Reliable Computing 1 (3) (1995), pp. 225–237

Ockham's razor in interval identification

Bo H. Friesen and Vladik Kreinovich

Since real-life measurements cannot be absolutely precise, we never know the precise value of a physical quantity, we only know an interval of its possible values. Due to this uncertainty, there are several different models that are consistent with the same measurement results. Which model should we choose? In this paper, we show that Ockham's razor principle (*Entities should not be multiplied unnecessarily*) can lead to a natural criterion for choosing a model. As an example, we apply this criterion to data processing related to a reasonably simple psychological problem.

Бритва Оккама в интервальной идентификации

Б. Х. Фризен, В. Крейнович

Практические измерения не могут быть абсолютно точными. Поэтому мы никогда не знаем точное значение физической величины, но лишь интервал, в котором заключены ее возможные значения. Благодаря этой неопределенности могут существовать несколько различных моделей, совместимых с одними и теми же результатами измерений. Какую из них выбрать? В настоящей работе показывается, что принцип бритвы Оккама («сущности не следует умножать без необходимости») может привести к естественному критерию выбора модели. В качестве примера этот критерий применяется к обработке данных в достаточно простой психологической задаче.

1. Introduction

1.1. The need for identification in interval computations

A typical application of interval computations (see, e.g., [17]) is as follows: we want to know the value of a physical quantity y, and it is either impossible, or difficult to measure y directly. So, to estimate y, we measure other parameters $x_1, \ldots, x_n$ that are easy to measure, and then try to use the measurement results $\tilde{x}_1, \ldots, \tilde{x}_n$ to reconstruct y.

To be able to do that, we must find an algorithm f that transforms the results $\tilde{x}_i$ of measuring x_i into an estimate $\tilde{y} = f(\tilde{x}_1, \ldots, \tilde{x}_n)$ for y.

Since measurements are not absolutely precise, their results $\tilde{x}_i$ are different from the actual values x_i. Hence, the resulting estimate $\tilde{y}$ is different from the actual value of y. In measurement, we usually know the upper bound for an error, i.e., we know Δ_i such that $|\tilde{x}_i - x_i| \leq \Delta_i$. Interval computations help to find an interval of possible values of y, i.e., help to find Δ such that $|\tilde{y} - y| \leq \Delta$.

To apply these methods, we need to know f, i.e., in other words, we must *identify* the real-life object that we are analyzing.

1.2. Identification problem: general formulation[1]

In some situations, the dependency between x_i and y is already known: either from some approved theory, or from some previous experiments. But in many real-life situations, it is not known, so we must reconstruct it from the experimental data.

In other words, for every object that we want to be identified, we must measure x_i and y in several situations, and then reconstruct f from the measurement results. Let us denote the number of measurements by N, the results of k-th ($1 \le k \le N$) measurement by $\tilde{x}_1^{(k)}, \ldots, \tilde{x}_n^{(k)}$, $\tilde{y}^{(k)}$, and the accuracy of these measurements correspondingly by Δ_i and Δ.

Definition 1. *Let's fix an integer n. It will be called the number of variables. By measurement accuracies we mean a tuple $(\Delta_1, \ldots, \Delta_n, \Delta)$ of positive real numbers. By a measurement result we mean a tuple $(\tilde{x}_1, \ldots, \tilde{x}_n, y)$ of $n+1$ real numbers. By data D we mean a finite set of measurement results $(\tilde{x}_1^{(k)}, \ldots, \tilde{x}_n^{(k)}, \tilde{y}^{(k)})$, $1 \le k \le N$ (here, N denotes the number of measurement results). We say that a function $f(x_1, \ldots, x_n)$ is consistent with the data D if for every k from 1 to N, there exist values $x_i^{(k)}$ such that $|x_i^{(k)} - \tilde{x}_i^{(k)}| \le \Delta_i$ for $1 \le i \le n$, and $|\tilde{y}^{(k)} - f(x_1^{(k)}, \ldots, x_n^{(k)})| \le \Delta$.*

In these terms, the problem is to find a function f that is consistent with the data. Even if we measure x_i and y with absolute precision, this condition only restricts the value of f for N combinations $\vec{x} = (x_1, \ldots, x_n)$. For other values $\vec{x}$, there are no restrictions on $f(x_1, \ldots, x_n)$. Therefore, there are many different functions f that satisfy the above condition. Which of them to choose?

At first glance, it looks like a problem that cannot be solved. However, in real life, we usually have some idea of how y must depend on x_i. For example, we may know that the dependency of y on x_i is linear, i.e., that $y = C_1 x_1 + \cdots + C_n x_n + C_{n+1}$ for some coefficients C_i. Or, we can assume that f is quadratic, i.e., $y = \sum C_{ij} x_i x_j$. In general, we know a function $y = f(C_1, \ldots, C_p, x_1, \ldots, x_n)$, where p parameters C_i characterize an object. This function is usually called a *model*.

Definition 2. *By a model we mean a function $f(C_1, \ldots, C_p, x_1, \ldots, x_n)$ of $n+p$ variables, where $p \ge 0$. Variables $C_1, \ldots, C_p$ are called parameters of the model. We say that a function $g : R^n \to R$ is a particular case of the model f, if it can be obtained from f by fixing some values of C_i.*

Definition 3. *Suppose that we have a finite set of data $(D_1, \ldots, D_M)$. We say that a model f is adequate with respect to this set of data if for each j from 1 to M, there exists a particular case of this model that is consistent with the data D_j.*

Comment. In other words, a model f is adequate if for every object that we analyze, we can find the values of C_i for which the correspondent function $f(\vec{C}, \vec{x})$ is consistent with the measurement results for this particular object.

1.3. What if several models are adequate? Ockham's razor

What if data are consistent with several models? Which of them to choose?

One of the cases when this happens is as follows: suppose that we have a model with p parameters $C_1, \ldots, C_p$. Some objects are not consistent with this model, so a generalization is being developed, that has more parameters. Of course, if the data is consistent with the original

[1] For more details see [5, 10–12, 14, 15, 18, 21, 22, 24–31].

model, then it is also consistent with the generalized model. So, if the data is consistent with the original model, then we actually have two models that fit the data: the original model and the generalized one. In such a situation, it makes no sense to consider a generalized model, with the larger number of parameters, since the simpler one suffices.

This principle was first proposed by William of Ockham (also spelled Occam) around 1320, who said that *entities should not be multiplied unnecessarily*. This principle is called *Ockham's razor*.

1.4. The existing applications of Ockham's razor

This principle has been applied a lot in physics (for a brief survey of physical applications, see, e.g., [9]). We will just mention (see, e.g., [9, 16]) that when General Relativity first appeared, it contained one additional parameter Λ (also called *cosmological constant*). However, since all the experiments were consistent with the assumption that $\Lambda = 0$, Einstein decided to use only the model with $\Lambda = 0$. Later on, in 1961, another generalization of General Relativity was proposed by C. Brans and R. Dicke [2] under the name of a *scalar-tensor theory*. This theory contained an additional parameter $1/\omega$. Again, all future experiments were consistent with this parameter being equal to 0, therefore at present, the mainstream viewpoint is that we must use only the simplest model, i.e., General Relativity itself.

These and similar applications are applications to the situation when one of the models has more parameters than another one (e.g., it is a generalization of another one). In the general interval framework, this situation was described (with numerous examples) in [5, 12, 24, 25, 27, 28].

What to do if we have several competing models with the same number of parameters? This situation was analyzed only for the case when we have one object (and hence one data D). For probabilistic errors, criteria for choosing a model were analyzed in [1, 9, 20, 32], and for interval data in [5, 12, 24]. So, we arrive at the following problem:

1.5. Formulation of the main problem

Suppose that we have several models with the same number of parameters, and all of these models are consistent with the experimental data about several objects. Which of these models should we choose?

1.6. What we are planning to do

In this paper, first, we will describe the selection of a model as a mathematical problem. In the simplest case, when each model has a single physically meaningful parameter, the natural invariance conditions lead to a unique choice criterion (Section 2). This criterion can be interpreted in terms of Ockham's razor (Section 3). This interpretation enables us to generalize this criterion to the case of several parameters. A psychological example is given in Section 4.

In a special Appendix, we will also illustrate the difficulties of applying Ockham's razor.

2. Selecting a model as a mathematical problem: the simplest case

2.1. Main idea

Ockham's razor, intuitively speaking, can be understood as follows: the more information we need to provide in order to specify a particular case of the model, the worse the model. Ideal model should require the smallest possible amount of information to specify a particular case. In the above-mentioned applications of the Ockham's razor idea, we chose a model with the smallest number of parameters; this model is considered to be the simplest and therefore, the one chosen. This simple principle does not work if we compare several models with the same number of parameters. In this case, to describe the *simplicity* of a model, we must take into consideration not only how many parameters must be specified in order to select a unique model, but also, how different it is to specify these parameters. Crudely speaking, if in one model, we have narrow intervals for parameters C_i, then this model is much easier to specify this model than a competitive model in which an interval of possible values of parameters is much larger. In this section, for a simplest case, we will describe this idea in mathematical terms.

First, let us describe what we mean by a "simplest case".

2.2. What we mean by "the simplest case"

Since we have identified complexity with the number of parameters, the model is the simplest if it contains exactly one parameter. Such models, in their turn, can be (crudely) divided into two groups:

- In some one-parameter models, the parameter has no direct physical meaning.
- In some other models, the parameter has a direct physical meaning: it actually represents the value of some physical quantity. For example, if we consider a linear model $V = C \cdot I$ for the dependency of voltage V on the current I, then the parameter C has the known meaning of *resistance*.

If we compare two models for which parameters have no direct physical meaning, then we usually have no intuition on whether the interval of possible values is "large" or "narrow". In case the parameters have a direct physical meaning, we often have some understanding of whether the accuracy is good or not. This intuition definitely helps in choosing a model, so, we would like to formalize it.

2.3. Unit-invariance: a way to formalize physical intuition

One important feature of physical quantities that we will use is that usually, the choice of a *unit* in which we measure this quantity is rather arbitrary: for example, we can measure length in centimeters or in inches. If in a model, we have resistance measured in ohms, then it is reasonable to demand that the same model, but with resistance expressed in kilohms, will be of the same quality.

The idea of such "invariance relative to the choice of a unit" has been successfully used in physics, starting from the pioneer work [23]. Together with other physical conditions, unit-invariance can explain the fundamental physical equations such as Maxwell's equation that describe electromagnetism, Schroedinger's equations that describe quantum mechanics, and Einstein's equations that describe space-time geometry [7, 8, 13]. We will see that in our problem, unit invariance also leads to a unique comparison criterion.

Let us describe unit-invariance in mathematical terms. If we change a unit to a one that is λ times smaller, then the resulting numerical values are multiplied by λ. For example, if instead of inches, we consider cm, that are ≈ 2.54 times smaller, then, instead of 2 in, we get $2 \cdot 2.54$ cm. So, if we have two intervals $[a^-, a^+]$ and $[\lambda \cdot a^-, \lambda \cdot a^+]$, and we do not know what units were used to describe these intervals, then they could be one interval, but expressed in two different units. Therefore, in this case, we have no reasons to choose one of these intervals as "narrower". On the other hand, if one interval is a proper subset of another, then the first interval is clearly narrower. So, we arrive at the following definition.

2.4. Definitions and the main result

Definition 4. *Let I denote the set of all positive intervals, i.e., intervals $\mathbf{a} \subseteq (0, \infty)$. By a pre-ordering, we mean a transitive reflexive relation $\preceq$ on the set I. We will use the following denotations:*

- $\mathbf{a} \sim \mathbf{b}$ *if* $\mathbf{a} \preceq \mathbf{b}$ *and* $\mathbf{b} \preceq \mathbf{a}$.
- $\mathbf{a} \prec \mathbf{b}$ *if* $\mathbf{a} \preceq \mathbf{b}$ *and* $\mathbf{b} \not\preceq \mathbf{a}$.

Let a pre-ordering be given.

- *We say that the pre-ordering is natural if* $\mathbf{a} \subset \mathbf{b}$ *implies* $\mathbf{a} \prec \mathbf{b}$.
- *We say that the pre-ordering is unit-invariant if for every* $a^- \le a^+$ *and for every* $\lambda > 0$, $[a^-, a^+] \sim [\lambda a^-, \lambda a^+]$.

Proposition 1. *There exists exactly one natural unit-invariant pre-ordering:* $[a^-, a^+] \preceq [b^-, b^+]$ *iff* $d(a) \le d(b)$, *where* $d(\mathbf{a}) = (a^+ - a^-)/(a^+ + a^-)$ *and* $d(\mathbf{b}) = (b^+ - b^-)/(b^+ + b^-)$.

Proof. The parameter $d(\mathbf{a})$ can be rewritten as

$$\frac{(a^+/a^-) - 1}{(a^+/a^-) + 1} = 1 - \frac{2}{(a^+/a^-) + 1}.$$

The function $1 - 2/(x+1)$ is strictly increasing, so, $d(\mathbf{a}) \le d(\mathbf{b})$ iff $a^+/a^- \le b^+/b^-$. Hence, to prove Proposition 1, it is sufficient to prove that $[a^-, a^+] \preceq [b^-, b^+]$ iff $a^+/a^- \le b^+/b^-$. Let us consider three possible cases:

- Let $a^+/a^- = b^+/b^-$. Then, if we define $\lambda = b^-/a^-$, we get $\lambda \cdot a^\pm = b^\pm$. Hence, due to unit-invariance, $\mathbf{a} \sim \mathbf{b}$.
- Now, let $a^+/a^- < b^+/b^-$. Let us again take $\lambda = b^-/a^-$. Then, $\lambda \cdot a^- = b^-$, and $\lambda \cdot a^+ < b^+$. Since $\preceq$ is unit-invariant, we get $\mathbf{a} = [a^-, a^+] \sim [b^-, \lambda \cdot a^+]$. From $\lambda \cdot a^+ < b^+$, we conclude that $[b^-, \lambda \cdot a^+] \subset [b^-, b^+]$ and hence, that $[b^-, \lambda \cdot a^+] \prec [b^-, b^+] = \mathbf{b}$. So, $\mathbf{a} \sim [b^-, \lambda \cdot a^+] \prec \mathbf{b}$, and $\mathbf{a} \prec \mathbf{b}$.

- Similarly, from $a^+/a^- > b^+/b^-$, we conclude that $\mathbf{b} \prec \mathbf{a}$.

From these three cases, we conclude that $[a^-, a^+] \preceq [b^-, b^+]$ iff $a^+/a^- \leq b^+/b^-$. □

2.5. Resulting recommendation

So, in this simplest case, we choose a model for which the "relative width" of the interval $\mathbf{a}$ of possible value of the parameter is the smallest.

2.6. This idea is only applicable if the parameters of the model have direct physical meaning

We deduced this idea in the assumption that the parameter of the model has a direct physical meaning. Let us show that this idea is not always applicable to the situations in which the parameter of the model is *not* directly physically meaningful (we are thankful to the anonymous referee who provided us with the idea of this example).

Let us consider the dependency between the voltage and the current. We assume that V is a linear function of I. In natural physical terms, this assumption can be described by a model $V = C \cdot I$ with a physically meaningful parameter $C = R$. Instead of this physically meaningful parameter, we can reformulate the model in a mathematically equivalent form $V = (C_1 - 10) \cdot I$. The new parameter C_1 does not have any direct physical meaning, so, the formal transformation $C_1 \to \lambda C_1$ does not correspond to any physically meaningful "change of a unit". If, in spite of that fact, we apply the above-describe criterion to compare the two mathematically equivalent models, we will arrive at the absurd conclusion that the second model is mush better: indeed, if, e.g., $C \in \mathbf{C} = [0.9, 1.1]$, then $C_1 \in \mathbf{C}_1 = [10.9, 11.1]$, so $d(\mathbf{C}_1) = 0.01 \ll d(\mathbf{C}) = 0.1$.

Let us now interpret this result in terms of Ockham's razor.

3. Ockham's razor as a criterion for choosing a model: a heuristic idea

3.1. Main idea

As we have already mentioned, Ockham's razor can be understood as follows: the more information we need to provide in order to specify a particular case of the model, the worse is the model. Ideal model should require the smallest possible amount of information to specify a particular case. In the above-mentioned physical applications of the Ockham's razor idea, we estimated this amount of information as the number of parameters. With this estimate in mind, the absence of "unnecessary entities" means that we take a model with the smallest possible number of parameters. This estimate is, however, too crude to distinguish between the two models with exactly the same number of parameters. In this case, it is reasonable to take into consideration not only how many parameters we must fix to specify a particular case of a model, but also how many *bits* we must specify (i.e., how many binary digits (0's and 1's) we must use to get a computer description of a specification).

For example, if we have two models, both with one parameters, and in one of them C_1 ranges from 0 to 1, in another from 0.99 to 1, then it sounds reasonable to conclude that the first model has an unnecessarily wide parameter range, and therefore, the second model is preferable.

What if we have two intervals, $[0.9, 1.1]$ and $[99, 101]$? The *absolute* range (i.e., the length of the interval of possible values of C_1) is larger for the second model, but intuitively, it sound reasonable to conclude that the second model is preferable, because it has a smaller *relative* range: in the second model, we already know the parameter with the precision of 1%, while in the first model, the accuracy of an a priori knowledge of this parameter is 10%.

So, it is reasonable to compare *relative ranges*, i.e., to compare the percentages with which the values of these parameters can deviate from the average.

3.2. For models with one parameter, how to choose a model?

If the interval of possible values of some parameter C_1 is $[C_1^-, C_2^+]$, then the average is $(C_1^- + C_1^+)/2$, the absolute deviation from the average is $(C_1^+ - C_1^-)/2$, and the relative deviation d from the average equals $[(C_1^+ - C_1^-)/2]/[(C_1^+ + C_1^-)/2] = (C_1^+ - C_1^-)/(C_1^+ + C_1^-)$. For models that have one parameter, we will use this value d as a criterion for choosing a model: namely, we choose a model with $d \to \min$.

This is exactly the criterion that we came up with in Section 2.

3.3. Analogy

To justify our reasoning, let us invoke the following analogy: when we speak about measuring devices, we can say that one of them is more accurate (or more precise) than another. For example, a complicated system that measures distance from Earth to Moon with a centimeter precision (relative accuracy about 10^{-10}) is certainly much more precise than a ruler that enables its user to measure distances from 0 to 10 cm with a millimeter precision (i.e., with relative precision 1%). So, when we compare precisions, we do not usually compare absolute precisions, we compare relative ones.

3.4. How is d related to the number of bits

We started with the idea of using the number of bits as a criterion, and then "jumped" to relative deviation from the average d. Is there a formal relationship between these two notions? Heuristically, yes.

For real-life objects, values of the parameters C_i will be obtained from measurement results. If we make all the measurements with a relative accuracy δ (i.e., if $\Delta_i/|\tilde{x}_i| \le \delta$ and $\Delta/|\tilde{y}| \le \delta$), then we get the resulting values C_i also with a similar relative precision. Strictly speaking, this is not always true, but in general, if we start with the real numbers that are known with 3 decimal digits (i.e., with precision 0.1%), then we get the results with 3 (or in the worst case 2) valid decimal digits (unless, of course, the algorithm is really badly numerically unstable). So, every specification of the model is obtained with relative accuracy δ and hence with absolute accuracy $\approx d_{\text{abs}} = [(C_1^+ + C_1^-)/2]\delta$. Therefore, specifications that differ by this amount may really describe the same object. Therefore, there are only $(C_1^+ - C_1^-)/d_{\text{abs}}$ different specifications: the ones that correspond to the values $C_1 = C_1^-, C_1^- + d_{\text{abs}}, C_1^- + 2d_{\text{abs}}, \ldots, C_1^- + jd_{\text{abs}}, \ldots, C_1^+$.

The more possible specifications, the more bits we must use to describe a specification. Namely, with one bit, we can describe two possible cases (corresponding to 0 and 1), with b bits, we can describe 2^b different binary numbers, and therefore, 2^b cases. Hence, to describe S specifications, we need b bits, where $2^b = S$, i.e., we need $b = \log_2 S$ bits. Since $S \approx (C_1^+ - C_1^-)/d_{\text{abs}} \approx [(C_1^+ - C_1^-)/(C_1^+ + C_1^-)]/(0.5\delta)$, we thus need $b \approx \log_2\left((C_1^+ - C_1^-)/(C_1^+ + C_1^-)\right) - \log_2 \delta + 1 = \log_2 d - \log_2 \delta + 1$ bits. So, the smaller d, the fewer bits we need.

3.5. For models with several parameters, how to choose a model? An idea

If we have a model with p parameters $C_1, \ldots, C_p$, and the range of i-th parameter is $[C_i^-, C_i^+]$, then for each parameter, we have $S_i \approx (C_i^+ - C_i^-)/\delta \approx d_i/(\delta/2)$ different possible specifications, where by d_i, we denoted the relative deviation $d_i = (C_i^+ - C_i^-)/(C_i^+ + C_i^-)$ of i-th parameter C_i from its average value. Then, totally, we have $S = S_1 \times S_2 \times S_3 \times \cdots \times S_p$ different possible specifications. Since $S \approx d_i/(\delta/2)$, we have $S \approx (d_1 \ldots d_p)/(\delta/2)^p$. The bigger the product $d_1 \ldots d_p$, the bigger S and therefore, the bigger the number of bits $b \approx \log_2 S$ bits that we need to specify a particular case of the model. Therefore, it is reasonable to choose a mode for which the product $d_1 \ldots d_p$ is the smallest possible.

3.6. For models with several parameters, how to choose a model? A proposed method

For every model $f(\vec{C}, \vec{x})$, and for each of its parameters C_i, let us denote by C_i^-, the smallest possible value of C_i that is consistent with one of the data (i.e., with one of the objects). By C_i^+, we will denote the biggest possible value of C_i for all vectors $\vec{C}$ for which this model is consistent with one of the objects. By a *relative range* d_i of i-th parameter C_i, we mean a value $(C_i^+ - C_i^-)/(C_i^+ + C_i^-)$. For each model, we can thus compute the product $d_1 \ldots d_p$. We recommend to choose a model for which this product is the smallest possible.

Comment. Our arguments were based on approximate equalities. Therefore, if the product computed for one model is only slightly smaller than the product computed for another model, it can well be that the second model is actually better. In other words, the proposed choice criterion is really convincing only if for some model the product is really much smaller than for other competing models.

4. Example: computational complexity in the human mind

The problem on which we want to show the use of the criteria proposed in the previous section is motivated by the desire to know how the human mind works. One of the ways to find out exactly what algorithm the human brain is using to solve problems from some problem set is to measure the time spent by a human brain for different problems from this set.

There are two main groups of algorithms (see, e.g. [4]): polynomial-time algorithms and exponential-time algorithms. Polynomial time means that the time required to solve a

problem of size n is limited by a polynomial of n. Usually, this time grows as Cn^k for some integer k. Exponential time means that the computation time grows as a^n for some $a > 0$. Polynomial-time algorithms are usually considered feasible because even for reasonably large n (e.g., $100 \leq n \leq 1000$), Cn^k is still within our reach. Exponential-time algorithm are usually considered infeasible because, e.g., 2^{300} already exceeds the lifetime of the Universe.

It is thus interesting to find out whether the human mind uses a polynomial or an exponential algorithm to solve a certain problem.

As an example of such a problem, we took the Tower of Hanoi problem. In this problem, there are three pads, and n disks of different size. Initially, all the disks are on the first pad in the order of their sizes: the largest disk is at the bottom, the smallest one if on the top. On each step, we can take a top disk from one of the pads and place it on top of some other pad. The objective is to rearrange the disks in such a way that all the disks are located on the third pad (in the same order as they were initially located on the first pad).

This problem is well known to be exponential-time: there is an algorithm that solves this problem in time $2^n - 1$, and it can be proved that no algorithm can solve it faster (see, e.g., [19]). The algorithm is simple, so a person who knows the algorithm can make the moves real fast.

In this analysis, we tested ten subjects who did not know the algorithm. They were three females and seven males, ages from 14 to 66, with educational background from high school to Master's degree. As a result, for $n = 3$, we got the following times (in seconds): 74, 43, 37, 61, 126, 61, 38, 31, 70, 25. The smallest time was 25 sec, the biggest 126 sec.

We tested these results against two classes of models: exponential-time model $t(n) = a^n$, and polynomial-time models $t(n) = Cn^k$ for different k. We have $t \in [25, 126]$. Since we have only one value of n ($n = 3$), both models are evidently consistent with the experimental data: for any t, we can take $a = t^{1/3}$ and $C = t/3^k$.

For an exponential model, $a = t^{1/3}$. Therefore, the interval of possible values of a is $[25^{1/3}, 126^{1/3}] \approx [3, 5]$. Hence, the midpoint is ≈ 4, and the relative range is $\approx 1/4 = 0.25$.

For a polynomial-time model, $C = t/3^k$. Therefore, the interval of possible values of C is $[25/3^k, 126/3^k]$, the midpoint is $75.5/3^k$, and the relative range is $50.5/75.5 \approx 2/3$.

Since $1/4 \ll 2/3$, according to our criterion, this means that our data support the (correct) exponential-time model.

This same example also shows that choosing *relative* range as opposed to *absolute* was a good idea, because the absolute accuracy of the polynomial model $50.5/3^k$ tends to 0 as $k \to \infty$, and therefore, is smaller than for the correct exponential-time model.

Warning. The above example is only given as an illustration. As we have mentioned in Section 2, our simple choice of the model is reasonable only if we have already made a pre-section of the models, and we are already left only with the models in which the parameters have direct physical meaning. The following example, proposed by the referee, illustrates this warning: Suppose that in addition to the above-described two models, we consider the model $t(n) = a^{kn}$ for different k, then we would have $a^- = 25^{1/(3k)}$, $a^+ = 125^{1/(3k)}$. As $k \to \infty$, we have $a_k^+ \to 1$, $a_k^- \to 1$ and hence, $d([a_k^-, a_k^+]) \to 0$. So, we end up with a meaningless conclusion that models with large k are better than the original exponential model (that is mathematically absolutely equivalent to each of them).

5. Appendix: problems with Ockham's razor

Let us show on two examples that there are some problems with using Ockham's razor. The first example will be about real numbers, the second one about logic in general.

5.1. An example with real numbers

As we have already mentioned, one of the reasonable applications of Ockham's razor idea is as follows [27, 28]: if we have a model, and the data is consistent with the assumption that one of its parameters is equal to 0, then we can assume that this value is 0.

The problem appears if this condition is satisfied for two different parameters C_i. To illustrate it, let us consider the simplest possible case: a linear model with two parameters $y = C_1x_1 + C_2x_2$. Let us assume that we have only one measurement result $(\tilde{x}_1, \tilde{x}_2, \tilde{y}) = (1, 1, 1.5)$, and that the accuracies are $(0, 0, 0.5)$. This means that the measurements of x_i were very precise, so $x_i = \tilde{x}_i$, and the interval of possible values of y is $[1, 2]$. This data is evidently consistent with the assumption that $C_1 = 0$, so we can take $C_1 = 0$. On the other hand, this same data is consistent with the assumption that $C_2 = 0$, so we can take $C_2 = 0$. We have two different models: $y = C_1x_1$ and $y = C_2x_2$. If we try to equate both C_1 and C_2 to 0, we get a model $f = 0$, that is not consistent with the data.

In general, the problem is as follows: by applying this principle to different C_i, we get different models; and if we try to equate both C_i to 0, we may end up with a wrong model.

5.2. Second example: general logic

A natural way to reformulate Ockham's razor in terms of logic and set theory is as follows. In these terms, an *entity* can be understood as a set. So, the idea is: if for two sets X and Y that are described by different formulas, it is possible to assume that $X = Y$, then we should take $X = Y$.

Let us formalize this seemingly natural formalization and show that it leads to a contradiction.

Definition 5. *Let ZF denote a standard axiomatic of set theory (see, e.g., [6]). We say that a model M of ZF is an Ockham model if for every two formulas $\phi(x)$ and $\psi(x)$, for which the sets $\{x|\phi(x)\}$ and $\{x|\psi(x)\}$ exist, and it is consistent with ZF that $\{x|\phi(x)\} = \{x|\psi(x)\}$, this equality holds in M.*

Proposition 2. *There exist no Ockham models.*

Proof. Indeed, since ZF is incomplete, there exists an undecidable formula F, i.e., a formula for which neither F, nor its negation $\neg F$ can be deduced from ZF. This implies that F is consistent with ZF, and that $\neg F$ is also consistent with ZF. Let us take $X^+ = \{x \in \{0\}|F\}$, $X^- = \{x \in \{0\}|\neg F\}$, and $Y = \{0\}$. The formula F is true if and only if $X^+ = Y$. The formula F is false iff $X^- = Y$. Since F is consistent with ZF, it is therefore consistent with ZF that $X^+ = Y$. Hence, in an Ockham model, we would have $X^+ = Y$, and thus F is true. Similarly, from the fact that $\neg F$ is consistent with ZF, we will conclude that in an Ockham model, $X^- = Y$, and thus, F is false. So, in an Ockham model, F is simultaneously true and false. This contradiction shows that there are no Ockham models. □

Comment. As one can easily see, our arguments apply not only to ZF but practically to all known axiomatic set theories.

Acknowledgments

This work was sponsored by NSF grant No. CDA–9015006, NASA Research Grants No. 9–482 and NAG 9–757, and a Grant No. PF90–018 from the General Services Administration (GSA), administered by the Materials Research Institute. The authors are greatly thankful to the anonymous referees for their helpful remarks.

References

[1] Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. *Occam's razor*. Information Processing Letters **24** (1987), pp. 377–380; reprinted in: Shavlik, J. W. and Dietterich, T. G. (eds) "Readings in Machine Learning", Morgan Kaufmann, San Mateo, CA, 1990, pp. 201–204.

[2] Brans, C. and Dicke, R. H. *Mach's principle and a relativistic theory of gravitation*. Physics Review **124** (1961), p. 925.

[3] Constable, S. C., Parker, R. L., and Constable, C. G. *Occam's inversion: a practical algorithm for generating smooth modela from electromagnetic sounding data*. Geophysics **52** (1987), pp. 289–300; reprinted in: Lines, L. R. (ed.) "Inversions of Geophysical Data. Geophysics reprint series No. 9", Society of Exploration Geophysicists, 1988, pp. 529–540.

[4] Cormen, Th. H., Leiserson, Ch. L., and Rivest, R. L. *Introduction to algorithms*. MIT Press, Cambridge, MA, 1990.

[5] Dyvak, N. P. *Design of saturated experiment in interval model building*. In: "Proceedings of the International Conference on Interval and Stochastic Methods in Science and Engineering INTERVAL'92", Moscow, 1992, Vol. 1, pp. 42–45 (in Russian; English abstract Vol. 2, p. 23).

[6] Enderton, H. B. *A mathematical introduction to logic*. Academic Press, N.Y., 1972.

[7] Finkelstein, A. M. and Kreinovich, V. *Derivation of Einstein's, Brans-Dicke and other equations from group considerations* In: Choque-Bruhat, Y. and Karade, T. M. (eds) "On Relativity Theory. Proceedings of the Sir Arthur Eddington Centenary Symposium, Nagpur, India 1984", Vol. 2, World Scientific, Singapore, 1985, pp. 138–146.

[8] Finkelstein, A. M., Kreinovich, V., and Zapatrin, R. R. *Fundamental physical equations uniquely determined by their symmetry groups*. Lecture Notes in Mathematics **1214**, Springer-Verlag, Berlin—Heidelberg—N.Y., 1986, pp. 159–170.

[9] Garrett, A. J. M. *Ockham's razor*. In: Grandy, W. T. and Schick, L. H. (eds) "Maximum Entropy and Bayesian Methods", Kluwer, Amsterdam, 1991, pp. 357–364.

[10] Ivshina, A. V., Kuznetsov, V. A., and Kadagidze, Z. G. *Dymanics of biological populations*. Gorky University, Gorky, 1985 (in Russian).

[11] Jaulin, L. and Walter, E. *Estimation of the parameters of nonlinear models from experimental data via interval analysis*. In: "Abstracts for an International Conference on Numerical Analysis

with Automatic Result Verification: Mathematics, Application and Software, February 25—March 1, 1993", Lafayette, LA, 1993, p. 38.

[12] Khalygov, M. A. *Simulation of engineering systems under interval uncertainty.* In: "Urgent Problems of Applied Mathematics, Proceedings of the USSR National Conference, Saratov, May 20–24, 1991, Part 1", pp. 171–176 (in Russian).

[13] Kreinovich, V. *Derivation of the Schrödinger equations from scale invariance.* Teoreticheskaya i Mathematicheskaya Fizika **26** (3) (1976), pp. 414–418 (in Russian); English translation: Theoretical and Mathematical Physics **8** (3) (1976), pp. 282–285.

[14] Kuznetsov, V. A. and Pankov, P. S. In: "Proceedings of the 1st Republican Conference on Young Scientists, Section "Biological Sciences", Frunze, Kyrgyzstan, 1981 (in Russian).

[15] Kuznetsov, V. A., Pankov, P. S., and Kenenbaeva, G. M. *Automatic checking of consistency between the theoretical dependency and the experimental data when the measurement accuracy is known.* In: "Automation of Scientific Research, Abstracts of a 21 USSR National Workshop, Cholpon-Ata, September 1987", Ilim, Frunze, 1987, p. 91 (in Russian).

[16] Misner, C. W., Thorne, K. S., and Wheeler, J. A. *Gravitation.* Freeman, San Francisco, 1973.

[17] Moore, R. E. *Methods and applications of interval analysis.* SIAM, Philadelphia, 1979.

[18] Pankov, P. S. *Algorithms for proving stability results and for global optimization in a bounded domain.* VINITI, Publ. No. 5250–84, Moscow—Frunze, 1984 (in Russian).

[19] Rawlins, G. J. E. *Compared to what?* Computer Science Press, Freeman, N.Y., 1992.

[20] Risanen, J. *Stochastic complexity and modeling.* Ann. Stat. **14** (3) (1986), pp. 1080–1100.

[21] Schweppe, F. C. *Recursive state estimation: unknown but bounded errors and system inputs.* IEEE Transactions on Automatic Control **13** (1973), p. 22.

[22] Schweppe, F. S. *Uncertain dynamic systems.* Prentice Hall, Englewood Cliffs, NJ, 1973.

[23] Sedov, L. I. *Similarity and dimensional methods in mechanics.* Academic Press, N.Y., 1959.

[24] Sidulov, M. V. and Trusov, V. A. *An approach to design of interval models based on the principle of indistinguishability of solutions.* In: "Abstracts of the 3rd USSR National Conference on Prospective Methods of Experiment Analysis and Design for Random Fields and Processes, Grodno, 1988", Moscow, 1988, pp. 177–178 (in Russian).

[25] Voshchinin, A. P. *Some questions of application of interval mathematics in parameter estimation and decision making.* In: "Proceedings of the International Conference on Interval and Stochastic Methods in Science and Engineering INTERVAL'92", Moscow, 1992, Vol. 2, pp. 32–33.

[26] Voshchinin, A. P., Bochkov, A. F., and Sotirov, G. P. *Interval analysis of data as an alternative to regression analysis.* In: "9th USSR National Conference on Experiment Design and Automation in Science", Moscow, 1989, Part. 1, pp. 24–25 (in Russian).

[27] Voshchinin, A. P., Bochkov, A. F., and Sotirov, G. P. *Method of data analysis with a nonstatistical interval error.* Industrial Laboratory **56** (7) (1990), pp. 854–860 (in Russian).

[28] Voshchinin, A. P. and Sotirov, G. R. *Optimization in case of uncertainty*. Tehnika, Moscow—Sofia, 1989 (in Russian).

[29] Walter, E. and Piet-Lahanier, H. *Robust nonlinear parameter estimation in the bounded noise case*. In: "Proceedings of the 25th IEEE Conference on Decision and Control", IEEE, N.Y., 1986, p. 1037.

[30] Walter, E. and Piet-Lahanier, H. *Estimation of non-uniquely identifiable parameters via exhaustive modeling and membership set theory*. Mathematics and Computers in Simulation **28** (6) (1986).

[31] Walter, E. and Piet-Lahanier, H. *Estimation of parameter bounds from bounded-error data: a survey*. Mathematics and Computers in Simulation **32** (1990), pp. 449–468.

[32] Zemel, R. S. *A minimum description length framework for unsupervised learning*. University of Toronto, Department of Computer Science, Ph. D. Dissertation.

Received: June 15, 1993
Revised version: April 21, 1995

Computer Science Department
University of Texas at El Paso
El Paso
TX 79968
USA

Current address of B. H. FRIESEN:
Texas Instruments
Dallas, TX 75265
USA
E-mail: bfriesen@dsk92.itg.ti.com