VII. QUANTUM INFORMATION THEORY

A. Introduction

Information theory has developed over the past five or six decades in parallel to computer science. Its roots are in communication theory, that is, in the theory of transmission of information by telephone or radio. Of course, all parts of our course deal with information theory in a wider sense, but as the subfields have developed, questions of computation and algorithm development have been separated from information theory in a narrower sense. In this chapter we will restrict ourselves to some problems dealing with the *transmission of information*.

The most fundamental questions of course are, what *is* information, or, more precisely, how can it be quantified? These questions were dealt with in the pioneering contributions of Claude Shannon [60] in the late 1940s. The historical (or socio-economic) context was the rapid growth of communication by telephone lines. Consequently the problem was formulated as the problem of effectively transmitting information through a given "channel". The channel, for example a telephone line, may connect two points in space, but it may also connect two points in time, in which case we are dealing with effective data *storage*. As every channel has physical limits, there is an obvious interest in precisely determining these limits and extending them if possible. To do that, a measure of the information content of a communication must be developed and related to the capacity of the channel. That is the content of *Shannon's noiseless channel coding theorem*. Of course channels are always noisy, and questions of error-correction immediately come to mind. Actually there is a large subfield of classical information theory dealing with the development of error-correcting codes. The fundamental limits are fixed by *Shannon's noisy channel coding theorem*.

In contrast to the theory of quantum (or classical) algorithms, here we are not dealing with a small number of (qu-)bits which must be processed, but with large quantities of transmitted data. From the point of view of the communications engineer these data form a random sequence of symbols about which only some statistical properties may be known. It turns out (not unexpectedly) that some key concepts from statistical mechanics, such as entropy are useful also in information theory, both classical and quantum.

After discussing some notions of classical information theory we will try to generalize the concepts to the quantum regime. Unfortunately it turns out that the use of qubits does not significantly speed up the transmission of classical information (such as this text) through a noiseless channel. Nevertheless it is interesting to study how the notion of classical information may be generalized to quantum information, how strongly quantum information may be compressed (looking for the quantum analogs of Shannon's theorems), and how quantum noise (i.e. *continuous* fluctuations in both amplitude and phase in contrast to mere bit flips) may affect the transmission.

B. A few bits of classical information theory

1. Information content and entropy

The first question is, how to quantify information. Imagine I tell you

$$X = 2.$$

How much information do you gain? That depends on your previous knowledge: if you knew already that X was 2, you learn nothing. If you only knew that X was determined by throwing a die you gain information. The information content of X is a measure of your *ignorance*: how much information would you gain if you learned the value of X? That depends obviously on the number of values x of the random variable X and their probabilities $p(x)$. The general formula for the information content of X is

$$S(X) \equiv S(\{p(x)\}) = -\sum_x p(x) \log_2 p(x).$$

Since $0 \leq p(x) \leq 1$, $S(X) \geq 0$. Let us look at more examples to see if this definition makes sense:

- $p(x) = \delta_{x,2}$ (for integer x) $\Rightarrow S(X) = 0$.
  Nothing is learned if we know already that $X = 2$.
- $p(x) = \frac{1}{N}$ for $x = 1, \ldots, N$ and zero otherwise $\Rightarrow S = \log_2 N$.
  $N = 6 \Rightarrow S = 2.58$ (the fair die)
  $N = 2^m \Rightarrow S = m$: m bits must be specified to convey the information
- $p(6) = \frac{1}{2}, p(1) = \ldots = p(5) = \frac{1}{10} \Rightarrow S = 2.16$
  (a loaded die)

The comparison between the fair die and the loaded die shows that the potential information gain decreases if the information about the probability distribution increases. The uniform probability distribution is the one with "maximal ignorance". Obviously S is closely related to the entropy well-known from Statistical Mechanics, and it is indeed often called information entropy or *Shannon entropy*. A simple but important special case is a binary variable ($X = 0$ or 1, say), with $p(1) = p \Rightarrow p(0) = 1 - p$. $S(x)$ is then a function of p only:

$$S(X) = H(p) = -p \log_2 p - (1-p) \log_2 (1-p).$$

The binary entropy function $H(p)$ assumes its maximum value 1 at $p = \frac{1}{2}$.

2. Mutual information and the data processing inequality

Consider two random variables X and Y; then we can define the *conditional probability* $p(y|x)$ that $Y = y$

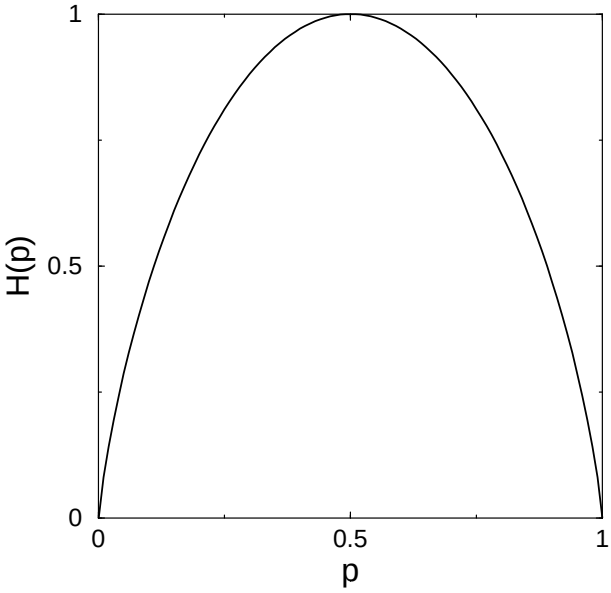

FIG. VII.1: The binary entropy function $H(p)$.

given that $X = x$, and the *conditional entropy*

$$S(Y|X) = -\sum_x p(x) \sum_y p(y|x) \log_2 p(y|x).$$

(Note that the ("simultaneous") probability $p(x, y)$ that $X = x$ and $Y = y$ is $p(x, y) = p(x)p(y|x)$.) Since $-\sum_y p(y|x) \log_2 p(y|x)$ is the information content of Y for given value of X, $S(Y|X)$ is the average information content remaining in Y if we were to learn the value of X. (Where the average is performed over the possible values of X.)
The *mutual information* content of X and Y is defined as

$$I(X : Y) = \sum_x \sum_y p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}.$$

By using the fact that

$$p(x|y) = \frac{p(x, y)}{p(x)p(y)}$$

we see

$$I(X : Y) = \sum_x \sum_y p(x, y) \log_2 p(x|y) - \sum_x \sum_y \log_2 p(x)$$

$$= -S(X|Y) + S(X)$$

due to $p(x) = \sum_y p(x, y)$. From that equality and its analog for $p(y)$ we also see that

$$I(X : Y) = -S(X, Y) + S(X) + S(Y)$$

where $S(X, Y)$ is the information content of the "vector" random variable (X, Y). This shows (as did the definition already) that

$I(X : Y)$ is symmetric with respect to X and Y. If X and Y are independent random variables,

$$p(x, y) = p(x)p(y) \Rightarrow I(X : Y) = 0$$

and this indicates that $I(X : Y)$ in fact measures how much X and Y "know about each other".
Let us now view data processing as a stochastic process (a Markov chain) of three random variables $X \to Y \to Z$ where successive variables are connected by conditional probabilities $p(y|x)$ and $p(z|y)$ and where the simultaneous probability $p(x, y, z) = p(x)p(y|x)p(z|y)$. The the *data processing inequality* says

$$S(X) \geq I(X : Y) \geq I(X : Z),$$

that is, Z cannot know more about X than Y knew which is less than the information content of X. This highly plausible inequality (a corollary to which is the well-known rule "garbage in, garbage out") can be deduced from the properties of the various entropy functions discussed above. (Compare, for example, [3], Chap. 11).

3. Data compression and Shannon's noiseless channel coding theorem

The basic idea of data compression is very simple and very old, too: Determine which sequences of symbols or words occur most frequently and use abbreviations for them, that is, code these symbols in short strings of the symbols (bits, for example) used for data transmission. We illustrate this principle with a very simple example. Suppose we wish to transmit information from a source X with a four-letter alphabet with unequal probabilities. Four symbols can be distinguished by using two bits and there is a "natural" (or naïve) way to do this. In the table we show both the naïve code and a "clever" code which we analyze below.

symbol	probability	naïve code	clever code
1	$\frac{1}{2}$	00	0
2	$\frac{1}{4}$	01	10
3	$\frac{1}{8}$	10	110
4	$\frac{1}{8}$	11	111

Note that in the naïve code all symbols are stored in two bits each. The clever codes uses bit strings of variable length, but nevertheless the boundaries of the symbols are always well defined: after a "0" or after at most three bits. The average length of the cleverly coded string in bits per symbol then is

$$\frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{2}{8} \cdot 3 = \frac{7}{4} < 2.$$

Let us compare this to the entropy of the source:

$$S(X) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{2}{8} \log_2 \frac{1}{8}$$

$$= \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{2}{8} \cdot 3 = \frac{7}{4}.$$

The fact that the two numbers are equal is no coincidence. Also, no compression scheme can be constructed which works with a smaller number of bits per symbol on average. This is the contents of Shannon's noiseless channel coding theorem.

To illustrate the idea a little more generally (but without going into full generality) we consider a source sending a stream of binary symbols: $X = 0, 1$; $p(1) = p, p(0) = 1 - p$ with $p \neq \frac{1}{2}$. (Remember: the central elements of data compression were the fact that not all strings are equally probable, and the use of short codes for frequent symbols.) We will not encode individual symbols but blocks of n symbols with n large. In the typical case such a block will contain np ones and $n(1-p)$ zeros. (Let us postpone for a moment the discussion of what "typical" really means.) There are many blocks of n symbols np of which are ones. The probability of any such sequence of zeros and ones is

$$p_{typ} = p^{np}(1-p)^{n(1-p)}.$$

Now note that

$$\log_2 p_{typ} = np \log_2 p + n(1-p)\log_2(1-p) = -nH(p)$$

$$\Rightarrow p_{typ} = 2^{-nH(p)}$$

where $H(p)$ is the binary entropy function defined earlier. As these typical sequences all have equal probability $2^{-nH(p)}$, their total number is $2^{nH(p)}$, and they can be numbered, from 1 to $2^{nH(p)}$. To communicate which one of the $2^{nH(p)}$ possible typical sequences are transmitted, only $nH(p)$ bits are needed, not n bits as in the case where bits are transmitted one by one. It is not possible to distinguish the typical sequences by sending fewer than $nH(p)$ bits, since they are all equally probable, so the compression from n to $nH(p)$ is optimal.

So, how typical is typical, and why is the above argument relevant? Why do we really encounter (almost) only typical sequences? It turns out that the answer to these questions is provided by one of the "laws of large numbers" arguments which are familiar from elementary statistical mechanics. There it is shown, for example, that in the "canonical ensemble" the energy per particle may be allowed to fluctuate arbitrarily, but nevertheless the total energy of a *large* number of particles practically does not deviate from its mean value.

Recall that a typical sequence was one with np ones. The probability of finding m ones in a sequence of n symbols is

$$p(m) = \left(\begin{array}{c} n \\ m \end{array} \right) p^m (1-p)^{n-m},$$

the binomial distribution. For fixed p and large n, the binomial distribution is excellently approximated by a Gaussian distribution. (To see this expand $\ln p(m)$ to second order about its maximum at $m = np$, using Stirling's formula for the binomial coefficient.)

$$p(m) \approx \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(m-np)^2}{2\sigma^2}}$$

with $\sigma = \sqrt{np(1-p)}$, the standard deviation. Note that while the mean value np grows linearly with the sequence length n, the standard deviation only grows as $\sqrt{n}$. That is, the *relative* fluctuations of the number of ones in a sequence becomes smaller as the sequence grows longer and for long enough sequences we can be pretty sure that almost all sequences are typical.

Thus we only have to transmit $H(p) < 1$ bits per symbol for our binary source. More generally for a source producing random variables X (capable of d values so that coding the symbols one by one would require $\log_2 d$ bits per symbol) with an information content $S(X)$ we need $nS(X) < n \log_2 d$ bits to communicate n values of X. This fact about the compressibility of data is known as Shannon's noiseless channel coding theorem.

For practical purposes it is of course not always possible to wait until a large n number of symbols have accumulated before starting the transmission. However, there are near-optimal coding schemes for blocks of a few (say, four) symbols only. They are based on the same idea as the example we started with: use shorter transmission codes for the most frequently occurring blocks of symbols. An example for such a scheme is the *Huffman code* (compare [5]).

4. The binary symmetric channel and Shannon's noisy channel coding theorem

We have to think about signal transmission in the presence of noise, because noise is unavoidable in real-world systems. Depending on the physical nature of the signal and the transmission channel, different types of noise are possible. We will concentrate on the important and simple case of binary digital transmission (of zeros and ones, that is) and symmetric bit-flip noise. That means that every single bit is flipped with a certain probability p on its way down the channel, regardless of its value (0 or 1) and regardless of the fate of all other bits. Such a channel is called a binary symmetric channel, and we want to know its capacity, measured in (useful) bits transmitted per bits in. It turns out (see [5] for details) that for the maximum information content of the source, $S(X) = 1$ (that is, 0 and 1 are equally probable in the input bit stream) the channel capacity is

$$C(p) = 1 - H(p)$$

where $H(p)$ is again the binary entropy function defined earlier. For a noisy channel one must use some redundancy, that is, one must employ *error-correcting codes*. Shannon's noisy channel coding theorem tells us that for any given channel capacity $C(p)$ there exist error-correcting codes which allow transmission with an arbitrarily small error probability.

Unfortunately the theorem is an existence theorem and does not tell us immediately how such a code may be constructed, but fortunately, a variety of clever error-correcting codes have been constructed (see [7] for some examples), for example for the transmission

of image data from satellites traveling the solar system to Jupiter and beyond.

C. Some bits of quantum information theory

1. *The von Neumann entropy*

It turns out that a useful quantum analog to Shannon's entropy (information content) for a classical set of probabilities p_i (which characterize the possible values x_i of a random variable X)

$$S(\{p_i\}) = -\sum_i p_i \log_2 p_i$$

is the *von Neumann entropy*

$$S(\rho) = -\mathrm{Tr}\rho\log_2\rho$$

which is defined for any density operator, that is, any operator ρ with $\rho = \rho^\dagger \geq 0$, $\mathrm{Tr}\rho = 1$. Any such ρ can be decomposed in normalized pure states,

$$\rho = \sum_i p_i |\phi_i\rangle\langle\phi_i| \quad (p_i \geq 0; \sum_i p_i = 1).$$

This is possible in many ways for any given ρ, and to any of these possibilities we can assign a (classical) Shannon entropy $S(\{p_i\})$; and it can be shown that

$$S(\{p_i\}) \geq S(\rho),$$

with equality if and only if the vectors $|\phi_i\rangle$ are pairwise orthogonal. (Take, for example, the eigenstates of ρ.) This inequality has a fairly obvious interpretation in terms of the distinguishability of two quantum states. Imagine a person (Alice) sending a string of classical symbols x_i down a line to another person (Bob), according to probabilities p_i. We have learned that the information content of this transmission is $S(\{p_i\})$. Now let us assume that Alice is a dedicated follower of fashion and goes into the quantum communication business. Instead of sending classical symbols x_i she sends quantum states $|\phi_i\rangle$. While Bob can easily distinguish all possible x_i, he can only distinguish two states with certainty if they are orthogonal to each other. This is also related to the no-cloning theorem: imagine Bob *could* clone arbitrary unknown quantum states. He then could make many copies of the incoming state and perform *many* measurements comparing clones of Alice's state to clones of all possible states and determine Alice's state with high probability.
It is instructive to consider a simple example involving a two-dimensional Hilbert space spanned by the vectors $|\alpha\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $|\gamma\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Let us define a third vector

$$|\beta\rangle := \cos\phi|\gamma\rangle + \sin\phi|\alpha\rangle$$

and the density matrix

$$\rho := p|\alpha\rangle\langle\alpha| + (1-p)|\beta\rangle\langle\beta|$$

$$= \begin{pmatrix} p + (1-p)\sin^2\phi & (1-p)\cos\phi\sin\phi \\ (1-p)\cos\phi\sin\phi & (1-p)\cos^2\phi \end{pmatrix}.$$

The easiest way to calculate the von Neumann entropy $S(\rho)$ is via the eigenvalues λ_i of ρ:

$$S(\rho) = -\sum_i \lambda_i \log_2 \lambda_i.$$

The eigenvalues of the above density matrix are

$$\lambda = \frac{1}{2} \pm \sqrt{\frac{1}{4} - p(1-p)\cos^2\phi}.$$

For $\phi = 0$, $\lambda = p, 1-p \Rightarrow S(\rho) = H(p)$ (the binary entropy function) and for $\phi \neq 0$, $S(\rho)$ is strictly smaller, as seen in the figure.

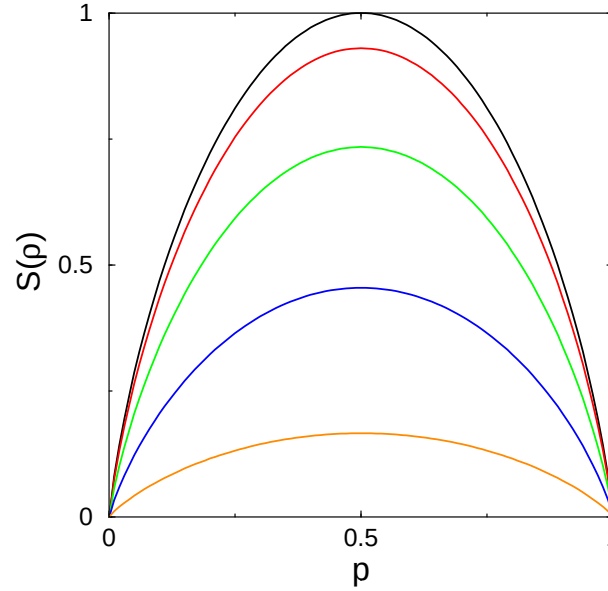

FIG. VII.2: The von Neumann entropy for a simple two-dimensional density matrix. Curves are for $\phi = 0, 0.1\pi, 0.2\pi, 0.3\pi$, and 0.4π, respectively (top to bottom). See text for details.

The quantum entropy has some non-classical properties. Whereas classical random variables X, Y always fulfill

$$S(X) \leq S(X,Y),$$

that is, the entropy of a subsystem is never greater than that of the total system, this *is* possible for a quantum system. Consider two qubits A, B in the (pure!) state

$$|\phi\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle); \quad \rho_{AB} = |\phi\rangle\langle\phi| \Rightarrow S(\rho_{AB}) = 0.$$

However, the reduced density matrix of subsystem A (obtained from ρ_{AB} by performing the trace over the Hilbert space of B) is $\rho_A = \frac{1}{2}\mathbf{1} \Rightarrow S(\rho_A) = 1$.
Evidently this related to the entanglement between A and B. In general A and B can be considered entangled if and only if

$$S(\rho_{AB}) < S(\rho_A) \text{ (or } S(\rho_B)),$$

where, of course, ρ_A is again the reduced density matrix.
Most theorems concerning entropy which are relevant to quantum information theory can be derived from a few fundamental properties which are discussed, proved and applied in [3] and which we just quote here for the sake of completeness:
i) *concavity*

$$\lambda_1 S(\rho_1) + \lambda_2 S(\rho_2) \leq S(\lambda_1\rho_1 + \lambda_2\rho_2)$$

$(\lambda_{1,2} \geq 0, \lambda_1 + \lambda_2 = 1)$ (This property is related to stability in the context of Statistical Mechanics.)
ii) *strong subadditivity*

$$S(\rho_{ABC}) + S(\rho_B) \leq S(\rho_{AB}) + S(\rho_{BC})$$

iii) *triangularity*

$$|S(\rho_A) - S(\rho_B)| \leq S(\rho_{AB} \leq S(\rho_A) + S(\rho_B)$$

All of these inequalities hold also (in appropriately modified form) for the Shannon entropy, except the first one in iii).

2. The accessible information and Holevo's bound

We are still dealing with the transmission of classical data through a quantum channel. Let Alice have a classical information source X, that is, a random variable with values x_i, probabilities $p_i (i = 0, \ldots, n)$. According to the value x_i to be transmitted, Alice prepares one quantum state ρ_i from a fixed set of (mixed, in general) states $\rho_0, \ldots, \rho_n$ and gives it to Bob who measures the state and gets a result which can be treated like a classical random variable Y capable of values $y_0, \ldots, y_m$. Let us discuss Bob's measurement a little more precisely. Bob has a set of *measurement operators* $\mathbf{M}_i (i = 0, \ldots, m)$ which he can apply to any incoming state vector $|\psi\rangle$ (and also, with appropriate changes in notation, to mixed states). The probability for finding the result i is

$$p_i = \langle\psi|\mathbf{M}_i^\dagger \mathbf{M}_i|\psi\rangle$$

and the state immediately after the measurement is

$$\frac{\mathbf{M}_i|\psi\rangle}{\sqrt{\langle\psi|\mathbf{M}_i^\dagger \mathbf{M}_i|\psi\rangle}}.$$

The operators $\mathbf{E}_i := \mathbf{M}_i^\dagger \mathbf{M}_i$ are positive, and if $\sum_{i=0}^m \mathbf{E}_i = \mathbf{1}$ they are called POVM elements (positive operator valued measure elements). (If the sum is smaller than one Bob's measurement misses some possibilities of the incoming $|\psi\rangle$.) An extremely simple example for a set of POVM elements are the projectors $\mathbf{P}_i$ on the states of a basis.
Turning back to the result Y of Bob's measurement (described by POVM elements $\mathbf{E}_0, \ldots, \mathbf{E}_m$), it is clear that what Bob can learn about Alice's message is $I(X : Y)$, the mutual information, which depends on the cleverness of his measurement strategy. The *accessible information* is the maximum of $S(X : Y)$ over all measurement strategies. There is no prescription to calculate the accessible information, but there is a *bound* by Holevo (also often spelled Kholevo). Under the conditions described above, and with $\rho := \sum_i p_i\rho_i$, we have

$$S(X : Y) \leq S(\rho) - \sum_i p_i S(\rho_i) =: \chi$$

where χ is sometimes called the Holevo information. (For the simplest possible example compare section 12.1.2 of [3].)

3. Schumacher's noiseless channel coding theorem

Consider a "quantum alphabet" of states $|\phi_i\rangle$ (not necessarily orthogonal to each other) with probabilities p_i. Such an alphabet can be described by a density operator

$$\rho = \sum_{i=1}^{|A|} p_i |\phi_i\rangle\langle\phi_i|.$$

A message is a sequence of n "quantum characters": $|\phi_{i_1}\rangle|\phi_{i_2}\rangle \ldots |\phi_{i_n}\rangle$. The ensemble of n-symbol messages is described by the density operator $\rho^{\otimes n}$ which lives in a Hilbert space $H^{\otimes n}$ of dimension

$$|A|^n = 2^{n \log_2 |A|}$$

(or smaller, if the alphabet states are not linearly independent).
Is it possible to compress the information contained in $\rho^{\otimes n}$? Schumacher's 1995 theorem provides an affirmative answer. For sufficiently large n, $\rho^{\otimes n}$ is compressible to a state in a Hilbert space of dimension $2^{nS(\rho)}$ (that is, in $nS(\rho)$ qubits) with a fidelity (probability that after decompression the original state is recovered) approaching 1. This means that $S(\rho)$ is the number of qubits of essential quantum information, per character of the alphabet.
The proof rests on the same ideas as that of Shannon's noiseless channel coding theorem, namely typical sequences and the laws of large numbers. The density operator ρ can be decomposed in its eigenstates $|x\rangle$ (which are orthonormal), with eigenvalues $p(x)$:

$$\rho = \sum_x p(x)|x\rangle\langle x|.$$

Then the von Neumann entropy is equal to the Shannon entropy

$$S(\rho) = S(\{p(x)\}).$$

We can then define a typical sequence

$$x_1, x_2, \ldots, x_n$$

of classical symbols x_i and associate with it a typical state

$$|x_1\rangle|x_2\rangle \ldots |x_n\rangle$$

in the Hilbert space $H^{\otimes n}$. The typical states span the *typical subspace* and by the laws of large numbers a few facts can be shown about the typical subspace for sufficiently large n which are very similar to the properties of the typical sequences leading to Shannon's noiseless channel coding theorem. (See [3] for a nice parallel treatment of both theorems.)

- $\rho^{\otimes n}$ has almost all of its weight in the typical subspace:

  $$\mathrm{Tr}P(n)\rho^{\otimes n} \geq 1-\delta \quad (\delta \to 0)$$

  where $P(n)$ is the projector on the typical subspace.

- The dimension of the typical subspace is asymptotically $2^{nS(\rho)}$:

  $$\mathrm{Tr}P(n) \approx 2^{nS(\rho)},$$

  implying that compression is possible.

- The weight of $\rho^{\otimes n}$ in any *smaller* subspace is negligible: Let $Q(n)$ be a projector on any subspace of $H^{\otimes n}$ of dimension at most 2^{nR} with $R < S(\rho)$. Then for any $\delta > 0$ and n sufficiently large

  $$\mathrm{Tr}Q(n)\rho^{\otimes n} \leq \delta$$

  implying that compression is limited: if one tries to press too hard, the data will get lost.

4. Classical information over noisy quantum channels

This is a subject of ongoing research (as is, even more so, the subject of quantum information over noisy quantum channels). The usage of quantum states for information transfer offers many possibilities which do not exist classically. Many of these possibilities are related to entanglement. For example, two or more successive qubits transmitted may be entangled, and there may also be entanglement between transmitter and receiver. (This leads to the fascinating possibilities of quantum cryptography and teleportation which we discussed in last semester's course.) Many of the schemes involving entanglement between the transmitted qubits are not explored very well. The simplest case is that of product state transmission, that is, the n-symbol quantum message is just a product state of n factors (*no* entanglement). For that case an analogy of Shannon's noisy channel coding theorem has been shown which gives a *lower bound* for the capacity of a noisy quantum channel. That lower bound is known as the Holevo-Schumacher-Westmoreland (HSW) bound. Some researchers suspect that the bound is in fact the exact value of the capacity, but this has not yet been proved. Details on the HSW theorem together with some simple examples can be found in [3].