CO781 / QIC890 Fall 2016 Lec 6.

Topic 2: Entropy & data compression.

Reference: NC Sec 12.2, Preskill Sec 10.1.1, 10.3, 10.4. Cover & Thomas

Goal: part 1 of quantifying info & redundancy.

- X: random variable
- Ω: sample space, $|\Omega| = m$ = # outcomes of X

} X takes values in Ω

- p: prob distribution of X

$$p: \Omega \to [0,1]$$
$$x \mapsto p(x)$$

s.t. $\sum_{x \in \Omega} p(x) = 1$.

capital = rv
lowercase = outcomes

eg. biased coin

$\Omega = \{0,1\}$, $p(0) = 0.1$, $p(1) = 0.9$.

- A discrete "information source" is a sequence of rv's $X_1, X_2, X_3, \ldots$ with a common sample space Ω (source alphabets).

(n draws have m^n outcomes)

eg. can toss the above biased coin as many times as wished

eg. Weather every day, $\Omega = \{$sun, cloud, rain, snow$\}$

In general X_i need not be independent, or identically distributed.

- If X_i's are independent and identically distributed, we call $X_1 X_2 \ldots$ an "iid source".

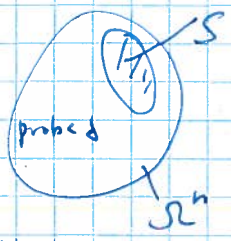
(*)

Typicality & the asymptotic equipartition thm:

Idea: Consider $X^n = X_1 X_2 \dots X_n$.

For large n, $\exists$ large probability subset S in Ω^n with low cardinality.

Furthermore, elements of S have similar probabilities.



Why?

Consider an arbitrary $x^n = x_1 x_2 \dots x_n$

$p(x^n) = p(x_1)\,p(x_2) \cdots p(x_n)$ (by independence)

$= 2^{\log p(x_1)}\, 2^{\log p(x_2)} \cdots 2^{\log p(x_n)}$ (log base 2)

$= 2^{-n\left[\frac{1}{n}\sum_{i=1}^{n}(-)\log p(x_i)\right]}$

($\frac{1}{n}\sum_{i=1}^{n}(-)\log p(x_i)$: empirical average of $-\log p(x)$)

$\downarrow$ LLN

$2^{-n\left[-\sum_{x\in\Omega} p(x)\log p(x)\right]}$ ($\mathbb{E}_p[-\log p(x)] =: H(x)$)

So as $n \to \infty$, $p(x^n) \to 2^{-nH(x)}$. "typical"

If $p(x^n) \approx 2^{-nH(x)}$, then $x^n \in S$.

Def: [Shannon Entropy]

$H(X)$ or $H(p) := -\sum_{x \in \Omega} p(x) \log p(x)$.

eg. for the biased coin,

$H(X) = -0.1 \log 0.1 - 0.9 \log 0.9 = 0.469$

Def: [typical sequence]

x^n is δ-typical if $\left| -\frac{1}{n} \log p(x^n) - H(X) \right| \le \delta$ $\quad \left(p(x^n) \approx 2^{-nH(X)} \right)$

Def: [typical set]

$T_{n,\delta} = \{ x^n : x^n \text{ is } \delta\text{-typical} \}$.

eg. For the biased coin, $n = 100$, $\delta = 0.1$

If there are t 0's & $n-t$ 1's,

then $-\frac{1}{n} \log p(x^n) = -\frac{t}{n} \log 0.1 - \frac{n-t}{n} \log 0.9$

and $-\frac{1}{n} \log p(x^n) \in [0.369, 0.589]$ for $7 \le t \le 13$.

So $T_{100, 0.1}$ = all 100-bit strings with 7 to 13 0's.

Idea: $T_{n,\delta}$ is a large prob set with low cardinality,

because $-\frac{1}{n} \log p(x^n) \to H(X)$ as $n \to \infty$.

eg. Prob $(T_{100, 0.1}) = 0.75897$

$|T_{100,0.1}| = 8.3 \times 10^{15}$

$|\Omega^{100}| = 1.3 \times 10^{30}$

$\Big\}$ $\dfrac{|T_{100,0.1}|}{|\Omega^{100}|} \ll 1$.

Asymptotic Equipartition Thm (AEP):

$\forall \varepsilon > 0 \;,\; \forall \delta > 0 \;,\; \exists n_0$ s.t. $\forall n \geq n_0$

① $p(|T_{n,\delta}|) \geq 1 - \varepsilon$

② $(1-\varepsilon)\, 2^{n(H(X)-\delta)} \leq |T_{n,\delta}| \leq 2^{n(H(X)+\delta)}$

③ $\forall A \subseteq \Omega^n, \; p(A) \geq 1-\varepsilon \Rightarrow |A| \geq (1-2\varepsilon)\, 2^{n(H(X)-\delta)}$

NB: ①: large prob, ② small size, and how small ③ large prob set can't be smaller

Bonus: $T_{n,\delta} \approx$ equiprobable by definition

See Preskill for full motivating example for the biased coin.

Pf ① We apply LLN on $Y = \log p(X)$, an induced r.v s.t $\forall x \in \Omega$, $Y = \log p(x)$ wp $p(x)$.

So $\mathbb{E}Y = \sum_x p(x) \log p(x) = -H(X)$.

Since X^n is i.i.d, so is $Y^n = Y_1 \cdots Y_n$.

Let $x^n = x_1 x_2 \ldots x_n \in \Omega^n$, $y_i = \log p(x_i)$

Then $x^n \notin T_{n,\delta} \iff \left| \frac{1}{n} \sum_{i=1}^n y_i - \mathbb{E}Y \right| > \delta$ (*)

To bound the prob of (*), recall Chebyshev's ineq for a r.v Z:

$$\Pr\{ |z - \mathbb{E}Z| \geq K \sqrt{\mathrm{Var}\, Z} \} \leq \frac{1}{K^2}.$$

(z: outcome, Z: rv)

Choose • $Z = \frac{1}{n} \sum_{i=1}^n Y_i$.

• $n_0 = \frac{\mathrm{Var}\, Y}{\delta^2 \varepsilon}$ so if $n \geq n_0$, $\mathrm{Var}\, Z = \frac{1}{n} \mathrm{Var}\, Y \leq \frac{1}{n_0} \mathrm{Var}\, Y = \delta^2 \varepsilon$

• $K = \frac{\delta}{\sqrt{\mathrm{Var}\, Z}}$ so $K \sqrt{\mathrm{Var}\, Z} = \delta$.

So Chebyshev's ineq gives:

$$\Pr\{ \left| \tfrac{1}{n} \textstyle\sum_{i=1}^n y_i - \mathbb{E}Y \right| \geq \delta \} \leq \frac{\mathrm{Var}\, Z}{\delta^2} \leq \varepsilon \quad (**)$$

(z = $\frac{1}{n} \sum y_i$, $\mathbb{E}Z$ = $\mathbb{E}Y$, δ^2 = K^{-2})

dropping "=" decreases Pr

$\therefore$ Prob $(x \notin T_{n,\delta}) = \Pr\{ \left| \frac{1}{n} \sum_{i=1}^n y_i - \mathbb{E}Y \right| > \delta \} \leq \varepsilon$

(* = left, ** = right)

$\therefore \Pr(T_{n,\delta}) > 1 - \varepsilon$.

Pf ② $1-\varepsilon \overset{①}{\leq} p(T_{n,\delta}) \leq 1$

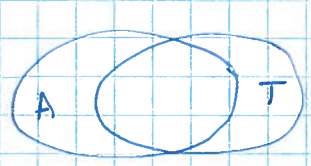
$$|T_{n,\delta}|\, 2^{-n(H(X)+\delta)} \leq \sum_{x^n \in T_{n,\delta}} p(x^n) \leq |T_{n,\delta}| \cdot 2^{-n(H(X)-\delta)}$$

$\min p(x^n)$; $\max p(x^n)$ for $x^n \in T_{n,\delta}$

$$\therefore\ |T_{n,\delta}| \leq 2^{n(H(X)+\delta)} \quad \& \quad (1-\varepsilon)\, 2^{n(H(X)-\delta)} \leq |T_{n,\delta}|$$

NB: $\dfrac{|T_{n,\delta}|}{|\Omega^n|} \leq 2^{-n\underbrace{[\log|\Omega| - H(X) - \delta]}_{+ve \text{ for most } X}}$ ← exp decaying in n.

Pf ③ Let $A \subseteq \Omega^n$ have $p(A) > 1-\varepsilon$, $T = T_{n,\delta}$.



$$\begin{aligned} P(A \cap T) &= P(A) - P(A \setminus T) \\ &\geq P(A) - P(\Omega \setminus T) \\ &\geq 1 - \varepsilon - \varepsilon \\ &= 1 - 2\varepsilon. \end{aligned}$$

$A \cap \Omega$; $P(\Omega \setminus T) \leq \varepsilon$

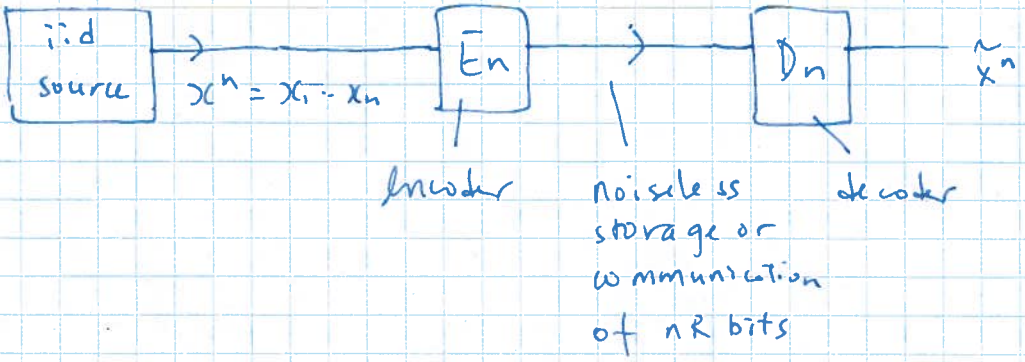
$$|A| \geq |A \cap T| \geq \frac{P(A \cap T)}{\max_{a \in A \cap T} p(a)} \geq \frac{1-2\varepsilon}{2^{-n(H(X)-\delta)}}$$

$\geq$: uniformity

Application: data compression of i.i.d sources.

Aka: Shannon's noiseless coding theorem.

Setting:



Goal: min R while keeping prob $(\tilde{x}^n \neq x^n)$ negligible.

Idea: transmit only typical sequences & ignore the rest

Simplest "code book" or E_n:

Assign to each $x^n \in T_{n,\delta}$ a unique label of $n(H(x)+\delta)$ bits.

$$E_n: x^n \longmapsto b(x^n) \quad \text{if } x^n \in T_{n,\delta}.$$
$$x^n \longmapsto \text{ERR} \quad \text{otherwise}$$

The decoder D_n simply invert b if the received message $\neq$ ERR

b = bijection agreed upon by Alice & Bob.

$$\Pr(\tilde{x}_n \neq x_n) = \Pr(x_n \notin T_{n,\delta}) \leq \varepsilon \qquad \forall n > n_0 = \frac{\mathrm{Var}[\log p(x)]}{\delta^2 \varepsilon}$$

Shannon's noiseless coding theorem:

Summary: compression rate R
$= \begin{cases} H(X) + o(1) & \text{achievable} \\ H(X) - \Omega(1) & \text{not.} \end{cases}$

Let $X_1 X_2 \ldots$ be an iid source.

① $\forall \varepsilon > 0 \quad \forall R > H(X)$

$\exists n_0$ s.t. $\forall n \geq n_0 \quad \exists E_n, D_n$

s.t. $\Pr\left(D_n \circ E_n(x^n) \neq x^n \right) < \varepsilon$.

② $\forall R < H(X)$

$\exists n_0$ s.t. $\forall n \geq n_0 \quad \forall E_n, D_n$

$$\Pr\left(D_n \circ E_n(x^n) = x^n \right) \leq \varepsilon + 2^{-n\left[\frac{H(X)-R}{2}\right]}$$

Pf ①: The E_n, D_n described in "Idea" work.

Pf ②: WLOG, E_n is deterministic.

(Otherwise, E_n succeeds w/ high prob over the random coins and $\exists$ coin values r for which E_n succeeds whp. Replace E_n by its restriction to r.)

So 2^{nR} symbols are encoded, the rest turned to ZERO.

Let A = these 2^{nR} symbols, $\delta = (H(X)-R)/2 > 0$, $T = T_{n,\delta}$

$$\begin{aligned} \text{Prob}(A) &= \text{Prob}(A \setminus T) + \text{Prob}(A \cap T) \\ &\leq \varepsilon + |A| \max_{x^n \in T} P(x^n) \\ &\leq \varepsilon + 2^{nR} \cdot 2^{-n(H(X)-\delta)} \\ &= \varepsilon + 2^{-n\left[(H(X)-R)-\delta\right]} \\ &= \varepsilon + 2^{-n\left[\frac{H(X)-R}{2}\right]} \end{aligned}$$

Comments:

- Allowing an arbitrarily small error ε reduces the compression rate from $n \log|\mathcal{X}|$ to $nH(X)$ bits.

- Note wp $1-\varepsilon$, the ENTIRE x^n correct.

- Data compression gives $H(X)$ an operational meaning.
  - How much space needed to rep each symbol ASYMPTOTICALLY. ↑ larger limit
  - How uncertain a symbol is.

- We considered "block codes" where n is fixed.
  We were not concerned about complexity of E_n, D_n.
  See Cover & Thomas for other codes.
  Ex: Hoffmann code is exactly but variable length with bruteforce E_n / D_n. Same EXPECTED rate.

- There are universal compression algorithm that requires only an upper bound on $H(X)$ without knowing what X is!
  (Generalizations useful in Q setting.)