the average error magnitude of *all* the defining sets. Occam's razor as defined in this paper refers to the difference in the average error magnitude between those defining sets with the lowest simplicity measure and those defining sets with the largest simplicity measure. The question of whether or not the guess of the simplest defining sets is better than the average defining set's guess will be addressed later in this paper. Although intimately related to the Occam error, the guessing error of the simplest defining sets is not determinable from the Occam error alone.

$(M, \phi)(q) = f(q)\ \forall\ q \in X,\ \phi \in f_\Phi$. By (1.1'), f_Φ is nonempty. Define $\sigma \equiv \min(S(f_\Phi))$, the minimal S value of the elements of f_Φ, and define N to be the cardinality of X. The set of defining sets that reproduce a learning set generated from $f(.)$ shrinks as that learning set grows. This culminates with N element learning sets, which fully specify $f(.)$ and which are compatible only with the defining sets of f_Φ. If we are fortunate in $f(.)$ and in our choice of simplicity measure, then this shrinking of $R_{M,\theta}(\Phi)$ eventually results in there being an integer $n < N$ such that for any learning set θ of cardinality n, every defining set $\in R_{M,\theta}(\Phi)$ but $\notin f_\Phi$ has simplicity measure $> \sigma$. For such a case, any defining set with simplicity $S_{\min}(\theta)$ is contained in f_Φ, and therefore guesses $f(q)$ exactly. This means that every term in the average over q and θ_i in equation (1.4) is negative, for this case, for any learning set cardinality $\geq n$. Therefore the entire Occam error is negative for such learning set cardinalities. Since we can allow some terms in the average over q and θ_i in equation (1.4) to be positive and still have a negative Occam error, there will often in fact be an integer m slightly less than n such that the Occam error is negative for any learning set cardinality $\geq m$.[8]

1.5 Convergent guessing

Now on to convergent guessing, the property of one's guessing becoming more and more accurate as the learning set size increases. We need a measure of the average guessing error we would expect for a method M guessing a parent function $f(x)$, for an n-point learning set chosen from the parent function. Define the *expected guessing error* E_{guessing} of an architecture (M, Φ, X, Y), a restriction R, a parent function $f(x)$, and a learning set cardinality n, as the average of the absolute value of the guessing error for an average question. The average is over all n-point learning sets chosen from $f(x)$, and, for each of the learning sets θ, over all questions from X_θ and over all defining sets that are contained in $R_{M,\theta}(\Phi)$:

$$E_{\text{guessing}} = \langle \langle |(M, \phi_{ij})(q) - f(q)| \rangle_{(\phi_{ij} \in R_{M,\theta_i}(\Phi))} \rangle_{(\theta_i \subset f(x), q \in X_{\theta_i})} \tag{1.5}$$

As in (1.4), i indexes all the learning sets going into the average, and j indexes the set of defining sets contained in $R_{M,\theta_i}(\Phi)$. All θ_i are assumed to contain n elements.

We can rewrite the expected guessing error E_{guessing} in the form

$$E_{\text{guessing}} = \left\langle \frac{\sum_{\phi_{ij} \in R_{M,\theta_i}(\Phi)} \{|(M, \phi_{ij})(q) - f(q)|\}}{\sum_{\phi_{ij} \in R_{M,\theta_i}(\Phi)} \{1\}} \right\rangle_{(\theta_i \subset f(x), q \in X_{\theta_i})}$$

$$= \left\langle \frac{\sum_{y \in Y} \sum_{\phi_{ij} \in R_{M,\theta_i}(\Phi)} \{\delta((M, \phi_{ij})(q), y)|y - f(q)|\}}{\sum_{\phi_{ij} \in R_{M,\theta_i}(\Phi)} \{1\}} \right\rangle_{(\theta_i \subset f(x), q \in X_{\theta_i})}$$

[8] Note that this argument does not by any means constitute a proof of Occam's razor; it simply describes some very limited circumstances in which Occam's razor might apply. This argument is very similar to the conventional machine learning arguments concerned with constructing a learning set large enough to force unique generalization when the parent function is known beforehand to be from a particular "concept class."

When the y are discrete-valued, $\delta(.,.)$ is the normal Kronecker delta function. When the index j on the ϕ_{ij} is continuous-valued, the sum over defining sets is replaced by an integral. When the y are continuous-valued as well, $\delta(.,.)$ is replaced by the Dirac delta function, and the sum over y is replaced by an integral. (Due to (1.1'), the y cannot be continuous-valued if the index j on the ϕ_{ij} is discrete-valued.) We can rewrite the denominator inside the outermost brackets:

$$\sum_{\phi_{ij} \in R_{M,\theta_i}(\Phi)} \{1\} = \sum_{y \in Y} \sum_{\phi_j \in R_{M,\theta_i}(\Phi)} \{\delta((M, \phi_{ij})(q), y)\}$$

Therefore, if we multiply and divide the formula for E_{guessing} by the total number of possible guesses and then by the total number of defining sets, we get the following:

$$E_{\text{guessing}} = \left\langle \frac{\langle O_{\theta,q}(y) | y - f(q) | \rangle_{(y \in Y)}}{\langle O_{\theta,q}(y) \rangle_{(y \in Y)}} \right\rangle_{(\theta \subset f(x),\ q \in X_\theta)} \tag{1.6}$$

where when the index j on the ϕ_{ij} is discrete,

$$O_{\theta,q}(y) \equiv \frac{\sum_{\phi \in R_{M,\theta}(\Phi)} \{\delta((M, \phi)(q), y)\}}{\sum_{\phi \in R_{M,\theta}(\Phi)} \{1\}}$$

The subscript i has been dropped from the learning sets θ_i in (1.6) since there is no longer any ϕ_{ij} whose indices have to be matched to those of the learning set. For this case of discrete valued j, $O_{\theta,q}(y)$ is the fraction of defining sets from $R_{M,\theta}(\Phi)$ that, for method M, question q, and learning set θ, make the guess y.

For continuous-valued j and discrete-valued y, we have sums over outputs and integrals over defining sets, so (1.6) holds for

$$O_{\theta,q}(y) \equiv \frac{\int_{\phi \in R_{M,\theta}(\Phi)} d\phi\ \delta(y, (M, \phi)(q))}{\int_{\phi \in R_{M,\theta}(\Phi)} d\phi}$$

In this case $O_{\theta,q}(y)$ is the probability of a randomly chosen defining set from $R_{M,\theta}(\Phi)$ making the guess y for the question q. If y as well as j is continuous-valued, we have integrals over outputs as well as defining sets, so

$$O_{\theta,q}(y) \equiv \frac{\int_{\phi \in R_{M,\theta}(\Phi)} d\phi\ \delta(y - (M, \phi)(q))}{\int_{\phi \in R_{M,\theta}(\Phi)} d\phi}$$

In such a situation $O_{\theta,q}(y)$ is the *differential* probability of a defining set from $R_{M,\theta}(\Phi)$ making the guess y for the question q; $\int O_{\theta,q}(y) dy = 1$. We always assume that for this case of continuous-valued y $O_{\theta,q}(y)$ is finite, i.e. for no $y \in Y$ is the Lebesgue measure of the support of the function $\delta(y - (M, \theta)(q))$ over the space of defining sets nonzero.

Equation (1.6) has the advantage that it, like the original formula for E_{guessing}, is given entirely in terms of expectation values. Nonetheless, in

practice it will often be useful to use a formula for E_{guessing} that does not share this property:

$$E_{\text{guessing}} = \left\langle \sum_{y \in Y} \{O_{\theta,q}(y)|y - f(q)|\} \right\rangle_{(\theta \subset f(x),\ q \in X_\theta).} \tag{1.7}$$

As usual, the sum over y is replaced by an integral for continuous-valued y.

A *generalizer* is a mapping taking a learning set to an input–output function (see [29]). Examples of generalizers are backpropagation [21], memory-based reasoning [25], and hyperplanar HERBIEs [28]. In convergent guessing we are interested in how the guessing accuracy of a particular generalizer varies as the size of the learning set grows. However, just specifying an architecture does not specify a full generalizer. Such a full specification necessitates also choosing a means of deciding amongst the elements of the (restricted) set of defining sets. Such a means of choosing a defining set tells us how to guess in response to a particular learning set and question. For the purpose of defining convergent guessing, we assume no knowledge of this scheme by which the generalizer chooses amongst the defining sets. Equivalently, we assume a worst case where the scheme behaves randomly (when averaged over all learning sets chosen from the parent curve). The guessing accuracy of such random behavior is given to us by E_{guessing}.[9]

A method M along with its associated architecture and restriction is said to exhibit *convergent guessing* for a parent function $f(x)$ if the following condition holds: For all errors $\Delta > 0$, there exists an integer m such that $E_{\text{guessing}} < \Delta$ for all learning set cardinalities n that exceed m. The requirement of convergent guessing for a parent function $f(x)$ is a restriction on the method used in the generalization. Intuitively speaking, M exhibits convergent guessing for a parent function $f(x)$ if you would expect that the function is compatible with (and reflected in) the method. Convergent guessing is roughly equivalent to Pearl and Cover's "ambiguity" [16, p. 259]). Convergent guessing is also related to cross-validation [6]. More precisely, convergent guessing of an architecture and parent function implies that for large enough learning sets the "average" generalizer implied by the architecture has small strong self-guessing error for learning sets chosen from the parent function [30]. In the analysis below, it might be possible to replace the definition of convergent guessing given here with a weaker one, but due to its intuitive appeal, we will stick with this definition.

2. Uniform simplicity measures

The previous section outlined the conventional ways of exploiting Occam's razor, which use simplicity measures only dependent on the defining set. As

[9]The word "random" can be interpreted differently than it is interpreted here. For example, the averaging over learning sets and questions could be modified so that it is weighted, perhaps according to the entropy of the distribution of the guesses induced by a given learning set and question. Such alternative definitions of "random" will not be considered in this paper.

was mentioned in the introduction, however, there are some major flaws in this conventional approach. This section starts by exploring these problems inherent in simplicity measures that are only a function of the defining set (e.g., all the examples in table 1). To avoid these problems it is necessary for S to be a function of the method, restriction, and question, as well as of the defining set. This section presents a formalism for dealing with such a modified simplicity measure. (It is through such modifications to the simplicity measures that this paper goes beyond the conventional ways of exploiting Occam's razor.) This section then introduces an optimality condition for such a modified simplicity measure. Any measure obeying this optimality condition is called a "uniform simplicity measure." This section ends by detailing the many advantages of uniform simplicity measures (e.g., it is when used with uniform simplicity measures that Occam's razor can be proven equivalent to convergent guessing).

2.1 The problems with the simplicity measures of section 1

There are two major problems with the simplicity measures introduced in section 1 and their dependence solely on the defining set. The first is that they are architecture-specific. In general, using Occam's razor with one architecture from the list in section 1 (along with its associated simplicity measure) rather than with another leads to different generalization. Yet there is no reason to assume that one particular architecture together with one particular simplicity measure should be used over all others. However, without such an ad hoc assumption, there is no answer to the question "If I'm given a learning set and a question and nothing else, what is the best guess for the output corresponding to the question, according to Occam's razor?"

The second major problem is that even if the architecture is given beforehand, there is no reason to believe that the associated simplicity measures given in section 1 are optimal. There is no reason to believe they will result in negative Occam error. All those simplicity measures are either ad hoc or, at best, are designed to address an issue (like coding efficiency) that is not directly related to the efficacy of Occam's razor (i.e., that is not directly related to the sign of the Occam error).

These problems have many different guises. One is the fact that infinitesimal changes in the method can change the guessing recommended by Occam's razor in an arbitrary manner. Using method M, if ϕ and ϕ' are both members of $R_{M,\theta}(\Phi)$ and $S(\phi) < S(\phi')$, then Occam's razor says that we should generalize with the function (M, ϕ). Now let M' be a new method identical with M except for one difference: if the defining set is ϕ (ϕ'), M' first replaces it with ϕ' (ϕ) and then performs the same mapping M does. In a certain sense this is the smallest possible modification to the method M. However since the simplicity measure has not been changed by this modification, using M' we would be led to generalize with $(M', \phi) = (M, \phi')$, getting the exact *opposite* results from when we used M. (This problem arises even when we average over all defining sets having given simplicity values, as in

calculating E_{Occam}.) If we had some way to pick between M and M' we would not have this problem. Barring such a way to choose, we cannot derive unique generalization of our learning set based only on the principle of Occam's razor.

In a similar manner, we can fix the architecture, but make an infinitesimal change in the simplicity measure. If the original simplicity measure is S, then define S' to be the same measure as S, except that $S'(\phi) = S(\phi')$ and $S'(\phi') = S(\phi)$. Assuming S meets (1.2′) and (1.3′), it is often the case that S' does as well. However, just as when we changed from M to M', the tiny change to S resulting in S' does not change the method and means that we would not pick ϕ' rather than ϕ and generalize with (M, ϕ').

Other problems with conventional simplicity measures arise from the fact that many of the simplicity measures in section 1 simply count the number of free variables in the defining set. Any finite set of real- or integer-valued variables can always be bijectively mapped to a single variable. A new method, entirely equivalent in its guessing to the original method, can then be used for such a compacted defining set. (This new method simply uncompacts its defining set and then runs the original method.) Equivalently, an original defining set of n variables can be bijectively mapped to a set of $m > n$ variables. Unless one wants to somehow set restrictions on the continuity of methods, or restrict to a finite set the number of allowed values of the variables in a defining set, there is nothing to disallow such "compacting" or "expanding" of defining sets. Yet if such compacting and expanding are allowed, freedom to change methods means that the common simplicity measure of "cardinality of the defining set" is meaningless. It results in different generalizing for "defining set compacted" and "defining set noncompacted" methods. Similar problems afflict "coding length" measures of simplicity (e.g., [19,20]). Different coding schemes give different guesses.

To get around these problems, S must not depend only on the defining set. S must also be a function of M; $S = K(M)$. This way S cannot be changed without changing M, and to the degree that K is one-to-one, M cannot be changed without changing S. To negate the "second major problem" completely, K should be designed so that using M with the associated S results in a negative Occam error. Unfortunately, no function K whatsoever is used to determine the simplicity measures listed in table 1, never mind a K designed to result in negative Occam error.

To prove anything at all about the efficacy of Occam's razor when used with a particular architecture, what is important is the *guessing distribution*. This is the distribution of which defining sets from $R_{M,\theta}(\Phi)$ make which guesses, for a given learning set θ and a given question $q \in X_\theta$. It is through the guessing distribution that simplicity measures can be related to methods. Unfortunately, none of the simplicity measures listed in section 1 take the guessing distribution into account. In fact, they are completely independent of the guessing distribution, despite the fact that it (and it alone) determines the efficacy of Occam's razor. This ignoring of the guessing distribution is

the underlying reason why these conventional simplicity measures have the problems expounded above.

2.2 Simplicity distributions and uniform simplicity measures

The only way to avoid these difficulties is to have the simplicity measure take into account the guessing distribution and therefore be an implicit function of the architecture, question, learning set, and restriction, as well as of the defining set. In other words, the (very reasonable) stipulation is made that to exploit Occam's razor the simplicity measure should depend on everything that affects the guessing, not just on the defining set. To investigate how the simplicity measure should depend on these other factors, we must introduce the concept of a "simplicity distribution" and how to optimize one.

We are provided with an architecture, a restriction, a learning set θ, a question $\in X_\theta$, and therefore a guessing distribution. First assume that both Y and D contain an uncountably infinite number of values. The *simplicity distribution* ρ is the differential probability distribution of defining sets in the space $Y \times D$. Its value at a given point (y, d) is the normalized differential probability density of a randomly chosen defining set (from the restricted set of defining sets) having simplicity measure d and making guess y in response to question q. Formally, with $\langle G \rangle_{(z)}$ meaning the probabilistic average of G subject to the constraint z, we require that

$$\langle F[(M,\phi)(q), S(\phi)] \rangle_{(\phi | \phi \in R_{M,\theta}(\Phi), y_1 \leq (M,\phi)(q) \leq y_2, s_1 \leq S(\phi) \leq s_2)} = \frac{\int_{y_1}^{y_2} dY \int_{s_1}^{s_2} dD \rho(y,d) F[y,d]}{\int_{y_1}^{y_2} dY \int_{s_1}^{s_2} dD \rho(y,d)}$$

for all functions $F[y, d]$ and for all $y_1, y_2 \in Y$ and $s_1, s_2 \in D$. (When y_1, y_2, s_1, and s_2 extend to the limits of Y and D respectively, then the integral in the denominator is assumed to equal 1.) As one would expect, $\int \rho(y,d) dD = O_{\theta,q}(y)$, the function introduced just below (1.5) that is completely determined by the guessing distribution. To see this, simply choose a function F independent of its second argument, $S(\phi)$, and verify that the formula for $\langle F \rangle$ given here agrees with the formula given in section 1 when $\int \rho(y,d) dD = O_{\theta,q}(y)$.

If the space Y only contains a finite number of elements, then we require that

$$\langle F[(M,\phi)(q), S(\phi)] \rangle_{(\phi | \phi \in R_{M,\theta}(\Phi), y_1 \leq (M,\phi)(q) \leq y_2, s_1 \leq S(\phi) \leq s_2)} = \frac{\sum_{y=y_1}^{y_2} \int_{s_1}^{s_2} dD \rho(y,d) F[y,d]}{\sum_{y=y_1}^{y_2} \int_{s_1}^{s_2} dD \rho(y,d)}$$

for all functions $F[y, d]$ and for all $y_1, y_2 \in Y$ and $s_1, s_2 \in D$. If D is also discrete, then the same formula holds except that now all integrals are replaced by sums. In general, the simplicity distribution $\rho(y, d)$ is implicitly determined by the architecture, restriction, learning set, and question.

We can now extend (1.1′) through (1.3′) to the case of infinite Φ. First, note that as (1.1′) is defined in section 1 the maximum value over all D of $\rho(y, d)$ could be infinitesimally close to zero for certain values of y. For example, assume that there are an infinite number of defining sets making any guess except that one guess, y_1, is only guessed by a finite number of defining sets. In such a case, $\rho(y_1, d)$ would have to equal 0 $\forall$ $d \in D$, despite the fact that (1.1′) is being obeyed. In general, such pathological cases where the Cantor cardinality of $\int \rho(y, d) dD$ varies with y are excluded from consideration for the same kinds of reasons that justified (1.1′). Formally, we modify (1.1′) to be the statement that for all $y \in Y$, $\exists$ $d \in D$ such that $\rho(y, d) \neq 0$. Similarly, we modify (1.2′) to be the statement that for all $y \in Y$, and d_1 and $d_2 \in D$, if $d_1 > d_2$, then $\rho(y, d_1) \geq \rho(y, d_2)$. (1.3′) is now the statement that $\int \rho(y, d)\ dY$ is a nondecreasing function of d over the space D ($\sum$'s replace $\int$'s for discrete Y). This new version of (1.3′) follows from the new version of (1.2′) regardless of the cardinality of Φ.

The (differential) probability of defining sets making a given guess y, $\int \rho(y, d) dD$, is given by the guessing distribution. Equations (1.1′), (1.2′), and (1.3′) put further constraints on the allowed form of $\rho(y, d)$. Due to (1.1′), for all $y \in Y$, $\exists$ $d \in D$ such that $\rho(y, d) \neq 0$. Write the smallest such d as $d_{\min}(y)$. Then if the upper bound on the set D is d^*, (1.2′) means that $\forall d \geq d_{\min}(y)$, $\rho(y, d) \neq 0$. In other words, for any value of Y the distribution of $\rho(y, d)$ over D is nonzero and nondecreasing everywhere above a certain D value, and zero everywhere below it (D being viewed as the vertical axis and Y being viewed as the horizontal axis).

An *Occam transformation* of a guessing distribution and/or simplicity measure is any transformation that either leaves the simplicity distribution unchanged, scales all the simplicity values by an identical positive nonzero constant (i.e., sends $\rho(y, d)$ to $\rho(|k|y, d)$), or translates all those values by an identical constant. An interchanging of both the guesses and the simplicity values of two defining sets is an Occam transformation, as is permutation of the simplicity values of the defining sets all making the same guess. (Both operations leave the simplicity distribution unchanged.) The *Occam invariance* of a guessing distribution together with a simplicity measure is the invariance, under Occam transformations, of how Occam's razor says to guess. For example, if the defining sets ϕ and ϕ' result in the same guess for the value of the output, then interchanging their simplicity values will not affect Occam's razor's recommendation for what guess to make. Similarly, scaling all simplicity values by the same positive constant will not affect the recommendations of Occam's razor. Note that there are transformations that leave the Occam error as defined in (1.4) unchanged but are not Occam transformations. This is because the Occam error as defined here ignores the simplicity distribution for simplicity values in between d^* and d_*, d_* being the minimal d value over all Y and D such that $\rho(y, d) \neq 0$.

To determine how best to relate the simplicity measure to the guessing distribution we do not need to set the simplicity distribution uniquely — setting it up to Occam transformations will do. Although they are rather severe

restrictions, equations (1.1′) through (1.3′) are not quite strong enough to set the simplicity measure uniquely, even up to Occam transformations. (If they were strong enough, then since all the examples in table 1 obey (1.1′) through (1.3′), they would all have equivalent simplicity distributions and therefore — when using Occam's razor — they would all result in the same guess, on average.) To fix the distribution uniquely up to Occam transformations, we return to the justification for (1.3′): the higher the simplicity value, the less restricted should be the guessing. In accordance with this principle, we require that for the highest simplicity value d^*, $\rho(y, d^*)$ is minimally restricted. In other words, we require that the entropy of the distribution $\rho(y, d^*)$ over all Y is maximal (subject to the constraint that $\int \rho(y, d^*) dY$ equals some unimportant constant). From (1.1′) and (1.2′), we know that $\rho(y, d^*)$ is nonzero for all y. This new constraint says that in fact $\rho(y, d^*)$ must be flat across Y. (The fact that such a flat distribution has higher entropy than any other distribution serves as further justification of (1.1′).) Accordingly, define $\rho^* \equiv \rho(y, d^*)$. Given the guessing distribution, to get uniqueness of the simplicity distribution $\rho(y, d)$ it suffices to add the constraint that $\rho(y, d)$ is as informative as possible, i.e. the entropy of $\rho(y, d)$ over all Y and D is as small as possible. Given the guessing distribution, this constraint sets $\rho(y, d)$ uniquely, up to Occam transformations:

(2.1) Given a guessing distribution, the optimal simplicity distribution $\rho(y, d)$ is one that everywhere has either the value 0 or the value ρ^*, ρ^* being an arbitrary but fixed positive and nonzero real-valued constant. For any $y \in Y$, $\rho(y, d) = 0$ for all d values less than a threshold $d_{\min}(y)$ and equals ρ^* for all d values above that threshold. $d_{\min}(y)$ is fixed by the guessing distribution together with ρ^*.

Proof. Hypothesize that for the distribution with minimal entropy there is a region in $Y \times D$ where $p \notin (\rho^*, 0)$. Everywhere outside of this region $\rho \in (\rho^*, 0)$. Assume that this region is simply connected in $\mathbf{R} \times \mathbf{R}$. (Arguments similar to those presented here work when this assumption does not hold.) $\rho(y, d^*) = \rho^*$ and $\rho(y, d)$ is everywhere nondecreasing. Therefore $\rho(y, d) \leq \rho^* \; \forall \; (y \in Y, d \in D)$, and in particular $0 < \rho < \rho^*$ over the hypothesized region. Moreover, for all $y \in$ this hypothesized region, the highest D value within the region bounds from below a region where $\rho = \rho^*$. Take any two very small distinct subregions of the hypothesized anomalous region, σ_1 and σ_2, neither of which contact the boundary of the hypothesized anomalous region. Because they are arbitrarily small, ρ is constant across the subregions σ_1 and σ_2 (assuming $\rho(y, d)$ is a continuous function over σ_1 and σ_2). The corresponding ρ values are written as $\rho(\sigma_1)$ and $\rho(\sigma_2)$. Without loss of generality, assume $\rho(\sigma_1) \geq \rho(\sigma_2)$. Adjust the sizes of the two subregions so that $(\rho^* - \rho(\sigma_1))V_1 = \rho(\sigma_2)V_2$, where V_1 is the area of σ_1 and V_2 is the area of σ_2. Now consider a new distribution $\rho'(y, d)$, identical to $\rho(y, d)$ except that $\rho'(\sigma_1) = \rho^*$ and $\rho'(\sigma_2) = 0$. Changing distributions from ρ to ρ' does not modify the value of the integrated probability, nor does it run afoul of (1.1′). The difference in entropy between ρ and ρ' is

$-\rho^* \ln[\rho^*]V_1 + \{\rho(\sigma_1)\ln[\rho(\sigma_1)]V_1 + \rho(\sigma_2)\ln[\rho(\sigma_2)]V_2\}$. This can be rewritten as $-V_1\{\rho^*\ln[\rho^*] - \{\rho(\sigma_1)\ln[\rho(\sigma_1)] + (\rho^* - \rho(\sigma_1))\ln[\rho(\sigma_2)]\}\} \leq -V_1\{\rho^*\ln[\rho^*] - \{\rho(\sigma_1)\ln[\rho(\sigma_1)] + (\rho^* - \rho(\sigma_1))\ln[\rho(\sigma_1)]\}\} = -V_1\{\rho^*\ln[\rho^*/\rho(\sigma_1)]\}$. Since $\rho^* > \rho(\sigma_1)$, this quantity is always negative, which means that the entropy difference is always negative. This new ρ' distribution is not legal, however, since it violates (1.2′). To take care of this, we modify ρ' by "shuffling it around" in D space. For all y in σ_1 we carry out the following procedure (carry out a similar procedure for σ_2): Let $[d_1, d_2]$ be the range in D for the subregion σ_1 along the axis $Y = y$. Let $d_{\max}(y)$ be the highest d value for which $\rho'(y, d)$ does not have the value ρ^*. ($d_{\max}(y)$ is the boundary of our hypothesized region within which $\rho \notin \{\rho^*, 0\}$.) Change ρ' by setting $\rho'(y, d)$ to the original value of $\rho'(y, d + (d_2 - d_1))$, $\forall d \in [d_1, d_{\max}(y) - (d_2 - d_1)]$. Then set $\rho'(y, d)$ for the interval from $d = d_{\max}(y) - (d_2 - d_1))$ to $d = d_{\max}(y)$ to the value ρ^*, the original value of $\rho'(\sigma_1)$. This shuffling procedure does not affect the entropy of the simplicity distribution, nor does it violate normalization of the distribution or requirement (1.1′). When carried out for all y in σ_1, however, it results in a distribution that obeys (1.2′) and therefore is fully legal. This legal distribution has smaller entropy than the original distribution $\rho(y, d)$, however, contrary to hypothesis. Therefore all ρ values must either be ρ^* or 0. ∎

$d_{\min}(y)$ is fixed by the guessing distribution and ρ^* via $\int \rho(y, d)dD = \rho^*(d^* - d_{\min}(y))$. Given the guessing distribution, the optimal simplicity distribution is defined completely by ρ^*. However, changing ρ^* does not change how Occam's razor says to guess; such a change is just an Occam transformation.[10] Therefore, up to such (irrelevant) transformations, the optimal simplicity distribution is set uniquely by the guessing distribution.

Given an architecture and a restriction, a simplicity measure that results in an optimal simplicity distribution for all learning sets and all questions is called a *uniform* simplicity measure. (The adjective "uniform" refers to the fact that $\rho(y, d)$ has the same value over all of its support.) The dependence of a uniform simplicity measure on the method occurs through the guessing distribution. Although still written as $S(\phi)$, such measures are implicitly understood to be functions of the guessing distribution in addition to the defining set.

For a uniform simplicity measure, the function $d_{\min}(y)$ is simply a (negative-valued) scaling and translating of the function $O_{\theta,q}(y) = \int \rho(y, d)dD$, introduced below equation (1.6). Up to such a scaling and translating, the simplicity distribution is set by the guessing distribution, for a uniform simplicity measure.

[10] Note, however, that for a fixed guessing distribution, lower ρ^* means higher values $\{d^* - d_{\min}(y)\}$. This effect means that the guessing distribution sets a lower bound on ρ^*; ρ^* smaller than this lower bound would mean that d_*, the lowest value of $d_{\min}(y)$, is less than 0, the value of the smallest element of D.

2.3 Advantages of uniform simplicity measures

In addition to their resulting in an optimal simplicity distribution, there are many other advantages to uniform simplicity measures. First, uniform simplicity measures reflect the Occam invariance of the associated guessing distribution; for a given guessing distribution, changing from one uniform simplicity measure to another is simply an Occam transformation and therefore does not affect how Occam's razor says to guess. The freedom in choosing a uniform simplicity measure corresponds exactly to the invariances in the effects of applying Occam's razor. As a result, any analysis of Occam's razor that holds for one uniform simplicity measure also holds for another uniform simplicity measure. No such property characterizes the simplicity measures of table 1 — those measures completely ignore the invariances inherent in Occam's razor.

Another advantage of uniform simplicity measures is that they are impervious to the "guises of the major problems" of conventional simplicity measures concerning tiny modifications to the method or the simplicity measure. For example, replacing M with M' as in section 2.1 does not change $O_{\theta,q}(y)$ and therefore does not change the uniform simplicity measure, up to an Occam transformation. So when a uniform simplicity measure is used it does not change the recommendation of Occam's razor as to how to guess.[11]

Uniform simplicity measures also have the advantage that they focus attention on the guessing behavior of an architecture rather than on the details of how the architecture works. Moreover, since Occam's razor is dependent solely on this guessing behavior, uniform simplicity measures allow an analysis of whole categories of architectures at once. This contrasts with the conventional approach, which relies on a case-by-case analysis of architectures, as was carried out in [16] for example.

Some might argue that the simplicity measures in section 1 are uniquely specified by the architecture, more or less, as the most "reasonable" measures for those architectures. However, for certain architectures it might not be clear how to construct a "most reasonable" simplicity measure. On the other hand, given an architecture, a learning set, and a restriction, the set of uniform simplicity measures is always uniquely specified, no matter how peculiar the architecture. Even if we can not calculate a uniform simplicity distribution analytically, we can always build one by using Monte Carlo techniques to estimate $O_{\theta,q}(y)$. Therefore the domain of applicability of uniform simplicity measures far exceeds that of constructing a "most reasonable" simplicity measure (the procedure implicitly employed in formulating the examples of table 1).

[11] Note however that we still haven't really addressed in full the "first major problem," presented earlier in this section, of how to deduce an *a priori* optimal architecture. An architecture must still be assumed beforehand in order to use Occam's razor. However, the "second major problem" has been partly solved by using uniform simplicity measures, in the sense that we now have a unique choice of simplicity measure given an architecture. (It is the task of further analysis to determine if and when this simplicity measure results in negative Occam error.)

If for some reason the simplicity measure is fixed beforehand, the results of this section can be used to help determine the architecture. For example, using Rissanen's MDL scheme the simplicity measure is determined by a parameterized conditional probability working over the provided learning set. Different parameterizations of the conditional probability will result in different simplicity distributions. The results of this section suggest a way to decide amongst a set of parameterizations; pick that parameterization that, for the learning set at hand, is closest to being uniform.

The final advantage of uniform simplicity measures is that they allow the analysis of the rest of this paper; it is with them that Occam's razor is equivalent to convergent guessing.

Note that uniform simplicity measures appear as very peculiar beasts when viewed from the perspective of the conventional simplicity measures listed in table 1. All those measures work by examining the details of how the associated architecture is implemented and then trying to minimize something roughly equivalent to the coding length of the implementation. This procedure seems to reflect how human beings are accustomed to using Occam's razor. However as was mentioned previously, the implementation issue is actually irrelevant; only the guessing distribution is important for Occam's razor. Accordingly, uniform simplicity measures ignore the implementation issue altogether. They view the issue of which particular defining set (out of those making a certain guess) gets which simplicity value as a red herring. Indeed, they explicitly treat different simplicity measures as equivalent if they are Occam transformations of one another.

Despite the theoretical rationale for using uniform simplicity measures, the fact that they contrast so strongly with the conventional kinds of simplicity measures might be a bit worrisome. Presumably humans use Occam's razor as they do because, by and large, it works well when used that way. This would seem to imply that uniform simplicity measures *cannot* work well. Fortunately, this conundrum can easily be resolved. First note that in the Occam error all we are interested in is the simplicity distribution for $d = d^*$ and for $d = d_*$. Now conventional simplicity measures obey (1.1′) through (1.3′). In particular, whenever a simplicity measure is something like "coding length," one would expect (1.2′) to be obeyed. Therefore, everything else being equal, one would expect conventional simplicity measures to be rough approximations to uniform simplicity measures. In particular, one would expect the y value corresponding to d_* for a conventional simplicity measure to be near to the mode of the guessing distribution (i.e., close to d_* for a uniform simplicity measure).[12] By this line of reasoning, when

[12]Unfortunately, it is extremely difficult to measure with any precision just how closely any of the simplicity distributions listed in table 1 exhibit this behavior. Just calculating *guessing* distributions for those cases, never mind simplicity distributions, is a calculational nightmare. For example, even the problem of how to calculate what defining sets reproduce a learning set (i.e., the problem of how to calculate the restricted set of defining sets) has yet to be solved analytically for the neural net architecture. (That is why people use techniques like backpropagation [21] and simulated annealing [11] to construct neural nets to reproduce provided learning sets.) In practice, it seems likely that Monte Carlo

determining a simplicity measure for a particular means of implementing a particular architecture, humans are really just applying heuristics for how to construct a useful simplicity measure. On average, these heuristics result in a simplicity measure that is approximately uniform and therefore whose minimum occurs near the mode of the guessing distribution. The results of the next section then tell us that if we believe that convergent guessing applies, then minimizing one of these heuristic simplicity measures will on average result in minimal guessing error.

3. Why convergent guessing implies Occam's razor

In this section it is proven that for uniform simplicity measures convergent guessing implies Occam's razor when there are a finite number of possible guesses (i.e., when the cardinality of the output space Y is finite). This restriction to finite Y obtains, for example, whenever the architecture is emulated on a finite digital computer. This section starts by deriving a new formula for E_{Occam} that applies when the simplicity measure is uniform. Then Occam's razor is related to convergent guessing for the case of two possible guesses. If Y is the set $\{\alpha + si | i \in [a, b]\}$ (α and s are real-valued constants, a and b both $\in \mathbb{Z}$, and $[a, b]$ indicates the closed interval of integers between a and b) then the analysis is more complicated; this case is dealt with in appendices A and B. If elements are removed from this set Y (i.e., if Y is changed to be a proper subset of $\{\alpha + s[a, b] | a$ and b both $\in \mathbb{Z}\}$), a variation of the analysis of appendices A and B can be used. In this way all cases are covered in which Y is finite and the differences between the elements of Y are rational number products of one another. However any finite set of real numbers Y can be viewed as a limit of a sequence of sets, all of which have the differences between their elements rationally related to one another. Therefore the analysis of this section in fact applies to any finite set Y.

After dealing with these cases where Y is finite, this section discusses the situation where there are an uncountably infinite number of possible guesses and what conditions are sufficient to have Occam's razor imply convergent guessing for such a case. The section then ends with a cursory discussion of alternative definitions of Occam's razor and how they relate to convergent guessing.

3.1 An intuitive rationale for the relationship between Occam's razor and convergent guessing

Consider the case of the feedforward neural net architecture. Given any learning set θ, question $q \in X_\theta$, and output y, there exists a set of weights (i.e., a defining set) such that the resultant net both reproduces θ and makes guess y in response to question q. Therefore running a procedure like back-propagation to derive a net that reproduces a learning set cannot, by itself,

techniques would have to be used to calculate the guessing and simplicity distributions for any nontrivial architecture.

have anything to do with generalization. To put it another way, there exists a global minimum in weight space (for the problem of reproducing a given learning set θ) corresponding to any output y in response to any question $q \in X_\theta$. So the procedure of simply finding such a minimum can result in any guess at all, independent of the correct guess $f(q)$.

There are two natural ways to modify things so that backpropagation can, theoretically at least, have some relation to expected generalization accuracy. The first is to constrain the net over which backpropagation is being run, for example by fixing the number of hidden neurons. Now, for a sufficiently severe constraint, it is no longer necessarily true that there exists a set of weights both reproducing any learning set q and making guess y in response to an arbitrary question $q \in X_\theta$. We have constrained the allowed guessing. In this scheme backpropagation will result in accurate guessing if and only if there is a close correspondence between the guess made in response to the question q, when the constraint is in force, and the correct guess $f(q)$.

The second way to modify things is to examine the average guess made by backpropagation, without any constraints imposed on the net. Here backpropagation will result in accurate guessing if and only if there is a close correspondence between the average guess made in response to the question q and the correct guess $f(q)$.

Assume that the "severe constraint" picks out nets approximately lying at the D-space minimum of the support of a uniform simplicity measure, in accordance with Occam's razor. Then these two "natural ways to modify things" are roughly equivalent. This is because for a uniform simplicity measure the mode of the Y-space projection of the simplicity distribution has the same Y value as the D-space minimum of the support of the distribution. However, under the assumption that the average and the mode of the Y-space projection of the simplicity distribution are close to one another, then the "second way to modify things" picks out the mode of the Y-space projection. On the other hand, the D-space minimum of the distribution is picked out by the "first way to modify things" (under our assumption concerning the effects of the "severe constraint"). Therefore when one way of modifying things results in accurate guessing, so does the other.

In addition, both of these ways of modifying things result in accurate guessing only if they enjoy a correspondence with the correct guessing. The second way to modify things enjoys such a correspondence if convergent guessing holds (and if the learning set is of sufficient size). The first way to modify things enjoys such a correspondence if Occam's razor holds. Therefore convergent guessing and Occam's razor are related.[13]

[13]Note that we can actually test if these two ways of modifying things are equivalent. Test if convergent guessing applies (for a reasonably chosen parent function, it should). Then test if allowing the size of the net to expand for a given size learning set results in worse generalization. If the size of the net is indeed a uniform simplicity measure, then the one effect will always occur when the other does. Moreover, if the guessing accuracy does indeed improve with increasing learning set size, then the improvement gained by using a small net should become more pronounced with increasing learning set size. This is because if convergent guessing applies, then the larger the learning set the more tightly

The rest of this section is nothing more than a formal statement of this relationship. To carry it out, we first need to perform some preliminary analysis.

3.2 Preliminaries

For a uniform simplicity measure, $S_{\max}(\theta) = d^*$ for all θ. Furthermore, for all allowed guesses $y \in Y$ the (differential) probability that a defining set $\phi \in R_{M,\theta}(\Phi)$ with simplicity measure d^* results in guess y in response to a question $q \in X_\theta$ is constant and independent of y (the probability is ρ^*, in fact). Therefore the second term inside the outer brackets in equation (1.4) is simply the average over all allowed guesses of (the absolute value of) the difference between that guess and $f(q)$. This term is independent of the method and learning set. As a result, we can write this term as $\langle |y - f(q)| \rangle$, where the expectation is over all possible guess values $y \in Y$.

For a uniform simplicity measure S, $\{(M,\phi)(q) | \phi \in R_{M,\theta}(\Phi),\ S(\phi) = S_{\min}(\theta)\} = \{\text{argmax}\ (O_{\theta,q}(y)\}$. Therefore we can rewrite the Occam error for a uniform simplicity measure as

$$E_{\text{Occam}} = \langle \ \langle |y - f(q)| \rangle_{(y \in \text{argmax}\ [O_{\theta,q}(.)])} \ \rangle_{(\theta \subset f(x),\ q \in X_\theta)} - \qquad (3.1)$$

$$\langle \qquad \langle |y - f(q)| \rangle_{(y \in Y)} \qquad \rangle_{(\theta \subset f(x),\ q \in X_\theta)}$$

The subscript i on the learning sets q has been dropped in going to equation (3.1) from equation (1.4) because there is no longer any ϕ_{ij} whose indices have to be matched to those of the learning set. For the rest of this paper the definition of E_{Occam} given by equation (3.1) will be taken as axiomatic, i.e., for the rest of this paper any reference to a simplicity measure implicitly means a uniform simplicity measure.

It should be clear from comparing equations (1.7) and (3.1) that, for sufficiently small expected guessing error, there is a limit on how big the Occam error can be. The size of the Occam error is determined by the first term in (3.1), the expected difference between the mode of $O_{\theta,q}(y)$ and $f(q)$. The expected guessing error is instead determined by the $O_{\theta,q}(y)$-weighted average of the difference between the elements of Y and $f(q)$. When this average difference gets small, the mode of $O_{\theta,q}(y)$ must approach $f(q)$. Indeed, as we will shortly see, for sufficiently small expected guessing error, the Occam error has to be less than zero and therefore Occam's razor must apply. This is the sense in which convergent guessing implies Occam's razor.

To facilitate the analysis define the *partial* expected guessing error as the expected guessing error for a fixed learning set and question (i.e., the expected guessing error where only the defining set is allowed to vary). The partial expected guessing error is the term inside the expectation value brackets in equation (1.7). Similarly, define the *partial* Occam error as the Occam error for a fixed learning set and question (again, this meaning that only

$O_{\theta,q}(y)$ gets wrapped about a point ($f(q)$) and the closer the mode of $O_{\theta,q}(y)$ is to its average.

the defining set is allowed to vary). Define the term $\langle |y - f(q)| \rangle_{(y \in Y)}$ from equation (3.1) as ε (the averaging is over all $y \in Y$ and *not* over different learning sets or different questions). Equation (3.1) tells us that the partial Occam error equals the average of the values $|y - f(q)|$ such that $O_{\theta,q}(y)$ is maximized, minus ε. Call that average μ; the partial Occam error $= \mu - \varepsilon$.

3.3 Two possible outputs

First we will deal with the case where there are only two allowed guesses, a and $b > a$. This is the case, for example, when we are dealing with a neural net outputting to a single neuron. Without loss of generality, set $a = 0$ (i.e., translate all outputs in both learning sets and guesses by $-a$). ε for this case of only two possible outputs $= (b - a)/2 = b/2$, regardless of $f(q)$.

Without loss of generality, assume that $f(q) = a = 0$. Call $O_{\theta,q}(0)$ m, and call $O_{\theta,q}(b)$ n; $m + n = 1$. The partial expected guessing error equals bn. Since ε is independent of the learning set and question, the partial Occam error for a particular learning set and question is completely determined by μ, and therefore by whether $m > n$ or vice versa. More precisely, the partial Occam error $= -b/2$ if $m > n$, 0 if $m = n$, and $b/2$ if $m < n$. If $m > n$, the partial expected guessing error $< b/2$. If $m < n$, the partial expected guessing error $> b/2$. Therefore for the partial expected guessing error $< b/2$, the partial Occam error $= -b/2$, and for the partial expected guessing error $> b/2$ the partial Occam error $= b/2$. For the partial expected guessing error $= b/2$, the partial Occam error $= 0$.

Now we must move from partial errors to the total errors of section 1, i.e. to the full errors, averaged over all allowed learning sets and questions. If $\langle$ partial expected guessing error $\rangle$, E_{guessing}, $= x$, what is an upper bound on $\langle$ partial expected Occam error $\rangle$, E_{Occam}? In other words, what limits on the (total) Occam error are set by the (total) expected guessing error? To answer this question, we write

$$E_{\text{guessing}} = \int_0^b dE \rho(E)\{E\}, \quad \text{where} \quad \int_0^b dE \rho(E) = 1 \tag{3.2}$$

$\rho(E)$ is the (differential) probability density of learning set/question pairs that have partial expected guessing error E. (It is not to be confused with the $\rho(y, d)$ of section 2.) Because of (1.1′), there is not a legitimate distribution of defining sets for which either $m = 0$ or $n = 0$. Therefore there is not a legitimate distribution of defining sets for which the partial expected guessing error equals 0, nor is there one for which it equals b. However for the case where an uncountable infinity of defining sets meet the restriction of the learning sets, the partial expected guessing error can get arbitrarily close to the limits 0 and b. Hence the limits on the integrals in (3.1) must be set to 0 and b. The total Occam error can be written as

$$E_{\text{Occam}} = \int_0^{b/2} dE \rho(E) \left\{ \frac{-b}{2} \right\} + \int_{b/2}^b dE \rho(E) \left\{ \frac{b}{2} \right\} \tag{3.3}$$

The limits on the integrands written as "$b/2$" are implicitly understood to actually be $b/2 - |\delta|$, where δ is infinitesimal. This takes care of the case where the partial expected guessing error $= b/2$ exactly and the corresponding partial Occam error $= 0$ by causing the expression for E_{Occam} in equation (3.3) to actually be an overestimate of the true E_{Occam} corresponding to a distribution $\rho(E)$.

We must now find the distribution $\rho(E)$ meeting the requirements of equation (3.2) for a particular value of E_{guessing} that maximizes E_{Occam} as given by equation (3.3). To do this first assume that there exists an E value, e, greater than $b/2$ and less than b, such that $\rho(e) \equiv D \neq 0$. Now if we reduce $\rho(e)$ to 0, increase $\rho(b/2)$ by $D + C$, and decrease $\rho(0)$ by C, where $C \equiv D(e - b/2)/(b/2))$, then we are still meeting the requirements of equation (3.2). Now E_{Occam} has been increased however. (In doing this rearranging, we might have forced $\rho(0)$ to be negative. But we are only interested in bounds, so the meaninglessness of such a value for $\rho(0)$ is of no concern.) As a result, to maximize E_{Occam} we should have $\rho(E) = 0$ for all $E > b/2$. This means that the limits on the second integral can be replaced with $b/2 - |\delta|$ and $b/2$, rather than $b/2 - |\delta|$ and b.

Now assume that there exists an E value, e, lying between 0 and $b/2 - |\delta|$, such that $\rho(e) \equiv D \neq 0$. As above, set $\rho(e)$ down to 0, but this time increase $\rho(0)$ by $D(b/2 - e)/(b/2)$, and $\rho(b/2)$ by $De/(b/2)$. Again, although this does not violate the requirements of equation (3.2) it increases the Occam error. Therefore $\rho(E) = 0$ for E lying between 0 and $b/2 - |\delta|$. We can now write down the $\rho(E)$ which maximizes the expression in equation (3.3), subject to the constraints of equation (3.2):

$$\rho(E) = \left(\frac{b/2 - E_{\text{guessing}}}{b/2}\right)\delta(E) + \left(\frac{E_{\text{guessing}}}{b/2}\right)\delta(E - b/2)$$

($\delta(.)$ is here the Dirac delta function, regardless of the cardinality of Y and/or Φ.) This allows us to write down the maximum possible value of E_{Occam}, given E_{guessing}:

The maximum of E_{Occam}, given E_{guessing}, $=$

$$\frac{b}{2}\frac{E_{\text{guessing}}}{(b/2)} - \frac{b}{2}\left(1 - \frac{E_{\text{guessing}}}{(b/2)}\right) = 2E_{\text{guessing}} - \frac{b}{2} \tag{3.4}$$

Now due to our assumption of convergent guessing, for any positive Δ there exists a cardinality m such that for all learning set cardinalities greater than m, $E_{\text{guessing}} < \Delta$. In particular, there exists such an m such that $E_{\text{guessing}} < b/4$. From equation (3.4), for such an E_{guessing}, the maximum of E_{Occam} is less than zero. This concludes the proof of the following theorem:

Theorem 1. *Assume we have a restriction and an architecture whose output space Y consists of two elements. Assume further that the architecture and restriction exhibit convergent guessing for a parent surface $f(x)$. Then there*

exists a positive integer m such that, for the parent surface $f(x)$ and for learning sets of cardinality greater than m, the architecture obeys Occam's razor for any uniform simplicity measure.

For cases where there is a limit L to the maximum size of the learning set (e.g., when the architecture is emulated on a finite digital computer so that both X and Y are finite), we must modify our definition of convergent guessing. For such cases, theorem 1 holds if convergent guessing is taken to mean that "for all errors $\Delta > \Sigma$, where Σ is a positive real number $< b/4$, there exists an integer $m < L$ such that $E_{\text{guessing}} < \Delta$ for all learning set cardinalities that exceed m."

3.4 More than two possible outputs

To deal with cases where the cardinality of Y is finite but greater than two, first we address the case where $Y \equiv \{\alpha + si | i \in [a, b]\}$ (α and s are real-valued constants, a and $b \in \mathbb{Z}$, $[a, b]$ is the closed interval of integers between a and b, and $b - a \geq 2$). The case of two possible outputs is a special case of this one, except that there $b - a = 1$. Without loss of generality, we can take $\alpha = 0$. (This simply amounts to translating $f(.)$ along with all the elements of Y by $-\alpha$, an operation that clearly will not affect partial expected guessing errors or partial Occam errors.)

As when there are two possible guesses, the analysis for more than two possible guesses starts by showing how a given partial expected guessing error sets an upper limit on the corresponding partial Occam error. Unlike the case when Y has cardinality 2, however, here we cannot immediately write down a simple relation between a partial Occam error and a partial expected guessing error. This is because $O_{\theta,q}(y)$ now has more than one degree of freedom, and therefore is not uniquely specified by a single number like the partial expected guessing error.

We will find how partial expected guessing errors set upper limits on the corresponding partial Occam errors by working in reverse, so to speak. We will find the minimum partial expected guessing error which can correspond to a given partial Occam error first, and then we will show that an increase in this minimum partial expected guessing error necessitates an increase in the corresponding partial Occam error, as well as vice versa. Appendix A is a rigorous proof that this "working in reverse" is a valid way to relate a partial expected guessing error and the maximum possible corresponding partial Occam error.

As before, define the term $\langle |y - f(q)| \rangle_y$ from equation (3.1) as ε:

$$\varepsilon \equiv s \left\{ \frac{[b - f(q)][b - f(q) + 1]}{2} + \frac{[f(q) - a][f(q) - a + 1]}{2} \right\} / \{b - a + 1\}$$

As before, equation (3.1) tells us that the partial Occam error equals the average of the values $|y - f(q)|$ such that $O_{\theta,q}(y)$ is maximized, minus ε. Again as before, call that average μ; the partial Occam error $= \mu - \varepsilon$. Call

the maximum of $O_{\theta,q}(y)$ m. Assume further that there are k values of y such that $O_{\theta,q}(y) = m$. By (1.1'), $mk \leq 1$. Delineate the set of y values such that $O_{\theta,q}(y) = m$ by $\{g_i\}$. There are k elements in $\{g_i\}$. Delineate the set of y values not contained in $\{g_i\}$ by $\{G_i\}$. The total number of possible guesses $= b - a + 1 \equiv N$. The cardinality of $\{G_i\} = N - k$.

Since we are "working in reverse," our task is to vary the question and the guesses of the individual defining sets so as to minimize the partial expected guessing error, subject to a given partial Occam error. Appendix B shows how this is done. In a manner similar to the analysis for the case of two possible outputs, appendix B then uses the relation between the partial errors to derive the maximum (total) Occam error that can correspond to a particular (total) expected guessing error. The result is the following:

The maximum of E_{Occam}, given E_{guessing},

$$\begin{aligned} &\leq s\frac{(b-a)}{2}\frac{E_{\text{guessing}}}{E'} \\ &\quad + \left[1 - \frac{E_{\text{guessing}}}{E'}\right][s]\left[\frac{1}{4} - \frac{(b-a)^2 + 2(b-a)}{4((b-a)+1)}\right] \\ &= s\frac{E_{\text{guessing}}}{E'}\left[\frac{3(b-a)^2 + 3(b-a) - 1}{4((b-a)+1)}\right] - s\left[\frac{(b-a)^2 + (b-a) - 1}{4((b-a)+1)}\right] \end{aligned} \tag{3.5}$$

E' is a positive constant defined in appendix B. It lies in the real-value interval $(0,\ s(b-a)]$ and is set by a, b, and s. Note that because $(b-a) \geq 2$, both terms in the square brackets in the last line of (3.5) are positive. Under the assumption of convergent guessing, for any positive Δ there exists a cardinality m such that for all learning set cardinalities greater than m $E_{\text{guessing}} < \Delta$. In particular, there exist an m such that $E_{\text{guessing}} < \{E'\}\{(b-a)^2 + (b-a) - 1\}/\{3(b-a)^2 + 3(b-a) - 1\}$. From equation (3.5), for such an E_{guessing}, the maximum of E_{Occam} is less than zero; E_{Occam} must be negative. In conjunction with theorem 1, this concludes the proof of the following theorem:

Theorem 2. *Assume we have a restriction and an architecture whose output space Y consists of $\{\alpha + si | i \in [a,b]\}$ (a and $b \in Z$, α and s real-valued constants, and $b > a$). Assume further that the architecture and restriction exhibit convergent guessing for a parent surface $f(x)$. Then there exists a positive integer m such that, for the parent surface $f(x)$ and for learning sets of cardinality greater than m, the architecture obeys Occam's razor for any uniform simplicity measure.*

If there is a limit L to the maximum size of the learning set (i.e., if L, the cardinality of X, $< \infty$), we must modify our definition of convergent guessing for theorem 2 to hold. For such cases, convergent guessing implies that Occam's razor applies if convergent guessing is taken to mean the following: "For all errors $\Delta > \Sigma$, where Σ is a positive real number $< \{E'\}\{(b-a)^2 +$

$(b-a)-1\}/\{3(b-a)^2+3(b-a)-1\}$, there exists an integer $m < L$ such that $E_{\text{guessing}} < \Delta$ for all learning set cardinalities that exceed m." Note that the expected guessing error is not defined for learning sets of size L, since for no such learning set does there exist a question that is not an input component of one of the elements of the learning set.

3.5 Discontiguous guess ranges

The results of section 3.4 can be extended to the case where the range of allowed guesses is not the full set of numbers as, $(a+1)s, \ldots bs$, but is instead a subset of this set. For example, if $s = 1$, one such "discontiguous set" would be the set of numbers $\{[a, c], [d, b]\}$, where $(d - c) > 1$. It is assumed that Y still consists of more than two elements (if this isn't the case, then theorem 1 applies).

Equations (1.7) and (3.1) are not changed if the set of allowed guesses is discontiguous. In addition, the analysis of appendices A and B goes through unchanged through the conclusion in appendix B that $k = 1$, through the "squeezing" procedure used to set the $O_{\theta,q}(G_i)$, through the conclusion that $m' = 0$, through equation (B.1), and through equation (B.2). Throughout all of this portion of the analysis we can ignore the fact that there are certain guesses where, due to the discontiguity of the set of allowed guesses, we are not allowed to place the free defining sets. This is because taking account of these discontiguities could only serve to increase the partial expected guessing error over what it would be otherwise.

Now, however, ε, the term $\langle |y - f(q)| \rangle_y$ from equation (3.1), is different. In particular, the lowest possible value of ε is now lower than it is for the case of no discontiguities (although it is still greater than 0). Taking this fact into account, equation (B.3) can be rewritten as $\partial E_{\text{Occam}} = s[E_{\text{guessing}}/s + 1/2]^2 - \varepsilon_{\min}$, $\varepsilon_{\min}$ being the lowest possible value of ε. The variable E', which is defined in terms of equation (B.3), is now different (although it is still between 0 and $s(b - a)$.) Other than this, the analysis still continues as before to equation (B.7); the only difference is that the two two terms in the square brackets in equation (B.7) have to be rewritten in terms of $\varepsilon_{\min}$:

The maximum of E_{Occam}, given E_{guessing},

$$= s\left\{\frac{E_{\text{guessing}}}{E'}\left[\frac{(b-a)}{2} - \frac{1}{4} + \frac{\varepsilon_{\min}}{s}\right] - \left[\frac{\varepsilon_{\min}}{s} - \frac{1}{4}\right]\right\}$$

Since $(b - a) \geq 2$, the term in the first set of square brackets is still necessarily positive, just as in equation (3.5). Since there are at least three possible guesses, the term in the second pair of square brackets is also necessarily positive. Therefore the conclusion stated in theorem 2 still holds: convergent guessing implies Occam's razor, for large enough learning sets, even if the allowed set of guesses is dicontiguous. (Note that this means, in turn, that even if restriction (1.1′) does not hold, convergent guessing still

implies Occam's razor, for large enough learning sets.) Given the relationship between rational and irrational numbers outlined in the preface to this section, this means that convergent guessing implies Occam's razor for any finite output space Y having two or more elements.

3.6 Continuous guesses

We can treat the case of continuous-valued outputs (i.e., uncountably infinite Y) by letting b in equation (3.5) go to infinity, while keeping a constant (0) and having s equal t/b for some constant t. Such a procedure gives us the maximum of E_{Occam}, given a particular E_{guessing}, for the case where $Y = [\alpha, \alpha + t]$ $(\alpha, t \in \mathbf{R})$. To deduce the resultant relationship we must first write E' in terms of b and s:

$$E' = s \left\{ \sqrt{\frac{b}{2} + \frac{b^2 + 2b}{4(b+1)}} - \frac{1}{2} \right\};$$

$$\text{as } b \to \infty, E' \to t \left\{ \sqrt{\frac{3}{4b}} - \frac{1}{2b} \right\}$$

Therefore the maximum of E_{Occam}, given E_{guessing}, becomes

$$E_{\text{guessing}} \sqrt{\frac{3b}{4}} - \frac{t}{4}$$

Since b is becoming infinite, any value of E_{guessing} above 0 will result in a positive Occam's error and Occam's razor will not apply. It is not hard to see why this is so. Imagine $\alpha = 0.0$ and $t = 1.0$, so the possible outputs lie in $[0.0, 1.0]$. Imagine further that $f(q) = 0.0$, and that $O_{\theta,q}(y)$ is infinitesimal everywhere except for at $y = 1.0$, where it is m, and for the range from 0.0 to Δ, where it is $m - \delta$. The partial Occam error is positive, but the partial expected guessing is $\Delta/2$, and since we can make Δ as small as we like, we can have the partial expected guessing error be as small as we like while still having a positive partial Occam error.[14]

As a result, to have convergent guessing imply Occam's razor for the case of continuous-valued outputs, we have to strengthen our assumptions concerning the nature of the guessing distribution when the cardinality of the learning set is large. This should not be too surprising; the requirement

[14] At first glance, it might appear that our formula for maximal E_{Occam} given E_{guessing} for continuous-valued Y must be wrong, since for $E_{\text{guessing}} \neq 0$ it says E_{Occam} can be infinite, when in fact its maximum is $t/2$. This overestimate for the maximal value of E_{Occam} is due to our forcing $\rho(E)$ to be two delta functions centered at 0 and E' (see appendix B). For the limit we are taking, E' approaches 0, and the delta function at 0 is weighted by an infinite negative number. A more careful analysis than the one in appendix B would conclude that the maximum of E_{Occam} is $t/2$. However, the conclusion that infinitesimal E_{guessing} can result in positive E_{Occam} must still be maintained, as is shown by the example where $O_{\theta,q}(y)$ is infinitesimal everywhere except at 1.0 and in a Δ-neighborhood of 0.0.

of convergent guessing is an extremely weak one, and in fact it is somewhat surprising that there exist any situations at all where it can set limits on the Occam error.

There are a number of ways to go about this "strengthening of assumptions." For example, to directly take care of the case where $O_{\theta,q}(y)$ is 1 everywhere except for at $y = 1.0$ (where it is m) and for the range from 0.0 to Δ (where it is $m - \delta$), we could require that for sufficiently large learning sets, $O_{\theta,q}(y)$ is allowed to have at most one local maximum within the allowed range of guesses.

Although this extra assumption will probably recover the relationship between convergent guessing and Occam's razor even when the guesses are continuous-valued, proving this relationship from this extra assumption would entail a calculation at least as complicated as that contained in appendix B. What is more, it is not clear that this assumption is really all that reasonable — it is a pretty strong claim to say that $O_{\theta,q}(g)$ must be roughly bell-shaped for *all* learning sets and for *all* questions.

Instead, we will make the simpler assumption that as the sizes of the learning sets grows larger, the magnitude of the difference between $\langle \mu \rangle$ and E_{guessing} grows smaller (the averages being over all learning sets θ and over all questions $\in X_\theta$). More precisely, define the *mode-average difference* $D(O, f(.), n)$ as follows:

$$D(O, f(x), n) \equiv \Big\langle \{\langle |y - f(q)| \rangle_{(y \in \text{argmax } [O_{\theta,q}(.)])} - \sum_{y \in Y} \{O_{\theta,q}(y)|y - f(q)|\} \} \Big\rangle_{(\theta \subset f(x),\ q \in X_\theta)} \quad (3.6)$$

All learning sets θ in the average are required to have n elements. The first term on the right-hand side is recognizable from equation (3.1) as $\langle \mu \rangle$. The second term is recognizable from equation (1.7) as the expected guessing error, E_{guessing}.

For a parent curve $f(x)$, a method M exhibits *convergence between the expected guessing error and the expected guessing mode* if, for all real numbers $\delta > 0$, there exists a positive integer n such that for all learning set cardinalities exceeding n, $|D(O, f(x), n)| < \delta$. Roughly speaking, convergence between the expected guessing error and the expected guessing mode means that we expect that, for large enough learning sets, the mode of the probability distribution over $|y - f(q)|$ values occurs near the average of that distribution (the distribution being set by $O_{\theta,q}(y)$, averaged over all learning sets θ and questions $q \in X_\theta$). Certainly for situations where $O_{\theta,q}(y)$ becomes approximately gaussian for large learning sets for any question, this assumption should be satisfied.

We can now write

$$E_{\text{Occam}} = D(O, f(x), n) + E_{\text{guessing}} - \langle \langle |y - f(q)| \rangle_{(y \in Y)} \rangle_{(\theta \subset f(x),\ q \in X_\theta)} \quad (3.7)$$

For convergent guessing and convergence between the expected guessing error and the expected guessing mode, there exists an n such that $D(O, f(x), n)$

$+E_{\text{guessing}}$ is an arbitrarily small (positive) number. What's more, for any learning set and question, $\langle |y - f(q)| \rangle_{(y \in Y)}$ has its lower bound set by the range of allowed guesses, and that lower bound is necessarily a positive number (i.e., regardless of the correct guess $f(q)$, $\langle |y - f(q)| \rangle_{(y \in Y)}$ is positive definite). Therefore we have the following theorem:

Theorem 3. *If an architecture and restriction exhibits convergent guessing for a parent function $f(x)$, and if it exhibits convergence between the expected guessing error and the expected guessing mode for that parent function $f(x)$, then there exists a positive integer n such that, for the parent function $f(x)$ and for learning sets of cardinality greater than n, M obeys Occam's razor for any uniform simplicity measure.*

Theorem 3 holds regardless of the cardinality of Y. It is trivial to prove the converse of theorem 3:

Theorem 4. *We are provided with an architecture and restriction that exhibits convergence between the expected guessing error and the expected guessing mode for a parent surface $f(x)$. Assume further that for any real number $\delta > 0$ there exists an integer n such that the Occam error of the architecture and restriction (calculated with a uniform simplicity measure) for the parent surface $f(x)$ is less than $\delta - \langle \langle |y - f(q)| \rangle_{(y \in Y)} \rangle_{(\theta \subset f(x), q \in X_\theta)}$ for all learning set cardinalities $> n$. Then the architecture and restriction exhibit convergent guessing for the parent surface $f(x)$.*

Note that this converse of "convergent guessing implies Occam's razor" depends critically on the assumption of convergence between the expected guessing error and the expected guessing mode; if such an assumption is not made (as it is not in the analysis leading up to equation (3.5)), this converse does not necessarily hold.

As in the discussion following theorem 2, if there is a limit L to the sizes of the allowed learning sets, then theorems 3 and 4 hold for a modified definition of convergent guessing and/or convergence between the expected guessing error and the expected guessing mode.

3.7 Alternative measures of Occam's razor

Equation (3.5) tells us that if convergent guessing applies, your learning set θ is large enough, and you have two randomly chosen defining sets ϕ_1 and ϕ_2 both $\in R_{M,\theta}(\Phi)$ where $S(\phi_2) < S(\phi_1)$ (but there is no defining set ϕ' such that $S(\phi_2) < S(\phi') < S(\phi_1)$), then you should use ϕ_2 rather than ϕ_1 to generalize. This means that guessing with the mode of $O_{\theta,q}(y)$ should give better generalizing than guessing with a randomly chosen $y \in Y$. Now, having chosen ϕ_2 over ϕ_1, should we choose a defining set ϕ_3 with lower simplicity than ϕ_2 over the defining set ϕ_2? Equation (3.5) does not answer this question, since we are no longer picking the simplicity value of the defining set with higher simplicity at random. (We are constraining $S(\phi_2)$ to be strictly less than d^* since $S(\phi_1) \leq d^*$ and $S(\phi_2) < S(\phi_1)$.) Therefore (3.5) does not

answer the question of whether we should implement a procedure of picking defining sets with lower simplicity all the way down until we get a defining set with simplicity d_*.

We can therefore ask the following novel question: If convergent guessing applies and you have two defining sets ϕ_1 and ϕ_2 both $\in R_{M,\theta}(\Phi)$ where $S(\phi_2)$ is d_*, the minimal S value over all defining sets $\in R_{M,\theta}(\Phi)$ (ϕ_1 being a randomly chosen defining set $\in R_{M,\theta}(\Phi)$), then should you use ϕ_2 rather than ϕ_1? This question is equivalent to asking if guessing with the mode of $O_{\theta,q}(y)$ should give better generalizing than guessing with a randomly chosen $\phi \in R_{M,\theta}(\Phi)$. (This contrasts with the question that is answered by (3.5), which assumes a randomly chosen $y \in Y$ rather than a randomly chosen $\phi \in R_{M,\theta}(\Phi)$.)

Intuitively, we would expect that if E_{guessing} is small enough, then $O_{\theta,q}(y)$ is very close to $f(q)$, whereas due to (1.1′) E_{guessing} must always take into account the error associated with all $y \in Y$. Therefore we would expect that for small enough E_{guessing} the magnitude of the guessing error of the mode of $O_{\theta,q}(y)$ would have to be less than the average guessing error magnitude over all of the guessing distribution, i.e. $\langle \mu \rangle$ would be smaller than E_{guessing}. To try to prove this, we could start arguing in a manner similar to that followed by the proof in appendix B. Following along with appendix B, we would figure out the limits on $\partial E_{\text{guessing}}$ set by $\partial\mu$. ("$\partial E_{\text{guessing}}$" is short hand for the partial expected guessing error and similarly for $\partial\mu$; see appendix B.) Unfortunately, $\partial E_{\text{guessing}}$ will only set limits on $\partial\mu$ if $\partial E_{\text{guessing}}$ is very small (e.g., $< s/2$). This is precisely the domain in which the approximations of appendix B result in the greatest overestimation of what $\partial\mu$ can correspond to a given $\partial E_{\text{guessing}}$. (This overestimation is so bad in fact that it leads to the conclusion that $\partial E_{\text{guessing}}$ must be negative to force $\partial\mu < \partial E_{\text{guessing}}$.)

Even if we make a more careful analysis, however, we can only come to a conclusion concerning partial errors. *For a given guessing distribution* (i.e., for a given question and learning set) there are (partial) expected guessing errors low enough to force an even lower guessing error magnitude for the mode of $O_{\theta,q}(y)$. However, the relationship between these two errors is not linear. In fact when we average over all questions and learning sets to get total errors it turns out that the expected guessing error can never be so low as to force the average guessing error of the mode of $O_{\theta,q}(y)$ to be even lower. E_{guessing} can force μ to be less than ε, but it cannot force it to be less than E_{guessing}.

As an example, imagine that Y consists of the three elements $\{0, 1, 2\}$ and that we have $n + 1$ {question, learning set} pairs. For n of those pairs, let $O_{\theta,q}(y)$ be infinitesimal for all y except $f(q)$. For the remaining pair, let $f(q) = 0$, have $O_{\theta,q}(y)$ be infinitesimal for $y = 1$, and have $O_{\theta,q}(2)$ infinitesimally greater than $O_{\theta,q}(0)$. For all of those first n pairs $\partial\mu = \partial E_{\text{guessing}} = 0$, $\varepsilon \geq 2/3$ and $\partial\mu - \varepsilon \leq -2/3$. For the $(n+1)$th pair, however, $\partial\mu = 2$ while $\partial E_{\text{guessing}} = \varepsilon = 1$ and $\partial\mu - \varepsilon = 1$. The total μ, averaged over all $n + 1$ pairs, $= 2/(n + 1)$. However, $E_{\text{guessing}} = 1/(n + 1)$ and can be arbitrarily small (for large enough n). For this example small enough E_{guessing} forces n

large and therefore forces the average of $\partial\mu - \varepsilon$ (i.e., E_{Occam}) to be negative. Nonetheless, no matter how small E_{guessing} is, the total μ averaged over all pairs exceeds it.

There are two ways around this impasse. One is to make additional assumptions concerning how much the guessing distribution can vary from one {question, learning set} pair to another. Another solution is to not average over all {question, learning set} pairs, but instead to work in terms of a single (provided) question and learning set. This second solution involves replacing the assumption of convergent guessing with the assumption that the particular learning set at hand is big enough so that for it and the provided question the partial expected guessing error is particularly low (i.e., $< s/2$). Under such an assumption we can conclude that finding the defining set $\in R_{M,\theta}(\Phi)$ with lowest simplicity will give better guessing than simply picking one of the defining sets $\in R_{M,\theta}(\Phi)$ at random.

There are a number of other questions similar to this "novel question" concerning the preferability of guessing with a defining set chosen from the mode of $O_{\theta,q}(y)$ rather than with a random defining set. To deal with these questions, one might prefer to define the Occam error as the difference in guessing accuracy averaged over *all* legitimate $S' < S''$ (i.e. over all $S' < S''$ that correspond to at least one guess), not just as the average difference in guessing accuracy when S' is the highest legitimate simplicity value $< S''$. In other words, where the a_i are the partial Occam errors corresponding to a given value of S'' (as in the derivation of (1.4)), one might want to investigate the alternate Occam error E_{OA}:

$$E_{\text{OA}} \equiv \left\langle \sum_{i=0}^{i=n-1} \sum_{j=i+1}^{j=n} (a_i - a_j) \right\rangle$$

as opposed to the definition from equation (1.4),

$$E_{\text{Occam}} \equiv \left\langle \sum_{i=0}^{i=n-1} \sum_{j=i+1}^{j=i+1} (a_i - a_j) \right\rangle = \left\langle \sum_{i=0}^{i=n-1} (a_i - a_{i+1}) \right\rangle$$

(The expectations are over all learning sets θ with a given cardinality and all questions $\in R_{M,\theta}(\Phi)$, as usual.)

When there are two possible guesses (i.e., $n = 1$), as in subsection 3.3, $E_{\text{OA}} = E_{\text{Occam}}$. When there are three possible guesses, $E_{\text{OA}} = 2E_{\text{Occam}}$. However, when there are four or more possible guesses, specifying E_{Occam} does not uniquely specify E_{OA}. For example, for $n = 3$, $E_{\text{OA}} = \langle \{a_0 - a_3 + a_1 - a_3 + a_2 - a_3 + a_0 - a_2 + a_1 - a_2 + a_0 - a_1\} \rangle = \langle \{3(a_0 - a_3) + (a_1 - a_2)\} \rangle = 3(E_{\text{Occam}}) + \langle (a_1 - a_2) \rangle$. Since the value of $(a_1 - a_2)$ is not set by E_{Occam}, the value of E_{OA} is not set by E_{Occam} either.

Unfortunately, carrying out the calculation of the bounds set on E_{OA} by a given expected guessing error is a much more complicated enterprise than performing the calculations already done in this section. (That is why this alternate definition of Occam's error has not been used in this paper.) Note, however, that a very simple and reasonable assumption allows us to

go directly from the conclusions concerning E_{Occam} reached previously to conclusions concerning E_{OA}: just assume that as the size of the learning set grows, terms like $\langle (a_1 - a_2) \rangle$ grow very small, at least in relation to terms like E_{Occam}. This assumption simply says that for large learning sets, the distribution $O_{\theta,q}(y)$ should become approximately bell-shaped, on average, so that $a_2 \approx a_1$, on average. In point of fact, due to convergent guessing we would expect that for large learning sets a_1 would actually become smaller than a_2, which would make Occam's razor even more pronounced than it would be if instead a_1 becomes equal to a_2 (in the limit of large learning sets). So this extra assumption that terms like $\langle a_1 - a_2 \rangle$ grow small (and that as a result convergent guessing implies Occam's razor when the Occam error is defined to be E_{OA}) is really very weak.

4. Various measures of "complexity" and Occam's razor

There has recently emerged a variant of the field of computational complexity [9] that is interested in broader measures of the "complexity" of an entity than its demand on time and computational resources [2,4,13,24]. At present there is still uncertainty as to what this broader "complexity" is supposed to measure, however. What some researchers seem to mean by "complexity" is a measure of the computational randomness of an entity (e.g., a sequence of integers). However, for a given size entity, the "randomness" of that entity can be viewed as a measure of how well one would expect to be able to generalize from that entity. The less "random" an entity, the more accurately one would expect to be able to generalize from it. For example, if the entity is a sequence of integers, then we can view it as a learning set of elements in $\mathbf{Z}^+ \times \mathbf{Z}^+$; the inputs of the elements of the learning set are the successive counting numbers, while the outputs of the elements of the learning set are the numbers in the sequence. For such a case, "generalizing from that entity" means extrapolating the sequence of integers. The less "random" the sequence, the more accurately one would expect to be able to extrapolate from it.

Implicitly acknowledging this relation between randomness and expected accuracy of generalization, Occam's razor is often assumed and the randomness then calculated with a simplicity measure. The lower the randomness (i.e., the lower the simplicity measure), the better the extrapolating should be, according to Occam's razor. For example, one common complexity measure of a sequence of bits is the size of the smallest Turing machine that can reproduce that sequence [4, 24]. This measure is said to give the "randomness" of the sequence. What it gives, in fact, is an estimate of how well the sequence can be extrapolated; according to Occam's razor, the smaller the simplicity measure of the extrapolator (i.e., the smaller the size of the Turing machine reproducing the sequence), the more accurate you would expect those extrapolations to be.[15]

[15]Strictly speaking, as usually formulated, algorithmic information complexity does not deal with TMs mapping $\mathbf{Z}^+$ to $\mathbf{Z}^+$ but rather with TMs generating a sequence of numbers in $\mathbf{Z}^+$. I am being a bit loose with the terminology here.

Even complexity measures that do not directly make use of the idea of "randomness" often exploit Occam's razor. In fact, this reliance on Occam's razor is often explicitly acknowledged. For example, Lloyd and Pagels [13] define the "absolute depth" of a system directly in terms of Occam's razor. There are other cases, however, where the reliance on Occam's razor, though pronounced, is not explicit. For example, Bennett's "logical depth" [2] is defined as the number of operations made when reproducing a sequence of numbers by the *minimal* Turing machine that can reproduce the sequence. But here, just as with algorithmic information complexity, it is not clear why one should be interested in things like *minimal* Turing machines unless it is because of what Occam's razor has to say about them. And in turn, it is not clear that there is *anything* Occam's razor can tell us, except how well you should expect to be able to generalize.

Whether or not it is only generalization accuracy we are interested in, it seems that expected generalization accuracy is closely related to the notion of "complexity." As a result, to investigate complexity properly we should have a theoretical structure telling us, from first principles and in an architecture-independent manner, when and how we should and should not expect to be able to generalize well.

Along with [29] and [30], this paper constitutes a first foray at creating such a theoretical structure. Insofar as the "complexity" of an entity is related to simplicity measures and therefore expected generalization accuracy, the results of this paper bear directly on the question of how to measure the "complexity" of that entity. For example, this paper provides a more sophisticated measure of the generalization accuracy expected from use of Occam's razor than, for example, the size of the minimal Turing machine: this more sophisticated accuracy is simply the (uniform simplicity measure-based) Occam error resulting from guessing part of the learning set when reproducing the rest of it. Such an error is an estimate of the Occam error when taught with the entire learning set. The lower this Occam error for the entire learning set, the more likely the guessing is to be convergent. In addition, the lower this error, the more evidence we have that picking a defining set with approximately minimal uniform simplicity measure (e.g., a Turing machine of minimal size) generalizes accurately for the parent function at hand, since the guessing accuracy of that defining set for the parent function is directly reflected in the Occam error.

If you are provided with a full-blown generalizer rather than just an architecture (i.e., if you are provided with a single-valued mapping sending learning sets to guessed input/output functions), there are other means in addition to Occam's razor for estimating the accuracy of a particular generalization of a particular learning set. One such alternative measure is the cross-validation error of the generalizer over the learning set, i.e. how accurately the generalizer guesses proper subsets of the learning set when taught with the rest of the learning set ([6]; and for a more sophisticated analysis of cross-validation, [30]). For example, if your generalizer is a surface-fitter (e.g., the hyperplanar HERBIE of [28], or a memory-based reasoner [25],

or a backpropagated neural net [21]), then the cross-validation error is one measure of how accurately you expect to extrapolate from the entire learning set (when the surface-fitter is taught with the entire learning set). Choosing a surface-fitter with minimal cross-validation error and then extrapolating with it, with the cross-validation error being a measure of your confidence in the extrapolation, is exactly analogous to choosing a Turing machine with minimal size and then extrapolating with it, with the size of that minimal Turing machine being a measure of your confidence in the extrapolation.

As Bennett's logical depth illustrates, however, "complexity" is not always directly associated with the generalizing accuracy. This suggests several alternatives to measuring the complexity of an entity as the estimated accuracy in generalizing from that entity. One such alternative is based on the notion of testing generalization accuracy by teaching a system with a subset of the learning set and seeing how well it guesses the rest of the learning set. Under this idea one would define complexity as the rate of fall-off of a quantity connected with generalization accuracy as the size of the portion of the learning set used for training increases. Such a complexity measure is referred to as a "differential" complexity measure. As an example, rather than the raw cross-validation error, the rate of fall-off of the cross-validation error as the proportion of the learning set used for teaching increases could be used as a measure of the "complexity" of the learning set. In this case, one would search for the surface-fitter with the greatest fall-off rate (i.e., derivative) in the cross-validation error.[16] Both completely random sequences and completely regular ones have a low fall-off rate, because no additional information pertinent to prediction is gained by increasing the size of the training set. Therefore, although this measure would view the brain as complex, a gas or a crystal would be viewed as noncomplex.[17] (This is a property that some researchers (e.g., [13]) deem desirable in a complexity measure.)

Such differential complexity measures do not have to use generalization and cross-validation. Nor do they have to involve first derivatives. For example, the complexity of a sequence of bits could be defined as the rate of change of {the change in its algorithmic information complexity as you add successive bits (perhaps averaged over all orderings of which bits get added first, perhaps divided by the number of bits) and evaluated when the entire sequence is in place}. Just as with the measure of differential cross-validation error, this differential algorithmic information complexity would view a regular sequence and a purely random one as having the same complexity. This is because for both such sequences the derivative of the algorithmic information complexity is essentially a straight line (when plotted against the number of bits of the sequence examined so far).

[16] Note that to use this definition of complexity, a standard measure is needed for "number of elements of the learning set," i.e., the independent variable in the derivative. For example, if the learning set is a sequence of integers, one such measure might be the number of integers in the sequence.

[17] Presumably, one should somehow incorporate the size of the entity into these differential complexity measures. This would prevent the measure from deeming a salamander's brain to have the same complexity as a human one, for example.

5. Conclusions

Occam's razor constitutes one of the enigmas of the scientific method. It is used extremely often in scientific research, but it is hard (if not impossible) to justify it from first principles. So the question arises, "Why does adherence to Occam's razor so often give good results when used in our universe?" Convergent guessing is also a property that is central to much of modern science and that has not been proven from first principles. This paper shows that the two principles are not independent; convergent guessing implies Occam's razor and vice versa. We therefore have reduced by one the number of principles in the foundations of science that, although of fundamental importance, are unproven. It seems likely that cross-validation can be worked into this framework as well, as might Rissanen's minimum description length principle and perhaps the technique of entropy maximization [10]. If indeed all these techniques (as well as all other fundamentally important but unproven techniques resident in the scientific method) can be incorporated into this framework, and are all seen to be equivalent, then the entire scientific method will be on a much firmer footing. Instead of a whole series of "fundamentally important but unproven techniques" disturbing the foundations of science, we would have only one.

In addition to shoring up the foundations of science, there are other advantages that accrue from investigating the equivalence of these techniques. Once we see how these techniques are related to one another, we can apply them in more sophisticated and effective manners. For example, the analysis of the relationship between Occam's razor and convergent guessing presented in this paper provides us a means of applying Occam's razor where before we might not have been able to.

Finally, there is another, more speculative, advantage that might accompany a complete and successful application of the kind of analysis carried out in this paper. Currently science is conducted by requiring that its theories possess as much objectivity and rigor as possible. The theories should be mathematically correct, agree with all of the data, be consistent with other well-established results, etc. In short, current science requires that its theories be completely rigorous. On the other hand, the means scientists currently use for choosing between competing (rigorous) theories is completely subjective, relying on aesthetics as much as anything else. No mathematically precise standards of rigor have been established for this meta-realm (the imprecise musings of some philosophers notwithstanding). Scientists strive for one set of standards *within* the theories and are content with another (much weaker) set of standards *amongst* the theories.

If the problem of inductive inference ever gets solved (whether according to the approach of this paper or according to some other approach), this unsatisfying and rather curious double-standard on the part of science will be removed. We will have a means for determining the optimal generalization of any learning set, i.e., a means for determining the optimal theory to explain any set of data. This goal of determining a means to find the unique

optimal generalization of any learning set is analogous to Hilbert's program for algorithmically codifying mathematics (a program which Gödel proved was impossible). Both programs, if successful, would take humans out of the loop.

Appendix A.

In this appendix it is proven that finding the lowest partial expected guessing error for a given partial Occam error is equivalent to finding the largest partial Occam error for a given partial expected guessing error.

For a given learning set, define the function $\Omega(E, O)$, from $\mathbf{R}^2$ to $\mathbf{R}$, to equal 1 if there exists a method and a parent function with partial expected guessing error equal to E and partial Occam error equal to 0. $\Omega(E, O) = 0$ otherwise. The range of possible guesses is as in section 3.4.

$\max(E|\Omega(E, O) = 1) \leq b - a$. (This maximum occurs when $f(q) = a$ for all q and M has infinite defining sets reproducing any learning set, all but an infinitesimal fraction of which guess b for all q and all learning sets.) $\min(E|\Omega(E, O) = 1) \geq 0$. (This minimum occurs in the same situation as the one resulting in the maximum of $(E|\Omega(E, O) = 1)$, except that the guess with (almost) all of defining sets is a, not b.) In a similar way, there exist finite bounds on $\max(O|\Omega(E, O) = 1)$ and $\min(O|\Omega(E, O) = 1)$.

We want to find the function from E to $\max(O|\Omega(E, O) = 1)$. Call this function $g(E)$. What we are going to be calculating is a function from O to a value which happens to be $\leq \min(E|\Omega(E, O) = 1)$ for those O values where $\min(E|\Omega(E, O) = 1)$ is defined and finite (there might be O values for which $\Omega(E, O) = 0$ for all E). Call this function $j(O)$, and call the function from O to $\min(E|\Omega(E, O) = 1)$ $h(O)$; $j(O) \leq h(O)$ for all O where $h(O)$ is defined.

We will see when $j(O)$ is calculated that $j(O)$ is defined everywhere, is everywhere continuous, and has infinite domain (unlike $g(E)$ and $h(O)$). Furthermore, $j'(O)$, the derivative of j with respect to O, is everywhere greater than 0 (and nowhere infinitesimal). As a result, $j(O)$ is invertible.

The proposition to be proved is that $g(E) \leq j^{-1}(E)$ for all E for which $g(E)$ is defined. The proof of this proposition involves two parts. First, it is shown that $g(E) \leq \max(O|h(O) \leq E)$ for all E for which $g(E)$ is defined. Then it is shown that, for all such E, $\max(O|h(O) \leq E) \leq j^{-1}(E)$. The transitivity of the $\leq$ relation then completes the proof.

Given any value of O and any fixed value of E, $\Omega(E, O)$ can equal 1 only if the minimum over E' of $(E'|\Omega(E', O) = 1)$ is defined and is $\leq E$. Therefore, for fixed E, the set of O's such that the minimum over E' of $(E'|\Omega(E', O) = 1) \leq E \supseteq$ the set of O's such that $\Omega(E, O) = 1$ (whether or not either of these sets are empty). Therefore $\max(O|\Omega(E, O) = 1) \leq \max(O|\{\min(E'|\Omega(E', O) = 1) \leq E\})$ (assuming both sides of the inequality are defined, i.e. assuming there is an E value e, such that $\Omega(e, O) = 1$ for some O value). In other words, $g(E) \leq \max(O|h(O) \leq E)$ for all E for which both sides of the inequality are defined. It is possible, however, to have the right-hand side defined and not the left (but not vice versa). To

take care of this case, it suffices to note that we are only interested in those E for which $g(E)$ is defined. So for our purposes we can simply state that $g(E) \leq \max(O|h(O) \leq E)$, as required.

Now examine any point (E', O') lying on $\max(O|h(O) \leq E')$. Because $j(O) \leq h(O)$ for all O where $h(O)$ is defined, $j(O') \leq E'$. By definition, there is no O value exceeding O' such that $h(O) = E'$. However, the derivative of j is positive definite and nowhere infinitesimal, so there *is* a value of O exceeding O' such that $j(O) = E'$. Since j is invertible, we see that $j^{-1}(E') \geq O'$. Therefore $\max(O|h(O) \leq E) \leq j^{-1}(E)$, as required. ∎

Appendix B.

This appendix shows how to minimize the partial expected guessing error, given the partial Occam error, for the case when the output space Y is as is section 3.4. The variables a, b, s, ε, μ, k, m, $\{g_i\}$, $\{G_i\}$, and N are all defined in section 3.4. The set of $\{g_i\}$ cannot be empty, although the set of $\{G_i\}$ might be.

To this end of minimizing the partial expected guessing error, first assume that $f(q)$ and m are fixed to the values that will minimize the partial expected guessing error. ε is fixed by Y, μ is fixed by ε and the partial Occam error, while k, the set $\{g_i\}$, and the $O_{\theta,q}(y)$ have yet to be set.

Now we will show that we can always take $k = 1$, so that $\{g_i\}$ contains a single element, and as a result $\{G_i\}$ is nonempty. Let $k \neq 1$ and assume that the average of the magnitudes of the distances of the $\{g_i\}$ from $f(q)$, added to $f(q)$, lies on a point in Y, labeled as g_{ave}; $\mu = g_{\text{ave}} - f(q)$ and $g_{\text{ave}} \in Y$. Assume also that $O_{\theta,q}(g_{\text{ave}}) = 0$. (More accurately, due to (1.1′) assume that $O_{\theta,q}(g_{\text{ave}})$ is infinitesimally close to 0. Whenever this appendix refers to setting an $O_{\theta,q}(y)$ value to "0," what is really meant is that it is set to an infinitesimal positive number.) Now modify $O_{\theta,q}(y)$ by setting $O_{\theta,q}(g_i)$ to 0 for all the $\{g_i\}$. Simultaneously, set $O_{\theta,q}(g_{\text{ave}})$ to mk. This procedure leaves μ and therefore the partial Occam error unchanged. It also leaves the partial expected guessing error unchanged and preserves the normalization of $O_{\theta,q}(y)$. Therefore for this case we can assume that the set of $\{g_i\}$ consists of the single element g_{ave}, i.e., we can replace the old value of k with the value 1. (Note that despite our assumption of optimal m this procedure gives us a new (larger) value for m, namely the old value multiplied by the old k. Either there exists more than one m minimizing E_{guessing}, or our original assumption that $k \neq 1$ resulted in a contradiction.)

Now examine the case where $O_{\theta,q}(g_{\text{ave}}) \neq 0$ prior to the process of squeezing the $O_{\theta,q}(g_i)$ into $O_{\theta,q}(g_{\text{ave}})$, although it is still true that $g_{\text{ave}} \in Y$. This can in general happen in one of two ways; $g_{\text{ave}} \in \{g_i\}$, or $g_{\text{ave}} \notin \{g_i\}$, but $O_{\theta,q}(g_{\text{ave}})$ is not infinitesimal. In the first scenario setting all the $O_{\theta,q}(g_i)$ to 0 and $O_{\theta,q}(g_{\text{ave}})$ to mk is still a valid process, i.e., it does not change the partial Occam error, the partial expected guessing error, or the normalization of $O_{\theta,q}(y)$. In the second scenario, however, the partial expected guessing error has been changed. This can be fixed though — simply set the new $O_{\theta,q}(g_{\text{ave}})$

to the old (i.e., pre-"squeezing") noninfinitesimal value + mk. The rest of the argument follows, and we can take $k = 1$.

The remaining possible case is where $g_{\text{ave}} \notin Y$. For this case, for the calculations of all quantities except ε, modify Y to equal the old $Y \cup \{g_{\text{ave}}\}$. ($\varepsilon$ is calculated according to the original set Y.) With this new Y we can always fix $O_{\theta,q}(g_{\text{ave}})$ to 0, and thereby get the exact same values (as when Y did not include g_{ave}) for all quantities, except ε, which go into the calculation of the partial Occam error and the partial expected guessing error. Since ε is calculated with the old Y, however, this fixing of $O_{\theta,q}(g_{\text{ave}})$ to 0 shows that, with this new system, there exists a set of $O_{\theta,q}(y)$ resulting in the old value for the minimum of the partial expected guessing error given the partial expected Occam error. As a result, the minimum calculated this new way is bounded above by the true minimum. Since we are only interested in such a lower bound, we can carry on with this new Y. (The fact that one element of this new $Y \notin \{si | i \in [a, b]\}$ will not affect the analysis of this appendix, given that ε is calculated with the original Y.) With this new Y, $O_{\theta,q}(g_{\text{ave}}) = 0$ before the squeezing, and so we can carry on the analysis exactly as in the preceding discussion. When we squeeze the $O_{\theta,q}(g_i)$ to $O_{\theta,q}(g_{\text{ave}})$, we do not change the normalization of $O_{\theta,q}(y)$, the value of the partial expected guessing error, or the value of the partial Occam error (so long as ε is calculated with the original Y). Therefore we conclude that k can be set to 1. This takes care of all possible cases, so we conclude that k can always be taken to equal 1 (although we might have to augment Y to do so).

m is still the optimal value for Y; if Y has been augmented, we assume that m has been modified to be the optimal value for the new Y (though k is still fixed to 1). This modification can not result in a bound on the partial expected guessing error greater than the real bound for the original Y. Even if it was necessary to augment Y, $f(q)$ is still assumed to $\in$ the original set Y. Like m, $f(q)$ is assumed to be modified if need be (i.e., if Y is augmented) so as to minimize the partial expected guessing error.

Since $k = 1$, $O_{\theta,q}(g_{\text{ave}})$is fixed (to m). Now we will vary the values of the remaining $O_{\theta,q}(y)$ (i.e., the values for $y \in \{G_i\}$) so as to minimize the partial expected guessing error. Once this is done, we will calculate the optimal values of m and $f(q)$, the remaining undetermined parameters.

To minimize over the $O_{\theta,q}(G_i)$ we carry out the following procedure: Start with $O_{\theta,q}(G_i)$ infinitesimal for all y except for g_{ave}. Add $m - \delta$ to $O_{\theta,q}(f(q))$, then to $O_{\theta,q}(f(q) + s)$ and to $O_{\theta,q}(f(q) - s)$, then to $O_{\theta,q}(f(q) + 2s)$ and to $O_{\theta,q}(f(q) - 2s)$, etc. δ is an infinitesimal positive constant. The procedure is continued until we run out of free defining sets, i.e., we continue until m', defined to equal the difference between 1 and the sum of m with all the additions to the $O_{\theta,q}(.)$ made so far by this procedure, is less than $2m$. At this point we set each of the next two outputs in line (one for each side of $f(q)$) to the value $m'/2$. The procedure, which will be referred to as "the $O_{\theta,q}$ procedure," is now over. This $O_{\theta,q}$ procedure put as much of the $O_{\theta,q}(y)$ distribution as close to $f(q)$ as possible while preserving g_{ave} (and therefore the partial Occam error) as well as the value of k. As a result, this $O_{\theta,q}$

procedure has resulted in minimal partial expected guessing error, given the constraints.

Strictly speaking, in carrying out the $O_{\theta,q}$ procedure we should worry about the effects of the boundary of the allowed range of the guesses. If $f(q)$ is too close to such a boundary, then we cannot keep adding $m - \delta$ to points arrayed symmetrically about $f(q)$, since that would mean some defining sets result in guesses outside of Y. For calculational purposes however, we always assume that in carrying out the $O_{\theta,q}$ procedure we can range over a subset of $\{G_i\}$ which is symmetric about $f(q)$. We can make this assumption because we are only interested in finding a lower bound for the partial expected guessing error, and boundary effects can only serve to increase the partial expected guessing error. As when setting $k = 1$ for $g_{\text{ave}} \notin Y$, it might be necessary to augment Y (for calculations not involving ε) to meet this assumption. This augmented Y consists of all $\{a + si | i \in [(a-b), (b-a)]\}$. If this augmentation is necessary, then we assume that m is chosen to minimize the partial expected guessing error for this augmented Y, as usual. Also as usual, this augmentation will not change the applicability of our lower bound calculations to the architecture with the original Y.

Another caveat is needed to take care of the fact that the $O_{\theta,q}$ procedure as defined might add $m - \delta$ to $O_{\theta,q}(g_{\text{ave}})$. Since m is already optimal (by assumption), we must require that the $O_{\theta,q}$ procedure skip over g_{ave} if it comes to it. For calculational simplicity however, if the $O_{\theta,q}$ procedure comes to g_{ave} and $g_{\text{ave}} \in$ the original Y (i.e., if $g_{\text{ave}} = a + si$ for some $i \in \mathbb{Z}^+$), then we will calculate the partial expected guessing error as though g_{ave} were not skipped and $m - \delta$ were added to $O_{\theta,q}(g_{\text{ave}})$. Again, we can do this since it results in an underestimate of the true lower bound on the partial expected guessing error. (Note, however, that no augmenting of Y was necessary to carry out this particular approximation.)

As a result of these considerations, we can write down the minimal partial expected guessing error: it is simply $\mu m + s\{(m-\delta)(0) + 2(m-\delta)(1) + 2(m-\delta)(2) + \ldots + 2(m-\delta)(c) + m'(c+1)\}$, for some constant c. We must now find the m, c, and m' which minimize this sum, subject to the constraint that probability is conserved. To 0th order in δ, this constraint means that $m[1 + 1 + 2c] = 2m[c + 1] = 1 - m'$. (Using the constraints that $c \in \mathbb{Z}^+$ and $0 < m' < 2m$, if we are given m we can calculate m' and c, as well as vice versa.) To 0th order in δ, the partial expected guessing error equals $m[\mu + s\{c^2 + c + m'(c+1)/m\}]$.

As a first step to solving this minimization problem, we will establish that $m' = 0$. Imagine a distribution, $h(y)$, of real-valued numbers over Y. Given the distribution $h(y)$, $O_{\theta,q}(y)$ is calculated by dividing any particular $h(y)$ by the sum of all $h(y)$ values; $O_{\theta,q}(y)$ is $h(y)$ normalized. We start with $h(y)$ infinitesimal for all y except for g_{ave}, for which it equals 1. It is assumed that $\mu = g_{\text{ave}} - f(q)$, so that the $O_{\theta,q}(y)$ calculated from this starting $h(y)$ has the correct partial Occam error. We are now going to carry out the $O_{\theta,q}$ procedure in a gradual manner. First we gradually increase $h(y)$ to the value $1 - \delta$ at the point $f(q)$ (δ is an infinitesimal positive constant.) Then we do

the same thing with the point $f(q)+s$, and so on. At any point in this process we can calculate $O_{\theta,q}(y)$ and therefore both partial errors. Once the partial expected guessing error starts to increase in this process, it will never again decrease. (This is because the average error will start to increase only when the point y whose $O_{\theta,q}$ value is currently being increased lies further from $f(q)$ than the current value of the average error. At no point in carrying out the $O_{\theta,q}$ procedure beyond such a point can the average error ever decrease.) Since the partial Occam error does not change in the process, the value of the partial expected guessing error at this point where it first starts to increase is the smallest partial expected guessing error for the given partial Occam error. Note, however, that if at any point in the process an addition to the value of a certain $O_{\theta,q}(y)$ causes an increase in the partial expected guessing error, then *any* addition to the original infinitesimal value of that $O_{\theta,q}(y)$ occuring in this process must cause such an increase. Therefore, if $O_{\theta,q}(y)$ is not infinitesimal when the procedure ends, it equals $1-\delta$; $m'=0$.

Since $m'=0$, we can write the minimal partial expected guessing error as $[\mu+s\{c^2+c\}]/[2c+2]$. ($m=1/(2c+2)$.) Differentiating with respect to c and setting the result to 0, we get the following:

$$c=\sqrt{\frac{\mu}{s}}-1 \tag{B.1}$$

(The zero corresponding to the negative square root can be discarded because c must be nonnegative.) If $\mu/s>1$, this value of c is nonnegative and we know that it has to correspond to a minimum since $\{\mu+sc(c+1)\}/\{2c+2\}=\mu/2$ for $c=0$, it equals ∞ for $c=\infty$, and the derivative of $\{\mu+sc(c+1)\}/\{2c+2\}$ with respect to c is negative for $c=0$ and $\mu/s>1$. Assume for the moment that μ/s is indeed >1. Since c is not a real number but an integer, we have two numbers to check, namely the two values of $\{\mu+sc(c+1)\}/\{2c+2\}$ for each of the two integers bracketing the c given by equation (B.1). Call these two values E_1 and E_2. The lower bound on the partial guessing error is then given by the minimum of E_1, and E_2. But as usual we are only interested in lower bounds, and since the value of $\{\mu+sc(c+1)\}/\{2c+2\}$ is minimized by the c given by equation (B.1), even if this is not a valid c (i.e., is not an integer), the value of $\{\mu+sc(c+1)\}/\{2c+2\}$ given by plugging in this value of c is a valid lower bound. Therefore we can dispense with E_1 and E_2 and simply plug equation (B.1) into the expression $\{\mu+sc(c+1)\}/\{2c+2\}$. Using the fact that the partial Occam error $=\mu-\varepsilon$, we can write the lower bound on the partial expected guessing error for $\mu/s>1$:

$$\partial E_{\text{guessing}}=s\frac{2\sqrt{(\partial E_{\text{Occam}}+\varepsilon)/s}-1}{2} \tag{B.2}$$

("∂" indicates a partial error here. Technically speaking though, a partial error is not a differential.) If $\mu/s\leq 1$, so that the c given by equation (B.1) is negative, the derivative of $\{\mu+sc(c+1)\}/\{2c+2\}$ with respect to c is

positive for all $c \geq 0$, so that setting $\partial E_{\text{guessing}}$ as in equation (B.2) is an underestimation of the true error. Therefore we can always use equation (B.2), without any concern for the value of μ/s.

Since we do not know $f(q)$, we do not know ε, in point of fact. Therefore we must find a lower bound on ε, given a and b. (This is equivalent to minimizing with respect to the only remaining free variable, $f(q)$.) The lower bound on ε occurs when $f(q)$ lies midway between sa and sb:

$$\varepsilon = s\{b - (b+a)/2\}\{b - (b+a)/2 + 1\}/\{b - a + 1\}$$

for $b - a$ even, and

$$\begin{aligned}\varepsilon \;=\; & s\{b - (b+a-1)/2\}\{b - (b+a-1)/2 + 1\}/2\{b - a + 1\} + \\ & s\{b - (b+a+1)/2\}\{b - (b+a+1)/2 + 1\}/2\{b - a + 1\}\end{aligned}$$

for $b - a$ odd.

Since the first expression is always lower than the second, we can use it for getting the lower bound on $\partial E_{\text{guessing}}$. The result is

$$\partial E_{\text{guessing}} = s\frac{2\sqrt{\{\partial E_{\text{Occam}} + s\frac{a^2+b^2-2(ab+a-b)}{4(b-a+1)}\}/\{s\}} - 1}{2}$$

Looking at this equation, it is clear that as $\partial E_{\text{guessing}}$ increases, $\partial E_{\text{Occam}}$ must also increase. Therefore this equation is not only a lower bound on $\partial E_{\text{guessing}}$ given $\partial E_{\text{Occam}}$; it is also an upper bound on $\partial E_{\text{Occam}}$ given $\partial E_{\text{guessing}}$ (see appendix A). In other words, the maximum of the partial Occam error is given by

$$\partial E_{Occam} = s\left\{\left[\frac{\partial E_{\text{guessing}}}{s} + 1/2\right]^2 - \frac{a^2 + b^2 - 2(ab + a - b)}{4(b - a + 1)}\right\} \qquad \text{(B.3)}$$

The largest possible value of $\partial E_{\text{Occam}} = s(b-a)/2$. It occurs when $f(q) = a$ and the mode of $O_{\theta,q}(y)$ is at $y = b$. Note, however, that depending on the values of a, b, and s, this worst possible $\partial E_{\text{Occam}}$ might actually be *less* than the $\partial E_{\text{Occam}}$ given by formula (B.3) for certain (completely legal) values of $\partial E_{\text{guessing}}$. This state of affairs occurs due to the (numerous) instances in the derivation of equation (B.3) in which the minimum possible partial expected guessing error is approximated by an expression that is always lower than that minimum possible partial expected guessing error.

Now we must move from partial errors to full errors, averaged over all allowed learning sets and questions. If $\langle$partial expected guessing error$\rangle$, E_{guessing}, $= x$, what is an upper bound on $\langle$partial expected Occam error$\rangle$, E_{Occam}? We write

$$E_{\text{guessing}} = \int_0^{(b-a)s} dE\rho(E)\{E\}, \text{ where } \int_0^{(b-a)s} dE\rho(E) = 1 \qquad \text{(B.4)}$$

As in the analysis for two possible outputs, here $\rho(E)$ is the (differential) density of learning set/question pairs which have partial expected guessing error E. To simplify the mathematical bookkeeping, we now translate a and b by $-a$ and redefine b as $b - a$ (i.e., we set the lower and upper limits of the range of possible guesses as 0 and bs, respectively) and assume that all learning sets and guesses are translated accordingly. Clearly this translating will not affect the relationship between the expected guessing error and the total Occam error. Now the lower limit on the integrals in (B.4) are still 0, but the upper limits are bs.

With a equalling 0, $\{a^2 + b^2 - 2(ab + a - b)\}/\{4(b - a + 1)\} = \{b^2 + 2b\}/\{4(b+1)\}$. Therefore, given $\rho(E)$, the average value of the partial Occam errors (i.e., the full Occam error, as defined in equation (3.1)), can be written as

$$E_{\text{Occam}} = \int_0^{bs} dE\rho(E)s\left\{\left(\frac{E}{s} + 1/2\right)^2 - \frac{b^2 + 2b}{4(b+1)}\right\} \tag{B.5}$$

We want to maximize the expression in equation (B.5) subject to the constraints of equation (B.4).

For the purposes of our maximizing E_{Occam} subject to the constraints of equation (B.4), we do not want to allow the quantity inside the curly brackets in the integrand in (B.5) to exceed the maximal possible $\partial E_{\text{Occam}}$ value of $sb/2$. (This can be viewed as an attempt to mitigate the cumulative effects of all our underestimates of the minimal $\partial E_{\text{guessing}}$ that can correspond to a given $\partial E_{\text{Occam}}$.) Therefore we cap the quantity in the curly brackets in the integrand of equation (B.5), i.e., we replace that function with the function $z(E) : z(E) = s\{(E/s+1/2)^2-(b^2+2b)/(4b+b)\}$ for $E < E'$, and $z(E) = sb/2$ for $E \geq E'$. E' is the value of E such that $s\{(E/s + 1/2)^2 - [b^2 + 2b]/[4(b + 1)]\} = sb/2$, unless such a value exceeds sb, in which case $E' \equiv sb$. $E' > 0$ always. E' is the lowest value of $\partial E_{\text{guessing}}$ that can correspond to the largest possible $\partial E_{\text{Occam}}$; see equation (B.2).

Now we must find the $\rho(E)$ that maximizes E_{Occam} subject to the constraints of equation (B.4). First hypothesize that for this $\rho(E)$ there is a value of $E > E'$, e, such that $\rho(e)$ is not 0. Define $\rho(e) \equiv D$. Now modify $\rho(E)$ by setting $\rho(e)$ to 0, $\rho(E')$ to its old value $+ D + C$, and $\rho(0)$ to its old value minus C, where $C \equiv D(e - E')/E'$. This modification preserves the constraints of equation (B.4), but it increases E_{Occam}. (Note that this rearranging might result in $\rho(0) < 0$. As usual, we are only interested in bounds, and therefore such negative densities are of no concern.) Therefore, for the purposes of finding the maximum of E_{Occam}, we can set $\rho(E)$ to 0 for all $E > E'$.

Now hypothesize that the optimal $\rho(E)$ is nonzero for some E lying between 0 and E'. Call this E value e, and call $\rho(e)$ D. Modify $\rho(E)$ by setting $\rho(e)$ to 0 and incrementing $\rho(E')$ by De/E'. Also increment $\rho(0)$ by $D(E' - e)/E'$. In carrying out this modification we have not violated the requirements on $\rho(E)$, but now E_{Occam} equals its old value minus $sD\{(e/s+1/2)^2-[b^2+2b]/[4(b+1)]\}$, plus $s\{D(E'-e)/E'\}\{(0+1/2)^2-[b^2+$

$2b]/[4(b+1)]\}$, plus $s\{De/E'\}\{(E'/s+1/2)^2 - [b^2+2b]/[4(b+1)]\}$. Adding up these terms, since $e < E'$, $D > 0$, and $s > 0$, this rearranging ends up increasing E_{Occam}. Therefore, to maximize E_{Occam}, one should set $\rho(E) = 0$ for all values of E other than 0 and E'.

The requirements on $\rho(E)$ then give us

$$\rho(E) = \left(\frac{E' - E_{\text{guessing}}}{E'}\right)\delta(E) + \left(\frac{E_{\text{guessing}}}{E'}\right)\delta(E - E') \tag{B.6}$$

where $\delta(x)$ is the Dirac delta function of x. Since we know that $z(E') = bs/2$, this allows us to write down the maximum possible value of E_{Occam}, given E_{guessing}:

The maximum of E_{Occam}, given E_{guessing},

$$\leq s\frac{b}{2}\frac{E_{\text{guessing}}}{E'} + \left[1 - \frac{E_{\text{guessing}}}{E'}\right][s]\left[\frac{1}{4} - \frac{b^2+2b}{4(b+1)}\right] \tag{B.7}$$

$$= s\frac{E_{\text{guessing}}}{E'}\left[\frac{3b^2+3b-1}{4(b+1)}\right] - s\left[\frac{b^2+b-1}{4(b+1)}\right] \tag{B.8}$$

Acknowledgments

I would like to thank J.D. Farmer and M. Casdagli for helpful comments on early versions of this manuscript. This work was done under the auspices of the Department of Energy.

References

[1] V.V. Anshelevich, B.R. Amirikian, A.U. Lukashin, and M.D. Frank-Kamenetskii, "On the ability of neural networks to perform generalization by induction," *Biological Cybernetics*, **61** (1989) 125–128.

[2] C.H. Bennett, In *Emerging Syntheses in Science*, D. Pines, ed. (Addison-Wesley, Santa Fe, NM, 1988).

[3] N. Cercone, ed. *Computational Intelligence*, **3** (1987), special issue on machine learning.

[4] G. Chaitin, "Randomness and mathematical proof," *Scientific American*, **232** (1975) 47–52.

[5] J. Denker, D. Schwartz, B. Wittner, S. Solla, R. Howard, L. Jackel, and J. Hopfield, "Automatic learning, rule extraction, and generalization," *Complex Systems*, **1** (1987) 877–922.

[6] B. Efron, "Computers and the theory of statistics: Thinking the unthinkable, *SIAM Review*, **21** (1979) 4670–4480.

[7] M. Feder, "Maximum entropy as a special case of the minimum description length criterion," *IEEE transactions on information theory*, **IT-32**, no. 6 (1986) 847–849.

[8] J. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence* (University of Michigan Press, Ann Arbor, 1975).

[9] J.E. Hopcrot and J.D. Ullman, *Introduction to Automata Theory, Languages, and Computation* (Addison-Wesley, 1979).

[10] E.T. Jaynes, "On the rationale of maximum-entropy methods," *Proceedings of the IEEE*, **70** (1982) 939–952.

[11] S. Kirkpatrick, C.D. Gellatt Jr., and M.P. Vecchi, "Optimization by simulated annealing," *Science*, **220** (1983) 671–680.

[12] N. Littlestone, "Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm," *Machine Learning*, **2** (1988) 285–318.

[13] S. Lloyd and H. Pagels, "Complexity as thermodynamic depth," *Annals of Physics* (1988) 186–213.

[14] T.M. Mitchell, "Generalization as search," *Artificial Intelligence*, **18** (1982) 203–226.

[15] S.M. Omohundro, "Efficient Algorithms with neural network behavior," *Complex Systems*, **1** (1987) 273–347.

[16] J. Pearl, "On the connection between the complexity and credibility of inferred models," *International Journal of General Systems*, **4** (1978) 255–264.

[17] T. Poggio and MIT AI Lab staff, "MIT progress in understanding images," published in *Proceedings of the Image Understanding Workshop*, L. Bauman, ed. (SAI Corporation, McLean, VA, 1988).

[18] T. Poggio, V. Torre, and C. Koch, "Computer vision and regularization theory," *Nature*, **317** (1985) 314–319.

[19] J.R. Quinlan and R.L. Rivest, "Inferring decision trees using the minimum description length principle," *Information and Computation*, **80** (1989) 227–248.

[20] J. Rissanen, "Stochastic complexity and modeling," *The Annals of Statistics*, **14** (1986) 1080–1100.

[21] D.E. Rumelhart and D.E. McClelland, *Explorations in the Microstructure of Cognition*, Vols. I and II (MIT Press, Cambridge, MA, 1986).

[22] T.J. Sejnowski and C.R. Rosenberg, "NETtalk: A parallel network that learns to read aloud," Johns Hopkins University Electircal Engineering and Computer Science Department Technical Report JHU/EECS-86/01 (1986).

[23] C.E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1949).

[24] R. Solomonoff, "A formal theory of inductive inference: I and II," *Information and Control*, **7**(1) (1964) 224.

[25] C. Stanfil and D. Waltz, "Toward memory-based reasoning," *Communications of the ACM*, **29** (1986) 1213–1228.

[26] L.G. Valinat, "A theory of the learnable," *Communications of the ACM*, **27** (1984) 1134–1142.

[27] S. Watanabe, "Inductive ambiguity and the limit of artificial intelligence," *Computational Intelligence*, **3** (1987) 304–309.

[28] D.H. Wolpert, "A benchmark for how well neural nets generalize," *Biological Cybernetcs*, **61** (1989) 303–315.

[29] D.H. Wolpert, "A mathematical theory of generalization: Part I," *Complex Systems*, **4** (1990a) 151–200.

[30] D.H. Wolpert, "A mathematical theory of generalization: Part II," *Complex Systems*, **4** (1990b) 201–249.