Entropy and Compression

In his 1948 paper, Shannon related the entropy *S* of a message source to the number of "typical" messages that source could send. Recall the source is modeled in terms of independent and identically distributed messages from an alphabet $\{1, ..., m\}$, where each message sends k with probability p_k .

For a large number of messages $N \gg 1$, each letter k is sent approximately $p_k N$ times. As we counted previously, the number of strings with those letter frequencies is:

$$\frac{N!}{\prod_{k=1}^{m}(p_k N)!} \sim \frac{N^N}{\prod_{k=1}^{m}(p_k N)^{p_k N}} = 2^{NS} \quad , \quad S = -\sum_{k=1}^{m} p_k \log p_k$$

We described Huffman codes, which substitute each letter with a variable-length bit string ("codeword") based on the letter frequencies, in order achieve a compression rate set by the Shannon entropy.

Shannon's original idea was conceptually simpler and easy to fully analyze, though less explicit and more dependent on the asymptotic limit. The idea is called block coding and it is based on typical sequences.

If you only cared to send typical messages then how would you compress the information?

Entropy and Compression

Block coding is based on the "strong law of large numbers", which expresses the fact that estimators based on samples from a probability distribution will eventually converge to their expectation values.

Let $x_1, ..., x_N$ be iid events (observations) with probability distribution *p ("sampled from p").* Let f be a function defined on this space of events (an observable), and define the estimator

$$\bar{f} = \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

(Strong) law of large numbers: for all $\epsilon, \delta > 0$ there exists an N_0 such that

$$\Pr\left[|\bar{f} - \langle f \rangle| \leq \delta\right] \geq 1 - \epsilon$$

For all $N \geq N_0$.

Note that the LLN does not tell us anything about the rate of the convergence with the number of samples, only that convergence is eventually guaranteed.

Entropy and Compression

We can apply the LLN to the random variable $f(x) = \log \frac{1}{p(x)}$

In this case the true expectation value is the entropy of the source $\langle f \rangle = S_X$, while the estimator is

$$\bar{f} = \frac{1}{N} \sum_{i=1}^{N} f(x_i) = -\frac{1}{N} \log p(x_1, ..., x_N)$$

Therefore the LLN implies that for all $\delta, \epsilon > 0$ we can choose N such that

$$S_X - \delta \leq -\frac{1}{N} \log p(x_1, ..., x_N) \leq S_X + \delta$$

With probability at least $1 - \epsilon$. We say a sequence satisfying the above is "δ-typical." Therefore if $\mathbf{x}$ is a length N message that is δ- typical then the probability of this message satisfies

$$p_{\min} = 2^{-N(S_X + \delta)} \leq p(\mathbf{x}) \leq 2^{-N(S_X - \delta)} = p_{\max}$$

Entropy and Compression

For given values of these parameters, the number $T = T(N, \delta, \epsilon)$ of typical sequences must satisfy

$$T \cdot p_{\min} \leq 1 \quad , \quad T \cdot p_{\max} \geq 1 - \epsilon$$

Therefore

$$(1 - \epsilon)2^{N(S_X - \delta)} \leq T(N, \epsilon, \delta) \leq 2^{N(S_X + \delta)}$$

From the upper bound, we can see that a block code with $N(S_X + \delta)$ bits suffice to encode all of these typical messages (simply assign a distinct positive integer to each typical message).

Since every typical message is assigned a distinct bit string on $N(S_X + \delta)$ bits, the probability of a successful transmission (encoding, decoding) is 1 as long as the message is typical. If the message is atypical then it may not be decoded correctly, but this only ever happens with probability at most ϵ.

Entropy and Compression

To see that the entropy characterizes the compression rate of the optimal asymptotic block coding scheme, suppose we try to compress the message to only

$$S_X - \delta'$$

Bits per letter, where δ' is a constant independent of N. Using this number of bits we can uniquely encode

$$2^{N(S_X - \delta')}$$

distinct typical messages. Perhaps the hope is that these messages are somehow "the most typical" and we'll get away with only encoding these. But the probability of each typical message is at most

$$p_{\max} = 2^{-N(S_X - \delta)}$$

And so the probability of success is at most

$$p_{\text{success}} \leq 2^{-N(S_X - \delta')} 2^{-N(S_X - \delta)} = 2^{-N(\delta' - \delta)}$$

Which for constant δ' will eventually become arbitrarily small at large N.

Entropy and Compression

More formally, a compression scheme with rate R maps the string $\mathbf{x} = (x_1, ..., x_n)$ to a string on $\lfloor nR \rfloor$ bits, Denoted $C(\mathbf{x})$. The corresponding decompression string maps back from the $\lfloor nR \rfloor$ bits to the original alphabet. The compression-decompression scheme (C,D) is *reliable* if

$$\lim_{n \to \infty} \Pr\left[D(C(\mathbf{x})) = \mathbf{x}\right] = 1$$

Shannon's source-coding theorem: suppose $\{x_i\}$ is an iid information source with entropy S. If R > S then there exists a reliable compression scheme of rate R for the source. If R < S then there can be no reliable compression scheme with rate R for the source.

A compression rate of $S(X) + o(1)$ is *achievable,* and a compression rate of $S(X) - \Omega(1)$ is not achievable.

Entropy and Compression

Benjamin Schumacher, who is co-credited with coining the term “qubit”, generalized Shannon’s noiseless coding theorem to the quantum setting in 1995.

Consider an iid quantum information source that sends various pure states with various probabilities.

$$\{|\psi_1\rangle, ..., |\psi_k\rangle\} \qquad \{p_1, ..., p_k\}$$

We can model this by saying that each use of the channel sends the density matrix:

$$\rho = \sum_{i=1}^{k} p_k |\psi_k\rangle\langle\psi_k|$$

And so after n uses of the channel, we have sent the state $\rho^{\otimes n} = \rho \otimes \rho \otimes ... \otimes \rho$

Entropy and Compression

How much redundancy is contained in the message $\rho^{\otimes n} = \rho \otimes \rho \otimes \cdots \otimes \rho$? How much can it be compressed?

Schumacher's answer is that there is a reliable compression-decompression scheme which compresses the state $\rho^{\otimes n}$ down to a state in a Hilbert space of dimension

$$\dim \mathcal{H} = 2^{n(S(\rho)+o(1))}$$

And that no such reliable scheme exists when $\dim \mathcal{H} \leq 2^{n(S(\rho)-\Omega(1))}$.

The proof of Schumacher's quantum source-coding theorem closely parallel's Shannon's, replacing the notion of "typical sequences" with that of "typical subspaces." Therefore the most difficult part of this result from a historical perspective was the conceptual leap required to consider quantum states as information.

Entropy and Compression

Before moving to the general proof, we will do a small example to gain intuition for relating von Neumann entropy to compression.

Suppose we have a channel that sends $|0\rangle, |+\rangle$ with equal probability, and so it sends the density matrix

$$\rho = \frac{1}{2}\left(|0\rangle\langle 0| + |+\rangle\langle +|\right)$$

Classically, if we have an alphabet of two symbols that are each sent with equal probability, than the source outputs 1 bit of entropy per letter and no reliable compression is possible.

But the key point for this example is that $|0\rangle, |+\rangle$ are nonorthogonal states. Intuitively, the nonzero overlap between these vectors creates a new kind of quantum redundancy that has no classical analog.

Diagonalizing the density matrix turns it into a probability distribution over a set of orthogonal quantum states, and the von Neumann entropy is the Shannon entropy of this distribution. This is the main idea which relates Schumacher's proof to Shannon's.

Entropy and Compression

In our example, the eigenstates and eigenvalues of ρ are

$$|0'\rangle = \cos\left(\frac{\pi}{8}\right)|0\rangle + \sin\left(\frac{\pi}{8}\right)|1\rangle \qquad \lambda_0 = \cos^2\left(\frac{\pi}{8}\right) = 0.8535$$

$$|1'\rangle = \sin\left(\frac{\pi}{8}\right)|0\rangle - \cos\left(\frac{\pi}{8}\right)|1\rangle \qquad \lambda_1 = \sin^2\left(\frac{\pi}{8}\right) = 0.1465$$

In this diagonal form, we see that there was quite a lot of redundancy in this channel after all, because it sends the state $|0'\rangle$ over 85% of the time.

The idea of Schumacher coding is to exploit this redundancy in terms of typical sequences. In this example, after three uses of the channel the typical sequences could be chosen to be

$$|0'0'0'\rangle, |0'0'1'\rangle, |0'1'0'\rangle, |1'0'0'\rangle$$

An encoding of basis states also suffices to encode ever state in the “typical subspace” spanned by these states.

Entropy and Compression

More generally, consider a quantum channel sending quantum letters with the following diagonal form:

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

Therefore, it also makes sense to speak of a δ - typical sequence $x_1, ..., x_n$ which satisfies,

$$\left| \frac{1}{n} \log \left(\frac{1}{p(x_1, ..., x_n)} \right) - S(\rho) \right| \leq \delta$$

Now the δ - typical subspace is define as the span of all δ - typical sequences of quantum states. The projector onto this typical subspace is

$$P(n, \delta) = \sum_{(x_1, ..., x_n): \delta\text{-typical}} |x_1\rangle\langle x_1| \otimes |x_1\rangle\langle x_1| \otimes ... \otimes |x_n\rangle\langle x_n|$$

Entropy and Compression

Repeating Shannon's application of the large of large numbers also gives us

$$\mathrm{tr}\left(P(n,\delta)\rho^{\otimes n}\right) \geq 1-\epsilon$$

For sufficiently large n. Similarly, the dimension of the typical subspace is now

$$(1-\epsilon)2^{n(S(\rho)-\delta)} \leq \dim \mathcal{H} \leq 2^{n(S(\rho)+\delta)}$$

The strategy for Schumacher coding again parallel's Shannon's block coding scheme. Each typical sequence is assigned a codeword (e.g. a computational basis state) on $n(S(\rho)+\delta)$ qubits.

Asymptotically, the probability for an atypical sequence is vanishingly small, so it won't matter what coding strategy is applied to atypical sequences.

Entropy and Compression

More specifically, starting from the message $|\phi(x)\rangle = |\phi(x_1)\rangle \otimes ... \otimes |\phi(x_n)\rangle$, the first step applies a quantum channel to measure whether it is typical or atypical (without revealing any other information),

$$|\phi(x)\rangle\langle\phi(x)| \mapsto \rho(x) = P(n,\delta)|\phi(x)\rangle\langle\phi(x)|P(n,\delta) + \rho_{\text{junk}}(x)\langle\phi(x)|(1-P(n,\delta)|\phi(x)\rangle$$

If the state is typical, Alice will perform a unitary scheme to compress it

$$U|\psi_{\text{typ}}\rangle = |\psi_{\text{compressed}}\rangle \otimes |0_{\text{rest}}\rangle$$

Alice can then send this state $|\psi_{\text{compressed}}\rangle$ to Bob, who can append $|0_{\text{rest}}\rangle$ onto it and apply U^{-1} to recover the original message. If the sequence is atypical he receives a junk state, but this happens with arbitrarily low probability and so his reconstructed state is close in trace distance to the Alice's original message.

Entropy and Compression

Conversely, if we tried to compress beyond the limit set by the von Neumann entropy than the scheme will not be reliable. To see this, note that regardless of Bob's unitary decoding strategy, the decoded states must be in a Hilbert space of dimension

$$\dim \mathcal{H}' \leq 2^{n(S(\rho)-\delta')}$$

Let P' be the projector onto this subspace. Since $\rho \succeq 0$, Fan's dominance principle in matrix analysis implies:

$$\mathrm{tr}\left(\rho^{\otimes n} P'\right) \leq \lambda_{\max}(\rho^{\otimes n}) \cdot \mathrm{rank}(P')$$

Where $\lambda_{\max}(\rho^{\otimes n})$ is the largest eigenvalue of $\rho^{\otimes n}$. But the δ- typical eigenvalues are no smaller than $2^{-n(S(\rho)-\delta)}$, and so the overlap of the original message with Bob's subspace of limited dimension is:

$$\mathrm{tr}\left(\rho^{\otimes n} P'\right) \leq 2^{n(S(\rho)-\delta')} \cdot 2^{-n(S(\rho)-\delta)} = 2^{-n(\delta'-\delta)}$$

And so the probability for Bob to recover the original message becomes arbitrarily small.

Entropy and Compression

Schumacher's noiseless channel coding theorem: Let $\{\rho\}$ be an iid quantum information source with von Neumann entropy S per letter. If R > S, there exists a reliable compression-decompression scheme of rate R for the source $\{\rho\}$. If R < S, then no such scheme with rate R can be reliable.

Schumacher's compression scheme gives an operational meaning to the von Neumann entropy, and expresses the ultimate limits for the compressibility of quantum communication.

We now understand the limits on classical compression of classical information, and quantum compression of quantum information. But what about quantum compression of classical information?

This question was addressed by Holevo in 1973, "Bounds for the Quantity of Information Transmitted by a Quantum Communication Channel." Informally, Holevo's theorem states that **you cannot reliably store n bits in fewer than n qubits.**

Entropy and Compression

Suppose Alice wants to encode a random variable x that takes values in $\{1, ..., n\}$. When she sees the value i , she records this outcome and produces density matrix ρ_X^i , thereby generating the state

$$\rho_{CX} = \sum_i p_i |i\rangle\langle i| \otimes \rho_X^i$$

The subscripts C, X reflect the bipartition of the system into a classical part and a quantum part. Density matrices of this form are sometimes called classical-quantum states.

The goal will be for Alice to generate a state of this form and send the marginal ρ_X to Bob so that he can perform some measurements to extract the classical information.

We considered states of this form previously in order to show the concavity of von Neumann entropy based on the nonnegativity of the quantum mutual information. In particular, we computed

$$S(\rho_{CX}) = S(p_i) + \sum p_i S(\rho_X^i)$$

Entropy and Compression

(recall that we computed this by using the fact that $\langle i|j\rangle = \delta_{ij}$). Therefore the mutual information is

$$I(C:X) = S(\rho_C) + S(\rho_X) - S(\rho_{CX})$$

$$= S(\rho_X) - \sum_i p_i S(\rho_X^i)$$

This mutual information between the classical and quantum parts of a CQ state is called the Holevo quantity

$$\chi = I(C:X) = S(\rho_X) - \sum_i p_i S(\rho_X^i)$$

Suppose that the classical messages are on n bits, while the quantum messages are written with $k \leq n$ qubits. Therefore $S(\rho_X) \leq k$, and so

$$I(C:X) \leq k$$

Furthermore, by the data processing inequality there is no further post-processing that Bob could do that would raise this mutual information.

Entropy and Compression

This mutual information between the classical and quantum parts of a CQ state is called the Holevo quantity

$$\chi = I(C:X) = S(\rho_X) - \sum_i p_i S(\rho_X^i)$$

Suppose that the classical messages are on n bits, while the quantum messages are written with $k \leq n$ qubits. Therefore $S(\rho_X) \leq k$, and so

$$I(C:X) \leq k$$

Therefore, at most k bits of information about Alice's notebook are encoded in the state that Bob receives. Furthermore, by the data processing inequality there is no further post-processing that Bob could do that would raise this mutual information (this is the historical reason that Holevo's result followed closely after the proof of strong subadditivity).