Updated 9/17/2024

Part 1: Theory L04: Compression (Algorithmic Derivation of Entropy via Compression)

Javed Aslam, Wolfgang Gatterbauer

cs7840 Foundations and Applications of Information Theory (fa24)

https://northeastern-datalab.github.io/cs7840/fa24/

9/16/2024

Last time

- Expectation
- Variance
- Markov Chains

- Intuitive Derivation of Entropy

Hartley ⇒ Shannon

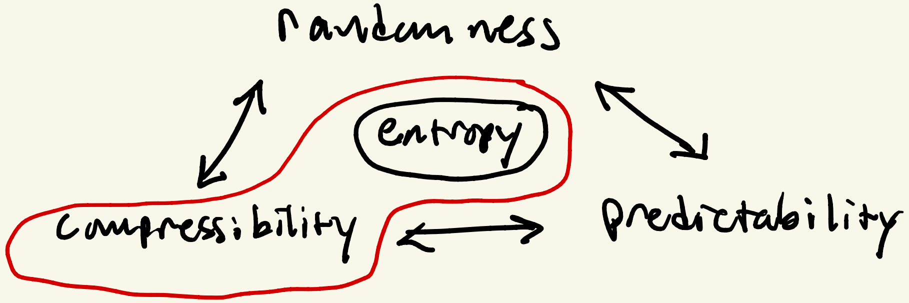
$$H(X) = H(\vec{p}) = \sum_i p_i \lg 1/p_i$$

Today

- Algorithmic Derivation of Entropy via Compression

Next time

- Fundamental concepts in Information Theory



Today: Motivate Entropy via Compression

- Consider codes with codewords of length $l_1, l_2, l_3, \ldots$
- Kraft's Inequality: $\sum_i 2^{-l_i} \leq 1$ or generally $\sum_i D^{-l_i} \leq 1$

(binary codes) (D-ary codes)

$$\begin{tikzcd}
\text{instantaneous (prefix-free) code} \arrow[r, Leftrightarrow, "①"] & \text{Kraft's Inequality} \arrow[d, Rightarrow, "②"'] & \text{uniquely decodable codes} \arrow[l, Rightarrow, "③"'] \\
 & l_i^* = \lg \frac{1}{p_i} &
\end{tikzcd}$$

Note:

$\lg \triangleq \log_2$

$$E[L] = \sum_i p_i \cdot l_i \geq \sum_i p_i \cdot l_i^* = \sum_i p_i \lg \frac{1}{p_i} = H(X)$$

Compression setup:

- A source code C for r.v. X is a mapping from $\mathcal{X}$, the range of X, to $\mathcal{D}^*$, the set of strings over encoding alphabet $\mathcal{D}$.
- The expected length $L(C) = \sum_{x \in \mathcal{X}} p(x) \cdot \ell(x)$
- A code is non-singular if every element of $\mathcal{X}$ maps to a unique string in $\mathcal{D}^*$, i.e.,

$$x_i \neq x_j \Rightarrow C(x_i) \neq C(x_j)$$

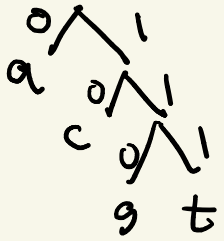
- The extension C^* of code C is a mapping from finite length strings from $\mathcal{X}$ to finite length strings from $\mathcal{D}$.
- A code is uniquely decodable if its extension is non-singular

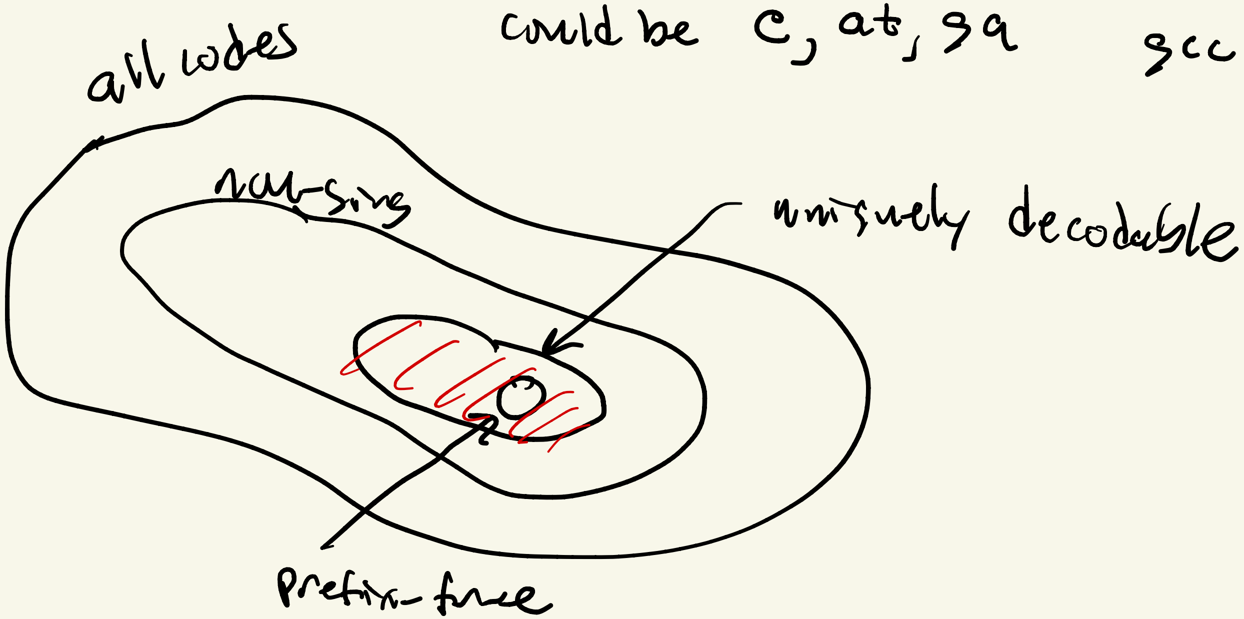
Classes of codes

X	Singular	Non-singular but not uniquely decodable	uniquely decodable but not instantaneous	instantaneous (prefix-free)
a	0	0	10	0
c	0	010	00	10
g	0	01	11	110
t	0	10	110	111

↓ e.g. 010 could be c, at, ga

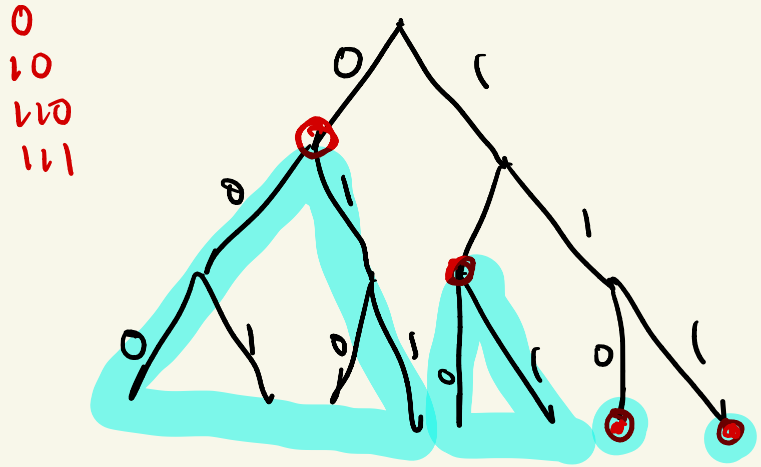
↓ 110000 gcc





Claim: Instantaneous (prefix-free) code $\Longleftrightarrow$ Kraft's Inequality

Pf ($\Rightarrow$): Let l_{max} be longest code word



prefix-free property:
internal node code makes unavaible all possible codes in subtree below.

- code of length l_i wipes out how many leaves? $2^{l_{max}-l_i}$

- tree only has $2^{l_{max}}$ leaves

$$\sum_i 2^{l_{max}-l_i} \le 2^{l_{max}}$$

→ dividing both sides by $2^{l_{max}}$

$\Rightarrow$ $\sum_i 2^{-l_i} \le 1$

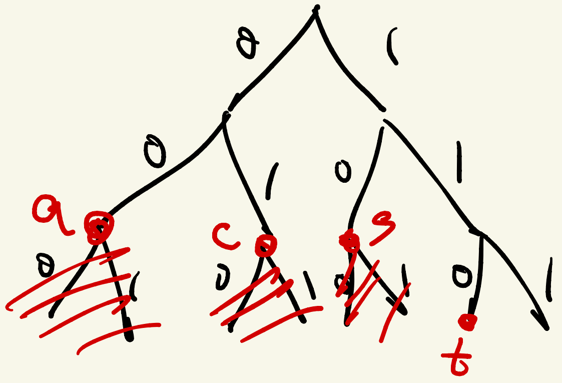
($\Leftarrow$)

		l_i
a	10	2
c	00	2
g	11	2
t	110	3

- Sort by length

  $l_1 \le l_2 \le \ldots \le l_n$

- assign source symbol associated w/ l_1 to first code lexicographically available of length l_1
- remove all children as possible codes
- repeat for $l_2, l_3, \ldots, l_n$



a	00
c	01
g	10
t	110

Proof sketch:

- subtrees assigned contiguously left-to-right
- If code lengths satisfy Kraft's inequality, never run out of subtrees

Setup: Find l_i where $\min L(C) = \sum_i p_i \cdot l_i$

s.t. $\sum_i 2^{-l_i} \le 1$

① $f(x) = x^2$ $\quad \min_x f(x)$?

univariate optimization

$$\frac{df}{dx} = 2x = 0 \Rightarrow x = 0 \qquad f(0) = 0 \ \checkmark$$

② $f(x,y) = x^2 + y^2$ $\quad \min_{x,y} f(x,y)$?

multi variate optimization

$$\frac{\partial f}{\partial x} = 2x = 0 \qquad x = 0$$

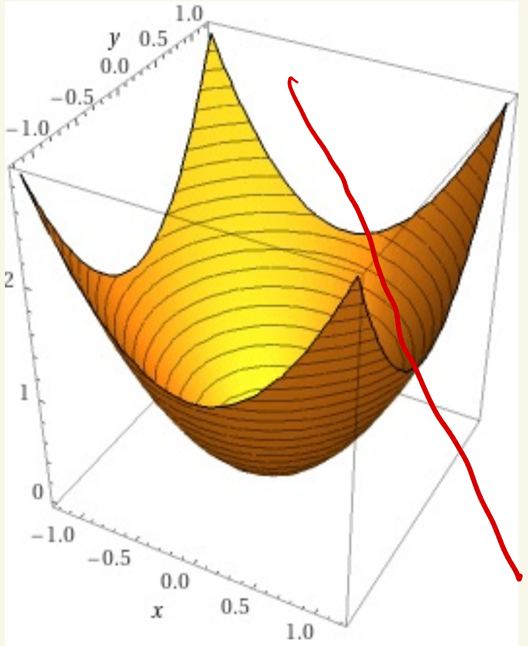
$$\frac{\partial f}{\partial y} = 2y = 0 \qquad y = 0$$

$$f(0,0) = 0$$

$$\min \; f(x,y) = x^2+y^2$$

$$\text{s.t} \quad \boxed{x+y=1}$$

Constrained optimization via Lagrange multipliers



$$J(x,y,\lambda) = x^2+y^2+\lambda(x+y-1)$$

$$\frac{\partial J}{\partial x} = 2x+\lambda = 0$$

$$\frac{\partial J}{\partial y} = 2y+\lambda = 0$$

subtract

$$2x-2y=0$$

$$x-y=0$$

$$\frac{\partial J}{\partial \lambda} = x+y-1 = 0$$

$$x+y=1$$

add

$$2x=1$$

$$x=1/2$$

$$y=1/2$$

✓

Find l_i where min $L(C) = \sum_i p_i \cdot l_i$

s.t. $\sum_i 2^{-l_i} \le 1$

$$J(\vec{l}, \lambda) = \sum_i p_i l_i + \lambda\left(\sum_i 2^{-l_i} - 1\right)$$

$$2^{-l_i} = e^{-l_i \cdot \ln 2}$$

$$\forall_i \quad \frac{\partial J}{\partial l_i} = p_i + \lambda \cdot 2^{-l_i} \cdot (-\ln 2) = 0$$

$$\frac{\partial J}{\partial \lambda} = \sum_i 2^{-l_i} - 1 = 0 \Rightarrow \sum_i 2^{-l_i} = 1$$

$$\sum_i \left(p_i + \lambda \cdot 2^{-l_i} (-\ln 2)\right) = 0$$

$$\Rightarrow \sum_i p_i - (\ln 2) \cdot \lambda \sum_i 2^{-l_i} = 0$$

$$\sum_i p_i - (\ln 2) \cdot \lambda \cdot 1 = 0$$

$$1 - (\ln 2) \cdot \lambda = 0 \Rightarrow \boxed{\lambda = 1/\ln 2}$$

$$p_i + \frac{1}{\ln 2} 2^{-l_i} (-\ln 2) = 0$$

$$p_i - 2^{-l_i} = 0$$

$$2^{-l_i} = p_i$$

$$\Rightarrow \boxed{l_i = \lg 1/p_i}$$

So what is $\min L(C) = \sum_i p_i l_i$?

$$\sum_i p_i \cdot l_i^* = \sum_i p_i \cdot \lg 1/p_i = H(X)$$

Updated 9/24/2024

Part 1: Theory
L06: Compression
(uniquely decodable codes)

Javed Aslam, Wolfgang Gatterbauer

cs7840 Foundations and Applications of Information Theory (fa24)

https://northeastern-datalab.github.io/cs7840/fa24/

9/23/2024

Last time

- Basic results in Information Theory

Today

- Finish basic results
- Continue compression

Next time

- Continue compression

---

instantaneous (prefix-free) code $\overset{①}{\Longleftrightarrow}$ Kraft's Inequality $\overset{③}{\Longleftarrow}$ uniquely decodable codes

② $\Downarrow$

$$l_i^* = \lg \frac{1}{p_i}$$

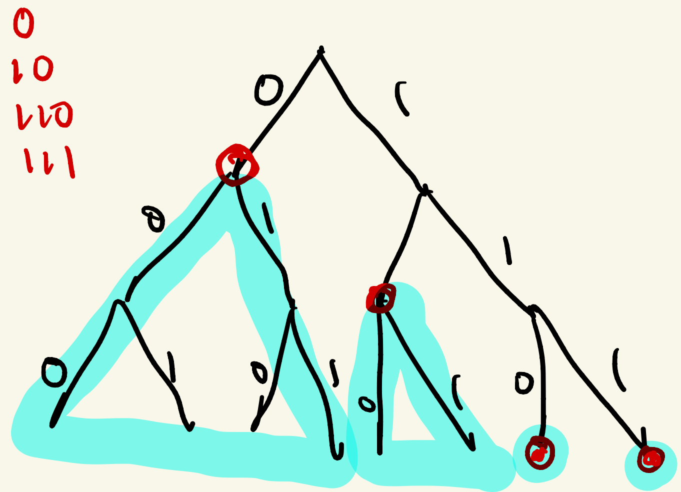
$$E[L] = \sum_i p_i \cdot l_i \ge \sum_i p_i \cdot l_i^* = \sum_i p_i \lg \frac{1}{p_i} = H(X)$$

Note:

$\lg \triangleq \log_2$

Claim: Instantaneous (prefix-free) code ⟺ Kraft's Inequality

Pf (⇒): Let l_{max} be longest code word



- code of length l_i wipes out how many leaves? $2^{l_{max}-l_i}$
- tree only has $2^{l_{max}}$ leaves

$$\sum_i 2^{l_{max}-l_i} \le 2^{l_{max}}$$

prefix-free property:
internal node code makes unavailable all possible codes in sub tree below.

→ dividing both sides by $2^{l_{max}}$

$$\Rightarrow \boxed{\sum_i 2^{-l_i} \le 1}$$

Recap for ①: counting argument

- What is $\left(\sum_{x\in\mathcal{X}} D^{-\ell(x)}\right)^k$?

Now 3

- Consider $\mathcal{X}=\{a,c,g,t\}$
  $\mathcal{D}=\{0,1\}$ and $k=2$

a → 0
c → 10
g → 110
t → 111

$x_1\in\mathcal{X}$ \ $x_2\in\mathcal{X}$	a	c	g	t
a	$D^{-\ell(a)}\cdot D^{-\ell(a)}$	$D^{-\ell(a)}\cdot D^{-\ell(c)}$	$\cdots$	
c				
g				
t				

- $\left(\sum_{x\in\mathcal{X}} D^{-\ell(x)}\right)^2$

$$= \sum_{x_1\in\mathcal{X}}\ \sum_{x_2\in\mathcal{X}} D^{-\ell(x_1)}\cdot D^{-\ell(x_2)}$$

$$= \sum_{x_1x_2\in\mathcal{X}^2} D^{-\ell(x_1)}\cdot D^{-\ell(x_2)} = \sum_{x_1x_2\in\mathcal{X}^2} D^{-(\ell(x_1)+\ell(x_2))}$$

$\mathcal{X}^2=\mathcal{X}\times\mathcal{X}$

$$= \sum_{x^2\in\mathcal{X}^2} D^{-\ell(x^2)}$$

$x^2\in\mathcal{X}^2=\mathcal{X}\times\mathcal{X}$

$\ell(aa)=2$
$\ell(ac)=3$
$\vdots$
$\ell(tt)=6$

$$= \sum_{m=1}^{2\cdot\ell_{max}} a(m)\cdot D^{-m} \le \sum_{m=1}^{2\cdot\ell_{max}} D^{m}\cdot D^{-m} = 2\cdot\ell_{max}$$

$a(m) = \#\ x^2\in\mathcal{X}^2$ where $\ell(x^2)=m$

unique decodability ⇒ $a(m)\le D^m$!

In general, $\left(\sum_{x\in\mathcal{X}} D^{-\ell(x)}\right)^k \le k\cdot\ell_{max}$

$$\left(\sum_{x \in \mathcal{X}} D^{-\ell(x)} \right)^k \leq k \cdot \ell_{max}$$

$$\Rightarrow \quad \sum_{x \in \mathcal{X}} D^{-\ell(x)} \leq \left(k \cdot \ell_{max} \right)^{1/k}$$

must hold for all k, e.g., $k \to \infty$

$$= e^{\ln\left[\left(k \cdot \ell_{max}\right)^{1/k}\right]}$$

$$= e^{\frac{\ln(k \cdot \ell_{max})}{k}}$$

$$\lim_{k \to \infty} = ? \quad \Rightarrow \quad e^{0} = 1$$

$$\Rightarrow \quad \sum_{x \in \mathcal{X}} D^{-\ell(x)} \leq 1$$

Updated 9/27/2024

Part 1: Theory
L07: Compression
(uniquely decodable codes continued)

Javed Aslam, Wolfgang Gatterbauer

cs7840 Foundations and Applications of Information Theory (fa24)

https://northeastern-datalab.github.io/cs7840/fa24/

9/25/2024

Last time

- Basic results in Information theory
- Kraft's Inequality for uniquely decodable codes

Today

- Implications of Kraft's Inequality
- Shannon codes
- Block coding
- Asymptotic Equipartion Property (AEP)

Next time

- Continue...

$$\begin{tikzcd}
\text{instantaneous (prefix-free) code} \arrow[r, Leftrightarrow, "1"] & \text{Kraft's Inequality} \arrow[d, Rightarrow, "2"] & \text{uniquely decodable codes} \arrow[l, Rightarrow, "3"'] \\
& l_i^* = \lg \frac{1}{p_i} &
\end{tikzcd}$$

Note: $\lg \triangleq \log_2$

$$E[L] = \sum_i p_i \cdot l_i \geq \sum_i p_i \cdot l_i^* = \sum_i p_i \lg \frac{1}{p_i} = H(X)$$

Bounds on n! : Stirling's Approximation (Mathematical Digression)

$n! = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1$

Approximations:

good: $(n/e)^n$

better: $(n/e)^n \sqrt{2\pi n}$

even better: $(n/e)^n \sqrt{2\pi n}\,(1+O(1/n))$

best: $(n/e)^n \sqrt{2\pi n}\; e^{\lambda_n}$

where $\frac{1}{12n+1} < \lambda_n < \frac{1}{12n}$

Example

		lg
$10!$	$= 3,628,800$	21.79
$\left(\frac{n}{e}\right)^n$	$= 453,999$	18.79
$\left(\frac{n}{e}\right)^n \sqrt{2\pi n}$	$= 3,598,696$	21.78
$\left(\frac{n}{e}\right)^n \sqrt{2\pi n}\, e^{\frac{1}{12n}}$	$= 3,628,810$	21.79

Kraft's Inequality: $\sum_i 2^{-l_i} \leq 1$

- consider $r_i = \frac{2^{-l_i}}{\sum_j 2^{-l_j}}$ - this is a distribution

- $L - H(X) = \sum_i p_i \cdot l_i - \left(-\sum_i p_i \lg p_i\right)$

$$= \sum_i p_i \cdot \lg \frac{1}{2^{-l_i}} + \sum_i p_i \lg p_i$$

$$= \sum_i p_i \cdot \lg \frac{1}{r_i \cdot \sum_j 2^{-l_j}} + \sum_i p_i \lg p_i$$

$$= \sum_i p_i \lg \frac{1}{r_i} + \sum_i p_i \cdot \lg \frac{1}{\sum_j 2^{-l_j}} + \sum_i p_i \lg p_i$$

$$= \sum_i p_i \lg \frac{p_i}{r_i} + \lg \frac{1}{\sum_j 2^{-l_j}}$$

$$= \underbrace{D(\vec{p} \,\|\, \vec{r})}_{\geq 0} + \underbrace{\lg \frac{1}{\sum_j 2^{-l_j}}}_{\geq 0}$$

$= 0$ iff

$2^{-l_i} = p_i$ and $\sum_j 2^{-l_j} = 1$

$\Leftrightarrow l_i = \lg \frac{1}{p_i}$

To show: $H(x) \leq L^* \leq H(x)+1$ L^* is opt code

Shannon codes: $l_i = \lceil \lg \frac{1}{p_i} \rceil$

Claim: these l_i satisfy Kraft's inequality

Pf: $\sum_i 2^{-\lceil \lg \frac{1}{p_i} \rceil} \leq \sum_i 2^{-\lg \frac{1}{p_i}} = \sum_i 2^{\lg p_i} = \sum_i p_i = 1$

$\Rightarrow$ valid code lengths; can easily be turned into prefix-free code by earlier results

Now: $\lg \frac{1}{p_i} \leq l_i = \lceil \lg \frac{1}{p_i} \rceil < \lg \frac{1}{p_i} + 1$

$$\sum_i p_i \lg \frac{1}{p_i} \leq \sum_i p_i \cdot l_i < \sum_i p_i \left(\lg {}^1\!/_{p_i} + 1\right)$$

$$H(x) \leq L < H(x) + 1$$

Block Coding : encode blocks of length n at a time

- induced distribution $p(x_1, x_2 \dots x_n) = p(x_1) \cdot p(x_2) \dots p(x_n)$
- $L_n = \frac{1}{n} \sum p(x_1 \dots x_n) \cdot l(x_1 \dots x_n)$
- From last slide

$$H(X_1 \dots X_n) \leq E\, l(X_1 \dots X_n) < H(X_1 \dots X_n) + 1$$

$$\Longrightarrow \quad n\, H(X) \leq n \cdot L_n < n\, H(X) + 1$$

$$\Longrightarrow \quad H(X) \leq L_n < H(X) + 1/n$$

$\therefore$ Can drive inefficiency down arbitrarily small by using larger and larger blocks.

- Issue : Code book grows exponentially with block length

2 other codes ... Example:

$H(\vec{p}) = 2.23$

	A	B	C	D	E	
p_i	.35	.17	.17	.16	.15	
$\lg \frac{1}{p_i}$	1.51	2.56	2.56	2.64	2.74	
$\lceil \lg 1/p_i \rceil$	2	3	3	3	3	Shannon

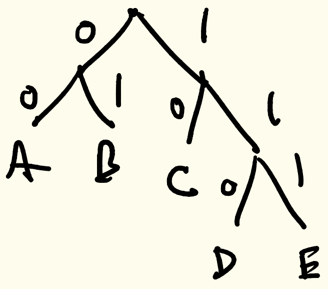
$L_S = 2.65$ bpc

(bpc - bits per character on average)

Fano:

A .35 | B .17 ‖ C .17 | D .16 | E .15

- sort prob.
- split as close to 50/50 as possible
- recurse



A	00
B	01
C	10
D	110
E	111

2.31 bpc

$L_F = 2.31$ bpc

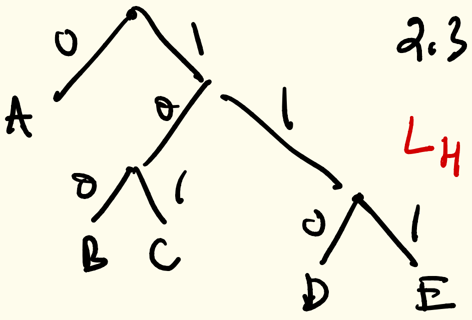
Huffman

- combine least probable events bottom-up, creating combined events along the way

A	B	C	{DE}
.35	.17	.17	.31

A	{BC}	{DE}
.35	.34	.31

A	0
B	100
C	101
D	110
E	111



2.3 bpc

$L_H = 2.3$ bpc

AEP : Asymptotic Equipartion Property

Consider biased coin $\left\{ \begin{array}{l} Pr(H) = 1/3 \\ Pr\{T\} = 2/3 \end{array} \right.$

Flip coin n times.

- What is most probable outcome? TTT--T $\quad Pr = (2/3)^n$
- typical sequence has about $1/3$ H $2/3$ T $\quad Pr = (1/3)^{\frac{n}{3}} \cdot (2/3)^{2n/3}$
- Let $Pr[\text{"typical sequence"}] = x = (1/3)^{n/3} \cdot (2/3)^{2n/3}$
- What is x?

$$\begin{aligned} \lg x &= \frac{n}{3} \lg(1/3) + \frac{2n}{3} \lg(2/3) \\ &= n \cdot \left[\frac{1}{3} \lg 1/3 + \frac{2}{3} \lg 2/3 \right] \\ &= -n \cdot H(x) \end{aligned}$$

$\Rightarrow$ $Pr[\text{"typical sequence"}] = x = 2^{-n H(x)}$

How many strings of $1/3$ H & $2/3$ T ?

$\binom{n}{n/3}$ typical sequences

$$\binom{n}{n/3} = \frac{n!}{\left(\frac{n}{3}\right)!\,\left(\frac{2n}{3}\right)!} \sim \frac{\left(n/e\right)^{n}}{\left(\frac{n/3}{e}\right)^{n/3}\cdot\left(\frac{2n/3}{e}\right)^{2n/3}} = \frac{\cancel{\left(n/e\right)^{n}}}{\cancel{\left(\frac{n}{e}\right)^{n/3}}\cdot\left(\frac{1}{3}\right)^{n/3}\cdot\cancel{\left(\frac{n}{e}\right)^{2n/3}}\cdot\left(\frac{2}{3}\right)^{2n/3}}$$

$$= \frac{1}{\left(1/3\right)^{n/3}\cdot\left(2/3\right)^{2n/3}} = 2^{nH(x)}$$

- There are about $2^{nH(X)}$ typical sequences
- Each has about $2^{-nH(x)}$ probability

$\Rightarrow$ almost everything one is likely to see is <u>typical</u> and <u>equally likely</u>

Updated 9/30/2024

Part 1: Theory
L08: Compression
(AEP = Asymptotic Equipartition Property)

Javed Aslam, Wolfgang Gatterbauer

cs7840 Foundations and Applications of Information Theory (fa24)

https://northeastern-datalab.github.io/cs7840/fa24/

9/30/2024

Last time

- Implications of Kraft's Inequality
- Shannon, Fano, Huffman Codes
- Block Coding
- Motivating the AEP

Today

- AEP
  - formal defs. & proofs
  - consequences

- Finish fundamentals

Next time

- Continue Fundamentals

(Weak) Law of Large Numbers

(Mathematical Digression)

Roughly: Sample mean converges to true mean (in probability)

Technically: Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$; $E[X] = \mu$

- $\overline{X}_n \xrightarrow{P} \mu$ as $n \to \infty$
- $\lim_{n\to\infty} \Pr(|\overline{X}_n - \mu| < \varepsilon) = 1$ for any $\varepsilon > 0$
- $(\forall \varepsilon > 0)(\forall \delta > 0)(\exists n_0)\ \Pr(|\overline{X}_n - \mu| < \varepsilon) > 1-\delta \quad \forall n \geq n_0$

Thm: (AEP) If $X_1, X_2, \dots X_n$ are i.i.d. $\sim p(x)$, then

$$-\frac{1}{n} \lg p(X_1 X_2 \dots X_n) \longrightarrow H(X) \text{ in probability.}$$

Intuition: Consider an actual sequence $x_1 x_2 \dots x_n$

$$-\frac{1}{n} \lg p(x_1 x_2 \dots x_n) = \underbrace{\frac{1}{n} \boxed{\lg \frac{1}{p(x_1 x_2 \dots x_n)}}}_{\text{avg. length per message}} \sim \text{length of Shannon code for block}$$

→ converges to $H(X)$ as $n \to \infty$ by block coding

Formal Proof: Since X_i are i.i.d., then so are r.v. $\lg p(X_i)$ & $\lg \frac{1}{p(X_i)}$

By LLN:
$$-\frac{1}{n} \lg p(X_1 X_2 \dots X_n) = -\frac{1}{n} \sum_i \lg p(X_i)$$
$$= \frac{1}{n} \sum_i \lg \frac{1}{p(X_i)}$$
$$\to E\left[\lg \frac{1}{p(X)}\right] \text{ in probability}$$
$$= H(X)$$

Informal Relationship

$$-\frac{1}{n} \lg p(X_1 X_2 \dots X_n) \sim H(X)$$

$$\therefore \iff p(X_1 X_2 \dots X_n) \sim 2^{-n H(X)}$$

Def: The *typical set* $A_\varepsilon^{(n)}$ with respect to $p(x)$ is the set of all sequences $(x_1 x_2 \dots x_n) \in \mathcal{X}^n$ such that

$$2^{-n(H(X)+\varepsilon)} \le p(x_1, x_2 \dots x_n) \le 2^{-n(H(X)-\varepsilon)}$$

← empirical probability of sequence

Thm: ① If $(x_1 x_2 \dots x_n) \in A_\varepsilon^{(n)}$, then $H(X) - \varepsilon \le -\frac{1}{n} \lg p(x_1 \dots x_n) \le H(X) + \varepsilon$

② $\Pr[A_\varepsilon^{(n)}] > 1 - \varepsilon$ for n sufficiently large

③ $|A_\varepsilon^{(n)}| \le 2^{n(H(X)+\varepsilon)}$

④ $|A_\varepsilon^{(n)}| \ge (1-\varepsilon) 2^{n(H(X)-\varepsilon)}$ for n sufficiently large

Pf: ① Immediate from definition of typical set

② By LLN $(\forall \varepsilon > 0)(\forall \delta > 0)(\exists n_0)$

$$\Pr\Big[\underbrace{\left|-\tfrac{1}{n} \lg p(x_1 x_2 \dots x_n) - H(X)\right| < \varepsilon}_{\text{typical set by ①}}\Big] > \underset{\substack{\uparrow \\ \text{choose } \delta = \varepsilon}}{1 - \delta} \quad \forall n > n_0$$

③ $|A_\varepsilon^{(n)}| \leq 2^{n(H(x)+\varepsilon)}$

Let $\vec{x} = (x_1, x_2 - x_n)$

④ $|A_\varepsilon^{(n)}| \geq (1-\varepsilon)\, 2^{n(H(x)-\varepsilon)}$ for n sufficiently large

Pf ③ :

$$
\begin{aligned}
1 &= \sum_{\vec{x} \in \mathcal{X}^n} p(\vec{x}) \\
&\geq \sum_{\vec{x} \in A_\varepsilon^{(n)}} p(\vec{x}) \\
&\geq \sum_{\vec{x} \in A_\varepsilon^{(n)}} 2^{-n(H(X)+\varepsilon)} \\
&= 2^{-n(H(x)+\varepsilon)} \cdot |A_\varepsilon^{(n)}|
\end{aligned}
$$

$$\Rightarrow \quad |A_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}$$

Pf ④ : For n sufficiently large, $\Pr[A_\varepsilon^{(n)}] > 1-\varepsilon$ so...

$$
\begin{aligned}
1-\varepsilon &< \Pr[A_\varepsilon^{(n)}] \\
&= \sum_{\vec{x} \in A_\varepsilon^{(n)}} p(\vec{x}) \\
&\leq \sum_{\vec{x} \in A_\varepsilon^{(n)}} 2^{-n(H(x)-\varepsilon)} \\
&= 2^{-n(H(x)-\varepsilon)} \cdot |A_\varepsilon^{(n)}|
\end{aligned}
$$

$$\Rightarrow \quad |A_\varepsilon^{(n)}| \geq (1-\varepsilon) \cdot 2^{n(H(x)-\varepsilon)}$$

Upshot:

- $2^{-n(H(X)+\varepsilon)} \leq p(x_1, x_2 \ldots x_n) \leq 2^{-n(H(X)-\varepsilon)}$

a typical sequence has empirical probability $\sim 2^{-nH(X)}$

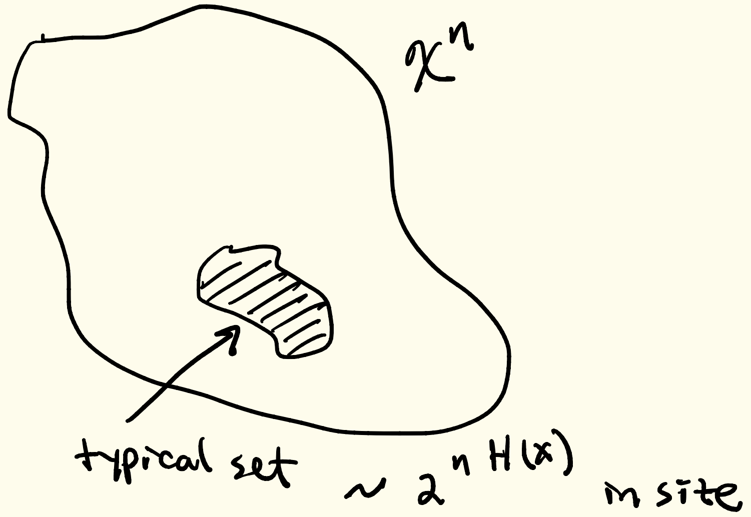
- $|A_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}$
- $|A_\varepsilon^{(n)}| \geq (1-\varepsilon)\, 2^{n(H(X)-\varepsilon)}$

there are $\sim 2^{nH(X)}$ typical sequences

- $\Pr[A_\varepsilon^{(n)}] > 1-\varepsilon$ for n sufficiently large

typical sequences contain almost all probability

Immediate Consequence for compression



$\Rightarrow$ high prob events are typical sequences which use $nH(x)+1$ bits or $H(x)+1/n$ per encoded message, on average.

Block coding compression method (roughly):

- If $\vec{x}$ is <u>typical</u>, start with a 0 and encode the exact typical sequence in straight binary using $\lg(2^{nH(x)}) = nH(x)$ additional bits
- If $\vec{x}$ is <u>not</u> <u>typical</u>, start with a 1 and encode the exact atypical sequence in straight binary using $\lg(|\mathcal{X}|^n) = n\lg|\mathcal{X}|$ additional bits

(see text for more careful treatment taking into acount ϵ, etc.