CS 766/QIC 820 Theory of Quantum Information (Fall 2011)

Lecture 9: Entropy and compression

For the next several lectures we will be discussing the von Neumann entropy and various concepts relating to it. This lecture is intended to introduce the notion of entropy and its connection to compression.

9.1 Shannon entropy

Before we discuss the von Neumann entropy, we will take a few moments to discuss the Shannon entropy. This is a purely classical notion, but it is appropriate to start here. The *Shannon entropy* of a probability vector $p \in \mathbb{R}^{\Sigma}$ is defined as follows:

$$H(p) = -\sum_{\substack{a \in \Sigma \\ p(a) > 0}} p(a) \log(p(a)).$$

Here, and always in this course, the base of the logarithm is 2. (We will write $\ln(\alpha)$ if we wish to refer to the natural logarithm of a real number α.) It is typical to express the Shannon entropy slightly more concisely as

$$H(p) = -\sum_{a \in \Sigma} p(a) \log(p(a)),$$

which is meaningful if we make the interpretation $0 \log(0) = 0$. This is sensible given that

$$\lim_{\alpha \to 0^+} \alpha \log(\alpha) = 0.$$

There is no reason why we cannot extend the definition of the Shannon entropy to arbitrary vectors with nonnegative entries if it is useful to do this—but mostly we will focus on probability vectors.

There are standard ways to interpret the Shannon entropy. For instance, the quantity $H(p)$ can be viewed as a measure of the amount of uncertainty in a random experiment described by the probability vector p, or as a measure of the amount of information one gains by learning the value of such an experiment. Indeed, it is possible to start with simple axioms for what a measure of uncertainty or information should satisfy, and to derive from these axioms that such a measure must be equivalent to the Shannon entropy.

Something to keep in mind, however, when using these interpretations as a guide, is that the Shannon entropy is usually only a meaningful measure of uncertainty in an asymptotic sense—as the number of experiments becomes large. When a small number of samples from some experiment is considered, the Shannon entropy may not conform to your intuition about uncertainty, as the following example is meant to demonstrate.

Example 9.1. Let $\Sigma = \{0, 1, \ldots, 2^{m^2}\}$, and define a probability vector $p \in \mathbb{R}^{\Sigma}$ as follows:

$$p(a) = \begin{cases} 1 - \frac{1}{m} & a = 0, \\ \frac{1}{m} 2^{-m^2} & 1 \leq a \leq 2^{m^2}. \end{cases}$$

It holds that $H(p) > m$, and yet the outcome 0 appears with probability $1 - 1/m$. So, as m grows, we become more and more "certain" that the outcome will be 0, and yet the "uncertainty" (as measured by the entropy) goes to infinity.

The above example does not, of course, represent a paradox. The issue is simply that the Shannon entropy can only be interpreted as measuring uncertainty if the number of random experiments grows and the probability vector remains fixed, which is opposite to the example.

9.2 Classical compression and Shannon's source coding theorem

Let us now focus on an important use of the Shannon entropy, which involves the notion of a *compression scheme*. This will allow us to attach a concrete meaning to the Shannon entropy.

9.2.1 Compression schemes

Let $p \in \mathbb{R}^{\Sigma}$ be a probability vector, and let us take $\Gamma = \{0, 1\}$ to be the binary alphabet. For a positive integer n and real numbers $\alpha > 0$ and $\delta \in (0, 1)$, let us say that a pair of mappings

$$f : \Sigma^n \to \Gamma^m$$
$$g : \Gamma^m \to \Sigma^n,$$

forms an (n, α, δ)-*compression scheme* for p if it holds that $m = \lfloor \alpha n \rfloor$ and

$$\Pr\left[g(f(a_1 \cdots a_n)) = a_1 \cdots a_n\right] > 1 - \delta, \tag{9.1}$$

where the probability is over random choices of $a_1, \ldots, a_n \in \Sigma$, each chosen independently according to the probability vector p.

To understand what a compression scheme means at an intuitive level, let us imagine the following situation between two people: Alice and Bob. Alice has a device of some sort with a button on it, and when she presses the button she gets an element of Σ, distributed according to p, independent of any prior outputs of the device. She presses the button n times, obtaining outcomes $a_1 \cdots a_n$, and she wants to communicate these outcomes to Bob using as few bits of communication as possible. So, what Alice does is to *compress* $a_1 \cdots a_n$ into a string of $m = \lfloor \alpha n \rfloor$ bits by computing $f(a_1 \cdots a_n)$. She sends the resulting bit-string $f(a_1 \cdots a_n)$ to Bob, who then *decompresses* by applying g, therefore obtaining $g(f(a_1 \cdots a_n))$. Naturally they hope that $g(f(a_1 \cdots a_n)) = a_1 \cdots a_n$, which means that Bob will have obtained the correct sequence $a_1 \cdots a_n$.

The quantity δ is a bound on the probability the compression scheme makes an error. We may view that the pair (f, g) *works correctly* for a string $a_1 \cdots a_n \in \Sigma^n$ if $g(f(a_1 \cdots a_n)) = a_1 \cdots a_n$, so the above equation (9.1) is equivalent to the condition that the pair (f, g) works correctly with high probability (assuming δ is small).

9.2.2 Statement of Shannon's source coding theorem

In the discussion above, the number α represents the average number of bits the compression scheme needs in order to represent each sample from the distribution described by p. It is obvious that compression schemes will exist for some numbers α and not others. The particular values of α for which it is possible to come up with a compression scheme are closely related to the Shannon entropy $H(p)$, as the following theorem establishes.

Theorem 9.2 (Shannon's source coding theorem)**.** *Let Σ be a finite, non-empty set, let $p \in \mathbb{R}^{\Sigma}$ be a probability vector, let $\alpha > 0$, and let $\delta \in (0,1)$. The following statements hold.*

1. *If $\alpha > H(p)$, then there exists an (n, α, δ)-compression scheme for p for all but finitely many choices of $n \in \mathbb{N}$.*
2. *If $\alpha < H(p)$, then there exists an (n, α, δ)-compression scheme for p for at most finitely many choices of $n \in \mathbb{N}$.*

It is not a mistake, by the way, that both statements hold for any fixed choice of $\delta \in (0,1)$, regardless of whether it is close to 0 or 1 (for instance). This will make sense when we see the proof.

It should be mentioned that the above statement of Shannon's source coding theorem is specific to the somewhat simplified (fixed-length) notion of compression that we have defined. It is more common, in fact, to consider variable-length compressions and to state Shannon's source coding theorem in terms of the average length of compressed strings. The reason why we restrict our attention to fixed-length compression schemes is that this sort of scheme will be more natural when we turn to the quantum setting.

9.2.3 Typical strings

Before we can prove the above theorem, we will need to develop the notion of a *typical string*. For a given probability vector $p \in \mathbb{R}^{\Sigma}$, positive integer n, and positive real number ε, we say that a string $a_1 \cdots a_n \in \Sigma^n$ is *ϵ-typical (with respect to p)* if

$$2^{-n(H(p)+\varepsilon)} < p(a_1) \cdots p(a_n) < 2^{-n(H(p)-\varepsilon)}.$$

We will need to refer to the set of all ε-typical strings of a given length repeatedly, so let us give this set a name:

$$T_{n,\varepsilon}(p) = \left\{ a_1 \cdots a_n \in \Sigma^n \,:\, 2^{-n(H(p)+\varepsilon)} < p(a_1) \cdots p(a_n) < 2^{-n(H(p)-\varepsilon)} \right\}.$$

When the probability vector p is understood from context we write $T_{n,\varepsilon}$ rather than $T_{n,\varepsilon}(p)$.

The following lemma establishes that a random selection of a string $a_1 \cdots a_n$ is very likely to be ε-typical as n gets large.

Lemma 9.3. *Let $p \in \mathbb{R}^{\Sigma}$ be a probability vector and let $\varepsilon > 0$. It holds that*

$$\lim_{n\to\infty} \sum_{a_1\cdots a_n \in T_{n,\varepsilon}(p)} p(a_1) \cdots p(a_n) = 1$$

Proof. Let $Y_1, \ldots, Y_n$ be independent and identically distributed random variables defined as follows: we choose $a \in \Sigma$ randomly according to the probability vector p, and then let the output value be the real number $-\log(p(a))$ for whichever value of a was selected. It holds that the expected value of each Y_j is

$$E[Y_j] = -\sum_{a\in\Sigma} p(a)\log(p(a)) = H(p).$$

The conclusion of the lemma may now be written

$$\lim_{n\to\infty} \Pr\left[\left| \frac{1}{n} \sum_{j=1}^{n} Y_j - H(p) \right| \geq \varepsilon \right] = 0,$$

which is true by the weak law of large numbers. □

Based on the previous lemma, it is straightforward to place upper and lower bounds on the number of ε-typical strings, as shown in the following lemma.

Lemma 9.4. *Let $p \in \mathbb{R}^{\Sigma}$ be a probability vector and let ε be a positive real number. For all but finitely many positive integers n it holds that*

$$(1-\varepsilon)2^{n(H(p)-\varepsilon)} < |T_{n,\varepsilon}| < 2^{n(H(p)+\varepsilon)}.$$

Proof. The upper bound holds for all n. Specifically, by the definition of ε-typical, we have

$$1 \geq \sum_{a_1 \cdots a_n \in T_{n,\varepsilon}} p(a_1) \cdots p(a_n) > 2^{-n(H(p)+\varepsilon)} \, |T_{n,\varepsilon}|,$$

and therefore $|T_{n,\varepsilon}| < 2^{n(H(p)+\varepsilon)}$.

For the lower bound, let us choose n_0 so that

$$\sum_{a_1 \cdots a_n \in T_{n,\varepsilon}} p(a_1) \cdots p(a_n) > 1 - \varepsilon$$

for all $n \geq n_0$, which is possible by Lemma 9.3. For all $n \geq n_0$ we have

$$1 - \varepsilon < \sum_{a_1 \cdots a_n \in T_{n,\varepsilon}} p(a_1) \cdots p(a_n) < |T_{n,\varepsilon}| \, 2^{-n(H(p)-\varepsilon)},$$

and therefore $|T_{n,\varepsilon}| > (1-\varepsilon)2^{n(H(p)-\varepsilon)}$, which completes the proof. □

9.2.4 Proof of Shannon's source coding theorem

We now have the necessary tools to prove Shannon's source coding theorem. Having developed some basic properties of typical strings, the proof is very simple: a good compression function is obtained by simply assigning a unique binary string to each typical string, with every other string mapped arbitrarily. On the other hand, any compression scheme that fails to account for a large fraction of the typical strings will be shown to fail with very high probability.

Proof of Theorem 9.2. First assume that $\alpha > H(p)$, and choose $\varepsilon > 0$ so that $\alpha > H(p) + 2\varepsilon$. For every choice of $n > 1/\varepsilon$ we therefore have that

$$m = \lfloor \alpha n \rfloor > n(H(p) + \varepsilon).$$

Now, because

$$|T_{n,\varepsilon}| < 2^{n(H(p)+\varepsilon)} < 2^m,$$

we may defined a function $f : \Sigma^n \to \Gamma^m$ that is 1-to-1 when restricted to $T_{n,\varepsilon}$, and we may define $g : \Gamma^m \to \Sigma^n$ appropriately so that $g(f(a_1 \cdots a_n)) = a_1 \cdots a_n$ for every $a_1 \cdots a_n \in T_{n,\varepsilon}$. As

$$\Pr\left[g(f(a_1 \cdots a_n)) = a_1 \cdots a_n\right] \geq \Pr[a_1 \cdots a_n \in T_{n,\varepsilon}] = \sum_{a_1 \cdots a_n \in T_{n,\varepsilon}} p(a_1) \cdots p(a_n),$$

we have that this quantity is greater than $1 - \delta$ for sufficiently large n.

Now let us prove the second item, where we assume $\alpha < H(p)$. It is clear from the definition of an (n, α, δ)-compression scheme that such a scheme can only work correctly for at most $2^{\lfloor \alpha n \rfloor}$

strings $a_1 \cdots a_n$. Let us suppose such a scheme is given for each n, and let $G_n \subseteq \Sigma^n$ be the collection of strings on which the appropriate scheme works correctly. If we can show that

$$\lim_{n\to\infty} \Pr[a_1 \cdots a_n \in G_n] = 0 \tag{9.2}$$

then we will be finished.

Toward this goal, let us note that for every n and ε, we have

$$\begin{aligned}\Pr[a_1 \cdots a_n \in G_n] &\leq \Pr[a_1 \cdots a_n \in G_n \cap T_{n,\varepsilon}] + \Pr[a_1 \cdots a_n \notin T_{n,\varepsilon}] \\ &\leq |G_n|\, 2^{-n(H(p)-\varepsilon)} + \Pr[a_1 \cdots a_n \notin T_{n,\varepsilon}].\end{aligned}$$

Choose $\varepsilon > 0$ so that $\alpha < H(p) - \varepsilon$. It follows that

$$\lim_{n\to\infty} |G_n|\, 2^{-n(H(p)-\varepsilon)} = 0.$$

As

$$\lim_{n\to\infty} \Pr[a_1 \cdots a_n \notin T_{n,\varepsilon}] = 0$$

by Lemma 9.3, we have (9.2) as required. □

9.3 Von Neumann entropy

Next we will discuss the *von Neumann entropy*, which may be viewed as a quantum information-theoretic analogue of the Shannon entropy. We will spend the next few lectures after this one discussing the properties of the von Neumann entropy as well as some of its uses—but for now let us just focus on the definition.

Let $\mathcal{X}$ be a complex Euclidean space, let $n = \dim(\mathcal{X})$, and let $\rho \in \mathrm{D}(\mathcal{X})$ be a density operator. The von Neumann entropy of ρ is defined as

$$S(\rho) = H(\lambda(\rho)),$$

where $\lambda(\rho) = (\lambda_1(\rho), \ldots, \lambda_n(\rho))$ is the vector of eigenvalues of ρ. An equivalent expression is

$$S(\rho) = -\operatorname{Tr}(\rho \log(\rho)),$$

where $\log(\rho)$ is the Hermitian operator that has exactly the same eigenvectors as ρ, and we take the base 2 logarithm of the corresponding eigenvalues. Technically speaking, $\log(\rho)$ is only defined for ρ positive definite, but $\rho \log(\rho)$ may be defined for all positive semidefinite ρ by interpreting $0 \log(0)$ as 0, just like in the definition of the Shannon entropy.

9.4 Quantum compression

There are some ways in which the von Neumann entropy is similar to the Shannon entropy and some ways in which it is very different. One way in which they are quite similar is in their relationships to notions of compression.

9.4.1 Informal discussion of quantum compression

To explain quantum compression, let us imagine a scenario between Alice and Bob that is similar to the classical scenario we discussed in relation to classical compression. We imagine that Alice has a collection of identical registers $\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_n$, whose associated complex Euclidean spaces are $\mathcal{X}_1 = \mathbb{C}^\Sigma, \ldots, \mathcal{X}_n = \mathbb{C}^\Sigma$ for some finite and nonempty set Σ. She wants to *compress* the contents of these registers into $m = \lfloor \alpha n \rfloor$ qubits, for some choice of $\alpha > 0$, and to send those qubits to Bob. Bob will then *decompress* the qubits to (hopefully) obtain registers $\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_n$ with little disturbance to their initial state.

It will not gnenerally be possible for Alice to do this without some assumption on the state of $(\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_n)$. Our assumption will be analogous to the classical case: we assume that the states of these registers are independent and described by some density operator $\rho \in \mathrm{D}(\mathcal{X})$ (as opposed to a probability vector $p \in \mathbb{R}^\Sigma$). That is, the state of the collection of registers will be assumed to be $\rho^{\otimes n} \in \mathrm{D}(\mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_n)$, where

$$\rho^{\otimes n} = \rho \otimes \cdots \otimes \rho \quad (n \text{ times}).$$

What we will show is that for large n, compression will be possible for $\alpha > S(\rho)$ and impossible for $\alpha < S(\rho)$.

To speak more precisely about what is meant by quantum compression and decompression, let us consider that $\alpha > 0$ has been fixed, let $m = \lfloor \alpha n \rfloor$, and let $\mathsf{Y}_1, \ldots, \mathsf{Y}_m$ be qubit registers, meaning that their associated spaces $\mathcal{Y}_1, \ldots, \mathcal{Y}_m$ are each equal to $\mathbb{C}^\Gamma$, for $\Gamma = \{0, 1\}$. Alice's compression mapping will be a channel

$$\Phi \in \mathrm{C}(\mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_n, \mathcal{Y}_1 \otimes \cdots \otimes \mathcal{Y}_m)$$

and Bob's decompression mapping will be a channel

$$\Psi \in \mathrm{C}(\mathcal{Y}_1 \otimes \cdots \otimes \mathcal{Y}_m, \mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_n).$$

Now, we need to be careful about how we measure the accuracy of quantum compression schemes. Our assumption on the state of $(\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_n)$ does not rule out the existence of other registers that these registers may be entangled or otherwise correlated with—so let us imagine that there exists another register Z, and that the initial state of $(\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_n, \mathsf{Z})$ is

$$\xi \in \mathrm{D}(\mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_n \otimes \mathcal{Z}).$$

When Alice compresses and Bob decompresses $\mathsf{X}_1, \ldots, \mathsf{X}_n$, the resulting state of $(\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_n, \mathsf{Z})$ is given by

$$\Big(\Psi\Phi \otimes \mathbb{1}_{\mathrm{L}(\mathcal{Z})}\Big)(\xi).$$

For the compression to be successful, we require that this density operator is close to ξ. This must in fact hold for all choices of Z and ξ, provided that the assumption $\mathrm{Tr}_{\mathcal{Z}}(\xi) = \rho^{\otimes n}$ is met. There is nothing unreasonable about this assumption—it is the natural quantum analogue to requiring that $g(f(a_1 \cdots a_n)) = a_1 \cdots a_n$ for classical compression.

It might seem complicated that we have to worry about all possible registers Z and all $\xi \in \mathrm{D}(\mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_n \otimes \mathcal{Z})$ that satisfy $\mathrm{Tr}_{\mathcal{Z}}(\xi) = \rho^{\otimes n}$, but in fact it will be simple if we make use of the notion of *channel fidelity*.

9.4.2 Quantum channel fidelity

Consider a channel $\Xi \in \mathrm{C}(\mathcal{W})$ for some complex Euclidean space $\mathcal{W}$, and let $\sigma \in \mathrm{D}(\mathcal{W})$ be a density operator on this space. We define the *channel fidelity* between Ξ and σ to be

$$\mathrm{F}_{\text{channel}}(\Xi, \sigma) = \inf\{\mathrm{F}(\xi, (\Xi \otimes \mathbb{1}_{\mathrm{L}(\mathcal{Z})})(\xi))\},$$

where the infimum is over all complex Euclidean spaces $\mathcal{Z}$ and all $\xi \in \mathrm{D}(\mathcal{W} \otimes \mathcal{Z})$ satisfying $\mathrm{Tr}_{\mathcal{Z}}(\xi) = \sigma$. The channel fidelity $\mathrm{F}_{\text{channel}}(\Xi, \sigma)$ places a lower bound on the fidelity of the input and output of a given channel Ξ provided that it acts on a part of a larger system whose state is σ when restricted to the part on which Ξ acts.

It is not difficult to prove that the infimum in the definition of the channel fidelity may be restricted to pure states $\xi = uu^*$, given that we could always purify a given ξ (possibly replacing $\mathcal{Z}$ with a larger space) and use the fact that the fidelity function is non-decreasing under partial tracing. With this in mind, consider any complex Euclidean space $\mathcal{Z}$, let $u \in \mathcal{W} \otimes \mathcal{Z}$ be any purification of σ, and consider the fidelity

$$\mathrm{F}\left(uu^*, \left(\Xi \otimes \mathbb{1}_{\mathrm{L}(\mathcal{Z})}\right)(uu^*)\right) = \sqrt{\left\langle uu^*, \left(\Xi \otimes \mathbb{1}_{\mathrm{L}(\mathcal{Z})}\right)(uu^*)\right\rangle}.$$

The purification $u \in \mathcal{W} \otimes \mathcal{Z}$ of σ must take the form

$$u = \mathrm{vec}\left(\sqrt{\sigma} B\right)$$

for some operator $B \in \mathrm{L}(\mathcal{Z}, \mathcal{W})$ satisfying $BB^* = \Pi_{\mathrm{im}(\sigma)}$. Assuming that

$$\Xi(X) = \sum_{j=1}^{k} A_j X A_j^*$$

is a Kraus representation of Ξ, it therefore holds that

$$\mathrm{F}\left(uu^*, \left(\Xi \otimes \mathbb{1}_{\mathrm{L}(\mathcal{Z})}\right)(uu^*)\right) = \sqrt{\sum_{j=1}^{k} \left|\left\langle \sqrt{\sigma} B, A_j \sqrt{\sigma} B \right\rangle\right|^2} = \sqrt{\sum_{j=1}^{k} \left|\left\langle \sigma, A_j \right\rangle\right|^2}.$$

So, it turns out that this quantity is independent of the particular purification of σ that was chosen, and we find that we could alternately have defined the channel fidelity of Ξ with σ as

$$\mathrm{F}_{\text{channel}}(\Xi, \sigma) = \sqrt{\sum_{j=1}^{k} \left|\left\langle \sigma, A_j \right\rangle\right|^2}.$$

9.4.3 Schumacher's quantum source coding theorem

We now have the required tools to establish the relationship between the von Neumann entropy and quantum compression that was discussed earlier in the lecture. Using the same notation that was introduced above, let us say that a pair of channels

$$\Phi \in \mathrm{C}(\mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_n, \mathcal{Y}_1 \otimes \cdots \otimes \mathcal{Y}_m),$$
$$\Psi \in \mathrm{C}(\mathcal{Y}_1 \otimes \cdots \otimes \mathcal{Y}_m, \mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_n)$$

is an (n, α, δ)-*quantum compression scheme* for $\rho \in \mathrm{D}(\mathcal{X})$ if $m = \lfloor \alpha n \rfloor$ and

$$\mathrm{F}_{\text{channel}}\left(\Psi\Phi, \rho^{\otimes n}\right) > 1 - \delta.$$

The following theorem, which is the quantum analogue to Shannon's source coding theorem, establishes conditions on α for which quantum compression is possible and impossible.

Theorem 9.5 (Schumacher)**.** *Let $\rho \in \mathrm{D}(\mathcal{X})$ be a density operator, let $\alpha > 0$ and let $\delta \in (0, 1)$. The following statements hold.*

1. *If $\alpha > S(\rho)$, then there exists an (n, α, δ)-quantum compression scheme for ρ for all but finitely many choices of $n \in \mathbb{N}$.*
2. *If $\alpha < S(\rho)$, then there exists an (n, α, δ)-quantum compression scheme for ρ for at most finitely many choices of $n \in \mathbb{N}$.*

Proof. Assume first that $\alpha > S(\rho)$. We begin by defining a quantum analogue of the set of typical strings, which is the *typical subspace*. This notion is based on a spectral decomposition

$$\rho = \sum_{a \in \Sigma} p(a) u_a u_a^*.$$

As p is a probability vector, we may consider for each $n \geq 1$ the set of ε-typical strings $T_{n,\varepsilon} \subseteq \Sigma^n$ for this distribution. In particular, we form the projection onto the *typical subspace*:

$$\Pi_{n,\varepsilon} = \sum_{a_1 \cdots a_n \in T_{n,\varepsilon}} u_{a_1} u_{a_1}^* \otimes \cdots \otimes u_{a_n} u_{a_n}^*.$$

Notice that

$$\left\langle \Pi_{n,\varepsilon}, \rho^{\otimes n} \right\rangle = \sum_{a_1 \cdots a_n \in T_{n,\varepsilon}} p(a_1) \cdots p(a_n),$$

and therefore

$$\lim_{n \to \infty} \left\langle \Pi_{n,\varepsilon}, \rho^{\otimes m} \right\rangle = 1,$$

for every choice of $\varepsilon > 0$.

We can now move on to describing a sequence of compression schemes that will suffice to prove the theorem, provided that $\alpha > S(\rho) = H(p)$. By Shannon's source coding theorem (or, to be more precise, our proof of that theorem) we may assume, for sufficiently large n, that we have a classical (n, α, ε)-compression scheme (f, g) for p that satisfies

$$g(f(a_1 \cdots a_n)) = a_1 \cdots a_n$$

for all $a_1 \cdots a_n \in T_{n,\varepsilon}$. Define a linear operator

$$A \in \mathrm{L}\left(\mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_n, \mathcal{Y}_1 \otimes \cdots \otimes \mathcal{Y}_m\right)$$

as

$$A = \sum_{a_1 \cdots a_n \in T_{n,\varepsilon}} e_{f(a_1 \cdots a_n)} (u_{a_1} \otimes \cdots \otimes u_{a_n})^*.$$

for each $a_1 \cdots a_n \in \Sigma^n$. Notice that

$$A^* A = \Pi_{n,\varepsilon}.$$

Now, the mapping defined by $X \mapsto AXA^*$ is completely positive but generally not trace-preserving. However, it is a *sub-channel*, by which it is meant that there must exist a completely positive mapping Ξ for which

$$\Phi(X) = AXA^* + \Xi(X) \tag{9.3}$$

is a channel. For instance, we may take

$$\Xi(X) = \langle \mathbb{1} - \Pi_{n,\varepsilon}, X \rangle \, \sigma$$

for some arbitrary choice of $\sigma \in \mathrm{D}(\mathcal{Y}_1 \otimes \cdots \otimes \mathcal{Y}_m)$. Likewise, the mapping $Y \mapsto A^*YA$ is also a sub-channel, meaning that there must exist a completely positive map Δ for which

$$\Psi(Y) = A^*YA + \Delta(Y) \tag{9.4}$$

is a channel.

It remains to argue that, for sufficiently large n, that the pair (Φ, Ψ) is an (n, α, δ)-quantum compression scheme for any constant $\delta > 0$. From the above expressions (9.3) and (9.4) it is clear that there exists a Kraus representation of $\Psi\Phi$ having the form

$$(\Psi\Phi)(X) = (A^*A)X(A^*A)^* + \sum_{j=1}^{k} B_j X B_j^*$$

for some collection of operators $B_1, \ldots, B_k$ that we do not really care about. It follows that

$$\mathrm{F}_{\text{channel}}(\Psi\Phi, \rho^{\otimes n}) \geq \left| \left\langle \rho^{\otimes n}, A^*A \right\rangle \right| = \left\langle \rho^{\otimes n}, \Pi_{n,\varepsilon} \right\rangle.$$

This quantity approaches 1 in the limit, as we have observed, and therefore for sufficiently large n it must hold that (Φ, Ψ) is an (n, α, δ) quantum compression scheme.

Now consider the case where $\alpha < S(\rho)$. Note that if $\Pi_n \in \mathrm{Pos}(\mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_n)$ is a projection with rank at most $2^{n(S(\rho)-\varepsilon)}$ for each $n \geq 1$, then

$$\lim_{n \to \infty} \left\langle \Pi_n, \rho^{\otimes n} \right\rangle = 0. \tag{9.5}$$

This is because, for any positive semidefinite operator P, the maximum value of $\langle \Pi, P \rangle$ over all choices of orthogonal projections Π with $\mathrm{rank}(\Pi) \leq r$ is precisely the sum of the r largest eigenvalues of P. The eigenvalues of $\rho^{\otimes n}$ are the values $p(a_1) \cdots p(a_n)$ over all choices of $a_1 \cdots a_n \in \Sigma^n$, so for each n we have

$$\left\langle \Pi_n, \rho^{\otimes n} \right\rangle \leq \sum_{a_1 \cdots a_n \in G_n} p(a_1) \cdots p(a_n)$$

for some set G_n of size at most $2^{n(S(\rho)-\varepsilon)}$. At this point the equation (9.5) follows by similar reasoning to the proof of Theorem 9.2.

Now let us suppose, for each $n \geq 1$ and for $m = \lfloor \alpha n \rfloor$, that

$$\Phi_n \in \mathrm{C}(\mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_n, \mathcal{Y}_1 \otimes \cdots \otimes \mathcal{Y}_m),$$
$$\Psi_n \in \mathrm{C}(\mathcal{Y}_1 \otimes \cdots \otimes \mathcal{Y}_m, \mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_n)$$

are channels. Our goal is to prove that (Φ_n, Ψ_n) fails as a quantum compression scheme for all sufficiently large values of n.

Fix $n \geq 1$, and consider Kraus representations

$$\Phi_n(X) = \sum_{j=1}^{k} A_j X A_j^* \qquad \text{and} \qquad \Psi_n(X) = \sum_{j=1}^{k} B_j X B_j^*,$$

where

$$A_1, \ldots, A_k \in \mathrm{L}\left(\mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_n, \mathcal{Y}_1 \otimes \cdots \otimes \mathcal{Y}_m\right),$$
$$B_1, \ldots, B_k \in \mathrm{L}\left(\mathcal{Y}_1 \otimes \cdots \otimes \mathcal{Y}_m, \mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_n\right),$$

and where the assumption that they have the same number of terms is easily made without loss of generality. Let Π_j be the projection onto the range of B_j for each $j = 1, \ldots k$, and note that it obviously holds that

$$\operatorname{rank}(\Pi_j) \leq \dim(\mathcal{Y}_1 \otimes \cdots \otimes \mathcal{Y}_m) = 2^m.$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\mathrm{F}_{\text{channel}}(\Psi_n \Phi_n, \rho^{\otimes n})^2 &= \sum_{i,j} \left| \left\langle \rho^{\otimes n}, B_j A_i \right\rangle \right|^2 \\
&= \sum_{i,j} \left| \left\langle \Pi_j \sqrt{\rho^{\otimes n}}, B_j A_i \sqrt{\rho^{\otimes n}} \right\rangle \right|^2 \\
&\leq \sum_{i,j} \left\langle \Pi_j, \rho^{\otimes n} \right\rangle \left(\operatorname{Tr} B_j A_i \rho^{\otimes n} A_i^* B_j^* \right).
\end{aligned}$$

As

$$\operatorname{Tr}\left(B_j A_i \rho^{\otimes n} A_i^* B_j^* \right) \geq 0$$

for each i, j, and

$$\sum_{i,j} \operatorname{Tr}\left(B_j A_i \rho^{\otimes n} A_i^* B_j^* \right) = \operatorname{Tr}(\Psi\Phi(\rho^{\otimes n})) = 1,$$

it follows that

$$\mathrm{F}_{\text{channel}}(\Psi_n \Phi_n, \rho^{\otimes n})^2 \in \operatorname{conv}\left(\left\{ \left\langle \Pi_j, \rho^{\otimes n} \right\rangle \,:\, j = 1, \ldots, k \right\} \right).$$

As each Π_j has rank at most 2^m, it follows that

$$\lim_{n \to \infty} \mathrm{F}_{\text{channel}}(\Psi_n \Phi_n, \rho^{\otimes n}) = 0.$$

So, for all but finitely many choices of n, the pair (Φ_n, Ψ_n) fails to be an (n, α, δ) quantum compression scheme. □