

Quantum source coding and data compression¹

Dénes Petz²

Department for Mathematical Analysis
Budapest University of Technology and Economics
H-1521 Budapest XI., Hungary

This lecture is intended to be an easily accessible first introduction to quantum information theory. The field is large and it is not completely covered even by the recent monograph [15]. Therefore the simple topic of data compression is selected to present some ideas of the theory. Classical information theory is not a prerequisite, we start with the basics of Shannon theory to give a feeling for Shannon entropy and for the informational divergence or relative entropy. The aim is to present Schumacher's compression theorem and to demonstrate that the von Neumann entropy, introduced in the 1920's by thermodynamical considerations, is a measure of quantum information exactly in the way as the Shannon entropy is that for classical information. Our discussion makes clear that the compression theorem depends heavily on the existence of the high-probability subspace.

At the end of the lecture quantum sources with memory and some related questions are briefly discussed. This part could be skipped by new-comers in the field.

1 Classical source coding

Let X be a random variable with a finite range \mathcal{X} . A **source code** C for X is a mapping from \mathcal{X} to the set of finite length strings of symbols of a D -ary alphabet which is assumed to be the set $\{0, 1, 2, \dots, D-1\}$. Let $C(x)$ denote the codeword corresponding to x and let $\ell(x)$ denote the length of $C(x)$. If $p(x)$ is the probability of $x \in \mathcal{X}$, then the **expected length** of a source code C is given by

$$L(C) := \sum_x p(x) \ell(x).$$

Since the transmission of lengthy codewords could be costly, the aim of source coding is to make the expected code-length as small as possible. It is obvious that to meet

¹This text is a good written approximation of the first talk of a series given at the *Volterra-CIRM International School on Quantum information and quantum computing* in Trento, July, 20001.

²Partially supported by OTKA T032662 and T032374.

this requirement the most frequent outcome of X must have the shortest codeword. For example in the Morse code the letter e (which is the most frequent one both in the English and Hungarian language) is represented by a single dot. (The **Morse code** uses an alphabet of four symbols: a dot, a dash, a letter space and a word space.) The extension of a code C from the finite length strings of \mathcal{X} is defined by

$$C^*(x_1x_2 \dots x_n) = C(x_1)C(x_2) \dots C(x_n),$$

where the right hand side is the concatenation of the corresponding codewords.

A code C is **uniquely decodable** if $C^*(x_1x_2 \dots x_n) = C^*(x'_1x'_2 \dots x'_m)$ implies that $x_1x_2 \dots x_n = x'_1x'_2 \dots x'_m$, that is $n = m$ and $x_i = x'_i$ for all $1 \leq i \leq n$. A code is called **prefix code** if no codeword is a prefix of any other. In case of a prefix code the end of a codeword is immediately recognised and hence such a code is uniquely decodable. For example, if 0, 10, 110 and 111 are the binary codewords (of a prefix code), then the binary string 1011001101110 is easily decomposed into 6 codewords: 10,110,0,110,111,0.

Theorem 1 (Kraft–MacMillan). *The codeword lengths $\ell(x)$ of a uniquely decodable code over an alphabet of size D satisfy the inequality*

$$\sum_x D^{-\ell(x)} \leq 1.$$

Conversely, given a set of codeword lengths that satisfy this inequality, there exists a prefix code with these codewords lengths.

The proof is available in several standard books, for example [4]. It follows from the theorem that a uniquely decodable code could be always replaced by a prefix code which has the same codeword lengths.

Let $\lceil t \rceil$ denote the smallest integer $\geq t \in \mathbb{R}$. The codeword lengths $\ell(x) := \lceil -\log_D p(x) \rceil$ satisfy the Kraft inequality

$$\sum_x D^{-\ell(x)} \leq \sum_x p(x) = 1.$$

According to the theorem there exists a prefix code with this codeword length. (Such a code is called **Shannon code**.) Since $-\log_D p(x) \leq \ell(x) \leq -\log_D p(x) + 1$, we have

$$-\sum_x p(x) \log_D p(x) \leq L(C) \leq 1 - \sum_x p(x) \log_D p(x).$$

for the expected code-length $L(C)$. For the rest we assume that $D = 2$. Then the bounds are given in terms of the **Shannon entropy** $H(p(x)) := -\sum_x p(x) \log p(x)$ as

$$H(p(x)) \leq L(C) \leq H(p(x)) + 1.$$

According to the next theorem the Shannon code is close to optimal.

Theorem 2. *The expected code-length of any prefix code is greater than or equal to the Shannon entropy of the source.*

Proof. We want to show $L - H(p(x)) \geq 0$ and estimate as follows

$$\begin{aligned} L - H(p) &= \sum_x p(x)\ell(x) + \sum_x p(x) \log p(x) \\ &= -\sum_x p(x) \log 2^{-\ell(x)} + \sum_x p(x) \log p(x) \\ &= \sum_x p(x) \log \frac{p(x)}{r(x)} - \log c, \end{aligned}$$

where $r(x) = c^{-1}2^{-\ell(x)}$ and $c = \sum_x 2^{-\ell(x)}$. The **relative entropy** of two probability distributions is defined as

$$D(p\|r) := \sum_x p(x)(\log p(x) - \log r(x))$$

and this quantity is known to be positive and 0 if and only if $p = q$. In terms of the relative entropy we have

$$L - H(p) = D(p\|r) + \log \frac{1}{c}.$$

Since $D(p\|r) \geq 0$ and $c \leq 1$ from the Kraft-McMillan inequality, this shows $L - H(p) \geq 0$. \square

The Shannon code is close to optimal only if we know correctly the distribution of the source X . Assume that it is not the case and we associate to x the codeword length $\lceil -\log q(x) \rceil$, where q is another probability distribution on \mathcal{X} , possibly different from the true distribution p . One can compute that in this case

$$H(p) + D(p\|q) \leq L(C) \leq H(p) + D(p\|q) + 1. \quad (1)$$

For the use of the wrong distribution the relative entropy is the penalty in the expected length.

The optimal coding is provided by a procedure due to Huffman. The **Huffman code** is not easy to describe, therefore we show another coding due to Fano. The **Fano code** is nearly optimal, it satisfies the inequality

$$L(C) \leq H(p) + 2.$$

In the Fano coding we order the probabilities $p(x)$ decreasingly as $p_1 \geq p_2 \geq p_3 \geq \dots \geq p_m$. We choose k such that

$$\left| \sum_{i=1}^k p_i - \sum_{i=k+1}^m p_i \right|$$

is minimal. The division of the probabilities into the two classes divides the source symbols into two classes. A sign 0 for the first bit for the lower class and 1 for the first bit of the upper class. The two classes have nearly equal probabilities. Then we repeat the procedure for each of the two classes to determine the further bits of the code strings. This is Fano's scheme.

Up to now we have dealt with uniquely decodable codes. If the transmission of lengthy codewords is expensive, we might give up the exact decodability provided that the probability of mistake is small and long codewords can be avoided. This is a different approach to coding and decoding. Assume that the source emits the symbols $X_1, X_2, X_3, \dots, X_n$ (independently and according to the same distribution p , typical for the source). We fix a coding procedure and all the emitted symbols are coded by this procedure which could be the Fano code, for example. Let L_1, L_2, \dots, L_n be the code-length of X_1, X_2, \dots, X_n , respectively. Both X_1, X_2, \dots, X_n and L_1, L_2, \dots, L_n are identically distributed independent random variables, the expectation of L_i is $L(C)$. The law of large numbers tells us that the probability of the event

$$L_1 + L_2 + \dots + L_n \geq L(C) + \varepsilon \quad (2)$$

goes to 0 as $n \rightarrow \infty$. When x_1, x_2, \dots, x_n is a string of source symbols such that the corresponding code string is shorter than $n(L(C) + \varepsilon)$, then we code the string $x_1 x_2 \dots x_n$ perfectly, otherwise we use always the same code string. If the latter case happens to occur, then we cannot recover the emitted symbol string from the code string. However, the probability of this error is exactly the probability of the event (2) which tends to 0. What did we win in this way? The number of source strings is $|\mathcal{X}|^n$ and the number of binary strings used in the coding is $2^{n(L(C) + \varepsilon)}$. When $L(C) < \log |\mathcal{X}|$, then

$$2^{n(L(C) + \varepsilon)} \ll |\mathcal{X}|^n.$$

Hence the cardinality of our code book is much smaller than the cardinality of the source strings if a small probability of error is allowed. We also say that that the data set \mathcal{X}^n is compressed to a set of binary strings of length $n(L(C) + \varepsilon)$. What we have is an example of data compression. Efficient data compression is the same as source coding by short binary code strings. Since we need $n(L(C) + \varepsilon)$ binary digit for a source string of length n , $L(C) + \varepsilon$ is called **code rate**. (It is the number of binary digits needed for a single source symbol, in the average.) Using the Shannon code, we can achieve a code rate $H(p) + \varepsilon$. However, if we mistake the distribution of the source and assume q instead of p , then the rate is higher, it is about $H(p) + D(p||q) + \varepsilon$. Hence the above method is very sensitive for the distribution of the source. To avoid this and to achieve slightly better code rate **block coding** can be used. Shortly speaking block coding means that the source string is not coded by letter by letter but the whole string gets a code string.

A **block code** $(2^{nR}, n)$ for a source X_1, X_2, \dots is given by two (sequences of) mappings:

$$f_n : \mathcal{X}^n \rightarrow \{1, 2, \dots, 2^{nR_n}\}, \quad \phi_n : \{1, 2, \dots, 2^{nR_n}\} \rightarrow \mathcal{X}^n.$$

Here f_n is the **encoder**, ϕ_n is the **decoder** and $R := \lim R_n$ is called the **rate of the code**. The **probability of error** of the code is

$$P_e^{(n)} := \text{Prob} (\phi_n \cdot f_n(X_1, \dots, X_n) \neq (X_1, \dots, X_n)).$$

Shannon's source coding theorem is the following.

Theorem 3. *Let H be the entropy of the source and $R > H$. There exists a sequence of $(2^{nR_n}, n)$ block codes with error probability $P_e^{(n)}$ such that $P_e^{(n)} \rightarrow 0$ and $R_n \rightarrow R$.*

More precisely, this is only the positive part of Shannon's theorem telling that any rate $\geq H + \varepsilon$ is achievable under an arbitrary small bound on the probability of error. (The negative part tells that rates $< H$ are not achievable under the same constraint.)

Before we enter the proof we give an outline of the method of types. Let $\mathbf{x} \in \mathcal{X}^n$. The **type** of $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathcal{X}^n$ is a probability mass function on \mathcal{X} . The mass of $x \in \mathcal{X}$ is the relative frequency of x in the sequence (x_1, x_2, \dots, x_n) :

$$P_{\mathbf{x}}(x) := \frac{1}{n} \#\{1 \leq i \leq n : x_i = x\}.$$

Let \mathcal{P}_n denote the set of all types and for $P \in \mathcal{P}_n$ the **type class** of P is the set of all sequences of type P :

$$T(P) := \{\mathbf{x} \in \mathcal{X}^n : P_{\mathbf{x}} = P\}.$$

Since the frequency of any $x \in \mathcal{X}$ in a sequence $\mathbf{x} = (x_1, x_2, \dots, x_n)$ is at most n , we obviously have

$$\#(\mathcal{P}_n) \leq (n+1)^{\#\mathcal{X}}.$$

The cardinality of a type class $T(P)$ is a multinomial coefficient but the following exponential bounds are useful:

$$\frac{1}{(n+1)^{\#\mathcal{X}}} 2^{nH(P)} \leq \#(T(P)) \leq 2^{nH(P)}.$$

(A proof could be based on Stirling's formula on factorial functions, see [4] p. 282 for other proofs.)

Assume that a probability measure Q is given on \mathcal{X} and let Q^n be the product measure on \mathcal{X}^n . The probability of a sequence $\mathbf{x} \in \mathcal{X}^n$ depends only on the type $P_{\mathbf{x}}$ of \mathbf{x} . A straight calculation gives that

$$Q^n(\{\mathbf{x}\}) = \prod_x Q(x)^{nP_{\mathbf{x}}(x)} = 2^{-nH(P_{\mathbf{x}}) + nD(P_{\mathbf{x}}\|Q)}.$$

The probability of a type class has exponential bounds:

$$\frac{1}{(n+1)^{\#\mathcal{X}}} 2^{-nD(P\|Q)} \leq Q^n(T(P)) \leq 2^{-nD(P\|Q)}$$

for $P \in \mathcal{P}_n$.

Proof of Theorem 3: Let $\{Q(x) : x \in \mathcal{X}\}$ be the probability distribution of the given source and assume that $R > H(Q)$. Following the idea of Csiszár and Körner [5], we set

$$R_n := R - \frac{\log(n+1)}{n}$$

and

$$A_n := \{\mathbf{x} \in \mathcal{X}^n : H(P_{\mathbf{x}}) \leq R_n\}.$$

Then

$$\begin{aligned} \#(A_n) &= \sum \#(T(P)) \leq \sum 2^{nH(P)} \leq \sum 2^{nR_n} \\ &\leq (n+1)^{\#\mathcal{X}} 2^{nR_n} = 2^{nR}, \end{aligned}$$

where all summations are over the set $\{P \in \mathcal{P}_n : H(P) \leq R_n\}$. We can easily define an encoding and a decoding such that elements of A are encoded correctly and the other sequences give an error. (We just use elements of A as codewords). Then the probability of error is

$$P_e^{(n)} = 1 - \text{Prob}(A_n) = \sum Q^n(T(P)),$$

where the summation is over all $P \in \mathcal{P}_n$ such that $H(P) > R_n$. Estimating the sum by the largest term we obtain

$$P_e^{(n)} \leq (n+1)^{\#\mathcal{X}} 2^{-n \min D(P\|Q)},$$

where \min is over all $P \in \mathcal{P}_n$ such that $H(P) > R_n$. When n is large enough then $R_n > H(Q)$ and $Q \notin \{P \in \mathcal{P}_n : H(P) \leq R_n\}$. The minimum in the exponent is strictly positive and we can conclude that the probability of error converges to 0 exponentially fast as $n \rightarrow \infty$.

The interesting feature of the block code constructed in the proof of the theorem is the fact that the distribution Q of the source does not appear, only its entropy $H(Q)$ should be known to construct the **universal encoding scheme**.

2 Quantum mechanical sources

A pure state of a quantum mechanical system is given by a unit vector of a Hilbert space. Assume that a quantum mechanical source emits the pure states $|\varphi_i\rangle$ with probability p_i ($1 \leq i \leq m$). The source is specified by $(p_i, |\psi_i\rangle)_{i=1}^m$ which is called an **ensemble of (pure) quantum states**. Pure states of a quantum system are infinitely many and possess a fine topological structure. If after encoding and decoding we arrive at a state $|\psi'_i\rangle$ instead of $|\psi_i\rangle$, our error could be small when the vectors $|\psi_i\rangle$ and $|\psi'_i\rangle$ are close enough. Hence the problem of source coding in the quantum setting is rather different from the theory of source coding for finite classical sources and it is conceptually closer to the rate distortion theory initiated by Shannon as well.

How close are two quantum states? There are many possible answers to this question. Restricting ourselves to pure states, we have to consider two unit vectors, $|\varphi\rangle$ and $|\psi\rangle$. Quantum mechanics has used the concept of transition probability $|\langle\varphi|\psi\rangle|^2$ for a long time (see cite, for example). This quantity is phase invariant, it lies between 0 and 1. It equals to 1 if and only if the two states coincide that is, $|\varphi\rangle$ equals to $|\psi\rangle$ up to a phase.

We call the square root of the transition probability **fidelity**:

$$F(|\varphi\rangle, |\psi\rangle) := |\langle\varphi|\psi\rangle|.$$

Shannon used a nonnegative distortion measure, and we may regard $1 - F(|\varphi\rangle, |\psi\rangle)$ as a distortion function on quantum states.

Under quantum operation a pure state could be transformed into a mixed state, hence we need extension of the fidelity:

$$F(|\varphi\rangle\langle\varphi|, D) = \sqrt{\langle\varphi|D|\varphi\rangle}.$$

(Some properties of fidelity are summarised in the Appendix.)

The n -shot source is $(p_I, |\psi_I\rangle)$ on the n -fold tensor product $\mathcal{H}^n = \mathcal{H} \otimes \mathcal{H} \otimes \cdots \otimes \mathcal{H}$, where

$$p_I = p_{i_1} p_{i_2} \cdots p_{i_n}, \quad |\psi_I\rangle = |\psi_{i_1}\rangle \otimes |\psi_{i_2}\rangle \otimes \cdots \otimes |\psi_{i_n}\rangle$$

when $I = (i_1, i_2, \dots, i_n) \in \{1, 2, \dots, m\}^n$. The product structure expresses that the generation of the quantum state is a process without memory. The states of different shots are statistically independent.

Now we are ready to define what we mean by a **reliable compression** of a source $(p_i, |\psi_i\rangle)$ on a Hilbert space \mathcal{H} . The compression scheme consist of two quantum operations $\mathcal{C}^n : \mathcal{B}(\mathcal{H}^n) \rightarrow \mathcal{B}(\mathcal{K}_n)$ and $\mathcal{D}^n : \mathcal{B}(\mathcal{K}_n) \rightarrow \mathcal{B}(\mathcal{H}^n)$. \mathcal{K}_n is a Hilbert space of dimension 2^{nR} . We assume that $\mathcal{K}_n \subset \mathcal{H}^n$ and $\mathcal{D}^n(D) = D \oplus 0$. This compression scheme is **reliable** and has **rate** R when

(i) $R_n \rightarrow R$

(ii) $\sum_I p_I F(|\psi_I\rangle\langle\psi_I|, \mathcal{D}^n \cdot \mathcal{C}^n(|\psi_I\rangle\langle\psi_I|)) \rightarrow 1$ as $n \rightarrow \infty$.

The first condition tells that asymptotically 2^R dimension is used for the compression of a single emission of the source on the average. (This dimension is equivalent to the use of R qubits) On the other hand, the second condition tells that the emitted state and the compressed one are close in the average, the expectation value of the fidelity is converging to 1. (Note that this definition of the reliable compression scheme is not the most general, since the form of \mathcal{D}^n is rather restrictive.)

The key role of the quantum extension of Shannon's first theorem is played by the **von Neumann entropy**. If D is a density matrix, then its eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_k$ are nonnegative and von Neumann set

$$S(D) := - \sum_i \lambda_i \log \lambda_i. \tag{3}$$

Concerning the von Neumann entropy, see the Appendix.

The positive part of **Schumacher's source coding theorem** is the following.

Theorem 4. *Let $(p_i, |\psi_i\rangle)$ be a source of pure states on a Hilbert space \mathcal{H} and let S be the von Neumann entropy of the density matrix $\sum_i p_i |\psi_i\rangle\langle\psi_i|$. If $R > S$, then there exists a reliable compression scheme of rate R .*

Proof. Let $\xi_1, \xi_2, \dots, \xi_k$ be the eigenvectors of the density matrix $\sum p_i |\psi_i\rangle\langle\psi_i|$ and $\lambda_1, \lambda_2, \dots, \lambda_k$ be the corresponding eigenvalues. $(\lambda_1, \lambda_2, \dots, \lambda_k)$ is a probability distribution on the set $\mathcal{X} := \{1, 2, \dots, k\}$ and $H(\lambda_1, \lambda_2, \dots, \lambda_k) = S$. We shall use the universal coding method of Csiszár and Körner (see the proof of Theorem 3).

Let

$$\mathcal{P}_n^\circ = \{P \in \mathcal{P}_n : H(P) \leq R_n\},$$

where $R_n = R - \frac{k}{n} \log(n+1)$. We showed above that for

$$A_n := \{I \in \mathcal{X}^n : P_I \in \mathcal{P}_n^\circ\}$$

we have

$$\#(A_n) \leq 2^{nR} \quad \text{and} \quad \sum_{I \in A_n} \lambda_I \rightarrow 1 \quad \text{as } n \rightarrow \infty$$

where $\lambda_I = \lambda_{i(1)} \lambda_{i(2)} \dots \lambda_{i(n)}$ for $I = (i(1), i(2), \dots, i(n))$. The Hilbert space \mathcal{K}_n for the compressing scheme will be a subspace of \mathcal{H}^n ,

$$\{\xi_I : I \in A_n\}$$

is a basis for \mathcal{K}_n . We have $\dim \mathcal{K}_n \leq 2^{nR}$. Next we give the quantum operations $\mathcal{C}^n : \mathcal{B}(\mathcal{H}^n) \rightarrow \mathcal{B}(\mathcal{K}_n)$ and $\mathcal{D}^n : \mathcal{B}(\mathcal{K}_n) \rightarrow \mathcal{B}(\mathcal{H}^n)$. Set

$$\mathcal{C}^n(\sigma) = P_n \sigma P_n + \sum_{I \notin A_n} A_I \sigma A_I^*$$

where P_n is the orthogonal projection $\mathcal{H}^n \rightarrow \mathcal{K}_n$, $A_I = |\xi\rangle\langle\xi_I|$ with a fixed vector $\xi \in \mathcal{K}_n$. For $\rho \in \mathcal{B}(\mathcal{K}_n)$ $\mathcal{D}^n(\rho)$ acts on $\mathcal{K}_n \subset \mathcal{H}^n$ and ρ and 0 on $\mathcal{H}^n \ominus \mathcal{K}_n$. Our task is to show that

$$F_n := \sum_I p_I F(|\psi_I\rangle\langle\varphi_I|, \mathcal{D}^n \cdot \mathcal{C}^n(|\psi_I\rangle\langle\psi_I|))$$

converges to 1. We give a lower estimate simply by neglecting the second term in the definition of $\mathcal{C}^n(\sigma)$:

$$\begin{aligned} F_n &\geq \sum_{I \in A_n} p_I \|P_n \psi_I\|^2 = \sum_{I \in A_n} p_I \sum_{J \in A_n} |\langle\psi_I | \xi_J\rangle|^2 \\ &= \sum_{J \in A_n} \lambda_J \end{aligned}$$

since

$$\sum_I p_I |\langle \psi_I | \xi_J \rangle|^2 = \langle \xi_J | D \otimes \cdots \otimes D | \xi_J \rangle = \lambda_J.$$

Above we observed that the lower bound goes to 1, hence $F_n \rightarrow 1$, in fact exponentially. \square

\square

Note that the pure states $|\psi_i\rangle$ compressed into mixed state in the scheme we have constructed. It is also remarkable that the statistical operator $\sum p_i |\psi_i\rangle\langle\psi_i|$ of the ensemble played a key role and not the ensemble itself. (Many different ensembles may have the same statistical operator.) To construct the compression we used the eigenbasis of the statistical operator and the value of its entropy. No further data was necessary.

The negative part of Schumacher's theorem depends on the **high probability subspace theorem** obtained by Hiai, Ohya and Petz ([16], [11]).

Theorem 5. *Let D be a density matrix acting on the Hilbert space \mathcal{H} . Then the n -fold tensor product $D_n := D \otimes D \otimes \cdots \otimes D$ acts on the n -fold product space $\mathcal{H}_n := \mathcal{H} \otimes \mathcal{H} \otimes \cdots \otimes \mathcal{H}$. For any $1 > \varepsilon > 0$ we have*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \inf \{ \log \text{Tr } Q_n : Q_n \text{ is a projection on } \mathcal{H}_n, \text{Tr } D_n Q_n \geq 1 - \varepsilon \} = S(D).$$

Roughly speaking the theorem tells that a projection Q_n of large probability has the dimension $\exp(nS(D))$.

Proof. First we construct projections of high probability and of small dimension. Fix $\delta > 0$ and let $P(n, \delta)$ be the spectral projection of $-\frac{1}{n} \log D_n$ corresponding to the interval $S(D) - \delta, S(D) + \delta$. It follows that

$$(S(D) - \delta)P(n, \delta) \leq \left(-\frac{1}{n} \log D_n \right) P(n, \delta) \leq (S(D) + \delta)P(n, \delta)$$

and hence

$$e^{-n(S(D)+\delta)} P_{(n,\delta)} \leq D_n P(n, \delta) \leq e^{-n(S(D)-\delta)} P_{(n,\delta)}.$$

From this we easily conclude

$$\frac{1}{n} \log \text{Tr } P(n, \delta) \leq S(D) + \delta.$$

and $\limsup_{n \rightarrow \infty} \leq S(D)$ follows concerning the limit in the statement.

Let now Q_n be a projection on \mathcal{H}_n such that $\text{Tr } Q_n D_n \geq 1 - \varepsilon$. This implies

$$\liminf_{n \rightarrow \infty} \text{Tr } D_n Q_n P(n, \delta) \geq 1 - \varepsilon$$

since

$$\begin{aligned}\mathrm{Tr} D_n Q_n P(n, \delta) &= \mathrm{Tr} D_n Q_n - \mathrm{Tr} D_n Q_n P(n, \delta)^\perp \\ &\geq \mathrm{Tr} D_n Q_n - \mathrm{Tr} D_n P(n, \delta)^\perp.\end{aligned}$$

Next we estimate as follows:

$$\begin{aligned}\mathrm{Tr} Q_n &\geq \mathrm{Tr} Q_n P(n, \delta) \\ &\geq \mathrm{Tr} D_n Q_n P(n, \delta) e^{n(S(D)-\delta)} \\ &= e^{n(S(D)-\delta)} \cdot \mathrm{Tr} D_n Q_n P(n, \delta)\end{aligned}$$

and

$$\frac{1}{n} \log \mathrm{Tr} Q_n \geq S(D) - \delta + \frac{1}{n} \log \mathrm{Tr} D_n Q_n P(n, \delta).$$

When $n \rightarrow \infty$ the last term of the right hand side converges to 0.

□

Now we turn back to Schumacher's theorem and present the negative part.

Theorem 6. *Let $(p_i, |\psi_i\rangle)$ be a source of pure states on a Hilbert space \mathcal{H} and let S be the von Neumann entropy of the density matrix $\sum_i p_i |\psi_i\rangle\langle\psi_i|$. If $R < S$, then reliable compression scheme of rate R does not exist.*

Proof. Assume that a reliable compression scheme of rate $R < S$ exists. Then

$$\begin{aligned}F_n &:= \sum_I p_I F(|\psi_I\rangle\langle\varphi_I|, \mathcal{D}^n \cdot \mathcal{C}^n(|\psi_I\rangle\langle\psi_I|)) \\ &= \sum_I p_I \sqrt{\langle\psi_I| \mathcal{C}^n(|\psi_I\rangle\langle\psi_I|) |\psi_I\rangle} \\ &\leq \sqrt{\sum_I p_I \langle\psi_I| \mathcal{C}^n(|\psi_I\rangle\langle\psi_I|) |\psi_I\rangle}\end{aligned}$$

by concavity. Moreover,

$$\begin{aligned}\sum_I p_I \langle\psi_I| \mathcal{C}^n(|\psi_I\rangle\langle\psi_I|) |\psi_I\rangle &= \sum_I p_I \mathrm{Tr} |\psi_I\rangle\langle\psi_I| P_n \mathcal{C}^n(|\psi_I\rangle\langle\psi_I|) \\ &\leq \sum_I p_I \mathrm{Tr} |\psi_I\rangle\langle\psi_I| P_n = \mathrm{Tr} D_n P_n\end{aligned}$$

for the projection P_n of \mathcal{H}^n onto \mathcal{K}_n . Since the compression is of rate R we have

$$\lim_n \frac{1}{n} \log \dim P_n \leq R.$$

On the other hand, the high probability subspace theorem tells us that in this case

$$\limsup_n \mathrm{Tr} D_n P_n \geq 1 - \varepsilon$$

is impossible for any $0 < \varepsilon < 1$. We have arrived at a contradiction with the assumption that the average fidelity is converging to 1. In fact, we have shown that for any compression scheme of rate R the average fidelity converges to 0.

□

3 Extension to sources with memory

Extension of Schumacher's source coding theorem is possible in several directions. One way would be to allow a source of mixed states. Not much is known about this direction and we refer to [2], where this problem is discussed. Our attention here will be focused on sources having some memory but producing pure states. In this case the optimal compression rate depends on the density matrix of the source only.

Let \mathcal{H} be a finite-dimensional Hilbert space and $\mathcal{H}_n := \mathcal{H} \otimes \mathcal{H} \otimes \cdots \otimes \mathcal{H}$. Let X^n denote the set of all messages of length n . If $\mathbf{x} \in X^n$ is a message, then a quantum state $|\psi(\mathbf{x})\rangle$ of the n -fold quantum system is corresponded with it. If \mathbf{x} appears with probability $p(\mathbf{x})$, then

$$D_n := \sum_{\mathbf{x}} p(\mathbf{x}) |\psi(\mathbf{x})\rangle \langle \psi(\mathbf{x})|$$

is the density matrix of the n -shot source. This general formulation allows the case

$$|\psi(\mathbf{x})\rangle = |\psi(x_1)\rangle \otimes |\psi(x_2)\rangle \otimes \cdots \otimes |\psi(x_n)\rangle \quad (\mathbf{x} = (x_1, x_2, \dots, x_n)),$$

but the scheme is more general. It turns out that the optimal compression rate will depend on the density matrices D_n only, hence we do not assume anything about the probability distributions $p(\mathbf{x})$, however we make some assumption on the sequence D_n of density matrices. We always assume that the von Neumann entropy density

$$h := \lim_{n \rightarrow \infty} \frac{1}{n} S(D_n) \quad (4)$$

exists. This holds in many examples. For $0 < \varepsilon < 1$, set

$$\beta_n(\varepsilon) := \inf \{ \log \text{Tr}(q) : q \text{ is a projection on } \mathcal{H}_n, \text{Tr } D_n q \geq 1 - \varepsilon \}.$$

We shall say that the high-probability subspace theorem holds if

$$\text{(HP)} \quad \lim_{n \rightarrow \infty} \frac{1}{n} \beta_n(\varepsilon) = h.$$

Since we want to let $n \rightarrow \infty$, it is reasonable to view all the n -fold systems as subsystems of an infinite one. Let an infinitely extended system be considered over the lattice \mathbf{Z} of integers. The observables confined to a lattice site $k \in \mathbf{Z}$ form the self-adjoint part of a finite-dimensional matrix algebra \mathcal{A}_k , that is the set of all operators acting on the finite-dimensional space \mathcal{H} . It is assumed that the local observables in any finite subset $\Lambda \subset \mathbf{Z}$ are those of the finite quantum system

$$\mathcal{A}_\Lambda = \bigotimes_{k \in \Lambda} \mathcal{A}_k.$$

The quasilocal algebra \mathcal{A} is the norm completion of the normed algebra $\mathcal{A}_\infty = \cup_\Lambda \mathcal{A}_\Lambda$, the union of all local algebras \mathcal{A}_Λ associated with finite intervals $\Lambda \subset \mathbf{Z}$.

A state φ of the infinite system is a positive normalised functional $\mathcal{A} \rightarrow \mathbb{C}$. It does not make sense to associate a statistical operator to a state of the infinite system in general. However, φ restricted to a finite-dimensional local algebra \mathcal{A}_Λ admits a density matrix D_Λ . We regard the algebra $\mathcal{A}_{[1,n]}$ as the set of all operators acting on the n -fold tensor product space $\mathcal{H}^{\otimes n}$. Moreover, we assume that the density D_n from the first part of this section is identical with $D_{[1,n]}$. Under this assumptions we call the state φ the state of the (infinite) channel. If φ happens to be a product, then we are in the memoryless setting discussed above. Now we want to allow memory effect and pose weaker conditions.

The right shift on the set \mathbf{Z} induces a transformation γ on \mathcal{A} . A state φ is called **stationary** if $\varphi \circ \gamma = \varphi$. The state φ is called **ergodic** if it is an extremal point in the set of stationary states. Moreover, φ is **completely ergodic** when it is an extreme point for every $m \in \mathbb{N}$ in the convex set of all states ψ such that $\psi \circ \gamma^m = \psi$.

A weaker form of property **HP** was proven in [11] for a completely ergodic stationary state. The weak high-probability subspace theorem holds if

$$(\mathbf{w-HP}) \quad \limsup_{n \rightarrow \infty} \frac{1}{n} \beta_n(\varepsilon) \leq h \text{ and } \liminf_{n \rightarrow \infty} \frac{1}{n} \beta_n(\varepsilon) \geq \frac{1}{1-\varepsilon} h - \frac{\varepsilon}{1-\varepsilon} \log d,$$

and the McMillan-type convergence holds if

$$(\mathbf{McM}) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \log D_n = h \cdot I$$

in a certain topology. Loosely speaking $\mathbf{McM} \implies \mathbf{HP} \implies \mathbf{w-HP}$ and all these properties imply that the extension of Schumacher's theorem holds and the optimal compression rate is the von Neumann entropy density h .

A nice class of stationary states is formed by the **quantum Markov states** (finitely correlated states, algebraic states, or generalised Markov chains are other names for the same thing, see [1, 7, 12]). For those states property **HP** was also proved [12].

Let $H_{[1,n]}$ be the Hamiltonian of **stationary interaction of finite range**, see [3], or Sect. 15 of [16] for the definitions. Assume that D_n is the density of the local Gibbs state, that is

$$D_n := \frac{e^{H_{[1,n]}}}{\text{Tr } e^{H_{[1,n]}}}$$

and let ψ be the equilibrium state of the infinite system. The equilibrium state of the finite system $\mathcal{H}^{\otimes n}$ with some specified interaction could be a model for storing information. If the interaction is stationary and of finite range then the asymptotically optimal compression rate is again the von Neumann entropy density h , since the **McM** property holds in the GNS space for the state ψ [10].

4 Appendix

4.1 Stochastic mappings as state transformations

Assume that \mathcal{H} is the Hilbert space of our quantum system which initially has a statistical operator D (acting on \mathcal{H}). When the quantum system is not closed, it is coupled to another system, called **environment**. The environment has a Hilbert space \mathcal{H}_e and statistical operator D_e . Before interaction the total system has density $D_e \otimes D$. The dynamical change caused by the interaction is implemented by a unitary and $U(D_e \otimes D)U^*$ is the new statistical operator and the reduced density \tilde{D} is the new statistical operator of the quantum system we are interested in. The affine change $D \mapsto \tilde{D}$ is typical for quantum mechanics and called **stochastic mapping**.

The above defined stochastic mapping can be described in several other forms, reference to the environment could be omitted completely. Assume that D is an $n \times n$ matrix and D_e is of the form $(z_k \bar{z}_l)_{kl}$ where (z_1, z_2, \dots, z_m) is a unit vector in the m dimensional space \mathcal{H}_e . (D_e is pure state.) All operators acting on $\mathcal{H}_e \otimes \mathcal{H}$ are written in a block matrix form, they are $m \times m$ matrices with $n \times n$ matrix entries. In particular, $U = (U_{pq})_{p,q=1}^m$ and $U_{pq} \in M_n$. The definition of the reduced density matrix gives

$$\begin{aligned} \tilde{D} &= \sum_{p,q,r} U_{pq}(D_e \otimes D)_{qr}(U^*)_{rp} = \sum_p \left(\sum_q z_q U_{pq} \right) D \left(\sum_r z_r U_{pr} \right)^* \\ &= \sum_p A_p D A_p^* \end{aligned}$$

where the operators $A_p := \sum_q z_q U_{pq}$ satisfy

$$\sum_p A_p A_p^* = I. \quad (5)$$

Theorem 7. *Any stochastic mapping $D \mapsto \tilde{D}$ can be written in the form*

$$\tilde{D} = \sum_p A_p D A_p^*,$$

where the operator coefficients satisfy (5). Conversely, all transformation of this form are stochastic.

The first part of the theorem was obtained above. To prove the converse part, we need to solve the equations

$$\sum_q z_q U_{pq} = A_p \quad (p = 1, 2, \dots, m).$$

Choose simply $z_1 = 1$ and $z_2 = z_3 = \dots = z_m = 0$ and the equations reduce to $U_{p1} = A_p$. This means that the first column is given from the block matrix U and we need to determine the other columns such a way that U should be a unitary. Thanks to the condition

(5) this is possible. Condition (5) tells us that the first column of our block matrix determines an isometry which extends to a unitary.

The coefficients A_p in the **operator-sum representation** are called the **operation elements** of the stochastic map. The term quantum (state) operation is also often used instead of stochastic map.

The stochastic maps form a convex subset of the set of all positive trace preserving linear transformations.

4.2 Von Neumann entropy

The above formula for the von Neumann entropy is equivalently written as

$$S(D) = \text{Tr } \eta(D), \text{ where } \eta(t) = -t \log t.$$

Since η is a concave function on \mathbb{R}^+ , the von Neumann entropy is a **concave functional** on the state space. The maximum is reached at the density whose all the eigenvalues are the same and the minimum is at pure states.

A density matrix D admits generally many convex decomposition into pure states: $D = \sum_j \mu_j |\psi_j\rangle\langle\psi_j|$. The von Neumann entropy is the infimum of all Shannon entropies corresponding to those decompositions

$$S(D) = \inf \left\{ H(\mu_j) : D = \sum_j \mu_j |\psi_j\rangle\langle\psi_j|, \mu_j \geq 0, \sum_j \mu_j = 1 \right\}.$$

When D_{12} is a density matrix of a composite system $\mathcal{H}_1 \otimes \mathcal{H}_2$ with reduced density matrices D_1 and D_2 , respectively, then the **subadditivity** of the von Neumann entropy holds:

$$S(D_{12}) \leq S(D_1) + S(D_2)$$

This property is responsible for the fact that for shift invariant states of an infinite tensor product the von Neumann entropy density (4) exists.

Several chapters of the book [16] are devoted to properties and extensions of the von Neumann entropy. For a historical approach to the von Neumann entropy, see [17].

4.3 Fidelity

The general formula for the fidelity of the density matrices D_1 and D_2 is

$$F(D_1, D_2) = \text{Tr} \sqrt{D_1^{1/2} D_2 D_1^{1/2}}. \quad (6)$$

This quantity was studied by Uhlmann in a different context and he proved that

$$F(D_1, D_2) = \min \left\{ \sqrt{\text{Tr} (D_1 G) \text{Tr} (D_2 G^{-1})} : G \text{ is positive and invertible} \right\} \quad (7)$$

([20] and see [8] for a rather detailed discussion). From this the symmetry of $F(D_1, D_2)$ is obvious and we can easily deduce the **monotonicity of the fidelity** under stochastic state transformation:

$$\begin{aligned}
F(\mathcal{C}(D_1), \mathcal{C}(D_2))^2 &\geq \text{Tr } \mathcal{C}(D_1)G \text{Tr } \mathcal{C}(D_2)G^{-1} - \varepsilon \\
&\geq \text{Tr } D_1 \mathcal{C}^*(G) \text{Tr } D_2 \mathcal{C}^*(G^{-1}) - \varepsilon \\
&\geq \text{Tr } D_1 \mathcal{C}^*(G) \text{Tr } D_2 \mathcal{C}^*(G)^{-1} - \varepsilon \\
&\geq F(D_1, D_2)^2 - \varepsilon,
\end{aligned}$$

where \mathcal{C}^* is the adjoint of \mathcal{C} with respect to the Hilbert-Schmidt inner product, $\varepsilon > 0$ is arbitrary and G is chosen to be appropriate. It is well-known that \mathcal{C}^* is unital and positive, hence $\mathcal{C}^*(G)^{-1} \geq \mathcal{C}^*(G^{-1})$. In this way the monotonicity

$$F(\mathcal{C}(D_1), \mathcal{C}(D_2)) \geq F(D_1, D_2) \tag{8}$$

is concluded.

From the definition (6) one observes that $F(D_1, D_2)$ is concave in D_2 . (Remember that \sqrt{t} is operator concave.) However, the monotonicity gives that $F(D_1, D_2)$ is **jointly concave** as well. Consider the stochastic mapping

$$\mathcal{C} : \begin{bmatrix} A & B \\ C & D \end{bmatrix} \mapsto A + D$$

Then

$$\begin{aligned}
\lambda F(D_1, D_2) + (1 - \lambda)F(D'_1, D'_2) &= F\left(\begin{bmatrix} \lambda D_1 & 0 \\ 0 & (1 - \lambda)D'_1 \end{bmatrix}, \begin{bmatrix} \lambda D_2 & 0 \\ 0 & (1 - \lambda)D'_2 \end{bmatrix}\right) \\
&\leq F(\lambda D_1 + (1 - \lambda)D'_1, \lambda D_2 + (1 - \lambda)D'_2)
\end{aligned}$$

as an application of monotonicity and the concavity is obtained.

Another remarkable formula is

$$\begin{aligned}
F(D_1, D_2) = \max \{ & \langle \psi_1 | \psi_2 \rangle \mid : \mathcal{C}(|\psi_1 \rangle \langle \psi_1|) = D_1, \\
& \mathcal{C}(|\psi_2 \rangle \langle \psi_2|) = D_2 \text{ for some stochastic mapping } \mathcal{C} \}
\end{aligned}$$

which has a certain **operational meaning** [6].

References

- [1] L. Accardi and A. Frigerio, Markov cocycles, Proc. R. Ir. Acad. **83A**, 251–269 (1983).
- [2] H. Barnum, C.M. Caves, C.A. Fuchs, R. Jozsa and B.W. Schumacher, On quantum coding for ensembles of mixed states, J. Phys. A **34**, 6767– 6785 (2001)

- [3] O. Bratteli and D.W. Robinson, *Operator Algebras and Quantum Statistical Mechanics II*, Springer-Verlag, New York-Heidelberg-Berlin, 1981
- [4] T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley (1991).
- [5] I. Csiszár and J. Körner, *Information theory. Coding theorems for discrete memoryless systems*, Akadémiai Kiadó, Budapest (1981).
- [6] J. L. Dold and M. A. Nielsen, A simple operational interpretation of fidelity, arXiv e-print quant-ph/0111053
- [7] M. Fannes, B. Nachtergaele and R. F. Werner, Finetely correlated states on quantum spin chains, *Comm. Math. Phys.* **144**, 443–490 (1992).
- [8] C. A. Fuchs, Ph. D. thesis, The university of New Mexico, Albuquerque, 1996, arXiv e-print quant-ph/9601020
- [9] R. M. Gray, *Entropy and Information Theory*, Springer, 1990.
- [10] F. Hiai, M. Ohya, D. Petz, McMillan type convergence for quantum Gibbs states, *Arch. der Math.* **65**, 154–158 (1995).
- [11] F. Hiai, D. Petz, The proper formula for relative entropy and its asymptotics in quantum probability, *Commun. Math. Phys.* **143**, 99-114 (1991).
- [12] F. Hiai, D. Petz, Entropy density for algebraic states, *J. Functional Anal.* **125**, 287–308 (1994).
- [13] A.S. Holevo, Quantum coding theorems, *Russian Math. Surveys*, **53**, 1295–1331 (1998).
- [14] R. Jozsa, B. Schumacher, A new proof of the quantum noiseless coding theorem, *J. Modern Optics*, **41**, 2343–2349 (1994).
- [15] M.A. Nielsen, I.L. Chuang, *Quantum Computation and Quantum Information*, Cambridge University Press, 2000.
- [16] M. Ohya, D. Petz, *Quantum Entropy and Its Use*, Springer-Verlag, Heidelberg, 1993.
- [17] D. Petz, Entropy, von Neumann and the von Neumann entropy, in *John von Neumann and the Foundations of Quantum Physics*, eds. M. Rédei and M. Stöltzner, Kluwer, 2001
- [18] D. Petz, M. Mosonyi, Stationary quantum source coding, *J. Math. Phys.* **42**, 4857–4864 (2001).
- [19] B. Schumacher, Quantum coding, *Phys. Rev. A* **51**, 2738–2747 (1995).
- [20] A. Uhlmann, *Rep. Math. Phys.* **9**, 273– (1976).