

Verisimilitude and Likelihood

Malcolm Forster, October 5, 2004.

Historical Introduction

Karl Popper viewed science as a series of conjectures and refutations (see Lakatos 1970 for a review of Popper's program and an extension to it designed to meet the skeptical challenge posed by Kuhn 1970). Since a refutation is the rejection of a false hypothesis, Popper took it to be obvious that the ultimate goal of science was truth (why reject falsehoods otherwise?). He also saw his philosophy of science as being an alternative to those based on probabilistic theories of confirmation, even though these also assumed the goal of science to be truth. After all, it is probability *of truth* that is valued. In chapter 10 of Popper (1963), saw a way of distinguishing between the two competing approaches. Popper argued that there was a general conflation of two things—the probability that a theory is true and the closeness of a theory to the truth. Without thinking about it very much, one is inclined to think that they go hand in hand. But a quick example is sufficient to show that this inclination is wrong.

Consider two hypotheses that attempt to predict the correct time to the nearest minute. Hypothesis A says that the time is that given by my watch, which is 3 minutes fast. Hypothesis B says that the time displayed on clock B. Clock B happens to be stopped. Both hypotheses are false. But our intuitions say that A is closer to the truth than B because A makes more accurate predictions than B, except for rare instances. Given this information, we want to base our choice on probability. The question is how should probabilities be used in the evaluation. Do we want to compare the probabilities that the hypotheses are true, or probabilistic expectation that the hypotheses are close to the truth. Hypotheses of type A are *always* false, so the probability of A being true is zero. However, hypotheses of type B are exactly right every twelve hours, so the probability of B being true are greater than the probability that A is true. The expected closeness to the truth and probability of truth do not coincide.

Popper's hope was that falsificationism would be turn out to be the methodology that is effective in bringing our theories closer to the truth. He also knew that he needed make the concept of verisimilitude precise in order to connect it with falsification. Unfortunately, Miller (1974) and Tichý (1974) showed that Popper not only failed in this hope, but also that by his definition, it was impossible for any false theory to have greater verisimilitude than any other false theory.

Other definitions of verisimilitude replaced Popper's (see Niiniluoto 1998 for a recent review). The problem with the current state of the Popperian program is that there are many different definitions of verisimilitude, none of which has gained universal acceptance. In my view, the major source of dissent still centers around the problem of language variance introduced by Miller (1974) and Miller (1975).

I argue that there is a language-invariant definition of verisimilitude that is an integral part of the philosophy of the quantitative sciences, even though it is not everything that realists had hoped for. It does not show that our best scientific theories provide at least

an approximately true picture of the world *behind the appearances*.

Popper's Problem of Progress

Popper (1963) noted that there is no general agreement on the answers to two very basic questions:¹

(A) Can we specify what scientific progress consists of?

(B) Can we show that science has actually made progress?

One quick answer to (A) is that

(1) Science aims at true theories and that progress consists of the fulfillment of this aim.

In answer to (B), we should add:

(2) Science has made progress in meeting this aim.

The problem now arises when we add a third plausible statement to the first two:

(3) Scientific theories have all been false.

In the history of planetary astronomy for example, Ptolemy's geocentric theory is false, Copernicus's version of the heliocentric theory is false, Kepler's laws are false, Newtonian gravitational theory is false, since it has been superseded by Einstein's general theory of relativity. And it would be naïve to suppose it won't be superseded by an even better theory some day. So, how have we made progress towards the truth at any time if all the theories known at that time are false. When Popper formulated the problem, he had in mind progress towards the truth about the world behind the observable phenomena.

The Problem of Verisimilitude

The now famous problem of verisimilitude flows from Popper's problem of progress (Musgrave, unpublished):

Realists...seem forced to give up *either* their belief in progress *or* their belief in the falsehood of all extant scientific theory. I say 'seemed forced' because Popper is a realist who wants to give up neither of them. Popper has the radical idea that the conflict between (1), (2), and (3) is only an apparent one, that progress with respect to truth is possible through a succession of falsehoods because one false theory can be *closer to the truth* than another. In this way Popper discovered the (two-fold) *problem of verisimilitude*;

(A*) Can we explain how one theory can be *closer to the truth*, or has *greater verisimilitude* than another?

(B*) Can we show that scientific change has sometimes led to theories which are closer to the truth than their predecessors?

Musgrave's point that the problem has two parts is important. In fact, I would like to extend Musgrave's point by describing the problem in a more detailed way. This task is undertaken in the following section.

¹ I owe this formulation of the problem to Alan Musgrave.

Methods and Goals and the Connection between them

Philosophers of science aim to *explain* how the methods of inference used in science succeed in achieving certain goals of inference. Popper's problem of progress is the problem of defining what scientific progress is and then arguing that progress has been made. I want to say that the problem is deeper than this. Philosophers of science also need to explain the success of scientific methods in achieving certain well defined goals. This requires more than some argument to show that science today is closer to the truth than science of yesterday. It requires that we explain why it is plausible to claim that certain *methods of science have been responsible for that progress*. There should be some plausible account of the causal connection between methods and the achievement of goals.

I am committed not only to the project of defining some objective sense in which science has made progress towards the truth, but I also want the definition of verisimilitude that shows that a curve fitting methodology of science is effective, to some extent, in achieving various goals of inference. I have already suggested that there are at least two goals here: (1) To produce model that predictively more accurate than their predecessors (and instrumentalist and realist goal). (2) To produce models that are good representations of the reality behind the phenomena (a realist goal).

Popper (1963) had also spent many pages describing the methodology of science in terms of falsificationism, so he attempted to define verisimilitude in a way that would help make sense of his methodology. Since his methodology is different in detail from the one I have described, it is good news for me that his definition failed. If I can replace it with something that fits in with a curve fitting methodology, then Popper's failure is grist for my mill.

Popper's Definition of Verisimilitude

An natural reaction to the fact that all our theories have been false is to point out that a false can have true logical consequences. Recall that the definition of entailment says that it is impossible for the consequence of a theory to be false if the theory is true, but it is possible for a consequence of a theory to be true if the theory a theory is false. So, it may be that our best scientific theories are partially true, and this is the idea that Popper tried to capture in his definition of verisimilitude:

DEFINITION: Theory A is closer to the truth than theory B if and only if (i) all the true consequences of B are true consequences of A , (ii) all the false consequences of A are consequences of B , and (iii) either and some true consequences of A are not consequences of B or some false consequences of B are not consequences of A . Clause (iii) is designed to rule out the possibility that A and B are equally close to the truth.

The intuitive idea behind the definition is that A is closer to the truth than B if it has more true consequences and fewer false consequences. That is, A not only says more about what is true, but by having fewer false consequences, is less falsifiable at the same time. From this we see how Popper wanted his definition to be such that the solution to part B^* of the problem of verisimilitude was a methodology of conjectures and refutations, which says that we judge the best theory to be one that is the most falsifiable out of those that are unfalsified. It is interesting to note that the definitions that have

superseded Popper's are designed to retain this connection with falsificationism. That is why the connection between the two parts of the problem is important to keep in mind.

The fatal flaw in Popper's definition was noticed independently by Tichý (1974) and Miller (1974). They showed that, according to Popper's definition, for *any* false theories *A* and *B* neither is closer to the truth than any other. This flaw is *fatal* because the philosophical motivation behind Popper's definition was to solve the problem of progress by showing that it is possible that some *false* theories are closer to the truth than other *false* theories.

Rather than repeating the proofs found in Tichý (1974) and Miller (1974), I shall provide the intuitive insight behind the proof within a context of the weather example introduced in these papers. A thorough understanding of the same example will then serve the dual purpose of evaluating the various corrections to Popper's definition.

The Weather Example

Suppose that the weather outside is hot (*h*) or cold ($\sim h$), rainy (*r*) or dry ($\sim r$), windy (*w*) or calm ($\sim w$). Suppose that the truth is that it's hot, rainy, and windy ($h \ \& \ r \ \& \ w$). Now consider two competing theories about the weather. *A* says that it's cold, rainy and windy ($\sim h \ \& \ r \ \& \ w$) and *B* says that it's cold, dry and calm ($\sim h \ \& \ \sim r \ \& \ \sim w$). Both theories are false, but intuitively, one might think that *A* is closer to the truth than *B*. Yet Popper's definition does not yield that result.

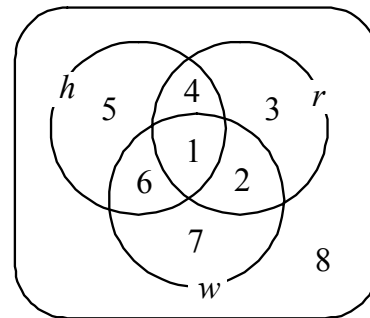
First, *B* has some true consequences that *A* does not have. For example, *B* implies that *A* is false, but *A* does not imply that.

Therefore, clause (i) is not satisfied, and so *A* is not closer to the truth than *B*. Similarly, *A* is not closer to the truth than *B*. More generally, it is possible to prove that no false theory is closer to the truth than any other false theory by Popper's definition.

Therefore Popper's definition cannot solve the problem of progress because the problem of progress requires that some false theories are closer to the truth than other false theories.

While this argument shows that there is a problem with the definition in this example, it is not a general proof. It is sufficient to see that the problem in this example is very general by examining a series of possible world diagrams.

Diagram 1: Each point inside the rectangle represents a possible world. There are 8 kinds of possible worlds, which can be described in terms a conjunction of *h*, *r*, and *w* or their negation. Any proposition expressible in the language is represented by the set of possible worlds in which the proposition is true. Note that the propositions, *h*, *r*, and *w* (called the *generators* of the language) are each true within four of the numbered regions.



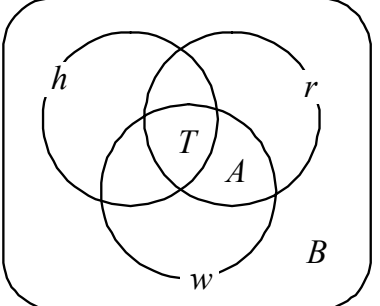
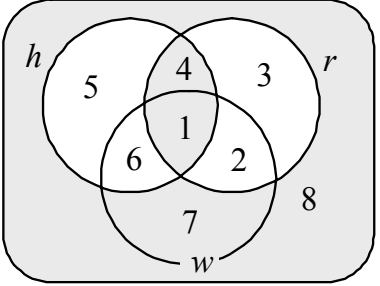
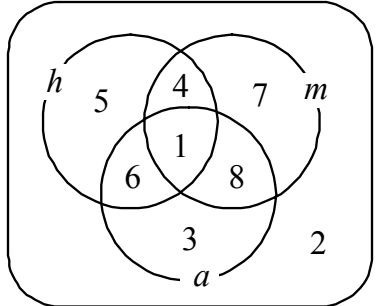
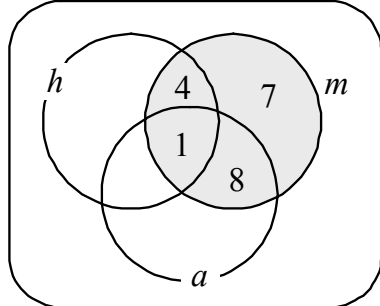
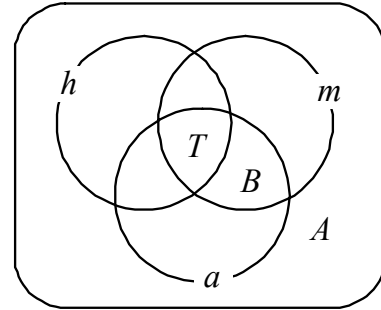
<p>Diagram 2: Any proposition is represented by the set of possible worlds at which it is true. Since T, A, and B are maximally strong propositions (called <i>state descriptions</i>), they are each represented by the single numbered regions. Logical entailment, is represented by the subset relation, because, by definition, $P \Rightarrow Q$ if and only if all P-worlds are Q-worlds. A appears to be closer to T than B is to T because A entails the truths r and w, while B does not.</p>	
<p>Diagram 3: Define the proposition m that the weather is Minnesotan by the shaded region. Notice that it is the combination of four numbered regions, 1, 4, 7, and 8. Each of h, r, and w is also represented by four numbered regions. Clearly, $T \Rightarrow m$ and $B \Rightarrow m$, although A does not entail m. In fact, $A \Rightarrow \sim m$. So, m is a truth that B entails that A does not. So, by Popper's definition, A cannot be closer to the truth than B, despite the contrary appearance given in Diagram 2.</p>	<p>The weather is Minnesotan</p> 
<p>Diagram 4: Suppose that we also define a (the weather is Arizonan) by the regions 1, 3, 6, and 8. If we reorganize the numbered regions according to whether they fall inside the areas h, m, and a, we are still able to express exactly the same propositions as before. That is, the set of generators $L' = \{h, m, a\}$ forms an alternative language, which is just as powerful as the original language (based on $L = \{h, r, w\}$). We can translate any statement in one language into a statement in the other.</p>	
<p>Diagram 5: In the new representation, based on $\{h, m, a\}$, the proposition m is expressed in a primitive way. But its meaning is the same as before, because it is true in exactly the same set of possible worlds. This is evident when you compare the numbers of the shaded region in this figure with the numbers of the shaded region in Diagram 3. The geographical arrangement of regions the diagram makes no difference in the meaning of proposition in the diagram.</p>	

Diagram 6: Likewise, the theories T , A , and B may be re-expressed in the new language, without there being any meaning change. Notice that the representation looks different, since B is now next to T . Compare this to Diagram 2, where A was closer to T than B was to T . This visual difference has cash value in the new language, in the precise sense that B now entails the true propositions a and m , whereas A entails neither of them.



The earlier proof that neither A nor B is closer to T according to Popper’s definition involved using the theories themselves, or their negations, as logical consequences of the theories. We may avoid this by using h , r , w , a , and m , as our set of consequences. All of these propositions are entailed by T , and are therefore true. A entails the set $\{r, w\}$, while B entails the set $\{a, m\}$. Therefore, neither is closer to T than the other because neither set of true consequences is a subset of the other. This fact cannot be changed by enlarging the set of consequences. Therefore clause (i) in Popper’s definition cannot be satisfied. A similar argument shows that clause (ii) in Popper’s definition cannot be satisfied.

We have shown that neither A nor B is closer to T according to Popper’s definition. This is not damaging by itself because Popper recognized that some false theories may be equally distant from the truth. The damage is done by the fact that the same is true of *any* two false theories. While we have not proven that fact in its fullest generality, the possible world diagrams allow us to see that in any case in which T , A , and B , are mutually exclusive, and therefore represented by disjoint regions in a possible worlds diagram, it is possible to find a true proposition that is entailed by B , but not by A ; namely the disjunction of T with B .

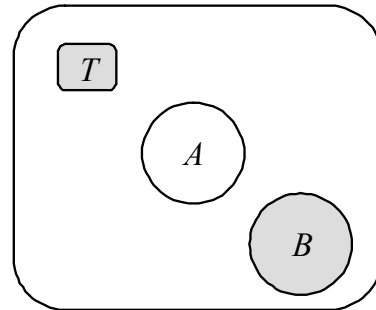


Figure 1: The shaded proposition is entailed by T and B , but not by A .

Recall that the intuitive idea behind Popper’s definition is that A is closer to the truth than B if A has more truths and fewer falsehoods in its set of deductive consequences than B . So, why not think of this statement purely in terms of the *number* of true and false consequences of A and B rather than requiring that A *includes* all the true consequences of B ? That is, why don’t we allow that A and B quite different sets of consequences? Unfortunately, this does not solve the problem, as is clearly shown by our possible world diagrams. For in the weather example, A and B have *exactly the same numbers of true consequences, and exactly the same number of false consequences*. Harris (1974) generalized the result to show that the idea can not work.

The second fact we learn from the possible world diagrams is that we can express any fact about the weather entirely in terms of the weather predicates $\{h, m, a\}$. To show this, it is sufficient to be able to defined r and w in terms of $\{h, m, a\}$. The following equivalencies establish the translatability from $\{h, r, w\}$ to $\{h, m, a\}$:

$$r \text{ if and only if either } (h \ \& \ m) \text{ or } (\sim h \ \& \ \sim m)$$

w if and only if either $(h \ \& \ a)$ or $(\sim h \ \& \ \sim a)$

In words, the weather is *rainy* if and only if it is either hot and Minnesotan or cold and not Minnesotan, and the weather is *windy* if and only if it is either hot and Arizonan or cold and not Arizonan.

	Language L	Language L' :
T	$h \ \& \ r \ \& \ w$	$h \ \& \ m \ \& \ a$
A	$\sim h \ \& \ r \ \& \ w$	$\sim h \ \& \ \sim m \ \& \ \sim a$
B	$\sim h \ \& \ \sim r \ \& \ \sim w$	$\sim h \ \& \ m \ \& \ a$

The translation in the other direction is given by:

m if and only if either $(h \ \& \ r)$ or $(\sim h \ \& \ \sim r)$
 a if and only if either $(h \ \& \ w)$ or $(\sim h \ \& \ \sim w)$

The translations are easier to remember if we introduce \leftrightarrow to stand for ‘if and only if’ in the languages L and L' . Then, $m = (h \leftrightarrow r)$, $a = (h \leftrightarrow w)$, $r = (h \leftrightarrow m)$, and $w = (h \leftrightarrow a)$. The translations of the three propositions, T , A , and B are summarized in the table below.

The symmetry between these alternative languages is important to Miller’s objection to Tichý’s modification of Popper’s definition.

Tichý’s Definition of Verisimilitude

Tichý (1974) presented a way of defining the verisimilitude that is immune from the objection to Popper’s definition. For our purposes, his idea is adequately illustrated in terms of the weather example. Tichý says that A is closer to the truth than B because A makes one mistake but gets two things right, while B is wrong on all three counts. More exactly, A gets h wrong and r and w right, whereas B gets all three wrong. One mistake is better than three mistakes, so A has greater verisimilitude or truthlikeness. Now it is possible for one false theory to be closer to the truth than another. The idea is to restrict the class of consequences to the set generators of the language, $\{h, r, w\}$, and their negations. From this we can anticipate Miller’s objection, because $\{h, m, a\}$ is an alternative set of generators.

Miller’s Problem of Language Variance

The objection to Tichý’s definition is philosophical rather than technical. In particular, the restriction to the proposition set $\{h, r, w\}$ seems quite natural given the way we talk about the weather. But if we considered the alternative set $\{h, m, a\}$, as defined above, then B is closer to T than A is to T , which reversed the verisimilitude ordering. The results are summarized in the table above. In summary, Popper’s definition has no problem of language variance because it refers to all consequences, but that is why it doesn’t work. But when one tries to restrict the class of consequences, there seems to be no principled way of doing this.

Truth is language invariant. That is, Newton’s theory is true when expressed in English if and only if it is also true when expressed in French. We should also expect that closeness to the truth is also language invariant. In fact, it language variance would introduce a subjective element into the concept that would render it unsuitable for solving

the problem of progress. But it appears that it depends on whether we take $\{h, r, w\}$ or $\{h, m, a\}$ as our primitive set of predicates (as Miller 1974 pointed out). Therefore Tichý's definition of verisimilitude does not solve the problem of progress, for it does not provide us with an *objective* measure of progress towards the truth.

	Verisimilitude in language L :				Verisimilitude in language L' :			
	h	r	w	Total	h	m	a	Total
A	0	1	1	2	0	0	0	0
B	0	0	0	0	0	1	1	2

Some readers may be unhappy with the view that Miller's argument has much force at all. For it may appear to them that there is a distinct asymmetry between the two languages. For isn't it true that the Arizonan and Minnesotan predicates have to be defined in terms of the standard predicates, hot, rainy, and windy? The standard reply to point out that the two languages are perfectly symmetric. That is, if we begin with the alternative language, then it is the standard weather predicates, r and w , that are defined in a gerrymandered way that is parasitic upon the predicates m and a . This worry is addressed by the argument from chauvinism in a later section.

The problem with Tichý's definition is that relative to the alternative language L' , B is closer to T than A is to T , which reverses the verisimilitude ordering (see the table above). In sum, Popper's definition avoids problem of language variance because it refers to all consequences, but it doesn't work. Unfortunately, when one tries to solve the problem by restricting the class of consequences used to define verisimilitude, the result is language dependent. There are several ways of addressing the problem within a strictly non-probabilistic framework.

Three Responses to Miller's Problem

There are three common responses to Miller's problem:

1. Abandon the thesis that one false theory can be closer to the truth than another. Give up on any sense of verisimilitude, and therefore give up on an objective notion of progress.

For example, Urbach states:

I shall argue, moreover, that the attempt to make sense of an objective notion of degrees of closeness to the truth or false theories is fundamentally and irretrievably misguided. (Urbach 1983 p. 267)

Also, Barnes (1991, p. 309) states: "truthlikeness" cannot supply a basis for an objective account of scientific progress.'

2. Embrace the idea that verisimilitude ordering depends on our language, and thereby embrace a form of subjectivism. It embraces the notion of verisimilitude at the expense of providing no objective sense of progress.
3. Explain why we should regard one particular representation as "privileged." That is, insist the set of predicates $\{\text{hot, rainy, windy}\}$ is the objectively fundamental because the properties of being hot, being rainy, and being windy are the properties that really exist in nature. The weather properties of being Minnesotan and being Arizonan are concocted, and artificial, because they do not capture the true "essence" of the world. Tichý and many who have defended variations of Tichý's concept of verisimilitude have opted for the third kind of response. Indeed it could lead to a solution to the problem of

progress, because it does avoid the problem of language variance, and defines verisimilitude in an objective way by appealing to “the way the world really is”. Yet, it has come under attack. This third option, according to Miller, must appeal to the fallacious and outmoded doctrine of *essentialism*, which is the doctrine that says that one set of predicates role in capturing the *essential* or fundamental properties of the world.² Not only is it controversial to assume that any particular set of predicates has such a privileged ontological status, but it also raises a problem that we would have to *know* what this status is before we could measure verisimilitude. That is, a solution to part (B) of the problem of progress—the task of saying whether science has actually made progress towards the truth—is harder to solve.

The best way of formulating the objection to solution (3) is by granting, for the sake of argument, that essentialism is true. The problem is that essentialism makes no observable difference to any consequence of any theory. If it has no testable consequences, but makes a crucial difference to verisimilitude orderings, then verisimilitude becomes epistemologically inaccessible to us. So, there is a sense in which solution (3) is like solution (2). Both fail to solve the problem of progress.

There is a more impressive way of arguing against solution (3): The claim of privilege of one language over another, made for any reason, introduces an undesirable kind of chauvinism according to which some apparently legitimate claims of progress must be denied. I call this the argument from cultural chauvinism.

The Argument from Cultural Chauvinism

Suppose we come across an alien culture living in a valley where they grow two kinds of corn: Minnesotan corn and Arizonan corn. Minnesotan corn grows in the lower part of the valley, while Arizonan corn grows in the upper part of the valley. Their daily decisions do not concern whether to wear warm clothes, take an umbrella, or wear a windbreaker. Their decisions are more important than that. Each day they need to know whether they need to tend their Minnesotan corn or their Arizonan corn (or both). They cannot afford the time and energy to travel to the lower valley or the high valley when it is unnecessary. They need to tend to the Minnesotan corn if and only if the weather is Minnesotan, while they need to tend to the Arizonan corn if and only if the weather is Arizonan. Fortunately, their science, though not perfect, has progressed considerably in recent decades. In times when their science was less advanced, they believed theory A, which was unsuccessful at predicting whether it was hot, whether it is Minnesotan, and whether it is Arizonan. Now theory A has been superseded by theory B, which is still bad at predicting temperature, but is now successful at predicting the two most important facets of the weather.

Should we say that their science is not progressive? Why should we deny that their achievement is genuine and objective simply because the questions that concerned them happen not to be about the “true objective essences” of the world? Wouldn’t that be an unsavory kind of cultural chauvinism? Why should scientific progress depend on the ‘true essences’ of the world (assuming that essentialism is not an outmoded doctrine) if essences make no difference to the success of the theory.

² For replies to Miller, see Oddie (1986) and Niiniluoto (1987) 13.2.

The argument does not assume that facts about essences can *never* make a difference to the truth and to the verisimilitude of theories. If a scientific theory made an explicit claim about which properties are fundamental, and was wrong, then it should receive a negative score for that false consequence. That is not at issue. The question is whether verisimilitude scores should depend on facts about the world of which rival theories say nothing. Such is the case in the weather theories *A* and *B*. They make no claims about which set of predicates is primitive. That is why theories are so straightforwardly translated from one language to the other, and back again.

It may appear that this story is supporting a kind of subjectivism, as in the response (2), which says that verisimilitude is relative to the language in which the theories are expressed. However, that is not the lesson of our story. In fact, quite the opposite is true. To *avoid* an objectionable kind of chauvinism, each culture should *acknowledge* the scientific achievements of the other. The conclusion recommended in response (2) is that *relative to the standard language*, the other scientific community in our story have *not* progressed. They, in turn, say the same about our weather science. The result is a failure of communication exactly because the verisimilitude ordering is not preserved when the same theories are translated into the other language. That is why subjectivism is objectionable.

Nor is it appropriate to deny that the languages are inter-translatable. To say that *A* is not logically equivalent to *A'* or *B* is not logically equivalent to *B'* would be another way of displaying the kind of chauvinism we wish to avoid. It seems to us that there is only one response to Miller's problem that achieves this result.

A Fourth Solution

A fourth solution is to insist that there is no verisimilitude relationship between a theory and the truth that can be measured on a single number scale. *Aspects* of verisimilitude are relative to a particular set of questions and there is some formula for weighing the relative importance of those questions.³ There is a truthlikeness relation, but it is not one-dimensional as solution (3) assumes.

Once upon a time, psychologists thought that there was an absolute measure of human intelligence that one could measure on a linear scale, which became known as the Intelligence Quotient, or IQ for short. The hope was a high IQ score would correctly predict high math skills, high language skill, and high analytical skills simultaneously. Unfortunately, the fact is that these aptitudes do not go together. People with high math skills can have poor language skills, and vice versa. People with poor language skills can have extremely good analytical skills, and so on. So, we now measure intelligence in a multi-dimensional way by recording independent scores for these aptitudes (and each of these is an imperfect measure as well).

According to the fourth solution, measuring verisimilitude is like measuring intelligence. There is no *a priori* assumption that the ordering on one set of test questions has to be same as on another set of test questions. However, it's important not to overstate the view in either case. Just as the multi-faceted view of intelligence *allows* for a degree of correlation between the different question sets, so too this view of verisimilitude allows that a theory that scores high on one set of questions will tend to

³ See Zwart (1998) for a discussion of issues that appear to anticipate many of the ideas behind solution (4).

score high relative to another relevantly similar set of questions. What counts as ‘relevantly similar’ depends on the theory under consideration, and what is true. When there is a correlation, it needs explaining, but only *to the extent* that it is true.

In the weather example, *A* and *B* get at least two scores: One with respect to Question Set 1, {Is it hot? Is it rainy? Is it windy?}, and the other with respect to Question Set 2, {Is it hot? Is the weather Arizonan? Is the weather Minnesotan?}. Both scores count. The example is an unusual case in that the orderings are opposite in each case. Nevertheless, both verisimilitude orderings are equally objective measures of progress, albeit measures of different kinds aspects of progress. So, if there did exist a culture whose science aimed at predicting whether the weather is Minnesotan or Arizonan, then we can say that *their* science has progressed along one dimension of verisimilitude and degenerated along another dimension.

According to the fourth solution, there is no contradiction between the two verisimilitude claims, and there is not problem of language variance. Nor is there any appeal to any privileged ontological status of one language over another.

In case this needs spelling out, here is the argument that solution (4) is language invariant: Consider an arbitrary theory *A* and a question about whether *x* is true. We want to show that if we translate from one language to another, by translating *both* the theory and the question, then the verisimilitude score of the theory relative the question does not change. Either *A* entails *x*, or it does not. When we translate both *A* and *x* to *A'* and *x'*, then the same entailment will hold between *A'* and *x'*. Moreover, the truth value of *x* and *x'* are the same, since they are logically equivalent. Therefore, the scores received by *A* and *A'* with respect to this question are the same, since the scores are determined by whether *A* entails *x*, and the truth of *x*, and whether *A'* entails *x'*, and the truth of *x'*. If the scores for all questions in a set are language invariant, then the total scores are invariant, given that they are weighted the same way. An immediate corollary of this result is that solution (4) does not reduce to solution (2).

So, the presence of different languages is irrelevant to verisimilitude. It could be that a different scientific community from ours uses exactly the same weather language in pursuit of different questions, such as whether it is either hot and rainy or cold and dry. There is no introduction of the new weather predicates in this version of the story, and the result is the same: Accepting theory *B* in place of *A* would be progress towards the truth with respect to the research question *they* address. And that fact is an objective fact.

Nor does solution (4) reduce to solution (3), for there is no attempt to pick out one set of questions as privileged according to ‘the way that world really is’.

Nevertheless, one may interject that this point, and claim that this prevents solution (4) from providing an *objective* account of scientific progress, for every set of questions in which *A* is closer to the truth than *B*, there may be another set of questions for which the opposite is true. The rejoinder is to say that the existence of the second set of questions does not undermine the claim that there is objective progress with respect to the first set of questions. Scientific communities may only be concerned with certain *aspects* of verisimilitude. Their focus may be pragmatically motivated, but the verisimilitude score that motivates their scientific endeavors is not determined by them—it depends on what is true.

I believe that solution (4) is the best solution out of those discussed so far, a solution that avoided any kind of relativization of progress to a set of questions while also

avoiding the problems with solution (3) would be even better. It is therefore pleasing and surprising to see that there is an alternative. The alternative is to consider probabilistic hypotheses. It is understandable that Popperians have not been quick to explore this possibility because probabilistic hypotheses are not falsifiable in any strictly logical sense, even though some generalized notion of falsification might apply. This solution takes us further from Popper's original idea.⁴

A Fifth Solution: The Likelihood Approach

The discussion so far has focused on hypotheses that do not make probabilistic claims. But something interesting happens when we consider probabilistic hypotheses. Then the difference between the two languages is naturally associated with a difference in the assumptions made about the *probabilistic independence* of atomic propositions, *and this reverses the verisimilitude ordering* in a language invariant way. Here are the details.

Consider the probabilistic hypothesis P_A defined by: $P_A(\sim h) = 1 - \varepsilon$, $P_A(r) = 1 - \varepsilon$, $P_A(w) = 1 - \varepsilon$, for some $\varepsilon < \frac{1}{2}$, where the probabilities of h , r , and w are probabilistically independent (according to A). To say that events are probabilistically independent is to imply that the probability of a joint event is calculated by multiplying to probability all the events together. Thus, probability of the joint event of three different coins landing heads is the product of each one of them landing heads; that is, $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}$.

The *likelihood* of a hypothesis relative to some state of affairs, or event, is by definition, the probability of that event given the hypothesis. The likelihood of a hypothesis relative to some fact is a measure of how well the hypothesis fits the facts. This should not be confused with the probability of the hypothesis given the fact. For example, the probability of someone using hard drugs given that they drink alcohol is small, but the probability that someone drinks alcohol given that they use hard drugs is high.

The truth is $T = h \& r \& w$. So the likelihood of P_A relative to the truth is $P_A(T) = P_A(h)P_A(r)P_A(w) = \varepsilon(1 - \varepsilon)^2$. Compare this with the likelihood of P_B defined by: $P_B(\sim h) = 1 - \varepsilon$, $P_B(\sim r) = 1 - \varepsilon$, $P_B(\sim w) = 1 - \varepsilon$, where these probabilities are again independent according to the hypothesis. Then $P_B(T) = P_B(h)P_B(r)P_B(w) = \varepsilon^3$. P_A has a greater likelihood than P_B because P_A gives two out of three events its higher probability, while P_B gives all of them its lowest probability. The idea is the same as Tichý's. The crucial difference is that the likelihood measure is language invariant. The reason is that a probabilistic hypothesis is defined by the set of probabilities that it assigns to the eight maximally strong propositions corresponding to the eight regions in the possible worlds diagram (see Fig. 2 below). Call these *state descriptions*. Given that probabilistic hypotheses are individuated in this way, it does not matter how the state descriptions are expressed.

The similarity with Tichý's definition is especially clearer when we replace likelihoods with log-likelihoods. This replacement is allowable because for any

⁴ Nevertheless, the idea was mentioned as appropriate way of defining verisimilitude for probabilistic hypotheses at the end of chapter 12 in Niiniluoto 1987.

hypotheses, A has greater likelihood than B if and only if A has greater log-likelihood than B . The replacement is useful because the logarithm transforms a product into a sum. Thus, $\log P_A(T) = \log \varepsilon + 2 \log(1 - \varepsilon)$ and $\log P_B(T) = 3 \log \varepsilon$. $\log(1 - \varepsilon)$ is the score for getting the answer to one of the questions right, while $\log \varepsilon$ is the score for getting a question wrong, and $\log(1 - \varepsilon) > \log \varepsilon$ if $\varepsilon < \frac{1}{2}$. The scores are added because the *hypotheses* assert that the events are probabilistically independent. Thus, log-likelihoods follow Tichý's idea because they yield the same verisimilitude ordering as in the non-probabilistic case.

It is true that P_A and P_B are more *naturally* expressed in terms of the language $L = \{h, r, w\}$ rather than $L' = \{h, m, a\}$. However, this is not essential, since P_A and P_B are also well defined in terms of L' . Nevertheless, there is a sense in which the language L is taken to be privileged. But this status is not built into the definition of verisimilitude itself—this is defined for any hypothesis in a language invariant way. Rather, the privileged status of h , r , and w is built into the *hypothesis* via the independent assumptions that it uses. It is asserted by the hypothesis, rather than assumed as a metaphysical given.

Moreover, if the truth itself were probabilistic, and we were able to observe many repeated instances, then there would be an objective way of testing whether the assumption of probabilistic independence were correct or not. Assumptions about probabilistic independence do have testable consequences.

It is instructive to calculate the log-likelihood of hypotheses constructed in the language L' . Consider the probabilistic hypothesis, $P_{A'}$, defined by: $P_{A'}(\sim h) = 1 - \varepsilon$, $P_{A'}(\sim m) = 1 - \varepsilon$, $P_{A'}(\sim a) = 1 - \varepsilon$, where $\varepsilon < \frac{1}{2}$ and the probabilities of h , m , and a are postulated to be independent. The truth is $T = h \& m \& a$. Then the log-likelihood of $P_{A'}$ relative to the truth is $\log P_{A'}(T) = \log(P_{A'}(h)P_{A'}(m)P_{A'}(a)) = 3 \log \varepsilon$. Compare this with the likelihood of $P_{B'}$, defined by: $P_{B'}(\sim h) = 1 - \varepsilon$, $P_{B'}(m) = 1 - \varepsilon$, $P_{B'}(a) = 1 - \varepsilon$, where the probabilities are independent. Then:

$$\log P_{B'}(T) = \log(P_{B'}(h)P_{B'}(m)P_{B'}(a)) = \log \varepsilon + 2 \log(1 - \varepsilon).$$

Therefore the verisimilitude orderings are reversed just as Tichý's definition dictates.

Notice that P_A and $P_{B'}$ have the same log-likelihoods, even though they are quite different hypotheses. What this means is a community of scientists that moves from $P_{A'}$ to $P_{B'}$ achieves the same amount of progress as our scientists achieve by moving from P_B to P_A . Both communities achieve a positive degree of progress (in seemingly opposite directions), and both their measures of verisimilitude are objective (both are defined in terms of the log-likelihood of the truth).

The philosophically important lesson is that *there is no problem of language variance* in the probabilistic case. The reason is that P_A and $P_{A'}$ are very different hypothesis despite the fact that they assign their highest probability to the same proposition, *viz.* A . Similarly for P_B and $P_{B'}$. The change in verisimilitude ordering is due to a difference in the extra-logical assumption of probabilistic independence.

In sum, the likelihood definition of verisimilitude follows Tichý's idea in spirit, but is immune to the usual objections to it. There is no language variance in any verisimilitude orderings, there is no appeal to an unknowable ontological status of predicates, and progress is one community of scientists who think in terms of language L is consistent with progress in another community of scientists immersed in L' . The probabilistic framework provides an attractive solution.

Note that we are presently assuming that the *truth* is *non-probabilistic*. If we were to assume that the truth is probabilistic, then the definition should be generalized to what is referred to as *predictive accuracy* in Forster and Sober (1994).⁵

Predictive Accuracy

The predictive accuracy of any hypotheses is defined as the average of all its possible log-likelihood scores, where the average is taken over all states that have non-zero probabilities according to the true hypothesis. The different cases are weighted according to the true probabilities of their occurrence. Let us denote the propositions corresponding to the eight regions in Fig. 1 s_1, s_2, \dots, s_8 . By definition, the predictive accuracy of any probability distribution P is equal to:

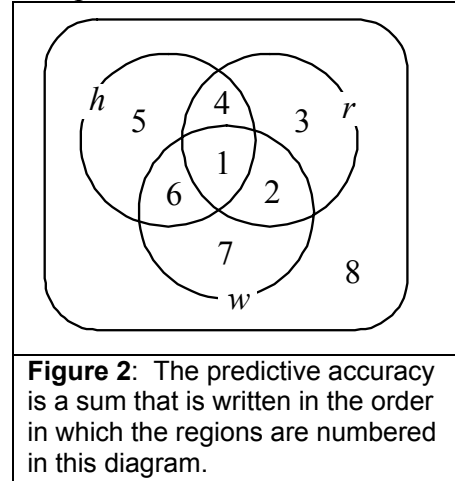
$$\sum_{i=1}^8 P_T(s_i) \log P(s_i),$$

where $P_T(s_i)$ is the true probability of s_i . Note that the predictive accuracy is language invariant because probabilities are language invariant. Also note that in the special case in which $P_T(s_i) = 1$, for some i , the definition reduces to the case discussed in the previous section.

The predictive accuracy of P is related to what is called the Kullback-Leibler discrepancy of P relative to P_T , defined by:

$$\Delta_{KL}(P, P_T) \equiv \sum_{i=1}^8 P_T(s_i) \log P_T(s_i) - \sum_{i=1}^8 P_T(s_i) \log P(s_i).$$

Note that the first term is equal to the predictive accuracy of true hypothesis. While this term is unknown, it is also constant. The second term is the predictive accuracy of P . Kullback and Leibler (1951) prove that $\Delta_{KL}(P, P_T) \geq 0$ with equality if and only if $P(s_i) = P_T(s_i)$ for all i . That is, the true hypothesis has the greatest predictive accuracy possible.⁶



⁵ I've heard the allegation that predictive accuracy applies only when the truth probabilistic. This is not true. It is required is that the hypotheses are probabilistic, for otherwise likelihoods are not objectively defined.

⁶ Rosenkrantz (1980) has used the same measure to define verisimilitude, and Niiniluoto (1987) endorses Rosenkrantz's definition as a special case of his notion of verisimilitude.

The only unintuitive feature of the definition arises in the case when $P(s_i) = 0$ and $P_T(s_i) > 0$ for some i . For then $P_T(s_i) \log P(s_i) = -\infty$, which implies that $\Delta_{KL}(P, P_T) = -\infty$. That is, if a hypothesis says that an event has zero probability when it does not, then the hypothesis has the lowest predictive accuracy possible, no matter how well it does at predicting other events. A sufficient condition for avoiding this problem is that all hypotheses under consideration assign every event a non-zero probability. There are no such restriction on the true hypothesis.

(This is the same solution commonly adopted by Bayesians in order to avoid the following problem. Suppose that an event with zero probability is actually observed. How is this evidence used to update the prior probability distribution, when the conditional probability on that event is undetermined.⁷ In fact, the condition is imposed for the exactly same reason because a Bayesian posterior can be defined as the distribution that has the least Kullback-Leibler discrepancy with the prior distribution out of all those distributions that assign the observed event a probability equal to one (see Williams 1980).)

The Kullback-Leibler discrepancy is a well known measure of the closeness of two probability distributions. When the second distribution is the true distribution, it is equal to the difference between the predictive accuracy of the hypothesis in question and the predictive accuracy of the true hypothesis. Therefore, predictive accuracy can be thought of in two ways: Either as closeness to the true probability distribution, or as a measure of how well the hypothesis would, on average, fit the data generated by the true hypothesis.

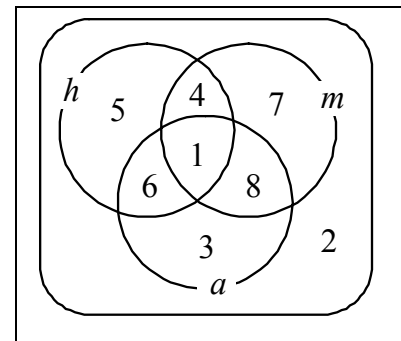


Figure 3: Notice the similarity between this diagram and that in Fig. 2. The key difference is the reversal in roles between regions 2 and 8.

Predictive Accuracy in the Weather Example

Consider the weather example again. Suppose that P_T is defined by:

$P_T(h) = P_T(r) = P_T(w) = 1 - \varepsilon^*$, where $\varepsilon^* < \frac{1}{2}$ and h, r , and w are probabilistically independent. According to the true hypothesis, the outcome $h \& r \& w$ is the most probable. However, because the events h, m , and a are not probabilistically independent, we expect that the true hypothesis will give the hypotheses P_A and P_B some advantage over $P_{A'}$ and $P_{B'}$. Now consider the predictive accuracies of the hypotheses $P_A, P_B, P_{A'}$ and $P_{B'}$, where these are the same as defined in a previous section. After some tedious calculations, the results are:

$$PredAcc(P_A) = (1 + \varepsilon^*) \log \varepsilon + (2 - \varepsilon^*) \log(1 - \varepsilon).$$

⁷ Note that I say ‘undetermined’ rather than ‘undefined’. Some Bayesians respond to the problem by saying that the probability conditional on an event of probability zero is well defined under the Popper axioms of probability (Popper 1959). But merely appealing to Popper’s axioms does not determine the value of the conditional probability. So this, by itself, does not solve the problem.

$$\text{PredAcc}(P_B) = (3 - 3\varepsilon^*) \log \varepsilon + 3\varepsilon^* \log(1 - \varepsilon).$$

$$\text{PredAcc}(P_{A'}) = (3 - 5\varepsilon^* + 4\varepsilon^{*2}) \log \varepsilon + (5\varepsilon^* - 4\varepsilon^{*2}) \log(1 - \varepsilon).$$

$$\text{PredAcc}(P_B) = (1 + 2\varepsilon^* - 2\varepsilon^{*2} - \varepsilon^{*3}) \log \varepsilon + (2 - 2\varepsilon^* + 2\varepsilon^{*2} + \varepsilon^{*3}) \log(1 - \varepsilon).$$

In the special case in which $\varepsilon^* = 0$, each expression reduces to the non-probabilistic log-likelihood values.

The results are intuitively pleasing. Our scientists have made progress because P_A is closer to the truth P_B . At the same time, other scientists have also made progress because P_B is closer to the truth than $P_{A'}$. However, there is a sense in which the alternative community is using the wrong predicates because they incorrectly assume that h , m , and a are probabilistically independent. Because of this mistake, P_A is closer to the truth than $P_{B'}$. That is, the other scientists pay a small price for making a false assumption, *but their progress is positive despite this mistake*.

Some people may find it surprising that P_B is closer to the truth than $P_{A'}$. After all, the other scientists are assigning the proposition B their highest probability, $(1 - \varepsilon)^3$, to the event that has the lowest true probability. However, the key point is that this mistake is not heavily penalized exactly because the true probability is low—the mistake doesn't arise very often, so it's not weighted very heavily. What counts more is the log-likelihood of events that occur more frequently. In particular, the most important term is the term associated with the most frequently occurring event $h \& r \& w$. That is why P_B does better than $P_{A'}$, as is especially clear in the non-probabilistic case, where that event is the only one that occurs.

The philosophical consequences of this solution are subtle. The fact that the events $\{h, r, w\}$ are independent provides privileged ontological status to those events. However, this privilege is not based on any special intrinsic feature of the properties themselves. Probabilistic independence is a relational property, which has observable consequences.

The privileged status of language L does not prevent scientists from working successfully in language L' in such a way that the direction of progress is as one would intuitively expect. The falsity of the independence assumption slows their progress without reversing its direction! Having the 'right' concepts is important, but that science can still progress with 'false' concepts. It is possible to bridge the gap between no science and perfect science. That is what verisimilitude should reflect. For our current theories are not only false, but the concepts they use also fail to carve nature at its joints.

Miller's "Accuracy of Predictions"

A year after publishing his arguments against the definition of Popper (1963) and Tichý (1974), Miller produces a "deeply skeptical" result to show that an "eminently plausible" suggestion for the accuracy of quantitative (numerical) theories is also useless. The proposal that Miller attacks postulates a *sufficient* condition for a verisimilitude ordering: If the numerical predictions of *A* are *uniformly* more accurate than those of *B*; that is, if *A*'s predictions are never further from the true values than the predictions of *B*, then *A* is closer to the truth than *B*.

	θ	ϕ	ψ	χ
<i>T</i>	0	2	2	8
<i>A</i>	7	2	9	15
<i>B</i>	8	0	8	8

Miller's challenge is very general, although it is most easily understood in terms of a simple numerical example. Let θ (*theta*) and ϕ (*phi*) be two physical magnitudes. Suppose that theory *A* asserts that $\theta = 7$ and $\phi = 2$, while *B* asserts that $\theta = 8$ and $\phi = 0$. Now suppose that the true value of θ is 0 and the true value of ϕ is 2. The prediction of *B* is therefore closer to the true values than the prediction of *A* for each magnitude. But consider instead new physical magnitudes ψ (*psi*) and χ (*chi*), defined by $\psi = \theta + \phi$ and $\chi = \theta + 4\phi$. The table shows predicted values for all four parameters.

In terms of the transformed quantities, theory *B* is *uniformly* closer to the truth than *A*. The verisimilitude ordering between *A* and *B* has been reversed. Miller has proved that the relationship of *univocal* accuracy in all quantities can be reversed by a simple change of language.

As in the weather example, there is a complete symmetry between the original and the transformed sets of magnitudes. The former define the latter and the latter can define the former by the transformations $\theta = (4\psi - \chi)/3$ and $\phi = (\psi - \chi)/3$. Adherents of the verisimilitude program either have to abandon the naive proposal that led to the problem, or somehow embrace the representation-dependence that Miller's example illustrates. In either case, the *theories A* and *B* must be *unambiguously* ordered, for the *theories* are the same theories in different representations.

The responses and replies to this problem are the same as those already described for the weather example. The best solution is to consider probabilistic hypotheses. Good (1975), in his response to Miller, anticipates this approach. He supposes that there is a *law of error* that assigns a definite probability, or probability density,⁵ to all possible values of the pair (θ, ϕ) . More specifically, Good supposes that this probability distribution is a Gaussian distribution with a peak value at $(\theta, \phi) = (0, 2)$, which means that the probability density drops off as we move away from the point corresponding to *T* according to a bell-shaped Gaussian curve. We may visualize the situation by imagining that the error distribution is represented by the height of the landscape above each point in the (θ, ϕ) -plane. The rate of "drop off" may occur more rapidly in some directions in

⁵ A probability density is assigned by a probability distribution to the possible values of a quantity that ranges over a continuum, as is the case for the quantities in Miller's example. The probability that a continuous quantity falling within a certain interval (region) of values is calculated as the area (or volume) under the distribution curve (or surface) within that interval (or region). Clearly, the probability that a continuous quantity takes on a particular point value is zero even if its probability density is not zero.

the (θ, ϕ) -plane than in others, so that the contour lines of equal probability density are ellipses. The geography of the landscape is depicted by the contour map (Fig. 4). The ellipses center on the point of maximum density (at the top of the hill).

Now suppose that part of what is true about the world is that observed values of (θ, ϕ) follow this error distribution. Good's idea is that since the probability density of points inside the ellipse is greater than those outside, we see may say that B is closer to T than A . This leads to the opposite verisimilitude ordering than the one described by Miller. This, therefore, undermines Miller's argument, because it motivates the idea that Miller's sufficient condition for verisimilitude ordering may be false.

More importantly, Good's criterion appears to be language invariant, because the point, marked B in Fig. 4, still lies inside the ellipse, and A continues to fall outside the ellipse no matter how we transform the coordinates.⁶ Since both A and B are false, this appears to be a counterexample to Miller's conclusion that no *false* theory can be more accurate than any other false theory.

In his rejoinder to Good, Miller (1975a, p.214) concedes that a law of error may provide a language-independent measure of the distance of any prediction from the true value *if* there are *probabilistic laws* governing the *errors of prediction*. But he denies the antecedent of the conditional. He cannot see "any discrimination of the sort envisaged as founded on any objective considerations." Miller also admits that Good's proposal is not entirely clear to him.

The Language Invariance of Predictive Accuracy

Miller is skeptical that there are *objective* probabilistic laws governing errors of prediction. The objectivity of probabilistic laws is a serious question, which I shall discuss in the section on curve fitting. In the meantime, it is worth making the point that predictive accuracy is perfectly well defined in the case in which the truth is non-probabilistic. What we want is something similar in spirit but different in substance from what Good proposes.

All that is required is that laws of error are attached to the hypotheses. When the error law is a part of the hypothesis, it is not subjective. It does not depend on how one describes the hypothesis. If the truth is non-probabilistic, then of course any probabilistic error law is false. But that is not a problem because we are meant to be defining the verisimilitude of false hypotheses. All that matters is that some postulated error laws may be closer to the truth than other error laws in some objective sense.

Just as in the weather example, log-likelihoods can be used to define the distance of a probabilistic hypothesis from the truth in an objective way. The idea is that there are two ellipses like the one drawn in Fig. 4, where one is centered at the point A and the other

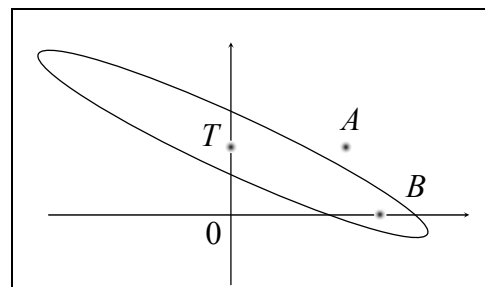


Figure 4: Good's (1975) response to Miller is that if there is a law of errors associated with T , then it may provide a closeness measure that is different from the naïve measure that Miller assumes.

⁶ I say "appears" because Good's proposal is not clearly language invariant. By the 'same' ellipse, Good means the ellipse that contains a certain proportion of the total probability inside the region it defines. But he provides no proof that points cannot move in and out of such a region upon arbitrary transformations. The language invariance of predictive accuracy, as I define it, will be proved later.

centered at the point B . If T lies within the ellipse centered at B but outside the ellipse centered at A , then B is closer to the truth than A .

The problem is different from the weather example because the quantities in this example take on a continuous range of values, rather than a finite set of discrete values. This means that likelihoods are defined in terms of probability *densities* rather than probabilities. Probabilities are language invariant, while probability densities are not. This can be seen in terms of a very simple example. Consider a random variable Φ that has a uniform probability distribution between 0 and 1. That is, all values within this range are equally probable, in some sense. The sense in which this is true needs to be carefully defined, for the probability that Φ has any particular point value is strictly speaking zero, and this is true no matter what the probability density is. What we need to consider is the *probability* that Φ lies between two values that lie within the interval $[0,1]$. Denote this interval by $[\theta, \theta + \Delta\theta]$, where θ now denotes an arbitrary value the random variable.⁸ The width of this interval is $\Delta\theta$. Then the probability that θ lies within an interval of this length is equal to $\Delta\theta$ times the probability density. Since the total probability must add up to 1, the probability density is equal to 1 everywhere within the interval $[0,1]$. Therefore, the probability that θ within the interval of length $\Delta\theta$ is equal to $\Delta\theta$ if the probability density is uniform.

Probability theory is constructed so that all probabilities are language invariant. So, introduce a new random variable Ψ defined by the equation $\Psi = 2\Phi$, which implies that arbitrary values obey the same equation: $\psi = 2\theta$. The probability distribution for Ψ is determined by the probability distribution for Φ because Ψ is defined in terms of Φ . Clearly, Ψ has a uniform distribution between 0 and 2, because when $\theta = 0$, $\psi = 0$, and when $\theta = 1$, $\psi = 2$. But since the total probability that Ψ lies in the interval $[0,2]$ must be 1, the probability *density* at all points ψ must be equal to $\frac{1}{2}$. Therefore, the probability density of ψ is different from the probability density of θ . Probability densities are not invariant under changes of coordinates.

In fact, we need this change to ensure that all probabilities are *unchanged*. In particular, the interval of θ values under consideration has length $\Delta\theta$. This corresponds to an interval of ψ values $[2\theta, 2(\theta + \Delta\theta)]$, which has length $2\Delta\theta$. The probability that ψ lies within an interval of length $2\Delta\theta$ is equal to this length times the probability density for ψ , which is $\frac{1}{2}$. So, the 2 cancels with the $\frac{1}{2}$ and the probability is again equal to $\Delta\theta$.

Now consider the arbitrary case in which the distribution for θ need not be uniform. Denote the probability density by $f(\theta)$. Now consider the probability that θ lies in the interval $[\theta, \theta + d\theta]$. I have written $d\theta$ in place of $\Delta\theta$ because we need to assume that $d\theta$ is very small. The probability that Φ is in the interval $[\theta, \theta + d\theta]$ is equal to $f(\theta)d\theta$. If we introduce a new random variable Ψ defined as an arbitrary one-to-one function of Φ , then the probability that Ψ is in the corresponding interval of points $[\psi, \psi + d\psi]$ is equal to $g(\psi)d\psi$, where $g(\psi)$ is the probability density of Ψ at ψ , such

⁸ Recall the a random variable is, by definition, any variable that has a probability distribution associated with it. The normal convention in statistics is that a random variable is denoted by an upper case letter, in this case Φ , while arbitrary values are denoted by the lower case letter, in this case θ .

that $f(\theta)d\theta = g(\psi)d\psi$. In general $d\theta \neq d\psi$, in which case the probability densities must change.

We are now ready to solve the problem for one variable. Suppose that the probability densities associated with hypotheses A and B are $f_A(\theta)$ and $f_B(\theta)$, respectively, and that true value of θ is 0. Then the log-likelihoods of the two probabilistic hypotheses are $\log[f_A(0)d\theta]$ and $\log[f_B(0)d\theta]$. Notice that these values depend on the value of $d\theta$, which is small, but arbitrary. But now recall that the log function has the property that the log of a product is the sum of the logs. Therefore the *difference* in the log-likelihoods is given by:

$$\log[f_A(0)d\theta] - \log[f_B(0)d\theta] = \log f_A(0) - \log f_B(0).$$

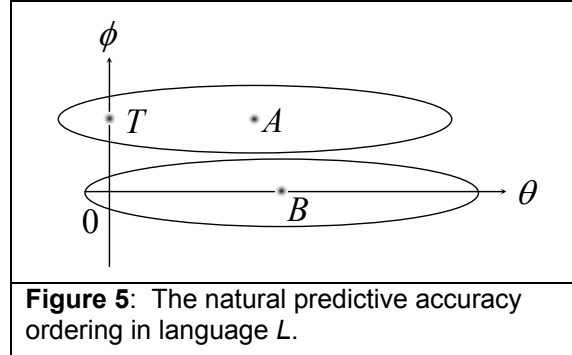
That is, the arbitrary factor $d\theta$ cancels out when we consider differences. Since the verisimilitude *comparison* of two false hypothesis is all that matters in the problem of progress, it is harmless to ignore the factor $d\theta$ and *define* the likelihoods in terms of the probability density alone.

Finally, we need to show that the differences in the log-likelihoods, so defined, is invariant under an arbitrary one-to-one transformation of coordinates. That is, we need to prove that

$$\log f_A(0) - \log f_B(0) = \log g_A(h(0)) - \log g_B(h(0)),$$

where $\psi \equiv h(\theta)$ for some one-to-one function h (which may be nonlinear). The proof is given in the Appendix, where the proof is also extended to the case of two variables.

The general proof of language invariance is algebraic in nature. So, it is worthwhile seeing how it works geometrically. The final part of this section will consider the special case in which the error distributions are independent Gaussian distributions. In this case, the ellipses of constant likelihood will transform in such a way that preserves the likelihood measure of distance. The first point is that the



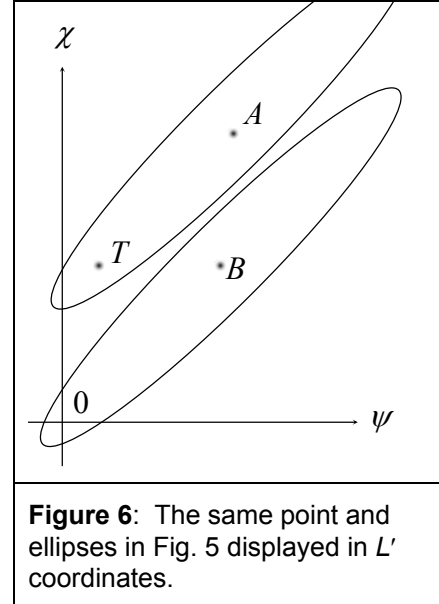
assumption that the error distributions for θ and ϕ are probabilistically independent means that the ellipses of constant likelihood (and therefore constant log-likelihood) have their axes of symmetry parallel to the axes in θ - ϕ plane. They have the same shape except that they are centered at the points A and B respectively. This is shown in Fig. 5.

In Fig. 6, the points T , A , and B and the points on the ellipses are represented in the ψ - χ coordinate system. The result is clearly the same. T is still inside the ellipse centered at A and outside the ellipse centered at B . The error distributions are no longer independent in the ψ - χ coordinate system.

At the same time, it is obvious that if a system of ellipses that does assume independence in the new coordinate system will be parallel to the new axes are centered at the same points, then the verisimilitude ordering will conform to the natural metric in the ψ - χ plane, and the verisimilitude ordering would be reversed. An algebraic proof of this feature is the topic of the following section.

The Importance of Probabilistic Independence

The example under consideration is completely analogous to the weather example. In this case, we have alternative languages, $L = \{\theta, \phi\}$ and $L' = \{\psi, \chi\}$, such that any claim formulated in L can be translated into L' and vice versa. The languages are equally powerful and the relationship between them is



symmetric. Nevertheless, the assumption of probabilistic independence that is naturally applied in each case makes a difference to the verisimilitude ordering. In the weather example, we found that the naturally formulated hypotheses followed a verisimilitude ordering identical to the intuitive ordering captured by Tichý's theory. Does the same thing happen in this case?

To show that it does, first consider the language L . For the sake of simplicity, suppose that the probabilistic hypotheses adopt the same error distributions for θ and ϕ except for the point at which they centered. That is, the error distribution assigned to θ by the probabilistic hypothesis p_A is $q(\theta - 7)$ while the distribution for the hypothesis p_B is $q(\theta - 8)$, where $q(\theta)$ is a probability density function for Φ . Similarly, let the error distribution assigned to ϕ by p_A be $r(\phi - 2)$ while that for p_B the density function is $r(\phi - 0)$. The next assumption is that $q(\theta)$ and $r(\phi)$ decrease monotonically from their central value. That is, the maximum values of $q(x)$ and $r(x)$ occur at $x = 0$, and decrease steadily on either side. A particular example of a density function with this property is the bell-shaped Gaussian distribution, but it is only one example. Nevertheless, it has the intuitively simple property that the log-likelihoods are proportional to the squared deviation from the central value.

The final assumption is that the joint probability distribution for θ and ϕ are probabilistically independent—that is, the joint distribution is formed by multiplying the densities of each variable.

Under these assumptions, the joint distributions that define the probabilistic hypotheses are:

$$p_A(\theta, \phi) = q(\theta - 7) r(\phi - 2),$$

$$p_B(\theta, \phi) = q(\theta - 8) r(\phi - 0).$$

It is now possible to prove that if verisimilitude is defined in terms of log-likelihoods, then the verisimilitude ordering conforms to the proposal very similar in spirit to the one

that Miller attacks: p_A is closer to the truth than p_B if the sum of log-likelihoods for all the *uniformly* more accurate than those of B ; that is, if A 's predictions are never further from the true values than the predictions of B . The proof is:

$$\begin{aligned}\log p_A(0,2) &= \log(q(0-7)) + \log(r(2-2)) \\ &> \log(q(0-8)) + \log(r(2-0)) \\ &= \log p_B(0,2).\end{aligned}$$

Therefore $\log p_A(0,2) > \log p_B(0,2)$. Clearly the proof depends only on the assumptions listed above, and so the result is quite general.

In particular, we can repeat the same construction in the language $L' = \{\psi, \chi\}$. If we refer to A' and B' as the points A and B in the ψ - χ plane, then we find that $\log p_{B'}(2,8) > \log p_{A'}(2,8)$. The verisimilitude ordering satisfies the necessary condition that Miller attacks without being affected by language variance, as the previous section proves. The reason is that the hypotheses, though each assigning their maximum probability density to same point (p_A assigns maximum probability to A , $p_{A'}$ assigns maximum probability density to A' , where A is the same point as A'), the density at other points (which is what matters for the verisimilitude of false hypotheses) is constructed in a way that depends on the language. There is no contradiction because $p_{A'}$ is not the same hypothesis as p_A even though A is the same point as A' . The analogy with the weather example is complete.

In the present example, the truth is non-probabilistic. Thus, there is a sense in which the independence assumptions made by the hypotheses are equally false, and none of the hypotheses are penalized for this anymore than the others. As in the weather example, this is not true if the truth is probabilistic.

Extension of the Definition to Probabilistic Truths

Miller was worried about the objectivity of any case in which the truth is described in probabilistic terms. Let us postpone the discussion of that worry until the next section. In this section, let us assume that the true hypothesis says that a number of states (expressible as pairs of values of (θ, ϕ)) can occur with a specific probability. Let us label these states as s_1, s_2, \dots, s_n , and denote their 'true' probabilities by P_1, P_2, \dots, P_n , respectively, where the probabilities must sum to one. Then the definition of predictive accuracy is extended to this case as the sum of the log-likelihood scores in each case weighted by the 'true' probabilities. That is, the predictive accuracy is the average log-likelihood, where the average is determined by the (unknown) true distribution. It is obvious from the previous section that difference in predictive accuracy of any two hypotheses is language invariant in the appropriate sense.

If the true distribution is continuous, rather than discrete, then the sums turn into integrals, and the language invariance of predictive accuracy is unaffected.

The Difference Between this Proposal and Good's Proposal

Good proposed that the error distribution whose likelihoods determine the verisimilitude ordering should be the one associated with the *true hypothesis*. This is a simpler proposal, but it is not the one I have adopted. The reason is very much to do with Miller's objection to Good's proposal: Miller was worried about the objectivity of such an error distribution. The worry is legitimate because Good requires that such an objective distribution must exist in every case, for otherwise there is no language invariant verisimilitude ordering. In contrast, I have shown that likelihoods define a language invariant ordering even when the truth is not probabilistic, so long as the hypotheses are (falsely) probabilistic.

It may be that this solution will still elicit a charge of subjectivity. To deflect such a charge, think of a non-probabilistic truth as a degenerate kind of probabilistic truth in which one state, say s_1 has probability 1, and all the others have probability 0. There can be no complaint that this description of the truth is subjective in any way. Then the general definition of predictive accuracy applies in this case, and measures, in effect, the Kullback-Leibler discrepancy between the truth distribution and the postulated error distributions. Both of the postulated distributions are false, yet one can be closer to the truth than the other. There is nothing subjective about this fact.

Furthermore, we might compare p_A above with a hypothesis that assign zero probability for every event except A , including zero probability to T . It seems to me quite intuitive to say that p_A is closer to the truth than this non-probabilistic alternative. This is a provable result in the present framework (the non-probabilistic hypothesis has a predictive accuracy of $-\infty$). The only unintuitive fact is that all such non-probabilistic hypotheses have a predictive accuracy of $-\infty$, and are therefore equally far from the truth. But why should we care about this 'defect' when we know how to construct hypotheses that do better than all of them. Isn't it better to move ahead in our investigation of the problem by focusing our attention on the class of probabilistic hypotheses? Isn't this exactly what *should* be done in a world in which uncertainty and error are a ubiquitous part of everyday life?

This advice seems especially prudent given that a large majority of models in the quantitative sciences are evaluated and compared according to statistical methods that make sense only with a probabilistic framework.

Appendix: Proof of Language Invariance

Let the probability densities associated with hypotheses A and B expressed in the new coordinate be $g_A(\psi)$ and $g_B(\psi)$, where the true value of ψ is now $h(0)$. Since $\psi \equiv h(\theta)$, $d\psi = |h'(\theta)|d\theta$, where $h'(\theta)$ is the derivative of h with respect to θ , and the absolute value enforces the convention that both $d\theta$ and $d\psi$ are positive. At the point $\theta=0$, $d\psi = |h'(0)|d\theta$. By the invariance of probabilities, we know that $f_A(0)d\theta = f_A(0)d\psi/|h'(0)|$, from which it follows that $g_A(h(0)) = f_A(0)/|h'(0)|$. We know that $h'(0) \neq 0$ because h is a one-to-one function. This ensures that $f_A(0)d\theta = g_A(h(0))d\psi$. The cancellation of $|h'(0)|$ is like the cancellation of the 2 in the elementary example of a uniform distribution. Clearly,

$$\log[f_A(0)d\theta] - \log[f_B(0)d\theta] = \log[g_A(h(0))d\psi] - \log[g_B(h(0))d\psi].$$

As a trivial consequence of this,

$$\log f_A(0) - \log f_B(0) = \log g_A(h(0)) - \log g_B(h(0)).$$

Now suppose that θ and ϕ have joint probability densities given by $f_A(\theta, \phi)$ and $f_B(\theta, \phi)$. Now assume that the new variables are defined by some one-to-one function h , such that $(\psi, \chi) \equiv h(\theta, \phi)$. Recall that the true value of (θ, ϕ) is $(0, 8)$. The complication in this case is that the rectangular area $d\theta d\phi$ is not only distorted in its length and height, but it is also transformed into a parallelogram in the ψ - χ plane. Without going into all the details (see e.g., Cramér 1946, section 22.2), all this is taken care of in the equation

$$d\theta d\phi = |J|d\psi d\chi,$$

where $|J|$ is the Jacobian of the inverse transformation at the point under consideration. The important point is that $|J|$ is a property of the transformation, and is therefore the same for both hypotheses. As expected, the invariance of the probabilities is expressed by the equation:

$$f_A(0, 8) d\theta d\phi = g_A(h(0, 8)) d\psi d\chi,$$

where

$$g_A(\psi, \chi) \equiv f_A(\theta, \phi)|J|.$$

The bottom line is therefore exactly the same as before: *Differences* in log-likelihoods are the same in each coordinate system:

$$\log f_A(0, 8) - \log f_B(0, 8) = \log g_A(h(0, 8)) - \log g_B(h(0, 8)).$$

Log-likelihoods are therefore language invariant in the required sense.

References

- Forster, Malcolm R. and Elliott Sober (1994): 'How to Tell when Simpler, More Unified, or Less *Ad Hoc* Theories will Provide More Accurate Predictions'. *British Journal for the Philosophy of Science* **45**: 1 - 35.
- Kullback, S. and R. A. Leibler (1951), "On Information and Sufficiency", *Annals of Mathematical Statistics* **22**: 79-86.
- Miller, David (1974): 'Popper's Qualitative Theory of Verisimilitude', *British Journal for Philosophy of Science* **25**, pp. 166-77.
- Miller, David (1975): 'The Accuracy of Predictions,' *Synthese* **30**, pp. 159-191.
- Miller, David (1975a): 'The Accuracy of Predictions: A Reply', *Synthese* **30**, pp. 207-219.
- Niiniluoto, Ilkka (1987): *Truthlikeness*, Kluwer Academic Publishing: Dordrecht.
- Niiniluoto, Ilkka (1998): 'Verisimilitude: The Third Period', *British Journal for the Philosophy of Science* **49**, pp. 1-29.
- Oddie, Graham (1986): *Likeness to the Truth*, The University of Western Ontario Series in Philosophy of Science, Kluwer Academic Publishing: Dordrecht.
- Oddie, Graham (1996): 'Truthlikeness', in Donald M. Borchert (ed.) *The Encyclopedia of Philosophy Supplement*, Macmillan Reference: New York, U. S. A.
- Popper, Karl (1959): *The Logic of Scientific Discovery*. London: Hutchinson.
- Popper, Karl (1963): *Conjectures and Refutations*, Routledge and Kegan Paul: London.
- Rosenkrantz, Roger D. (1980): "Measuring Truthlikeness," *Synthese* **45**: 463-487.
- Tichý, Pavel (1976): 'Verisimilitude Redefined', *British Journal for the Philosophy of Science* **27**, pp. 25-42.
- Tichý, Pavel (1974): 'On Popper's Definitions of Verisimilitude', *British Journal for the Philosophy of Science*. **25**, pp. 155-160.
- Williams, P. M. (1980): "Bayesian Conditionalization and the Principle of Minimum Information." *The British Journal for the Philosophy of Science* **31**: 131-144.