1 Entropy

The concept of entropy is not new in physics. In mathematics there exist lots of definitions of entropy in many connections. In information theory this definition is connected with our feeling of unexpectedness of the events. In every measurement we have the results occuring with some frequency $f(x)$ that we understand as probability of occurencies of the result x as the number of measurements N tends to infinity. After performing the measurement we obtain the sequence

$$(x_1, x_2, \ldots, x_N)$$

where $x_i \in \mathbf{A}$ and $\mathbf{A}$ is the set of all possible results of the experiments (all alternatives). In such a sequence the result x exists about $Np(x)$-times. The number of all possible sequencies that contain $Np(x)$-times the result $x \in \mathbf{A}$ is given by

$$\Omega(p) = \frac{N!}{\Pi_{x\in\mathbf{A}}(Np(x))!}. \tag{1}$$

Using the Stirling approximation

$$N! = \sqrt{2\pi N} N^N e^{-N} \tag{2}$$

$$\ln N! = N \ln N - N + \ln \sqrt{2\pi N} \approx N \ln N - N \tag{3}$$

we obtain that the number of possible sequencies after the N measurements is

$$\Omega(p) = \frac{\exp\left(-N\sum_{x\in\mathbf{A}} p(x)\ln p(x)\right)}{\sqrt{(2\pi N)^{n-1}\Pi_{x\in\mathbf{A}}p(x)}} = \frac{\exp NH(p)}{\sqrt{(2\pi N)^{n-1}det(p)}} \tag{4}$$

where $H(p) := -\sum_{x\in\mathbf{A}} p(x)\ln p(x)$, $det(p) := \prod_{x\in\mathbf{A}} p(x)$ and n is the number of different values in the sequence of outcomes. One can see the usage of the function $H(p)$ in this estimation of the number of possible sequencies that appear as a result of experiment. We often talk only about the probability distributions of experiments and forget the specific sequence. $\Omega(p)$ is the number of sequencies that leads to the same distribution. This number tells us how much knowledge we simply throw out.

After the measurement we have only the sequencies that fullfiled the given probability distribution and the number of such sequencies is approximatively given by $\Omega(p)$. For large enough N we can in logarithmic scales neglegted the last term in Stirling approximation to get

$$\ln \Omega(p) = NH(p) \tag{5}$$

which one often use in statistical physics. In statistical physics the entropy is defined like the logarithm of the number of microstates $\Omega(E)$ that represents the given macrostate with the energy E,i.e.

$$S = k \ln \Omega(E). \tag{6}$$

Because we do not have any knowledge about the microstate we suppose that each one of them is equally likely. It follows for the probability of microstate i $p(i) = 1/\Omega(E)$ and thus

$$H(p) = -\sum_{i\in\Omega(E)} p(i)\ln p(i) = \ln\Omega(E). \tag{7}$$

In both cases also the function $H(p)$, or the function S, represents our uncertainty about the given system. Once about the sequencies of repeatedly performed measurement and once about the actual microstate of the system. Both we call *entropy*.

Now it is the right time to talk about the information and entropy connection. Firstly we have to define some notions the information theory works with. The set of all alternatives (or outcomes of experiment) we will call *an alphabet*. An element of an alphabet is known as *letter*. The sequence of the letters gives us some *message*. Another important notion is *information source*. It is represented by some measurement, or observation. It describes the way how we gain the information. The information for us is given by a message. We have to put some meaning into these sequencies of letters. It means connect it with some other sequence that we understand. Every source of information we can describe by some probability distribution of the letters $p(x)$. After the N letters we deal with about $\Omega(p)$ possible different messages. Every message has some prior probability $p(x_1,\ldots,x_N)$. In the case when the production of letters is independent it follows that

$$p(x_1,\ldots,x_N) = \prod_{k=1}^{N} p(x_k). \tag{8}$$

For large enough N the probability $p(x)$ of the letter x is roughly equal to frequency $f(x)$ of the occurencies of this letter in a given message, i.e.

$$p(x) \approx f(x) := \frac{N(x)}{N}.$$

Remember that such a sequencies we call *statistically stable*. It implies that for such *typical sequencies*

$$\log p(x_1,\ldots,x_N) = \log\left[\prod_{k=1}^{N} p(x_k)\right] = \log\left[\prod_{x\in\mathbf{A}} p(x)^{f(x)N}\right] = \sum_{x\in\mathbf{A}} \log\left[p(x)^{p(x)N}\right]$$

$$= N\sum_{x\in\mathbf{A}} p(x)\log p(x) = -NH(p).$$

Note that in these relations we have free choice of the logarithm base. The standard choice is with base equal two. Thus for the probability of typical sequencies we obtain

$$p(x_1,\ldots,x_N) \approx 2^{-NH(p)}. \tag{9}$$

We have found the the probability of typical sequencies, but not all sequencies with such probability satisfy the condition of statistical stabilization. For example, the case of binary alphabet 0 and 1 with the probability distribution $p(0) = p(1) = 1/2$. Every message generated by such source has the same probability $1/2^N$.

We little fuzzy our condition of statistical stabilization to introduce the so-called *set of typical sequencies* $T(N, \varepsilon)$. For $\varepsilon > 0$ and $N = 1, 2, \ldots$ the set is defined as

$$T(N,\varepsilon) = \left\{ (x_1, \ldots, x_N) \in \mathbf{A}^N : \left| \frac{1}{N} \log \frac{1}{p(x_1, \ldots, x_N)} - H(p) \right| \leq \varepsilon \right\}. \quad (10)$$

In this set is contained all the messages that have probability of occurence about $K^{-N(H(p)\pm\varepsilon)}$, where K is the base of the logarithm we choose in the definition of the entropy function $H(p)$. The number of the sequencies in such set satisfies some useful inequalities

$$N > N' \Longrightarrow \Omega(T(N,\varepsilon)) > \Omega(T(N',\varepsilon))$$

$$\varepsilon < \varepsilon' \Longrightarrow \Omega(T(N,\varepsilon)) > \Omega(T(N,\varepsilon')).$$

In the following we show that for all ε the probability of whole this set tends to one if N tends to infinity.

Let us introduce the *convergence in measure*. A sequence of functions f_j converges to f in measure if for every $\varepsilon > 0$ there is some J such that for all $j > J$ the measure (or probability) of the set $\{x \in \mathbf{A} : |f_j(x) - f(x)| \geq \varepsilon\}$ is at most ε. The sequence f_j converges *in mean* to f if for every $x \in \mathbf{A}$ and every $\varepsilon > 0$ there is some J such that fo all $j > J$ $|f_j(x) - f(x)| \leq \varepsilon$. If we fix some $\varepsilon > 0$ and define the set $E_j := \{x \in \mathbf{A} : |f_j(x) - f(x)| \geq \varepsilon\}$ then we are able to show, that the convergence in mean implies the convergence in measure! Consider the following relations

$$\int_{\mathbf{A}} |f_j - f| dp \geq \int_{E_n} |f_j - f| dp \geq \varepsilon p(E_j).$$

Now, if the f_j converges in mean then the left side approaches zero for large j. It follows that also $p(E_J)$ approaches zero and thus f_j converges in measure.

Next we simply use this knowledge to prove that the probability of finding the sequence from the complement of typical set vanishes. In this casee we can write

$$\lim_{N\to\infty} Prob(T(N,\varepsilon)) = 1.$$

To see this, it is enough to consider that the sequence of function

$$f_N := \frac{1}{N} p(x_1, \ldots, x_N)$$

tends to entropy of the distribution in mean. This holds for the statistically stable messages as was shown above. One can now also estimate the number of the sequencies contained in the typical set. From the

$$1 \geq Prob(T(N,\varepsilon)) = \sum_{\vec{x} \in T(N,\varepsilon)} p(\vec{x}) \geq 2^{-N(H(p)+\varepsilon} \Omega(T(N,\varepsilon))$$

it follows that

$$\Omega(T(N,\varepsilon)) \leq 2^{N(H(p)+\varepsilon)}.$$

In the previous part of this section we often use the function $H(p)$. We have seen that this function one can take as some measure of uncertainty about the informational source (result of the observation) if we describe the source (or experiment) only via the probability distribution of different outcomes. On the other hand we can connected it with the information gain we are able to obtain from a given source (observation), because it determines the number of all possible messages what means the number of possible alternatives that might have some informational contents for us. The more the messages the more the information. Thus we can define a useful function called *entropy*

$$H(p) := -\sum_x p(x) \log p(x). \tag{11}$$

What is the quantum analog of entropy? As we have seen we can define the entropy to any possible measurement, respectively to any given probability measure. Let us define *von Neumann entropy* $S(\varrho)$ as the minimal value of entropy where the minimalization is taken over the set of all quantum measurement. One can easy show that measurement that minimalize the entropy for a given state ϱ is projective one in the eigenvectors of density matrix that fully describe the state of the system. It follos that (**vid diplomovku**)

$$S(\varrho) = -\mathrm{Tr}(\varrho \log \varrho).$$

1.1 Entropy-like functions

Now we introduce some other entropy functions. In the following by X and Y I mean classical systems described by the joint probability distributions $p(x, y)$. Similarly by A, B I mean the quantum subsystems described by joint state ϱ_{AB} of the whole system $A + B$.

- *Shannon entropy*

$$H(p) = H(X) := -\sum_x p(x) \log p(x) \tag{12}$$

- *von Neumann entropy*

$$S(\varrho) = S(A) := -\mathrm{Tr}(\varrho \log \varrho) \tag{13}$$

- *Relative entropy*

$$H(p||q) := \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{14}$$

- *Quantum relative entropy*

$$S(\varrho||\sigma) := \mathrm{Tr}(\varrho(\log\varrho - \log\sigma) \tag{15}$$

- *Joint probability entropy*

$$H(X,Y) := -\sum_{x,y} p(x,y)\log p(x,y) \tag{16}$$

- *Quantum composite system entropy*

$$S(A,B) := -\mathrm{Tr}(\varrho_{AB}\log\varrho_{AB}) \tag{17}$$

- *Conditional entropy*

$$H(Y|X) := H(X,Y) - H(X) \tag{18}$$

- *Quantum conditional entropy*

$$S(B|A) := S(A,B) - S(A) \tag{19}$$

- *Mutual information*

$$H(X:Y) := H(X) + H(Y) - H(X,Y) \tag{20}$$

- Quantum correlation

$$S(A:B) := S(A) + S(B) - S(A,B) \tag{21}$$

1.2 Properties of entropy

We only name the properties of quantum von Neumann entropy, but each of these properties is of course satisfied by so-called *Shannon entropy* $H(p)$.

1. **Pure state property.** For $\varrho = |\psi\rangle\langle\psi|$ be a pure state the entropy is zero,

$$S(|\psi\rangle\langle\psi|) = 0 \tag{22}$$

2. **Maximum entropy property.** Entropy achieves its maximum only for the maximally mixed state $\varrho = \frac{1}{d}\mathbf{1}$ where d is the dimension of Hilbert space $\mathcal{H}$.

$$S(\frac{1}{d}\mathbf{1}) = \log d \tag{23}$$

3. **Invariance under unitary operations.**

$$S(U\varrho U^\dagger) = S(\varrho) \tag{24}$$

4. **Useful inequality.**

$$\sum_k p_k S(\varrho_k) \leq S(\sum_k p_k \varrho_k) \leq H(p) + \sum_k p_k S(\varrho_k) \tag{25}$$

with the equality on the right achievable only for mutually commuting operators ϱ_k.

5. **Araki-Lieb inequality.**

$$|S(A) - S(B)| \leq S(A,B) \leq S(A) + S(B) \tag{26}$$

with the equality on the right only in the case of product state, i.e. $\varrho_{AB} = \varrho_A \otimes \varrho_B$, because entropy is aditive in the sense that

$$S(\varrho_A \otimes \sigma_B) = S(\varrho_A) + S(\sigma_B). \tag{27}$$

On the left we have the equality if ϱ_{AB} is a pure state?

6. **Strong subaditivity.**

$$S(A,B,C) + S(B) \leq S(A,B) + S(B,C) \tag{28}$$

7. **Klein's inequality**

$$S(\varrho||\sigma) \geq 0 \tag{29}$$

with equality if and only if $\varrho = \sigma$.

8. **Chaining rule.**

$$H(X_1, \ldots, X_l|Y) = \sum_{j=1}^{l} H(X_j|Y, X_1, \ldots, X_{j-1}) \tag{30}$$

1.3 Fano's inequality

Let X and Y be the unknown and known random variables. Put E be the "error" random variable. It means

$$\begin{array}{lll} E = 1 & if & X \neq \overline{X} \\ E = 0 & if & X = \overline{X} \end{array} \tag{31}$$

where $\overline{X} = f(Y)$ is our guess of the random variable X based on the knowledge of Y. Denote by $p_e = p(X \neq \overline{X})$ the probability of error. Then the entropy of the random variable E is given as

$$H(E) = H_{\text{bin}}(p_e) := -p_e \log p_e - (1 - p_e) \log (1 - p_e).$$

Applying the chain rule we can write

$$H(E, X|Y) = H(E|X, Y) + H(X|Y).$$

Using the fact that E is exactly determined if X and Y is known, i.e. $H(E|X,Y) = 0$, we obtain

$$H(E,X|Y) = H(X|Y). \tag{32}$$

The chain rule also implies that

$$H(E,X|Y) = H(X|E,Y) + H(E|Y) \leq H(X|E,Y) + H(E). \tag{33}$$

From the equations (32) and (33) it follows that

$$\begin{aligned} H(X|E,Y) &= p(E=0)H(X|E=0,Y) + p(E=1)H(X|E=1,Y) \\ &\leq p(E=0)0 + p_e \log(|X|-1) \end{aligned} \tag{34}$$

where the last inequality is fullfiled, because in the case when error occured, i.e. $E = 1$, the random variable X can take at most $|X| - 1$ different values, where $|X|$ denote the number of elements in the set X. In the case with no error, i.e. $E = 0$, the conditional entropy vanish. From the equations above it follows the Fano's inequality

$$H(X|Y) \leq H_{\mathrm{bin}}(p_e) + p_e \log(|X| - 1). \tag{35}$$

One can use this inequality to bound the reliability of the channels.

2 Entanglement

What is the entanglement? Does it have any physical meaning? Are we able to measure it?

In this section we are going to try to answer these and some other questions that is concerned with the phenomena of *entanglement*. Firstly we describe some examples which leads some people to contradiction with their intuition.

2.1 The EPR problem

2.2 Quantum teleportation

2.3 Measures of entanglement

3 Quantum channels and capacities

In this section we use the known facts in the case of information transmission, i.e. *communication*. The basics of the mathematical theory of communication was introduced in famous Shannon paper in 1948. Recently his ideas was developed in quantum theory of information, where one use the quantum state instead classical one to represent information. It has been found quantum generalization of classical theory. It has its own problems that just waited for solution.

Without more words come to a definition of the problem. Imagine the situation that we want to send some information via the channel. We gain the information from the so-called information source that is described by prior probability distribution. Denote it by $\pi(x)$ where x are letters from some given alphabet of source A. Next we use some device - *channel*- to tranmit the information from the source to other part (it must not to be a person). Physically the channel is some system that is able to carry the information using some carriers. Firstly we have to determine different carrier to each different letter. This process we call *encoding* of information. We do not go into the details of technical realization. We will consider the encoding with no errors. We just mention that different carriers means physically the different states of the same system, or different physical systems. The only condition is our ability to distinguish between different carriers, that we will sometimes called *signals*, too. At the end of the transmission we have to translate the received signals in the alphabet of receiver, denote this alphabet by B. Such a translation we call *decoding* and in some sense represents inverse operation to encoding. Also in this cae of decoding we do not pay an attention in its realization and we consider it as perfect. Every error of the communication we connect with the transmission over the channel alone.

In this game we have some input message and some output sequence of letters from alphabet B that represents the output message. Our goal is to choose communication diminishing all the mistakes that may occur during the transmisssion. Without loss of generality (because encoding is perfect) we can consider the input of the channel as the information source that generates the signals. The probability of these signals is given by $\pi(x)$. Physically the input signals x are illustrated by the specific physical states ϱ_x. The problem is how to describe the channel. Is it only some specific mapping on the set of states? Not only! Because of the unknown and uncontrolated errors at the end we have to make some *decison rule* that give us some message. Mathematically, after the transmission we receive a letter y when x was sended with some conditional probability $p(y|x)$. And such conditional probability determines the channel. In the quantum case to obtain this characterization is little bit more complicated, because in this case we do not know to find the state of the received signal only by performing single observation. Moreover we can in principle describe the transmission as superoperator \$. Denote by $\{E_y\}$ generalized quantum measurement, i.e. POVM. Such measurement represents our decision rule and the conditonal probability of sended x received y is given by relation

$$p(y|x) = \mathrm{Tr}[\$(\varrho_x)E_y]. \tag{36}$$

3.1 Classical channels.

We will use the procedure known as *block coding* to establish some reliable communication. In block coding we decise about the sended message not from the received letter, but from the received block of N letters. We know from the previous section that information source generates mostly the messages from the so-called typical set. Our task is to choose *code of size M* such that the probability of error vanishes. Code of size M means the choice of M messages composed from N letters from typical set, that the receiver is able to clearly decode received message with probability equal to one. The messages in the such a code we call *codewords.* We want to know the amount of information transmited. Let 2^{NR} be the number of codewords that we are allowed to use. The number R we call *rate of communication* and it characterizes the amount of information transmited. Define the amount of information transmited per signal as

$$I := \frac{1}{N} \log 2^{NR} = R. \tag{37}$$

Firstly, consider noiseless and memoryless classical channel. Noiseless means that the action of the channel is trivial. Memoryless means that sended letters do not depend on the letters already had sended. Classicality means that we do not care about the measurement performed. It follows that it is the same as bring the source of information to receiver. After he received about N letters from source he can obtain about $2^{NH(\pi)}$ different messages. The maximum of the transmited information is achieved in the case if $R = H(\pi)$ because then

$$I := \frac{1}{N} \log 2^{NH(\pi)}. \tag{38}$$

Remember, if $R > H(\pi)$ then the transmission is not perfect, because the probability of error doesn't vanish. We want to send also the messages that are impossible in the probabilistic sense, i.e. their probability is vanishingly small. Note that it is not true, that $\Omega(\pi) = 2^{NH(\pi)}$, but approximatively holds

$$\log \Omega(\pi) \approx NH(\pi). \tag{39}$$

To conclude we can say that in this special case of channel the transmited information is given by

$$I = H(\pi) = H(X). \tag{40}$$

Next we pay attention in memoryless classical channel without absence of noise. In this case the channel is described by conditional probability $p(y|x)$. Remark that noiseless of the channel implies $p(y|x) = \delta_{xy}$. We use the condition of memorylessness to write

$$p(y_1, \ldots, y_N | x_1, \ldots, x_N) = p(y_1|x_1) \ldots p(y_N|x_N). \tag{41}$$

According to the law of large number the following is true

$$\lim_{N\to\infty} -\frac{1}{N} \log p(y_1, \ldots, y_N | x_1, \ldots, x_N) = H(Y|X), \tag{42}$$

because

$$\begin{aligned} -\tfrac{1}{N} \log p(y_1, \ldots, y_N | x_1, \ldots, x_N) &= -\tfrac{1}{N} \log \textstyle\prod_{x,y} p(y|x)^{N(x,y)} \\ &= -\tfrac{1}{N} \textstyle\sum_{x,y} N(x,y) \log p(y|x) \\ &= -\textstyle\sum_{x,y} p(x,y) \log p(y|x) \\ &= -\textstyle\sum_{x,y} p(x,y) \log p(x,y) + \sum_{x,y} p(x,y) \log \pi(x) \\ &= H(X,Y) - H(X) = H(Y|X) \end{aligned} \tag{43}$$

where we used for the number of occurencies of the event (sended x, received y) the relation $N(x,y) = Np(x,y)$.

If one fix the input letter x then he will obtain the probability distribution $p_x(y) := p(y|x)$ of the outcome letters y. Thus if we use only the input letter x we end with the number of possible output messages around $2^{NH(p_x)}$. But we do not use only the letter x. It follows that we have to average the number of output messages over all possible input letters to obtain

$$\prod_x 2^{N(x)H(p_x)} = 2^{\sum_x N(x)H(p_x)} = 2^{N\sum_{x,y}\pi_x p(y|x)\log p(y|x)} = 2^{NH(Y|X)}. \quad (44)$$

We conclude that the number of possible outcomes for each sended message is in average about $2^{NH(Y|X)}$.

How many input messages we can distinguish with probability near to one based on the knowledge of the output messages? We know the probability distribution of the output letters, because the probability of the event (x,y) is given by $p(x,y) = p(y|x)\pi(x)$ and it follows that

$$p(y) := \sum_x p(y|x)\pi(x). \quad (45)$$

Thus we know the number of possible output messages $2^{NH(Y)}$. One can again compare this with the noiseless channel for which $H(X) = H(Y)$. We also know the opposite conditional probability

$$p(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(y|x)\pi(x)}{\sum_z p(y|z)\pi(z)}. \quad (46)$$

Similar as above we can estimate the number of the input messages for each received one. On average for N received letters there are about $2^{NH(X|Y)}$ possible input messages.

Consequently, our choice of code should with certainty determine the transmitted codeword just based on the knowledge of output. We know that the number of possible input messages is $2^{NH(X)}$. But around each codeword one can write the set of $2^{NH(X|Y)}$ messages that should not contain more than one codeword. Thus the maximal number of codeword is given by the fraction

$$\Omega(\mathit{codeword}) = 2^{N[H(X)-H(X|Y)]} = 2^{N[H(X)+H(Y)-H(X,Y)]} = 2^{NH(X:Y)}. \quad (47)$$

Here one can also see the reason why the mutual information $H(X:Y)$ is defined in the way used in the previous section. Follow the definition of the amount of transmitted information used in the noiseless case we obtain

$$I := \frac{1}{N}\log 2^{NH(X:Y)} = H(X:Y). \quad (48)$$

This formula also bounds the maximum rate R of the code we use.

3.1.1 More mathematical formulation

Let us define the communication problem in more precise way.

- By $x \in X$ we understand the input signals of the channel.
- By $y \in Y$ we understand the output signals of the channel.
- The knowledge of the informational source is represented by the input probability distribution $\pi(x)$.
- The code of size M is the choice of the M elements from the set of all typical messages $T(N, \varepsilon) \subset X^N = X \times \ldots \times X$ which allowed us to communicate with the vanishing probability of error $P_{error} < \varepsilon$. The elements used in the code we call codewords.
- The channel alone is described by the conditional probability $p(\text{received } y|\text{sended } x)$ in the memoryless case. In general it is described for $N = 1, 2, \ldots$ by the conditional probability $p(y_1, \ldots, y_N|x_1, \ldots, x_N)$.
- The transmisson rate R is the number of asymptoticaly transmited bits per one transmited signal using some specific code $\mathcal{C}$, i.e.

  $$R := \frac{1}{N} \log \Omega(N, \mathcal{C}) \tag{49}$$

  where $\Omega(N, \mathcal{C})$ is the number of codewords.
- **Our real task:** Find a code such that the transmission rate is as large as possible.
- **Our mathematical task:** Find some criterion that help us characterize the reliability of the channel. It means maximize the transmitted information over the set of all possible input distribution. We will call such a number *capacity of the channel* C which is defined as follows

  $$C := \lim_{M \to \infty} \frac{1}{M} I_M(X, Y) \tag{50}$$

  where $I_M(X, Y)$ is the mutual information for the input alphabet X^M.

We state the following theorem for the memoryless case. Note that one can generelize it, but often in practise we simulate the realistic channels to be memoryless.
Theorem.(*Shannon noisy coding theorem for memoryless channels*)
For a given clasical communication channel described by the conditional probabilities $p(y|x)$ for every $N = 1, 2, \ldots$ one can choose the code of size $M = 2^{NR}$ such that for every $\varepsilon > 0$ in the limit $N \to \infty$ the probability of error $P_{error} < \varepsilon$. Denote by R the rate of the transmission using a specific code. For any $R > C$

such a code does not exist, where C is called capacity of the channel and is defined via the following relation

$$C := \max_{\pi(x)} H(A:B) = \max_{\pi(x)} \sum_{x,y} \pi(x)p(y|x) \log \frac{p(y|x)}{\sum_z \pi(z)p(y|z)}. \tag{51}$$

We do not prove it here, because it have been already done in previous part of this section. We found there the formula that bounds the transmission rate R. Remember that $R < H(X:Y)$. Thus to find the maximum of this rate over the set of possible sources, i.e. probability distributions $\pi(x)$, we have to maximize the mutual information $H(X:Y)$ over this set. Then we obtain for the capacity of the memoryless channel the mentioned formula in the theorem.

3.2 Communication with the help of quantum states

In the introduction we showed the main difference between the communication using classical and quantum case. We have mentioned that in the quantum case the conditional probability that determines the channel is given by

$$p(x|y) = \mathrm{Tr}[\$(\varrho_x)E_y]. \tag{52}$$

One can see the dependencies on the mapping $\$$ described the action of the channel, and also on the performed measurement $\{E_y\}$. This action of the channel is nothing else as superoperator. The encoding in this case means the mapping $x \to \varrho_x \in \mathcal{H}$. The measurement (so-called decision rule) is given by the set positive operators $\{E_y\}$ such that $\sum_y E_y = \mathbf{1}$. By the memoryless channel we mean the mapping defined on the set of states of Hilbert space $\mathcal{H}^{\otimes N}$ such that $\$^N = \$ \otimes \ldots \otimes \$$. If we consider also the possibility of performing the quantum measurements on the whole Hilbert space of incoming messages, i.e. $\mathcal{H}^N$, then there is a difference between the notion of the memoryless channel in classical and quantum case. If we are allowed to perform such *entangled measurements* $\{E_{y_1 \ldots, y_N}\}$, then the condition for the memoryless channel satisfied in classical case is not valid, i.e.

$$p(y_1 \ldots, y_N | x_1, \ldots, x_N) \neq \prod_{k=1}^{N} p(y_k|x_k).$$

Such a condition is satisfied only in the case if we are performing only measurements on each received signal severally. It means that in this case the transmitted information depends also on the measurement. Thus we can define the capacity of the channel using the Shannon theorem, but we have to maximize over measurement, too. It follows

$$C_N(\$) := \lim_{N\to\infty} \frac{1}{N} \sup_{\pi, E_{\vec{y}}} \sum_{\vec{x},\vec{y}} \pi(\vec{x})p(\vec{y}|\vec{x}) \log \left(\frac{p(\vec{y}|\vec{x})}{\sum_{\vec{z}} \pi(\vec{z})p(\vec{y}|\vec{z})} \right) \tag{53}$$

where we use for the input message $x_1, \ldots, x_N$ the notation $\vec{x}$ and similarly for the output message $\vec{y} = y_1, \ldots, y_N$. The input probability distribution of the message $\vec{x}$ is given by

$$\pi(\vec{x}) = \prod_{k=1}^{N} \pi(x_k). \tag{54}$$

It means that we encode each letter independently and the input message is encoded as

$$\varrho_{\vec{x}} = \varrho_{x_1} \otimes \ldots \otimes \varrho_{x_N}. \tag{55}$$

Then it follows for the state of the output message

$$\$(\varrho_{\vec{x}}) = \$(\varrho_{x_1} \otimes \ldots \otimes \varrho_{x_N}). \tag{56}$$

The conditional probability is given by

$$p(\vec{y}|\vec{x}) = \mathrm{Tr}(\$^N(\varrho_{\vec{x}})E_{\vec{y}}). \tag{57}$$

It is nontrivial to compute the capacity of the quantum channel using this formula. Nevertheless, we can define different kinds of capacities, because in principle we have two possibilities of preparation of the messages and two possibilities how to perform the measurement. Let us call them *entangled* and *separated* preparations and measurements. Thus we can define four types of capacities $C_{ss}, C_{se}, C_{es}, C_{ee}$, where the first index represent the encoding preparation and the second one the measurement performed at the end of the transmission. But now we do not go into the details of these problems and we will use as definition of the capacity the C_{se} case. It means we investigate the one by one preparation and the possibility of the entangled measurements on the whole Hilbert space $\mathcal{H}^N$ representing the space of all messages of the length N.

3.2.1 Holevo bound and quantum coding theorem

In this section we introduce the relation that help us to enumerate the channel capacity. Let us start with the definition of sometimes called *Holevo information*

$$\chi(\pi(x), \varrho_x) := \sum_x \pi(x) S(\varrho_x || \overline{\varrho}) \tag{58}$$

where

$$\overline{\varrho} := \sum_x \pi(x) \varrho_x$$

and by $S(\sigma||\varrho)$ we understand the quantum relative entropy (15). It is easy to see that

$$\chi = S(\overline{\varrho}) - \sum_x \pi(x) S(\varrho_x).$$

The following quantum entropy bound holds

$$\sup_{E_{\vec{y}}} I_N(\pi(\vec{x}), E_{\vec{y}}) := \sup_{E_{\vec{y}}} \sum_{\vec{x},\vec{y}} \pi(\vec{x}) p(\vec{y}|\vec{x}) \log \left(\frac{p(\vec{y}|\vec{x})}{\sum_{\vec{z}} \pi(\vec{z}) p(\vec{y}|\vec{z})} \right) \leq \chi(\$^N(\varrho_{\vec{x}}), \pi(\vec{x})) \tag{59}$$

We denote by $I(\pi(\vec{x}), E_{\vec{y}})$ the amount of information transmitted if we fix the information source, i.e. input probability distribution $\pi(x)$, and are able to perform the measurement on the Hilbert space $\mathcal{H}^{\otimes N}$, i.e. entangled one. The prove of this inequality is based on the strong aditivity (28).

A code of size M in the quantum case we call a collection $(\varrho^1, E_1), \ldots, (\varrho^M, E_M)$, where ϱ^l are states and $\{E_l\}$ are positive operators in $\mathcal{H}^{\otimes N}$, such that $\sum_{l=1}^M E_l \leq \mathbf{1}$. If we measure $l \geq 0$, then we accept the decision that the state ϱ_k was transmited. Put

$$E_{error} := \mathbf{1} - \sum_{l=1}^{M} E_l \leq \mathbf{1}$$

interpreted as an evasion of accepting any definite decision. The mean probability of error P_{error} of such code is given by

$$P_{error} = \frac{1}{M} \sum_{l=1}^{M} [1 - \mathrm{Tr}\$(\varrho_l) E_l] \tag{60}$$

Put $M = 2^{NR}$ and denote by $\overline{P}_{error}(N, 2^{NR})$ the infimum of the probability of error over all codes of size M.

Theorem[**Holevo**](*Holevo quantum coding theorem)*

If $C(\$) < \infty$ and $R > C(\$)$, then $\overline{P}_{error}(N, 2^{NR}) \not\to 0$. On the other hand, if the channel satisfies the condition

$$\sup_{\varrho \in \mathcal{H}} S(\$(\varrho)) < \infty,$$

then $\overline{P}_{error}(N, 2^{NR}) \to 0$ for $R < \sup_{\pi} \chi(\$(\varrho_{\vec{x}}), \pi(\vec{x}))$, where $\vec{x} = x_1 \ldots, x_N$ and $\varrho_{\vec{x}}$ are the states from the given code of size M. In particular,

$$C(\$) = \lim_{N \to \infty} \frac{1}{N} \sup_{\pi} \chi(\$(\varrho_{\vec{x}}), \pi(\vec{x})). \tag{61}$$

The proof of this statement is not so easy and use complicated techniques like typical subspaces, a choice of optimal decision rule and a random coding. One can see this proof in [**Holevo**]. First part of the theorem is just the consequence of the Fano's inequality. Remember that for the one by one preparation of the sended letters due to the aditivity of the von Neumann entropy (27) it follows that

$$C = \sup_{\pi} \chi(\$(\varrho_x), \pi(x)) = \sup_{\pi} \left[S(\overline{\$(\varrho)}) - \sum_{x \in A} \pi(x) S(\varrho_{\$(x)}) \right]. \tag{62}$$

According to this theorem it is now much easier to calculate the channel capacity. Some examples?

3.2.2 Some differencies between the quantum and classical capacity

- Classical capacity is given as a supremum over all possible probability distribution of the signal states. Note that these distributions corresponds to classical states of the system. On the other hand the quantum capacity also depends only on these distributions, whch of course do not represents the quantum states of the system. One can define the capacity by maximizing over all posible quantum alphabets for the given system, i.e.

$$C(\$) = \sup_{\{\varrho_x, \pi(x)\}} \chi(\$(\varrho_x), \pi(x)). \tag{63}$$

- There is only one classical notion of capacity, but many possible definitions of the quantum one.
- Classical capacity is aditive in the sense that for any two channels Φ_1, Φ_2 the folowing holds

  $$C(\Phi_1 + \Phi_2) = C(\Phi_1) + C(\Phi_2).$$

  But it is questionable, whether it holds also for quantum case. This question, of course, depends on the definition of the quantum analogue of the capacity.

3.2.3 Remarks on Holevo information χ

Firstly we introduce very useful property of the Holevo information. It is known as *Donald's identity.* For any given ensemble of states $\{\varrho_x, \pi(x)\}$ and any state σ the following is true

$$\sum_x \pi(x) S(\varrho_x||\sigma) = \chi(\varrho_x, \pi(x)) - S(\sigma||\overline{\varrho}). \tag{64}$$

Proof

$$\begin{aligned}
\textstyle\sum_x \pi(x) S(\varrho_x||\sigma) &= \textstyle\sum_x \pi(x)[\mathrm{Tr}\varrho_x \log \varrho_x - \mathrm{Tr}\varrho_x \log \sigma] \\
&= \textstyle\sum_x \pi(x)[\mathrm{Tr}\varrho_x \log \varrho_x + \mathrm{Tr}\varrho_x \log \overline{\varrho} - \mathrm{Tr}\varrho_x \log \overline{\varrho} - \mathrm{Tr}\varrho_x \log \sigma] \\
&= \textstyle\sum_x \pi(x) S(\varrho_x||\overline{\varrho}) + \mathrm{Tr}[(\sum_x \pi(x)\varrho_x) \log \overline{\varrho}] - \mathrm{Tr}[(\sum_x \pi(x)\varrho_x) \log \sigma] \\
&= \textstyle\sum_x \pi(x) S(\varrho_x||\overline{\varrho}) + S(\overline{\varrho}||\sigma) = \chi(\varrho_x, \pi(x)) + S(\overline{\varrho}||\sigma).
\end{aligned}$$

where we use that $\overline{\varrho} = \sum_x \pi(x)\varrho_x$.

Let us define *optimal signal ensemble* as ensemble satisfying the following equality

$$\chi^\star := \max_{\{\sigma_x \in \mathcal{A}, \pi(x)\}} \sum_x \pi_x S(\sigma_x||\overline{\sigma})$$

where $\overline{\sigma} = \sum_x \pi(x)\sigma_x$ and $\mathcal{A}$ is the set of available state, for example at the output of the channel. Imagine an given ensemble $\{\sigma_x, \pi(x)\}$. Change a little this ensemble introducing the state ω in it such that we obtain the mixture

$$\overline{\sigma}' = \eta\omega + (1-\eta)\overline{\sigma}.$$

Then it follows for the Holevo information

$$\chi' = \eta S(\omega||\overline{\sigma}') + (1-\eta) \sum_x \pi(x) S(\sigma_x||\overline{\sigma}').$$

Applying the Donald equality to the ensemble $\overline{\sigma}$ we obtain

$$\sum_x \pi(x) S(\sigma_x||\overline{\sigma}') = \chi + S(\overline{\sigma}||\overline{\sigma}'). \tag{65}$$

We can write

$$\chi' = \eta S(\omega||\overline{\sigma}') + (1-\eta)\sum_x \pi(x) S(\sigma_x||\overline{\sigma}) + (1-\eta) S(\overline{\sigma}||\overline{\sigma}').$$

Then for the difference of the Holevo information holds

$$\Delta\chi := \chi' - \chi = \eta\left[S(\omega||\overline{\sigma}') - \chi\right] + (1-\eta) S(\overline{\sigma}||\overline{\sigma}').$$

Secondly we apply the Donald identity to the ensemble $\overline{\sigma}'$ to obtain

$$\Delta\chi = \eta[S(\omega||\overline{\sigma}) - \chi] - S(\overline{\sigma}'||\overline{\sigma}).$$

Since the relative entropy is never negative we can conclude that

$$\eta[S(\omega||\overline{\sigma}\prime) - \chi] \leq \Delta\chi \leq \eta[S(\omega||\overline{\sigma}) - \chi].$$

Thus we obtain the bounds for the change of Holevo information, if we mix our original ensemble with additional state ω. We say that the ensemble has the *maximal distance property* if for any state $\omega \in \mathcal{A}$ the following inequality holds

$$S(\omega||\overline{\sigma}) \leq \chi(\sigma_x, \pi(x)).$$

Now we can state that an ensemble is optimal if and only if it has the maximal distance property.

To prove this statement we use our bound on the change of Holevo information. We know from the previous that if the ensemble described by the mixture $\overline{\sigma}$ has maximal distance property then

$$\Delta\chi \leq \eta[S(\omega||\overline{\sigma}) - \chi] \leq 0.$$

Thus we do not increase the Holevo information by mixing it with additional state ω and the ensemble have to be the optimal one, and $\chi = \chi^\star$.

But there remains still one possibility to increase the Holevo information by taking off one of the states creating the ensemble with maximal distance property. Suppose $\{\sigma_x, \pi(x)\}$ to be the original ensemble with maximal distance property. If we cancel one of the state $x = 0$ contained in it, we obtain the ensemble denoted by $\overline{\sigma}'$

$$\{\sigma_x, p(x) = \frac{\pi(x)}{1-\pi(0)}\}$$

where $x \neq 0$. Thus for the original ensemble $\overline{\sigma}$ we can write

$$\chi = \sum_x \pi(x) S(\sigma_x||\overline{\sigma}) = \pi(0) S(\sigma_0||\overline{\sigma}) + (1-\pi(0)\sum_{x\neq 0} p(x)) S(\sigma_x||\overline{\sigma}).$$

Applying the Donald identity on the second term we get the equation

$$\begin{aligned}\chi \quad &= \pi(0)S(\sigma_0||\overline{\sigma}) + (1-\pi(0)\left[\sum_{x\neq 0} p(x))S(\sigma_x||\overline{\sigma}') + S(\overline{\sigma}'||\overline{\sigma})\right] \\ &= \pi(0)S(\sigma_0||\overline{\sigma}) + (1-\pi(0))[\chi' + S(\overline{\sigma}'||\overline{\sigma})].\end{aligned}$$

It follows that for the change of the Holevo information by cancelation of one state from a given ensemble we have

$$\Delta\chi = \chi' - \chi = \pi(0)[\chi' - S(\sigma_0||\overline{\sigma})] - (1-\pi(0))S(\overline{\sigma}'||\overline{\sigma}). \tag{66}$$

The maximal distance property implies $S(\sigma_x||\overline{\sigma}) \leq \chi$ for each x. Then whenever $\pi(x) > 0$ we must have

$$S(\sigma_x||\overline{\sigma}) = \chi,$$

otherwise we do not obtain χ after summing up over all x. Using such relation for $x = 0$ in the (66) one comes to

$$\begin{aligned}\Delta\chi &= \pi(0)\Delta\chi - (1-\pi(0))S(\overline{\sigma}'||\overline{\sigma}) \\ (1-\pi(0))\Delta\chi &= -(1-\pi(0))S(\overline{\sigma}'||\overline{\sigma}) \\ \Delta\chi &= -S(\overline{\sigma}'||\overline{\sigma}).\end{aligned}$$

But, we know that the relative entropy is nonnegative function. Thus

$$\Delta\chi = \chi' - \chi \leq 0$$

and we see that the Holevo information for the changed ensemble is lower than for the original one. It confirms the validity of the implications, that maximal distance property means also the optimality.

Conversely, suppose that the ensemble is optimal. If we are able to add some state $\omega \in \mathcal{A}$ such that $S(\omega||\overline{\sigma}) > \chi^\star$, then we can choose η small enough that

$$\Delta\chi \geq \eta[S(\omega||\overline{\sigma}') - \chi^\star] > 0,$$

because in the limit $\eta \to 0$ also the convergence $S(\omega||\overline{\sigma}) \to S(\omega||\overline{\sigma}')$ is fullfiled. And this inequality contradicts the fact that the original ensemble is optimal. Thus, the property of the optimal ensemble and maximal distant property are equivalent.

Combine this statement with the following property

- If supp$\omega \not\subseteq$ suppσ then $S(\omega||\sigma) = \infty$, where supp$\omega$ is the subspace spanned by the eigenvectors of ω belonging to non-zero eigenvalues.

we state that support of the average state of an optimal ensemble must contain the support of every state in $\mathcal{A}$. Otherwise one can find such state in $\mathcal{A}$ that contradicts the maximal distance property.

We have shown the equivallence between the maximal distance property and optimality for a given ensemble.

3.2.4 Additivity for quantum channels

We have already remarked that the additivity is fulfilled for the capacity of the classical channel. Because of the many possibilities how to define the capacity for the quantum channels we have many possibilities for investigating the additivity property. Without any reference to definition of the capacity C the additivity problem means to prove or contradict the following equation

$$C(\Phi_1 \otimes \ldots \Phi_K) = C(\Phi_1) + \ldots + C(\Phi_K) \tag{67}$$

where Φ_j denotes the channel acting on the jth letter. Let us introduce new definition of the capacity of the quantum channel. The little difference between the previous and present definition is in the possibility of maximizing the Holevo quantity over the set of all ensembles in a given Hilbert space $\mathcal{H}$. Thus

$$C(\Phi) = \max_{\{\pi(x),\varrho_x\}} \chi(\pi(x), \Phi(\varrho_x)) = \chi^\star(\mathcal{A}) \tag{68}$$

where $\mathcal{A} = \Phi(\mathcal{S}(\mathcal{H}))$ and $\mathcal{S}(\mathcal{H})$ is the set of all states of the physical system described by the Hilbert space $\mathcal{H}$. The question of additivity implies the question of equality in the following relation

$$\chi^\star(\mathcal{AB}) = \chi^\star(\mathcal{A}) + \chi^\star(\mathcal{B}) \tag{69}$$

where $\mathcal{AB} = \Phi\left(\mathcal{S}(\mathcal{H}_\mathcal{A} \otimes \mathcal{H}_\mathcal{B})\right)$. Do we know an answer? Only in some special cases. Firstly we have to introduce some types of the quantum channels. Namely *classical-quantum channel* and *quantum- classical channel.* The $c-q$ channel is defined as

$$\Phi(\sigma) = \sum_k \varrho_k \mathrm{Tr}\sigma E_k \tag{70}$$

where $E_k = |e_k\rangle\langle e_k|$, $\{|e_k\rangle\}$ is an orthonormal basis and $\{\varrho_k\}$ is a family of quantum states. Thus it is uniquely defined as mapping $k \to \varrho_k$ of the input alphabet X onto the signal quantum states ϱ_k. The second type , $q-c$ channel, is given by the same transformation (70), but $\{E_k\}$ is an arbitrary partition of unity and $\{\varrho_k = |e_k\rangle\langle e_k|\}$. Of course, there are many channels that are not one of these types. The simplest one is *unitary channel* which is given by

$$\Phi(\sigma) = U\sigma U^\dagger \tag{71}$$

where U is arbitrary unitary transfomation.

Let us go back to our problem of the aditivity for $\chi^\star$. It is clear that

$$\chi^\star(\mathcal{AB}) \geq \chi^\star(\mathcal{A}) + \chi^\star(\mathcal{B}),$$

because we can always combine the optimal separated ensembles together into a product ensemble for $\mathcal{AB}$, but the question is whether such product ensemble is optimal, too. It is the same question as to ask if the capacity C_{ee} is bigger than C_{se}. Or in the other words the question of the possibility of increasing of the information transmission by using the entangled states as the input messages.

One can show that the additivity holds for the both $c-q$ and $q-c$ channels. Also for the unitary channels the condition (69) is satisfied.

Suppose now the so-called *half-noisy channel* which is represented by an map of the form $\Phi^{\mathcal{A}} \otimes \mathcal{I}^{\mathcal{B}}$, where $\mathcal{I}^{\mathcal{B}}$ is the identity map on the set of states $\mathcal{S}(\mathcal{H}_{\mathcal{B}})$. Let $\overline{\sigma}^{\mathcal{A}}$ and $\overline{\sigma}^{\mathcal{B}}$ be the average states of the optimal ensembles for the subsystems individually. We will show that the product states $\overline{\sigma}^{\mathcal{A}} \otimes \overline{\sigma}^{\mathcal{B}}$ is optimal by showing the maximal distance property. Consider another state of the composite system $\omega_{\mathcal{AB}}$ from the set of possible output states

$$\mathcal{AB} = (\Phi^{\mathcal{A}} \otimes \mathcal{I}^{\mathcal{B}})(\mathcal{S}(\mathcal{H}_{\mathcal{A}} \otimes \mathcal{H}_{\mathcal{B}})).$$

Our aim is to prove tha following inequality

$$S(\omega_{\mathcal{AB}}||\overline{\sigma}^{\mathcal{A}} \otimes \overline{\sigma}^{\mathcal{B}}) \leq \chi^{\star}(\mathcal{A}) + \chi^{\star}(\mathcal{B}).$$

Since $\mathcal{B}$ involves according to identity map, it is easy to see that the state represented the optimal ensemble is given as

$$\overline{\sigma}^{\mathcal{B}} = \frac{1}{\dim \mathcal{H}_{\mathcal{B}}} \mathbf{1}_{\mathcal{B}}$$

and $\chi^{\star}(\mathcal{B}) = \dim \mathcal{H}_{\mathcal{B}}$. For any state $\omega_{\mathcal{B}}$ then follows

$$S(\omega_{\mathcal{B}}) + S(\omega_{\mathcal{B}}||\overline{\sigma}^{\mathcal{B}}) = \chi^{\star}(\mathcal{B}).$$

Let us turn back to our proof. For any $\omega_{\mathcal{AB}}$ one can write

$$\begin{aligned} S(\omega_{\mathcal{AB}}||\overline{\sigma}^{\mathcal{A}} \otimes \overline{\sigma}^{\mathcal{B}}) &= \mathrm{Tr}\left[\omega_{\mathcal{AB}} \log \omega_{\mathcal{AB}} - \omega_{\mathcal{AB}} \log \overline{\sigma}^{\mathcal{A}} \otimes \overline{\sigma}^{\mathcal{B}}\right] \\ &= -S(\omega_{\mathcal{AB}}) - \mathrm{Tr}\omega_{\mathcal{A}} \log \overline{\sigma}^{\mathcal{A}} - \mathrm{Tr}\omega_{\mathcal{B}} \log \overline{\sigma}^{\mathcal{B}} \\ &= S(\omega_{\mathcal{A}}) + S(\omega_{\mathcal{B}}) - S(\omega_{\mathcal{AB}}) + S(\omega_{\mathcal{A}}||\overline{\sigma}^{\mathcal{A}}) + S(\omega_{\mathcal{B}}||\overline{\sigma}^{\mathcal{B}}) \\ &= S(\omega_{\mathcal{A}}) - S(\omega_{\mathcal{AB}}) + S(\omega_{\mathcal{A}}||\overline{\sigma}^{\mathcal{A}}) + \chi^{\star}(\mathcal{B}) \end{aligned} \tag{72}$$

In general, the state $\omega_{\mathcal{AB}}$ is a mixed state, but we can always purify it by introducing another system $\mathcal{C}$ such that

$$\omega_{\mathcal{AB}} = \mathrm{Tr}_{\mathcal{C}} |\Omega_{\mathcal{ABC}}\rangle\langle\Omega_{\mathcal{ABC}}|.$$

Due to the Schmidt decomposition of the pure states $S(\omega_{\mathcal{AB}}) = S(\omega_{\mathcal{C}})$ Furthermore, imagine that we perform the measurement on the subsystem $\mathcal{B}$ with the outcomes k occuring with the probability p_k. After this measurement the remaining system $\mathcal{AC}$ wil be in a state $|\Omega^k_{\mathcal{AC}}\rangle$. Putting

$$\begin{aligned} \omega^k_{\mathcal{A}} &= \mathrm{Tr}_{\mathcal{C}} |\Omega^k_{\mathcal{AC}}\rangle\langle|\Omega^k_{\mathcal{AC}}| \\ \omega^k_{\mathcal{C}} &= \mathrm{Tr}_{\mathcal{A}} |\Omega^k_{\mathcal{AC}}\rangle\langle|\Omega^k_{\mathcal{AC}}| \end{aligned} \tag{73}$$

we know that $S(\omega^k_{\mathcal{A}}) = S(\omega^k_{\mathcal{C}})$ for all k. Because of the linearity of the Tr operation we have

$$\begin{aligned} \omega_{\mathcal{A}} &= \textstyle\sum_k p_k \omega^k_{\mathcal{A}} \\ \omega_{\mathcal{C}} &= \textstyle\sum_k p_k \omega^k_{\mathcal{C}}. \end{aligned} \tag{74}$$

Thus we have written both $\omega_{\mathcal{A}}$ and $\omega_{\mathcal{C}}$ as ensembles. We apply this in the following calculation

$$\begin{aligned} S(\omega_{\mathcal{A}}) - S(\omega_{\mathcal{AB}}) + S(\omega_{\mathcal{A}}||\overline{\varrho}^{\mathcal{A}}) &= S(\omega_{\mathcal{A}}) - \sum_k p_k S(\omega_{\mathcal{A}}^k) - S(\omega_{\mathcal{C}}) \\ &\quad + \sum_k p_k S(\omega_{\mathcal{C}}^k) + S(\omega_{\mathcal{A}}||\overline{\varrho}^{\mathcal{A}}) \\ &= \chi(\omega_{\mathcal{A}}^k, p_k) - \chi(\omega_{\mathcal{C}}^k, p_k) + S(\omega_{\mathcal{A}}||\overline{\varrho}^{\mathcal{A}}) \\ &= \sum_k p_k S(\omega_{\mathcal{A}}^k||\overline{\varrho}^{\mathcal{A}}) - \chi(\omega_{\mathcal{C}}^k, p_k) \end{aligned} \tag{75}$$

where we use Donald's identity in the last equality. Thus we can write

$$S(\omega_{\mathcal{AB}}||\overline{\sigma}^{\mathcal{A}} \otimes \overline{\sigma}^{\mathcal{B}}) = \chi^\star(\mathcal{B}) + \sum_k p_k S(\omega_{\mathcal{A}}^k||\overline{\varrho}^{\mathcal{A}}) - \chi(\omega_{\mathcal{C}}^k, p_k) \tag{76}$$

Note, that the states $\omega_{\mathcal{A}}^k$ are the available output states of the $\Phi^{\mathcal{A}}$ channel. These states are obtained by making measurement on system $\mathcal{H}_{\mathcal{B}}$ when the joint system is in the state $\omega_{\mathcal{AB}}$. The map $\Phi^{\mathcal{A}} \otimes \mathcal{I}^{\mathcal{B}}$ commutes with the measurement performed on a subsystem $\mathcal{A}$. Thus the measurement result k determine the input state which would evolve into $\omega_{\mathcal{A}}^k$. Taking into account that the $\overline{\sigma}^{\mathcal{A}}$ has maximal distance property, i.e.

$$S(\omega_{\mathcal{A}}^k||\overline{\sigma}^{\mathcal{A}}) \leq \chi^\star(\mathcal{A})$$

and combining the equations (72) and (76), we can write the desired inequality

$$S(\omega_{\mathcal{AB}}||\overline{\sigma}^{\mathcal{A}} \otimes \overline{\sigma}^{\mathcal{B}}) \leq \chi^\star(\mathcal{A}) + \chi^\star(\mathcal{B}). \tag{77}$$

It follows that the product state $\overline{\sigma}^{\mathcal{A}} \otimes \overline{\sigma}^{\mathcal{B}}$ is also optimal for the joint usage of both channels. It means that in the case of half-noisy channels the additivity property holds. It is still an open problem to find the solution of the additivity problem in general. Remmember that we are always talking only about the memoryless channels, because to solve the channels with memory is nontrivial and very difficult in general.

4 Coherent information

In this section we introduce a notion of *coherent information* that in some sense plays a role of information in quantum case and is very useful. We shall try to give the reasons why.

We start with defining the distance measures between quantum states. We do not talk about the distance in mathematical sense, but about a distance that measures our possibility to distinguish between two states. These quantities are derived from the statistical distances between probability distribution. The most used ones are

- **Absolute distance**
$$D(p,q) = \sum_x |p_x - q_x|$$

- **Fidelity**

$$F(p,q)=\sum_x \sqrt{p_x q_x}$$

where p_x, q_x are probability distribution defined on the set X. The quantum analogues of these functions are defined for an arbitrary states ϱ, σ and projection P

- **Absolute distance**

$$D(\varrho,\sigma)=\mathrm{Tr}|\varrho-\sigma| = 2\max_P[\mathrm{Tr}P(\varrho-\sigma)]$$

- **Fidelity**

$$F(\varrho,\sigma)=\mathrm{Tr}\sqrt{\sqrt{\sigma}\varrho\sqrt{\sigma}}.$$

4.1 Properties of absolute distance

1. *Metric properties*
   It is clear that $D(\varrho,\sigma)=0$ if and only if $\varrho = sigma$. The symmetry of absolute distance is also easy to see. The triangle inequality follows from the fact that there exist projector P such that

   $$D(\varrho,\sigma)=2\mathrm{Tr}P(\varrho-\sigma)=2\mathrm{Tr}P(\varrho-\tau)+2\mathrm{Tr}P(\tau-\sigma)\leq D(\varrho,\tau)+D(\tau,\sigma).$$

2. *Double convexity*
   Using similar technique as before one can write

   $$D(\sum_k p_k\varrho_k, \sum_k q_k\sigma_k)\leq \sum_k p_k D(\varrho_k,\sigma_k)+2D(p_k,q_k).$$

3. *Nonincreasing under evolution*
   Let mapping Φ describe the evoluiton. Then

   $$D(\Phi(\varrho),\Phi(\sigma))\leq D(\varrho,\sigma).$$

4. *Fannes inequality*

   $$|S(\varrho)-S(\sigma)|\leq D(\varrho,\sigma)\log d+\eta(D(\varrho,\sigma)),$$

   where $\eta(x)=-x\log x$ and d is the dimension of underlying Hilbert space.

4.2 Properties of fidelity

1. *Fidelity and purification*
   Consider $|\psi\rangle, |\phi\rangle$ be the states of the composite system AB such that $\varrho=\mathrm{Tr}_B|\psi\rangle\langle\psi|$ and $\sigma=\mathrm{Tr}_B|\phi\rangle\langle\phi|$. It can be shown that

   $$F(\varrho,\sigma)=\max_{|\psi\rangle,|\phi\rangle}|\langle\psi|\phi\rangle|.$$

2. *Concavity*
$$F(\sum_k p_k \varrho_k, \sum_k q_k \sigma_k) \geq \sum_k \sqrt{p_k q_k} F(\varrho_k, \sigma_k)$$
Let $|\psi_k\rangle, |\phi_k\rangle$ be the purifications of ϱ_k and σ_k such that $F(\varrho_k, \sigma_k) = \langle\psi_k|\phi_k\rangle$. It is easy to check that the states
$$|\psi\rangle = \sum_k \sqrt{p_k}|\psi_k\rangle \otimes |k\rangle$$
$$|\phi\rangle = \sum_k \sqrt{q_k}|\phi_k\rangle \otimes |k\rangle$$
are purifications of the states $\varrho = \sum_k p_k \varrho_k$ and $\sigma = \sum_k q_k \sigma_k$. In the following calculation we derive the concavity of the fidelity
$$\begin{aligned} F(\textstyle\sum_k p_k \varrho_k, \sum_k q_k \sigma_k) &\geq |\langle\psi|\phi\rangle| \\ &= \textstyle\sum_k \sqrt{p_k q_k}\langle\psi_k|\phi_k\rangle \\ &= \textstyle\sum_k \sqrt{p_k q_k} F(\varrho_k, \sigma_k). \end{aligned} \tag{78}$$

3. *Non-decreasing under evolution*
$$F(\Phi(\varrho), \Phi(\sigma)) \geq F(\varrho, \sigma)$$

These measures we call *static* ones. If we consider also the evolution we can define the so-called *dynamic* measures that describes the change of states during the evolution.

4.3 Dynamic measures

The purification of the state ϱ of the system Q we make by introducing another system R where the state of the joint system RQ is already pure. After such process we create some correlation between our system Q and the formal system R. Because the whole system is in pure state we have only quantum correlations - *entanglement*. Thus our systems RQ are described by pure entangled state $|RQ\rangle$. The system Q is subjected to dynamics described by map Φ. How well the initial entanglement between the systems Q and R is preserved by the evolution of the system Q according to map Φ? Note that the mapping Φ describes in the language of subsystem Q the possible interaction of this subsystem with the enviroment. These mutual interaction leads to creation of the entanglement between the enviroment and subsystem Q to the prejudice of the initial entanglement between the subsystems R and Q. Let the state after the evolution will be given as
$$\varrho_{R'Q'} = (\mathcal{I}_R \otimes \Phi_B)|RQ\rangle\langle RQ|.$$
We can use the mentioned static measures to define dynamic measures in the following way

- **Dynamic distance**

$$D(\varrho, \Phi) := D(|RQ\rangle, \varrho_{R'Q'})$$

- **Dynamic fidelity**

$$F(\varrho, \Phi) := [F(|RQ\rangle, \varrho_{R'Q'})]^2$$

- **Dynamic error**

$$E(\varrho, \Phi) := 1 - F(\varrho, \Phi)$$

Note that all these quantities do not depend on the purification procedure. In what manner these quantities describe the conservation of entanglement? Stricktly speaking they describe only the change of the change of the state during the evolution. To illustrate the entanglement changes we need to define more appropriate quantity.

4.4 Entropy exchanges

We know that the entanglement is well preserved in the case when the entanglement between the enviroment E and subsystem Q will not arise. We consider three systems R, Q and E which is initialy in the pure state $|RQE\rangle$. The entanglement shared between the system and the enviroment is zero. During the map Φ represented on the whole system RQE by unitary map of the type $U = U_R \otimes U_{QE}$ the entanglement shared between the enviroment and the system RQ can increase. It implies that the entanglement between R and Q could decrease. Because the final state is again pure one, $|R'Q'E'\rangle$, we can express the entanglement between the enviroment and the rest of the system as

$$E_{RQ-E} = S(\varrho_{E'}).$$

This quantity better illustrate the change of the entanglement between R and Q and is called *entropy exchange.*

Thus the entropy exchange is given by

$$S(\varrho, \Phi) := S(\varrho_{R'Q'}) = S(\varrho_{E'}).$$

Note that it is not depending on the purification of the state ϱ of the subsystem Q. Let us assume that

$$\Phi(\varrho) = \sum_k M_k \varrho M_k^\dagger.$$

Then the final state of the enviroment is given by

$$\begin{aligned}
\varrho_{E'} &= \mathrm{Tr}_{RQ}[U(|RQ\rangle\langle RQ| \otimes |0_E\rangle\langle 0_E|)U^\dagger] \\
&= \mathrm{Tr}_{\mathrm{Q}}[U_{QE}(\varrho \otimes |0_E\rangle\langle 0_E)|U_{QE}^\dagger] \\
&= \mathrm{Tr}_{\mathrm{Q}}[(\mathbf{1}_Q \otimes \textstyle\sum_k |k_E\rangle\langle k_E|)U_{QE}(\varrho \otimes |0_E\rangle\langle 0_E|)U_{QE}^\dagger(\mathbf{1}_Q \otimes \textstyle\sum_l |l_E\rangle\langle l_E|)] \\
&= \mathrm{Tr}_{\mathrm{Q}}[\textstyle\sum_{kl} |k_E\rangle\langle k_E|U_{QE}|0_E\rangle\varrho\langle 0_E|U_{EQ}|l_E\rangle\langle l_E|] \\
&= \textstyle\sum_{kl}(\mathrm{Tr}_Q[M_k \varrho M_l^\dagger])|k_E\rangle\langle l_E| \\
&= \textstyle\sum_{kl} W_{kl}|k_E\rangle\langle l_E|.
\end{aligned} \tag{79}$$

Then for the entropy exchange we can write

$$S(\varrho, \Phi) = -\mathrm{Tr} W \log W$$

where

$$W_{kl} := \mathrm{Tr} M_k \varrho M_l^\dagger.$$

The map Φ has many operator-sum representation and it is possible to choose such one, that the matrix W is diagonal. In particular, the sets of operators M_k and N_l generate the same evolution if and only if

$$N_l = \sum_k u_{lk} M_k$$

where u_{lk} is unitary matrix. According to this transformation of operators the matrix W will change in the way

$$W' = u W u^\dagger.$$

This is also the way how we can diagonalize the matrix W. The set of operators that leads to diagonal form of the matrix W we call *canonical representation of* Φ.

The entropy exchange quantifies the entanglement change between the subsystems R and Q during the process Φ only intuitively, because we do not have good measure for multiparticle entanglement yet. It describes also the noiseness of quantum operation. In some sense it is a generalization of the conditional entropy in quantum case.

Consider the the basis $\{|\psi_k\rangle\}_{k=1}^{d^2}$ of the system RQ described by d dimensional Hilbert space and let $|\psi_1\rangle = |RQ\rangle$. Define the numbers

$$p_k := \langle \psi_k | \varrho_{R'Q'} | \psi_k \rangle.$$

It follows from the definition of the von Neumman entropy that

$$S(\varrho_{R'Q'}) \leq H(p_1, \ldots, p_{d^2}).$$

The Shannon entropy functions satisfies the property

$$H(p_1, \ldots, p_{d^2}) = H_{\mathrm{bin}}(p_1) + (1 - p_1) H(\frac{p_2}{1 - p_1}, \ldots, \frac{p_{d^2}}{1 - p_1}).$$

Combining the following facts

1. $H(\frac{p_2}{1-p_1}, \ldots, \frac{p_{d^2}}{1-p_1}) \leq \log(d^2 - 1)$
2. $p_1 = F(\varrho, \Phi)$

we obtain the *quantum Fano inequality*

$$S(\varrho, \Phi) \leq H_{\mathrm{bin}}(F(\varrho, \Phi)) + (1 - F(\varrho, \Phi)) \log(d^2 - 1).$$

Comparing this result with the classical Fano inequality (35) we can see the quantum analoques of classical notions of probability of error and conditional entropy.

4.5 Coherent information

Following the analogues between the classical and quantum Fano inequality we can define the coherent information in the similar way as mutual information. The transmited information is given in classical case as a difference between the entropy of the channel's output and the conditional entropy which is given by a channel and describes the noise in it. Similarly, in quantum case we are look for the reliability of the transmission of the quantum states, or more physicaly for the change of the states during some evolution. For a given state ϱ and given evolution Φ we can define the coherent nformation as

$$I(\varrho, \Phi) := S(\Phi(\varrho)) - S(\varrho, \Phi).$$

One can easy see the analogies between the definition of the mutual information, which can we understand as the quantity how well the entropy of information source is preserved during the transmission.

Let us look at the mutual information from the point of view of entropy exchange. Put the source X describe by density matrix composed from mutualy orthogonal pure states

$$\varrho_X = \sum_x p_x |x\rangle\langle x|.$$

The decision rule will be described by the operators

$$E_{xy} := |y\rangle\langle x|.$$

The channel we can describe in a way

$$\Phi(\varrho_X) = \varrho_Y = \sum_{xy} p_{y|x} E_{xy} \varrho_X E_{xy} = p_y |y\rangle\langle y|.$$

The entropy exchange for such evolution we find with the help of the matrix W that is equal to

$$W_{xy,x'y'} = \delta_{xx'} \delta_{yy'} p_x p_{y|x} = \delta_{xx'} \delta_{yy'} p_{xy}.$$

It means that the matrix W is diagonal with the joint probabilities p_{xy}. For the mutual information we can write

$$H(X:Y) = H(X) + H(Y) - H(X,Y) = S(\varrho_X) + S(\varrho_Y) - S(\varrho_X, \Phi) = S(\varrho_X) - I(\varrho_X, \Phi).$$

The coherent infomation describe the change of the quantum state during the evolution. We can use it also for qualitatively measure the reliability of *quantum error correcting codes*. For more see **nielsen**.