Information gain when measuring an unknown qubit

Gunnar Björk

Department of Applied Physics, Royal Institute of Technology (KTH) AlbaNova University Center, SE-106 91 Stockholm, Sweden

E-mail: gbjork@kth.se

Received 7 April 2017, revised 26 September 2017
Accepted for publication 20 October 2017
Published 18 December 2017

Abstract
In quantum information the fundamental information-containing system is the qubit. A measurement of a single qubit can at most yield one classical bit. However, a dichotomous measurement of an unknown qubit will yield much less information about the qubit state. We use Bayesian inference to compute how much information one progressively gets by making successive, individual measurements on an ensemble of identically prepared qubits. Perhaps surprisingly, even if the measurements are arranged so that each measurement yields one classical bit, that is, the two possible measurement outcomes are \textit{a priori} equiprobable, it takes almost a handful of measurements before one has gained one bit of information about the gradually concentrated qubit probability density. We also show that by following a strategy that reaps the maximum information per measurement, we are led to a mutually unbiased basis as our measurement bases. This is a pleasing, although not entirely surprising, result.

Keywords: quantum information, quantum measurement, qubit, mutually unbiased basis, Shannon entropy, Bayesian inference

(Some figures may appear in colour only in the online journal)

1. Introduction
In recent years the field of quantum information has attracted considerable attention. A main driving force has been the concern about personal and national information integrity. We do
not want personal or national information to be read by unwanted and perhaps even unknown parties. At the same time we are all aware of the fact that both nations and individuals spend considerable resources trying obtain such information [1]. This concern has led to the development of quantum key distribution [2, 3], a system of distributing cryptographic keys that rely for its safety on quantum mechanics rather than on the belief that factorisation of large numbers (on which public key cryptographic systems rely on) is computationally hard [4, 5]. However, in 1995 it was shown by Shor [6] that a quantum mechanical computer could factorise a product in a number of steps scaling polynomially rather than exponentially with the number of digits of the product [7]. The latter is what the best, publicly known factorisation algorithms do.

Another driving force is the increased sensitivity quantum measurements offer compared to classical measurement when using the same amount of resources, such as energy. In the most advanced measuring systems constructed so far, such as the Laser Interferometer Gravitational-Wave Observatory, quantum enhanced measurement techniques are already in use [8]. However, to extract the most information from the measurement system and data, proper understanding of statistics and classical information theory and its relation to quantum measurements is needed.

A basic building block of quantum information is the qubit, just as the bit is the basic building block for contemporary information society. However, in spite of the seemingly similar notions of a bit and a qubit, there are fundamental differences between the two. A (classical) bit is a unit of information. A qubit, on the other hand, is not a unit but simply designates a two-state quantum system. A single qubit can be prepared to transfer one bit of classical information, but this is only possible if the the involved parties first agree on a (two-state) basis, and then an equiprobable superposition or mixture between the two basis states is prepared. If no designated basis is chosen \textit{a priori}, then the analogy between the bit and the qubit is incomplete. The reason is that probabilities in classical physics is replaced by complex-valued probability amplitudes in quantum mechanics. This allows for a richer set of possible measurements, and hence outcomes, for a quantum system.

As quantum information is becoming increasingly ubiquitous in physics and informatics, it is important that contemporary physics students learn some (quantum) information theory and that computer science students learn to appreciate the difference between classical bits and quantum qubits early on. Unfortunately, it is not so easy to find such information in quantum mechanics or quantum information textbooks, as they tend to focus on quantum information, without regard to its relation to classical information theory. The present manuscript makes no claim of offering a complete exposition of the difference between the two. However, we do show, in an intentionally pedagogic manner, at least how an unknown bit and qubit, with equiprobable \textit{a priori} state probabilities or probability density, respectively, differ in how much information a measurement will give. With the qubit as an example of a quantum system which allows for many different kind of measurements by choice of measurement basis, the concept of mutually unbiased basis naturally emerges. Hopefully, this will not only inform students of an important concept in contemporary quantum mechanics, but may also pique the students’ interest in quantum information science, a rapidly growing and exciting field.

2. Qubits and the Bloch sphere

In the following we shall assume that the qubit is encoded on a spin 1/2 system. This system should be familiar to most students that have come into contact with quantum mechanics. We
use the eigenstates $\downarrow$ and $\uparrow$ (spin down and spin up) to the spin $z$ observable and associate the states with the logical qubit states $|0\rangle$ and $|1\rangle$, respectively. A pure qubit state $|\psi\rangle$ can then be written

$$|\psi\rangle = \cos(\theta/2)|\uparrow\rangle + \exp(i\phi)\sin(\theta/2)|\downarrow\rangle,$$

where $0 \leq \theta \leq \pi$, $0 \leq \phi < 2\pi$, and $i$ is the imaginary unit. Through $\theta$ and $\phi$ we can associate every qubit to a point on the Bloch sphere [9, 10], see figure 1. Hence, the state $|\uparrow\rangle$ ($|\downarrow\rangle$) lies on the north (south) pole of the sphere.

As the spin of a light particle, a photon, is manifested in terms of the photon’s polarisation, students familiar with optics can alternatively look at the quantum state of the qubit as a polarisation state of a photon. The Bloch sphere is then replaced by the Poincaré sphere, and the states $|\downarrow\rangle$ and $|\uparrow\rangle$ can be interpreted as a left circularly and a right circularly polarised photon, respectively. By analogy, the states $(|\uparrow\rangle \pm |\downarrow\rangle)/\sqrt{2}$ then represent vertical and horizontal linear polarisation, and the states $(|\uparrow\rangle \pm i|\downarrow\rangle)/\sqrt{2}$ represent diagonal and anti-diagonal linear polarisation of a photon.

We shall also assume that we have two protagonists that we call Alice and Bob. Alice will prepare an ensemble of identical, pure qubits specified by $\theta$ and $\phi$. However, she will chose the two angles without telling Bob what values she has chosen. Bob’s task is to try to estimate what values Alice has chosen, through von Neumann measurements.

Since the state of the qubits Alice prepared is completely unknown to Bob, it is reasonable for him to assign a uniform prior probability density $p(\theta, \phi)$ on the Bloch sphere. (If the measurement outcome space is unknown, assigning a prior is somewhat more involved, see [11]. However, a spin qubit measurement results in either spin up or spin down.) Thus, $p(\theta, \phi) = (4\pi)^{-1}$. In this manner the probability that the qubit was prepared by Alice with the parameters $\theta$ and $\phi$ within a small surrounding solid angle $\Omega$ anywhere on the Bloch sphere is equal to $\Omega/(4\pi)$.

To motivate this scenario, suppose Bob wants to characterise a thick, transparent, birefringent material where neither the material thickness nor the material’s fast and slow axes are known. If a photon of a known polarisation (or spin) state is sent through the material, the polarisation state of the photon will in general change. This change will let Bob characterise the material. However, since Bob has no clue of the how the material will affect the photon, he has good reason to assign a uniform probability density as his prior. Thus, ‘Alice’ should not be taken literally as a person, but rather as some procedure to prepare the unknown quantum state, in this case by propagating a known state through an unknown material. Here,
it is perhaps prudent to point out that in order to characterise an unknown material in this way it is not sufficient to use only one input polarisation. If, e.g. the input state is linearly polarised in a direction parallel to either of the fast or slow axis of the material, the polarisation state will not change. This will allow Bob to draw the conclusion about the birefringence axis direction, but not about the birefringence ‘strength’ (for the used wavelength). To get information about the latter one could, e.g. use a diagonally linearly polarised photons and measure the material’s influence on such photons. The material’s influence on different input states will eventually disclose the material’s birefringence properties. The generalisation of such a procedure is called quantum process tomography [12].

Going back to the spin states, it is not difficult to prove that if we have two states $\langle \psi_1 \rangle$ and $|\psi_2\rangle$, represented on the Bloch sphere by the unit vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ through (1) then the absolute squared overlap $|\langle \psi_1 | \psi_2 \rangle|^2$ (the projection probability of $|\psi_1\rangle$ onto $|\psi_2\rangle$ or vice versa) equals $\cos^2(\gamma/2)$, where $\gamma$ is the angle subtended between $\mathbf{v}_1$ and $\mathbf{v}_2$ so that $\mathbf{v}_1 \cdot \mathbf{v}_2 = \cos(\gamma)$. Using a standard trigonometric identity we can now rewrite the absolute squared overlap

$$
|\langle \psi_1 | \psi_2 \rangle|^2 = \cos^2(\gamma/2) = \frac{1 + \cos(\gamma)}{2} = \frac{1 + \mathbf{v}_1 \cdot \mathbf{v}_2}{2}.
$$

The last step allows us to use the Bloch sphere representation exclusively in the following when computing state projection probabilities. From the equation we can immediately see that in order to get the projection probability $1/2$, the two states’ Bloch vectors must be perpendicular ($\gamma = \pm \pi/2$ or $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$).

3. Information and Bayes’ rule

Suppose Bob makes a measurement of the unknown spin qubit along the $z$ direction. It is clear that the von Neumann entropy of the qubit state is zero since the state is pure. However, the (Shannon) entropy of Bob’s measurements is not going to be zero, for only if he knows the qubits’ eigenbasis (i.e. he knows $\theta$ and $\phi$) can he make a measurement that will have a predictable outcome. From the symmetry of the prior it follows that the two possible measurement outcomes $|\uparrow\rangle$ and $|\downarrow\rangle$ are a priori equiprobable. Suppose Bob gets the result $|\uparrow\rangle$. He has now gained one bit of classical information because according to Shannon [13], the information gain is the difference between the measurement outcome’s a priori and a posteriori entropies. The entropy function $H$ is defined

$$
H = -\sum_k p(k) \log p(k),
$$

where $p(k)$ is the probability of outcome $k$, and if the logarithm is taken in basis 2, the result will be in units of bits. The entropy function is a measure of the uncertainty of the outcome $k$. Suppose that we have $n$ equiprobable outcomes, so that the probability of obtaining outcome $k$ is $p(k) = 1/n$. Then, $H$ is a monotonically increasing function with $n$. The larger the uncertainty of the outcome, the larger is the entropy. As is clearly explained in [13], the entropy function as defined in (3) is the unique function that fulfils this condition, that is continuous in the probabilities $p(k)$, and that is additive. The last condition can mathematically be stated as follows: assume that a process is generating successive outcomes $k$ and $l$, where the probability $p(k, l)$ for both $k$ and $l$ to happen is expressed $p(k, l) = p(k)p(l/k)$ where $p(l/k)$ is the conditional probability of outcome $l$ given the outcome $k$. Then additivity means that the entropy of $p(k, l)$ is the sum of the entropy of $p(k)$ and the probability weighted entropy of $p(l/k)$. That is, $H(p[k, l]) = H(p[k]) + \sum_k p(k)H(p[l/k])$. 

\[4\]
In Bob’s case the initial entropy of \( p(k) \) is
\[
- \frac{1}{2} \log_2(1/2) - \frac{1}{2} \log_2(1/2) = 1 \text{ bit}
\] (4)
because \textit{a priori} \( p(|\uparrow\rangle) = p(|\downarrow\rangle) = 1/2 \), and the final, \textit{a posteriori} entropy is
\[
-1 \log_2(1) = 0 \text{ bit},
\] (5)
because \textit{a posteriori} \( p(|\uparrow\rangle) = 1 \) and \( p(|\downarrow\rangle) = 0 \) so through the measurement he has gained
\( 1 - 0 = 1 \) bit of classical information. If he re-measures the same qubit in the same basis Bob can now predict the outcome with certainty. However, 1 bit is by no means how much information Bob has gained about what state the qubit was prepared in by Alice. To compute the latter we need two things: the first is Bayes’ rule for statistical inference \[14\]. It can be written
\[
p(a|b) = \frac{p(b|a)p(a)}{p(b)},
\] (6)
where, e.g. \( p(a) \) is the probability density of observing outcome \( a \) irrespective of the outcome \( b \). Bayes’ rule lets us update the probability density of getting an outcome \( a \), given that a measurement or observation resulted in outcome \( b \).

The other tool we need is the differential entropy function for a continuous probability density \( p(a) \). To this end we replace the sum in (3) by an integral, viz.
\[
h = - \int p(a) \log p(a) \, da.
\] (7)
The differential entropy differs from the Shannon entropy in two important respects. First, it depends on the unit in which \( p(a) \) is expressed. In our case we will express the probability density of \( \theta \) and \( \phi \) over the Bloch sphere in the unit \( \text{sr}^{-1} \). However, if we would have used some other unit, so that instead of using \( p(a) \) we would have used \( p'(ka + m) \), where \( k \) and \( m \) are constants and \( p' \) is the rescaled and translated probability density, then
\[
h[p'(ka + m)] = h[p(a)] + \log[k].
\] However, the information gained from a measurement that transforms an initial probability density \( q(a) \), through the use of Bayes’ theorem, into an updated probability density \( p(a) \) is the difference in differential entropy between the initial and updated densities, viz.
\[
\int p(a) \log p(a) \, da - \int q(a) \log q(a) \, da.
\] (8)
Hence, while the values of the differential entropies depend on what units we chose to express the probability density in, the information gain, i.e. the differential entropy difference between two probability densities, is independent on the unit chosen.

Secondly, the differential entropy can be negative, and it approaches minus infinity as \( p(a) \) approaches a (possibly multi-dimensional) delta function. This makes intuitive sense, because to specify precisely the parameter(s) \( a \) where the delta-like function is non-zero requires a large number of digits. The narrower the probability density gets, the larger is number of digits (=the more information) we need to specify \( a \). Since the information gained about the parameter to be estimated equals the differential entropy difference between an initial (wide) probability density and a delta-like final probability density, the information gain will be positive, just as one intuitively expects.

The entropy of the initial, flat prior probability density \( p(\theta, \phi) = (4\pi)^{-1} \) on the Bloch sphere is
As we have discussed above, the obtained numerical value is irrelevant as such, as it depends on the chosen units. What matters is that the information gained by an observation is the difference between the a priori probability density entropy (9) and the a posteriori probability density entropy. This number is independent on our choice of units. A flowchart over the estimation procedure and the computation of the associated information gain described above can be found in figure 2.

4. Measurements and information gain

Equipped with Bayes’ rule and differential entropy we can now see how much information about $\theta$ and $\phi$ Bob gains from one measurement of one of the qubits prepared by Alice. From

$$h_0 = \int_0^{2\pi} \int_0^\pi \frac{\log_2(4\pi)}{4\pi} \sin(\theta) \, d\theta \, d\phi$$

$$= 2 + \log_2(\pi) \approx 3.6515 \text{ bit.}$$ (9)

**Figure 2.** A flowchart outlining the estimation and information gain procedure.
the spherical symmetry of the problem it does not matter what his first von Neumann measurement is. We will therefore assume that Bob makes a first measurement along the $z$-axis on the Bloch sphere. Such a measurement will ‘project’ the measured qubits onto the states $|\uparrow\rangle$ and $|\downarrow\rangle$. For brevity we will denote these outcomes $+z$ and $-z$, respectively. The probability of measuring $+z$ given the state $|\psi_{qf}\rangle$ is

$$
p_{z,+} = \frac{1}{4\pi} \int_0^{2\pi} \int_0^{\pi} \cos^2(\theta/2) \sin(\theta) \, d\theta \, d\phi = \frac{1}{2}.
$$

Hence, the two mutually exclusive outcomes $+z$ and $-z$ are equiprobable, and the measurement therefore defines exactly one bit of classical information (or reduction of randomness) as discussed above.

Using Bayes’ rule (6), the prior $p(\theta, \phi) = (4\pi)^{-1}$, and the results above, Bob can get the updated qubit probability distribution

$$
p(\theta, \phi|z_+) = \frac{p(z_+|\theta, \phi)p(\theta, \phi)}{p(z_+)} = \frac{\cos^2(\theta/2)}{2\pi}.
$$

From this follows that the post measurement probability is no longer uniform over the sphere, but is displaced towards the northern hemisphere as is expected and is illustrated in figure 3(b).

How much information did Bob gain from this measurement? Naïvely one may guess one bit, since the measurement outcome was binary and equiprobable. However, this is incorrect. The entropy of the a posteriori probability density distribution is

$$
h_1 = -\frac{1}{2\pi} \int_0^{2\pi} \int_0^{\pi} \cos^2(\theta/2) \log_2 \left( \frac{\cos^2(\theta/2)}{2\pi} \right) \sin(\theta) \, d\theta \, d\phi
$$

$$
= \log_2(2\pi) + \frac{1}{2\ln(2)} \approx 3.3728 \text{ bit}.
$$

The information gained about the state preparation, as specified by $p(\theta, \phi|z_+)$, is

$$
h_0 - h_1 = 1 - \frac{1}{2\ln(2)} \approx 0.2787 \text{ bit}.
$$
This result was found by, e.g. Terashima and Ren and Fan [15, 16]. We see that the information gained is far less than a bit of information. What does this mean in practice? It means that we are far from being able to exclude with reasonable certainty even half of the possible outcomes. Given the inferred probability density \((11)\) we have e.g. only a \(3/4\) probability of correctly guessing on which hemisphere, north or south, the measured qubit state was prepared, and we can say nothing at all about \(\phi\). As can be seen from figure 3(b), the probability density is still spread over most of the Bloch sphere. Only the point \(\theta = \pi\) has vanishing density.

A re-measurement of the just measured qubit will yield no additional information, because according to the von Neumann projection postulate, the post-measurement state of the qubit is \(|1\rangle\) (if the qubit was not destroyed by the measurement). Hence, the just measured qubit retains no information about the parameters \(\theta\) and \(\phi\) that specified its preparation.

5. Repeated measurement of one observable

To gain more information about the qubit ensemble state, Bob has to measure another of the identically prepared qubits. If he uses the same observable, the possible outcomes must be the same: he either gets the result \(z_+\) or \(z_-\). The probability for the first outcome, given that the first measurement along the \(z\)-axis gave the result \(z_+\), is equal to

\[
\frac{\int_0^{2\pi} \int_0^{\pi} \cos^4(\theta/2) \sin(\theta) \, d\theta \, d\phi}{2/3}.
\]

(14)

Thus, the complementary outcome \(z_-\) happens with probability \(1/3\). The classical measurement entropy of this second measurement is hence 0.918 bit. The updated probability distribution \(p(\theta, \phi|z_+z_+)\) obtained from Bayes’ rule is

\[
p(\theta, \phi|z_+z_+) = \frac{\cos^2(\theta/2)\cos^2(\theta/2) / (2\pi)}{2/3} = \frac{3\cos^4(\theta/2)}{4\pi},
\]

(15)

and in case of the complementary outcome one gets

\[
p(\theta, \phi|z_-z_+) = \frac{3\sin^2(\theta/2)\cos(\theta/2)^2}{2\pi}.
\]

(16)

As expected, neither density depends on \(\phi\). The two densities are depicted in figures 4(b) and (c).

The average differential entropy of the ensuing updated probability distribution is

\[
h_2^{(c)} = -\frac{1}{3} \int_0^{2\pi} \int_0^{\pi} 3\cos^4(\theta/2) \log\left(\frac{3\cos^4(\theta/2)}{4\pi}\right) \sin(\theta) \, d\theta \, d\phi
\]

\[
- \frac{1}{3} \int_0^{2\pi} \int_0^{\pi} 3\sin^2(\theta/2)\cos(\theta/2)^2
\]

\[
\times \log\left(\frac{3\sin^2(\theta/2)\cos(\theta/2)^2}{2\pi}\right) \sin(\theta) \, d\theta \, d\phi \approx 3.1759 \text{ bit.}
\]

(17)

Thus, the average information gained about \(\theta\) by measuring a second qubit with the same observable is \(h_1 - h_2^{(c)} = 0.1969 \text{ bit}\). Thus, even less information was gained on average by the second measurement.

At this point it is worth mentioning that if Bob gets the measurement result \(z_-\) from his second measurement, then the relative entropy of the ensuing updated probability distribution \(p(\theta, \phi|z_-z_+)\) is approximately 3.4710 bit. That is, the entropy has increased by 0.0982 bit.
from 3.3728 bit. In physical terms this means that the qubit state ‘uncertainty’ has increased through a second measurement. This may seem paradoxical, but is not strange at all. We know from thermodynamics that heat can be transferred from a colder system to a warmer. If the system is large, this is very unlikely to be observed, but for small systems this is certainly a possibility. Consequently, when calculating the mean entropy of interacting systems we include all these unlikely events, weighted by their respective (small) probabilities. The result is almost universally that the average expected entropy decreases because these entropy increasing events have small probabilities.

In a qubit system the same thing happens. Certain possible outcomes increase the estimated differential entropy (or lowers the estimated information gain) but on the average, every subsequent measurement will decrease the entropy. If one studies the measurement outcome sequence \( + - + - + \) one finds that it is only after the second outcome the differential entropy of the estimated probability distribution increases. Already with the third measurement, the entropy is on a decreasing slope (even though it then takes a few measurements to get below the estimated entropy after the very first measurement).

Bob can subsequently go on and make a third measurement of the same observable. In total, there may be four outcomes, given that the first measurement yielded the result \( + \): \( z_+z_+z_+z_+ \), \( z_+z_-z_+z_+ \), \( z_-z_+z_+z_+ \), and \( z_-z_-z_+z_+ \). Following the same methodology as has been used above we find that the result \( z_+z_+ \) the probabilities for \( z_+z_+z_+ \) is 3/4 and for \( z_-z_+z_+ \) is 1/4, while given the result \( z_-z_- \) the probabilities for \( z_+z_+z_+ \) is 1/2 and for \( z_-z_-z_+ \) is 1/2. (On average, the classical measurement entropy of the third measurement is therefore 0.874 bit.) The use of Bayes’ rule leads to the following updated probability distributions:

\[
p(\theta, \phi|z_+z_+) = \frac{\cos^2(\theta/2)}{\pi}, \tag{18}
\]

\[
p(\theta, \phi|z_-z_+) = p(\theta|z_+z_-z_+) = \frac{3\sin^2(\theta/2)\cos^2(\theta/2)}{\pi}, \tag{19}
\]

\[
p(\theta, \phi|z_-z_-) = \frac{3\sin^4(\theta/2}\cos^2(\theta/2)}{\pi}. \tag{20}
\]

The three density functions are illustrated in figures 4(d)–(f).
Computing the weighted average entropy of these updated probability densities (where the weighing factors are 1/2, 1/6, 1/6, and 1/6, respectively) one arrives at an average differential entropy of $h_3^{(3)} = 3.0231$ after three measurements. That is, the average information gained by Bob’s third measurement is a mere $h_3^{(3)} - h_0^{(3)} = 0.1528$. That is, after three measurements Bob has gained less than one bit of information ($h_0^{(3)} - h_2^{(3)} \approx 0.6284$ bit) about $\theta$ and $\phi$.

One can continue to compute the information gained by considering all possible outcomes of $4, 5, \ldots, N$ measurements. The trend is clear, each successive measurement results in a smaller and smaller information gain about $\theta$ and no information about $\phi$.

In the next section we shall show that, as expected, the optimal information gain about the qubit state (that is, about both $\theta$ and $\phi$) is not obtained by measuring along the same axis since no information is obtained about $\phi$. However, in the case Bob is only interested in the spin tilt-angle $\theta$, it is not hard to guess, nor to prove, that repeated $z$-axes measurements are indeed optimal. In figure 5, the marginal conditional probability densities for $\theta$ obtained by integrating the joint density over the parameter $\phi$ are plotted. The differential entropy differences between these one-dimensional densities are identical to those calculated above, although the entropy of each one is reduced by a fixed amount since they don’t account for the uncertainty of the parameter $\phi$.

6. Measuring $\theta$ and $\phi$ optimally

Measuring the identically prepared qubits repeatedly using the same observable is not the optimal way to best estimate the state of the qubit ensemble Alice prepared. Instead, it is better to use either a set of mutually unbiased measurement bases (MUB) [17, 18], a symmetric, informationally complete positive operator valued measurement (SIC-POVM) [19, 20], a collective measurement of several qubits in the ensemble [21], or a sequence of adaptive measurements, where each new measurement is decided based on the measurement...
outcome history [22]. In some cases, depending on how the gain (or cost) of estimating the state correctly (or incorrectly), a collective measurement of the whole ensemble gives no advantage over a sequence of measurements on individual states [23]. Since we want to delineate how much information is gained by each successive measurement, and for simplicity, we will in the following use sequential von Neumann measurements corresponding to observables. In fact, as we shall see, optimising the two measurements following the first will give a MUB basis.

Two observables are mutually unbiased if all their respective eigenvectors \(|e_k\rangle\) and \(|e_l\rangle\) fulfil
\[
|\langle e_k | e_l \rangle|^2 = 1/N \quad \forall \; k, \; l = 1, \ldots, N,
\]
where \(N\) is the Hilbert space dimension [17, 18]. In our case \(N = 2\), and it is not difficult to see that the states
\[
|x_\pm\rangle = (|\uparrow\rangle \pm |\downarrow\rangle)/\sqrt{2}
\]
constitute an unbiased basis relative to \(|\uparrow\rangle\) and \(|\downarrow\rangle\). In terms of the Pauli operators \(\sigma_x\), \(\sigma_y\), and \(\sigma_z\), if the measurement along the \(z\)-axis corresponds to \(\sigma_z\) then the vectors \(|x_+\rangle\) and \(|x_-\rangle\) are the eigenvectors to \(\sigma_z\) with eigenvalues \(\pm 1/2\). The Bloch sphere vectors corresponding to \(|x_+\rangle\) and \(|x_-\rangle\) point along the positive and negative \(x\)-axis, respectively, and the unbiasedness condition (21) together with (2) imply that these Bloch sphere vectors are orthogonal to the \(z\) axis.

Suppose once more that a first measurement is made along the \(z\)-axis and that the outcome was \(\uparrow\), so that the prior probability density is given by (11). We note that the prior is independent of \(f\) so that in choosing the second measurement, without loss of generality, Bob can choose to measure in the \(xz\)-plane. Suppose Bob chooses the projection vectors defined by \(\theta = \alpha, \phi = 0\) and \(\theta = \pi - \alpha, \phi = \pi\) for the measurement outcomes we shall call \(\alpha_+\) and \(\alpha_-\), respectively. With this choice one finds that
\[
p(\alpha_\pm, \theta, \phi) = \left[ \cos^2(\theta/2)\cos^2(\alpha/2) + \sin^2(\theta/2)\sin^2(\alpha/2) \right] \\
\quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \}
\]

Moreover,
\[
p(\alpha_\pm) = \int_{0}^{2\pi} \int_{0}^{\pi} p(\alpha_\pm, \theta, \phi)p(\theta, \phi|x_\pm\rangle\sin(\theta)) \; d\theta \; d\phi \\
= \frac{3 \pm \cos(\alpha)}{6}.
\]

From these relations and from (11) one can compute the average differential entropy Bob would get as a function of the projection angle \(\alpha\). We note that since the prior (11) is independent of \(\phi\), the differential entropy must be symmetric around the the projection angle \(\alpha = \pi/2\). From figure 6 we see that for \(\alpha = 0\) Bob retrieves the result from (17). On the other hand, the minimum differential entropy, resulting in the highest information gain, is obtained for \(\alpha = \pi/2\) and is \(h_1 \approx 3.0942\). We also notice that this choice of projectors coincides with a measurement of the Pauli operator \(\sigma_y\). That is, the measurement that maximises the information gain is a measurement which is unbiased with respect to the first measurement. Thus, given that the second measurement yields the result \(\alpha_+ \equiv x_+\), Bayes’ rule gives the updated probability distribution
The density is depicted in figure 3(c).

The information gained by this second measurement by a MUB is $h_1 - h_2 = 0.2786$. That is, Bob gains the same amount of information by this second, unbiased measurement as by the first. Of course, had the optimal measurement instead given the result $\alpha_- \equiv x_-$, then the updated probability distribution would have been different, the factor $1 + \sin(\theta)\cos(\phi)$ would instead have become $1 - \sin(\theta)\cos(\phi)$, but the information gained would have been identical.

Now assume Bob makes a third measurement. If he follows the same procedure we just have outlined, and computes the average differential entropy obtained as a function of the axis direction, he finds that the optimum direction to measure along, given that he has already measured along the $z$ and $x$ axes, is the $y$ axis. The derivation is straightforward but a bit tedious. Since the result is intuitive we will not write down the whole derivation but concentrate on the results. To minimise the average probability density entropy Bob should measure along the $y$-axis. The measurement results $y_\pm$ are equiprobable. Suppose he obtains the result $y_+$. The conditioned probability density $p(y_+, \theta, \phi) = (1 + \sin(\theta)\sin(\phi))/2$, which together with the prior (25) gives the probability $1/2$ to get the assumed result $y_+$.

The entropy of the post measurement probability density

$$p(\theta, \phi | y_+, x_+ z_+ ) = \cos^2(\theta/2)[1 + \sin(\theta)\cos(\phi)]/[2\pi]$$

that is depicted in figure 3(d) is $h_3 \approx 2.8155$, so that the information gained about $\theta$ and $\phi$ by this measurement is once more $h_1 - h_2 = 0.2787$ bit. The gained information would have been identical if Bob would have obtained the measurement result $y_-$. Hence, the three measurements have given Bob $h_0 - h_3 \approx 0.8360$ bit of information towards estimating $\theta$ and $\phi$, somewhat less than bit of information, but certainly more than the $h_0 - h_3 \approx 0.6284$ bit he got on average by making three subsequent measurements using the same observable to estimate $\theta$ only.
7. Adaptive measurements

So far, the information Bob has gained about the preparation of the qubit ensemble would have been the same, irrespective of the measurement sequence along the $x$-, $y$-, and $z$-axes, and irrespective of the measurement results. In fact, given what we have discovered in section 6 Bob could have chosen to measure along any three mutually orthogonal directions on the Bloch sphere, and irrespective of the order or outcomes of the three measurements he would have obtained 0.8360 bits of information.

However, after having used the three mutually unbiased measurement basis, a good way to proceed is to use the measurement outcomes, and based on the outcome history devise a measurement that reaps maximal information [22]. Given the assumed measurement outcome history $y,x,z$, the a posteriori probability density is symmetric with respect to the exchange of any two Bloch sphere coordinate axis. Hence, the maximum probability density must be found in the direction $(1, 1, 1)/\sqrt{3}$ on the Bloch sphere. This direction is characterised by $\theta = \arccos(1/\sqrt{3}) + \pi/2$ and $\phi = \pi/4$. This is the direction along which Bob is most confident about the probability density, so measurement directions perpendicular to this direction should be the ones where he knows the least and therefore can expect the highest information gain.

This intuitive reasoning turns out to be correct, because if Bob chooses his measurement projection axis for the fourth measurement defined by $\theta$ and $\phi$ and plot the mean differential entropy of the post measurement, inferred probability density he obtains figure 7. The figure is only plotted for $0 \leq \phi \leq \pi$ since the projection axis defined by $(\theta, \phi)$ is equivalent (anti-parallel) to the direction $(\pi - \theta, \phi + \pi)$. The angular coordinate pairs corresponding to the measurement projection direction perpendicular to $(1, 1, 1)/\sqrt{3}$ all result in an average differential entropy of $h_4 = 2.597$ bit. For these measurement projection directions the

Figure 7. The mean differential entropy of the probability density after a fourth measurement is made along the projection direction defined by $\theta$ and $\phi$. The blue line gives the (constant) entropy for the angular coordinates of the measurement directions perpendicular to $(1, 1, 1)/\sqrt{3}$. The red dot indicates the direction $\theta = \arccos(1/\sqrt{3}) + \pi/2$ and $\phi = \pi/4$. The light blue dot indicates the projection direction $\theta = \pi/2$ and $\phi = 3\pi/4$. 
measurement outcome is totally random, and thus the measurement gives the maximum, classical entropy, one bit. Bob can, e.g. use the measurement projection direction $\theta = \arccos(1/\sqrt{3}) + \pi/2$ and $\phi = \pi/4$ (and its antiparallel vector). If the fourth measured qubit is projected onto the state vector corresponding to the latter of these two vectors we call the measurement outcome $w_4$. Using (2) we get

$$p(w_4|\theta, \phi) = \frac{1}{2} \left( 1 - \frac{\sin(\phi)\sin(\theta)}{\sqrt{6}} - \frac{\cos(\phi)\sin(\theta)}{\sqrt{6}} + \frac{1}{\sqrt{3}} \cos(\theta) \right)$$

and $p(w_4) = 1/2$. Thus we get the post measurement probability density

$$p(\theta, \phi|w_+y,x,z) = \frac{\cos^2(\theta/2)}{2\pi} \left( 1 - \frac{\sin(\phi)\sin(\theta)}{\sqrt{6}} - \frac{\cos(\phi)\sin(\theta)}{\sqrt{6}} + \frac{1}{\sqrt{3}} \cos(\theta) \right) \times [1 + \sin(\theta)\cos(\phi)][1 + \sin(\theta)\sin(\phi)].$$

The probability density is plotted in figure 3(e). It has its maximum in the direction $\theta \approx 0.5220$ and $\phi = \pi/4$. Bob gained $h_3 - h_4 \approx 0.2185$ bit of information by this measurement, somewhat less than through each of the first three MUB measurements.

To continue, one would guess that Bob should still measure perpendicularly to the direction where $p(\theta, \phi|w_+y,x,z)$ has its maximum, but at the same time Bob should also make sure that the projection direction is perpendicular to the direction defining $w_4$. This fifth direction is given by $\theta = \pi/2$ and $\phi = 3\pi/4$ (or its antiparallel vector). (A minimisation of the post measurement probability density differential entropy confirms that this measurement projection axis indeed gives the unique, global minimum.) We shall call the projection onto the latter vector as measurement result $u_+$. Using (2) one can readily compute the conditional probability

$$p(u_+|\theta, \phi) = \frac{1}{2} \left( \frac{\sin(\phi)\sin(\theta)}{2\sqrt{2}} + \frac{\cos(\phi)\sin(\theta)}{2\sqrt{2}} \right)$$

and using (28) and (29) we get $p(u_+) = 1/2$. Bayes’ rule now gives us the a posteriori probability density (which we do not write out explicitly due the the equation’s length)

$$p(\theta, \phi|u_+w_+y,x,z) = 2p(\theta, \phi|w_+y,x,z) p(u_+|\theta, \phi).$$

The function is plotted in 3(f). The entropy of this distribution is $h_5 \approx 2.3883$ bit, so the measurement reaped $h_4 - h_5 \approx 0.2087$ bit which is somewhat less than the fourth measurement gave Bob. The maximum of the probability density now is found at $\theta \approx 0.6419$ rad and $\phi \approx 0.1318$ rad, that is, the distribution has moved closer to the $x$-axis due to the projection on the vector closer to the positive $x$- rather than the positive $y$-axis. In total, the five measurements have given Bob $h_0 - h_5 \approx 1.2632$ bit of information, or an average of $\approx 0.2526$ bit per measurement.

8. Discussion

Above, we have used differential entropy to quantify the information gain when making successive measurements of an ensemble of qubits. In the literature one can find several other ways of estimating the state of a measured qubit and the associated ‘information gain’. The simplest, but crudest way is to use the fidelity of the qubit ensemble given to Bob to a target qubit $|\tau\rangle$. The fidelity $F$ between the pure qubits $|\psi\rangle$ and $|\tau\rangle$ is defined $F = |\langle \psi |\tau \rangle|^2$ (although
some authors use the square root of this number) so that $0 \leq F \leq 1$. However, the price for the simplicity of the definition is firstly that any state $\sqrt{F} |\tau\rangle + \sqrt{1 - F} \exp(i\varphi)|\tau_\perp\rangle$, where $|\tau_\perp\rangle$ is a state orthogonal to $|\tau\rangle$, has the same fidelity $F$ to $|\tau\rangle$ irrespective of $\varphi$. Secondly, therefore the fidelity $F$ and fidelity $1 - F$ specifies equally well where on the Bloch sphere the state $|\psi\rangle$ is located. For example, both zero and unit fidelity allows one to predict $|\psi\rangle$ perfectly. Thirdly, the average fidelity $F$ between two randomly chosen pure states is $1/2$. Thus, when measuring the fidelity $1/4$ Bob gets more information about the state $|\psi\rangle$ than if he measures the fidelity $1/2$. Thus, fidelity as a measure of ‘information gain’ is of most use when it is close to unity (or to zero).

A popular method of state estimation is quantum tomography, and to a large extent this procedure follows the measurement strategy outlined in section 6. One measures the qubit ensemble using a number of linearly independent measurement bases (where the mutually unbiased bases represent an optimal choice). One could then in principle use minimisation of the measurement entropy to optimally estimate the state $|\psi\rangle$, but due to the logarithm function in the definition of the entropy this is mathematically difficult. Instead the standard method is to use the maximum likelihood method from statistics, combined with the constraints that the result should fulfil physical boundary conditions and be normalised. Quantum tomography does usually not provide any quantitative value of the information gained through the tomography, but in principle the (multidimensional) likelihood function contains such information. The more peaked it is, the more information gained.

We remark that instead of using differential entropy we could have used some other measures of information gain, e.g. the Kullback–Liebler (K–L) divergence [24]. For an a priori and an a posteriori probability density $q(a)$ and $p(a)$, respectively, the K–L divergence is defined as

$$\int p(a) \log \frac{p(a)}{q(a)} \, da = \int p(a) \log p(a) \, da - \int p(a) \log q(a) \, da.$$  \hspace{1cm} (31)

We see that the K–L divergence is similar, but not identical, to the differential entropy difference (8). However, as we have seen, each measurement results in a rather ‘gentle’ updating of the post measurement probability density in the considered scenario, so that the measurement a priori and a posteriori probabilities will not differ substantially. Therefore, the differential entropy and the K–L divergence will give rather similar numerical values in the considered case. Specifically, the estimated information gain through the first measurement is identical for the two measures since the a priori probability density $q$ in (31) is a constant in this case. Thus, if we use the K–L divergence as a measure of information gain, we would obtain quantitatively slightly different results than when using differential entropy, but the main messages will be the same: with each successive measurement substantially less than one bit of information is reaped and the average obtained information per measurement is non-increasing.

Another information measure of statistical distributions is the Fisher information [25]. When measuring, e.g. along the $z$-axis, the Fisher information quantifies the amount of information that the outcome of the $z$-measurement carries about $\theta$. The Fisher information does not have the unit bit, but depends on the parametrisation of $\theta$ through $z$. However, the Cramér–Rao bound [26, 27] states that the inverse of the Fisher information of a probability density provides a lower bound for the variance of the estimate of $\theta$ through the measurement of $\theta$. Thus, a high Fisher information translates to the possibility of a rather accurate estimate of $\theta$. Moreover, finding an estimation procedure that saturates the Cramér–Rao bound guarantees the procedure’s optimality. Unfortunately, there exists no direct relation between the
Fisher information and the obtained information in bits as they have different units and have different uses.

9. Conclusions

We have used von Neumann measurements on the simplest quantum system, a pure qubit, to show that caution and care must be excised when making or discussing information gain as a consequence of quantum measurements. We have discussed three different measurement strategies on individual members of an ensemble of identically prepared qubits and showed that while each measurement outcome may result in a classical bit of information, the information gained about the preparation of the qubit by using Bayes’ inference rule is much smaller. For the initial measurements it is on the order of a quarter of a bit. Thus, it takes four optimal measurements before one has gained at least one bit of information.

We have also shown that the concept of mutually unbiased basis follows naturally from the requirement of maximal information gain. Irrespective of how Bob chooses his first measurement, the three first optimal measurements will always define a set of unitarily equivalent MUBs. When the mutually unbiased bases are exhausted, then our analysis shows that an adaptive measurement strategy yields more information than repeated measurements in the MUB bases. For example, the differential entropy for the choice $\theta = \pi$ (which represent a measurement along the $z$-axis) in figure 7 is higher than for the optimal directions.

It is also clear from our example that information is relative. Bob cannot, at all, predict the outcome, + or −, of any of the five optimal measurements, so that the five measurement outcomes represent to him five classical bits. Alice, on the other hand, knowing what state she has prepared, can make more educated guesses about the outcomes of Bob’s five measurements if she knows his measurement basis. Should, for instance, any of Bob’s measurement have the prepared state as an eigenstate, then Alice can predict with certainty what this measurement’s outcome will be. Hence, the sequence of the five measurement outcomes will give Alice less than five classical bits of randomness since it is impossible that all Bob’s five measurements will result in two equally probable outcomes from Alice’s perspective.

We also have seen that the average information gained by each successive, optimal measurement is non-increasing, and generally it is decreasing. This is well in line with any measurement process void of systematic errors. If each individual measurement has some intrinsic uncertainty, e.g. quantified by a variance $\sigma$, then from $N$ independent measurements the expected estimated measurement value has the variance $\sigma/\sqrt{N}$. Thus, the expected reduction of the variance, and hence the decrease of the differential entropy, will be smaller and smaller for each successive measurement. Only the MUB bases are such that a measurement in one basis gives no information about the likelihood of obtaining any of the outcomes of a measurement in another MUB basis.

Although the information gain per successive measurement approaches zero as the number of measurements tend to infinity, the accumulated information, contained in the inferred probability density, tends to infinity. This is because there is (at least in principle) no limit to how precisely we can get to know the angles $\theta$ and $\phi$ that define the preparation of the state. As the probability density approaches a delta-function on the Bloch sphere, the differential entropy that initially was finite, tends to negative infinity.

It can be worth noticing that up to the second measurement, the procedure outlined above can be mimicked by a classical system having two degrees of freedom, e.g. two unfair coins, each having an unknown probability (for Bob) to give the outcome ‘tails’ when flipped. However, classical systems that are prepared in a statistical mixture of four definite states can
only be (sensibly) measured using a ‘basis’ defined by these, mutually exclusive outcomes. Hence, a classical system will not give Bob the opportunity to measure in, e.g. mutually unbiased bases. We have seen that MUBs result in a more rapid gain of information about how the system is prepared than doing repeated measurements in the same basis.

We have not discussed the optimal measurement strategy for an ensemble of $N$ identically prepared quantum systems, which, in general but not always, is to make a joint measurement of all $N$ systems. When making joint measurements one can use interference between different qubits. Interference lies at the heart for most quantum information tasks and can often render them effectively parallelised. This is of course not possible if each qubit in the ensemble is measured separately.

A kind of complement to our study is a recent work by Wootters [28] where the goal is to transfer the value of a continuous variable, exemplified by $\theta$, via the outcomes of $N$ identical von Neumann measurements by preparing the state optimally.

Acknowledgments

This investigation was spurred by a discussion with Mr Ömer Bayraktar. The work was supported by the Swedish Research Council (VR) through grant 621-2014-5410 and through its support of the Linneaus Excellence Center ADOPT.

ORCID iDs

Gunnar Björk https://orcid.org/0000-0002-2082-9583

References

[1] Mazzetti M and Sangeroct D E 2013 Tap on Merkel provides peek at Vast Spy Net New York Times
[2] Bennett C H and Brassard G 1984 Proc. IEEE Int. Conf. on Computers, Systems and Signal Processing vol 1 (New York) p 175
[3] Ekert A 1991 Phys. Rev. Lett. 67 661
[4] Diffie W and Hellman M 1976 IEEE Trans. Inf. Theory IT-22 644
[5] Rivest R, Shamir A and Adleman L 1978 Commun. ACM 21 120
[6] Shor P 1997 SIAM J. Comput. 26 1484
[7] Aaronson S 2005 ACM SIGACT News 36 30
[8] Aasi J et al 2013 Nat. Photon. 7 613
[9] Bloch F 1946 Phys. Rev. 70 460
[10] Nielsen M A and Chuang I L 2004 Quantum Computation and Quantum Information (Cambridge : Cambridge University Press) pp 13
[11] Jaynes E T 2003 Probability Theory: The Logic of Science (Cambridge: Cambridge University Press) pp 372–82
[12] D’Ariano G M, Paris M G A and Sacchi M F 2003 Adv. Imaging Electron Phys. 128 205
[13] Shannon C and Weaver W 1949 A Mathematical Theory of Information (Champaign, IL: University of Illinois Press) pp 49–58
[14] Papoulis A 1984 Probability, Random Variables, and Stochastic Processes 2nd edn (New York: McGraw-Hill) pp 78–81, 112–114
[15] Terashima H 2011 Phys. Rev. A 83 032111
[16] Ren X-J and Fan H 2014 J. Phys. A: Math. Theor. 47 305302
[17] Ivanović I D 1981 J. Phys. A: Math. Gen. 14 3241
[18] Wootters W K and Fields B D 1989 Ann. Phys. 191 363
[19] Renes J M, Blume-Kohout R, Scott A J and Caves C 2004 J. Math. Phys. 45 2171
[20] Appleby D M 2005 J. Math. Phys. 46 052107
[21] Massar S and Popescu S 1995 *Phys. Rev. Lett.* **74** 1259
[22] Wiseman H M, Berry D W, Bartlett S D, Higgins B L and Pryde G J 2009 *IEEE J. Sel. Top. Quantum Electron.* **15** 1661
[23] Brody D and Meister B 1996 *Phys. Rev. Lett.* **76** 1
[24] Kullback S 1959 *Information Theory and Statistics* (New York: Wiley)
[25] Fisher R A 1925 *Math. Proc. Camb. Phil. Soc.* **22** 700
[26] Cramér H 1946 *Mathematical Methods of Statistics* (Princeton, NJ: Princeton University Press)
[27] Rao C R 1945 *Bull. Calcutta Math. Soc.* **37** 81
[28] Wootters W K 2013 *Entropy* **15** 3130