CSCI5370 Quantum Computing November 25, 2013

Lecture 11: Quantum Information III - Source Coding

Lecturer: Shengyu Zhang *Scribe: Hing Yin Tsang*

11.1 Holevo's bound

Suppose Alice has an information source X that generates symbol x with probability p_x, her goal is to send the information to Bob. Classically, what Alice can do is to encode the symbol x as a codeword of certain length, namely $C_x \in \{0,1\}^n$, then sends it to Bob and finally Bob decodes it to figure out what x is. Let Y be the symbol Bob gets after decoding, then $I(X;Y)$ essentially measures how much information of X Bob can extract. It is clear that $I(X;Y) \leq I(C_X;Y) = H(C_X) - H(C_X|Y)$ where the first inequality follows by the data processing inequality. It means that if we want Y to contain many information about X, then we need a long codeword. But what if we have a quantum channel that allows us to transmit qubits? In this case, Alice can actually encode x using qubits rather than classical bits, and the codeword she can send is now a quantum state of dimension 2^n where n is the number of qubits. After receiving the quantum state, Bob can perform measurement on the state to get the variable Y and figure out what X. It is natural to ask whether we can encode X using a small number of qubits? Again we are interested in the quantity $I(X;Y)$. From the above discussion, we know that $I(X;Y)$ is upper bounded by the number of different states Alice can send. And the issue about classical bit is that a single bit can only represent 2 different states. However, a qubit is much more powerful since it can represent infinitely many different states! It seems that we can do much better in the quantum case. But after a moment of thought, we will realize that encoding X using one qubit does not solve the problem since Bob cannot distinguish the states perfectly unless the states are orthogonal! In general, the less qubits Alice uses, the more difficult Bob can distinguish them and hence extracting the information about X. So, there is a trade-off between the number qubit Alice use and how well Bob can distinguish the states. And how $I(X;Y)$ is related to the quantum states is not clear. It turns out that one can upper bound $I(X;Y)$ in terms of the von Neumann entropies of the some quantum states.

Theorem 11.1 (Holevo's bound)**.** *Let X be a random variable such that $\Pr[X = x] = p_x$. Suppose Alice sends a state ρ_x to Bob if $X = x$, then for any measurement described by POVM elements $\{E_y\}$ on the state Bob receive and let Y be the measurement outcome,*

$$I(X;Y) \leq S(\rho) - \sum_x p_x S(\rho_x).$$

This theorem tells us what we can hope for if we want Y to contain much information about X. Indeed, since $S(\rho)$ is at most the logarithm of the dimension of the quantum state, which is nothing but the number of qubits sent, it is now clear that quantum does not buy us much.

Proof. The idea is to view everything as a quantum system consisting of three parts P, Q and M, where Q represents the quantum system ρ_x Alice sends to Bob, and P, M represent the classical information X and Y. Then the result will follow by applying suitable inequalities for the von Neumann entropy.

Formally, we consider the quantum state

$$\rho_{PQM} = \sum_x p_x \underbrace{|x\rangle\langle x|}_{P} \otimes \underbrace{\rho_x}_{Q} \otimes \underbrace{|0\rangle\langle 0|}_{M}.$$

The joint system PQ corresponds to the situation that Alice prepares the state ρ_x with probability p_x. Then Bob performs a measurement with POVM elements $\{E_y\}$ on the system Q and stores the measurement outcome into M without touching P. After this step the state would become

$$\rho_{P'Q'M'} = \sum_x p_x \underbrace{|x\rangle\langle x|}_{P'} \otimes \sum_y \underbrace{\sqrt{E_y}\rho_x\sqrt{E_y}}_{Q'} \otimes \underbrace{|y\rangle\langle y|}_{M'}.$$

Now consider the quantity $S(P;Q)$. It is clear that $S(P;Q) = S(P;Q,M)$ since M is independent of P and Q. We also have $S(P;Q,M) \geq S(P';Q',M')$ since the measurement only apply to the system QM, it does not increase the mutual information between P and QM. Finally observe that $S(P';Q',M') \geq S(P';M')$ since discarding system does not increase mutual information. So we have the inequality

$$S(P;Q) \geq S(P';M').$$

We claim that $S(P';M') = I(X;Y)$ and $S(P;Q) = S(\rho) - \sum_x p_x S(\rho_x)$. The first one follows by the fact

$$\rho_{P'M'} = \sum_x p_x |x\rangle\langle x| \otimes \sum_y p_{y|x} |y\rangle\langle y| = \sum_{x,y} p_{x,y} |x\rangle\langle x| \otimes |y\rangle\langle y|.$$

For the second one, by definition we have $S(P;Q) = S(\rho_P) + S(\rho_Q) - S(\rho_{PQ})$. Observe that $S(\rho_P) = H(X)$, $S(\rho_Q) = S(\rho)$ and by the inequality for von Neumann entropy, we have $S(\rho_{PQ}) = H(X) + \sum_x p_x S(\rho_x)$. Putting all together yields the equality $S(P;Q) = S(\rho) - \sum_x p_x S(\rho_x)$ and hence complete the proof. □

11.2 Application: quantum random access codes

The Holevo's bound says that sending n bits of classical information requires n qubits. However, in some cases we are not very interested in knowing every bit of X (for instance when we are querying some entries in a database). Note that Holeve's bound does not say we need many qubits as $I(X;Y_i)$ is small (which is at most 1) for each fixed i. Classically, it is known that the codeword must have at least $(1 - H(\varepsilon))m$ bits where m is the length of the original message and ε is the success probability (this bound is almost tight as there exists classical random access code of length $(1 - H(\varepsilon))m + O(\log m)$). Can we do better using quantum mechanics? It turns out that the answer is *no*. Nayak gave a very simple proof using Holevo's bound. First let us define what a quantum random access code is.

Definition 11.2 (Quantum random access code)**.** Let m, n be positive integers, $\varepsilon > 0$, a (m,n,ε)-***quantum random access code*** is a function C that maps a m-bit string to a quantum state of dimension 2^n such that for each $1 \leq i \leq m$, there exists a measurement $\{M_i^0, M_i^1\}$ such that the probability that the measurement outcome is b is at least ε when the i-th bit of the string is b.

Theorem 11.3 (Nayak's bound)**.** *Let $\varepsilon \in [1/2, 1]$, if (m,n,ε)-quantum random access code exists, then $n \geq (1 - H(\varepsilon))m$.*

Proof. We use the following lemma, which follows from Holevo's bound and Fano's inequality:

Lemma 11.4. *Let $\rho = \frac{1}{2}(\rho_0 + \rho_1)$, if there exists measurement $\{M_0, M_1\}$ such that $\mathrm{Tr}(M_b \rho_b) \geq \varepsilon \in [1/2, 1]$, then*

$$S(\rho) \geq (1 - H(\varepsilon)) + \frac{1}{2}\big(S(\rho_0) + S(\rho_1)\big).$$

Now, suppose we have a (m,n,ε)-quantum random access code. Let ρ_x be the encoding of x and $\rho = \sum_x \frac{1}{2^m}\rho_x = \frac{1}{2}(\rho_0 + \rho_1)$ where $\rho_b = \sum_{x:x_1=b} \frac{1}{2^{m-1}}\rho_x$. By the definition of quantum random access code, there exists measurement $\{M_1^0, M_1^1\}$ such that for every ρ_x with $x_1 = b$, we have $\mathrm{Tr}(M_b \rho_x) \geq \varepsilon$. It follows by the linearity of trace that $\mathrm{Tr}(M_b \rho_b) \geq \varepsilon$. So by Lemma 11.4, we have $S(\rho) \geq (1 - H(\varepsilon)) + \frac{1}{2}\big(S(\rho_0) + S(\rho_1)\big)$. And it is not difficult to see ρ_b still satisfy the condition of Lemma 11.4 and we have

$$S(\rho) \geq 2(1 - H(\varepsilon)) + \frac{1}{4}\big(S(\rho_{00}) + S(\rho_{01}) + S(\rho_{10}) + S(\rho_{11})\big),$$

and by induction we have $S(\rho) \geq m(1 - H(\varepsilon)) + \frac{1}{2^m}\sum_b S(\rho_b) \geq m(1 - H(\varepsilon))$, as desired. □

Proof of Lemma 11.4. Let X be a uniform random bit, and let Y be the measurement outcome after applying the measurement $\{M_0, M_1\}$ to the state $\rho = \frac{1}{2}(\rho_0 + \rho_1)$. Then by Holevo's bound, we have

$$S(\rho) \geq I(X;Y) + \frac{1}{2}\big(S(\rho_0) + S(\rho_1)\big).$$

Since we have $\mathrm{Tr}(M_b\rho_b) \geq \varepsilon$, $\Pr[X = Y] \geq \varepsilon$ and hence $I(X;Y) = H(X) - H(X|Y) = 1 - H(X|Y)$. By Fano's inequality, let $p = \Pr[X \neq Y]$, we have $H(X|Y) \leq H(p) + p\log_2(|\mathcal{X}| - 1) \leq H(\varepsilon)$ since $\mathcal{X} = 2$ and $p \geq \varepsilon \geq 1/2$. It follows that $I(X;Y) \geq 1 - H(\varepsilon)$ and the proof is completed. □

For completeness, we also give a proof of Fano's inequality. The following proof is quite standard that can be found in many textbooks on information theory such as [**?**].

Lemma 11.5 (Fano's inequality)**.** *Let X, Y be random variables with support $\mathcal{X}$, if $\Pr[X \neq Y] = p$, then*

$$H(X|Y) \leq H(p) + p\log_2(|\mathcal{X}| - 1).$$

Proof. Let Z be a random variable that

$$Z = \begin{cases} 1, & \text{if } X = Y, \\ 0, & \text{if } X \neq Y. \end{cases}$$

Now, let us consider $H(X|Y)$. Since $H(Z|X,Y) = 0$, we have

$$\begin{aligned} H(X|Y) &= H(X,Z|Y) \\ &= H(Z|Y) + H(X|Y,Z) \\ &\leq H(Z) + H(X|Y,Z) \\ &= H(p) + pH(X|Y,Z=0) + (1-p)H(X|Y,Z=1) \\ &= H(p) + pH(X|Y,Z=0) \\ &\leq H(p) + p\log_2(|\mathcal{X}| - 1) \end{aligned}$$

where the first inequality follows since conditioning does not increase entropy, and the second inequality follows by the fact that if $X \neq Y$ then the number of possible outcomes of X is at most $|\mathcal{X}| - 1$. □

11.3 Source coding theorems

11.3.1 Classical source coding theorem

Suppose Alice have an information source (a random variable) X that generates symbol x with probability p_x. For the purpose of efficient storage or transmission, she would want to encode the information using some short codes. So that when she sends the codeword to Bob, Bob can decode it and extract the information back. Ideally, Alice would want to have such a encoding and decoding scheme so that the probability that Bob can get back the symbol x is arbitrarily close to 1. Intuitively, $H(X)$ bits seem to be sufficient since the entropy of X is just $H(X)$. However, for any fix X, it is impossible to achieve arbitrarily small error probability if the codeword is shorter than $\log_2 |\mathcal{X}|$. Nevertheless, if we consider the asymptotic version of this problem, i.e. Alice has n independent copies of X, namely X^n, whose generates symbol $x^n = x_1 \ldots x_n$ with probability $\prod_{i=1}^n p_{x_i}$, and she want to assign a codeword for each x^n so that Bob can decode the codeword to get back x^n with probability tends to 1 as $n \to \infty$. It turns out that we can achieve it with rate (defined by the length of the codeword divided by n) arbitrarily close to $H(X)$. And moreover, it is also necessary.

Theorem 11.6 (Shannon's source coding theorem)**.** *Let X^n be n i.i.d. copies of X. Then the following holds:*

1. *If $R > H(X)$, there exist $C_n : \mathcal{X}^n \to \{0,1\}^{nR}$ and $D_n : \{0,1\}^{nR} \to \mathcal{X}^n$ such that*

$$\lim_{n\to\infty} \Pr_{x^n \sim X^n}[D_n(C_n(x^n)) = x^n] = 1.$$

2. *If $R < H(X)$, for any $C_n : \mathcal{X}^n \to \{0,1\}^{nR}$ and $D_n : \{0,1\}^{nR} \to \mathcal{X}^n$, we have*

$$\lim_{n\to\infty} \Pr_{x^n \sim X^n}[D_n(C_n(x^n)) = x^n] < 1.$$

The proof of this theorem uses a notion called *typicality*. The idea is that, if we look at the sequence $x^n = x_1 \dots x_n$ generated by X^n, since each x_i is an independent copy of X, with very high probability (when n sufficiently large) the number of i such that $x_i = x$ for each $x \in \mathcal{X}$ is roughly p_x. Sequences with such property is said to be typical. The main observation is that actually the set of typical sequence is roughly $2^{nH(X)}$. Thus we can simply encode all the typical sequences into distinct codewords, which takes around $nH(X)$ bits, and encode all the other sequences into arbitrary codewords. Then the decoder can successfully decode the message if it is typical, which happens with very high probability.

Lemma 11.7 (Typical sequence theorem)**.** *Let X^n be i.i.d. copies of X with support $\mathcal{X}$. Let $\varepsilon, \delta > 0$, and $T_\varepsilon^n = \{x^n \in \mathcal{X}^n : 2^{-n(H(X)+\varepsilon)} \le p(x^n) \le 2^{-n(H(X)-\varepsilon)}\}$, then for n sufficiently large, the following holds:*

1.
$$\Pr[X^n \in T_\varepsilon^n] \ge 1 - \delta, \tag{1}$$

2.
$$(1-\delta)2^{H(X)-\varepsilon)} \le |T_\varepsilon^n| \le 2^{n(H(X)+\varepsilon)} \tag{2}$$

Proof. For the first item, by the definition of T_ε^n, we have

$$\begin{aligned}
\Pr[X^n \in T_\varepsilon^n] &= \Pr[2^{-n(H(X)-\varepsilon)} \le p(X^n) \le 2^{-n(H(X)+\varepsilon)}] \\
&= \Pr\left[\left|\frac{1}{n}\sum_{i=1}^n \log\frac{1}{p(X_i)} - H(X)\right| \le \varepsilon\right] \\
&= \Pr\left[\left|\frac{1}{n}\sum_{i=1}^n \log\frac{1}{p(X_i)} - \mathbb{E}\Big[\log\frac{1}{p(X)}\Big]\right| \le \varepsilon\right] \\
&\ge 1 - \delta
\end{aligned}$$

for sufficiently large n, where the inequality follows by the weak law of large number and the fact that $-\log p(X_i)$ are i.i.d.

For the second item, by (1), we have

$$1 - \delta \le \Pr[X^n \in T_\varepsilon^n] \le 1,$$

which implies

$$1 - \delta \le \sum_{x^n \in T_\varepsilon^n} p(x^n) \le 1.$$

Now, it is immediate from the definition of T_ε^n that

$$1 - \delta \le |T_\varepsilon^n| 2^{-n(H(X)-\varepsilon)}$$

and

$$|T_\varepsilon^n| 2^{-n(H(X)+\varepsilon)} \le 1,$$

as desired. □

Proof of Theorem 11.6. For the first item, let $\varepsilon > 0$ such that $R - \varepsilon > H(X)$, the C_n simply assign each sequence $x \in T_\varepsilon^n$ a distinct nR bit-string other sequences an arbitrary nR bit-string. The decoder output the typical sequence in $C_n^{-1}(y^n)$. If x is typical, there is no error. So, the error probability is at most the probability that $x \notin T_\varepsilon^n$, which can be made to be arbitrary small by (1).

For the second item, let $A_n = \{x^n \in \mathcal{X}^n : D_n(C_n(x^n)) = x^n\}$. It is clear that $|A_n| \leq 2^{nR}$ since $A_n \subseteq D_n(\{0,1\}^{nR})$. It suffices to show $\Pr[X^n \in A_n]$ is small. Now, take $\varepsilon > 0$ such that $R + \varepsilon < H(X)$. Consider $\Pr[X^n \in A_n] = \Pr[X^n \in A_n \cap T_\varepsilon^n] + \Pr[X^n \in A_n \setminus T_\varepsilon^n]$. Take $\delta > 0$ small, then for n sufficiently large, the second term is at most δ by (1), and the first term can be upper bounded by

$$\Pr[X^n \in A_n \cap T_\varepsilon^n] \leq |A_n| \cdot 2^{-n(H(X)-\varepsilon)} \leq 2^{-n(H(X)-R-\varepsilon)}$$

which tends to 0 as $n \to \infty$. It follows that the error probability does not tends to 1 as $n \to \infty$, and the proof is completed. □

11.3.2 Quantum source coding theorem

In the quantum case, Alice are given a state $\rho^{\otimes n}$, which is a product of n copies of ρ. And her goal is to encode it using as less qubits as possible while at the same Bob are able to get back a state which is close to $\rho^{\otimes}$ after performing measurement. Here the closeness is measured by fidelity. We have the following theorem:

Theorem 11.8 (Schumacher's quantum source coding theorem)**.** *Let $\rho \in H$ be an i.i.d. quantum source. Then the following holds:*

1. *If $R > S(\rho)$, there exist $C_n : H^{\otimes n} \to \mathbb{C}^{2^{nR}}$ and $D_n : \mathbb{C}^{2^{nR}} \to H^{\otimes n}$ such that*

$$\lim_{n\to\infty} F(\rho^{\otimes n}, D_n(C_n(\rho^{\otimes n})) = 1$$

2. *If $R < S(\rho)$, for any $C_n : H^{\otimes n} \to \mathbb{C}^{2^{nR}}$ and $D_n : \mathbb{C}^{2^{nR}} \to H^{\otimes n}$, we have*

$$\lim_{n\to\infty} F(\rho^{\otimes n}, D_n(C_n(\rho^{\otimes n})) < 1$$

Its proof is similar to the proof of Theorem 11.6. The main difference is that we use the *typical subspace theorem*, which is an analogue of the typical sequence theorem.