Shannon Entropy & Data Compression → $\boxed{\text{IIT'19}}$ LECTURE NOTES

Last month: Formalism & toolbox of QIT. From now we will discuss IT prope...
... starting w/ classical data compression.

$P(\Sigma) := \{p : \Sigma \to \mathbb{R}_{\geq 0} \text{ prob. distribution}\}$. NOTATION: $X \sim p$ for RV X.

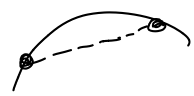
Shannon entropy of $p \in P(\Sigma)$: $\boxed{H(p) = \sum_x p(x) \log \frac{1}{p(x)}}$ BASE 2 $\quad 0 \cdot \log \frac{1}{0} = 0$

* $0 \leq H(p) \leq \log |\{x : p(x) > 0\}| \leq \log |\Sigma|$ * $H(p) = \mathbb{E}\left[\log \frac{1}{p(X)}\right]$ if $X \sim p$

$q \cdot \log \frac{1}{q} \geq 0$ Pf: Apply Jensen's inequality to concave log.

Jensen's inequality: $p \in P(\Sigma)$, $a \in \mathbb{R}^\Sigma$, f concave

$$\sum_x p(x) f(a(x)) \leq f\left(\sum_x p(x) a(x)\right)$$



If f strictly concave: "=" iff $\forall x, y$: $p(x)p(y) > 0 \Rightarrow a(x) = a(y)$

* $= 0$ iff p deterministic
  $= \log \#\Sigma$ iff uniform
* concave in p
  [$q \cdot \log \frac{1}{q}$ is concave]
* Subadditivity & monotonicity → $\boxed{\text{HW}}$

Today's goal: Show that $H(X)$ = optimal compression rate for IID source

Compression

Consider a data source modeled by a RV $X \sim p$. WANT:

$$X \longrightarrow \boxed{E} \longrightarrow \{0,1\}^\ell \longrightarrow \boxed{D} \longrightarrow \hat{X} = X$$

Raw bit content: $H_0(X) := H_0(p) := \log |\{x : p(x) > 0\}|$

* Can compress into ℓ bits $\Leftrightarrow$ $\ell \geq H_0(X)$ ☹

Pf: Need one bitstring for each possible x, i.e. $|\{0,1\}|^\ell \geq |\{x : p(x) > 0\}|$. □

How to do better? Two options:

Ⓐ Lossy Compression

Allow small probability of error: $\Pr(\hat{X} \neq X) \leq \delta$

e.g.

x	$p(x)$
A	0.98
B	0.01
C	0.01

$E(A) = 0$
$E(B) = 1$ $\Big\}\ \underline{\underline{\ell = 1}}$
$E(C) = $ arbitrary

$\hookrightarrow \Pr(\hat{X} \neq X) \leq \underline{\underline{0.01}} = \delta$

... but typically no significant saving for small δ.

PRACTICE ⟶ PROBLEM

Ⓑ Lossless Compression

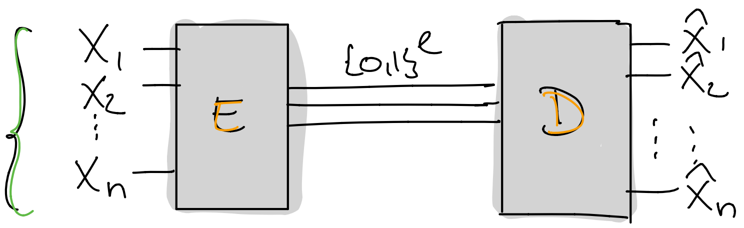
Use different length for different symbols & minimize average #bits

$E(A) = 0$
$E(B) = 10$
$E(C) = 11$

$\leadsto \bar{\ell} = 0.98 + 2 \cdot 0.02 = \underline{\underline{1.02}}$

Key idea: What if we compress blocks of symbols $X_1, X_2, \dots, X_n \overset{IID}{\sim} p$?

assume that IID



s.th.
$$\Pr(\hat{X}^n = X^n) \geq 1 - \delta \quad ?$$

NOTATION: $X^n = (X_1, \dots, X_n) = X_1 \cdots X_n$

Def: (n, R, δ)-code for $p \in \mathcal{P}(\Sigma)$: Functions

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

s.th. $\underbrace{\Pr(D(E(X^n)) = X^n)}_{} \geq 1 - \delta$ for $X^n \overset{IID}{\sim} p$

$$\sum_{x^n \in \Sigma^n,\ D(E(x^n)) = x^n} p(x_1) \cdots p(x_n) = \sum_{x^n \in \Sigma^n,\ D(E(x^n)) = x^n} p(x^n)$$ where $p(x^n) = p(x_1) \cdots p(x_n)$

Shannon's Source Coding Theorem: Let $0 < \delta < 1$:

① If $R > H(p)$: $\exists n_0$: $\forall n \geq n_0$: $\exists\ (n, R, \delta)$-code

② If $R < H(p)$: $\exists n_0$: $\forall n \geq n_0$: $\nexists\ (n, R, \delta)$-code

---

Thus: $H(P)$ is "optimal" compression rate for an IID source (independent of $0 < \delta < 1$!!!)

Why should be true? For "typical" x^n: $\#\{k : x_k = x\} \sim n \cdot p(x)$

$$\Rightarrow p(x^n) := p(x_1)\cdots p(x_n) \sim \prod_x p(x)^{np(x)} = 2^{-nH(p)}$$

i.e. typical strings have $\frac{1}{n}\log\frac{1}{p(x^n)} \approx H(p)$, so there should be $\sim 2^{nH(p)}$ many

Let's try to formalize this:

$\boxed{\text{Typical set:}}$

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

Properties:

⓪ $2^{-n(H(p)+\varepsilon)} \le p(x^n) \le 2^{-n(H(p)-\varepsilon)}$ (by definition)

① $|T_{n,\varepsilon}| \le 2^{n(H(p)+\varepsilon)}$

Pf: $1 \ge \Pr(X^N \in T_{N,\varepsilon}) \ge |T_{N,\varepsilon}| \cdot 2^{-n(H(p)+\varepsilon)}$ □

② $\Pr(X^N \notin T_{n,\varepsilon}) \le \frac{\sigma^2}{n\varepsilon^2} \to 0$, where $\sigma^2 = \mathrm{Var}\left(\log\frac{1}{p(X)}\right)$ for $X \sim p$

Pf: Let $R_k = \log\frac{1}{p(X_k)}$. Then: $R_1, \dots, R_n$ IID with mean $\mu = E[R_k] = H(X_k) = H(P)$.

$$\Rightarrow \Pr(X^N \notin T_{n,\varepsilon}) = \Pr\Big(\Big|\underbrace{\frac{1}{n}\sum_{k=1}^{n} R_k}_{\mathrm{Var} = \frac{1}{n^2} n\sigma^2 = \sigma^2/n} - \mu\Big| > \varepsilon\Big) \underset{\text{Chebyshev inequality}}{\le} \frac{\sigma^2}{n \cdot \varepsilon^2} \quad \square$$

PS: $\to 0$ is the (weak) law of large numbers!

"Asymptotic Equipartition Property" (AEP)

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

Proof of Shannon's theorem, part ①: Choose $\varepsilon = \frac{R - H(p)}{2} > 0$. Then:

$$|T_{n,\varepsilon}| \overset{①}{\le} 2^{n(H(p)+\varepsilon)} = 2^{n(R-\varepsilon)} \le 2^{\lfloor nR \rfloor} \quad \text{for } n \ge \frac{1}{\varepsilon}$$

$\Rightarrow \exists$ injective $E: T_{n,\varepsilon} \to \{0,1\}^{\lfloor nR \rfloor}$ w/ left inverse $D: \{0,1\}^{\lfloor nR \rfloor} \to T_{n,\varepsilon}$

Extend arbitrarily to Σ^n. Then:

$$\Pr(D(E(X^n)) \neq X^n) \leq \Pr(X^n \notin T_{n,\varepsilon}) \overset{(2)}{\leq} \frac{\sigma^2}{n \cdot \varepsilon^2} \longrightarrow 0,$$

hence $\leq \delta$ for n large enough. $\square$

<u>Remark:</u> $T_{n,\varepsilon}$ 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 $n \to \infty$!

How to use this in practice? $\boxed{\text{EX}}$

SCENARIO: want to compress IID (memoryless) data source P
(we <u>know</u> p, but NOT which samples will be emitted)

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

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

ENCODER: <u>Input</u>: A string $x^n = x_1 \cdots x_n$

* If $x^n \notin T_{n,\varepsilon}$: Return $0 \ldots 0$ (or fail).
* Determine index k of x^n in $T_{n,\varepsilon}$.
* Return k in binary.

DECODER

<u>Input</u>: A binary string s

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

Discussion $\boxed{\text{EX}}$

* How to make it **LOSSLESS**? Send atypical x^n uncompressed!

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

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

* Disadvantages:
  * Need to see <u>all</u> of x^n before we can start compressing.
  * <u>IID</u> assumption — what if p changes? what if local correlations?