Quantum Computation **(CMU 18-859BB, Fall 2015)**

Lecture 18: Quantum Information Theory and Holevo's Bound

November 10, 2015

Lecturer: John Wright *Scribe: Nicolas Resch*

1 Question

In today's lecture, we will try to answer the following question:

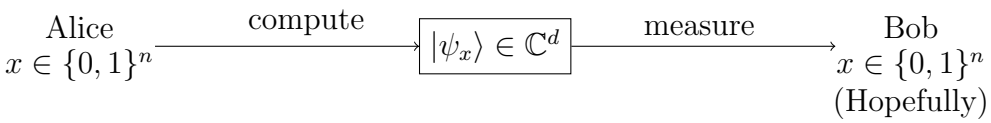
How many bits are in a qubit?

What's the answer? Well, we'll see... In reality, there's more than one way to answer this question. So we'll have to be a little bit more precise if we want to say anything interesting. In this lecture, we will motivate and formally state Holevo's Theorem [Hol73], which does a good job of answering our question.

We begin by considering the following scenario. We have two parties Alice and Bob, and Alice has a string $x \in \{0,1\}^n$ that she would like to transmit to Bob. Classically, we would do the following: Alice would store the string in some shared memory (say, a RAM), and then Bob would read the string from the RAM.

The above scheme works assuming the RAM is big enough, i.e. it can store at least n bits.

Now, since this is a quantum computation course, it is natural to ask if we can do better quantumly, perhaps with respect to the number n? The previous scheme looks as follows in the quantum setting:



In the above, Alice does some sort of computation to create the state $|\psi_x\rangle \in \mathbb{C}^d$, which depends on her input x. The scheme works only if $d \geq 2^n$. Otherwise there will exist non-orthogonal vectors $|\psi_x\rangle$ and $|\psi_y\rangle$ for $x \neq y$. In the previous lecture, we showed that we are only able to discriminate quantum states with probability 1 if they are orthogonal, so we cannot have this if we want Bob to be guaranteed to recover the correct string x. Since $d \geq 2^n$, we require n qubits.

So, it seems like we've answered our question: we need n qubits to represent n bits. Are we done? Well, we can say a little more.

2 More General Scenario

First of all, it might happen that we don't actually need n bits to represent x, even in the classical setting. As a very simple case, it could very well happen that Bob already knows x, so Alice needs to store 0 bits! To deal with this, we need some way to quantify how much Bob knows about Alice's input x. And to do this, we will need to quantify how much uncertainty there is about Alice's input. We will use the following uncertainty model:

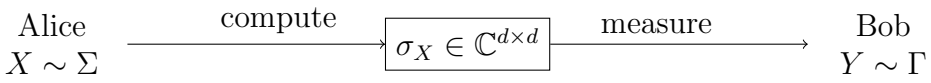
$$\boxed{\text{Alice samples random message } x \in \{0,1\}^n \text{ with probability } p(x).}$$

If all the $p(x)$'s are positive, then we still need n bits to encode all the possible options classically and, just as before, we require n qubits in the quantum setting. But, it is possible that we could do better.

Let's now try to turn our question around slighty. We will now think of d, the dimensionality of the quantum state prepared by Alice, as fixed in advance. Our question will now be: *How much information can Bob learn about x?*

Before continuing, we remark that Alice need not send a vector. As we've seen in recent lectures, quantum states are in the most general case represented by a density matrix, so Alice may send $\sigma_x \in \mathbb{C}^{d\times d}$, a mixed state. Also, Bob can perform the more general quantum measurements that we discussed recently. Hence, the scenario we will analyze is the following:[1]

- Alice samples $X \in \Sigma \subseteq \{0,1\}^n$, where $X = x$ with probability $p(x)$.
- Alice sends $\sigma_X \in \mathbb{C}^{d\times d}$.
- Bob picks POVM's $\{E_y\}_{y\in\Gamma}$, where $\Gamma \subseteq \{0,1\}^n$.
- Bob measures σ_X, and receives output "$Y \in \Gamma$", where $Y = y$ given $X = x$ with probability $\mathrm{tr}(E_y\sigma_x)$.
- Bob tries to infer X from Y.



The most natural thing to do is to put $\Gamma = \Sigma$, and if Bob observes "y" he guesses "y". This is essentially without loss of generality, but we will still consider the situation where Σ and Γ may be different.

[1] We now think of Alice's input, and therefore Bob's measurement, as random variables. Hence the capitalization.

3 Analysis

What's Bob's perspective in this game? He sees σ_x with probability $p(x)$, and he's tasked with discriminating between the different σ_x's. This is very reminiscent of the state discrimination problem we saw last time. However, in this case, we can try to come up with σ_x's that are very easy to discriminate from each other (recall that Alice and Bob are working together).

Bob sees the mixed state

$$\begin{cases} \sigma_{x_1} & \text{with prob. } p(x_1), \\ \sigma_{x_2} & \text{with prob. } p(x_2), \\ & \vdots \end{cases}$$
$$\equiv \sum_{x \in \Sigma} p(x)\sigma_x =: \rho_B.$$

Since each σ_x is a probability distribution over pure states, ρ_B is itself a probability distribution over probability distributions of pure states. We do therefore obtain a probability distribution over pure states, as one would expect.

Alice sees

$$\begin{cases} |x_1\rangle & \text{with prob. } p(x_1), \\ |x_2\rangle & \text{with prob. } p(x_2), \\ & \vdots \end{cases}$$
$$\equiv \sum_{x \in \Sigma} p(x)|x\rangle\langle x| =: \rho_A.$$

Note that ρ_A is precisely the diagonal matrix whose diagonal entries are given by $p(x_1), p(x_2), \ldots$.

To determine the joint state, note that the parties see $|x\rangle\langle x| \otimes \sigma_x$ with probability $p(x)$. Hence, the joint mixed system is

$$\rho := \sum_{x \in \Sigma} p(x)|x\rangle\langle x| \otimes \sigma_x.$$

Recalling an earlier Piazza post, one can indeed verify that $\rho_B = \mathrm{tr}_A(\rho)$ and $\rho_B = \mathrm{tr}_B(\rho)$.

Now that we're dealing with random variables and knowledge, we'll turn to:

4 Classical Information Theory

For further reading on this subject, see [CT12]. In classical information theory, we typically have some random variable X distributed according to P on some set Σ. The most basic question one can ask is:

How much information do you learn from seeing X?

Example 4.1. *Suppose that* $\Sigma = \{0,1\}^n$.

- *If* P *is the uniform distribution, then one gets* n *bits of info from seeing* X.
- *If* P *has all its probability on a single string* $x_0 \in \{0,1\}^n$, *i.e.* $X = x_0$ *with probability 1 and* $X = x$ *with probability 0 for all* $x \neq x_0$, *then we get 0 bits of information from seeing* X.

We can formalize this with the following definition:

Definition 4.2 (Shannon Entropy)**.** The *shannon entropy* of a random variable X distributed on a set Σ is

$$H(X) = \sum_{x \in \Sigma} p(x) \log \frac{1}{p(x)},$$

where $p(x) = \mathbf{Pr}[X = x]$.

Example 4.3. *Let's verify that this definition matches our intuition, at least for the previous two examples.*

- *If* X *is uniform, i.e.* $p(x) = 1/2^n$ *for all* $x \in \{0,1\}^n$, *then*

$$H(X) = \sum_{x \in \{0,1\}^n} \frac{1}{2^n} \log\left(\frac{1}{1/2^n}\right) = 2^n \frac{1}{2^n} \log(2^n) = n.$$

- *If* X *has all its probability mass on a single string* x_0, *then*

$$H(X) = 1 \cdot \log(1/1) + (2^n - 1) \cdot 0 \cdot \log(1/0) = 0 + 0 = 0,$$

where we define

$$0 \log(1/0) := \lim_{x \to 0+} x \log(1/x) = 0.$$

We record here a couple important properties of the shannon entropy function:

- $0 \leq H(X) \leq \log |\Sigma|$.
- H is concave.

So, returning to our earlier scenario, the largest amount of information Bob could hope to learn is $H(X)$. How much does he actually learn? We have two correlated random variables X and Y. We want to know how much knowing Y tells us about X.

In general, if we have random variables X and Y supported on the sets Σ and Γ respectively with joint distribution $P(x,y) = \mathbf{Pr}[X = x, Y = y]$, we have

$$H(X,Y) = \sum_{x \in \Sigma, y \in \Gamma} P(x,y) \log \frac{1}{P(x,y)}.$$

Example 4.4. *Let* $\Sigma = \Gamma = \{0,1\}^n$.

- *Suppose* X *and* Y *are independent, uniform random variables. Then* (X,Y) *is a random* $2n$*-bit string. So* $H(X,Y) = 2n$.
- *Suppose* X *is a uniform random variable and* $Y = X$. *Then* Y *is also a uniform random variable. However,* (X,Y) *is basically a random* n*-bit string. So* $H(X,Y) = n$.

In general, we note that if X and Y are independent, then $H(X,Y) = H(X) + H(Y)$. This seems reasonable, as seeing one of the random variables tells us nothing about the other, so seeing half of the pair (X,Y) only decreases tells us the shannon entropy of the random variable that we observe, but the other random variable still has all of its entropy.

Conversely, if X and Y perfectly correlated, then $H(X,Y) = H(X) = H(Y)$. Indeed, seeing half of the pair (X,Y) immediately tells us what the other half of the pair is, so the amount of entropy in the pair is the same as the amount of entropy in the random variables themselves.

We can formalize the notion of "how much does seeing one random variable tell me about the other" as follows:

Definition 4.5 (Mutual Information)**.** The *mutual information* $I(X;Y)$ between two random variables X and Y is

$$I(X;Y) = H(X) + H(Y) - H(X,Y).$$

This is supposed to represent the amount of information you learn about X from knowing what Y is. Since the definition is symmetric in X and Y, it also represents the amount of information you learn about Y from knowing X.

Example 4.6. *Let's return to our earlier examples.*

- *If* X *and* Y *are independent then we see that*

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = H(X,Y) - H(X,Y) = 0.$$

  This intuitively makes sense, as seeing X *tells us nothing about* Y *(and vice versa) since* X *and* Y *are independent.*

- *If* X *and* Y *are perfectly correlated, then*

$$I(X;Y) = H(X) = H(Y).$$

  In this case, seeing X *tells us everything there is to know about* Y *(and vice versa), so the mutual information between the random variables is a large as possible.*

The symmetry of this definition might be a bit surprising. However, the following example might help explain exactly why this should be the case.

Example 4.7. *Suppose* $\Sigma = \{0,1\}^n$ *and* $\Gamma = \{0,1\}^{2n}$. *Let* Y *be uniformly random on the set* Γ, *and let* $X = (Y_1, \ldots, Y_n)$, *i.e. it is the first* n *bits of* Y. *Then*

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = n + 2n - 2n = n.$$

This makes sense regardless of the way that we look at it.

- *Suppose I know* X. *Then, I know the first* n *bits of* Y, *i.e. I have* n *bits of information about* Y.
- *Suppose I know* Y. *Then, I know all of* X, *and since* X *is an* n*-bit string, I have* n *bits of information about* X.

5 Quantum Information Theory

With this understanding of classical information theory, we can now rephrase our earlier question. Indeed, the mutual information of X and Y is precisely what we are looking for! Recall that we wanted to know how much information Bob can learn about X. Since Alice and Bob see the joint distribution (X,Y), Bob learns $I(X;Y)$ bits of information about X. Note that $I(X;Y)$ depends on the σ_x's for $x \in \Sigma$ and the E_y's for $y \in \Gamma$.

As we are now looking at this scenario quantumly, we will in fact be studying quantum information theory. For further reading on this subject, see [Wat11].

We now provide an important definition:

Definition 5.1 (Accessible Information)**.** The *accessible information* is

$$I_{\text{acc}}(\sigma, p) = \max_{\substack{\text{over all} \\ \text{POVMs} \\ \{E_y\}_{y\in\Gamma}}} I(X;Y).$$

This represents the best Bob can do given Alice's choice of the σ_x's and the distribution p.

The best overall that the parties can do is therefore

$$\max_{\{\sigma_x\}_{x\in\Sigma}} I_{\text{acc}}(\sigma, p),$$

which can be upper bounded by $H(X) \leq \log|\Sigma|$. Our new goal is to relate $I_{\text{acc}}(\sigma,p)$ to the amount of "quantum" information in the σ's. Recall that Bob "sees" the mixed state ρ_B. But how much information is in a mixed state? To answer this question, we need to develop the quantum analogue of Shannon's classical coding theory.

So suppose we have the mixed state

$$\begin{cases} |\psi_1\rangle & \text{with prob. } p_1, \\ |\psi_2\rangle & \text{with prob. } p_2, \\ \quad\vdots \end{cases}$$

One might initially be inclined to define the quantum entropy to be $H(p)$, where p is the distribution over the $|\psi_j\rangle$'s. However, that would be wrong! This is due to the non-uniqueness of the representation of mixed states. For example, the above mixed state could be indistinguishable from the mixed state

$$\begin{cases} |\varphi_1\rangle & \text{with prob. } q_1, \\ |\varphi_2\rangle & \text{with prob. } q_2, \\ \vdots \end{cases}$$

even if we have $H(p) \neq H(q)$.

Here is the "correct" way to define the quantum analogue of shannon entropy, due to John von Neumann:

Definition 5.2 (Quantum Entropy)**.** Given a mixed state, let ρ be the density matrix, and suppose it has eigenvalues $\alpha_1, \ldots, \alpha_d$ with corresponding eigenvectors $|v_1\rangle, \ldots, |v_d\rangle$. We define

$$H(\rho) := \sum_{i=1}^{d} \alpha_i \log \frac{1}{\alpha_i} = H(\alpha).$$

This quantity is often referred to as the von Neumann entropy, and is sometimes denoted $S(\rho)$.

Remark 5.3. One can equivalently define

$$H(\rho) = \operatorname{tr}(\rho \log(1/\rho)).$$

While the matrix $\rho \log(1/\rho)$ might look a little scary, it can be simply thought of as the matrix that has the same eigenvectors $|v_1\rangle, \ldots, |v_d\rangle$ as ρ, but the corresponding eigenvalues are now $\alpha_1 \log(1/\alpha_1), \ldots, \alpha_d \log(1/\alpha_d)$ (with the same convention that $0 \log(1/0) = 0$). So

$$\rho \log(1/\rho) = \sum_{i=1}^{d} \alpha_i \log \frac{1}{\alpha_i} |v_i\rangle\langle v_i|.$$

Example 5.4. • *Suppose*

$$\rho = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

Then $H(\rho) = 0$.

- *Suppose*

$$\rho = \begin{bmatrix} 1/d & 0 & \cdots & 0 \\ 0 & 1/d & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/d \end{bmatrix}$$

Then $H(\rho) = \log d$. As a "cultural note" we remark that this state is often referred to as the "maximally mixed state".

Finally, we can answer the question: *how much quantum information does Bob get?*

Definition 5.5 (Quantum Mutual Information)**.** If ρ is the joint state of two quantum systems A and B then the *quantum mutual information* is

$$I(\rho_A; \rho_B) = H(\rho_A) + H(\rho_B) - H(\rho).$$

Example 5.6. *Suppose $\rho = \frac{1}{d}\sum_{i=1}^{d} |i\rangle\langle i| \otimes |i\rangle\langle i|$. Then we can mechanically compute ρ_A as follows:*

$$\begin{aligned} \rho_A = \mathrm{tr}_B(\rho) &= \mathrm{tr}_B\left(\frac{1}{d}\sum_{i=1}^{d} |i\rangle\langle i| \otimes |i\rangle\langle i|\right) \\ &= \frac{1}{d}\sum_{i=1}^{d} \mathrm{tr}_B(|i\rangle\langle i| \otimes |i\rangle\langle i|) \\ &= \frac{1}{d}\sum_{i=1}^{d} |i\rangle\langle i|\mathrm{tr}_(|i\rangle\langle i|) \\ &= \frac{1}{d}\sum_{i=1}^{d} |i\rangle\langle i|. \end{aligned}$$

At a more conceptual level, we could have immediately determined that $\rho_A = \frac{1}{d}\sum_{i=1}^{d} |i\rangle\langle i|$, the maximally mixed state, as follows. If Bob observes the state $|i\rangle\langle i|$, Alice's system collapses to $|i\rangle\langle i|$. Since Bob observes $|i\rangle\langle i|$ with probability $1/d$ for $i = 1, \ldots, d$, we conclude that Alice's mixed state is precisely given by the ensemble

$$\begin{cases} |1\rangle & \text{with prob. } 1/d, \\ |2\rangle & \text{with prob. } 1/d, \\ & \vdots, \end{cases}$$

which is represented by the density matrix $\frac{1}{d}\sum_{i=1}^{d} |i\rangle\langle i|$.

Similarly, we have $\rho_B = \frac{1}{d}\sum_{i=1}^{d} |i\rangle\langle i|$. We thus have

$$H(\rho_A) = H(\rho_B) = \log d.$$

Moreover, observe that

$$H(\rho) = \log d,$$

as ρ is essentially two perfectly correlated copies of the maximally mixed state. Hence,

$$I(\rho_A; \rho_B) = \log d + \log d - \log d = \log d.$$

Example 5.7. *If $\rho = \rho_A \otimes \rho_B$, then $I(\rho_A; \rho_B) = 0$.*

We can now define how much quantum information Bob gets from seeing Alice's state: it is precisely $I(\rho_A; \rho_B)$. This is such an important quantity that it gets its own name.

Definition 5.8 (Holevo Information)**.** The *Holevo information* is

$$\chi(\sigma, p) := I(\rho_A; \rho_B).$$

Next time, we will prove Holevo's Theorem:

Theorem 5.9 (Holevo [Hol73])**.** $I_{\text{acc}}(\sigma, p) \leq \chi(\sigma, p) \leq \log d$.

References

[CT12] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[Hol73] Alexander Semenovich Holevo. Bounds for the quantity of information transmitted by a quantum communication channel. *Problemy Peredachi Informatsii*, 9(3):3–11, 1973.

[Wat11] John Watrous. Theory of quantum information. *University of Waterloo Fall*, 2011.