International Journal of Theoretical Physics, Vol. 33, No. 1, 1994

Maximum Predictive Power and the Superposition Principle

Johann Summhammer[1]

Received July 22, 1993

We are looking for a way of combining experimentally determined probabilities that should yield maximum predictive power. This concept is defined as permitting calculation of the accuracy of future experimental results solely on the basis of the number of runs whose data will serve as input for making the prediction. Each probability is transformed to an associated variable whose uncertainty interval depends only on the amount of data and strictly decreases with it. We find that for a probability which is a function of two other probabilities maximum predictive power is achieved when linearly summing their associated variables and transforming back to a probability. This recovers the quantum mechanical superposition principle.

1. INTRODUCTION

Quantum theory is as yet not understood to the same degree as are, for instance, classical mechanics or special relativity. Classical mechanics coincides well with our intuition and is therefore rarely questioned. Special relativity runs counter to our immediate insight, but can easily be derived by assuming constancy of the speed of light for every observer. And the assumption of a frame-independent maximal velocity may be made plausible by epistemological arguments (Whitrow, 1984). Quantum theory, on the other hand, demands that we accept two premises. First, it wants us to give up determinism for the sake of a probabilistic view. This, in fact, might be acceptable in a fundamental theory of prediction, because any communicable observation can be decomposed into a finite number of bits, and hence predictions are limited to finite accuracy. More disturbing is the second premise: Quantum theory wants us to give up the sum rule of probabilities by requiring interference instead. However, this sum rule is

[1]Atominstitut der Österreichischen Universitäten, A-1020, Vienna, Austria.

0020-7748/94/0100-0171$07.00/0

deeply ingrained in our thought, not so much because it works in everyday life, but because of its roots in counting and the definition of sets: Define sets with no common elements, then define the set which joins them all. The number of elements in this latter set is just the sum of the elements of the individual sets. When deriving the notion of probability from the relative frequency of events we are thus immediately led to the sum rule, such that any other rule appears inconceivable. And this may be the reason why we have difficulties accepting the quantum-theoretic rule, where probabilities are summed by calculating the square of the sum of the complex square roots of the probabilities. In this situation two views are possible. We may either consider the quantum-theoretic rule as a pecularity of nature. Or, we may conjecture that the quantum-theoretic rule has something to do with how we organize data from observations into quantities that are physically meaningful to us. We want to adopt the latter position. Therefore we seek to establish a grasp of the quantum-theoretic rule with the general idea in mind that, given the probabilistic paradigm, there may exist an optimal strategy of prediction, quite independent of traditional physical concepts, but resting on what one can deduce from a given amount of information. We will formulate elements of such a strategy with the aim of achieving maximum predictive power.

2. REPRESENTATION OF KNOWLEDGE FROM PROBABILISTIC DATA

Any investigative endeavor rests upon one natural assumption: More data from observations will lead to better knowledge of the situation at hand. Let us see whether this holds in quantum experiments. The data are relative frequencies of events. From these we deduce probabilities from which in turn we derive the magnitude of physical quantities. As an example take an experiment with two detectors, where a click is registered in either the one or the other. (We exclude simultaneous clicks for the moment.) Here, only one probability is measurable, e.g., the probability p_1 of a click in detector 1. After N runs we have n_1 counts in detector 1 and n_2 counts in detector 2, with $n_1 + n_2 = N$. The probability p_1 can thus be estimated as

$$p_1 = \frac{n_1}{N} \tag{1}$$

with the uncertainty interval[2]

$$\Delta p_1 = \left[\frac{p_1(1-p_1)}{N}\right]^{1/2} \tag{2}$$

[2]The uncertainty interval is derived from Chebyshev's inequality; see, e.g., Feller (1968, p. 233). For reasons of simplicity we are using only the approximate form valid for large N.

From p_1 the physical quantity χ, which we are actually interested in, is derived by

$$\chi = \chi(p_1) \tag{3}$$

$$\Delta\chi = \left|\frac{\partial\chi}{\partial p_1}\right| \Delta p_1 = \left|\frac{\partial\chi}{\partial p_1}\right| \left[\frac{p_1(1-p_1)}{N}\right]^{1/2} \tag{4}$$

The accuracy is given by the inverse of the uncertainty interval $\Delta\chi$. With the above assumption we expect it to increase with each additional run, because we get additional data. Therefore, for any N, we expect

$$\Delta\chi(N+1) < \Delta\chi(N) \tag{5}$$

However, this inequality cannot be true for an arbitrary functional relation $\chi(p_1)$. In general $\Delta\chi$ will fluctuate and only decrease *on the average* with increasing N. To see this take a theory A which relates physical quantity and probability by $\chi_A = p_1$. In an experiment of $N = 100$ runs and $n_1 = 90$ we get $\Delta\chi_A(100) = .030$. By taking into account the data from one additional run, where detector 2 happened to click, we have $\Delta\chi_A(101) = .031$. The difference may appear marginal, but nevertheless the accuracy of our estimate for χ_A has *decreased* although we incorporated additional data. So our original assumption does not hold. This is worrisome, as it implies that a prediction based on a measurement of χ_A may be more accurate if the data of the last run are not included. Let us contrast this to theory B, which connects physical quantity and probability by $\chi_B = p_1^6$. With N and n_1 as before we have $\Delta\chi_B(100) = .106$. Incorporation of the data from the additional run leads to $\Delta\chi_B(101) = .104$. Now we obviously do not question the value of the last run, as the accuracy of our estimate has increased.

The lesson to be learnt from the two examples is that the specific functional dependence of a physical quantity on the probability (or several probabilities if it is derived from a variety of experiments) determines whether our knowledge about the physical quantity will increase with additional experimental data, and that this also applies to the accuracy of our predictions. This raises the question of what quantities we should be interested in to make sure that we get to know them more accurately by doing more experiments. From a statistical point of view the answer is straightforward: choose variables whose uncertainty interval strictly decreases, and simply *define* them as physical. And from a physical point of view? Coming from classical physics we may have a problem, as concepts like mass, distance, angular momentum, energy, etc., are suggested as candidates of physical quantities. But when coming from the phenomenology of quantum physics, where all we ever get from nature is random clicks and count rates, a definition of physical quantities according to statistical

criteria may seem more reasonable, simply because there is no other guideline as to which random variables should be considered physical.

Pursuing this line of thought, we want to express experimental results by random variables whose uncertainty interval strictly decreases with more data. When using them in predictions, which are also expressed by variables with this property, predictions should automatically become more accurate with more data input. Now a few trials will show that there are many functions $\chi(p_1)$ whose uncertainty interval decreases with increasing N [equation 4]. We want to choose the one with maximum predictive power. The meaning of this term becomes clear when realizing that in general $\Delta\chi$ depends on N *and* on n_1 (via p_1). These two numbers have a very different status. The number of runs N is controlled by the experimenter, while the number of clicks n_1 is solely due to nature. Maximum predictive power then means to eliminate nature's influence on $\Delta\chi$. For then we can know $\Delta\chi$ *even before* having done any experimental runs, simply upon deciding how many we will do. From equation (4) we thus get

$$\sqrt{N}\,\Delta\chi = \left|\frac{\partial\chi}{\partial p_1}\right|[p_1(1-p_1)]^{1/2} = \text{const} \tag{6}$$

which results in

$$\chi = C\arcsin(2p_1 - 1) + D \tag{7}$$

where C and D are real constants. The inverse is

$$p_1 = \frac{1}{2}\left[1 + \sin\left(\frac{\chi - D}{C}\right)\right] \tag{8}$$

showing that the probability is periodic in χ. Aside from the linear transformations provided by C and D, any other smooth functions $\alpha(\chi)$ in real or complex spaces will also fulfill requirement (6) when equally sized intervals in χ correspond to equal line lengths along the curve $\alpha(\chi)$. One particular curve is

$$\alpha(\chi) = \sin\left(\frac{\chi}{2}\right)e^{i\chi/2} \tag{9}$$

which is a circle in the complex plane with center at $i/2$. It exhibits the property

$$p_1 = |\alpha|^2 \tag{10}$$

known from quantum theory. But note that, for instance, the function $\beta = \sin(\chi/2)$ does not fulfill the requirement that the accuracy only

depend on N. Therefore the complex phase factor in equation (9) is necessary.[3]

3. DISTINGUISHABILITY

We have now found a unique transformation from a probability to another class of variables exemplified by χ in equation (7). These unique variables always become better known with additional data. But can they be considered physical? We should first clarify what a physical variable is. A physical variable can assume different numerical values, where each value should not only imply a different physical situation, but should most of all lead to a different measurement result in a properly designed experiment. Within the probabilistic paradigm two measurement results are different when their uncertainty intervals do not overlap. This can be used to define a variable which counts the principally distinguishable results of the measurement of a probability. Comparison of that variable to our quantity χ should tell us how much χ must change from a given value before this can be noticed in an experiment. Following Wootters (1981) and Wheeler (1982), the variable θ counting the statistically distinguishable results at detector 1 in N runs of our above example is given by

$$\theta(n_1) = \int_0^{p_1(n_1)} \frac{dp}{\Delta p(p)} = \sqrt{N}\left[\arcsin(2p_1 - 1) + \frac{\pi}{2}\right]_{p_1 = n_1/N} \tag{11}$$

where Δp is defined as in equation (2). When dividing θ by $N^{1/2}$ it becomes identical to χ when in equation (7) we set $C = 1$ and $D = \pi/2$. This illuminates the meaning of χ: It is a continuous variable associated with a probability, with the particular property that anywhere in its domain an interval of fixed width corresponds to an equal number of measurement results distinguishable in a given number of runs. With Occam's dictum of not introducing more entities than are necessary for the description of the subject matter under investigation, χ would be *the* choice for representing physical situations and can rightly be called physical.

4. A SIMPLE PREDICTION: THE SUPERPOSITION PRINCIPLE

Now we return to our aim of finding a strategy for maximum predictive power. We want to see whether the unique class of variables represented by χ indicates a way beyond representing data and perhaps affords special predictions. For the sake of concreteness we think of the

[3]More details and variables that can be used to represent several probabilities can be found in Summhammer (1988). The statistical properties of χ are analyzed in Summhammer (1989).

double-slit experiment. A particle can reach the detector by two different routes. We measure the probability that it hits the detector via the left route p_L by blocking the right slit. In L runs we get n_L counts. In the measurement of the probability with only the right path available, p_R, we get n_R counts in R runs. From these data we want to make a prediction about the probability p_{tot} when both paths are open. Therefore we make the hypothesis that p_{tot} is a function of p_R and p_L. What can we say about the function $p_{\text{tot}}(p_L, p_R)$ when we demand maximum predictive power from it? This question is answered by reformulating the problem in terms of the associated variables χ_L, χ_R, and χ_{tot}, which we derive according to equation (7) by setting $C = 1$ and $D = \pi/2$. The function $\chi_{tot}(\chi_L, \chi_R)$ must be such that a prediction for χ_{tot} has an uncertainty interval $\delta\chi_{\text{tot}}$, which only depends on the number of runs L and R and decreases with both of them. (We use the symbol $\delta\chi_{\text{tot}}$ to indicate that it is not derived from a *measurement* of p_{tot}, but from other measurements from which we want to predict p_{tot}.) In this way we can predict the accuracy of χ_{tot} by only deciding the number of runs L and R. No actual measurements need to have been done. Because of

$$\delta\chi_{\text{tot}} = \left(\left| \frac{\partial\chi_{\text{tot}}}{\partial\chi_L} \right|^2 \frac{1}{L} + \left| \frac{\partial\chi_{\text{tot}}}{\partial\chi_R} \right|^2 \frac{1}{R} \right)^{1/2} \tag{12}$$

maximum predictive power is achieved when

$$\left| \frac{\partial\chi_{\text{tot}}}{\partial\chi_L} \right| = \text{const} \tag{13}$$

and

$$\left| \frac{\partial\chi_{\text{tot}}}{\partial\chi_R} \right| = \text{const} \tag{14}$$

We want to have a real function $\chi_{\text{tot}}(\chi_L, \chi_R)$, and therefore we get

$$\chi_{\text{tot}} = a\chi_L + b\chi_R + c \tag{15}$$

where a, b, and c are real constants. Furthermore, we must have $c = 0$ and the magnitude of both a and b equal to 1 when we wish to have χ_{tot} equivalent to χ_R or to χ_L when either the one or the other path is blocked. So there is an ambiguity of sign with a and b. When rewriting this in terms of the probability we get

$$p_{\text{tot}} = \sin^2\left(\frac{\chi_L \pm \chi_R}{2} \right) \tag{16}$$

This does not look like the sum rule of probability theory. Only for $p_L + p_R = 1$ does it coincide with it. We may therefore conclude that the

sum rule of probability theory does not afford maximum predictive power. But neither does equation (16) look like the quantum mechanical superposition principle. However, this should not be surprising, because our input was just two real-valued numbers χ_L and χ_R from which we demanded to derive another real-valued number. A general phase as is provided in quantum theory could thus not be incorporated. But let us see what we get with complex representatives of the associated variables of probabilities. We take $\alpha(\chi)$ from equation (9). Again we define in an equivalent manner α_L, α_R, and α_{tot}. From p_L we have, for instance [from (9) and (8) with $C = 1$ and $D = \pi/2$]

$$\alpha_L = p_L^{1/2}[p_L^{1/2} + i(1 - p_L)^{1/2}] \tag{17}$$

and

$$\Delta\alpha_L = \left|\frac{\partial\alpha_L}{\partial p_L}\right| \Delta p_L = \frac{1}{2\sqrt{L}} \tag{18}$$

If we postulate a relationship $\alpha_{\text{tot}}(\alpha_L, \alpha_R)$ according to maximum predictive power, we expect the predicted uncertainty interval $\delta\alpha_{\text{tot}}$ to be independent of α_L and α_R and to decrease with increasing number of runs L and R. Analogous to (13) and (14), we must have

$$\left|\frac{\partial\alpha_{\text{tot}}}{\partial\alpha_L}\right| = \text{const} \tag{19}$$

and

$$\left|\frac{\partial\alpha_{\text{tot}}}{\partial\alpha_R}\right| = \text{const} \tag{20}$$

yielding

$$\alpha_{\text{tot}} = s\alpha_L + t\alpha_R + u \tag{21}$$

where s, t, and u are complex constants. Now u must vanish and s and t must both be unimodular when p_{tot} is to be equivalent to either p_L or p_R when the one or the other route is blocked. We then obtain

$$p_{\text{tot}} = |\alpha_{\text{tot}}|^2 = |s\alpha_L + t\alpha_R|^2 = p_L + p_R + 2(p_L p_R)^{1/2} \cos\phi \tag{22}$$

where ϕ is an arbitrary phase factor containing the phases of s and t. This is exactly the quantum mechanical superposition principle. What is striking is that with a theory of maximum predictive power we can obtain the general form of this principle, but *cannot* at all predict p_{tot} even when we have measured p_L and p_R, because of the unknown phase ϕ. So there is room for physical laws.

5. CONCLUSION

We have tried to obtain insight into the quantum mechanical superposition principle and set out with the idea that it might follow from a most natural assumption of experimental science: that more data should on the one hand provide a more accurate representation of the matter under investigation and on the other hand afford more accurate predictions. From this we defined the concept of maximum predictive power. In essence this concept demands laws to be such that the uncertainty of a prediction is solely dependent on the number of experiments on which the prediction is based, and not on the specific outcomes of these experiments. Applying this to the observation of two probabilities and to possible predictions about a third probability therefrom, we arrived at a general form combining the probabilities. This turned out to be the same as the superposition principle. Our result may be an indication that the laws of nature are such that from more observations more accurate predictions must be derivable.

REFERENCES

Feller, William (1968). *An Introduction to Probability Theory and Its Applications*, 3rd ed., Wiley, New York.

Summhammer, J. (1988). *Foundations of Physics Letters*, **1**, 113.

Summhammer, J. (1989). *Physics Letters A*, **136**, 183.

Wheeler, J. A. (1982). *International Journal of Theoretical Physics*, **21**, 557.

Whitrow, G. J. (1984). *The Natural Philosophy of Time*, 2nd ed., Clarendon Press, Oxford.

Wootters, W. K. (1981). *Physical Review D*, **23**, 357.