Schumacher Compression

Kim Boström
Institut für Physik, Universität Potsdam, 14469 Potsdam, Germany [*]

The compression scheme raised by Schumacher [1, 2] is a quantized version of Shannon's source coding theorem [3]. Say, we have an alphabet $\mathcal{A} = \{a_1, \ldots, a_K\}$ of K letters. Alice composes a "quantum message" by assigning to each classical letter a_k a quantum state $|a_k\rangle$ taken from some Hilbert state $\mathcal{H}$,

$$a_k \mapsto |a_k\rangle. \tag{1}$$

Doing so, the alphabet $\mathcal{A}$ is mapped to a *quantum alphabet* $\mathcal{Q}_\mathcal{A}$ consisting of K *quantum letters* $|a_k\rangle$,

$$\mathcal{Q}_\mathcal{A} = \{|a_1\rangle, \ldots, |a_K\rangle\}. \tag{2}$$

The span of the quantum alphabet $\mathcal{Q}_A$ is the letter space

$$\mathcal{H}_A := \mathrm{Span}\mathcal{Q}_A, \tag{3}$$

which is a subspace of $\mathcal{H}$. Alice composes a random message by choosing the quantum letter $|a_k\rangle$ with *a priori* probability $p_k = p(a_k)$. We may associate a *letter matrix* $\hat\rho$ corresponding to the ensemble of letter states,

$$\hat\rho = \sum_{x\in\mathcal{A}} p(x)\,|x\rangle\langle x| = \sum_{k=1}^{K} p_k\,|a_k\rangle\langle a_k|. \tag{4}$$

If Alice composes a random message $\boldsymbol{x} \equiv x_1 \cdots x_N$ of length N from the message set $\mathcal{M} \equiv \mathcal{A} \times \cdots \times \mathcal{A}$ by choosing N letters independently from the same letter ensemble. The resulting quantum message has the form

$$|\boldsymbol{x}\rangle \equiv |x_1\rangle \otimes \cdots \otimes |x_N\rangle, \tag{5}$$

where each letter $|x_n\rangle$ is an element of $\mathcal{Q}_\mathcal{A}$ and the entire message is a vector from the Hilbert space

$$\mathcal{H}_M := \mathcal{H}_A^{\otimes N} \equiv \mathcal{H}_A \otimes \cdots \otimes \mathcal{H}_A. \tag{6}$$

The message $|\boldsymbol{x}\rangle$ appears with probability $p(\boldsymbol{x}) = p(x_1) \cdots p(x_N)$, such that the total message ensemble can be represented by the *message matrix*

$$\hat{\boldsymbol{\rho}} \;=\; \hat\rho^{\otimes N} = \sum_{\boldsymbol{x}\in\mathcal{M}} p(\boldsymbol{x})\,|\boldsymbol{x}\rangle\langle\boldsymbol{x}|. \tag{7}$$

Before sending her block message to Bob, Alice has to convert the message into a sequence of qubits, because the channel to Bob only accepts qubits. Therefore, she wants to build an *encoder* $\hat C$ that unitarily maps the source space $\mathcal{H}_M$ to a code space $\mathcal{H}_C$ of qubits,

$$\hat C : \mathcal{H}_M \to \mathcal{H}_C. \tag{8}$$

In order for $\hat C$ to be unitary, the dimension of $\mathcal{H}_C$ must be equal to the dimension of the source space $\mathcal{H}_M$. The dimension of the alphabet space is at most K, but the quantum letters $|a_k\rangle$ do not have to be *mututally orthogonal*, yet they do not even have to be *linearly independent*, so the dimension of $\mathcal{H}_A$ can in fact be *smaller* than the number of alphabet letters, which gives

$$\log(\dim\mathcal{H}_M) = N\log(\dim\mathcal{H}_A) \le N\log K, \tag{9}$$

where logs are binary, here and in the following. A message $|\boldsymbol{x}\rangle$ is encoded into the binary message $|c(\boldsymbol{x})\rangle$ by applying the encoder $\hat C$,

$$|c(\boldsymbol{x})\rangle \;:=\; \hat C|\boldsymbol{x}\rangle. \tag{10}$$

Note that while $|\boldsymbol{x}\rangle$ is by construction a product state, the code state $|c(\boldsymbol{x})\rangle$ can be highly entangled. Let the alphabet space $\mathcal{H}_A$ have dimension L, then in order to encode every message in $\mathcal{H}_M$ we need a qubit space of dimension

$$\dim\mathcal{H}_C = \dim\mathcal{H}_M = L^N. \tag{11}$$

where we assume that L and N are chosen such that the above number is a power of two. In other words, we need $N\log L$ qubits to encode each message in $\mathcal{H}_M$ with perfect fidelity. The decoding procedure is represented by the inverse operator $\hat D : \mathcal{H}_C \to \mathcal{H}_M$,

$$\hat D := \hat C^\dagger. \tag{12}$$

Since the use of a quantum channel is very expensive, we want to save qubits for the transmission. Let us look for an encoder $\hat C$ that is restricted to a proper subspace $\Lambda \subset \mathcal{H}_M$ with a dimension significantly smaller than L^N, such that we still achieve asymptotically faithful decoding. First, we perform a diagonalization of the letter matrix, resulting in

$$\hat\rho = \sum_{l=1}^{L} q_l\,|\lambda_l\rangle\langle\lambda_l|. \tag{13}$$

The number L of $\hat\rho$-eigenstates coincides with the dimension of the alphabet subspace $\mathcal{H}_A$. We have

$$\hat\rho\log\hat\rho = \sum_{l=1}^{L} q_l\log q_l\,|\lambda_l\rangle\langle\lambda_l|, \tag{14}$$

such that

$$\mathrm{Tr}\{\hat\rho\log\hat\rho\} \;=\; \sum_{l=1}^{L} q_l\log q_l = H(Y), \tag{15}$$

[*] Electronic address: bostroem@qipc.org

where Y denotes the ensemble of $\hat{\rho}$-eigenstates. Defining the *von-Neumann entropy* of $\hat{\rho}$ as

$$S(\hat{\rho}) := \mathrm{Tr}\{\hat{\rho}\log\hat{\rho}\}, \tag{16}$$

we see that the von-Neumann entropy of $\hat{\rho}$ equals the Shannon entropy of the ensemble of $\hat{\rho}$-eigenstates,

$$S(\hat{\rho}) = H(Y). \tag{17}$$

One can show that the von-Neumann entropy is bounded from above by the Shannon entropy of the letter ensemble X,

$$S(\hat{\rho}) \leq H(X), \tag{18}$$

where equality holds in the case of mutual orthogonal letter states.

Quantum mechanics tells us that the scenario where Alice sends the ensemble Y cannot by any experiment be distinguished from the actual scenario where Alice sends the ensemble X. However, sending the ensemble Y corresponds to a classical situation. Consider the sequence $|\boldsymbol{y}\rangle \equiv |y_1\cdots y_N\rangle$ of basis states $|y_n\rangle \in \mathcal{B}_A$, which appear with probability $q(\boldsymbol{y}) = q(y_1)\cdots q(y_N)$. Just like in Shannon's noiseless coding theorem we introduce a typical subset T of messages $\boldsymbol{y}$ appearing with probability

$$2^{-N(S+\delta)} \leq q(\boldsymbol{y}) \leq 2^{-N(S-\delta)}, \tag{19}$$

where we have used the fact that $H(Y) = S(\hat{\rho}) \equiv S$. Then we define the *typical subspace* $\Lambda \subset \mathcal{H}_A$ as the space spanned by the typical messages,

$$\Lambda := \mathrm{Span}\{|\boldsymbol{y}\rangle \mid \boldsymbol{y} \in T\}. \tag{20}$$

Exploiting Shannon's theorem we know that for any fixed $\epsilon, \delta > 0$ there is a big enough N such that

$$P_\Lambda \geq 1 - \epsilon, \tag{21}$$

where P_Λ is the total probability of all members of T,

$$P_\Lambda = \sum_{\boldsymbol{y}\in T} q(\boldsymbol{y}). \tag{22}$$

Since the typical subspace Λ is spanned by the typical messages $|\boldsymbol{y}\rangle$ where $\boldsymbol{y} \in T$, the dimension of Λ is given by the size of T, so Shannon's theorem implies that

$$(1-\epsilon)2^{N(S-\delta)} \leq \dim\Lambda \leq 2^{N(S+\delta)}. \tag{23}$$

In the asymptotic limit $N \to \infty$, the dimension of the subspace approaches

$$\dim\Lambda \to 2^{NS}. \tag{24}$$

Because we have

$$S(\hat{\rho}) = \sum_{l=1}^{L} q_l \log q_l \leq \log L, \tag{25}$$

the dimension of Λ is smaller than or equal to the dimension of the space of all messages,

$$\dim\Lambda = 2^{NS(\hat{\rho})} \leq 2^{N\log L} = \dim\mathcal{H}_M. \tag{26}$$

In practice, except for the case of uniformly distributed letters, the typical subspace will have a *dramatically* smaller dimension (for large N). Hence we can save resources by encoding only the component of $|\boldsymbol{x}\rangle$ that lies in the typical subspace Λ. To this aim we need the projector onto the typical subspace, which is given by

$$\hat{\Pi}_\Lambda = \sum_{\boldsymbol{y}\in T} |\boldsymbol{y}\rangle\langle\boldsymbol{y}|. \tag{27}$$

Now we restrict the encoder to the typical subspace, $\hat{C} : \Lambda \to \mathcal{H}_C$, where it shall be a unitary operator. *Schumacher compression* goes as follows. First, Alice projects her source message $|\boldsymbol{x}\rangle$ onto the typical subspace Λ. With probability

$$P_\Lambda(\boldsymbol{x}) := \langle\boldsymbol{x}|\hat{\Pi}_\Lambda|\boldsymbol{x}\rangle = \mathrm{Tr}\{|\boldsymbol{x}\rangle\langle\boldsymbol{x}|\hat{\Pi}_\Lambda\}. \tag{28}$$

such projection will be successful and results in the state

$$|\phi(\boldsymbol{x})\rangle := \frac{1}{\sqrt{P_\Lambda(\boldsymbol{x})}}\hat{\Pi}_\Lambda|\boldsymbol{x}\rangle, \tag{29}$$

The average probability of a successful projection thus reads

$$\begin{aligned}
\langle P_\Lambda(\boldsymbol{X})\rangle &= \sum_{\boldsymbol{x}\in\mathcal{M}} p(\boldsymbol{x})P_\Lambda(\boldsymbol{x}) && (30)\\
&= \sum_{\boldsymbol{x}\in\mathcal{M}} p(\boldsymbol{x})\mathrm{Tr}\{|\boldsymbol{x}\rangle\langle\boldsymbol{x}|\hat{\Pi}_\Lambda\} && (31)\\
&= \mathrm{Tr}\Big\{\sum_{\boldsymbol{x}\in\mathcal{M}} p(\boldsymbol{x})|\boldsymbol{x}\rangle\langle\boldsymbol{x}|\hat{\Pi}_\Lambda\Big\} && (32)\\
&= \mathrm{Tr}\{\hat{\boldsymbol{\rho}}\,\hat{\Pi}_\Lambda\}, && (33)
\end{aligned}$$

that is,

$$\langle P_\Lambda(\boldsymbol{X})\rangle = \mathrm{Tr}\{\hat{\boldsymbol{\rho}}\hat{\Pi}_\Lambda\}. \tag{34}$$

Let us proceed,

$$\begin{aligned}
\mathrm{Tr}\{\hat{\boldsymbol{\rho}}\hat{\Pi}_\Lambda\} &= \sum_{\boldsymbol{y}\in T}\langle\boldsymbol{y}|\hat{\boldsymbol{\rho}}|\boldsymbol{y}\rangle && (35)\\
&= \sum_{\boldsymbol{y}\in T} q(\boldsymbol{y}) \equiv P_\Lambda \geq 1-\epsilon, && (36)
\end{aligned}$$

where in the last step we used Shannon's theorem. We arrive at

$$\langle P_\Lambda(\boldsymbol{X})\rangle = P_\Lambda \geq 1 - \epsilon, \tag{37}$$

hence the projection will be 100% successful in the asymptotic limit of infinitely long messages. After projection, Alice can encode the resulting state $|\phi(\boldsymbol{x})\rangle$ by $\hat{C}$ and send it to Bob, who then applies the inverse operation $\hat{C}^\dagger$ to obtain the state $|\phi(\boldsymbol{x})\rangle$. If the overlap of $|\phi(\boldsymbol{x})\rangle$

with the original message is big enough, it was an approximately faithful transmission. If the projection was not successful, Alice prepares some garbage state $|\phi_0\rangle \in \Lambda$, encodes it by $\hat{C}$ and sends it to Bob. In this case, the overlap with the orininal message $|\boldsymbol{x}\rangle$ is hopefully very small. To put this more precisely, we describe the statistical ensemble of successful and unsuccessful projections by a density matrix. The probability that the projection is *not* successful reads

$$1 - P_\Lambda(\boldsymbol{x}) = \langle \boldsymbol{x}|(\mathbb{1} - \hat{\Pi}_\Lambda)|\boldsymbol{x}\rangle. \tag{38}$$

So after the projection procedure the message will be in the mixed state

$$\begin{aligned} \hat{\rho}_{\boldsymbol{x}} &= P_\Lambda(\boldsymbol{x})|\phi(\boldsymbol{x})\rangle\langle\phi(\boldsymbol{x})| + (1 - P_\Lambda(\boldsymbol{x}))|\phi_0\rangle\phi_0| & (39)\\ &= \hat{\Pi}_\Lambda|\boldsymbol{x}\rangle\langle\boldsymbol{x}|\hat{\Pi}_\Lambda + (1 - P_\Lambda(\boldsymbol{x}))|\phi_0\rangle\langle\phi_0|. & (40) \end{aligned}$$

The subsequently performed encoding procedure by $\hat{C}$ maps the state $\hat{\rho}_{\boldsymbol{x}}$ to the qubit state $\hat{C}\hat{\rho}_{\boldsymbol{x}}\hat{C}^\dagger$, which is then send to Bob through the quantum channel. After receiving the code message, Bob applies the decoder $\hat{D} = \hat{C}^\dagger$ to it and since $\hat{C}$ is unitary, he recovers the state $\hat{\rho}_{\boldsymbol{x}}$. Originally, the message was given by the pure state $|\boldsymbol{x}\rangle\langle\boldsymbol{x}|$. The fidelity between original and decoded message is given by

$$\begin{aligned} F(\boldsymbol{x}) &= \langle\boldsymbol{x}|\hat{\rho}_{\boldsymbol{x}}|\boldsymbol{x}\rangle & (41)\\ &= \langle\boldsymbol{x}|\hat{\Pi}_\Lambda|\boldsymbol{x}\rangle\langle\boldsymbol{x}|\hat{\Pi}_\Lambda|\boldsymbol{x}\rangle + r_{\boldsymbol{x}} & (42)\\ &= P_\Lambda^2(\boldsymbol{x}) + r_{\boldsymbol{x}} & (43)\\ &\geq P_\Lambda^2 \geq 2\,P_\Lambda(\boldsymbol{x}) - 1, & (44) \end{aligned}$$

where we used

$$r_{\boldsymbol{x}} := \langle\boldsymbol{x}|\big\{(1 - P_\Lambda(\boldsymbol{x}))|\phi_0\rangle\langle\phi_0|\big\}|\boldsymbol{x}\rangle \geq 0 \tag{45}$$

together with the inequality $x^2 \geq 2x - 1$, which holds for all real numbers x. So the average fidelity for the ensemble $\boldsymbol{x}$ of source messages reads

$$\begin{aligned} F &= \sum_{\boldsymbol{x}\in\mathcal{M}} p(\boldsymbol{x})\,F(\boldsymbol{x}) & (46)\\ &\geq \sum_{\boldsymbol{x}\in\mathcal{M}} p(\boldsymbol{x})\,\big(2\,P_\Lambda(\boldsymbol{x}) - 1\big) & (47)\\ &= 2\,\mathrm{Tr}\{\hat{\boldsymbol{\rho}}\hat{\Pi}_\Lambda\} - 1 \geq 1 - 2\epsilon, & (48) \end{aligned}$$

where we used (36). We arrive at the important conclusion: The average fidelity of the decoded states with the original messages tends to unity in the limit of infinitely long messages. States that do not survive the projection will be all encoded by the same junk state, which thus cannot be faithfully decoded to give the original message. Happily, the probability of such erroneous decoding vanishes in the limit of infinitely long messages. Since the dimension of the typical space approaches $d \to 2^{NS}$, we need $I_N = NS(\hat{\rho})$ qubits to encode each typical message, hence per source letter we need

$$I = S(\hat{\rho}) \tag{49}$$

qubits in the limit of infinitely long messages, which represents a significant compression in most practical cases. Now let us investigate if we can achieve a compression below $S(\rho)$ qubits. Just like in the classical case, we fix some $\epsilon' > 0$ and project the source message on a "subtypical subspace" $\Lambda' \subset \Lambda$ whose dimension is

$$\dim\Lambda' \leq (1 - \epsilon)2^{N(S-\delta-\epsilon')} < 2^{N(H-\delta-\epsilon')}. \tag{50}$$

Let the space Λ' be spanned by the messages in a "subtypical set" $T' \subset T$,

$$\Lambda' := \mathrm{Span}\{|\boldsymbol{y}\rangle \mid \boldsymbol{y} \in T'\}, \tag{51}$$

so the dimension of Λ' equals the size of T',

$$\dim\Lambda' = |T'|. \tag{52}$$

The probability that a given message $|\boldsymbol{x}\rangle$ is successfully projected onto Λ' reads

$$\begin{aligned} P_{\Lambda'}(\boldsymbol{x}) &= \langle\boldsymbol{x}|\hat{\Pi}_{\Lambda'}|\boldsymbol{x}\rangle & (53)\\ &= \mathrm{Tr}\{|\boldsymbol{x}\rangle\langle\boldsymbol{x}|\hat{\Pi}_{\Lambda'}\}, & (54) \end{aligned}$$

and the projected state is then given by

$$|\phi'(\boldsymbol{x})\rangle := \frac{1}{\sqrt{P_{\Lambda'}(\boldsymbol{x})}}\hat{\Pi}_{\Lambda'}|\boldsymbol{x}\rangle. \tag{55}$$

The average probability that a message is successfully projected onto Λ' yields

$$\begin{aligned} P_{\Lambda'} &= \sum_{\boldsymbol{x}\in\mathcal{M}} p(\boldsymbol{x})P_{\Lambda'}(\boldsymbol{x}) & (56)\\ &= \sum_{\boldsymbol{x}\in\mathcal{M}} p(\boldsymbol{x})\,\mathrm{Tr}\{|\boldsymbol{x}\rangle\langle\boldsymbol{x}|\hat{\Pi}_{\Lambda'}\} & (57)\\ &= \mathrm{Tr}\{\hat{\boldsymbol{\rho}}\hat{\Pi}_{\Lambda'}\} = \sum_{\boldsymbol{y}\in T'}\langle\boldsymbol{y}|\hat{\boldsymbol{\rho}}|\boldsymbol{y}\rangle & (58)\\ &= \sum_{\boldsymbol{y}\in T'} q(\boldsymbol{y}) & (59)\\ &\leq q_{max}|T'| \leq 2^{-N(S-\delta)}2^{N(S-\delta-\epsilon')} & (60)\\ &= 2^{-N\epsilon'}, & (61) \end{aligned}$$

which vanishes for $N \to 0$. So already the projection will fail in the limit of long messages. This implies that the state $\hat{\rho}_{\boldsymbol{x}}$ after the projection will contain a vanishing component of the original message,

$$\begin{aligned} \hat{\rho}_{\boldsymbol{x}} &= P_{\Lambda'}(\boldsymbol{x})|\phi(\boldsymbol{x})\rangle\langle\phi(\boldsymbol{x})| + (1 - P_{\Lambda'}(\boldsymbol{x}))|\phi_0\rangle\phi_0| & (62)\\ &= \hat{\Pi}_{\Lambda'}|\boldsymbol{x}\rangle\langle\boldsymbol{x}|\hat{\Pi}_{\Lambda'} + (1 - P_{\Lambda'}(\boldsymbol{x}))|\phi_0\rangle\langle\phi_0|. & (63) \end{aligned}$$

The fidelity of $\hat{\rho}_{\boldsymbol{x}}$ with $|\boldsymbol{x}\rangle$ will also vanish,

$$\begin{aligned} F(\boldsymbol{x}) &= \langle\boldsymbol{x}|\hat{\rho}_{\boldsymbol{x}}|\boldsymbol{x}\rangle & (64)\\ &= P_{\Lambda'}^2(\boldsymbol{x}) + r_{\boldsymbol{x}} & (65)\\ &\leq P_{\Lambda'}(\boldsymbol{x}) + r_{\boldsymbol{x}} & (66)\\ &= \mathrm{Tr}\{|\boldsymbol{x}\rangle\langle\boldsymbol{x}|\hat{\Pi}_{\Lambda'}\} + r_{\boldsymbol{x}}, & (67) \end{aligned}$$

where we used $P_{\Lambda'}(\boldsymbol{x}) \leq 1$ and defined

$$r_{\boldsymbol{x}} := \langle \boldsymbol{x} | \big\{ (1 - P_{\Lambda'}(\boldsymbol{x})) |\phi_0\rangle\langle\phi_0| \big\} |\boldsymbol{x}\rangle \geq 0 \qquad (68)$$

The average fidelity becomes

$$\begin{aligned} F &= \sum_{\boldsymbol{x}\in\mathcal{M}} p(\boldsymbol{x})\, F(\boldsymbol{x}) & (69)\\ &\leq \mathrm{Tr}\{\hat{\boldsymbol{\rho}}\hat{\Pi}_{\Lambda'}\} + \sum_{\boldsymbol{x}\in\mathcal{M}} p(\boldsymbol{x})\, r_{\boldsymbol{x}} & (70)\\ &= P_{\Lambda'} + r & (71)\\ &\leq 2^{-N\epsilon'} + r, & (72) \end{aligned}$$

where

$$r := \sum_{\boldsymbol{x}\in\mathcal{M}} p(\boldsymbol{x})\, r_{\boldsymbol{x}}. \qquad (73)$$

In the limit $N \to \infty$, the average fidelity will approach

$$F \to r, \qquad (74)$$

which is just the average overlap of the source message ensemble with the garbage state $|\phi_0\rangle$. So even if the coding fails, there is still a chance to accidentally decode the correct message. However, such chance has nothing to do with faithful decoding, because the garbage state does not contain *any* information about the original message. Bob could simply *guess* the correct message with non-zero probability. We can get rid of r by choosing $|\phi_0\rangle$ orthogonal to all source messages.

Concluding, we arrive at the quantum analog of Shannon's source coding theorem: A letter ensemble $\hat{\rho}$ can be compressed to $S(\rho)$ qubits in the limit of infinitely long messages. Compressing to fewer than $S(\rho)$ qubits results in a loss of all information in the limit of infinitely long messages. A good review on the issue can also be found in [4, 5].

[1] R. Jozsa and B. Schumacher. A new proof of the quantum noiseless coding theorem. *J. Mod. Opt.*, **41**, 2343 (1994).

[2] B. Schumacher. Quantum coding. *Phys. Rev. A*, **51**, 2738–2747 (1995).

[3] C. E. Shannon and W. Weaver. A mathematical Theory of communication, *The Bell System Technical Journal*, **27**, 379–423,623–656, (1948).

[4] D.J.C. MacKay. Information theory, inference, and learning algorithms, http://wol.ra.phy.cam.ac.uk/mackay/itprnn/book.html, (1995-2000).

[5] J. Preskill. Lecture notes. http://www.theory.caltech.edu/people/preskill/ph219/, (1997-1999).