Introduction to Quantum Information Processing

QIC 710 / CS 768 / PH 767 / CO 681 / AM 871

Lecture 16 (2019)

Richard Cleve
QNC 3129
cleve@uwaterloo.ca

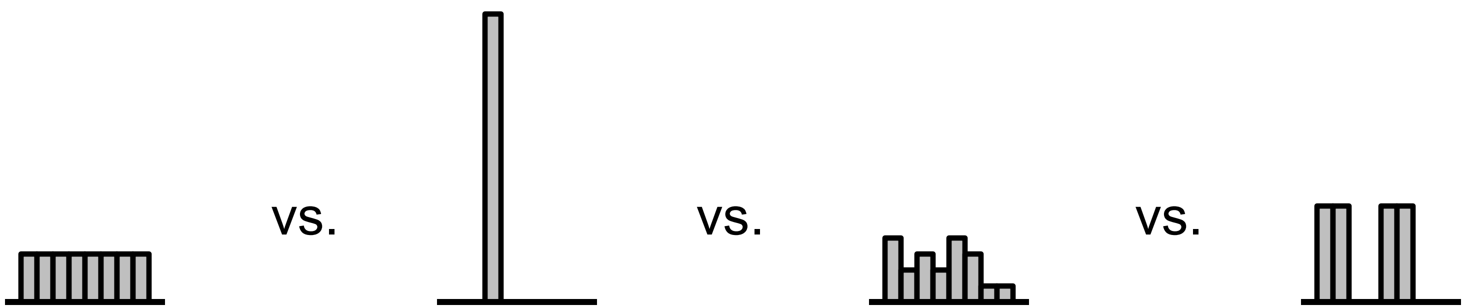
Entropy and compression

Shannon entropy

Let $\boldsymbol{p} = (p_1, \ldots, p_d)$ be a probability distribution on a set $\{1, \ldots, d\}$

Then the (Shannon) ***entropy*** of $\boldsymbol{p}$ is $H(p_1, \ldots, p_d) = -\sum_{j=1}^{d} p_j \log p_j$

Intuitively, this turns out to be a good measure of how much "randomness" (or "uncertainty") is there is in $\boldsymbol{p}$:



$H(\boldsymbol{p}) = \log d$ $\quad$ $H(\boldsymbol{p}) = 0$

We'll see that, operationally, $H(\boldsymbol{p})$ is the number of bits needed to store the outcome (in a sense that will be made formal)

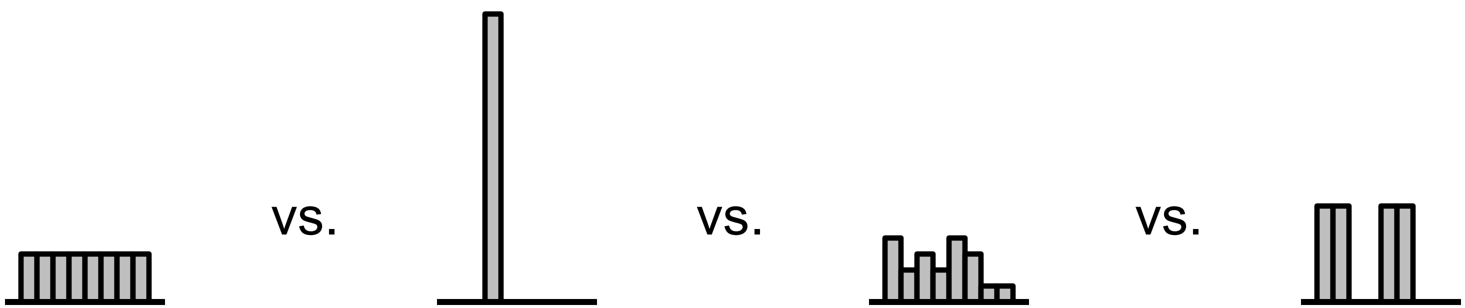
Shannon entropy

Then the (Shannon) ***entropy*** of $\boldsymbol{p}$ is $H(p_1, \ldots, p_d) =$

Let $\boldsymbol{p} = (p_1,\ldots, p_d)$ be a probability distribution on a set $\{1,\ldots,d\}$

Then the (Shannon) ***entropy*** of $\boldsymbol{p}$ is $H(p_1, \ldots, p_d) = \sum_{j=1}^{d} p_j \log p_j$

Intuitively, this turns out to be a good measure of how much “randomness” (or “uncertainty”) is there is in $\boldsymbol{p}$:



$H(\boldsymbol{p}) = \log d$ $H(\boldsymbol{p}) = 0$

We’ll see that, operationally, $H(\boldsymbol{p})$ is the number of bits needed to store the outcome (in a certain sense)

Von Neumann entropy

For a density matrix ρ, it turns out that $S(\rho) = -\mathrm{Tr}\rho \log\rho$ is a good quantum analogue of entropy

Note: $S(\rho) = H(p_1,\ldots, p_d)$, where $p_1,\ldots, p_d$ are the eigenvalues of ρ (with multiplicity)

Operationally, $S(\rho)$ is the number of ***qubits*** needed to store ρ (in a sense that will be made formal later on)

Both the classical and quantum compression results pertain to the case of large blocks of n independent instances of data:

- probability distribution $\boldsymbol{p}^{\otimes n}$ in the classical case, and
- quantum state $\rho^{\otimes n}$ in the quantum case

Classical compression (1)

Let $\boldsymbol{p} = (p_1,\ldots, p_d)$ be a probability distribution on a set $\{1,\ldots, d\}$ where n independent instances are sampled:
$(j_1,\ldots, j_n) \in \{1,\ldots, d\}^n$ (d^n possibilities, $n \log d$ bits to specify one)

Theorem* (Shannon data compression): for all $\varepsilon > 0$, for sufficiently large n, there is a scheme that compresses the specification to $n(H(\boldsymbol{p}) + \varepsilon)$ bits while introducing an error with probability at most ε

Example: an n-bit binary string with each bit distributed as $\Pr[0] = 0.9$ and $\Pr[1] = 0.1$ can be compressed to $\approx 0.47n$ bits

Intuitively, there is a subset $T \subseteq \{1,\ldots, d\}^n$, called the "typical sequences", that has size $2^{n(H(\boldsymbol{p}) + \varepsilon)}$ and probability $1 - \varepsilon$ of occurring

Note that, in the above example, $|T| \ll 2^n$ even though $\Pr[T] \geq 1 - \varepsilon$

* "Plain vanilla" version that ignores, for example, the tradeoffs between n and ε

Classical compression (2)

A nice way to prove the theorem, is based on two cleverly defined random variables …

Define the random variable $f : \{1, \ldots, d\} \to \mathbb{R}$ as $f(j) = -\log p_j$

Note that $E[f] = \sum_{j=1}^{d} p_j f(j) = -\sum_{j=1}^{d} p_j \log p_j = H(p_1, \ldots, p_d)$

Define $g : \{1, \ldots, d\}^n \to \mathbb{R}$ as $g(j_1, \ldots, j_n) = \dfrac{f(j_1) + \cdots + f(j_n)}{n}$

Thus $E[g] = H(p_1, \ldots, p_d)$

Also, $g(j_1, ..., j_n) = -\dfrac{1}{n} \log(p_{j_1} \ldots p_{j_n}) = -\dfrac{1}{n} \log\big(\Pr[(j_1, \ldots, j_n)]\big)$

which implies $\Pr[(j_1, ..., j_n)] = 2^{-ng(j_1, ..., j_n)}$

Classical compression (3)

By standard results in statistics*, as $n \to \infty$, the observed value of $g(j_1,\ldots,j_n)$ approaches its expected value, $H(\boldsymbol{p})$, in this sense:

$$\Pr\big[|g(j_1,\ldots,j_n) - H(p)| \le \varepsilon\big] \ge 1 - \varepsilon \text{ for all } \varepsilon > 0, \text{ for sufficiently large } n$$

[recall that $g(j_1,\ldots,j_n)$ is an average of independent $f(j)$]

Define $(j_1,\ldots,j_n) \in \{1,\ldots,d\}^n$ to be ***ε-typical*** if

$$\left|g\left(j_1,\ldots,j_n\right) - H\left(p\right)\right| \le \varepsilon$$

Then, the above implies, for all $\varepsilon > 0$, for sufficiently large n,

$$\Pr\big[(j_1,\ldots,j_n) \text{ is } \varepsilon\text{-typical}\big] \ge 1 - \varepsilon$$

We can also bound the ***number of*** these ε-typical sequences:

•By definition, each such sequence has probability $\ge 2^{-n(H(\boldsymbol{p}) + \varepsilon)}$

•Therefore, there can be at most $2^{n(H(\boldsymbol{p}) + \varepsilon)}$ such sequences (otherwise, the sum of probabilities would exceed 1)

---

* The weak law of large numbers

Classical compression (4)

In summary, the compression procedure is as follows:

The input data is $(j_1,\ldots,j_n) \in \{1,\ldots,d\}^n$, each independently sampled according the probability distribution $\boldsymbol{p} = (p_1,\ldots, p_d)$

The compression procedure is to leave $(j_1,\ldots,j_n)$ intact if it is ε-typical and otherwise change it to some fixed ε-typical sequence, say, some $(j_k,\ldots, j_k)$ (which will result in an error)

Since there are at most $2^{n(H(\boldsymbol{p}) + \varepsilon)}$ ε-typical sequences, the data can then be converted into $n(H(\boldsymbol{p}) + \varepsilon)$ bits

The error probability is at most ε, the probability of an input that is not typical arising

Quantum compression (1)

The scenario: n independent instances of a d-dimensional state are randomly generated according some distribution:

$$\begin{cases} |\varphi_1\rangle & \text{prob. } p_1 \\ \vdots & \vdots \quad\quad \vdots \\ |\varphi_r\rangle & \text{prob. } p_r \end{cases}$$

Example: $\begin{cases} |0\rangle & \text{prob. } \frac{1}{2} \\ |+\rangle & \text{prob. } \frac{1}{2} \end{cases}$

Goal: to "compress" this into as few qubits as possible so that the original state can be reconstructed "with small error"

What's a good formal definition of error in a quantum compression scheme?

Define a quantum compression scheme to be ***ε-good*** if no procedure can distinguish between these two states

a)the state resulting from compressing and then uncompressing the data

b)the original state

with probability more than $\frac{1}{2} + \frac{1}{4}\varepsilon$

Quantum compression (2)

Define $\rho = \sum_{i=1}^{r} p_i |\varphi_i\rangle\langle\varphi_i|$

Theorem (Schumacher data compression): for all $\varepsilon > 0$, for sufficiently large n, there is a scheme that compresses the data to $n(S(\rho) + \varepsilon)$ qubits, that is $\sqrt{2\varepsilon}$-good

For the aforementioned example, $\approx 0.6n$ qubits suffices

$$\begin{cases} |0\rangle & \text{prob. } \tfrac{1}{2} \\ |+\rangle & \text{prob. } \tfrac{1}{2} \end{cases}$$

The compression method:

Express ρ in its eigenbasis as $\rho = \sum_{j=1}^{d} q_j |\psi_j\rangle\langle\psi_j|$

With respect to this basis, we will define an ε-typical subspace of dimension $2^{n(S(\rho) + \varepsilon)} = 2^{n(H(\boldsymbol{q}) + \varepsilon)}$

Quantum compression (3)

The ***ε-typical subspace*** is that spanned by $|\psi_{j_1},\ldots,\psi_{j_n}\rangle$
where $(j_1,\ldots,j_n)$ is ε-typical with respect to $(q_1,\ldots,q_d)$

Define: Π_{typ} as the projector into the ε-typical subspace

By the same argument as in the classical case, the subspace has dimension $\leq 2^{n(S(\rho)+\varepsilon)}$ and $\text{Tr}(\Pi_{\text{typ}}\,\rho^{\otimes n}) \geq 1-\varepsilon$

Why? Because ρ is the density matrix of $\begin{cases} |\psi_1\rangle & \text{prob. } q_1 \\ \vdots & \vdots \quad \vdots \\ |\psi_d\rangle & \text{prob. } q_d \end{cases}$ ← "eigenstate" mixture

and $$\text{Tr}\left(\Pi_{\text{typ}}\rho^{\otimes n}\right) = \text{Tr}\Big(\Pi_{\text{typ}} \sum_{j_1\cdots j_n} q_{j_1}\ldots q_{j_n}|\psi_{j_1}\ldots\psi_{j_n}\rangle\langle\psi_{j_1}\ldots\psi_{j_n}|\Big)$$

$$= \sum_{j_1\cdots j_n} q_{j_1}\ldots q_{j_n}\langle\psi_{j_1}\ldots\psi_{j_n}|\Pi_{\text{typ}}|\psi_{j_1}\ldots\psi_{j_n}\rangle$$

$$= \sum_{j_1\cdots j_n} q_{j_1}\ldots q_{j_n}\chi_{[j_1\ldots j_n \text{ is typical}]} \quad \geq \quad 1-\varepsilon$$

Quantum compression (4)

We would now be done if our ***actual mixture*** was an ***eigenstate mixture***

actual mixture:

$$\begin{cases} |\varphi_1\rangle & \text{prob. } p_1 \\ \vdots & \vdots \quad \vdots \\ |\varphi_r\rangle & \text{prob. } p_r \end{cases}$$

eigenstate mixture:

$$\begin{cases} |\psi_1\rangle & \text{prob. } q_1 \\ \vdots & \vdots \quad \vdots \\ |\psi_r\rangle & \text{prob. } q_r \end{cases}$$

Calculation of the "expected fidelity" for our actual mixture:

$$\begin{aligned} \sum_I p_I \langle\phi_I|\Pi_{\text{typ}}|\phi_I\rangle &= \sum_I p_I \operatorname{Tr}\big(\Pi_{\text{typ}}|\phi_I\rangle\langle\phi_I|\big) \\ &= \operatorname{Tr}\Big(\sum_I p_I \Pi_{\text{typ}}|\phi_I\rangle\langle\phi_I|\Big) \\ &= \operatorname{Tr}\left(\Pi_{\text{typ}}\rho^{\otimes n}\right) \\ &\geq 1-\varepsilon \end{aligned}$$

Abbreviations used

$$I = i_1 i_2 \ldots i_n$$

$$p_I = p_{i_1} p_{i_2} \ldots p_{i_n}$$

$$|\phi_I\rangle = |\phi_{i_1}\phi_{i_2}\ldots\phi_{i_n}\rangle$$

Does this mean that the scheme is ε′-good for some ε′?

Quantum compression (5)

The ***true data*** is of the form $(I, |\phi_I\rangle)$ where the I is generated with probability p_I

The ***approximate data*** is of the form $\left(I, \frac{1}{\gamma_I}\Pi_{\mathrm{typ}}|\phi_I\rangle\right)$ where I is generated with probability p_I

$\gamma_I = \sqrt{\langle\phi_I|\Pi_{\mathrm{typ}}|\phi_I\rangle}$ is a normalization factor

Above two states ***at least*** as hard to distinguish as these two purifications:

$$|\Phi\rangle = \sum_I \sqrt{p_I}|I\rangle \otimes |\phi_I\rangle \qquad |\Phi'\rangle = \sum_I \sqrt{p_I}|I\rangle \otimes \frac{1}{\gamma_I}\Pi_{\mathrm{typ}}|\phi_I\rangle$$

Fidelity: $\langle\Phi|\Phi'\rangle = \sum_I p_I \frac{1}{\gamma_I}\langle\phi_I|\Pi_{\mathrm{typ}}|\phi_I\rangle \geq \sum_I p_I\langle\phi_I|\Pi_{\mathrm{typ}}|\phi_I\rangle \geq 1-\varepsilon$

Trace distance: $\|\,|\Phi\rangle - |\Phi'\rangle\,\|_{\mathrm{tr}} \leq \sqrt{2\varepsilon}$

Therefore the scheme is $\approx\sqrt{2\varepsilon}$-good