# Introduction to Quantum Information Processing

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

## Lecture 16 (2016)

**Jon Yard**

QNC 3126

jyard@uwaterloo.ca

http://math.uwaterloo.ca/~jyard/qic710

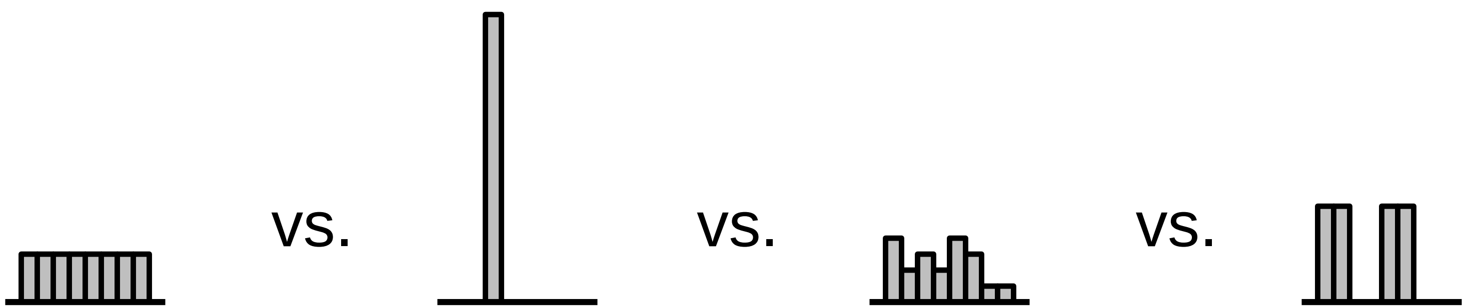
# Entropy and compression

# Shannon entropy

Let $p(x)$ be a probability distribution on a set $\{1,2,\ldots,d\}$.
A **random variable** $X$ takes values according to those probabilities, i.e. $\Pr[X = x] = p(x)$.

The (Shannon) ***entropy*** of $X$ is $H(X) = -\sum_{x=1}^{d} p(x) \log p(x)$.

Intuitively, this turns out to be a good measure of how much "randomness" (or "uncertainty", or "information") is there is in $X$:




$H(X) = \log d$ $\quad$ $H(X) = 0$

We'll see that, operationally, $H(X)$ is the number of bits needed to store the outcome (in a certain formal sense).

# Von Neumann entropy

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

**Note:** $S(\rho) = -\sum_x p(x)\log p(x)$, where the $p(x)$ are the eigenvalues of $\rho$ (with multiplicity), i.e. if

$$\rho = \sum_x p(x)|\psi_x\rangle\langle\psi_x| \quad \text{for orthonormal } |\psi_x\rangle.$$

Operationally, $S(\rho)$ is the number of ***qubits*** needed to store $\rho$ (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 $p(x_1, \dots, x_n) = p(x_1) \cdots p(x_n)$ for i.i.d. (independent and identically distributed) random variables $(X_1, \dots, X_n) \sim p(x)$
- Tensor power state $\rho^{\otimes n}$ in the quantum case

# Classical compression (1)

Let $(X_1, \ldots, X_n)$ be a sequence of i.i.d. random variables, drawn according to a probability distribution $p(x)$ on $\{1,2,\ldots,d\}$.
Then $(X_1, \ldots, X_n)$ can equal any $(x_1, x_2, \ldots, x_n) \in \{1, \ldots, d\}^n$ ($d^n$ possibilities, $n \log d$ bits to specify such a sequence)

**Theorem* (Shannon data compression):** for all $\epsilon > 0$ and all sufficiently large $n$, there is a scheme that compresses $(X_1, \ldots, X_n)$ to $n(H(X) + \epsilon)$ bits, while introducing an error with probability at most $\epsilon$.

For 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.

Proof constructs a subset $T_\epsilon^n \subset \{1, \ldots, d\}^n$ of "typical sequences" with $|T_\epsilon^n| \leq 2^{n(H(X)+\epsilon)}$ and $\Pr(X^n \in T_\epsilon^n) \geq 1 - \epsilon$.



* This version of the theorem ignores, for example, the tradeoffs between $n$ and ε

# Classical compression (2)

We prove the theorem by defining some other random variables.

First consider the random variable $\log \frac{1}{p(X)}$, where $X \sim p(x)$.

Note that $\mathbb{E}\left[\log \frac{1}{p(X)}\right] = -\sum_x p(x) \log p(x) = H(X)$

Next $(X_1, \ldots, X_n)$ be i.i.d. random variables $\sim p(x)$
and consider the random variable

$$\frac{1}{n} \log \frac{1}{p(X_1, \ldots, X_n)} = \frac{1}{n}\left(\log \frac{1}{p(X_1)} + \cdots + \log \frac{1}{p(X_n)}\right)$$

Because it is an average of i.i.d random variables $\log \frac{1}{p(X_i)}$,
the (weak) law of large numbers implies that $\frac{1}{n} \log \frac{1}{p(X_1, \ldots, X_n)}$
approaches its expected value $H(X)$ in the following formal sense:
For any $\epsilon > 0$, $\Pr\left[\left|\frac{1}{n} \log \frac{1}{p(X_1, \ldots, X_n)} - H(X)\right| \leq \epsilon\right] \to 1$ as $n \to \infty$.

# Classical compression (3)

Define $(x_1, \ldots, x_n) \in \{1, \ldots, d\}^n$ to be ***ϵ-typical*** if

$$\left| -\frac{1}{n} \log p(x_1, \ldots, x_n) - H(X) \right| \leq \epsilon.$$

Let $T_\epsilon^n$ denote the set of all $\epsilon$-typical sequences.

The results on the last slide imply the following:

For all $\epsilon > 0$ and all sufficiently large $n$,

$$\Pr[(X_1, \ldots, X_n) \in T_\epsilon^n] \geq 1 - \epsilon.$$

We can also bound the ***size*** $|T_\epsilon^n|$ of the typical set:

- By definition, each such sequence has probability $\geq 2^{-n(H(X)+\epsilon)}$
- Therefore, there can be at most $2^{n(H(X)+\epsilon)}$ such sequences

# Classical compression (4)

In summary, the compression procedure is as follows:

The input data is $(X_1, \dots X_n) \in \{1, \dots, d\}^n$, each independently sampled according the probability distribution $p(x)$

The compression procedure is to leave $(x_1, \dots, x_n)$ intact if it is $\epsilon$-typical and otherwise change it to some fixed $\epsilon$-typical sequence, say, some $(x_1, \dots, x_n)$ (which will result in an error)

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

The error probability is at most $\epsilon$, 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:

$$\begin{cases} |\varphi_1\rangle & \text{prob. } q(1) \\ \vdots & \vdots \quad \vdots \\ |\varphi_r\rangle & \text{prob. } q(r) \end{cases}$$

Example: $\begin{cases} |0\rangle & \text{prob. } ½ \\ |+\rangle & \text{prob. } ½ \end{cases}$

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

A formal definition of the notion of error is in terms of being

**$\epsilon$-good:**

No procedure can succeed at distinguishing between the following two states with probability better than $\frac{1}{2} + \frac{\epsilon}{4}$:

(a) compressing and then uncompressing the data

(b) the original data left as is

# Quantum compression (2)

Define $\rho = \sum_y q(y)|\varphi_y\rangle\langle\varphi_y|$

**Theorem (Schumacher data compression):** For all $\epsilon > 0$ and all sufficiently large $n$, there is a scheme that compresses the data to $n(S(\rho) + \epsilon)$ qubits, that is $2\sqrt{2\epsilon}$-good. If $\epsilon \leq \frac{1}{2}$, the scheme is $2\epsilon$-good.

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

**The compression method:**

Express $\rho$ in its eigenbasis as $\rho = \sum_x p(x)|\psi_x\rangle\langle\psi_x|$

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

# Quantum compression (3)

The ***$\epsilon$-typical subspace*** is that spanned by
$|\psi_{x^n}\rangle := |\psi_{x_1}\rangle|\psi_{x_2}\rangle \cdots |\psi_{x_n}\rangle$ where $(x_1, \ldots, x_n) \in T_\epsilon^n$.

**Define:** $\Pi_\epsilon^n$ as the projector into the $\epsilon$-typical subspace

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

Why? Because $\rho$ is the density matrix of $\begin{cases} |\psi_1\rangle & \text{prob. } p(1) \\ \vdots & \vdots \quad\quad \vdots \\ |\psi_d\rangle & \text{prob. } p(d) \end{cases}$

$$\mathrm{Tr}\Pi_\epsilon^n \rho^{\otimes n} = \mathrm{Tr}\Pi_\epsilon^n \sum_{x^n} p(x^n)|\psi_{x^n}\rangle\langle\psi_{x^n}| = \sum_{x} p(x^n)\langle\psi_{x^n}|\Pi_\epsilon^n|\psi_{x^n}\rangle$$

$$= \sum_{x^n \in T_\epsilon^n} p(x^n) \geq 1 - \epsilon.$$

# Quantum compression (4)

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

$$\sum_{y^n} q(y^n)\langle\varphi_{y^n}|\Pi_\epsilon^n|\varphi_{y^n}\rangle = \sum_{y^n} q(y^n)\mathrm{Tr}\Pi_\epsilon^n\big|\varphi_{y^n}\rangle\langle\varphi_{y^n}\big|$$

$$= \mathrm{Tr}\Pi_\epsilon^n \sum_{y^n} q(y^n)\,\big|\varphi_{y^n}\rangle\langle\varphi_{y^n}\big|$$

$$= \mathrm{Tr}\Pi_\epsilon^n \rho^{\otimes n}$$

$$\geq 1-\epsilon$$

**Does this mean that the scheme is $\epsilon$-good for some $\epsilon$?**

# Quantum compression (5)

The ***true data*** is of the form $(y^n, |\varphi_{y^n}\rangle)$, where $y^n$ is generated with probability $q(y^n)$.

The ***approximate data*** is of the form $(y^n, |\varphi'_{y^n}\rangle)$,

where $\left|\varphi'_{y^n}\right\rangle = \frac{1}{c_{y_n}} \Pi_\epsilon^n |\varphi_{y^n}\rangle$, $c_{y^n} = \sqrt{\langle\varphi_{y^n}|\Pi_\epsilon^n|\varphi_{y^n}\rangle}$ is a normalization factor and $y^n$ is generated with probability $q(y^n)$.

We can bound the fidelity between them by defining purifications:

$$|\Phi\rangle = \sum_{y^n} \sqrt{q(y^n)}\, |y^n\rangle|\varphi_{y^n}\rangle \quad |\Phi'\rangle = \sum_{y^n} \sqrt{q(y^n)}\, |y^n\rangle|\varphi'_{y^n}\rangle$$

$$F\left(\rho^{\otimes n}, \sum_{y^n} q(y^n)\, |\varphi'_{y^n}\rangle\langle\varphi'_{y^n}|\right) \geq \langle\Phi|\Phi'\rangle$$

$$= \sum_{y^n} \frac{q(y^n)}{c_{y^n}} \langle\varphi_{y^n}|\Pi_\epsilon^n|\varphi_{y^n}\rangle \geq \sum_{y^n} q(y^n)\, \langle\varphi_{y^n}|\Pi_\epsilon^n|\varphi_{y^n}\rangle \geq 1-\epsilon$$

# Quantum compression (6)

But how well can we distinguish between these two states?

$$\rho^{\otimes n} = \sum_{y^n} q(y^n)|\varphi_{y^n}\rangle\langle\varphi_{y^n}| = \sum_{y^n} p(x^n)|\psi_{x^n}\rangle\langle\psi_{x^n}|,$$

$$\rho' = \sum_{y^n} q(y^n)\left|\varphi'_{y^n}\rangle\langle\varphi'_{y^n}\right| = \frac{1}{\mathrm{Tr}\Pi_\epsilon^n\rho^{\otimes n}}\Pi_\epsilon^n\rho^{\otimes n}\Pi_\epsilon^n$$

Can try to directly bound trace distance

$$\left\|\rho^{\otimes n} - \rho'\right\|_1 \leq \frac{1}{\mathrm{Tr}\Pi_\epsilon^n\rho^{\otimes n}} \sum_{x^n \notin T_\epsilon^n} p(x^n) \leq \frac{\epsilon}{1-\epsilon} \leq 2\epsilon \text{ if } \epsilon \leq \frac{1}{2}.$$

Get a bound for all $\epsilon$ using relation to fidelity:

$$\left\|\rho^{\otimes n} - \rho'\right\|_1 \leq 2\sqrt{1 - F(\rho^{\otimes n}, \rho')^2} \leq 2\sqrt{1-(1-\epsilon)^2} \leq 2\sqrt{2\epsilon}.$$

Therefore the scheme is $\epsilon$-good if $\epsilon \leq \frac{1}{2}$,

and it is $2\sqrt{2\epsilon}$-good for regardless of the value of $\epsilon$.