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 $\vec{p} \in \mathbb{R}^k$, the entropy of $\vec{p}$ is defined as

$$H(\vec{p}) = -\sum_{i=1}^{k} p_i \log p_i$$

Examples:

- $H((1,0,0,\dots,0)) = 0$
- $H(\frac{1}{k}, \frac{1}{k}, \dots, \frac{1}{k}) = \log_2 k$
- in general $0 \le H(\vec{p}) \le \log_2 k$
- $H(\vec{p} \otimes \vec{q}) = H(\vec{p}) + H(\vec{q})$
- in general, for a joint distribution $p_{AB} \in \mathbb{R}^k \otimes \mathbb{R}^\ell$,
  $$\max\{H(p_A), H(p_B)\} \le H(p_{AB}) \le H(p_A) + H(p_B)$$

2. Shannon entropy and classical data compression

Stirling's approximation:

$$\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n} = 1$$

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

Corollary: For $p \in [0,1]$,

$$\log \binom{n}{pn} = \log \frac{n!}{(pn)!\,[(1-p)n]!}$$
$$\approx n\log n - (\log_2 e)n - \left[(pn)\log(pn) - (\log_2 e)pn + (1-p)n\log((1-p)n) - \log_2(e)(1-p)n\right]$$
$$= n \cdot H(p).$$

→

More generally,

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

$$\log \frac{n!}{\prod_{i=1}^{k} (p_i \cdot n)!} \approx n \cdot H(\vec{p}).$$

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

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

$$\frac{n!}{\prod_{i=1}^{k} (p_i \cdot n)!}.$$

These "typical" strings can be compressed using $n \cdot H(\vec{p})$ bits of information.

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

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

$$\mathbb{P}[X \in C] \geq 1 - \varepsilon$$

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

Proof:

$$\mathbb{P}[x_1, x_2, \ldots x_n] = p(x_1)\, p(x_2) \cdots p(x_n)$$

$$\therefore \log \mathbb{P}[x_1, \ldots x_n] = \sum_{i=1} \log p(x_i)$$

$$\therefore \log \mathbb{P}[X] = \sum_i \log p(X_i)$$

(sum of independent r.v.s with finite mean & variance)

By the Central Limit Theorem, this sum concentrates to its expectation, $n \cdot \sum_{j=1}^{k} p_j \log p_j = -n H(\vec{p})$. We may say

$$\mathbb{P}\left[\left| -\frac{1}{n} \log \mathbb{P}[X] - H(\vec{p}) \right| \leq \delta\right] \geq 1 - \varepsilon$$

$\therefore$ Let $C = \{x : \mathbb{P}[x] \in [2^{-n(H+\delta)}, 2^{-n(H-\delta)}]\}$. Then

$$\mathbb{P}[X \in C] \geq 1 - \varepsilon$$

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

Moreover, this construction is asymptotically optimal, since with fewer than $n(H-\delta)$ bits we could not cover all typical sequences, if $\delta' > \delta$,

$$\mathbb{P}[\text{successful decoding}] \leq 2^{n(H-\delta')} 2^{-n(H-\delta)} + \varepsilon = 2^{n(\delta'-\delta)} + \varepsilon.$$

3. Von Neumann entropy S

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

$$S(\rho) = H(\vec{p}) = -\mathrm{Tr}(\rho \log \rho).$$

Properties:

a. $S(|\psi\rangle\langle\psi|) = 0$

b. $S(U\rho U^\dagger) = S(\rho)$

c. $S(\rho) \leq \log D$, with equality only for the maximally mixed state $\rho = \frac{1}{D}\mathbb{1}$.

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. Subadditive: $S(\rho_{AB}) \leq S(\rho_A) + S(\rho_B)$.

f. Strongly subadditive:

$$S(\rho_{ABC}) + S(\rho_B) \leq S(\rho_{AB}) + S(\rho_{BC})$$

(difficult to prove, but extremely useful)

* g. Δ inequality:

$$S(\rho_{AB}) \geq |S(\rho_A) - S(\rho_B)|.$$

Unlike for the Shannon entropy, it is not true that $S(\rho_{AB}) \geq \max\{S(\rho_A), S(\rho_B)\}$.

Example: If $\rho_{AB} = |\psi\rangle_{AB}\langle\psi|$ is a pure state, then $S(\rho_{AB}) = 0$, but

$$S(\rho_A) = S(\rho_B) \quad \text{may be } > 0.$$

(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 ρ, i.e. a classical random variable with $\mathbb{P}(Y=i) = \rho_{ii}$, then $H(Y) \geq S(\rho)$, with equality iff the measurement basis diagonalizes ρ.

i. Preparing the ensemble $\{p_x, |\varphi_x\rangle\}$, $H(X) \geq S\left(\sum_x p_x |\varphi_x\rangle\langle\varphi_x|\right)$