The Language of Classical Physics

Hans Halvorson

We will mostly deal with the discrete case, which hardly ever comes up in classical physics. However, the mathematical complications that come in with infinities, taking integrals, etc., are really beside the point for what we want to discuss here. In particular, we want to see how the language of classical physics differs from the language of quantum physics.

1 Probability on finite sets

Definition. Let X be a finite set; the elements of X may be thought of as **pure states**, i.e. complete, classical configurations of some system or world. A **probability measure** is a map $p : X \to [0, 1]$ such that $\sum_{x \in X} p(x) = 1$.

Example. When the space X is finite — as we have assumed — there is one probability measure that seems special or preferred, viz. the flat distribution:

$$p_0(x) \;=\; \frac{1}{n},$$

where n is the number of elements in X. However, p_0 is by no means the only probability distribution on X. For example, for each point $x \in X$, there is a probability measure that is concentrated on x:

$$p_x(y) \;=\; \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{if } x \neq y. \end{cases}$$

If p and q are probability measures on X, and $\lambda \in [0, 1]$, then $\lambda p + (1 - \lambda)q$ is a probability measure on X, called a **convex combination** of p and q.

Definition. If a state p (i.e. a probability measure) can be written as a non-trivial convex combination of other states, then we say that p is a **mixed state**. Otherwise we say that p is a **pure state**.

The following result shows that the pure states of X are precisely the point masses, and hence stand in one-to-one correspondence with elements of X.

Proposition 1. *If $p(x) > 0$, then $p = \lambda p_x + (1-\lambda)q$ for some probability measure q and some $\lambda > 0$.*

Proof. If $p(x) = 1$, then $p = p_x$ and we're finished. If $p(x) < 1$, then we may define

$$q \;=\; (1-\lambda)^{-1}(p - \lambda p_x),$$

where $\lambda = p(x)$. Then

$$\begin{array}{rcl} \sum_{y\in X} q(y) & = & (1-p(x))^{-1} \sum_{y\in X} (p(y) - \lambda p_x(y)) \\ & = & (1-p(x))^{-1}(1-p(x)) \\ & = & 1. \end{array}$$

Hence q is a probability measure, and by definition $p = \lambda p_x + (1-\lambda)q$. □

Proposition 2. *A probability measure p on X is pure iff $p = p_x$ for some $x \in X$.*

Proof. Suppose that p is pure. Let $x \in X$ such that $p(x) > 0$. By the previous result, $p = \lambda p_x + (1-\lambda)q$. Since p is pure, $p = p_x$.

Now we show that p_x is pure. Suppose that $p_x = \lambda p + (1-\lambda)q$. If $y \neq x$ then

$$0 \;=\; p_x(y) \;=\; (1-\lambda)p(y) + \lambda q(y).$$

Hence $p(y) = 0 = q(y)$. Since y was arbitrary, it follows that $p = p_x = q$. □

Definition. An **event** or **proposition** E is defined to be a subset of X.

The events/propositions on X form a **Boolean algebra** with the operations $\wedge$ (intersection), $\vee$ (union), and $\neg$ (complement). Any event E also has an associated probability, denoted $p(E)$ and defined by

$$p(E) \;=\; \sum_{x\in E} p(x).$$

The map p from subsets of X to probabilities is called a **probability measure**, and it satisfies some obvious equations such as

$$p(E \vee F) \;=\; p(E) + p(F),$$

when E and F are disjoint.

Definition. Suppose that $p(E) > 0$. Then we define the **conditional probability** of F given E as

$$p(F|E) \;=\; \frac{p(F \wedge E)}{p(E)}.$$

In fact, $F \mapsto p_E(F) = p(F|E)$ is the probability measure generated by the function

$$p_E(x) \;=\; \begin{cases} p(E)^{-1}p(x) & x \in E \\ 0 & x \notin E. \end{cases}$$

itself a probability measure, and it's the most conservative choice of a new probability measure once one learns that E holds. Indeed, define a distance between probability measures on X as follows:

$$\|p - q\| \;=\; \sum_{x \in X} |p(x) - q(x)|.$$

Now let $M_E(X)$ be the set of all probability measures on X with the feature that $q(E) = 1$. Clearly $p_E \in M_E(X)$, and it can be shown that

$$\|p - p_E\| \;\leq\; \|p - q\|,$$

for all $q \in M_E(X)$. We will leave the details of a proof to the reader, but intuitively, p_E is the only measure on E that results from uniformly stretching values $p(x)$ for $x \in E$. If that goal is to minimize the distance from p, then no measure q can do better than a uniform stretch. If q were closer to p at some point $x \in E$, then q would have to be that much further away from p at some other point $y \in E$.

Definition. A **random variable** is a function $f : X \to \mathbb{R}$. We will sometimes call f a **quantity**, or for the sake of comparsion with quantum theory, an **observable**.

Example. If X is the classical configuration space $\mathbb{R}^3$, then the function $f(x_1, x_2, x_3) = x_1$ represents the quantity "first coordinate of position."

Let $\mathbb{R}^X$ be the set of random variables, i.e. functions from X to $\mathbb{R}$. This set $\mathbb{R}^X$ naturally forms an algebra where the operations are defined pointwise. That is, given f, g, we define

$$\begin{aligned} (f+g)(x) &= f(x) + g(x), \\ (fg)(x) &= f(x)g(x), \\ (rf)(x) &= rf(x). \end{aligned}$$

Clearly this algebra has a multiplicative identity (the constant 1 function), and is commutative, i.e. $fg = gf$.

Definition. The **spectrum** of f, $\mathrm{spec}(f) \subseteq \mathbb{R}$, is the image of X under f, i.e.

$$\mathrm{spec}(f) = \{f(x) \in \mathbb{R} : x \in X\}. \tag{1}$$

Proposition 3. *For a quantity f, the following are equivalent.*

1. $\mathrm{spec}(f) \in \{0, 1\}$.
2. *f is the characteristic function of some subset E of X.*
3. $f^2 = f$.

Proof. Suppose first that $\mathrm{spec}(f) \in \{0, 1\}$. If $E = \{x \in X \mid f(x) = 1\}$ then f is the characteristic function of E. It's also clear that a characteristic function f has the property that $f^2 = f$. Hence (1) $\Rightarrow$ (2) $\Rightarrow$ (3).

Now suppose that $f^2 = f$. Then for any $x \in X$, $f(x)^2 = f(x)$, which implies that $f(x) = 0$ or $f(x) = 1$. Therefore $\mathrm{spec}(f) \in \{0, 1\}$. □

There is a natural probability density on $\mathrm{spec}(f)$ denoted by p_f and defined (for $\lambda \in \mathrm{spec}(f)$) by

$$p_f(\lambda) = \sum_{x \in f^{-1}(\lambda)} p(x). \tag{2}$$

More generally, for any n random variables $f_1, \ldots, f_n$, there is a discrete probability density on $\mathrm{spec}(f_1) \times \cdots \times \mathrm{spec}(f_n)$ given by

$$p_{f_1,\ldots,f_n}(\lambda_1, \ldots, \lambda_n) = \sum_x p(x) \tag{3}$$

where $= \bigcap_{i=1}^n X_i \lambda_i$.

Definition. Given a random variable $f : X \to \mathbb{R}$, we define the **expectation value** of f as

$$p(f) \;=\; \sum_{x \in X} p(x) f(x). \tag{4}$$

Definition. If $E \subseteq X$, then the **characteristic function** of E is the function $e : X \to \{0, 1\}$ that assigns 1 to x iff $x \in E$.

It follows that $p(E) = p(e)$, where $p(E) = \sum_{x \in E} p(x)$, and $p(e) = \sum_{x \in X} e(x)p(x)$. Hence, we can freely interchange application of p to a subset and that subset's characteristic function.

Exercise 1. Show that expectation value is linear, i.e. $p(f+g) = p(f)+p(g)$, and $p(rf) = rp(f)$. Show that expectation value is positive, i.e. $p(f) \geq 0$ for any function f such that $\text{spec}(f) \subseteq \mathbb{R}^+$. Show that expectation value is normalized, i.e. $p(1) = 1$, where the first "1" is the constant function on X. Show that expectation value is not necessarily multiplicative, i.e. $p(fg) \neq p(f)p(g)$.

Example. If p_a is the measure concentrated on $a \in X$, then

$$p_a(f) \;=\; \sum_{x \in X} p_a(x) f(x) \;=\; f(a),$$

for all $f \in \mathbb{R}^X$. Conversely, if $p(f) = f(a)$ for all $f \in \mathbb{R}^X$, then $p(e) = 1$ where e is the characteristic function of $\{a\}$, and it follows that $p = p_a$.

Proposition 4. *A state p is pure iff $p(E) \in \{0, 1\}$ for all events E.*

Proof. Suppose first that p is pure. By Prop 2, $p = p_a$ for some $a \in X$. Hence

$$p(E) \;=\; p_a(E) \;=\; \sum_{x \in E} p_a(x) \;=\; \begin{cases} 0 & x \notin E, \\ 1 & x \in E. \end{cases}$$

Suppose now that $p(E) \in \{0, 1\}$ for all events E. In particular, $p(\{x\}) \in \{0, 1\}$ for each $x \in X$. Since

$$1 \;=\; p(X) \;=\; \sum_{x \in X} p(\{x\}),$$

it follows that $p(\{a\}) = 1$ for some $a \in X$, and hence $p = p_a$. Therefore, p is pure. □

Definition. Let f be a quantity and let p be a state. The **dispersion** $\sigma(p, f)$ of f in p is defined by

$$\sigma(p, f) \;=\; p(f^2) - p(f).$$

We say that p is **dispersion free** on f just in case $\sigma(p, f) = 0$.

If f is a projection, then $f^2 = f$, and hence $\sigma(p, f) = p(f)^2 - p(f)$. Therefore $p(f) \in \{0, 1\}$ iff $\sigma(p, f) = 0$.

We now look at the **spectral decomposition** of a function. For any subset $\lambda \in \text{spec}(f)$, let

$$e(\lambda) \;=\; \{x \in X \mid f(x) = \lambda\}.$$

Using the correspondence between subsets of X and characteristic functions, it's obviously true that

$$f \;=\; \lambda_1 e(\lambda_1) + \cdots + \lambda_n e(\lambda_n).$$

This fact is completely trivial in the case we are dealing with. But it will be important to remember the analogy when we derive an analogous result for quantum probability spaces.

Proposition 5. *A state p is pure iff $\sigma(p, f) = 0$ for all quantities f.*

Proof. Suppose that p is pure. By Prop 4, $p(g) \in \{0, 1\}$ for all g such that $g^2 = g$. In particular, $p(e(\lambda_i)) \in \{0, 1\}$, for any spectral projection $e(\lambda_i)$ of f. Hence,

$$p(f) \;=\; \sum_i \lambda_i p(e(\lambda_i)) \;=\; \lambda_j,$$

for some $\lambda_j \in \text{spec}(f)$. A similar calculation shows that $p(f^2) = \lambda_j^2$.

Now suppose that $\sigma(p, f) = 0$ for all quantities f. In particular, $\sigma(f^2) = \sigma(f)^2$ for any idempotent f, and hence $p(f) \in \{0, 1\}$. By Prop 4, p is pure. □

The following result shows the precise sense in which there are always **hidden variables** for classical systems, i.e. any state whatsoever can be interpreted as an ignorance mixture of determinate (i.e. dispersion-free) states.

Proposition 6 (unique decomposition)**.** *Every state p on X decomposes uniquely as a convex combination of dispersion-free states.*

Proof. If $X = \{x_1, \ldots, x_n\}$ and $\lambda_i = p(x_i)$, then $p = \sum_i \lambda_i p_{x_i}$. To see that the decomposition is unique: if $p = \lambda p_x + (1 - \lambda)q$ where $q(\{x\}) = 0$, then $p(x) = \lambda$. □

If we let $M(X)$ denote the convex set of all probability measures on X, then the previous result tells us that $M(X)$ is a **simplex**.

Proposition 7. *Let $q : \mathbb{R}^X \to \mathbb{R}$ be a positive linear functional such that $q(1) = 1$. Then there is a unique probability measure p on X such that $q(f) = \sum_{x \in X} p(x) f(x)$, for each $f \in \mathbb{R}^X$.*

Proof. Let $e_1, \ldots, e_n$ be characteristic functions of all singleton subsets of X. Since q is linear and normalized, we have

$$1 \;=\; q(e_1 + \cdots + e_n) \;=\; q(e_1) + \cdots + q(e_n).$$

Since q is positive, $q(e_i) \in [0, 1]$. Hence if we define $p(x) = q(\{x\})$, then p is a probability measure on X. Now let f be an arbitrary element of $\mathbb{R}^X$, and let $e(\lambda_1), \ldots, e(\lambda_m)$ be its spectral decomposition, which means that $f(x) = \lambda_i$ iff $e(\lambda_i)(x) = 1$. Clearly we have

$$\sum_{x \in e(\lambda_i)} p(x) \;=\; \sum_{x \in e(\lambda_i)} q(\{x\}) \;=\; q(e(\lambda_i)),$$

and hence

$$\begin{array}{rcl} q(f) & = & \lambda_1 q(e(\lambda_1)) + \cdots \lambda_m q(e(\lambda_m)) \\ & = & \lambda_1 p(e(\lambda_1)) + \cdots \lambda_m p(e(\lambda_m)) \\ & = & \sum_{x \in X} p(x) f(x). \end{array}$$

□

2 Dynamics

For finite state spaces, the mathematical representation of dynamical evolution is not very interesting. It's much more interesting for an infinite space X that might have further structure — such as a topology, or a metric, or a symplectic form. In any case, whether X is finite or infinite, one might assume that dynamical evolution is a "flow" on X, which we can represent by a parameterized family $u_t : X \to X$, $t \in \mathbb{R}$ of automorphisms of X. Furthermore, it would be natural to require that $u_{t+s} = u_t u_s$ and $u_{-t} = u_t^{-1}$. Notice, however, that this representation presumes **determinism**, i.e. that the state of a system at one time fixes the state of the system of future times.

In other words, a family such as $\{u_t \mid t \in \mathbb{R}\}$ is a **deterministic dynamical law**.

In contrast, a **stochastic dynamical law** would specify a probability distribution over future states. For example, $p_t(x, -)$ could define a probability distribution on X. We would then want to specify some further properties of the map $t, x \mapsto p_t(x, -)$, but we will not pursue that here.

There is another way that one can specify a stochastic dynamical law, and that is as one-parameter family of morphisms on the space $M(X)$ of states (i.e. probability distributions) on X. If a pure state p_y is mapped to a mixed state q, then that could naturally be interpreted as a stochastic process, where the transition probability from p_y to p_x is given by $q(x)$.

3 Composite systems

Given two state spaces X and Y, the state space of the composite system is the Cartesian product

$$X \times Y \;=\; \{\langle x, y\rangle \mid x \in X, y \in Y\}.$$

In this case, Prop 4 implies that every pure state is of the form $p_{\langle x,y\rangle}$. (Note: The angle-bracket notation $\langle x, y\rangle$ is put to multiple use in these notes. We will trust context to disambiguate which way we're using it.)

The space $X \times Y$ has the feature that for any functions $f : X \to \mathbb{R}$ and $g : X \to \mathbb{R}$, there is a unique function $f \times g : X \times Y \to \mathbb{R}$ given by

$$(f \times g)(x, y) \;=\; f(x)g(y), \qquad (x \in X, y \in Y).$$

However, there are also functions on $X \times Y$ that do not decompose in this way. For example, let $X = Y = \{a, b\}$, and consider the function p such that

$$p(x, y) \;=\; \begin{cases} \frac{1}{2} & x = y, \\ 0 & x \neq y. \end{cases}$$

In fact, this function p is a probability measure on $X \times Y$. Intuitively, it's a state in which the two systems are **strictly correlated**: either both are in state a, or both are in state b. Nonetheless, each state on $X \times Y$ is a convex combination of pure states. In particular, $p = \sum_i \lambda_i p_i$, where each p_i is a state of the form $p_x \times p_y$. This mathematical fact corresponds to the physical fact that correlated states can be interpreted *epistemically*, e.g. as representing our ignorance of the real state of the system, which is a logical sum of the state of the individual subsystems.

3.1 Bell's inequality

For real numbers $a, b \in [-1, 1]$, we claim that

$$|a + b| + |a - b| \;\leq\; 2. \tag{5}$$

Indeed, if $a + b$ is positive, then $|a + b| + |a - b| = 2\max\{a, b\}$, and if $a + b$ is negative, then $|a + b| + |a - b| = 2\max\{-a, -b\}$.

Now we consider two systems with state spaces X and Y. Let $f_1, f_2 \in \mathbb{R}^X$ such that $\mathrm{spec}(f_i) \subseteq [-1, 1]$, and let $g_1, g_2 \in \mathbb{R}^Y$ such that $\mathrm{spec}(g_i) \subseteq [-1, 1]$. That is, f_1 and f_2 are quantities associated with system X, and g_1 and g_2 are quantities associated with system Y. Consider the quantity represented by the function

$$r \;=\; f_1 \times (g_1 + g_2) + f_2 \times (g_1 - g_2).$$

This r is called a **Bell observable**, and it could, in principle, be measured by two observers with systems X and Y. We then have

$$\begin{array}{rcl} |r(x, y)| & = & |f_1(x)g_1(y) + f_1(x)g_2(y) + f_2(x)g_1(y) - f_2(x)g_2(y)| \\ & \leq & |f_1(x) + f_2(x)| + |f_1(x) - f_2(x)| \\ & \leq & 2, \end{array}$$

where the final inequality follows from Eq. 5, since $f_1(x), f_2(x) \in [-1, 1]$.

Bell's Theorem. *If r is a Bell observable, then*

$$|p(r)| \leq 2, \tag{6}$$

for any classical probability measure p.

Equation 6 is called **Bell's inequality**, or to be more accurate the **CHSH** variant of Bell's inequality (in honor of Clauser, Horne, Shimony, and Holt).

Proof. The discussion above shows that $-2 \leq q(r) \leq 2$ for any pure state q. An arbitrary state p is a convex combination of pure states, and so the result holds for p as well. □

3.2 Physical significance of Bell's theorem

At the time when Bell proved his theorem, it was already known that QM would predict a violation of Bell's inequality. (The calculation is quite simple, as we will see in the the next chapter.) However, at that time, no experiment had been undertaken to verify QM's prediction. In the intervening years, many different experiments have confirmed that Bell's inequality is violated. (Most people think that the decisive experiment was the one undertaken by Alain Aspect in 1982.) Let's look at the significance of these two facts in reverse order.

First, what is the significance of the fact that there are experimental violations of Bell's inequality? In the first instance, the only significance of this fact is that the model we made above does not adequately describe those experimental situations. In other words, if we thought that there were two systems X and Y, and that classical probability theory was applicable in the simple way we described above, then we would derive a false prediction about the outcomes the experiments.

Some other philosophers and physicists have not been so modest in their claims about what these experiments show. For example, according to Tim Maudlin, the violation of Bell's inequality show quite simply that the physical universe has a feature called "non-locality".

> [John Bell] taught us something about the world we live in, a lesson that will survive even the complete abandonment of quantum theory. For what cannot be reconciled with locality is an observable phenomenon: the violations of Bell's inequality for 'measurements' performed at arbitrary distances apart, or at least at space-like separation. And this phenomenon has been verified, and continues to be verified, in the lab. Neither indeterministic nor deterministic theories can recover these predictions in a local way. Non-locality is here to stay. [Maudlin, 2014, p 22]

Similarly, Travis Norsen claims that, "nonlocality really is required to coherently explain the empirical data" [Norsen, 2016]. This is an interesting point of view, and there are a couple of different ways to read it — either in the material mode, or in the formal mode. (The material mode is speaking about the universe, and the formal mode is speaking about theories.) In the material mode, the claim seems to be that some possible universes are local, and others are non-local, but that any universe that displays violations of

Bell's inequalities is one of the non-local universes. But that claim doesn't look anything like what Bell actually proved. Bell didn't talk about varieties of universes, and he didn't give us any insight into what a non-local universe would look like.

To read Maudlin and Norsen's claims in the formal mode would have Bell showing something like this:

> There is no theory T with property Φ such that T predicts violations of Bell's inequalities

where, in this particular case, Φ is the property of being a local theory. Once again, the claim seems too strong. Bell didn't do any surveying of the space of all possible theories, so it's not clear how his result could show anything of this sort. Instead, what Bell showed is that a certain familiar kind of modelling strategy — classical probability — makes the wrong predictions for these kinds of experiments. We have a long way to go before we can say anything about all possible future theories.

In fact, in the decades immediately following Bell's theorem, there was a different consensus about the physical significance of the result. In particular, the common view was that Bell's theorem should be thought of as a derivation of an (experimentally testable) inequality from the conjunction of two premises:

realism The moon is there even when no-one is looking.

locality Things that happen in one place cannot have an instantaneous effect on things in another place.

(The classic "Jarrett analysis" of Bell's derivation can be found in [Jarrett, 1984].) I've purposely stated these premises in both a vague, and an overspecific, way. The point of doing so is that as soon as one starts explicating (i.e. formalizing) these premises, then one has to beg some questions about the framework. In standard analyses of Bell's theorem, one begins immediately to translate **locality** into a statement about conditional probabilities. But to apply classical probability theory to a complicated situation requires making quite a few physical assumptions about what's going on.

In the case at hand, note that the Bell observable

$$f_1 \times (g_1 + g_2) + f_2 \times (g_1 - g_2),$$

is built out of four different observables: f_1 and f_2 belong to the first experimenter, and g_1 and g_2 belong to the second experimenter. Hence, to successfully carry out a test of Bell's inequality, the first experimenter must perform at least two different measurements, and the second experimenter must also perform at least two different measurements. So, we're not talking about any single state of affairs, but a sequence of different experiments. If we assume that these four experiments can be jointly modelled in the way that classical physics suggests, then we get a false prediction (i.e. that Bell's inequality would be satisfied).

To be clear, to prove a claim of the form

$$\textbf{locality} \implies |p(r)| \leq 2 \ ,$$

one first has to make **locality** into a mathematically precise statement. So let's say that **i-locality** is our intuitive concept of locality, and let's say that **m-locality** is a mathematical precisification of **i-locality**. Then Bell's theorem is of the form

$$\textbf{m-locality} \implies |p(r)| \leq 2 \ ,$$

and the experimental result $p(r) > 1$ shows that **m-locality** doesn't hold. Does it follow that **i-locality** doesn't hold? Well, not unless the intuitive concept of locality demands a particular mathematical explication. Perhaps it does; we will have to think about that. (For a similar argument, see [Werner, 2014].)

References

[Jarrett, 1984] Jarrett, J. P. (1984). On the physical significance of the locality conditions in the Bell arguments. *Noûs*, 18:569–589.

[Maudlin, 2014] Maudlin, T. (2014). What Bell did. *Journal of Physics A: Mathematical and Theoretical*, 47(42):424010.

[Norsen, 2016] Norsen, T. (2016). Quantum solipsism and non-locality. In *Quantum Nonlocality and Reality*, pages 204–237. Cambridge University Press.

[Werner, 2014] Werner, R. F. (2014). Comment on 'What Bell did'. *Journal of Physics A: Mathematical and Theoretical*, 47(42):424011.