Module II, Part 1

Entropy, information and entanglement

PH 534 QIC

Himadri Shekhar Dhar
himadri.dhar@iitb.ac.in

A. Information and entropy

Before we delve into the world of quantum information, we will spend some time to understand what we mean by classical information theory. While this in itself is a vast topic, and it is impossible to do it full justice within the scope of our present course, we will limit ourselves to some basic definition and properties of information and its manipulation. In principle, our scope of enquiry would be to ask – what are the resources involved in communicating classical information and how does this change if we use quantum resources. As such, our foray into information theory will be limited to Shannon entropy and its quantum analogue, and their properties.

i) Shannon entropy

The cornerstone of classical information theory is Shannon entropy, which captures the disorder or uncertainty in any variable X. Alternatively, such an entropy also captures the amount of information contained in the variable X. This complementary view basically connects the uncertainty in not knowing a quantity with the information gained when we know about X. In other words, Shannon entropy provides a duality of how much we know ("information") or how much we don't know ("uncertainty").

If X takes values $\{x_1, x_2, \ldots, x_n\}$ with probabilities $\{p_1, p_2, \ldots, p_n\}$, the Shannon entropy is then defined as:

$$H(X) = -\sum_i p_i \log_2 p_i.$$

This highlights two important points about information: i) its amount is independent of the actual values of the variable but rather their probabilities. It does not matter if we interchange p_1 with p_2, and ii) the log function ensures that information is additive for two independent events occurring, i.e., $f(pq) = f(p) + f(q)$. The summation over the probabilities just allows us to capture the average information over all probabilities. We consider the log to be taken to the base 2, to accommodate classical information in terms of bits. Also, we assume that $0 \log_2 0 = 0$, which implies that $p = 0$ events are not included in the average.

Operationally, Shannon entropy quantifies the resources needed to store information. Suppose you have a source that generates a string $p_1, p_2, \ldots, p_n$, from a set of independent, random variables $P_1, P_2, \ldots, P_n$. The question now is what is the minimum resource (number of bits) required to store or communicate this information. The answer is enshrined in *Shannon's noiseless coding theorem* and is equal to $H(P_i)$ number of bits per symbol, where $H(X)$ is the Shannon entropy.

Example: Let us consider a farm that stocks these food items: bread, eggs, chicken and fish, with stock size proportional to 1/2, 1/4, 1/8, and 1/8.

Ideally, we need two bits of store this information: 00, 01, 10, 11, i.e., bread (00), eggs (01), chicken (10), fish (11). Each message requires two bit.

But all these items do not have the same probability, so we can compress this data: let's say, bread (0), eggs (10), chicken (110), fish (111)
Average length: 1/2 * 1 + 1/4 * 2 + 1/8 * 3 + 1/8 * 3 = 7/4
Shannon entropy: -1/2*log(1/2) –1/4*log(1/4) – 1/8*log(1/8) – 1/8*log(1/8) = **7/4**

Therefore, on average each message will require $H(X) = 7/4 < 2$ bits.

Suppose, you have an n-bit message: 000111…..110, where the bit 0 occurs with probability p and the bit 1 with probability $1 - p$. Now, how many bits are required to represent this message?

In the limit of large n, the minimum resource needed to express a message containing n bits, is $nH(p)$, which is equal to n only for $p = 1/2$.

The questions above concerns redundancy. It seeks to ask how can less resources be used to carry a message, on average. Shannon connected these questions to the idea of entropy, which then forms the cornerstone of information theory. All of major work related to classical information theory is about manipulation of *Shannon entropy*.

Exercise: Show that you cannot do better than this without losing distinguishability?

ii) Data compression – Shannon's noiseless coding theorem

Let us consider a message containing a sequence of n binary values 0 and 1, with probability p and $1-p$ (where $0 \leq p \leq 1$), respectively. The key is to divide the possible sequences into two types – sequences that are highly likely to occur known as *typical* sequences, and those that hardly occur, called *atypical* sequences. The question now is how can this be done? As n becomes large, we would expect that with high probability there would be np equal to 0 and $n(1-p)$ equal to 1. So, the typical sequences are those for which this assumption is true. The number of such typical message is the binomial:

$$\binom{n}{np} = \frac{n!}{np!\,(n-np)!}.$$

From Stirling's approximation of large numbers: $\log n! = n\log n - n$. Note we have used base *e* for the time being but you can also take base 2.

$$\begin{aligned}\log\binom{n}{np} &= \log\frac{n!}{np!\,(n-np)!}\\ &\cong n\log n - n - [np\log np - np + n(1-p)\log n(1-p) - n(1-p)]\\ &= nH(p), \end{aligned} \qquad (2)$$

where, $H(p) = -p\log p - (1-p)\log(1-p)$ is the Shannon entropy. Hence, the number of typical sequences or strings of n bits is of order $2^{nH(p)}$ instead of 2^n. To convey essentially all the information carried by a string of n bits, it suffices to choose a block code that assigns a positive integer to each of the typical strings. This block code has about $2^{nH(p)}$ values (all occurring with equal a priori probability), so we may specify any one of the letters using a binary string of length $nH(p)$. Since $0 \leq H(p) \leq 1$ for $0 \leq p \leq 1$, and $H(p) = 1$ only for $p = \frac{1}{2}$, the block code shortens the message for any $p \neq \frac{1}{2}$ (whenever 0 and 1 are not equally probable). This is Shannon's result. The key idea is that we do not need a codeword for every sequence of bits, only for the typical sequences. The probability that the actual message is atypical becomes negligible asymptotically, i.e., in the limit $n \to \infty$.

Exercise: Show that a message containing n values from the set $\{a_1, a_2, \ldots, a_k\}$, each occurring with probability $\{p_1, p_2, \ldots, p_k\}$. Using the law of large numbers, show that the information in a message of n values can be compressed to, $nH(X) = -n\sum_i p_i \log p_i$.

iii) Typical sequences

Note: This subsection is not necessary for exams. For the interested Reader, Chapter 12, Quantum Computation and Quantum Information, Nielsen and Chuang has a more detailed derivation of this.

The above argument can also be rephrased in a slightly different language. A message of n values can be written from the set of letters as $x_1\ x_2\ x_3\ \ldots\ x_n$. With each occurring with a priori probability, $P(x_1x_2x_3\ \ldots\ x_n) = p(x_1)p(x_2)p(x_3)\ \ldots\ p(x_n)$, and

$$\log P(x_1x_2x_3\ \ldots\ x_n) = \log\big(p(x_1)p(x_2)p(x_3)\ \ldots\ p(x_n)\big) = \sum_i \log p(x_i)$$

Applying the central limit theorem to the above relations, one can conclude that for most "typical" sequences the mean of the probability distribution is given by:

$$-\frac{1}{n}\log P(x_1x_2x_3\ \ldots\ x_n) \sim -\langle \log p(x)\rangle \equiv H(X)$$

Typically, the frequency of occurrence of any given letter x_i in the output sequence is close to the probability $p(x_i)$ of the letter in the source under the strong law of large numbers. This allows us to define the notion of a typical sequence with a bit more rigour. For $\epsilon, \delta > 0$, a sequence of letters $x_1x_2x_3\ \ldots\ x_n$ is an ϵ-typical sequence if,

$$\left|-\frac{1}{n}\log P(x_1x_2x_3\ \ldots\ x_n) - H(X)\right| \leq \epsilon$$

for $n \geq N$, with probability $1-\delta$. This gives us;

$$H(X) - \epsilon \leq -\frac{1}{n}\log P(x_1x_2x_3\ \ldots\ x_n) \leq H(X) + \epsilon.$$

Therefore, for $\epsilon, \delta > 0$, and $n \geq N$, any ϵ-typical sequence $x_1\ x_2\ x_3\ \ldots\ x_n$ occurs with probability, $p(x_1)p(x_2)p(x_3)\ \ldots\ p(x_n)$ such that:

$$P_{min} = 2^{-n(H(X)+\epsilon)} \leq p(x_1)p(x_2)p(x_3)\ \ldots\ p(x_n) \leq 2^{-n(H(X)-\epsilon)} = P_{max}.$$

The set of ϵ-typical sequences is denoted by $T(n,\epsilon)$, with its cardinality $n(T)$ bounded by: $n(T)P_{min} \leq \sum_{\text{typ}} p(x_1)\ldots p(x_n) \leq 1$ and $n(T)P_{max} \geq \sum_{\text{typ}} p(x_1)\ldots p(x_n) \geq 1-\delta$, which implies: $1/P_{min} \geq n\big(T(n,\epsilon)\big) \geq (1-\delta)/P_{max}$ or

$$2^{n(H(X)+\epsilon)} \geq n\big(T(n,\epsilon)\big) \geq (1-\delta)2^{n(H(X)-\epsilon)}$$

Therefore, all typical sets can be encoded with a block code $n(H(X) \pm \epsilon)$, with success probability greater than $1-\delta$. But if we try to encode with a code smaller than $n(H(X)-\epsilon)$, the success probability, $p \leq P_{max} \times 2^{n(H(X)-\epsilon')} = 2^{-n(\epsilon'-\epsilon)}$, where ϵ' is positive. As ϵ is arbitrarily small and n large, the probability is very small for $\epsilon' > \epsilon$.

iv) Relative entropy (Kullback-Leibler divergence)

Shannon entropy makes use of the fact that one needs to only use the typical sequences. However, what if the true probabilities are not known – how does that affect the optimal message length.

Let's rephrase, Shannon's coding theorem (dropping noiseless as we won't really get noisy here) says that the effective length of each message is $l_i = -\log_2 p_i$, for some probability distribution $\{p_i\}$. But what if these lengths are not known, i.e., $l'_i = -\log_2 q_i$, where $\{q_i\}$ is the approximate of the actual distribution. On average, the optimal message length differs from the actual by $-\sum_i p_i (\log_2 q_i - \log_2 p_i)$. This difference gives us the relative entropy of $\{p_i\}$ with respect to $\{q_i\}$.

$$R(P||Q) = -\sum_i p_i (\log_2 q_i - \log_2 p_i) = -\sum_i p_i \log_2 \frac{q_i}{p_i} = -\sum_i p_i \log_2 q_i - H(P),$$

where, $H(P)$ is the Shannon entropy. On a more general note, the relative entropy measures the closeness between two probability distributions $\{p_i\}$ and $\{q_i\}$ defined in the same probability space. You can think of $\{p_i\}$ being the experimentally measured probabilities of some classical or quantum system, as compared to $\{q_i\}$, which is the theoretical prediction.

An important condition is that $R(P||Q) \geq 0$, with equality if and only if $\{p_i\} = \{q_i\}$. The proof uses the condition that for any positive x, $\ln x \leq x - 1$, with equality for $x = 1$.

$$R(P||Q) = -\sum_i p_i \log_2 \frac{q_i}{p_i} \geq \frac{1}{\ln 2}\sum_i p_i \left(1 - \frac{q_i}{p_i}\right) = \frac{1}{\ln 2}\sum_i p_i - q_i = 0,$$

since, $\sum_i p_i - \sum_i q_i = 1 - 1 = 0$. $R(P||Q) = 0$, iff $\frac{q_i}{p_i} = 1, \forall\, i$. Relative entropy has no upper bound in general. Using the above relation, we easily get a bound on $H(P)$:

$$R(P||1/d) = -\sum_i p_i \log_2 (1/d) - H(P) = -\log_2 (1/d)\sum_i p_i - H(P) \geq 0.$$

$$-\log_2 (1/d)\sum_i p_i - H(P) = -\log_2 (1/d) - H(P) \geq 0;\ H(P) \leq \log_2 (d).$$

v) Mutual information and conditional entropy

An important question here is how are the information contained in some variable P connected to some other variable Q. In other words, how correlated are two messages

– if $H(P)$ is the resource required by Alice to encode information by a message drawn from the distribution P, but after transmission through a noisy channel Bob can only decode the message from a distribution Q (instead of the original message).

Let $\{p_i\}$ and $\{q_j\}$ be the probability of Alice sending and Bob receiving the message, respectively. The noisy channel can be characterised by the probability $\{q_{j|i}\}$, which is the probability of Bob receiving Q for Alice sending P. This is a property of the noisy channel. The joint probability for Alice and Bob is then given by $r_{i,j} = q_{j|i}p_i$. Now, Bob's lack of knowledge about the original information P, when he knows Q, will be given by $p_{i|j}$. Alternatively, this is the same as any additional information Bob may need to completely know P. Again $r_{i,j} = p_{i|j}q_j$. So, we now have,

$$p_{i|j} = \frac{q_{j|i}p_i}{q_j} = \frac{r_{i,j}}{q_j}.$$

The amount of information, in terms of Shannon entropy, Bob needs is the conditional entropy $H(P|Q)$, which is given by

$$H(P|Q) = -\sum_i p_{i|j} \log p_{i|j} = -\sum_{i,j} p_{i|j}q_j \log p_{i|j} = -\sum_{i,j} r_{i,j} \log p_{i|j};$$

$$-\sum_{i,j} r_{i,j} \log p_{i|j} = -\sum_{i,j} r_{i,j} \log \frac{r_{i,j}}{q_j} = -\sum_{i,j} r_{i,j} \log r_{i,j} + \sum_{i,j} r_{i,j} \log q_j$$

$$\therefore H(P|Q) = -\sum_{i,j} r_{i,j} \log r_{i,j} + \sum_j q_j \log q_j$$

(where $\sum_i r_{i,j} = q_j$ and $\sum_j p_{i|j}q_j = p_{i|j}$)

$$\Rightarrow H(P|Q) = H(P,Q) - H(Q).$$

Similarly, $H(Q|P) = H(P,Q) - H(P)$. So, the information gained by Bob while receiving Q during the noisy transmission of P is given by the difference between $H(P)$ and $H(P|Q)$:

$$H(P:Q) = H(P) - H(P|Q) = H(Q) - H(Q|P) = H(P) + H(Q) - H(P,Q),$$

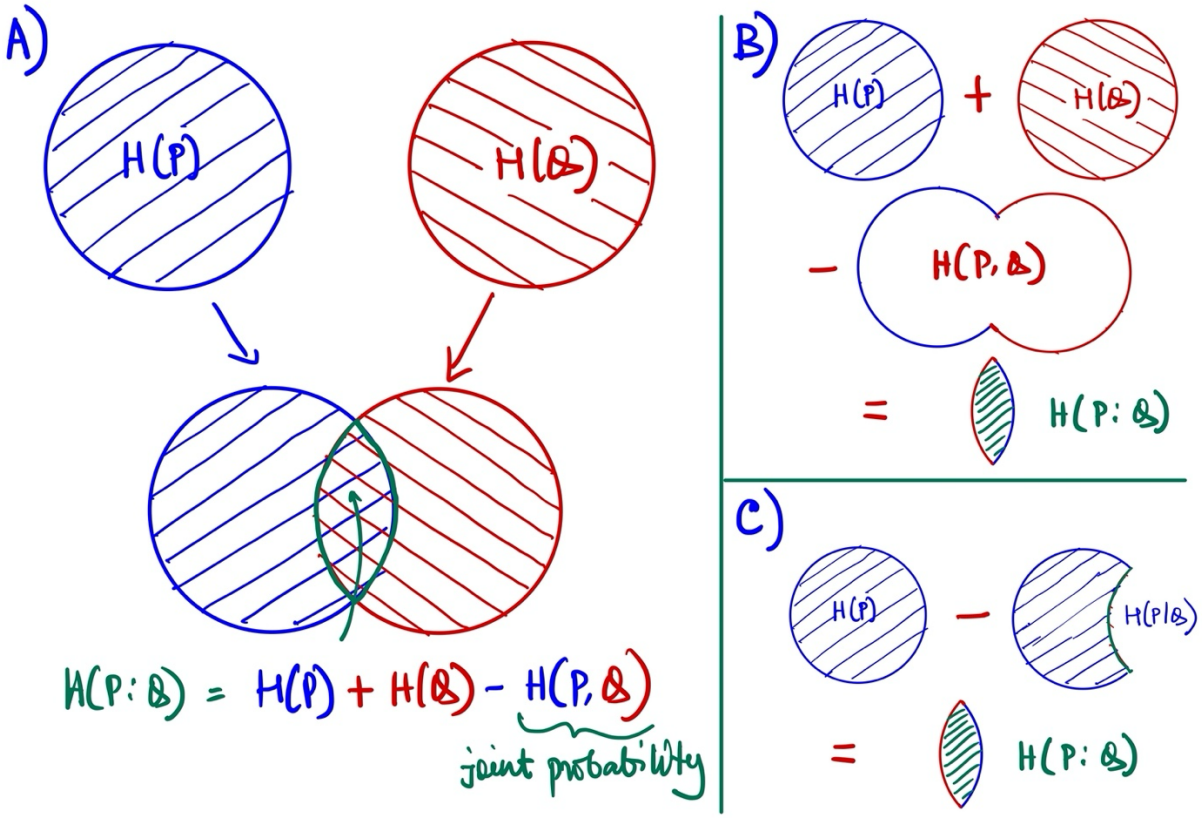
where $H(P:Q)$ is the mutual or common information shared by P and Q, and $H(P,Q)$ is the joint probability between P and Q.

The mutual information is closely connected to the rate at which information can be transmitted across noisy communication channels. The capacity C is the maximum possible rate of communication that can be achieved using a noisy channel, with probability of error tending to zero as the number of symbols (n) in the message goes to infinity. Shannon's noisy channel coding theorem states that (without proof),

$$C = \max_{\{p_i\}} H(P:Q),$$

where $H(P:Q)$ is the mutual information between P and Q. The maximization is over all possible inputs $\{p_i\}$ for a particular noisy channel characterised by $\{q_{j|i}\}$.

The mutual information is easy to visualise (but not always true) in terms of Venn diagrams (see (A) and (B) in the figure below).



vi) Properties of Shannon entropy

Both **joint entropy and mutual information are symmetric** by definition, i.e.,

$$H(P,Q) = H(Q,P); \qquad H(P:Q) = H(Q:P).$$

Conditional entropies are always positive. Let us say, the joint probability here is $r_{i,j} = p_i q_{j|i}$, where $q_{j|i} \geq 0$, is the probability of q when p is known, such that

$$H(P,Q) = -\sum_{i,j} r_{i,j} \log r_{i,j} = -\sum_{i,j} p_i q_{j|i} \log p_i q_{j|i} = -\sum_i p_i \log p_i - \sum_{i,j} q_{j|i} \log q_{j|i}$$

$$= H(P) - \sum_{i,j} q_{j|i} \log q_{j|i} = H(P) + H(Q|P).$$

We have used the relation: $\sum_j r_{i,j} = p_i$ and $\sum_i r_{i,j} = q_j$. Since, $-\log q_{j|i} \geq 0$ ($q_{j|i}$ is just probability here i.e., $q_{j|i} \geq 0$, $\sum_j q_{j|i} = 1$). Therefore, $H(Q|P) \geq 0$, with equality if and only if Q is completely determined by P.

This implies that $H(P|Q) = H(P) - H(P:Q) \geq 0$, and therefore, $H(P:Q) \leq H(P)$. Moreover, $H(P,Q) = H(P) + H(Q|P)$, which implies, $H(P,Q) \geq H(P)$.

Shannon entropy is subadditive: $H(P,Q) \leq H(P) + H(Q)$.

$$H(P,Q) - H(P) - H(Q) = -\sum_{i,j} r_{i,j} \log r_{i,j} + \sum_i p_i \log p_i + \sum_j q_j \log q_j,$$

$$= -\sum_{i,j} r_{i,j} \log r_{i,j} + \sum_{i,j} r_{i,j} (\log p_i + \log q_j) = \sum_{i,j} r_{i,j} \log \frac{p_i q_j}{r_{i,j}},$$

(where $\sum_{i,j} r_{i,j} = \sum_i p_i$ and $\sum_{i,j} r_{i,j} = \sum_j q_j$).

$\sum_{i,j} r_{i,j} \log \frac{p_i q_j}{r_{i,j}} \leq \sum_{i,j} r_{i,j} \left(\frac{p_i q_j}{r_{i,j}} - 1 \right) = \sum_{i,j} r_{i,j} - p_i q_j = 0$, using $\ln x \leq (x-1)$.

Therefore, $H(P,Q) - H(P) - H(Q) \leq 0$, which gives us, $H(P,Q) \leq H(P) + H(Q)$. Importantly, $H(P,Q) = H(P) + H(Q)$, when $r_{i,j} = p_i q_j$, which implies that the events are independent.

Subadditivity implies that mutual information is always positive:

$$H(P:Q) = H(P) + H(Q) - H(P,Q) \geq 0.$$

Also, by extension, $H(P:Q) \leq H(P)$ and $H(P:Q) = H(Q:P) \leq H(Q)$.

Shannon entropy is also strongly subadditive (without proof):

$$H(P,Q,X) + H(Y) \leq H(P,Q) + H(X,Y).$$

Mutual information can be written as relative entropy between the joint distribution $\{r_{i,j}\}$ and the completely uncorrelated $\{p_i q_j\}$: $H(P:Q) = R(\{r_{i,j}\}||\{p_i q_j\})$.

$$H(P:Q) = H(P) + H(Q) - H(P,Q) = \sum_{i,j} r_{i,j} \log r_{i,j} - \sum_{i,j} r_{i,j} (\log p_i + \log q_j),$$

$$H(P:Q) = -\sum_{i,j} r_{i,j} \log \frac{p_i q_j}{r_{i,j}} = R(\{r_{i,j}\}||\{p_i q_j\}).$$

vii) Von Neumann entropy

Shannon's noiseless coding theorem and Shannon entropy makes use of the fact that one needs to only use the typical sequences in the code and the rest can be ignored. This takes place in the asymptotic limit without loss of any significant fidelity. The question now is what is the quantum analogue of such a theorem.

Consider now a message consisting of *n* values, where each value is chosen from an ensemble of pure states $\{p_i, |\phi_i\rangle\}$. Here, $|\phi_i\rangle$'s need not be mutually orthogonal. So, each value or letter in the message is given by a density matrix: $\rho = \sum_i p_i |\phi_i\rangle\langle\phi_i|$.

The message with n values is given by the joint density matrix:

$$\rho^n = \rho \otimes \rho \otimes \rho \otimes \ldots \otimes \rho.$$

Similar to the classical case, we now ask how can we compress this information. Can we devise a quantum code that enables us to compress the data to a smaller Hilbert space, but again, without losing out on the fidelity of the encoded message? The optimal rate of compression was obtained by Ben Schumacher, and he provided a familiar answer. The best we can do, with very good fidelity, as $n \to \infty$, is to compress the message to a Hilbert space with $\log(\dim \mathcal{H}) = n\, S(\rho) = -n \operatorname{tr}(\rho \log \rho)$. So similar to the Shannon entropy for classical information, the corresponding quantity for qubits is the von Neumann entropy:

$$S(\rho) = -\operatorname{tr}(\rho \log \rho).$$

The most straightforward definition of the von Neumann entropy is in terms of its eigen decomposition. If $\rho = \sum_i \lambda_i |e_i\rangle\langle e_i|$, then we have,

$$S(\rho) = H(\{\lambda_i\}) = -\sum_i \lambda_i \log \lambda_i.$$

Therefore the von Neumann entropy reduces to the classical Shannon entropy over the probabilities $\{\lambda_i\}$, arising from the eigen decomposition of ρ, given by $\{\lambda_i, |e_i\rangle\}$.

An interesting thing to note here is that the original message consisted of sending a state from the ensemble $\{p_i, |\phi_i\rangle\}$, where $\{p_i\}$ is a set of probabilities. However, if $\{|\phi_i\rangle\}$ is not a mutually orthogonal set, then in general it can be shown that $S(\rho) < H(\{p_i\})$. So, the coding can be further compressed using quantum states. The catch here is that, since $\{|\phi_i\rangle\}$ is nonorthogonal, it cannot be perfectly distinguished by the receiver. On the other hand, if $\{|\phi_i\rangle\}$ is mutually orthogonal, then $\{p_i, |\phi_i\rangle\} = \{\lambda_i, |e_i\rangle\}$ is the eigen decomposition of ρ, and $S(\rho) = H(\{p_i\})$. So, for a set of orthogonal pure states, which are completely distinguishable from each other, the compression reduces to that for classical information. In fact, the encoder can get away with sending the classical information $\{p_i, P_i\}$, and for each P_i the receiver simply recreates the distinguishable state $|\phi_i\rangle$ from the set $\{|\phi_i\rangle\}$. In some sense, one can surmise that classical information was lost while compressing the non-distinguishable quantum states.

Importantly, von Neumann entropy captures both the classical and quantum aspects of information stored in a quantum state – we will later see an interesting example of finding how much classical information can be stored in quantum states. As such, von Neumann entropy forms the cornerstone of quantum information theory and is essential in understanding the encoding and transmission of both classical and quantum information over quantum channels and also in quantifying entanglement.

vii) Quantum relative entropy, mutual information and quantum discord

Similar to relative entropy between two probability distributions, the quantum relative entropy of ρ with respect to σ is defined as

$$R(\rho||\sigma) = \mathrm{Tr}[\rho \log \rho] - \mathrm{Tr}[\rho \log \sigma] = -\mathrm{Tr}[\rho \log \sigma] - S(\rho).$$

The quantum relative entropy is non-negative, as given by the **Klein's inequality**.

Using the respective eigen decomposition, we have $\rho = \sum_i p_i \, |i\rangle\langle i|$ and $\sigma = \sum_j q_i \, |j\rangle\langle j|$. If ρ commutes with σ, then they share the same eigenbasis and $R(\rho||\sigma) = R(\{p_i||q_i\})$. In general, $R(\rho||\sigma) = \sum_i p_i \log p_i - \sum_i \langle i|\rho \log \sigma \, |i\rangle$, where

$$\langle i|\rho \log \sigma \, |i\rangle = p_i \langle i| \log \sigma \, |i\rangle = p_i \langle i| \left(\sum_j \log q_i \, |j\rangle\langle j|\right) |i\rangle = p_i \sum_j \log q_i \;\; P_{ij},$$

where $P_{ij} = \langle i|j\rangle\langle j|i\rangle \geq 0$; $\sum_i P_{ij} = \sum_j P_{ij} = 1$. This gives us

$$R(\rho||\sigma) = \sum_i p_i \left(\log p_i - \sum_j P_{ij} \log q_i \right) \geq \sum_i p_i \, (\log p_i - \log r_i \,) \geq R(\{p_i||r_i\}),$$

where $r_i = \sum_j P_{ij} \; q_i$. Since log is a concave function, $\sum_j P_{ij} \log q_i \leq \log \sum_j P_{ij} \; q_i = \; \log r_i$. $R(\{p_i||r_i\}) \geq 0$, is the classical relative entropy. Therefore,

$$R(\rho||\sigma) \geq 0.$$

The equality holds if and only if $p_i = q_i \; \forall \, i$, and P_{ij} is a permutation matrix. This simplifies further to the statement that P_{ij} is an identity matrix, and $\rho = \sigma$.

On a similar note, the joint entropy for composite systems is given by the Von Neumann entropy of the composite density matrix. For subsystems A and B, the joint entropy, using notations similar to classical entropy, is given by

$$S(A, B) = S(\rho_{AB}) - \mathrm{Tr}[\rho_{AB} \log \rho_{AB}] = \; -\sum_{i,j} \lambda_{ij} \log \lambda_{ij},$$

where, $\rho_{AB} = \sum_{i,j} \lambda_{ij} |i\rangle\langle i| \otimes |j\rangle\langle j|$ is the density matrix of the joint state. An important corollary is the state $\rho_{AB} = \rho_A \otimes \rho_B$, for which, $S(A, B) = S(\rho_{AB}) = \; S(\rho_A) + S(\rho_B)$.

The quantum mutual information is similarly defined as,

$$S(A{:}\,B) = S(A) + S(B) - S(A, B) = S(\rho_A) + S(\rho_B) - S(\rho_{AB}).$$

It should be clear that, unlike relative entropy, all the von Neumann entropies in mutual information can be calculated in their respective eigenbasis. In other words, $S(\rho_A)$ and $S(\rho_B)$ can be computed in their respective eigenbasis, regardless of whether ρ_A commutes with ρ_B. Also, similar to the classical case, it is straightforward to show that $S(A{:}\,B) = R(\rho_{AB}||\rho_A \otimes \rho_B) \geq 0$. Also, $S(A{:}\,B) \leq 2 \log d$, when ρ_{AB} is pure and $\rho_A = \rho_A = \mathbb{I}_d/d$, where $\mathbb{I}_d$ is the $d \; \times d$ identity matrix.

The conditional entropy is a bit tricky in the quantum regime – the usual definition of $S(A|B) = S(A,B) - S(B) = S(\rho_{AB}) - S(\rho_B)$, can actually be negative. This is a positive quantity in the classical regime and intuitively implies that the disorder in a subsystem cannot be larger than the system. But this intuitiveness is thrown out of the window in the quantum regime. As discussed earlier in several instances, the reduced state of a pure state can be a mixed state – see the discussion on density operators and reduced states. Hence, it is quite natural for the global system to have zero entropy (pure state), with the subsystem having finite entropy. Moreover, what does conditional entropy really mean in the quantum context!

This issue arises strongly, while considering the quantum mutual information. While,

$$S(A:B) = S(\rho_A) + S(\rho_B) - S(\rho_{AB}) \neq S(\rho_A) - S(\rho_{A|B})$$
$$\implies I(A:B) \neq J(A:B).$$

The quantity $S(\rho_{A|B})$ is a quantum generalization of the conditional entropy, and implies the knowledge of A upon knowing B. This means the state B has to be measured to make any sense of the statement, and in general such measurements are not unique (we will run into similar problems all the time in quantum information). Optimizing over all such measurements (say $\{M\}$), gives you two different measures of $S(A:B)$, i.e., $I(A:B)$ and $J(A:B)$, and the difference between the two is called quantum discord $D(\rho_{AB})$, which is a measure of quantumness in the system (as the difference does not arise in the classical world). It is defined as follows

$$D(\rho_{AB}) = I(A:B) - \max_{\{M\}} J(A:B).$$

Exercise *Use the expression for $I(A:B)$ in the previous exercise to compute the quantum mutual information of the following bipartite quantum states.*

1. $\rho_{AB} = \frac{1}{2}\mathbb{1}_A \otimes \frac{1}{2}\mathbb{1}_B = \frac{1}{4}\mathbb{1}_{AB}$.
2. $\rho_{AB} = \frac{1}{2}(|00\rangle\langle 00| + |11\rangle\langle 11|)$. *(Maximal classical correlations)*
3. $\rho_{AB} = |\phi^+\rangle\langle\phi^+|$, *a Bell state of a two qubits.*
4. $\rho_{AB} = |\varphi\rangle\langle\varphi|$, *where* $|\varphi\rangle = (U_A \otimes U_B)|\phi^+\rangle$ *for some unitaries* U_A *and* U_B.

viii) Other properties of the von Neumann entropy

The **von Neumann entropy is non-negative**, which is clear from the definition, i.e., $S(\rho) \geq 0$. It is zero for pure states, which can always be written as rank-1 matrix with eigenvalue 1. Using the fact that $R(\rho||\sigma) \geq 0$, we have $R(\rho||\mathbb{I}_d/d) = -S(\rho) + \log d \geq 0$, where $\mathbb{I}_d$ is the d x d identity matrix. Therefore, $S(\rho) \leq \log d$.

From the Schmidt decomposition we know that for pure ρ_{AB}, the states $\rho_A = \mathrm{Tr}_B[\rho_{AB}]$ and $\rho_B = \mathrm{Tr}_A[\rho_{AB}]$ have the same eigenvalues. This implies $S(\rho_A) = S(\rho_B)$, if ρ_{AB} is a pure state. Importantly, $S(\rho)$ is invariant under unitary operations, $S(\rho) = S(U\rho U^{-1})$. This is true because the entropy only depends on the eigenvalues.

If $\rho = \sum_i p_i \rho_i$, such that ρ_i have support on orthogonal subspaces[1]. Then the von Neumann entropy of ρ is given by

$$S(\rho) = S\left(\sum_i p_i \rho_i\right) = -\sum_{i,j} p_i \lambda_{ij} \log p_i \lambda_{ij} = -\sum_i p_i \log p_i - \sum_i p_i \sum_j \lambda_{ij} \log \lambda_{ij},$$

$$\therefore\ S(\rho) = H(p_i) + \sum_i p_i S(\rho_i).$$

Here, $S(\rho_i) = \sum_j \lambda_{ij} \log \lambda_{ij}$. Note that, if ρ_i is a set of orthogonal projectors, we recover the definition of $S(\rho)$. Similarly, if we consider a $\rho = \sum_i p_i |i\rangle\langle i| \otimes \rho_i$, we get

$$S\left(\sum_i p_i |i\rangle\langle i| \otimes \rho_i\right) = H(p_i) + \sum_i p_i S(|i\rangle\langle i| \otimes \rho_i) = H(p_i) + \sum_i p_i \{S(|i\rangle\langle i|) + S(\rho_i)\}$$

$$= H(p_i) + \sum_i p_i \{0 + S(\rho_i\} = H(p_i) + \sum_i p_i S(\rho_i).$$

$$\therefore\ S\left(\sum_i p_i |i\rangle\langle i| \otimes \rho_i\right) = H(p_i) + \sum_i p_i S(\rho_i).$$

Subadditivity and the triangle inequality of von Neumann entropy

As shown earlier, $S(\rho_{AB}) = S(\rho_A) + S(\rho_B)$ if $\rho_{AB} = \rho_A \otimes \rho_B$. From Klein's inequality we know that, $S(\rho_{AB}||\rho_A \otimes \rho_B) \geq 0$, which implies that:

[1] Support of a matrix is the space spanned by its non-zero eigenvalues. The kernel of a matrix is the space spanned by its zero eigenvalues. Also see footnote 2.

$$S(\rho_{AB}||\rho_A \otimes \rho_B) = -S(\rho_{AB}) - \mathrm{Tr}[\rho_{AB}(\log \rho_A + \log \rho_B)]$$
$$= -S(\rho_{AB}) - \mathrm{Tr}[\rho_A \log \rho_A] - \mathrm{Tr}[\rho_{\boldsymbol{B}} \log \rho_B] = -S(\rho_{AB}) + S(\rho_A) + S(\rho_B) \geq 0.$$

This gives us the subadditivity:

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

Consider the purification of ρ_{AB} such that ρ_{ABE} is pure. Hence, $S(\rho_{AB}) = S(\rho_E)$ and $S(\rho_{AE}) = S(\rho_B)$. From subadditivity: $S(\rho_{AE}) \leq S(\rho_A) + S(\rho_E)$, which upon replacing, gives us: $S(\rho_B) \leq S(\rho_A) + S(\rho_{AB})$ or $S(\rho_{AB}) \geq S(\rho_B) - S(\rho_A)$. Along similar lines we have: $S(\rho_{BE}) \leq S(\rho_B) + S(\rho_E)$, which gives us $S(\rho_{AB}) \geq S(\rho_A) - S(\rho_E)$. Therefore, we end up with the triangle inequality:

$$S(\rho_{AB}) \geq |S(\rho_A) - S(\rho_B)|.$$

Concavity of von Neumann entropy

Let, $\rho = \sum_i p_i \rho_i$, where ρ_i is any density matrix occurring with probability p_i in the ensemble. If ρ_i belongs to system A (say), we can add an auxiliary system B, such that $\rho_{AB} = \sum_i p_i \, \rho_i \otimes |i\rangle\langle i|$, and $\rho_A = \mathrm{Tr}_B[\rho_{AB}] = \sum_i p_i \rho_i$ and $\rho_B = \mathrm{Tr}_A[\rho_{AB}] = \sum_i p_i |i\rangle\langle i|$.

Now, as discussed earlier: $S(\rho_{AB}) = H(p_i) + \sum_i p_i S(\rho_i)$. Also, $S(\rho_{AB}) \leq S(\rho_A) + S(\rho_B)$ from subadditivity. This gives us:

$$S(\rho_{AB}) = H(p_i) + \sum_i p_i S(\rho_i) \leq S\left(\sum_i p_i \rho_i\right) + S\left(\sum_i p_i |i\rangle\langle i|\right) = S(\rho) + H(p_i).$$

Therefore, $H(p_i) + \sum_i p_i S(\rho_i) \leq S(\rho) + H(p_i)$, which gives us

$$S(\rho) = S(\textstyle\sum_i p_i \rho_i) \geq \sum_i p_i S(\rho_i),$$

for $\rho = \sum_i p_i \rho_i$. Hence, von Neumann entropy is a concave function.

A more general statement about a density matrix $\rho = \sum_i p_i \rho_i$ (without proof):

$$\sum_i p_i S(\rho_i) \leq S(\rho) \leq \sum_i p_i S(\rho_i) + H(\{p_i\}).$$

If $\rho_i = |\phi_i\rangle\langle\phi_i|$, we have $S(\rho) \leq H(\{p_i\})$, which was discussed earlier.

Strong subadditivity of von Neumann entropy (without proof):

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

The proofs for the above are in Chapter 11, of Nielsen and Chuang.

ix) Von Neumann entropy and data compression

Note: This subsection is not necessary for exams.

Let us demonstrate the compression protocol using a simple example from the lecture notes of John Preskill. Let us assume that the qubits are taken from an ensemble:

$$\begin{aligned} |\uparrow_z\rangle &= \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad p = \tfrac{1}{2}, \\ |\uparrow_x\rangle &= \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \quad p = \tfrac{1}{2}, \end{aligned}$$

so for each letter, the density matrix is given by:

$$\begin{aligned} \boldsymbol{\rho} &= \frac{1}{2}|\uparrow_z\rangle\langle\uparrow_z| + \frac{1}{2}|\uparrow_x\rangle\langle\uparrow_x| \\ &= \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix}. \end{aligned}$$

The eigenstates and eigenvalues are:

$$\begin{aligned} |0'\rangle \equiv |\uparrow_{\hat{n}}\rangle &= \begin{pmatrix} \cos\frac{\pi}{8} \\ \sin\frac{\pi}{8} \end{pmatrix}, \\ |1'\rangle \equiv |\downarrow_{\hat{n}}\rangle &= \begin{pmatrix} \sin\frac{\pi}{8} \\ -\cos\frac{\pi}{8} \end{pmatrix}; \end{aligned}$$

$$\begin{aligned} \lambda(0') &= \frac{1}{2} + \frac{1}{2\sqrt{2}} = \cos^2\frac{\pi}{8}, \\ \lambda(1') &= \frac{1}{2} - \frac{1}{2\sqrt{2}} = \sin^2\frac{\pi}{8}; \end{aligned}$$

The eigenstate $|0'\rangle$ has relatively large and equal overlap with both the initial states and is therefore “likely”, while the state $|1'\rangle$ has relatively low but again equal overlap with the initial states, and is “unlikely”, as shown below:

$$\begin{aligned} |\langle 0'|\uparrow_z\rangle|^2 = |\langle 0'|\uparrow_x\rangle|^2 &= \cos^2\frac{\pi}{8} = .8535, \\ |\langle 1'|\uparrow_z\rangle|^2 = |\langle 1'|\uparrow_x\rangle|^2 &= \sin^2\frac{\pi}{8} = .1465. \end{aligned}$$

So if one does not know whether the state is $|\uparrow_z\rangle$ or $|\uparrow_x\rangle$, but simply by guessing $|\psi\rangle = |0'\rangle$, we can obtain maximal fidelity, $F = 0.8535$:

$$F = \frac{1}{2}|\langle\uparrow_z|\psi\rangle|^2 + \frac{1}{2}|\langle\uparrow_x|\psi\rangle|^2$$

Now if we need to send three qubits, is there a way to achieve this with lesser resource, say only two qubits. One way is to send two qubits (with fidelity equal to 1) and the

best the receiver can do for the third is guess $|0'\rangle$, so that the overall fidelity, $F = 0.8535$. But can this be improved!!

Let us consider the eigen basis of ρ^3, composed of the single qubit "likely" ($|0'\rangle$), and "unlikely" ($|1'\rangle$). So, if the state to send is $|\psi\rangle = |\psi_1\rangle|\psi_2\rangle|\psi_3\rangle$ (where each can be $|\uparrow_z\rangle$ or $|\uparrow_x\rangle$), then using the eigen basis we have:

$$
\begin{aligned}
&|\langle 0'0'0'|\psi\rangle|^2 = \cos^6\left(\frac{\pi}{8}\right) = .6219,\\
&|\langle 0'0'1'|\psi\rangle|^2 = |\langle 0'1'0'|\psi\rangle|^2 = |\langle 1'0'0'|\psi\rangle|^2 = \cos^4\left(\frac{\pi}{8}\right)\sin^2\left(\frac{\pi}{8}\right) = .1067,\\
&|\langle 0'1'1'|\psi\rangle|^2 = |\langle 1'0'1'|\psi\rangle|^2 = |\langle 1'1'0'|\psi\rangle|^2 = \cos^2\left(\frac{\pi}{8}\right)\sin^4\left(\frac{\pi}{8}\right) = .0183,\\
&|\langle 1'1'1'|\psi\rangle|^2 = \sin^6\left(\frac{\pi}{8}\right) = .0031.
\end{aligned}
$$

Now, we can divide the state space and identify a "likely" subspace spanned by the states $\{|0'0'0'\rangle, |0'0'1'\rangle, |0'1'0'\rangle, |1'0'0'\rangle\}$, and its orthogonal complement which is the "unlikely" subspace. If we now make a measurement that projects the signal into the "likely" subspace, the probability is

$$P_{likely} = .6219 + 3(.1067) = .9419,$$

and the probability of project into the "unlikely" subspace is

$$P_{unlikely} = 3(.0183) + .0031 = .0581.$$

The protocol can be carried out in two steps. The sender first performs a unitary transformation (U) that rotates the four "likely" and four "unlikely" subspace to the form, $|\cdot\rangle|\cdot\rangle|0\rangle$ and $|\cdot\rangle|\cdot\rangle|1\rangle$, respectively. If the sender measures the third qubit as $|0\rangle$, projects the state to the "likely" basis, and the compressed two qubit state $|\psi_{\text{com}}\rangle$ is sent to the receiver. The received qubits are appended with the state $|0\rangle$, and a U^{-1} is applied to decompress the state, i.e., $U^{-1}(|\psi_{\text{com}}\rangle|0\rangle)$. Now, if the sender measures $|0\rangle$, and projects the state to the "unlikely" basis, they sends the state that can be decompresses to $|0'0'0'\rangle$ by the receiver, i.e., $|0'0'0'\rangle = U^{-1}(|\psi_{\text{com}}\rangle|0\rangle)$. This achieves maximum fidelity.

So for the three qubit signal state $|\psi\rangle$, the sender sends two qubits and the receiver obtains the state:

$$|\psi\rangle\langle\psi| \to \boldsymbol{\rho}' = \mathbf{E}|\psi\rangle\langle\psi|\mathbf{E} + |0'0'0'\rangle\langle\psi|(1-\mathbf{E})|\psi\rangle\langle 0'0'0'|,$$

where $\mathbf{E}$ is the projection to the "likely" subspace. The fidelity achieved is

$$\begin{aligned} F &= \langle\psi|\boldsymbol{\rho}'|\psi\rangle = (\langle\psi|\mathbf{E}|\psi\rangle)^2 + (\langle\psi|(1-\mathbf{E})|\psi\rangle)(\langle\psi|0'0'0'\rangle)^2 \\ &= (.9419)^2 + (.0581)(.6219) = .9234. \end{aligned}$$

This protocol is better than simply sending two of three qubits, with perfect fidelity. The von Neumann entropy of the one qubit ensemble is

$$S(\boldsymbol{\rho}) = H\left(\cos^2\frac{\pi}{8}\right) = .60088$$

For increasing number of qubits the length can be compressed to 0.609 with very high fidelity and the 3 qubits can be compressed to roughly $0.609 \times 3 = 1.827 \approx 2$ qubits.

Exercise: Repeat the above problem, but now consider the initial probabilities for the state to be in $|\uparrow_z\rangle$ or $|\uparrow_x\rangle$ to be $p = \frac{2}{3}$ and $p = \frac{1}{3}$, respectively. Calculate the fidelities, the "likely" and "unlikely" subspace and the von Neumann entropy.

x) Schumacher's data compression – Typical subspace

Note: This subsection is not necessary for exams.

The compressibility of quantum information can be quantified by taking the forward the notion of a typical sequence to that of a typical subspace. We have already seen this in the context of "likely" and "unlikely" subspaces. So, the main aspect of Schumacher's noiseless quantum coding theorem is that we can use only the typical subspace and forget its orthogonal complement, without much loss of fidelity, in the limit of $n \to \infty$. In this section, as in the case of typical sequences, we look at a very simple proof of the above, without going into too much detail of the derivation. Interested Readers are advised to consult the supplementary reading material.

Suppose, the density operator associated with the signal has an orthonormal decomposition given by:

$$\rho = \sum_x p_x |x\rangle\langle x|,$$

where, $|x\rangle$ is the orthonormal basis and p_x are the corresponding eigenvalues, which follow the same rules as a probability distribution, i.e., they are non-negative and sum to unity. Furthermore, $S(\rho) = H(p_x)$, where the quantities are the von Neumann and Shannon entropies, respectively. Therefore, we do have an $\epsilon - \text{typical}$ sequence here, $x_1, x_2, x_3, \ldots\ x_n$, for which

$$\left|\frac{1}{n}\log\frac{1}{p(x_1)p(x_2)p(x_3)\ldots\ p(x_n)} - S(\rho)\right| \leq \epsilon, \tag{8}$$

in a similar manner to the classical definition. An $\epsilon - \text{typical}$ state is a state $|x_1\rangle|x_2\rangle \ldots |x_n\rangle$ for which the sequence $x_1, x_2 \ldots\ x_n$ is $\epsilon - \text{typical}$. The $\epsilon - \text{typical}$ subspace is the space spanned by all $\epsilon - \text{typical}$ states $|x_1\rangle|x_2\rangle \ldots |x_n\rangle$, denoted by $T(n, \epsilon)$ and the projector onto this subspace is defined by

$$P(n, \epsilon) = \sum_{x\ \epsilon-\text{typical}} |x_1\rangle\langle x_1| \otimes |x_2\rangle\langle x_2| \otimes \ldots \otimes |x_n\rangle\langle x_n|. \tag{9}$$

The theorem of typical subspaces then follow similar to the typical sequence theorem and can be found in the book by Nielsen and Chuang.

xi) Encoding classical information in quantum states

In the last two sections we discussed how messages using both classical information and quantum information can be compressed and encoded, which led us to the central idea of entropy as information.

But now we ask a different question. How much classical information can be packed into a qubit? Or in other words, if classical information is packed into quantum states how much of it is realistically accessible. We know that orthogonal states behave like classical registers, but can we communicate more or gain significant advantage by encoding classical bits in mutually nonorthogonal qubits.

At first it seems rather unwise to consider the prospect of storing classical information in mutually nonorthogonal quantum states, as they cannot be distinguished. However, this may be a consequence of circumstances rather than choice. For example, an

optical engineer may want to achieve higher output of classical information in an optical fibre, and therefore increases the number of photons transmitted. But this may lead to photon wavepackets overlapping and lose distinguishability. Other experiments may want to make couple quantum systems with optomechanical devices to make highly sensitive detectors and the knowledge of classical information stored in the quantum states can be important.

The qubit Bloch sphere is bigger than a classical bit state space, and has dimension $d^2 - 1 = 3$, so one may naively expect to be able to use a qubit to communicate more than a single bit – however Holevo's bound tells us that this is not the case. In the absence of entanglement, only a single qubit can only be used to communicate a single classical bit.

Let us again meet our friends Aditi and Bharat, who are from now on the quintessential partners in our study of quantum information. They often share quantum states, entanglement and information and help us study and build a better understanding of quantum systems.

Now let Aditi and Bharat share a quantum state ρ_{AB}:

The state is a classical-quantum state, where Aditi holds a classical register (a set of orthogonal states) and Bharat holds an arbitrary quantum state. We can view this as Aditi having encoded a classical label k into a quantum state ρ_k with probability p_k and then sending it to Bharat for him to do measurements and figure out k. The reduced states for Aditi and Bharat are,

$$\rho_A = \sum_k p_k |k\rangle\langle k| \text{ and } \rho_B = \sum_k p_k \rho_k.$$

xii) Accessible information and the Holevo bound

So, how can Bharat access the classical information encoded by Aditi in their joint state, ρ_{AB}. Bharat can perform a POVM, $\mathcal{M} = \{M_i = B_i^\dagger B_i\}$ with outcomes m_i. The outcome can then be saved in a classical register (since we are only interested in the classical content – or the classical mutual information between the Aditi and Bharat). So the map looks something like:

$$\mathcal{M}(X) = \sum_i \mathrm{Tr}\,[M_i X]|i\rangle\langle i|.$$

Therefore if Bharat's post POVM state looks like:

$$\sigma_{AB} = (\mathbb{I} \otimes \mathcal{M})\rho_{AB} = \sum_k p_k |k\rangle\langle k|_A \otimes \sum_i \mathrm{Tr}\,[M_i \rho_k]|i\rangle\langle i|_B$$

$$= \sum_{i,k} p_k\, \mathrm{Tr}[M_i \rho_k]|k\rangle\langle k|_A \otimes |i\rangle\langle i|_B.$$

This is now a classical – classical state shared by Aditi and Bharat, and is fully described by a classical distribution given by $\{p_k \mathrm{tr}(M_i \rho_k)\}$, which has the correlations between Bharat's measurement outcome i and Aditi's classical register k.

The classical mutual information between Bharat's measurement and Aditi's classical state, when Bharat uses the POVM $\{M_i = B_i^\dagger B_i\}$ for his measurements, is given by,

$$H(A{:}B) = R(\sigma_{AB}\,||\sigma_A \otimes \sigma_B).$$

Importantly, Bharat can maximize these correlations over all possible POVMs, and this maximal classical information he can obtain is called the accessible information, and is given by:

$$I_{acc} = \max_{\{M_i\}} H(A{:}B).$$

However, in general the accessible information is hard to compute because the optimization over the set of all POVMs. However, there is an interesting bound that comes from the monotonicity of quantum relative entropy (without proof). For any density matrix ρ and σ, and quantum operation $\mathcal{E}(X)$, which is completely positive and trace preserving, the following relation holds: $R(\rho||\sigma) \geq R(\mathcal{E}(\rho)\,||\mathcal{E}(\sigma))$.

As such, $R\big((\mathbb{I} \otimes \mathcal{M})\rho_{AB}\,||(\mathbb{I} \otimes \mathcal{M})\rho_A \otimes \rho_B\big) = R(\sigma_{AB}\,||\sigma_A \otimes \sigma_B) \leq R(\rho_{AB}\,||\rho_A \otimes \rho_B)$.

Now, $R(\rho_{AB}\,||\rho_A \otimes \rho_B) = S(A{:}B) = S(\rho_A) + S(\rho_B) - S(\sum_k p_k |k\rangle\langle k| \otimes \rho_k)$,

$$= S(\rho_A) + S(\rho_B) - S(\rho_A) - \sum_k p_k\, S(\rho_k).$$

$$\chi = S(\rho_B) - \sum_k p_k\, S(\rho_k),$$

is the Holevo quantity, and we obtain a very important bound on the accessible information called the Holevo bound, which is then given by the relation:

$$I_{acc} = \max_{\{M_i\}} H(A{:}B) \leq \chi.$$

So, the maximum information that Bharat can access from Aditi's encoded classical information using a classical-quantum state is bounded by the Holevo information.

How much classical information can be embedded in a qubit? We need to find the maximum Holevo quantity, χ, for $d = 2$. Let $\rho_k = |k\rangle\langle k|$, be pure for all k, and the maximum value of $S(\rho_B)$ is for $\rho_B = \mathbb{I}/2$, which gives us $\chi = 1$. Therefore, $I_{acc} = 1$, which means that no more than a single bit can be encoded in a qubit for communication.

xiii) Mixed state coding and Holevo information

From Schumacher's compression, we know that an n qubit message chosen from an ensemble of pure quantum states $\{p_i, |\phi_i\rangle\}$, can be compressed to $nS(\rho)$, where $\rho = \sum_i p_i |\phi_i\rangle\langle\phi_i|$. We also know that $S(\rho) = -\operatorname{tr}(\rho \log \rho) = H(\{\lambda_i\}) \leq H(\{p_i\})$, with the last equality holding if $\{|\phi_i\rangle\}$ is an orthonormal, completely distinguishable set of pure states, and the protocol reduces to communicating classical bits, and the compression is $S(\rho) = H(\{p_i\})$.

However, what if we chose from an ensemble of mixed states that were available to us, i.e., $\{p_i, \rho_i\}$ and $\rho = \sum_i p_i\, \rho_i$. In such a case, the optimal compression is not clearly known. Say, we have $p_1 = 1$ and $p_i = 0\ \forall\ i \neq 1$. Hence the message reduces to $\{\rho = \rho_1\}$, with unit probability, and the compression should ideally be 0. Why? If Alice has only ρ_1 to communicate, Bob can actually reproduce any n qubit message $({\rho_1}^{\otimes n})$ at his end without any communication. However, since ρ_1 is a mixed state, $S(\rho) > 0$.

Let us now look at this a bit more closely, if $\{\rho_i\}$ is an orthogonal set of density matrices i.e., it has a support that is orthogonal to each other[2] or $\mathrm{Tr}[\rho_i \rho_j] = 0, \forall\, i \neq j$.

Again, all orthogonal states mixed states are perfectly distinguishable, and therefore for any message chosen from $\{p_i, \rho_i\}$, the compression per qubit should reduce to the classical Shannon entropy, $H(\{p_i\})$. We also know that for $\rho = \sum_i p_i\, \rho_i$, where $\{\rho_i\}$ is an orthogonal set of density matrices, we have the relation:

$$S(\rho) = H(\{p_i\}) + \sum_i p_i S(\rho_i) \;\Rightarrow H(\{p_i\}) = S(\rho) - \sum_i p_i S(\rho_i) = \chi,$$

where χ is the Holevo information. You can see that the above relation removes the conundrum Alice faced earlier with $\rho = \rho_1$, and correctly gives the compression as 0.

On the other hand, if $\{\rho_i\} = \{|\phi_i\rangle\langle\phi_i|\}$ is pure, but not orthonormal, then we should be able to get back the optimal compression as $S(\rho)$. For general cases of mixed-state messages chosen from arbitrary $\{p_i, \rho_i\}$, the optimal compressibility I_p lies between the von Neumann entropy and the Holevo information, with several research trying to make the bounds tighter.

$$S(\rho) \geq I_p \geq S(\rho) - \sum_i p_i S(\rho_i) = \chi.$$

See: M. Koashi and N. Imoto, *Compressibility of Quantum Mixed-State Signals,* Phys. Rev. Lett. **87**, 017902 (2001).

[2] One way to check for orthogonality of mixed states is to see that their purifications form an orthonormal set, i.e., if $\rho_i \rightarrow |\psi_i\rangle_{AB}$, such that $\rho_i = \mathrm{Tr}_B[|\psi_i\rangle\langle\psi_i|_{AB}]$, then the set $\{|\psi_i\rangle_{AB}\}$ is orthonormal.