ENTROPY AND DATA COMPRESSION

→ compression software (ZIP, RAR, etc) exists

File (1 GB) —zip (compression encoding)→ New file (700 MB) —ZIP (decompres. decoding)→ File (1 GB)

New file (700 MB) —compress again→ 700 MB

① Classical (Shannon) entropy

→ Let Σ be a finite alphabet

$$\Sigma = \{x_1, x_2, \dots, x_d\} \qquad d = |\Sigma|$$

examples $\Sigma = \{0,1\}$; $\Sigma = \{a,b,c,\dots,z\}$, etc

→ Let $P(\Sigma)$ the set of probability dist. on Σ

$$P(\Sigma) = \left\{ p : \Sigma \to \mathbb{R}_+ : \sum_{x \in \Sigma} p(x) = 1 \right\}$$

$a \in \mathbb{R}, a \geq 0$

→ If Σ is numeric, X is a random variable

$$\mathbb{P}(X = x) = p(x) \qquad \forall x \in \Sigma$$

↳ $\mathbb{E}[X] = \sum_{x \in X} p(x) \cdot x$

↳ $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

$= \mathbb{E}[\underbrace{(X - \mathbb{E}(X))^2}_{\text{centered } X}]$

Definition The entropy of a prob. dist $p \in P(\Sigma)$ is given by

$$H(p) = -\sum_{x \in \Sigma} p(x) \log p(x)$$

base 2 log: $\log 2 = 1$

→ if $p(x) = 0$, then define $p(x) \log p(x) = 0$ since $\lim_{a \to 0^+} a \log a = 0$

→ we can write $H(p) = -\mathbb{E}[\log p(x)]$

$= \mathbb{E}[\underbrace{\log \frac{1}{p(x)}}_{\substack{\text{surprise} \\ \text{uncertainty}}}]$

Properties of H

→ $H(p) \geq 0 \ \forall p$ $\quad -a \log a \geq 0 \ \forall a \in [0,1]$

↳ equality if $\forall x \ p(x) \in \{0,1\}$

$H(p) = 0 \iff \exists x_0$ s.t. $p(x) = \begin{cases} 1 & \text{if } x = x_0 \\ 0 & \text{if } x \neq x_0 \end{cases}$

Such a p is called **deterministic**.

→ $H(p) \leq \log |\{x : p(x) > 0\}| \leq \log |\Sigma|$

↳ equality case:

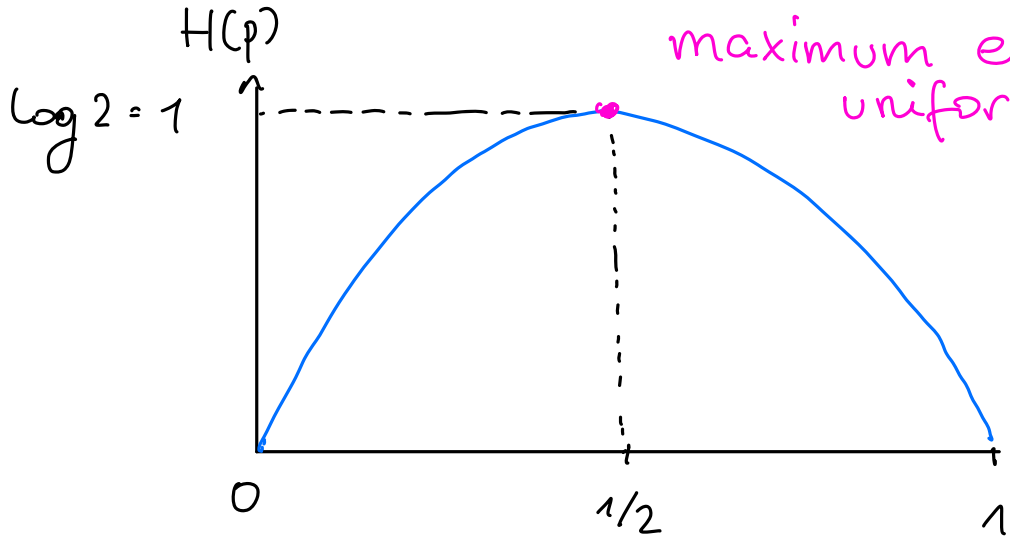
$H(p) = \log |\Sigma| \iff p(x) = \frac{1}{|\Sigma|} \quad \forall x$

Such a p is called **uniform**

→ **binary entropy** $|\Sigma| = 2$ (bits)

$p \in P(\Sigma)$ $\quad p = (1-p, p)$ $\quad p \in [0,1]$

$p(1) = p$ $\quad p(0) = 1 - p$.



maximum entropy for uniform p

↑ uniform random bit "coin flip"

② Data compression

→ we are given a **source** of random **iid.** letters from Σ

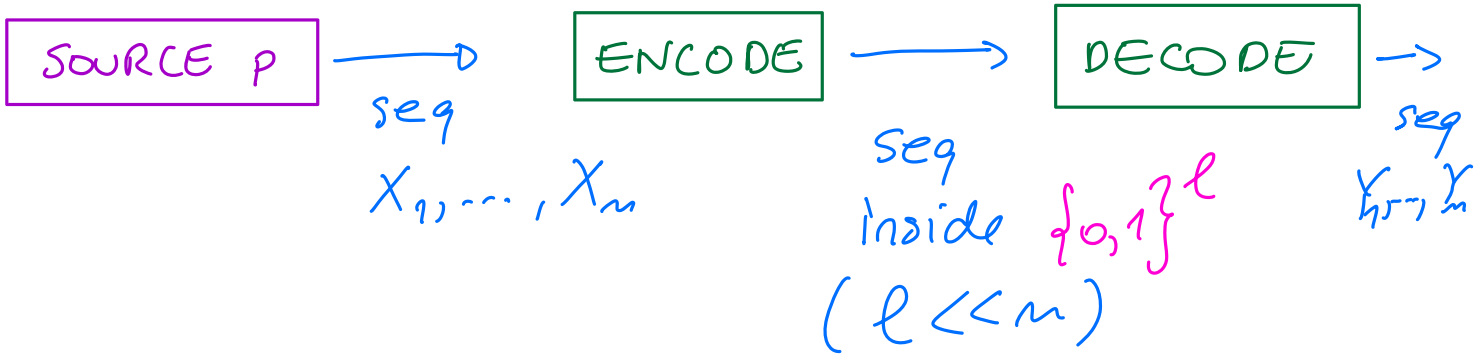
X random variable

$$\mathbb{P}(X = x \in \Sigma) = p(x) \qquad X \sim p.$$

→ a source is a device which produces letters from Σ

- independently : each letter is independent from the old ones
- each letter has distribution p

→ Goal : compress (as well as possible) and later to decompress sequences produced by the source



Want • $(Y_1, \dots, Y_m) \approx (X_1, \dots, X_m)$

• $\ell << m$

→ sequences of length 1, lossless

↑ we want $Y_1 = X_1$

↳ "best" ℓ :

$$\ell = H_0(p) = \log |\{x : p(x) > 0\}|$$

③ Shannon's source coding theorem

→ iid source with prob. $p \in P(\Sigma)$

→ **lossy** compression is allowed

$$\mathbb{P}\left[(Y_1, \dots, Y_m) \neq (X_1, \dots, X_m)\right] \leq \delta$$

for some small error prob. $\delta \in (0,1)$

→ **compression rate** $= \frac{\text{number of bits}}{\text{block length}} = \frac{\ell}{n} = R$

Definition A **(n, R, δ)-code** for a fixed $p \in P(\Sigma)$ is a pair of encoder-decoder functions

$$E : \Sigma^n \to \{0,1\}^{\lfloor nR \rfloor}$$

↑ floor of nR = largest integer $\leq nR$

$$D : \{0,1\}^{\lfloor nR \rfloor} \to \Sigma^n$$

such that $\mathbb{P}\left(D(E(\underbrace{X^n}_{(X_1, \dots, X_n)})) \neq X^n \right) \leq \delta$

for $X_1, \dots, X_n$ **iid** with dist p.

- n = block length (seq. length)
- R = compression rate
- δ = error probability

Example $\Sigma = \{a, b, c, d\}$

$p(a) = p(b) = p(c) = p(d) = \frac{1}{4}$

uniform probability

$(n=1, R=2, \delta=0)$ - code $= (E, D)$

$E(a) = 00$

$E(b) = 01$ $\quad (\ell = 2)$

$E(c) = 10$

$E(d) = 11$

and

$D(00) = a$

$D(01) = b$

$D(10) = c$

$D(11) = d$

↳ perfect code :

$\delta = 0 \iff \forall x \in \Sigma, \; D(E(x)) = x$

Theorem [Shannon's source coding theorem]

Let $p \in P(\Sigma)$ and $\delta \in (0,1)$. Then

(A) If $R > H(p)$, then there exists a (n, R, δ)-code for n large enough

(B) If $R < H(p)$, then, for n large enough, there does not exist a (n, R, δ)-code

→ this theorem is saying that the optimal asymptotic compression rate for an iid source is its entropy

→ probability of error δ

$$\mathbb{P}\left(D(E(X^n)) \neq X^n\right) = \sum_{\substack{x^n = (x_1,\dots,x_n) \in \Sigma^n \\ D(E(x^n)) \neq x^n}} p(x^n)$$

but iid $\Rightarrow p(x^n) = p(x_1)\cdot p(x_2)\cdots p(x_n)$

→ iid sequences from p (n large enough) have the typicality property

$$\forall y \in \Sigma \quad |\{i : x_i = y\}| \underset{\varepsilon}{\approx} n \cdot p(y)$$

Definition $p \in P(\Sigma)$, n block length, $\varepsilon > 0$. The typical set

$$T_{n,\varepsilon}(p) = \{ x^n \in \Sigma^n : \left| \frac{1}{n} \log \frac{1}{p(x^n)} - H(p) \right| \leq \varepsilon \}$$

Σ^n

$$= \{ x^n : \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{1}{p(x_i)} - H(p) \right| \leq \varepsilon \}$$

Properties of typical sets

→ $\forall x^n \in T_{n,\varepsilon}$

$$2^{-n(H(p)+\varepsilon)} \leq p(x^n) \leq 2^{-n(H(p)-\varepsilon)}$$

all sequences in a typical set have the same probability $\simeq 2^{-nH(p)}$

→ $|T_{n,\varepsilon}| \leq 2^{n(H(p)+\varepsilon)}$ ✓ (*)

typical sets are not too big

→ $\mathbb{P}[x^n \notin T_{n,\varepsilon}] \leq \frac{\sigma^2}{n\varepsilon}$ (**)

where $\sigma^2 = \mathrm{Var}\left(\log \frac{1}{p(x)}\right)$

↑ recall that $H(p) = \mathbb{E} \log \frac{1}{p(x)}$

typical sets are very likely

→ proof of (*)

$\mathbb{P}(x^n \in T_{n,\varepsilon}) \leq 1$

∨

$$|T_{n,\varepsilon}| \cdot \min_{x^n \in T_{n,\varepsilon}} p(x_n) \geq |T_{n,\varepsilon}| \, 2^{-n(H(p)+\varepsilon)}$$

④ Proof of part (A) of Shannon's theorem

→ We are given a rate $R > H(p)$.

→ define $\varepsilon = \frac{R - H(p)}{2} > 0$; consider $T_{n,\varepsilon}$

$\textcircled{*} \Rightarrow |T_{n,\varepsilon}| \leq 2^{n(H(p)+\varepsilon)} < 2^{\lfloor nR \rfloor}$

$= \{0,1\}^{\lfloor nR \rfloor}$

→ consider any $E : T_{n,\varepsilon} \longrightarrow \{0,1\}^{\lfloor nR \rfloor}$ injective

and its inverse

$D : \{0,1\}^{\lfloor nR \rfloor} \longrightarrow T_{n,\varepsilon}$ s.t.

$D(E(x^n)) = x^n \quad \forall x^n \in T_{n,\varepsilon}$

→ extend E arbitracily from $T_{n,\varepsilon}$ to Σ^n (we do not care about what happens outside $T_{n,\varepsilon}$; we shall make errors for words $\Sigma^n \setminus T_{n,\varepsilon}$ but these words are not very likely)

→ Let us bound the error probability of the code we just introduced

$$\mathbb{P}\left(D(E(x^n)) \neq x^n \right) \leq \mathbb{P}\left(x^n \notin T_{n,\varepsilon} \right)$$

$$\overset{\textcircled{**}}{\leq} \frac{\sigma^2}{n\varepsilon} \leq \delta \quad \text{for } n \text{ large enough}$$

Summary of the main proof idea

→ the encoder $E: \Sigma^n \longrightarrow \{0,1\}^{\lfloor nR \rfloor}$ checks whether $x^n \in \Sigma^n$ is an element of a typical set $T_{n,\varepsilon}$

- if YES, encode it perfectly
- if NO, we do not care, encode it any way you like

→ the decoder decodes any $z \in \{0,1\}^{\lfloor nR \rfloor}$ to an element $y \in T_{n,\varepsilon}$ (perfectly)

→ this pair (D, E) works because

- the sets $T_{n,\varepsilon}$ are small
- the sets $T_{n,\varepsilon}$ have large probability

Examples

① trivial source $p(x_0) = 1$ $p(x \neq x_0) = 0$

$H(p) = 0$

↝ rate $R = 0$ is possible

True, since the decoder can

always output "x_0"

② uniform source $p(x) = \frac{1}{|\Sigma|}$

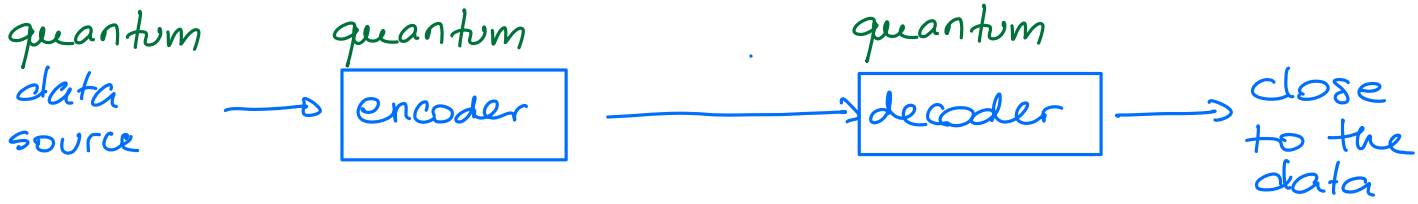
$$H(p) = \log |\Sigma|$$

best rate is $R = \log |\Sigma|$

→ no nontrivial encoding/decoding scheme is possible

↳ why passing a file through zip twice does not increase compression rate

QUANTUM DATA COMPRESSION



① Quantum data source

→ classical source : $p \in \mathcal{P}(\Sigma)$ ← finite alphabet $\{1, 2, \dots, d\}$

↳ probability vector

$p(x) \geq 0 \quad \forall x \in \Sigma$

$\sum_{x \in \Sigma} p(x) = 1$

↳ produces i.i.d. samples $X_1, \dots, X_n$

↑ independent identically distributed

$\mathbb{P}(X_i = x) = p(x)$

→ quantum data source

↳ replace finite alphabet $\Sigma = \{1, 2, \dots, d\}$

↓

Hilbert space $H = \mathbb{C}^d$

↳ a source is a collection

$p_1, (\rho_x)_{x=1}^d$ where ρ_x are density matrices on $H \simeq \mathbb{C}^k$

↳ classical sources

$p(x)$, $\rho_x = |x\rangle\langle x|$ ← basis vector

↳ iid: $\rho_{x_1} \otimes \rho_{x_2} \otimes \dots \otimes \rho_{x_n}$ ← used for independence

② Quantum fidelity

Definition Given two density matrices ρ, σ on H, their **fidelity** is defined by

$$F(\rho,\sigma) = \mathrm{Tr}\left[\sqrt{\sqrt{\sigma}\cdot\rho\cdot\sqrt{\sigma}}\right]$$

→ recall that, given a **PSD** matrix X having a sp. decomp

$$X = \sum_{i=1}^{d} \lambda_i \, |x_i\rangle\langle x_i|$$

($|x_i\rangle$: eigenvectors; λ_i: eigenvalues ≥ 0)

we define $$\sqrt{X} := \sum_{i=1}^{d} \sqrt{\lambda_i} \, |x_i\rangle\langle x_i|$$

Properties of the fidelity function

→ $F(\rho,\sigma) \in [0,1]$

- $F(\rho,\sigma) = 1 \iff \rho = \sigma$
- $F(\rho,\sigma) = 0 \iff \langle \rho, \sigma \rangle = 0$

"fidelity is a measure of **similarity**, rather than a distance"

→ Symmetry: $F(\rho,\sigma) = F(\sigma,\rho)$

→ classical states: $\rho = \mathrm{diag}(p)$; $\sigma = \mathrm{diag}(q)$

$$F(\rho,\sigma) = \sqrt{\sum p_i q_i}$$

→ If $\sigma = |\psi\rangle\langle\psi|$ is **pure**, then $\sqrt{\sigma} = |\psi\rangle\langle\psi|$ and thus

$$F(\rho, |\psi\rangle\langle\psi|) = \sqrt{\langle\psi|\rho|\psi\rangle}$$

↳ in particular, if ρ is also pure $\rho = |\varphi\rangle\langle\varphi|$, then

$$F(|\varphi\rangle\langle\varphi|, |\psi\rangle\langle\psi|) = |\langle\varphi,\psi\rangle|$$

③ Quantum encoders / decoders

→ encoder : compress data from $\Sigma^n \longrightarrow \{0,1\}^{\lfloor nR \rfloor}$

→ decoder : decompress $\{0,1\}^{\lfloor nR \rfloor} \longrightarrow \Sigma^n$

"classical functions"

↓

quantum channels

Definition An (n, R, δ)-quantum code is a pair of quantum channels

$$\mathcal{E} : \mathcal{B}(H^{\otimes n}) \longrightarrow \mathcal{B}\left((\mathbb{C}^2)^{\otimes \lfloor nR \rfloor}\right)$$

↑ linear operators on a Hilbert space

$$\mathcal{D} : \mathcal{B}\left((\mathbb{C}^2)^{\otimes \lfloor nR \rfloor}\right) \longrightarrow \mathcal{B}\left(H^{\otimes n}\right)$$

s.th.

$$F\left(\sigma_{A^n B}, \left[(\mathcal{D}\circ\mathcal{E})\otimes id_B\right](\sigma_{A^n B})\right) \geq 1-\delta$$

A is the system H ↑ reference state (accounts for ent.)

for all $\sigma_{A^n B}$ s.t. $\sigma_{A^n} = Tr_B\, \sigma_{A^n B} = \rho^{\otimes n}$

where $\rho = \sum_x p(x)\rho_x$ average state of the source

④ von Neumann entropy

Definition Given a density matrix $\rho \in B(\mathbb{C}^d)$, its von Neumann (quantum) entropy is

$$H(\rho) = -\sum \lambda_i \log \lambda_i$$

where $(\lambda_1, \dots, \lambda_d)$ are the eigenvalues of ρ.

→ $H(\rho) = -\mathrm{Tr}[\rho \cdot \log \rho]$

→ $H(\rho) = H(\lambda)$ $\quad \lambda$ = eigenvalue vector

($H(\rho)$: vN; $H(\lambda)$: Shannon)

→ inherits all the properties of the classical ent:

↳ $H(\rho) \in [0, \log d]$

- $H(\rho) = 0 \iff \rho$ is pure
- $H(\rho) = \log d \iff \rho = I/d$

⑤ Schumacher's coding theorem

Theorem Consider a data source with average state $\rho \in B(H)$, and $\delta \in (0,1)$.

(A) if $R > H(\rho)$, then there exists n_0 s.t. $\forall n \geq n_0$ $\exists$ (n, R, δ)-quantum code for ρ

(B) if $R < H(\rho)$, then exists some n_0 s.t. $\forall n \geq n_0$ $\nexists$ (n, R, δ)-quantum code for ρ

Proof idea Replace typical sequences $T_{n,\varepsilon}$ by typical subspaces $S_{n,\varepsilon}$

$$S_{n,\varepsilon} = \text{span}\{ |x_1\rangle \otimes \cdots \otimes |x_n\rangle : x^n = (x_1, \ldots, x_n) \in T_{n,\varepsilon} \}$$

↳ typical projections $\Pi_{n,\varepsilon}$ on $S_{n,\varepsilon}$

→ follow the classical proof by repacing

sets → subspaces

functions → channels