Advantages of Probability Amplitude Over Probability Density in Quantum Mechanics

Yoshimasa Kurihara[1] & Nhi My Uyen Quach[2]

[1] The High Energy Accelerator Organization (KEK), Tsukuba, Ibaraki 305-0801, Japan

[2] The Graduate University for Advanced Studies, Tsukuba, Ibaraki 305-0801, Japan

Correspondence: Yoshimasa Kurihara, The High Energy Accelerator Organization (KEK), Tsukuba, Ibaraki 305-0801, Japan. E-mail: yoshimasa.kurihara@kek.jp

Abstract

We discuss reasons why a probability amplitude, which becomes a probability density after squaring, is considered as one of the most basic ingredients of quantum mechanics. First, the Heisenberg/Schrödinger equation, an equation of motion in quantum mechanics, describes a time evolution of the probability amplitude rather than of a probability density. There may be reasons why dynamics of a physical system are described by amplitude. In order to investigate one role of the probability amplitude in quantum mechanics, specialized codeword-transfer experiments are designed using classical information theory. Within this context, quantum mechanics based on probability amplitude provides the following: i) a minimum error of the codeword transfer; ii) this error is independent of coding parameters; and iii) nontrivial and nonlocal correlation can be realized. These are considered essential advantages of the probability amplitude over the probability density.

Keywords: probability amplitude, probability density, quantum mechanics

1. Introduction

Quantum mechanics (QM) is considered the most basic theory of nature. All phenomena including those of the gravitational force are considered to be expressed by a language of QM. However, an essential understanding of the basic nature of QM yet to be realized, and efforts to look for more fundamental explanations continue. Of course, QM itself is a self-consistent theory and requires no fundamental reasoning to support its truths beyond what are gains from experiments. Still, it is worth pursuing more basic reasons which determine QM to be the most fundamental law of nature. For instance, Wheeler asked "Why the quantum?" and discussed the relation between QM and information theory (Wheeler, 1990, 1991). In this report we attempt to answer the same question from Wheeler's point of view. One of the most essential differences between quantum and classical mechanics is the former's need for a probabilistic treatment of theoretical predictions. One cannot avoid the probabilistic interpretation of a wave function proposed by Born (1926), which is now known as the *Copenhagen interpretation*. A fundamental equation of QM, the Heisenberg/Schrödinger equation, does not describe the behavior of a physical observable nor its probability density; rather, it describes the probability amplitude, which is a characteristic of QM and possesses no classical counterpart. (In a narrow sense,"quantum amplitude" is a complex number whose square of the absolute value is a probability. In this report, we use a word "quantum amplitude" not only for complex numbers, but also for vectors whose square of the absolute value is a probability.) This report considers reasons why fundamental laws of physics are described by probability amplitude instead of probability density, leaving aside the question of why probability itself is necessary. To clarify essential properties of probability amplitude, codeword-transfer experiments are designed on the basis of classical information theory. Taking into account the discussions on these experiments, three essential advantages of probability amplitude over probability density are pointed out in the following sections.

First, definition of quantum system and probability amplitude are given in Section 2 under a very general mathematical framework. Then, codeword-transfer experiments are designed within classical information theory to investigate the role of probability amplitude. Experiments using a stochastic algorithm cannot avoid statistical error due to sample number. In Section 3, we show that a coding method based on probability amplitude should minimize statistical error. Moreover, statistical errors of the codeword-transfer are independent of the parametrization allowing each character to be transferred; this is shown in Section 4. Another essential feature of

QM is its lack of local realism, which can be judged by Bell's inequality. This local realism and Bell's inequality are described using terminology of classical information theory, again as the codeword-transfer experiment. A method based on the probability amplitude can induce a violation of Bell's inequality, as shown in Section 5. Throughout this report, classical information theory is used to describe codeword-transfer experiments.

2. General Quantum System

A general framework to define the probability amplitude appearing in QM is considered in this section. Here we emphasis algebraic aspects of QM and ignore dynamical ones. The question which must be asked here is "What minimum set of assumptions makes a system look like quantum mechanics?" We propose the following elements as indispensable ingredients for QM.

Definition 2.1 (Quantum Space)

$\mathbb{K}$ is any field and V is a linear (vector) space on it. $\mathbb{K}$ is named as a base field and is associated to each point of a set, M. State vector and probability measure are introduced on these spaces as follows.

1) A map from a point on M to a tensor product of a vector space V,

$$\mathbf{\Psi}: M \to V^k = \underbrace{V \otimes \cdots \otimes V}_{k}: x \mapsto \mathbf{\Psi}(x) = \{\psi^1(x), \cdots, \psi^k(x)\}, \tag{1}$$

is named state vectors. Here, M is named the base set and x is a point on it.

2) A map from the state vector to a real number such as

$$\mu: V \to \mathbb{R}: \psi^i \to \mu(\psi^i) \in [0,1], i = 1, \cdots, k \tag{2}$$

is named a probability measure. The index i on ψ^i runs from 1 to k. The sequential map

$$\mu \circ \psi^i: M \to V \to \mathbb{R}: x \mapsto \mu^i(x) = \mu(\psi^i)(x) \tag{3}$$

is also called a probability measure and represented by the same symbol, μ, when V are obvious.

3) The probability measure must be normalized as

$$\int_{x \in \Gamma \subseteq M} \mu(\psi^i)(x) = 1 \tag{4}$$

for each i, where Γ is an appropriate subset of the base set M. Since the probability measure is considered as Lebesgue measure, the integral should be interpreted as summation when M is a discrete set.

4) The set $\{\mathbb{K}, V, \mathbf{\Psi}, \mu\}$ is named a "quantum space."

To construct QM, these conditions are necessary, but are not sufficient. For standard relativistic QM (or quantum field theory), we take Hilbert space as a vector-space V on a field of complex numbers $\mathbb{C}$. State vector can be constructed using square integrable functions on a given support. The state vector is associated with each point of the Minkowski manifold as a base set. (Sometimes a Fourier transformation of ψ^i defined in the momentum manifold is used instead of ψ^i itself. In that case, a corresponding Hilbert space is called "Fock space".) The probability measure is introduces as $\mu(\psi^i) = |\psi^i|^2$. For the normalization, Γ is taken as a hyper-surface on M such that any two points on Γ have a space-like distance each other. (Or it is normalized in the momentum space.) When the probability measure is defined as square of the absolute value of the state vector, the state vector is called a "*probability amplitude*" in this report, hereafter. In this report, simple quantum spaces are used since only algebraic aspects of QM are of interest here.

3. Minimization of Measurement Error

First, let us consider a statistical error for measurements of a single physical observable on the quantum space defined in the previous section. A codeword-transfer experiment simulating standard QM in a much simpler quantum space, retaining essential properties, is introduced here. In information theory, an encoding method which minimizes statistical error among methods using stochastic algorithms is known. The method using probability amplitude is shown to be an example of such an encoding method giving minimum errors. Terminology of classical information theory used here can be found in Cover and Thomas (1991) and Kurihara (2013) and Appendix 7.1.

Definition 3.1 (Stochastic Codeword-Transfer Experiment)

The experiment satisfying the following conditions is called a *stochastic codeword-transfer experiment*:

1) Alice ($\mathcal{A}$) transfers a set of m different codewords $\mathcal{W} = \{w_1, \cdots, w_m\}$ to Bob ($\mathcal{B}$) after converting them to state vectors $\psi \in \mathcal{V}$, where $\mathcal{V}$ is a m-dimensional vector space.

2) $\mathcal{B}$ receives a state vector sent from $\mathcal{A}$ and obtained one of codewords $\mathcal{W}$ by measuring them. Here meaning of "measuring" will be explain in following items.

3) The same probabilistic function of

$$\mu_c: \mathcal{V} \to \mathbb{R}: \mu_c(\psi)(\omega_i) \mapsto p_i \in [0,1]$$

is given for $\mathcal{B}$. The value p_i gives a probability to observe a codeword ω_i. Only one vector space $\mathcal{V}$ appears here, then the function $\mu_c(\psi)(\omega_i)$ will be written as $\mu_c(\omega_i)$, hereafter.

4) Probabilistic function μ_c is normalized as:

$$\sum_{i=1}^{m} \mu_c(\omega_i) = 1.$$

5) $\mathcal{A}$ can repeat to send a finite number (n times here) of the same state vectors to $\mathcal{B}$.

6) $\mathcal{B}$ obtains n independent codewords by measuring sets of state vectors sent from $\mathcal{A}$, such as $\mathcal{X} = \{x_1, \cdots, x_n\}$.

7) $\mathcal{B}$ has an unbiased estimator to obtain a set of real numbers $\bar{x}_i \in [0,1]$ from measured data as

$$\bar{x}_i = T_i(\mathcal{X}) = \frac{1}{n}\sum_{j=1}^{n} \Theta_i(x_j),$$

$$\Theta_i(x_j) = \begin{pmatrix} 1 & (x_j = \omega_i), \\ 0 & (x_j \neq \omega_i). \end{pmatrix}$$

Here $\bar{x}_i$ converges in probability to p_i when $n \to \infty$, thanks to the law of large numbers.

8) Finally $\mathcal{B}$ obtains a sequence of numbers $\{\bar{x}_1, \cdots, \bar{x}_m\}$, which $\mathcal{A}$ intended to send.

This codeword-transfer experiment is constructed on the quantum space $\{\mathcal{W}, \mathcal{V}, \psi, \mu_c\}$ as defined above. In this case, positions where $\mathcal{A}$ or $\mathcal{B}$ exists are not specified. No dynamical structure is assumed to transport a state vector from $\mathcal{A}$ to $\mathcal{B}$ here, however it is just assumed that these two points are separated from each other and there is no way to communicate other than the state-vector transfer. A question to ask here is how may one find the probability measure μ_c, which maps the state vector ψ to a real number $\mu_c(\psi)$ to minimize an error of this experiment for any ψ. The answer is already known as a theorem, which was first obtained by Fisher (1922). Wootters stated this theorem (Wootters, 1980) without any proof but later provided the same by introducing a statistical distance (Wootters, 1981). Recently Wootters discussed this subject again in Wootters (2013). Here we state the theorem clearly again and give an independent and much simpler proof using an information theory.
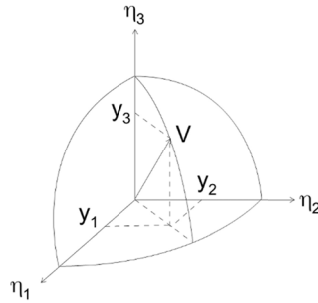

Figure 1. Example for $m = 3$: assignment of a vector $\mathcal{V}$ on two-dimensional sphere

Theorem 3.1 (Fisher–Wootters)

Among stochastic codeword-transfer experiments, that which employs the following probability measure gives the smallest error to measure a single codeword from a set of codewords:

1) $\mathcal{A}$ selects a set of codewords $\mathcal{W} = \{w_1, \cdots, w_m\}$ and a sequence of numbers $\mathcal{P} = \{p_1, \cdots, p_n\}$ which is intended to be sent to $\mathcal{B}$. The $\mathcal{P}$ is normalized as $\sum_{i=1}^{n} p_i = 1$.

2) $\mathcal{A}$ prepares an m-dimensional Euclid space $\mathbb{R}^m$ and orthonormal bases $\{\eta_1, \cdots, \eta_m\}$.

3) $\mathcal{A}$ sets a state vector ψ as to map each element of $\mathcal{P}$ at a point on a unit sphere S_1^{m-1} centered at the origin of $\mathbb{R}^m$ to be $p_i = |y_i|^2$, where y_i is an i's component of the position of ψ on S_1^{m-1} by the orthonormal bases defined above.

4) $\mathcal{B}$ obtains a codeword ω_i with the probability measure $\mu_c(\omega_i) = |y_i|^2$.

Proof.

1) The smallest error $\Rightarrow \mu_c(\omega_i) = |y_i|^2$:

Data after n independent measurements are expressed as $\mathcal{X} = (x_1, \cdots, x_n)$ with the probability $\mu_c(\omega_i) = |y_i|^2$. The probability density to obtain a set of data $\mathcal{X}$ is assumed to be expressed as $f(\mathcal{X}; \psi) = \mu_c(\omega_i)$, where $\mu_c(\omega_i)$ used defined as an equation (3). Then the Fisher information matrix (FIM) (Cover & Thomas, 1991) can be written as

$$J_{ij} = f(\mathcal{X}; \psi) \frac{\partial \log f(\mathcal{X};\psi)}{\partial \omega_i} \frac{\partial \log f(\mathcal{X};\psi)}{\partial \omega_j}$$

$$= \sum_{k=1}^{m} \mu_c(\omega_k) \frac{\partial \mu_c(\omega_k)/\partial \omega_i}{\mu_c(\omega_k)} \frac{\partial \mu_c(\omega_k)/\partial \omega_j}{\mu_c(\omega_k)}.$$

The functions $\mu_c(\omega_i)$ are not independent of each other owing to conservation of the total probability, $\sum_{i=1}^{m} \mu_c(\omega_i) = 1$. We can assume that all $\mu_c(\omega_i)(i \geq 2)$ are independent except $\mu_c(\omega_1) = 1 - \sum_{j=2}^{m} \mu_c(\omega_j)$ without any loss of generality. Since all other $\mu_c(\omega_{i \neq 1})$, except this correlation due to the conservation of probability, can be set to be independent after appropriate linear transformation of μ_c, the FIM can be taken to be a diagonal matrix. Here we use a short-hand expression, $\mu_c(\omega_i) = \mu_i$, $d\mu_c(\omega_j)/d\omega_i = \mu_{j,i}$, and $\sum_{j=2}^{m} \mu_c(\omega_j) = \bar{\mu}$; then the diagonal components of the FIM can be written as

$$J_{ii} = \sum_{k=1}^{m} \mu_k \left(\frac{\mu_{k,i}}{\mu_k}\right)^2$$

$$= \mu_1 \left(\frac{\mu_{1,i}}{\mu_1}\right)^2 + \sum_{k=2}^{m} \mu_k \left(\frac{\mu_{k,i}}{\mu_k}\right)^2$$

$$= (1 - \bar{\mu}) \left(\frac{\partial(1-\bar{\mu})/\partial\omega_i}{1-\bar{\mu}}\right)^2 + \mu_i \left(\frac{\mu_{i,i}}{\mu_i}\right)^2$$

$$= \frac{\mu_{i,i}^2}{1-\bar{\mu}} + \frac{\mu_{i,i}^2}{\mu_i}.$$

Here the independence of all $\mu_{k \geq 2}$ each other is used second line to third line in above calculations. The minimum value of J_{ii} is obtained when $\bar{\mu} = \mu_i$ within the allowed region of $\mu_i \leq \bar{\mu} \leq 1$. Then we get

$$\min\{J_{ii}\} = \tilde{J}_{ii} = \frac{\mu_{i,i}^2}{\mu_i(1-\mu_i)}.$$

On the other hand, measured data after n independent measurements must follow a multinomial distribution, whose covariance matrix σ is

$$\sigma_{ij} = \begin{pmatrix} n\tilde{p}_i\tilde{p}_j & (i \neq j), \\ n\tilde{p}_i(1 - \tilde{p}_i) & (i = j), \end{pmatrix}$$

where $\tilde{p}_i$ is measured probability of an ith codeword. Then, after n independent measurements through estimator T defined in **Definition 2.6**, a covariant matrix $\Sigma(\mathcal{X})$ can be expressed as

$$\Sigma_{ij}(\mathcal{X}) = \frac{1}{n}\sum_{k=1}^{n} (\Theta_i(x_k) - \bar{x}_i)(\Theta_j(x_k) - \bar{x}_j)$$
$$\approx \frac{1}{n}\sigma_{ij}.$$

Then, diagonal components of the covariant matrix become

$$\Sigma_{ii}(\mathcal{X}) \approx \frac{1}{n}\sigma_{ii}$$
$$= \tilde{p}_i(1 - \tilde{p}_i).$$

In general, measured probability $(\tilde{p}_i)$ differs from true probability (μ_i); however, it is certain that the error of $|\tilde{p}_i - \mu_i|$ will be less than any small value after a sufficient number of events accumulates, as a result of the law of large numbers and the assumption that the estimator is unbiased. Then, we use μ_i instead of $\tilde{p}_i$ in the discussions that follow. The probability μ_i that maximizes diagonal components of the covariant matrix is given as $\mu_i = 1/2$ due to $d\Sigma_{ii}/d\mu_i = (1 - 2\mu_i) = 0$. Then, the diagonal components of the covariant matrix are given as $\Sigma_{ii} = 1/4$. The Cramér–Rao inequality (Cover & Thomas, 1991; Cram`er, 1946; Rao, 1945) gives the lower bound of the covariant matrix as

$$\Sigma(\mu) \geq \mathbf{J}^{-1}.$$

A possible range of the inverse of the FIM is

$$\frac{\mu_i(1-\mu_i)}{\mu_{i,i}^2} \geq (\mathbf{J}^{-1})_{i,i} \geq \frac{1}{\tilde{J}_{i,i}} \geq 0,$$

where we use the FIM (**J**) is a diagonal matrix. Then a solution of the following differential equation gives the minimum variance in general:

$$\frac{\mu_{i,i}^2}{\mu_i(1-\mu_i)} = 4$$

$$\Rightarrow \mu_{i,i}^2 = 4\mu_i(1-\mu_i).$$

The solution of this equation can be obtained as

$$\mu_i = \cos^2(\omega_i + \phi_i),$$

where ϕ_i is an arbitrary phase factor. This phase factor corresponds to a rotation of the coordinate system prepared in **Theorem 3.1** and gives no essential effect on the result. Then we set $\phi_i = 0$ hereafter as $\mu_i = \cos^2\omega_i$. Each ω_i gives the same differential equation; then parametrization $y_i = \sqrt{\mu_i} = \cos\omega_i$ gives the lowest value of the variance, which is nothing other than the direction cosine of the vector V, whose endpoint is on the unit sphere S_1^{m-1}. Then, the method to give the minimum variance is: **i)** normalize the codeword ω_i to $0 \leq \omega_i \leq \pi/2$; **ii)** map on the S_1^{m-1} as ω_i to be an angle from axis η_i; set **iii)** the probability to observe the codeword ω_i to be $\cos^2\omega_i$, which are the same as the assumptions of the theorem.

2) $\mu_c(\omega_i) = |y_i|^2 \Rightarrow$ the smallest error:

When we set $\mu_i = |y_i|^2 = \cos^2\omega_i$, the diagonal components of a covariant matrix become

$$\Sigma_{ii}(\mu) = \mu_i(1-\mu_i)$$

$$= \cos^2\omega_i(1-\cos^2\omega_i)$$

$$= \cos^2\omega_i\sin^2\omega_i.$$

Then the minimum value of Σ_{ii} is obtained to be $1/4$ at $\omega_i = \pi/4$. On the other hand, the diagonal component of the FIM matrix can be

$$\tilde{J}_{ii} = |\mu_{i,i}|^2 \frac{1}{\mu_i(1-\mu_i)}$$

$$= 4\cos^2\omega_i\sin^2\omega_i/(\cos^2\omega_i\sin^2\omega_i)$$

$$= 4.$$

Then $\tilde{J}_{ii}^{-1} = 1/4$, which matches the minimum value of Σ_{ii}.

In the above decoding method, a relation between probability amplitude and density is algebraically the same as in the standard QM, which means the latter employs a coding method that minimizes statistical error among other stochastic methods. This is our first example outlining the advantage of the method using probability amplitude.
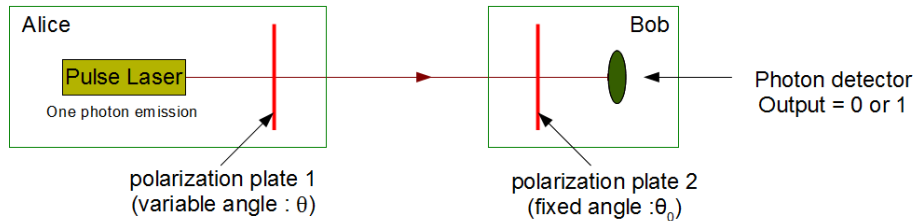

Figure 2. Code-transfer experiment realized by using polarized laser beam

4. Parametrization independence of a measurement error

Related to the **Theorem 3.1**, one can prove following theorem, which is also given by Wootters (Wootters, 1980, 1981) and is important to consider one role of the probability amplitude.

Theorem 4.1

The encoding rule given by the Fisher–Wootters theorem gives uniform errors independent of its parametrization.

Proof. A set of state vectors $\mathcal{V} = \{\omega_1, \cdots, \omega_m\}$ are encoded as $y_i = \cos\omega_i$, $\sum_{i=1}^m y_i^2 = 1$, according to the Fisher–Wootters theorem. Looking at an ith element ω_i, one sees a relation between an error of estimation $\delta\omega_i$ and an error to measure the parameter δy_i as $|\delta y_i / \delta \omega_i| = \sin\omega_i$. Under the normalization condition of $\sum_{i=1}^m y_i^2 = \sum_{i=1}^m \cos^2 \omega_i = 1$, the total error after measuring n independent data becomes

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^m \left| \frac{\delta y_i}{\delta \omega_i} \right|^2$$

$$= \frac{1}{n} \sum_{i=1}^m |\sin^2 \omega_i|$$

$$= \frac{1}{n} \sum_{i=1}^m |1 - \cos^2 \omega_i|$$

$$= \frac{m-1}{n}$$

$$\Rightarrow \sigma = \sqrt{\frac{m-1}{n}},$$

which means a mean-square error is determined by the statistics per degree of freedom and independent of the position on an m-dimensional sphere. A factor $\sigma^2 \propto 1/n$ follows from the central limit theorem.

Example 1 (Codeword-transfer experiment realized using a polarized laser beam)

Let us consider the codeword-transfer experiment defined in **Definition 3.1** for a realistic quantum system: the pulse laser has a polarizer ($\lambda/4$ plate). (See Fig. 2.) Alice ($\mathcal{A}$) has experimental equipment consisting of a pulse laser and a polarizer and can transfer a single photon with linear polarization with any polarization plane to Bob ($\mathcal{B}$). $\mathcal{B}$ has a $\lambda/4$ plate with fixed plane and photon detector with 100% efficiency. $\mathcal{A}$ knows the angle of the polariser plane of $\mathcal{B}$, say θ_0, and has a clock exactly synchronised to that of $\mathcal{B}$. $\mathcal{A}$ assigns codewords on equally separated points on a unit circle, and selects an integer, say j. Then $\mathcal{A}$ sets an angle of the polariser according to a codeword to be $\theta = \theta_0 + \alpha$, where $\alpha = j/2\pi$. $\mathcal{A}$ transfers one photon a second and n photons in total. $\mathcal{B}$ measures photons behind the $\lambda/4$ plate. If $\mathcal{B}$ observes a photon, he records "1" and if not, he records "0". As a result $\mathcal{B}$ obtains data $\mathcal{X}_n = \{X_1, X_2, \cdots, X_n\} = \{1,1,0,1,0,\cdots\}$, and decodes them to one real number with average $\bar{x} = \sum_i^n X_i / n$. According to quantum mechanics this number must be $\bar{x} = \sin\alpha$. Finally, $\mathcal{B}$ obtains a number which $\mathcal{A}$ intended to send. This codeword-transfer experiment satisfies **Definition 3.1**, which means quantum mechanics gives codeword-transfer experiments with the smallest errors, given by Theorem. 3.1.

5. Nonlocal realism

A point definitely distinguishing QM from classical mechanics is that QM does not have local realism. Related to this fact, there are two important theorems: violation of Bell's inequality (1964) and Kochen–Specker theorem. Both theorems are related to a correlation of two independent measurements. It is shown in this section that these two theorem can be realized again using the probability amplitude. The trick to introduce the nonlocal realism without breaking the special relativity is realized unsing a freedom in the map from the state vector (probability amplitude) to the probability measure defined in the equation (3). An additional information can be implemented into the probability amplitude as a phase factor of a $U(1)$ transformation, which keeps the probability measure unchanged locally. This is another example of the advantage of the probability amplitude over the probability density.

In order to discuss a correlation of two independent measurements, a double codeword-transfer experiment is designed.

Definition 5.1 (Stochastic double codeword-transfer experiment)

A stochastic double codeword-transfer experiment is defined by extending Definition 3.1 as follows:

1) Alice ($\mathcal{A}$) transfers two sets of m different codewords and state vectors, $\mathcal{W}_\alpha = \{\alpha_1, \cdots, \alpha_m\}$ and $\mathcal{W}_\beta = \{\beta_1, \cdots, \beta_m\}$, to Bob ($\mathcal{B}$) and Charley ($\mathcal{C}$) after converting them to state vectors $\psi_\alpha \in \mathcal{V}_\alpha$ and $\psi_\beta \in \mathcal{V}_\beta$, where $\mathcal{V}_\alpha$ and $\mathcal{V}_\beta$ are m-dimensional vector spaces.

2) ($\mathcal{B}$) and ($\mathcal{C}$) are placed opposite to $\mathcal{A}$ and receive state vectors sent from $\mathcal{A}$, stochastically choose one of the two sets to be measured. Neither $\mathcal{B}$ and $\mathcal{C}$ know which set is chosen by the other (independence of set selection).

3) Encoding is performed using the following probabilistic function:

$$\mu_{dc}:(\mathcal{V}_\alpha \oplus \mathcal{V}_\beta) \otimes (\mathcal{V}_\alpha \oplus \mathcal{V}_\beta) \to \mathbb{R}: (\gamma_i, \gamma_j) \mapsto \mu_{dc}(\gamma_i, \gamma_j) = p_{i,j} \in [0,1],$$

where $\mathcal{V}_\alpha \oplus \mathcal{V}_\beta = (\alpha_1, \cdots, \alpha_m, \beta_1, \cdots, \beta_m) = (\gamma_1, \cdots, \gamma_{2m})$ and $1 \le i, j \le 2m$. First (second) slot of μ_{dc} are for state vectors sent to $\mathcal{B}$ ($\mathcal{C}$), respectively.

4) $\mathcal{B}$ and $\mathcal{C}$ select for measurement one of the state vectors ψ_A or ψ_B, independently. Possible combinations of measured codewords are $\{(\alpha_k, \alpha_l), (\alpha_k, \beta_l), (\beta_k, \alpha_l), (\beta_k, \beta_l)\}$. Probabilistic function μ_{dc} is normalized as:

$$\forall \gamma_j, \sum_{i=1}^m \mu_{dc}(\alpha_i, \gamma_j) = 1, \quad \forall \gamma_j, \sum_{i=1}^m \mu_{dc}(\beta_i, \gamma_j) = 1,$$
$$\forall \gamma_i, \sum_{j=1}^m \mu_{dc}(\gamma_i, \alpha_j) = 1, \quad \forall \gamma_i, \sum_{j=1}^m \mu_{dc}(\gamma_i, \beta_j) = 1.$$

However it does not guarantee that all the probability measures, $\mu_{dc}(\alpha_k, \alpha_l)$, $\mu_{dc}(\alpha_k, \beta_l)$, $\mu_{dc}(\beta_k, \alpha_l)$, and $\mu_{dc}(\beta_k, \beta_l)$, exist at the same time.

5) $\mathcal{A}$ can send a finite number (n times here) of the same set of state vectors to $\mathcal{B}$ and $\mathcal{C}$.

6) Measurements:

(a) $\mathcal{B}$ obtains n independent codewords by measuring sets of state vectors sent from $\mathcal{A}$, such as $\mathcal{X}^\mathcal{B} = \{x_1^\mathcal{B}, \cdots, x_n^\mathcal{B}\}$, where $x_i^\mathcal{B} \in \mathcal{W}_\alpha \oplus \mathcal{W}_\beta$.

(b) For $\mathcal{C}$, the same as *(a) with a replacement* $\mathcal{B} \to \mathcal{C}$.

7) Estimator:

(a) $\mathcal{B}$ has an unbiased estimator to obtain a set of real numbers $\bar{x}_i \in [0,1]$ from measured data as

$$\bar{x}_i^\mathcal{B} = T_i(\mathcal{X}^\mathcal{B})$$

$$= \frac{\sum_{k=1}^n \Theta_i(x_k^\mathcal{B})}{\sum_{l=1}^{2m} \sum_{k=1}^n \Theta_l(x_k^\mathcal{B})}$$

$$\Theta_i(x_j) = \begin{cases} 1 & (x_j = \gamma_i), \\ 0 & (x_j \ne \gamma_i), \end{cases}$$

where i runs from 1 to $2m$.

(b) For $\mathcal{C}$, the same as above with a replacement $\mathcal{B} \to \mathcal{C}$.

8) After completing measurement, $\mathcal{B}$ and $\mathcal{C}$ make a table $\bar{x}_{i,j} = (\bar{x}_i^\mathcal{B}, \bar{x}_j^\mathcal{C})$, where $\bar{x}_{i,j}$ converges in probability to $p_{i,j}$ when $n \to \infty$, thanks to the law of large numbers.

Bell's inequality is a critical test to distinguish a nonlocal theory from a local one. This theorem can be expressed by the language of classical information theory (Braunstein & Caves, 1988). We state this theorem and give a proof in the context of **Definition 5.1**.

Theorem 5.1 (Bell)

Let us consider a case with a complete table to give the probability of observing any pair of codewords as

$$P(\alpha_{i_1}, \alpha_{i_2}, \beta_{j_1}, \beta_{j_2}) = \mu_{dc}(\alpha_{i_1}, \alpha_{i_2})\mu_{dc}(\beta_{j_1}, \beta_{j_2})$$
$$+ \mu_{dc}(\alpha_{i_1}, \beta_{j_2})\mu_{dc}(\beta_{j_1}, \alpha_{i_2})$$
$$+ \mu_{dc}(\beta_{j_1}, \alpha_{i_2})\mu_{dc}(\alpha_{i_1}, \beta_{j_2})$$
$$+ \mu_{dc}(\beta_{j_1}, \beta_{j_2})\mu_{dc}(\alpha_{i_1}, \alpha_{i_2}).$$

These measurements are performed as the stochastic double codeword-transfer experiment defined above. In this case, a conditional entropy follows the inequality

$$S(\alpha_{i_1}|\alpha_{i_2}) \le S(\alpha_{i_1}|\beta_{j_2}) + S(\beta_{j_2}|\beta_{j_1}) + S(\beta_{j_1}|\alpha_{i_2}).$$

Definitions and necessary formulae for following proof can be found in Cover and Thomas (1991) and summarized in **Appendix 7.2**.

Proof. On the assumption there exists a complete probability table, $P(\alpha_{i_1}, \alpha_{i_2}, \beta_{j_1}, \beta_{j_2})$, a joint entropy can be written as

$$S(\alpha_{i_1}, \alpha_{i_2}, \beta_{j_1}, \beta_{j_2}) = -\sum_{i_1, i_2, j_1, j_2} P(\alpha_{i_1}, \alpha_{i_2}, \beta_{j_1}, \beta_{j_2}) \log P(\alpha_{i_1}, \alpha_{i_2}, \beta_{j_1}, \beta_{j_2})$$
$$= S(\alpha_{i_2} \cap \beta_{j_1} \cap \beta_{j_2}, \alpha_{i_1})$$
$$= S(\alpha_{i_2} \cap \alpha_{i_1}, \beta_{j_1} \cap \beta_{j_2}).$$

Using the chain rule of entropy sequentially, one can get

$$S(\alpha_{i_2} \cap \beta_{j_1} \cap \beta_{j_2}, \alpha_{i_1}) = S(\alpha_{i_1}|\alpha_{i_2}, \beta_{j_1}, \beta_{j_2}) + S(\alpha_{i_2} \cap \beta_{j_1}, \beta_{j_2})$$

$$= S(\alpha_{i_1}|\alpha_{i_2}, \beta_{j_1}, \beta_{j_2}) + S(\beta_{j_2}|\alpha_{i_2} \cap \beta_{j_1}) + S(\alpha_{i_2}, \beta_{j_1})$$

$$= S(\alpha_{i_1}|\alpha_{i_2}, \beta_{j_1}, \beta_{j_2}) + S(\beta_{j_2}|\alpha_{i_2}, \beta_{j_1}) + S(\beta_{j_1}|\alpha_{i_2}) + S(\alpha_{i_2}).$$

On the other hand, this joined entropy satisfies

$$S(\alpha_{i_2} \cap \alpha_{i_1}, \beta_{j_1} \cap \beta_{j_2}) = S(\alpha_{i_2}, \alpha_{i_1}) + S(\beta_{j_1}, \beta_{j_2}|\alpha_{i_2}, \alpha_{i_1}),$$

$$\geq S(\alpha_{i_2}, \alpha_{i_1})$$

$$= S(\alpha_{i_2}) + S(\alpha_{i_1}|\alpha_{i_2}).$$

Inequality follows from nonnegativity of entropy. From the property of the probability measure in the probability space,

$$^\forall\alpha_1, ^\forall\alpha_2 \in \mathcal{V}_A, P(\alpha_1 \cap \alpha_2) \leq P(\alpha_1),$$

and the definition of joint entropy, the inequalities

$$S(\alpha_{i_1}|\alpha_{i_2}, \beta_{j_1}, \beta_{j_2}) \leq S(\alpha_{i_1}|\beta_{j_2}),$$

$$S(\beta_{j_2}|\alpha_{i_2}, \beta_{j_1}) \leq S(\beta_{j_2}|\beta_{j_1}),$$

follow. Then Bell's inequality is proved.

The necessary condition for Bell's inequality, the existence of the complete probability table $P(\alpha_{i_1}, \alpha_{i_2}, \beta_{j_1}, \beta_{j_2})$, corresponds to local realism in the physical terminology. Here, we give an example where Bell's inequality is not maintained.

5.3 Definition 5.2 (Stochastic double codeword-transfer experiment without a complete probability table)

Here, the number of codewords in the set is $m = 2$ for simplicity.

    1) Set $m = 2$ in **Definition 5.1-1** for two sets of codewords such as

$$\mathcal{W}_A = \{\alpha_1, \alpha_2\},$$

$$\mathcal{W}_B = \{\beta_1, \beta_2\},$$

$$\mathcal{W}_A \otimes \mathcal{W}_B = \{\alpha_1, \alpha_2, \beta_1, \beta_2\} = \{\gamma_1, \gamma_2, \gamma_3, \gamma_4\}$$

and for state vectors as

$$\mathcal{V}_A \ni \psi_A(\theta_\alpha) = (\cos\theta_\alpha, \sin\theta_\alpha),$$

$$\mathcal{V}_B \ni \psi_B(\theta_\beta) = (\cos\theta_\beta, \sin\theta_\beta),$$

where $0 \leq \theta_\alpha, \theta_\beta \leq \pi$. This parametrization configures an example of **Theorem 3.1**.

    2) The same as **Definition 5.1-2**.

    3) Encoding is performed using following probabilistic function:

$$\mu_{dc}(\gamma_i, \gamma_j) = |\gamma_i|^2 |\gamma_j|^2.$$

    4) $\mathcal{B}$ and $\mathcal{C}$ select for measurement one of the elements (codewords) in $\mathcal{W}_A$ or $\mathcal{W}_B$, independently. Before measurement, $\mathcal{B}(\mathcal{C})$ rotates a detector angle up to θ_b (θ_c). Neither knows the rotating angle of the other. $\mathcal{B}$ and $\mathcal{C}$ correct this rotation angle after completing all measurements. This rotation does not affect the error of the measurement, owing to **Theorem 4.1**.

        (a) If state vectors $\{\alpha_i\}$ and $\{\beta_i\}$ exist locally before the measurement for $\mathcal{B}$, the probability that $\mathcal{B}$ may obtain each codeword can be obtained after rotation as

$$\psi_\gamma(\theta_\gamma) \to \psi_\gamma(\theta_\gamma - \theta_b) = R(\theta_b)\psi_\gamma,$$

where $R(\theta)$ is a rotation matrix, $\psi_\gamma \in \mathcal{V}_\alpha \oplus \mathcal{V}_\beta$, and $\theta_\gamma = \theta_\alpha$ or θ_β depending on ψ_γ. The probability for $\mathcal{C}$ is similar to the above. In this case we do not observe any violation of Bell's inequality since we can prepare the complete probability table.

        (b) Suppose the angles θ_α and θ_β are not fixed before measurement and are fixed when $\mathcal{B}$ or $\mathcal{C}$ measure the code from $\mathcal{W}_A$ or $\mathcal{W}_B$ and the probability measure μ_{dc} depends on the result of their decision. Moreover we require that the probability measure does not follow the *functional composition condition* (FUNC) (Flori, 2013). In a context of the report, the FUNC is a requirement for any function f as arithmetic operations on vectors and real numbers as

$$\forall \mathrm{f}, \mu_{dc}(f(\psi_i, \psi_j), \psi_k) = f\big(\mu_{dc}(\psi_i, \psi_k), \mu_{dc}(\psi_j, \psi_k)\big).$$

A function f in l.h.s. maps real numbers to a real number. On the other hand, f in r.h.s from vectors to a real number. Here we consider a natural isomorphism between real numbers and vectors in operations of addition, subtraction, and (scalar) product, and represented the same symbol f. For a current example, the probability measure does not satisfy the FUNC, for example, as

$$\mu_{dc}(\alpha_1 + \alpha_2, \psi_k) = |\alpha_1 + \alpha_2|^2 |\psi_k|^2,$$

$$\mu_{dc}(\alpha_1, \psi_k) + \mu_{dc}(\alpha_2, \psi_k) = (|\alpha_1|^2 + |\alpha_2|^2)|\psi_k|^2,$$

$$\neq \mu_{dc}(\alpha_1 + \alpha_2, \psi_k).$$

Suppose $\mathcal{C}$ obtains α_1 (α_2). The angle θ_α for $\mathcal{B}$ is fixed as $\theta_\alpha = \theta_c$ ($\theta_\alpha = \pi/2 + \theta_c$), i.e., the probability table is now situation-dependent. The state vectors for $\mathcal{B}$ are now

$$\psi_A = \begin{pmatrix} \psi_A(\theta_c - \theta_b) & \mathcal{C}\, obtained\, \alpha_1 \\ \psi_A(\theta_c + 1/2 - \theta_b) & \mathcal{C}\, obtained\, \alpha_2, \\ \psi_A(\theta_\alpha - \theta_b) & \mathcal{C}\, obtained\, \beta_i. \end{pmatrix}$$

If $\mathcal{B}$ decided to measure a codeword from a set $\mathcal{W}_\beta$, nothing would happen. On the other hand, if $\mathcal{B}$ decided to measure a codeword from the same set as $\mathcal{C}$, then

$$\mu_{dc}(\alpha_1, \alpha_1 + \alpha_2)$$
$$= |\cos(\theta_c - \theta_b)\cos(\theta_\alpha - \theta_c) - \sin(\theta_c - \theta_b)\sin(\theta_\alpha - \theta_c)|^2$$
$$= \cos^2(\theta_\alpha - \theta_b),$$
$$\mu_{dc}(\alpha_2, \alpha_1 + \alpha_2)$$
$$= |\sin(\theta_c - \theta_b)\cos(\theta_\alpha - \theta_c) + \cos(\theta_c - \theta_b)\sin(\theta_\alpha - \theta_c)|^2$$
$$= \sin^2(\theta_\alpha - \theta_b).$$

Again the probability to obtain one of the α can be calculated using only local parameters on $\mathcal{B}$. In both cases, $\mathcal{B}$ can obtain a set of codewords that $\mathcal{A}$ intended to send. The probability table is situation-dependent and there is a possibility that Bell's inequality will be violated.

5) 6) The same as the **Definition 5.1**.

It is proved that the Kochen–Specker theorem is incompatible the FUNC (Flori, 2013). Above stochastic double codeword-transfer experiment is a model of the QM violating the FUNC to incorporate the Kochen–Specker theorem. In order to confirm a violation of Bell's inequality, it is tested numerically according to the above example. A correlation between measured codewords independently obtained by $\mathcal{B}$ and $\mathcal{C}$ is defined mimically like CHSH (Clauser, Horne, Shimony, & Holt, 1969) as

$$\Delta S = S(\alpha_{i_1}|\alpha_{i_2}) - \big(S(\alpha_{i_1}|\beta_{j_2}) + S(\beta_{j_2}|\beta_{j_1}) + S(\beta_{j_1}|\alpha_{i_2})\big).$$

According to the results of **Theorem 5.1**, ΔS is bounded by negative values when the complete probability table exists. If the theory is based on local realism, one can always prepare the complete table to observe codewords for both $\mathcal{B}$ and $\mathcal{C}$. In order to design the experiment that gives a stronger correlation ($\Delta S > 0$), one has to employ a rule for choosing the probability table, i.e., a choice that cannot be determined locally. Moreover, the rule must also satisfy requirements from special relativity, if one would like to interpret as physical law. The stochastic double codeword-transfer experiment defined by **Definition 5.2** is an example of such a rule. Under **Definition 5.2-4b**, for instance, $\mathcal{B}$ cannot know the probability table he is using because it depends on $\mathcal{C}$'s decision, and that cannot be known by $\mathcal{B}$. This lack of the complete probability table is deeply related to the Kochen–Specker theorem (KST). The KST insists of absence of a complete set of physical quantities without measurements in QM, and corresponds exactly to lack of the complete probability table introduced in **Definition 5.2**. Moreover, if we look at only $\mathcal{B}$'s results, we cannot extract any information about $\mathcal{C}$'s choices and results; that means $\mathcal{C}$'s information cannot transferred to $\mathcal{B}$ immediately, which is a requirement from special relativity. This coexistence of nonlocality and special relativity is realized by the rule of **Definition 5.2-4b** of the stochastic double codeword-transfer experiment. The probability tables, $\mu_B(\alpha_1)$ and $\mu_B(\alpha_2)$, include θ_c, though these are tables for $\mathcal{B}$, which is called "*entanglement*". However, $\mathcal{B}$ cannot extract a value of θ_c because θ_c appears only in phase of the unitary transformation and disappears after reaching the average. Violation of Bell's inequality can be judged by checking whether the correlation ΔS is greater than zero or not. Numerical results with employing rule of **Definition 5.2-4b** are calculated and shown in Fig. 3. One can clearly see the violation of Bell's inequality in some parameter regions.

This trick can be implemented because the probability table represents probability amplitude. For example, if $\mathcal{B}$ decided to measure a codeword from the same set as $\mathcal{C}$, say $\mathcal{V}_A$, state vectors for $\mathcal{B}$ is superposition of two possible sates depending on the result of measurement of $\mathcal{C}$ such as

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = R(\theta_b) \cdot \begin{pmatrix} \cos\theta_\alpha \\ \sin\theta_\alpha \end{pmatrix} |_{\theta_\alpha \to \theta_c} \times (1,0) \cdot R(\theta_c) \cdot \begin{pmatrix} \cos\theta_\alpha \\ \sin\theta_\alpha \end{pmatrix}$$

$$+ R(\theta_b) \cdot \begin{pmatrix} \cos\theta_\alpha \\ \sin\theta_\alpha \end{pmatrix} |_{\theta_\alpha \to \theta_c + \pi/2} \times (0,1) \cdot R(\theta_c) \cdot \begin{pmatrix} \cos\theta_\alpha \\ \sin\theta_\alpha \end{pmatrix},$$

$$= \begin{pmatrix} \cos(\theta_c - \theta_b)\cos(\theta_\alpha - \theta_c) - \sin(\theta_c - \theta_b)\sin(\theta_\alpha - \theta_c) \\ \sin(\theta_c - \theta_b)\cos(\theta_\alpha - \theta_c) - \sin(\theta_c - \theta_b)\cos(\theta_\alpha - \theta_c) \end{pmatrix},$$

where $R(\theta)$ is a rotation matrix. Then the probabilities of $\mu_{dc}(\alpha_i)$ are obtained as in **Definition 5.2-4b**. The nonlocal realism is induced by squaring the state vector after superposition of two possible states. This is another example outlining the advantage of the method using probability amplitude.
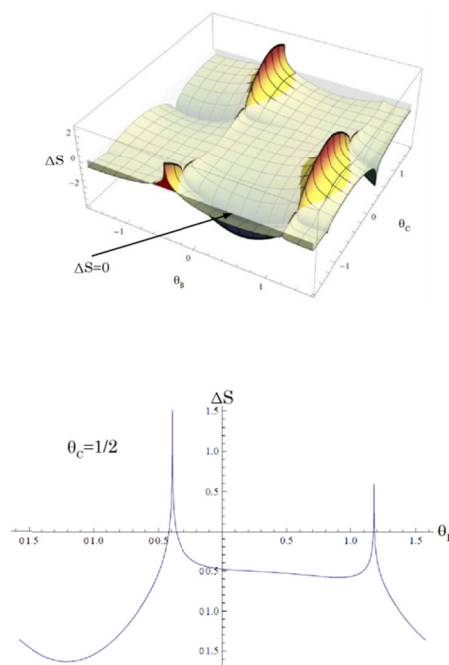

Figure 3. ΔS for the stochastic double codeword-transfer experiment without a complete probability table. It can be seen that Bell's inequality is broken in a part of parameter region

6. Summary

In this report, a basic definition of quantum mechanics concerning its static aspect is discussed. Here, a dynamic aspect of quantum mechanics is not treated. The simple codeword-transfer experiment which satisfies the definition of quantum mechanics is designed to investigate some of it aspects. Then it is proved that a method using probability amplitude gives the minimum error for the physical observables using information theory. Also, it is shown that the size of the error doesn't depend on parametrization of the coding. Nonlocal realism is one of the most essential parts of the nature of quantum mechanics. It is shown that quantum mechanics defined here can include nonlocal realism for the double codeword-transfer experiment introduced by extending a codeword-transfer experiment, above. We showed that the quantum mechanics defined here can violate Bell's inequality, thanks to the property of the probability amplitude. Here key property is that the probability amplitude has additional degree of freedom, such as a $U(1)$ phase factor for instance, over the probability density. Local experiments can measure only those based on the probability density and have noway to measure the phase factor itself. Effects from the phase appeared only after comparing two or more nonlocal observables.

Thus those nonlocal correlation stronger than the classical informational theory can be obtained without braking the special relativity.

In conclusion, the probability amplitude rather than probability density gives the minimum and independent mean-square errors from parametrization. Moreover, it allows one to obtain nontrivial and nonlocal correlation on two independent measurements which violate Bell's inequality incorporate with the Kochen–Specker theorem. It is worth pointing out that nonlocal realism can be realized without any complex-number valued amplitude here. The complex-number valued amplitude could be one of convenient representations for quantum mechanics, but not an indispensable ingredient of that. For example, the representation of $SU(2)$ (a rotation angle) can be used instead of that of $U(1)$ (a complex phase), which share the same Lie-algebra locally.

Acknowledgments

References

Bell, J. S. (1964). On the einstein-podolsky-rosen paradox. *Physics, 1*(3), 195-200.

Born, M. (1926). Quantenmechanik der stoßvorgänge. *Zeitschrift für Physik, 38*(11-12), 803-827. http://dx.doi.org/1926. 10.1007/BF01397184

Braunstein, S. L., & Caves, C. M. (1988). Information-theoretic Bell inequalities. *Physical review letters, 61*(6), 662.

Clauser, J. F., Horne, M. A., Shimony, A., & Holt, R. A. (1969). Proposed experiment to test local hidden-variable theories. *Physical review letters, 23*(15), 880.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York: Wiley.

Cram`er, H. (1946). *Mathematical Method of Statisticas*. Princeton University Press.

Fisher, R. A. (1922). On the dominance ratio. *Proc. R. Soc. Edinb., 42,* 321–341.

Flori, C. (2013). A first course in topos quantum theory. Lecture Notes in Physics. Springer, Berlin.

Kurihara, Y. (2013). Classical information theoretic view of physical measurements and generalized uncertainty relations. *Journal of Theoretical and Applied Physics, 7*(1), 28.

Rao, C. R. (1945). Information and accuracy obtainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc., 37*, 81.

Wheeler, J. A. (1990). *Information, physics, quantum: The search for links*. Physics Dept., University of Texas.

Wheeler, J. A. (1991). Sakharov revisited:"It from Bit". In *Proceedings of the First International AD Sakharov Memorial Conference on Physics, Moscow, USSR, 1991, May 27-31, ed M Man'ko, Nova Science Publishers, Commack, NY*.

Wootters, W. K. (1980). Information is maximised in photon polarization measurements. In A. R. Marlow (Ed.), *Quantum Theory and Gravitation* (pp. 13–26). Bostosn: Academic Press.

Wootters, W. K. (1981). Statistical distance and Hilbert space. *Physical Review D*, *23*(2), 357.

Wootters, W. K. (2013). Communicating through probabilities: Does quantum theory optimize the transfer of information? *Entropy, 15*(8), 3130-3147.

Appendix

7.1 Classical estimation theory

We define terms associated with physical measurement according to classical estimation theory (Cover & Thomas, 1991) as follows. Let $\mathcal{X}$ be a random variable for a given physical system described by the N-tuple $\theta = \{\theta_1, \cdots, \theta_N\}$, where θ_i is the *i th physical parameter*. The set of all possible values of $\theta_i \in \mathbb{R}$, denoted by Θ, is called the *parameter set*. The random variable $\mathcal{X}$ is distributed according to the probability density function $f(x; \theta) \geq 0$, which is normalized as $\int_{x \in \Omega} dx f(x; \theta) = 1$, where $x \in \mathbb{R}$ is one possible value of the whole event $(= \Omega)$. For physical applications, we introduce the *probability amplitude* defined by

$$|\omega(\mathrm{x}; \theta)|^2 = f(x; \theta).$$

A part of *experimental apparatus* is assumed to output numbers distributed according to the probability density. Any resulting set of numbers $\mathcal{X}_n = \{x_1, \cdots, x_n\}$, drawn independently and identically distributed (i.i.d.), is called the *experimental data*. The estimate of the physical parameter is called a *measurement*. Because experimental data are i.i.d., the corresponding probability density function can be expressed as a product:

$$f(\mathcal{X}_n; \theta) = \prod_{j=1}^{n} f(x_j; \theta).$$

A function mapping the experimental data to one possible value of the parameter set such as

$$T_i: \mathcal{X}_n \rightarrow \Theta: \{x_1, \cdots, x_n\} \mapsto \tilde{\theta}_i$$

is called an *estimator* for the ith physical parameter, denoted by $T_i(\mathcal{X}_n) = \tilde{\theta}_i$. The *experimental error* in the ith physical parameter is defined as the root mean square error:

$$\varepsilon_i = E[(T_i(\mathcal{X}_n) - \theta_i)^2]^{1/2},$$

where θ_i is the true value of the i th physical parameter. True values of physical parameters are typically unknown, but a mean-square error can be reduced below any desired value by accumulating a sufficiently large amount of experimental data, thanks to the law of large numbers. If the mean value of the experimental error converges to zero in probability, i.e.,

$$\lim_{n \to \infty} E_{\theta_i}[\tilde{\theta}_i - \theta_i] \rightarrow 0 \text{(inprobability)},$$

after accumulation of infinitely many statistics, that estimator is called an *unbiased estimator*. Among such estimators, the one giving the least error is called the *best estimator*.

7.2 Information Theory

For a probability space (Ω, A, P) and probability variable X defined on it, information entropy $S(X)$ is defined as

$$S(X) = -\sum_{x \in \Omega} P(x) \log P(x).$$

$S(X) \geq 0$ immediately follows from $0 \leq P \leq 1$. For two probability variable X, Y whose domains are Ω_x, Ω_y, where $\Omega_x, \Omega_y \subseteq \Omega$, a joint entropy is defined as

$$S(X, Y) = -\sum_{x \in \Omega_x} \sum_{y \in \Omega_y} P(x \cap y) \log P(x \cap y),$$

where $P(x \cap y)$ is a probability to observe x in X and y in Y, simultaneously. A conditional entropy is defined as

$$S(Y|X) = -\sum_{x \in \Omega_x} \sum_{y \in \Omega_y} P(x \cap y) \log P(y|x),$$

Where $P(y|x)$ is conditional probability to observe y in Y when x in X is obtained. On those entropies, following formulae are obtained:

$$S(X, Y) = S(X) + S(Y|X),$$
$$S(X|Y) \leq S(X).$$

Copyrights