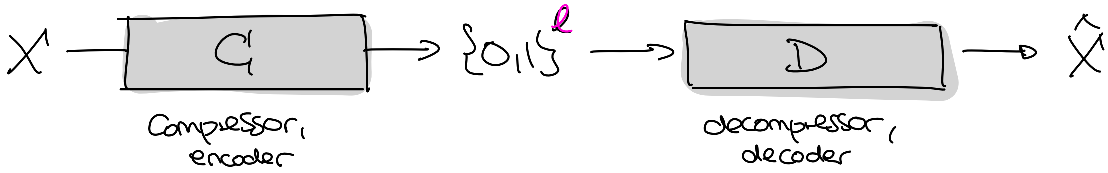
Lossy Compression & The Source Coding Theorem (§4)

Today we fix the number of bits but allow small error probability ("lossy"):



WANT: $\Pr(\hat{X} \neq X) \leq \delta$

How to achieve?

* Take set $S \subseteq \mathcal{A}$ with $\Pr(X \notin S) \leq \delta$.
* Then we can compress into $\ell = \lceil \log \# S \rceil$ bits with error probability $\leq \delta$. How? Simply define C by sending all $x \in S$ to distinct bitstrings. (For $x \notin S$, pick arbitrary, or fail.)

ex:

x	$P(x)$	$\delta=0$	$\delta=1/16$
a	1/4	000	00
b	1/4	001	01
c	1/4	010	10
d	3/16	011	11
e	1/64	100	—
f	1/64	101	—
g	1/64	110	—
h	1/64	111	—

arbitrary

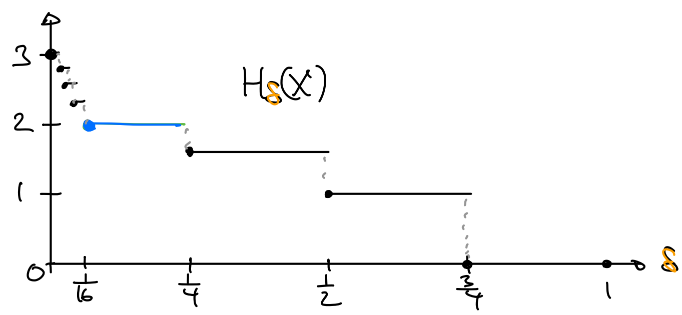
$\ell = 2$

Define **δ-essential bit content** by

$$H_\delta(X) = H_\delta(P) = \min \{ \log \# S \mid \Pr(X \notin S) \leq \delta \}$$

$\Rightarrow$ $\boxed{\underbrace{\lceil H_\delta(X) \rceil}_{\text{if not integer, need to round up!}} \text{ is minimal \# bits required to compress } X \text{ with error } \leq \delta}$

$H_\delta(X)$ is in general quite messy ... see here



Amazingly, it simplifies dramatically if we compress blocks of symbols!

Shannon's Source Coding Theorem: Let $X_1, X_2, X_3, \ldots \overset{\text{IID}}{\sim} P$ and $0 < \delta < 1$:

$$\lim_{N \to \infty} \frac{H_\delta(X_1, \ldots, X_N)}{N} = H(P)$$

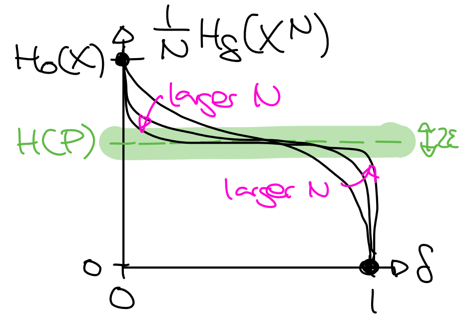
IID (memoryless) information source

optimal compression rate for block size N and error prob $\leq \delta$

optimal asymptotic compression rate ← independent of δ!

(i.e. $\forall \delta \in (0,1), \varepsilon > 0 \ \exists N_0 \ \forall N \geq N_0 : \left| \frac{H_\delta(X_1 \ldots X_N)}{N} - H(P) \right| \leq \varepsilon$)

* If $R > H(P)$: $\exists N_0\ \forall N \geq N_0$:
  CAN compress at rate R ($=$ into $\ell \leq RN$ bits)
* If $R < H(P)$: $\exists N_0\ \forall N \geq N_0$:
  CANNOT compress at rate R



Proof of the Source Coding Theorem

NOTATION: $x^N = x_1 \cdots x_N = (x_1, \dots, x_N)$ for strings of length N.

Typical set: $$T_{N,\varepsilon}(P) = \left\{ x^N \in \mathcal{A}_X^N : \left| \frac{1}{N} \log \frac{1}{P(x^N)} - H(P) \right| \leq \varepsilon \right\}$$
$$\overset{\text{iid}}{=} \left\{ x^N \in \mathcal{A}_X^N : \left| \frac{1}{N} \sum_{k=1}^{N} \log \frac{1}{P(x_k)} - H(P) \right| \leq \varepsilon \right\}$$

Properties:

⓪ $2^{-N(H(P)+\varepsilon)} \leq P(x^N) \leq 2^{-N(H(P)-\varepsilon)}$ (by definition)

① $\# T_{N,\varepsilon} \leq 2^{N(H(P)+\varepsilon)}$

Pf: $1 \geq \Pr(X^N \in T_{N,\varepsilon}) = \sum_{x^N \in T_{N,\varepsilon}} P(x^N) \geq \# T_{N,\varepsilon} \cdot 2^{-N(H(P)+\varepsilon)}$. □

② $\Pr(X^N \notin T_{N,\varepsilon}) \leq \frac{\sigma^2}{N\varepsilon^2} \longrightarrow 0$, where $\sigma^2 = \mathrm{Var}\left(\log \frac{1}{P(X_k)}\right)$.

Pf: Let $L_k = \log \frac{1}{P(X_k)}$ and $\mu := E[L_k] = H(X_k) = H(P)$. Then:
$$\text{LHS} = \Pr\left(\left| \frac{1}{N} \sum_{k=1}^{N} L_k - \mu \right| > \varepsilon \right) \leq \frac{\mathrm{Var}(L_k)}{N \varepsilon^2}. \quad \square$$

"Asymptotic Equipartition Property" (AEP)

"For large N... ...typical probabilities are $2^{-N(H(P) \pm \varepsilon)}$."

Proof of the theorem: Let $\delta \in (0,1)$ and $\varepsilon > 0$ be arbitrary.

(A) $$\Pr(X^N \in T_{N,\varepsilon}) \overset{②}{\geq} 1 - \frac{\sigma^2}{N\varepsilon^2} \geq 1 - \delta \quad \text{if } N \text{ large enough}$$
$$\Longrightarrow \frac{H_\delta(X^N)}{N} \leq \frac{\log \# T_{N,\varepsilon}}{N} \overset{①}{\leq} H(P) + \varepsilon \quad \text{for large } N. \qquad \infty$$

(13) Want to prove that $\frac{H_\delta(X^N)}{N} \geq H(P) - \varepsilon$ for N large.

If not: $\exists$ sets S_N for $N \to \infty$ s.th.

$$\Pr(X^N \in S_N) \geq 1-\delta \quad \text{and} \quad \# S_N < 2^{N(H(P)-\varepsilon)}.$$

$$\Longrightarrow 1-\delta \leq \Pr(X^N \in S_N) = \Pr(X^N \in S_N \cap T_{N,\varepsilon/2}) + \Pr(X^N \in S_N \setminus T_{N,\varepsilon/2})$$

$$\leq \underbrace{\Pr(X^N \in S_N \cap T_{N,\varepsilon/2})}_{\overset{(6)}{\leq} \# S_N \cdot 2^{-N(H(P)-\frac{\varepsilon}{2})} \leq 2^{-N\varepsilon/2} \to 0} + \underbrace{\Pr(X^N \notin T_{N,\varepsilon/2})}_{\to 0 \text{ by } (2)} \longrightarrow 0 \ ↯$$

$\square$

Remark: $T_{N,\varepsilon}$ is usually NOT the smallest set S_N w/ $\Pr(X^N \in S_N) \geq 1-\delta$...

... but small enough and easy to handle as $N \to \infty$! $\longrightarrow$ $\boxed{\text{EX CLASS}}$

How to use this in practice?

SCENARIO: want to compress IID (memoryless) data source P
(we know P, but NOT which string will be emitted)

FIX: * block size N
* parameter $\varepsilon > 0$
* a way to order the typical set $T_{N,\varepsilon}$

index	element
0	- - -
1	- - -
$\vdots$	- - -
$\# T_{N,\varepsilon} - 1$	- - -

COMPRESSOR: Input: A string $x^N = x_1 \cdots x_N$

* If $x^N \notin T^{(P)}_{N,\varepsilon}$: FAIL
* Determine index p of x^N in $T_{N,\varepsilon}$.
* Return p in binary.

DECOMPRESSOR:

Input: A binary string s

* Interpret s as integer p
* Return p-th element of $T_{N,\varepsilon}$.

This is a lossy compression protocol:

* Error probability: $\Pr(X^N \notin T_{N,\varepsilon}) \overset{\text{AEP}}{\leq} \frac{\sigma^2}{N\varepsilon^2} \longrightarrow 0$ as $N \to \infty$

* Rate $R = \frac{\#\text{bits required to represent } p}{N}$

$$\leq \frac{\log \# T_{N,\varepsilon} + 1}{N} \overset{\text{AEP}}{\leq} H(P) + \varepsilon + \underbrace{\frac{1}{N}}_{\to 0}$$

Variations

(A) How to make it LOSSLESS?

When $x^N \notin T_{N,\varepsilon}$, send uncompressed using $N \cdot \lceil \log \# \mathcal{A}_X \rceil$ bits.

0	as above
1	uncompressed

"flag" bit — prefix code!

$$\Rightarrow \text{average rate } \bar{R} \le \frac{1}{N} + \underbrace{\Pr(X^N \in T_{N,\varepsilon})}_{\to 1} \underbrace{\left(H(P) + \varepsilon + \frac{1}{N}\right)}_{\text{from above}} + \underbrace{\Pr(X^N \notin T_{N,\varepsilon})}_{\to 0} \cdot \lceil \log \# \mathcal{A}_X \rceil$$

$$\approx H(P) + \varepsilon \text{ for large } N$$

← agrees with symbol code discussion

(B) How to also make it UNIVERSAL? (IID, but we do NOT know P)

For simplicity: assume $\mathcal{A} = \{0,1\}$ i.e. data source of bits.

FIX: * block size N

* a way to order the sets

$B(N,k) := \{x^N \text{ with } k \text{ ones and } N-k \text{ zeros}\}$

$B(3,2)$

index	string
0	011
1	101
2	110

→ ex class, HW

COMPRESSOR: Input: A bitstring $x^N = x_1 \cdots x_N$

* Compute $k := \#$ones in x^N
* Determine index p of x^N in $B(N,k)$
* Return $\underbrace{k}_{\approx \log(N)+1}$ and $\underbrace{p}_{\approx \log \# B(N,k) + 1 \text{ bits}}$ in binary.

DECOMPRESSOR: Clear !? :

Key idea: $B(N,k)$ can be MUCH SMALLER than $\{0,1\}^N$ (e.g. imagine $k=1$)

Average rate $\bar{R}$? Assume that $X_1, \ldots, X_N \overset{\text{IID}}{\sim} P$. Then:

NOT used in protocol, only in the analysis !!!

depends on P, but only used in analysis!

$$x^N \in T_{N,\varepsilon} \xrightarrow[\text{since}]{} B(N,k) \subseteq T_{N,\varepsilon} \Rightarrow \# B(N,k) \le \# T_{N,\varepsilon} \quad (*)$$

since typicality only depends on # zeros and ones in x^N!

Thus we can argue as above:

$$\overline{R} = \frac{\#\text{bits required to represent } k + \#\text{bits required to represent } p}{N}$$

$$\lesssim \frac{\log(N)}{N} + \frac{\log \# B(n,k)}{N}$$

dropping some $\frac{1}{N}$ terms

$\to 0$, so can ignore

use ⊛ to obtain the following bound:

$$\le \underbrace{\Pr(X^N \in T_{N,\varepsilon})}_{\to 1} \cdot \underbrace{\frac{\log \# T_{N,\varepsilon}}{N}}_{\le H(P)+\varepsilon} + \underbrace{\Pr(X^N \notin T_{N,\varepsilon})}_{\to 0\text{, as before}} \underbrace{\frac{\log 2^N}{N}}_{=1}$$

$\approx H(P) + \varepsilon$ for large N!

$\boxed{\text{HW:}}$ Program this protocol & compress the donkey!

<u>Discussion:</u> Many disadvantages!

* Have to look at entire x^N to compress. Can we compress by looking at a few symbols at a time?

* Assume IID distribution... what if P <u>changes</u>? Or if we have <u>local</u> correlations?

Q $\xrightarrow{\text{frequent}}$ U

Q $\xrightarrow{\text{rare}}$ R

↳ Wednesday ☺