The Source Coding Theorem: Proof and Variations (§4)

Recall from Tuesday:

$$H_\delta(Y) = \log \min \{\#S : \Pr(Y \in S) \geq 1-\delta\} \leftarrow \delta\text{-essential bit content}$$

$\hat{=}$ minimum bits need to compress Y w/ error probability $\leq \delta$

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

IID (memoryless) information source

$$\lim_{N \to \infty} \frac{H_\delta(X^N)}{N} = H(P)$$

optimal compression rate for N samples

optimal asymptotic compression rate

Last time we motivated the following definition:

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

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

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 $2^{-N(H(P) \pm \delta)}$

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

($\leq$): $\Pr(X^N \in T_{N,\varepsilon}) \overset{②}{\geq} 1 - \frac{\sigma^2}{N\varepsilon^2} \geq 1 - \delta$ if N large enough

$$\Rightarrow \frac{H_\delta(X^N)}{N} \leq \frac{\log \# T_{N,\varepsilon}}{N} \overset{①}{\leq} H(P) + \varepsilon$$

($\geq$) 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.

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

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

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

$\square$

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 $N \to \infty$! $\to$ $\boxed{\text{EX CLASS}}$

How to use this in practice?

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

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

index	String
0	- - -
1	- - -
⋮	- - -
$\# T_{N,\varepsilon} - 1$	- - -

COMPRESSOR: Input: A string $x^N = x_1 \cdots x_N$
* If $x^N \notin T_{N,\varepsilon}$: 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: $\Pr(X^N \notin T_{N,\varepsilon}) \overset{\text{AEP}}{\leq} \frac{\sigma^2}{N\varepsilon^2} \to 0$ 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

How to make it LOSSLESS? Instead of failing, send x^N uncompressed!

↳ average rate $\bar{R} \leq \underbrace{\Pr(X^N \in T_{N,\varepsilon})}_{\to 1} \cdot \left(H(P) + \varepsilon + \frac{1}{N}\right) + \underbrace{\Pr(X^N \notin T_{N,\varepsilon})}_{\to 0} H_0(P)$

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

Undesirable that we need to know P... can we compress w/o knowing P?

"UNIVERSAL" SCENARIO: want to compress IID source, but do not know P

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

FIX: * block size

* a way to order the sets

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

$B(3,2)$

index	string
0	011
1	101
2	110

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

* Compute $k :=$ #zeros in x^N
* Determine index p of x^N in $B(N,k)$
* Return k and p in binary.

DECOMPRESSOR: clear !?

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

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

$$x^N \in T_{N,\varepsilon} \xRightarrow{\text{since}} B(n,k) \subseteq T_{N,\varepsilon} \Rightarrow \# B(n,k) \leq \# T_{N,\varepsilon} \quad (*)$$

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!

HW: Program this protocol & compress the chicken!

Discussion: Many disadvantages!

* Have to look at entire x^N to compress. Can we compress symbol by symbol?
* Assume IID distribution... what if P changes? Or if we have local correlations?

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

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

↳ next week ☺