Information Theory & Coding

Sept 29, 2020

Recall:

$$H(U) \le E[\text{best codeword length}] \le H(U)+1$$

$$\Rightarrow \frac{1}{n}H(U_1 \dots U_n) \le \underbrace{E\left[\frac{1}{n}\text{length } \hat{c}_n(U_1 \dots U_n)\right]}_{\#\text{ of bits/letter}} \le \frac{1}{n}H(U_1 \dots U_n) + \frac{1}{n}$$

Def: given a source $U_1 U_2 U_3 \dots$ (a stochastic process), we say the entropy-rate of the source is

$$\lim_{n\to\infty} \frac{1}{n} H(U_1 \dots U_n) =: \mathcal{H}(\{U_i\})$$

if the limit exists.

with this definition:

$$\lim_{n\to\infty} \frac{1}{n} E(\text{length } \hat{c}_n(U_1 \dots U_n)) = \mathcal{H}(\{U_i\})$$

Ex: if U_1, U_2, U_3 are i.i.d then

$$\mathcal{H}(\{U_i : i \in \mathbb{N}\}) = H(U_1)$$

chain rule

Pf: $H(U_1 \ldots U_n) = H(U_1) + H(U_2 | U_1) + \cdots$

$$+ H(U_n | U_1 \ldots U_{n-1})$$

$$= H(U_1) + H(U_2) + \cdots + H(U_n)$$

independence

ident dist

$$= n H(U_1)$$

$$\Rightarrow \frac{1}{n} H(U_1 \ldots U_n) = H(U_1) \quad //$$

Def: A stochastic process $U_1 U_2 \ldots$ is said to be stationary if for every $n \geq 1$ & every $k \geq 1$ and every $u_1 \ldots u_n$

$$\Pr(U_1 \ldots U_n = u_1 \ldots u_n)$$

$$= \Pr(U_{1+k}, U_{2+k} \ldots U_{n+k} = u_1 \ldots u_n).$$

Thm: if $U_1 U_2 U_3 \ldots$ is a stationary process then

$\lim_{n\to\infty} \frac{1}{n} H(U_1 \ldots U_n)$ exists and

equals $\boxed{\lim_{n\to\infty} H(U_n | U_1 \ldots U_{n-1})}$

Pf: Let $a_n = H(U_n | U_1 \ldots U_{n-1})$. 1st claim

$0 \le a_{n+1} \le a_n$. which will imply that

$a = \lim_n a_n$ exists. To show the claim:

$a_{n+1} = H(U_{n+1} | U_n U_{n-1} \ldots U_1)$

$\le H(U_{n+1} | U_n \ldots U_2)$ (cond. reduces entropy)

$= H(U_n | U_{n-1} \ldots U_1) = a_n$

$(U_1 \ldots U_n) \sim (U_2 \ldots U_{n+1})$ //

has the same stats as (stationarity).

Now observe

$$S_n \stackrel{\Delta}{=} \frac{1}{n} H(U_1 \dots U_n)$$

$$= \frac{1}{n}\left[H(U_1) + H(U_2|U_1) + \dots + H(U_n|U_1 \dots U_{n-1})\right]$$

$$= \frac{1}{n}\left[a_1 + a_2 \quad \dots \quad + a_n\right]$$

Fact: (Césaro): if x_1, x_2, x_3 is a $\mathbb{R}$-valued sequence with $x = \lim_{n\to\infty} x_n$ then

$y_n \stackrel{\Delta}{=} \frac{1}{n}(x_1 + \dots + x_n)$ also has a limit &

$$\lim_{n\to\infty} y_n = x.$$

All we need to do is to prove Césaro:

"$\lim_{n\to\infty} x_n = x$" $\equiv$ "$\forall \varepsilon > 0 \ \exists n_0(\varepsilon) \ \forall n \geq n_0(\varepsilon)$
$|x_n - x| < \varepsilon$".

We need to show that $\lim_n y_n = x$

$$y_n - x = \frac{1}{n}\left[(x_1 - x) + \dots + (x_n - x)\right]$$

$$|y_n - x| \leq \frac{1}{n}\left[|x_1 - x| + \dots + |x_n - x|\right]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n_0(\varepsilon)} |x_i - x| + \sum_{i=n_0(\varepsilon)+1}^{n} \overbrace{|x_i - x|}^{<\varepsilon}\right] \qquad n \geq n_0(\varepsilon)$$

$$\leq \varepsilon + \frac{1}{n} \sum_{i=1}^{n_0(\varepsilon)} |x_i - x|$$

for $n \geq \max\left\{ n_0(\varepsilon), \frac{\sum_{i=1}^{n_0(\varepsilon)} |x_i - x|}{\varepsilon} \right\} =: n_1(\varepsilon)$.

we have

$$|y_n - x| < \varepsilon + \varepsilon = 2\varepsilon.$$

$\Rightarrow \lim y_n = x.$ //.

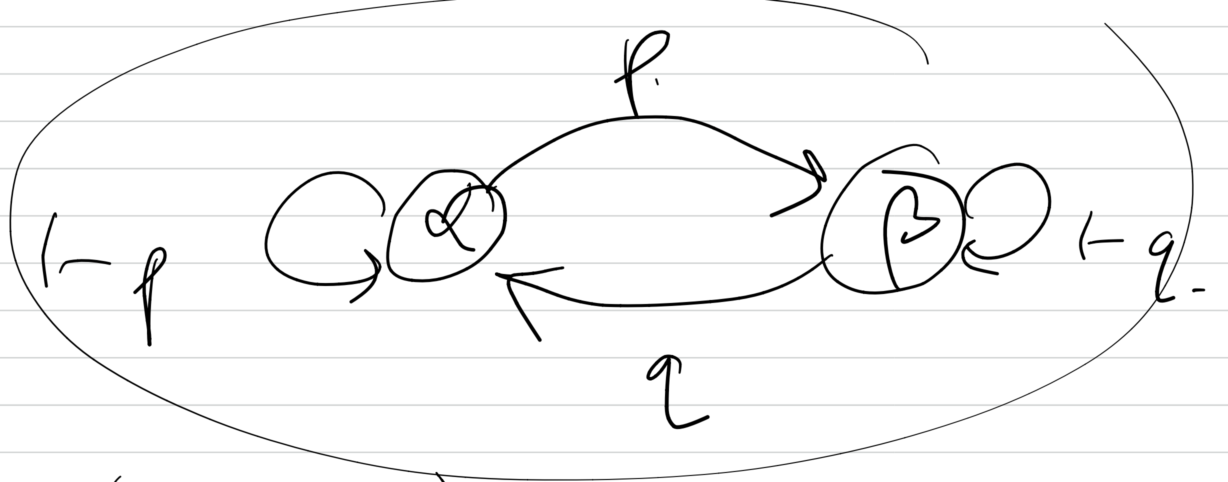
Ex: suppose $U_1, U_2, U_3 \ldots$ is a Markov process, & stationary; i.e.

$$\left.\begin{aligned} &p(U_n = u_n \mid U_{n-1} = u_{n-1}, \ldots U_1 = u_1) \\ &= \Pr(U_n = u_n \mid U_{n-1} = u_{n-1}). \end{aligned}\right\} \text{ Markovity.}$$

Then

$$\begin{aligned} \mathcal{H}(\{U_i : i \in \mathbb{N}\}) &= \lim_{n\to\infty} H(U_n \mid U_{n-1} \ldots U_1) \\ &= \lim_{n\to\infty} H(U_n \mid U_{n-1}) \\ &= \lim_{n\to\infty} H(U_2 \mid U_1) \\ &= H(U_2 \mid U_1). \end{aligned}$$

Example: $U_i \in \{\alpha, \beta\}$.



$\Pr(U_{n+1} = \beta \mid U_n = \alpha) = p$

$\Pr(\quad = \alpha \mid \quad = \beta) = q$

$$\underbrace{\Pr(U_{n+1} = \alpha)}_{\pi(\alpha)} = \Pr(U_n = \alpha)(1-p) + \underbrace{\Pr(U_n = \beta)}_{(1 - \Pr(U_n = \alpha))} q$$

$$\pi(\alpha) = \pi(\alpha)(1-p) + (1-\pi(\alpha))q$$

$$\pi(\alpha) = Pr(u_n = \alpha) = Pr(u_{n+1} = \alpha) = \cdots = Pr(u_1 = \alpha)$$

$$\pi(\alpha) = \frac{q}{p+q}, \quad \pi(\beta) = \frac{p}{p+q}.$$

$$\Rightarrow H(u_2 | u_1) = \frac{q}{p+q} \underbrace{H(u_2 | u_1 = \alpha)}$$

$$+ \frac{p}{p+q} H(u_2 | u_1 = \beta)$$

$$= \frac{q}{p+q}\left[p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p} \right)$$

$$+ \frac{p}{p+q}\left[q\log\frac{1}{q} + (1-q)\log\frac{1}{1-q} \right)$$

//

Back to where we left off yesterday: Remember

we were considering iid processes we had defined the notion of typicality:

Def: given a distribution p on $\mathcal{U}$, $\varepsilon > 0$, $n \in \{1,2,3\}$

we say a sequence $u_1 u_2 - u_n$ to be ε-typical w.r.t p if

$$\forall u \in \mathcal{U} \quad \frac{\#\{i : u_i = u\}}{n} \in \left[(1-\varepsilon)p(u), (1+\varepsilon)p(u)\right].$$

We let $T(n,\varepsilon,p)$ = set all ε-typical sequences wrt p of length n.

① We had seen that if $(u_1 \dots u_n) \in T(n,p,\varepsilon)$

$$\Pr(\underbrace{U_1 \dots U_n}_{\text{iid} \sim p} = u_1 \dots u_n) = 2^{-n H(U)[1 \pm \varepsilon]}$$

② $$1 \geq \Pr(\underbrace{(U_1 \dots U_n)}_{\text{iid} \sim p} \in T(n,p,\varepsilon))$$

$$= \sum_{(u_1 \dots u_n) \in T(\ )} \Pr(U_1 \dots U_n = u_1 - u_n) \geq \sum 2^{-n H(U)(1+\varepsilon)}$$

$$= |T(n,p,\varepsilon)| \cdot 2^{-n H(U)(1+\varepsilon)}$$

$$\Rightarrow \left(|T(n,p,\varepsilon)| \leq 2^{n H(U)(1+\varepsilon)} \right)$$

$\sim p.$

(3). We will show

$$\Pr((U_1 \ldots U_n) \in T(n, p, \varepsilon)) \approx 1.$$

Lemma: Fix p, $\varepsilon > 0$, then

$$\lim_{n \to \infty} \Pr((\underbrace{U_1 \ldots U_n}_{\text{iid} \sim p}) \in T(n, p, \varepsilon)) = 1.$$

Pf: $\{(U_1 \ldots U_n) \notin T(n, p, \varepsilon)\}$

$$= \bigcup_{u \in \mathcal{U}} \left\{ |\{i : U_i = u\}| \frac{1}{n} \notin [(1-\varepsilon)p(u), (1+\varepsilon)p(u)] \right\}.$$

$$\Pr((U_1 \ldots U_n) \notin T) \leq \sum_{u \in \mathcal{U}} \Pr(\downarrow)$$

We will show that for every $u \in \mathcal{U}$

$$\Pr\left(\frac{1}{n}|\{i : U_i = u\}| \notin [(1-\varepsilon)p(u), (1+\varepsilon)p(u)]\right)$$

$\to 0$ as n gets large

To do so, let $X_i = \begin{cases} 1 & U_i = u \\ 0 & U_i \neq u \end{cases}$

$$\Rightarrow \frac{1}{n}|\{i : U_i = u\}| = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Observe $X_1, X_2 \ldots X_n$ are iid.

$$E(X_1) = p(u), \quad Var(X_1) = p(u) - p(u)^2 = p(u)(1 - p(u)).$$

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - p(u)\right| > \varepsilon p(u)\right)$$

Recall Chebychev's Inequality:

$$\left(\Pr\left(|Y - E(Y)| > \alpha\right) \le \frac{Var(Y)}{\alpha^2}\right)$$

$$\le \frac{n\, Var(X_1)}{n^2 \varepsilon^2 p(u)^2} = \frac{Var(X_1)}{n\, \varepsilon^2 p(u)^2}$$

$$\Rightarrow \lim_{n\to\infty} \Pr\left(\frac{1}{n}\left|\{i : U_i = u\}\right| \notin \left[(1-\varepsilon)p(u), (1+\varepsilon)p(u)\right]\right) = 0.$$

//.

we have proved $\forall \varepsilon > 0, \exists n_0(\varepsilon)$ s.t. $\forall n \geq n_0(\varepsilon)$

$$\Pr(\underbrace{(U_1 \ldots U_n)}_{\sim \text{iid } p} \in T(n, p, \varepsilon)) > 1 - \varepsilon.$$

Corollary: for n large enough

$$(1-\varepsilon) < \Pr(\underset{\text{iid} \sim p}{(U_1 \ldots U_n)} \in T(n,p,\varepsilon)) = \sum_{(u_1 \ldots u_n) \in T}{}' \Pr(U_1 \ldots U_n = u_1 \ldots u_n)$$

$$\le \sum_{(u_1 \ldots u_n) \in T} 2^{-nH(U)(1-\varepsilon)} = |T(n,p,\varepsilon)| \cdot 2^{-nH(U)(1-\varepsilon)}$$

$$\Rightarrow |T(n,p,\varepsilon)| > (1-\varepsilon)\, 2^{nH(U)(1-\varepsilon)}.$$

Summary: $T(n, p, \varepsilon)$ has the following properties:

① $(u_1 \ldots u_n) \in T(n,p,\varepsilon)$ then

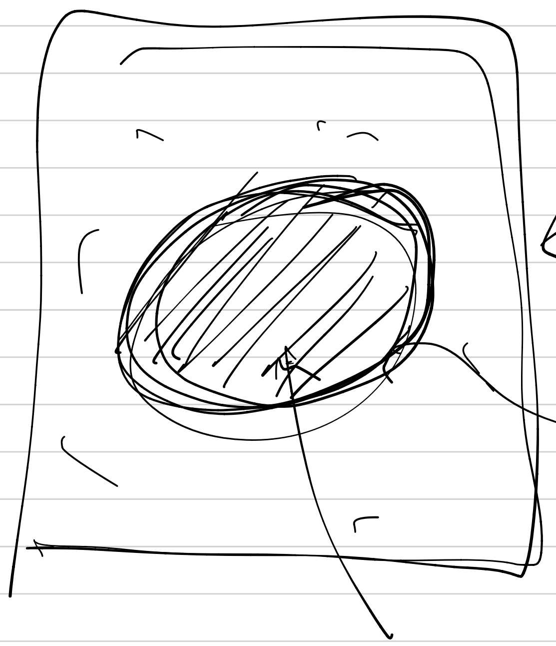
$$\Pr_{\substack{\text{iid} \sim p}}(U_1 \ldots U_n = u_1 \ldots u_n) = 2^{-n H(u)(1 \pm \varepsilon)}$$

② $|T(n,p,\varepsilon)| \leq 2^{n H(u)(1+\varepsilon)}$ $\sim p.$

③ for large n

$$|T(n,p,\varepsilon)| > (1-\varepsilon)\, 2^{nH(u)(1-\varepsilon)}$$

[Asymptotic equipartition Property].



$|\mathcal{U}^n| = 2^{n \log_2 |\mathcal{U}|}$

$|T(n,\varepsilon,p)| \approx 2^{nH(u)}$

each element has the $\simeq$ same probability of being produced by the iid $\sim p$ source.

So an "almost correct" view of the iid source is a probabilistic device that picks uniformly at random an element from $\mathbb{T}$.

Conceptual way for compressing iid sources:

- give each element of $T(n, \varepsilon, p)$ a binary representation. Since $|T| \le 2^{nH(u)(1+\varepsilon)}$, $\lceil n(1+\varepsilon)H(u) \rceil$ bits is enough.

- when the source produces $U_1 \dots U_n$ check if $(U_1 \dots U_n) \in T$, and if so emit the binary representation prefixed by 0

else ~~end universe~~

emit 1 followed by $\lceil n \log |\mathcal{U}| \rceil$ bit representation of $(U_1 \dots U_n)$.

So we describe n letters by

$$\le \begin{cases} \underline{nH(u)(1+\varepsilon) + 1 + 1} & \text{if } U_1 \dots U_n \in T \\ \underline{n \log |\mathcal{U}| + 2} & \text{else} \end{cases}$$

bits

In bits/letter we have

$$\leq \begin{cases} \boxed{H(u)(1+\varepsilon) + \tfrac{2}{n}} & u_1 \dots u_n \in T \\ \log|U| + \tfrac{2}{n} & \notin T \end{cases}$$

This gives a way to represent an iid source with $\approx H(u)$ bits/letter almost all the time