Entropy, Compression & Entanglement concentration

Shannon '48:

1. Noiseless coding theorem:
   How much can a message be compressed (asymptotically)?
   What is its information content?

2. Noisy channel coding theorem:
   At what rate can we communicate reliably over a noisy channel?

He solved both problems and today the applications are ubiquitous.
Quantumly, only the first problem is fully solved, so we will focus there.

1. Shannon entropy $H$:

   Definition: For a probability distribution $\mathbf{p} \in \mathbb{R}^k$, the entropy of $\mathbf{p}$ is defined as
   $$H(\mathbf{p}) = -\sum_{i=1}^{k} p_i \log p_i$$

   Examples:
   - $H(1,0,0,\ldots,0) = 0$
   - $H(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}) = \log_k k$
   - In general, $0 \leq H(\mathbf{p}) \leq \log_k k$
   - $H(\mathbf{p} \oplus \mathbf{q}) = H(\mathbf{p}) + H(\mathbf{q})$
   - In general, for a joint distribution $\mathbf{p}_{AB} \in \mathbb{R}^k \otimes \mathbb{R}^k$, $\max \mathbb{E} H(\mathbf{p}_A), H(\mathbf{p}_B) \leq H(\mathbf{p}_{AB}) \leq H(\mathbf{p}_A) + H(\mathbf{p}_B)$

2. Shannon entropy and classical data compression

   Stirling's approximation:
   $$\lim_{n \to \infty} \sqrt[n]{ \frac{n!}{\sqrt{2\pi n} \left( \frac{n}{e} \right)^n} } = 1$$

   $$\Rightarrow \ln(n!) = \ln \left( \frac{n}{e} \right)^n + O(\ln n) = n \ln n - n + O(\ln n)$$

   Corollary: For $n \in \mathbb{N}$,
   $$\log \left( \frac{n}{m} \right) = \log \frac{n!}{(m!)^{(\lfloor m/n \rfloor)!}}$$
   $$\approx n \log n - (\log e)n - \left[ (m) \log (m) + (\log e)n + (\log m) \log (\log e) \right] = n \cdot H(\mathbf{p})$$
More generally,

**Corollary:** For $\tilde{p}$ a distribution on $k$ elements,

$$\log \frac{n!}{\tilde{T}(\tilde{p}, n)!} \approx n \cdot H(\tilde{p}).$$

By the law of large numbers, if we draw $n$ independent samples from $\tilde{p}$ (i.e., one sample from $\tilde{p}^n$), with probability exponentially close to 1, every outcome $i$ will be observed about $\frac{n}{\tilde{T}(\tilde{p}, n)}$ times.

The number of strings in $\{1, 2, \ldots, k\}^n$ with $p_i$ in $i$'s is

$$\frac{n!}{\tilde{T}(\tilde{p}, n)!}.$$ These "typical" strings can be compressed using $n \cdot H(\tilde{p})$ bits of information.

More precisely, (for all $\varepsilon > 0$ and $n$ large enough)

**Noisy coding theorem:** There exists a set $C \subseteq \{1, \ldots, k\}^n$ of size $|C| \leq 2^{n(H(\tilde{p}) + \varepsilon)}$ such that

$$\Pr[X \in C] \geq 1 - \varepsilon$$

when sample $X$ is drawn from the distribution $\tilde{p}^n$.

**Proof:**

$$\Pr[X, x_2, \ldots, x_n] = p(x_1) p(x_2) \cdots p(x_n)$$

$$\implies \log \Pr[X, \ldots, x_n] = \sum_{x_1} \log p(x_1)$$

is a sum of independent r.v.s with finite mean and variance

$$\log \Pr[X] = \sum \log p(x_i)$$

by the Central Limit Theorem, this sum concentrates to its expectation, $n \cdot \sum p_i \log p_i = -n H(\tilde{p})$. We may say

$$\Pr[-\frac{1}{n} \log \Pr[X] - H(\tilde{p}) \leq \varepsilon] \geq 1 - \varepsilon$$

Let $C = \{x : \Pr[X] \in [2^{-n(H+\varepsilon)}, 2^{-n(H-\varepsilon)}]\}$, and

$$\Pr[X \in C] \geq 1 - \varepsilon$$

and, necessarily, $|C| \leq 2^{n(H+\varepsilon)}$ since the total probability is at most one.

Moreover, this construction is asymptotically optimal, since with fewer than $n(H-\varepsilon)$ bits we could not cover all typical sequences, i.e., $\forall \delta > 0$,

$$\Pr[\text{successful decoding}] \leq 2^{n(H-\varepsilon)} 2^{-n(H-\varepsilon)} + \varepsilon = 2^{n(H-\varepsilon)} + \varepsilon.$$
3. Von Neumann entropy $S$

**Definition:** For a quantum state $\rho$, that can be diagonalized as $\sum_i p_i |i\rangle \langle i|$, with $\langle i | j \rangle = \delta_{ij}$, the Von Neumann entropy of $\rho$ is

$$S(\rho) = H(\rho) = -\text{Tr}(\rho \log \rho).$$

**Properties:**

- a. $S(14\times41) = 0$
- b. $S(W_{\rho} W^*) = S(\rho)$
- c. $S(\rho) \leq \log D$, with equality only for the maximally mixed state $\rho = \frac{1}{D} I$.
- d. Concavity: for $\lambda \in [0, 1]$, $S(\lambda \rho_1 + (1-\lambda) \rho_2) \geq \lambda S(\rho_1) + (1-\lambda) S(\rho_2)$.

**Interpretation:** If we are more ignorant of how a state was prepared, then the entropy is higher.

- e. Subadditivity: $S(\rho_{AB}) \leq S(\rho_A) + S(\rho_B)$.
- f. Strongly subadditive:
  $$S(\rho_{ABC}) + S(\rho_{A}) \leq S(\rho_{AB}) + S(\rho_{BC}).$$
  (difficult to prove, but extremely useful)

- g. **Δ inequality:**
  $$S(\rho_A) > |S(\rho_A) - S(\rho_{AB})|.$$  
  Unlike for the Shannon entropy, it is not true that $S(\rho_{AB}) = \max \{S(\rho_A), S(\rho_B)\}$.

  **Example:** If $\rho_{AB} = 14\times41$ is a pure state, then $S(\rho_{AB}) = 0$, but $S(\rho_A) = S(\rho_B)$ may be $>0$.

  by Schmidt decomposition

  Unlike classically, discarding a subsystem can increase your uncertainty about the state.

**More operational properties:**

- h. If $Y$ is the result of any full measurement of $\rho$, ie.
  a classical random variable with $P(Y = i) = p_i$, then $H(Y) \geq S(\rho)$, with equality iff the measurement basis diagonalizes $\rho$.
- i. Preparing the ensemble $\{p_x, \{X, Y, Z\}\}$, $H(X) \geq S(\sum_x p_x |x\rangle \langle x|)$