QIC 820 / CO781 / CO 486 / CS 867 :

Theory of Quantum Information

Part 3, lecture 1

The asymptotic equipartition theorem, Shannon entropy and classical data compression

References:

Cover & Thomas: Chapter 3
Preskill Physic 219 lecture notes: Chapter 10.1.1
(April 2022 version)

What is uncertainty?

What is information?

What is redundancy?

How to quantify them?

X : random variable

Ω : sample space, $|\Omega| = m$

p : prob distribution of X

$$
\begin{aligned}
p : \Omega &\longrightarrow [0,1] \\
x &\longmapsto p(x)
\end{aligned}
$$

upper case: rv

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

lower case: outcome

e.g., biased coin

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

A "discrete information source" is a sequence of rvs X1, X2, X3, ... with a common sample space / source alphabet Ω.

e.g., can toss the biased coin as many times as wished

e.g., weather each day, Ω = {sun, cloud, rain}

With n draws, we get one out of m^n outcomes.

In general, the X_i's need not be independent or identically distributed. (e.g., weather)

If X_i's are independent and identically distributed, we call X1, X2, ... an "iid" source.

Focus on iid sources rest of this lecture.

Better than magic for iid sources:
-- typicality and asymptotic equipartition thm

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

For large n, $\exists$ a subset $S \subseteq \Omega^n$ with

(1) high prob, (2) low cardinality, (3) ~equiprobable elements

Why? Consider any $x^n = x_1 x_2 \cdots x_n$

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

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

$$= 2^{-n\left[\frac{1}{n}\sum_{i=1}^{n} (-)\log p(x_i)\right]}$$ empirical average of $(-)\log p(x)$ over n samples

$\downarrow$ LLN

$$= 2^{-n\left[\sum_{x\in\Omega} p(x)(-)\log p(x)\right]}$$ $\leftarrow \underset{p}{\mathbb{E}}\,(-)\log p(x) =: H(X)$ theoretical average

As $n \to \infty$, $p(x^n) \to 2^{-nH(X)}$, such x^n "typical".

Def: [Shannon entropy] H(X) or H(p) := $-\sum_{x \in \Omega} p(x) \log p(x)$

e.g., for 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 $\left| -\frac{1}{n} \log p(x^n) - H(X) \right| \leq \delta$

$\left(p(x^n) \approx 2^{-nH(X)} \right)$

Def: [typical set] $T_{\delta,n} = \{ x^n : x^n \text{ is } \delta\text{-typical} \}$

e.g., for biased coin, n = 100, δ = 0.1

if x^n has t 0's & $n-t$ 1's

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

$\in$ [0.369, 0.569] for t = 7, 8, ..., 13

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

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

e.g., $\text{Prob}(T_{100,0.1}) = 0.75897$

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

$$|\{0,1\}^{100}| = 1.3 \times 10^{30}$$

$$\frac{|T_{100,0.1}|}{|\{0,1\}^{100}|} \approx 6 \times 10^{-15}$$

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

Interpretations:

(1) says the typical set is a large prob set

(2) quantifies how small the typical set is

(3) says any large prob set can't be much smaller

Bonus: within typical set, elements are ~ equiprobable

(See Preskill for full motivating example for biased coin.)

Asymptotic equipartition theorem (AEP)

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

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

Proof: we upper bound $pr(x^n \notin T_{n,\delta})$

X induces a rv Y = log p(X)

i.e., $\forall x \in \Omega$, wp p(x), Y = log p(x)

$$\therefore \mathbb{E} Y = \sum_{x \in \Omega} p(x) \log p(x) = -H(X)$$

$\because X^n$ iid, so is $Y^n = Y_1 Y_2 \cdots Y_n$.

For $x^n = x_1 x_2 \cdots x_n$, let $y_i = \log p(x_i)$

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

use LLN on Y to bound prob of this

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

use LLN on Y to bound prob of this

By Chebyshev's inequality for a rv Z :

$$\Pr\left\{ \left| z - \mathbb{E}Z \right| \geq k\sqrt{\operatorname{Var} Z} \right\} \leq \frac{1}{k^2} \qquad (\text{rv } Z, \text{ outcome } z)$$

Choose $Z = \frac{1}{n}\sum_{i=1}^{n} Y_i$ $\left(\text{so } z = \frac{1}{n}\sum_{i=1}^{n} y_i,\ \mathbb{E}Z = \mathbb{E}Y,\ \operatorname{Var} Z = \frac{\operatorname{Var} Y}{n} \right)$

$k = \frac{\delta}{\sqrt{\operatorname{Var} Z}}$ $\left(\text{so } k\sqrt{\operatorname{Var} Z} = \delta \right)$

$n_0 = \left\lceil \frac{\operatorname{Var} Y}{\delta^2 \varepsilon} \right\rceil$ $\left(\text{so } \forall\, n \geq n_0,\ \frac{1}{k^2} = \frac{\operatorname{Var} Z}{\delta^2} = \frac{\operatorname{Var} Y}{n\,\delta^2} \leq \varepsilon \right)$

$$\therefore \Pr\left\{ \left| \frac{1}{n}\sum_{i=1}^{n} y_i - \mathbb{E}Y \right| \geq \delta \right\} \leq \varepsilon .$$

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

Asymptotic equipartition theorem (AEP)

$\forall \varepsilon > 0, \ \forall \delta > 0, \ \exists n_0 \text{ s.t. } \forall n \geq n_0$

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

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

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

$$\parallel$$

max $p(x^n)$ if $x^n \in T_{n,\delta}$

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

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

In particular, $\dfrac{|T_{n,\delta}|}{|\Omega^n|} \leq 2^{-n\underbrace{(\log|\Omega| - H(X) - \delta)}_{\text{+ve for most } X}}$

exponentially decreasing in n

Asymptotic equipartition theorem (AEP)

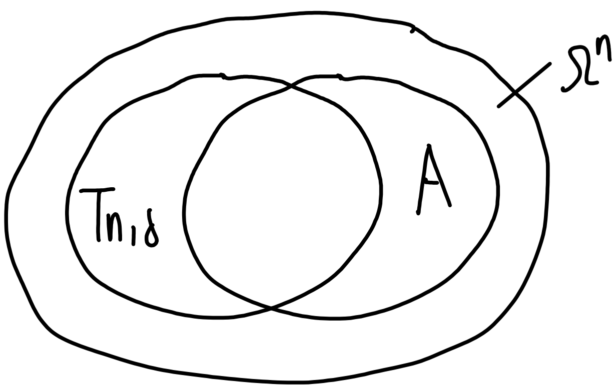
$\forall \varepsilon > 0,\ \forall \delta > 0,\ \exists n_0$ s.t. $\forall n \geq n_0$

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

Proof:

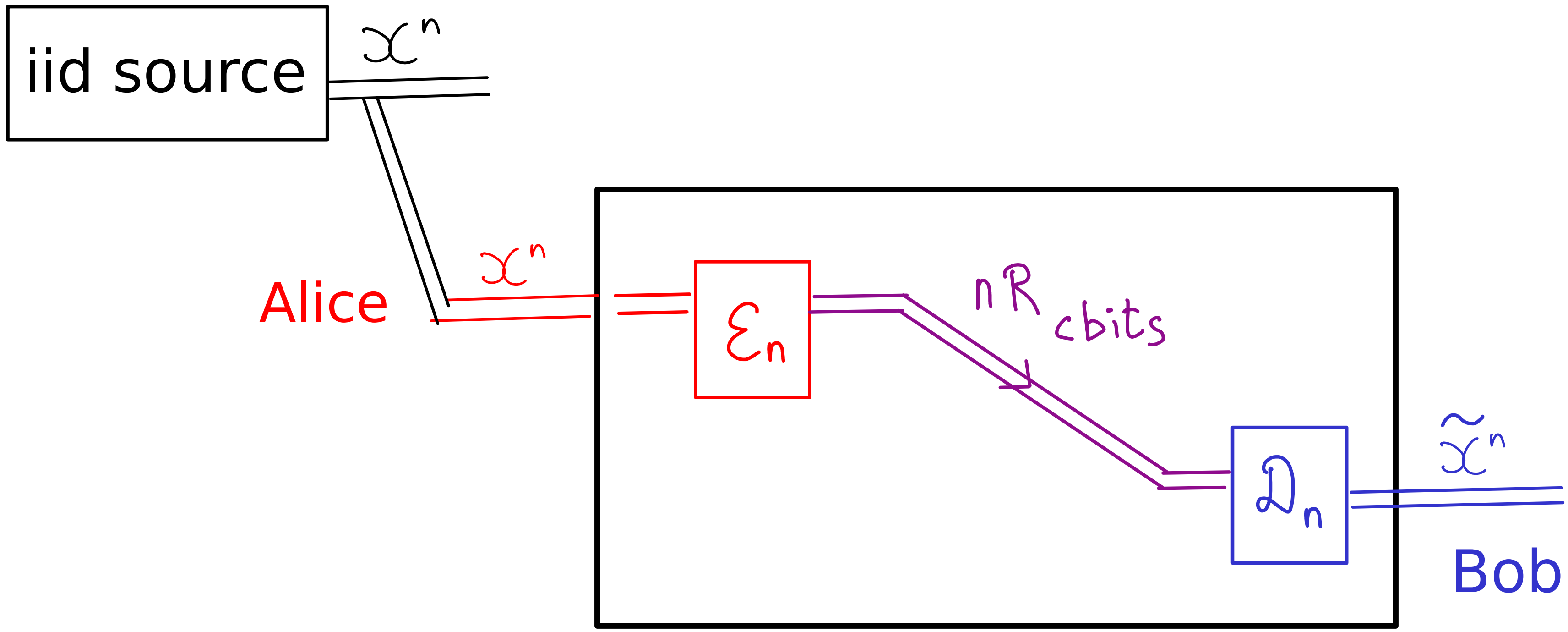


$$
\begin{aligned}
& p(A \cap T_{n,\delta}) \\
&= p(A) - p(A \setminus T_{n,\delta}) \\
&\geq p(A) - p(\Omega^n \setminus T_{n,\delta}) \\
&\geq 1-\varepsilon - \varepsilon = 1-2\varepsilon
\end{aligned}
$$

$$\therefore |A| \geq |A \cap T_{n,\delta}| \geq \frac{P(A \cap T_{n,\delta})}{\max\limits_{a \in A \cap T_{n,\delta}} p(a)} \geq \frac{1-2\varepsilon}{2^{-n(H(X)-\delta)}}$$

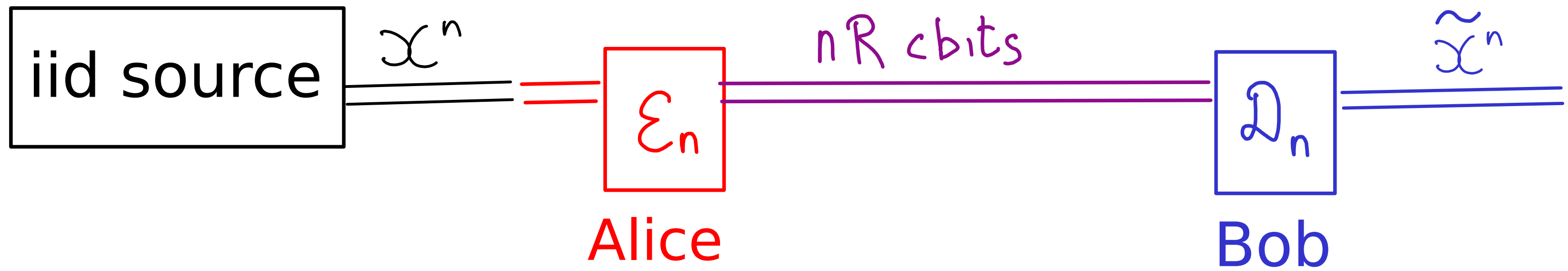
typicality

Application: data compression of iid sources



(if Bob = future Alice, nR cbits refer to storage space)

Application: data compression of iid sources



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

Shannon's noiseless coding theorem:

Let $X_1, X_2, \ldots, X_n$ be iid source

direct coding theorem - we can do ...

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

$\exists n_0$ s.t. $\forall n \geq n_0 \quad \exists \mathcal{E}_n, \mathcal{D}_n$

s.t. $\Pr(\mathcal{D}_n \circ \mathcal{E}_n(x^n) \neq x^n) \leq \varepsilon$

converse - cannot do better

② $\forall R < H(X)$

$\exists n_0$ s.t. $\forall n \geq n_0 \quad \forall \mathcal{E}_n, \mathcal{D}_n$

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

Proof of (1):

Idea: transmit only typical sequences, ignore the rest

For each $x^n \in T_{n,\delta}$,

let $b(x^n)$ be unique $n(H(X)+\delta)$ bit label for x^n

$$\mathcal{E}_n : \begin{cases} x^n \mapsto b(x^n) & \text{if } x^n \in T_{n,\delta} \\ x^n \mapsto \text{err} & \text{otherwise} \end{cases}$$

preagreed by Alice and Bob

$\mathcal{D}_n$: invert b if r not receive err
else output err

$$\Pr(\mathcal{D}_n \circ \mathcal{E}_n(x^n) \neq x^n) = \Pr(x^n \notin T_{n,\delta}) \leq \varepsilon$$

$$\text{for } n \geq n_0 = \frac{\mathrm{Var}[\log p(X)]}{\delta^2 \varepsilon}$$

Proof of (2):

By C2, at most 2^{nR} x^n's satisfies $\mathcal{D}_n \circ \mathcal{E}_n(x^n) = x^n$.

Let A = set of x^n's s.t. $\mathcal{D}_n \circ \mathcal{E}_n(x^n) = x^n$, $|A| \leq 2^{nR}$.

Let $\delta = \frac{1}{2}(H(X) - R) > 0$, $T = T_{n,\delta}$.

$$
\begin{aligned}
P(A) &= P(A \setminus T) + P(A \cap T) \\
&\leq \varepsilon + |A| \max_{x^n \in T} p(x^n) \\
&\leq \varepsilon + 2^{nR} \cdot 2^{-n(H(X)-\delta)} \\
&= \varepsilon + 2^{-n(H(X)-R-\delta)} \\
&= \underbrace{\varepsilon}_{\text{arbitrarily small as } n \uparrow} + \underbrace{2^{-n(H(X)-R)/2}}_{\exp \downarrow \text{ in } n}
\end{aligned}
$$

Comments:

* Allowing an arbitrarily small error ε reduces the compression cost from log $|\Omega|$ to H(X) cbits per symbol

* wp $1-\varepsilon$ the ENTIRE x^n correct !!

* data compression gives H(X) an operational meaning.
- how much space is needed to represent each symbol asymptotically (large n limit)?
- how much uncertainty is associated with each symbol?

* We considered "block codes" where n is fixed.

* We are not concerned about the computational complexity of $\mathcal{E}_n$, $\mathcal{D}_n$.

See Cover and Thomas for other codes, e.g., Hoffman code is exact, but variable-length, with expectation H(X) per symbol.